ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
AGRICULTURAL UNIVERSITY OF ATHENS

"ALEXANDER FLEMING"
Biomedical Sciences Research Center

ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ
ΤΜΗΜΑ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΒΙΟΛΟΓΙΑ ΣΥΣΤΗΜΑΤΩΝ"

Μεταπτυχιακή Διπλωματική Εργασία

Μέθοδοι λειτουργικής ανάλυσης στη Βιολογία Συστημάτων

Φωτεινή Π. Θανάτη
Βιολόγος, ΕΚΠΑ

Επιβλέπων καθηγητής:
Πολυδεύκης Χατζόπουλος, Καθηγητής ΓΠΑ

ΑΘΗΝΑ
2021

ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ
ΤΜΗΜΑ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

Μεταπτυχιακή Διπλωματική Εργασία

Μέθοδοι λειτουργικής ανάλυσης στη Βιολογία Συστημάτων

"Functional enrichment analysis methods in Systems Biology"

Φωτεινή Π. Θανάτη
Βιολόγος, ΕΚΠΑ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:
Πολυδεύκης Χατζόπουλος, Καθηγητής ΓΠΑ
Γεώργιος Παυλόπουλος, Ερευνητής Β', ΕΚΕΒΕ "Αλέξανδρος Φλέμινγκ''
Δημήτριος Στραβοπόδης, Αναπληρωτής Καθηγητής Βιολογίας Κυττάρου & Ανάπτυξης, ΕΚΠΑ

Μέθοδοι λειτουργικής ανάλυσης στη Βιολογία Συστημάτων

ΠΜΣ Βιολογία Συστημάτων
Τμήμα Βιοτεχνολογίας
Bioinformatics and Integrative Biology Lab, ΕΚΕΒΕ "Αλέξανδρος Φλέμινγκ"

ΠΕΡΙΛΗΨΗ

Η λειτουργική ανάλυση εμπλουτισμού της γονιδιακής έκφρασης είναι μια ευρέως χρησιμοποιούμενη μέθοδος για την ερμηνεία των πειραματικών αποτελεσμάτων, η οποία αποσκοπεί στον προσδιορισμό σημαντικά στατιστικών ομαδοποιήσεων πρωτεϊνών/γονιδίων σε κατηγορίες που σχετίζονται με ορισμένες βιολογικές διεργασίες, μονοπάτια, ασθένειες ή φαινότυπους. Παρά το γεγονός ότι σήμερα υπάρχει ένας μεγάλος αριθμός βιοπληροφορικών εργαλείων για λειτουργική ανάλυση, τα περισσότερα από αυτά μπορούν να επεξεργαστούν μία λίστα κάθε φορά, καθιστώντας έτσι μια πιο συνδυαστική ανάλυση περίπλοκη και επιρρεπή σε λάθη. Στην παρούσα πτυχιακή εργασία, παρουσιάζεται το FLAME, ένα διαδικτυακό εργαλείο για συνδυαστικές αναλύσεις δεδομένων, καθώς επιτρέπει τον συνδυασμό πολλών λιστών πριν από την ανάλυση εμπλουτισμού. Οι χρήστες μπορούν ως αρχεία εισόδου να χρησιμοποιήσουν αρκετές λίστες και να τις επεξεργαστούν με την χρήση διαδραστικών γραφημάτων UpSet, ως αποτελεσματικότερη εναλλακτική λύση έναντι των διαγραμμάτων Venn, με σκοπό την εύρεση τομών/ενώσεων κλπ μεταξύ όλων των επιθυμητών συνδυασμών των αρχείων. Ο εμπλουτισμός λειτουργικότητας και βιβλιογραφίας, καθώς και ενσωματωμένα εργαλεία για μετατροπές αναγνωριστικών των γονιδίων και εύρεση ορθόλογων γονιδίων προσφέρονται από τις εφαρμογές g:Profiler και aGOtool για 197 οργανισμούς. Στην παρούσα έκδοση, το FLAME μπορεί να αναλύει γονίδια/πρωτεΐνες και να τα εντάσσει σε γονιδιακές οντολογίες, μονοπάτια, ρυθμιστικά μοτίβα, λειτουργικές επικράτειες, ασθένειες, φαινότυπους και βιβλιογραφικές αναφορές, ενώ μπορεί επίσης να δημιουργήσει δίκτυα πρωτεϊνικών αλληλεπιδράσεων που προέρχονται από την STRING. Για τον έλεγχο της λειτουργικότητας του FLAME, πραγματοποιήθηκε μελέτη δεδομένων γονιδιακής έκφρασης που σχετίζονται με την ευαισθησία του περιφερικού τμήματος του παχέος εντέρου στον πειραματικό καρκίνο του παχέος εντέρου (Καρκίνος σε έδαφος κολίτιδας). Η εφαρμογή FLAME συνιστά μια διαδραστική φιλική προς το χρήστη πλατφόρμα που επιτρέπει τον εύκολο χειρισμό δεδομένων και αποτελεσμάτων, τα οποία μπορούν να απεικονιστούν ως διαδραστικοί και παραμετροποιήσιμοι πίνακες με τα αντίστοιχα διαγράμματα heatmaps, barcharts, διαγράμματα Manhattan και δίκτυα.
Διαθεσιμότητα: Η εφαρμογή FLAME βρίσκεται online: http://flame.pavlopouloslab.info
Κώδικας: https://github.com/PavlopoulosLab/FLAME

Επιστημονική περιοχή: Βιοπληροφορική

Λέξεις Κλειδιά: Λειτουργικός Εμπλουτισμός Γονιδίων/Πρωτεϊνών, Εμπλουτισμός βιβλιογραφίας, Δίκτυα, Διαδραστικές Απεικονίσεις, Διαγράμματα Upset

Functional enrichment analysis methods in Systems Biology

MSc Systems Biology
Department of Biotechnology
Bioinformatics and Integrative Biology Lab, BSRC Al. Fleming

ABSTRACT

Functional enrichment is a widely used method for interpreting experimental results by identifying classes of proteins/genes associated with certain biological functions, pathways, diseases or phenotypes. Despite the variety of existing tools, most of them can process a single list per time, thus making a more combinatorial analysis more complicated and prone to errors. In this thesis, we present FLAME, a web tool for combining multiple lists prior to enrichment analysis. Users can upload several lists of preference and use interactive UpSet plots, as an alternative to Venn diagrams, to handle unions or intersections among the given input files. Functional and literature enrichment along with gene conversions are offered by g:Profiler and aGOtool applications for 197 organisms. In its current version, FLAME can analyze genes/proteins for related articles, Gene Ontologies, pathways, annotations, regulatory motifs, domains, diseases, phenotypes while it can also generate protein-protein interactions derived from STRING. We have herein validated FLAME by interrogating gene expression data associated with the sensitivity of the distal part of the large intestine to experimental colitis-propelled colon cancer. The FLAME application comes with an interactive user-friendly interface that allows easy list manipulation and exploration, while results can be visualized as interactive and parameterizable heatmaps, barcharts, Manhattan plots, networks and tables.
Availability: FLAME application: http://flame.pavlopouloslab.info
Code: https://github.com/PavlopoulosLab/FLAME

Scientific area: Bioinformatics

Keywords: Functional Enrichment, Literature Enrichment, Network, Interactive Visualizations, Upset Plots

ΔΗΛΩΣΗ ΕΡΓΟΥ

Η κάτωθι υπογεγραμμένη, Θανάτη Φωτεινή δηλώνω ότι το κείμενο της μελέτης αποτελεί δικό μου, μη υποβοηθούμενο πόνημα. Υποβάλλεται σε μερική εκπλήρωση των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Βιολογία Συστημάτων του Γεωπονικού Πανεπιστημίου Αθηνών. Δεν έχει υποβληθεί ποτέ πριν για οποιοδήποτε λόγο ή για εξέταση σε οποιοδήποτε άλλο πανεπιστήμιο ή εκπαιδευτικό ίδρυμα της χώρας ή του εξωτερικού.

Θανάτη Π. Φωτεινή
Αθήνα, 31/10/2021

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στο Εργαστήριο “Bioinformatics and Integrative Biology Lab” στο Ερευνητικό Κέντρο Βιοϊατρικών Επιστημών "Αλέξανδρος Φλέμιγκ" / ΕΚΕΒΕ Α. Φλέμιγκ, κατά το χρονικό διάστημα Φεβρουαρίου 2021 - Σεπτεμβρίου 2021 υπό την επίβλεψη του Δρ. Γ. Παυλόπουλου, Ερευνητή Β΄.

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον Δρ. Γ. Παυλόπουλο, για την ανάθεση ενός τόσο ενδιαφέροντος και πολυδιάστατου θέματος, για την εμπιστοσύνη που μου έδειξε καθώς και για την επιστημονική υποστήριξή του κατά την εκπόνηση της Εργασίας. Οι ευκαιρίες που μου έδωσε σε κάθε επίπεδο ήταν πολύ σημαντικές μεταδίδοντας μου ταυτόχρονα έμπνευση, ενέργεια και θέληση για δημιουργικότητα και πολυεπίπεδη σκέψη.

Επιπλέον, η συμβολή του μεταδιδακτορικού Ερευνητή Δρ. Ευάγγελου Καρατζά, ως υπεύθυνο μου στο εργαστήριο, ήταν καταλυτική και χωρίς τη δική του καθοδήγηση, η παρούσα εργασία θα ήταν ημιτελής και ελλιπής. Για αυτό, θα ήθελα να τον ευχαριστήσω ιδιαίτερα για την υπομονή του, την εμπιστοσύνη, τη βοήθειά του και την άριστη συνεργασία που είχαμε όλο αυτό το διάστημα.

Δεν θα μπορούσα να μην εκφράσω τις ευχαριστίες μου στους Δρ. Δ. Στραβοπόδη και Δρ. Π. Χατζόπουλο, μέλη της τριμελούς μου επιτροπής, οι οποίοι δέχτηκαν να είναι στην επιτροπή μου, με εμπιστοσύνη στο πρόσωπό μου και βοήθεια σε κάθε προβληματισμό και σκέψη μου. Ο Δρ. Δ. Στραβοπόδης ήταν δίπλα μου σε κάθε επιστημονικό βήμα από το προπτυχιακό επίπεδο ως επιβλέπων της πτυχιακής μου και τον ευχαριστώ θερμα γι αυτό.

Θα ήθελα να ευχαριστήσω τα υπόλοιπα μέλη του εργαστηρίου και ιδιαίτερα τον μεταδιδακτορικό Ερευνητή Δρ. Φώτη Μπαλτουμά για την άριστη επιστημονική συνεργασία μας, την υποστήριξη, την αλληλοβοήθεια, την ομαδικότητα και το ευχάριστο κλίμα που επικρατούσε μεταξύ μας. Τέλος, ένα μεγάλο ευχαριστώ στην οικογένεια μου και τους φίλους μου, για τη στήριξή τους και την κατανόησή τους καθ’ όλο το διάστημα από την έναρξη της διπλωματικής εργασίας μέχρι και την ολοκλήρωσή της.

Όταν όλοι αυτοί οι άνθρωποι πιστεύουν σε σένα, μόνο δύναμη και θέληση να προχωράς και να εξελίσσεσαι έχεις.

Σας ευχαριστώ!!!

- Αντίο, είπε η αλεπού. Να το μυστικό μου. Είναι πολύ απλό:
μόνο με την καρδιά βλέπεις καλά. Την ουσία τα μάτια δεν τη βλέπουν.
- Την ουσία τα μάτια δεν τη βλέπουν, επανέλαβε ο μικρός πρίγκιπας για να το θυμάται.
- Είναι ο χρόνος που ξόδεψες για το τριαντάφυλλό σου που το κάνει τόσο σημαντικό.
- Είναι ο χρόνος που ξόδεψα για το τριαντάφυλλό μου… είπε ο μικρός πρίγκιπας για να το θυμάται.. "Αντουάν ντε Σαιντ-Εξυπερύ"

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1. ΘΕΩΡΗΤΙΚΗ ΕΙΣΑΓΩΓΗ

1.1 Βιολογικά δεδομένα και βάσεις

Η σύγχρονη, ραγδαία τεχνολογική πρόοδος σε τεχνικές υψηλής απόδοσης (high-throughput) έχει επιφέρει ως αποτέλεσμα την παραγωγή πληθώρας βιολογικών δεδομένων και των μεταξύ τους βιομοριακών αλληλεπιδράσεων. Για την εκμετάλλευση αυτού του όγκου πληροφορίας έχουν υλοποιηθεί διάφορες υπολογιστικές τεχνικές, διαδικτυακά εργαλεία και βάσεις δεδομένων με σκοπό την ανάλυση, οργάνωση και οπτικοποίηση των δεδομένων αυτών.

Τα βιολογικά συστήματα συντίθενται από διάφορες μοριακές οντότητες που αλληλεπιδρούν μεταξύ τους, όπως γονίδια, πρωτεΐνες, μεταβολίτες και άλλα στοιχεία τα οποία είναι απαραίτητα για τις κυτταρικές βιολογικές διεργασίες. Η ανάλυση και η ερμηνεία των αλληλεπιδράσεων των μοριακών οντοτήτων (βιομορίων) αποτελούν κομβικό σημείο για την κατανόηση της φυσιολογίας και της παθολογίας όλων των οργανισμών, τόσο σε μικροσκοπική κλίμακα, δηλαδή σε επίπεδο οργανισμού όσο και σε μακροσκοπική, δηλαδή έχοντας συμπεριλάβει όλες τις σχέσεις και τις αλληλεπιδράσεις μεταξύ των οργανισμών που συμβιώνουν σε ένα ενδιαίτημα/οικοσύστημα.

Οι βιολογικές βάσεις δεδομένων συνιστούν έναν κεντρικό πυλώνα της σύγχρονης Βιοπληροφορικής, καθώς αποτελούν τη κύρια πηγή άντλησης πληροφορίας και δεδομένων. Σήμερα οι βιολογικές βάσεις δεδομένων αναφέρονται σε ηλεκτρονικές βιβλιοθήκες-αποθετήρια που περιέχουν βιολογικά δεδομένα από διάφορες πηγές (π.χ. πειραματικά εργαστήρια, δημοσιευμένη βιβλιογραφία). Οι βιολογικές βάσεις δεδομένων, γενικά, μπορούν να διακριθούν σε επιμέρους κατηγοριοποιήσεις [1], [2] όπως περιγράφονται παρακάτω. Καταρχάς, υπάρχουν οι πρωτογενείς βάσεις δεδομένων, οι οποίες περιέχουν τα πρωτογενή πειραματικά δεδομένα και οι δευτερογενείς βάσεις δεδομένων, στις οποίες υπάρχουν κυρίως ταξινομήσεις των πρωτογενών δεδομένων, χρήσιμες για αναλυτικούς σκοπούς και σε τριτογενής (βάσεις πρόβλεψης) [1], [3]**.** Επίσης μπορούν να κατηγοριοποιηθούν ανάλογα με τον τύπο των δεδομένων και των αλληλεπιδράσεων και οι οποίες αναλύονται κυρίως σε:

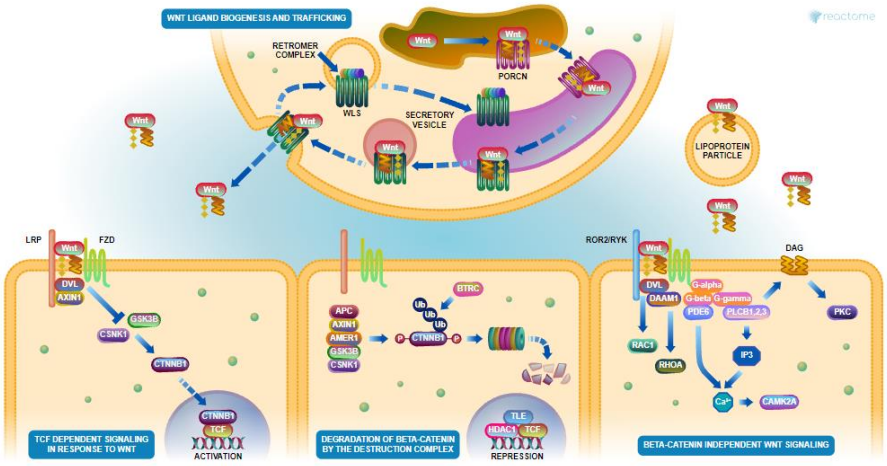
- Βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών
- Βάσεις δεδομένων αμινοξικών αλληλουχιών πρωτεϊνών
- Βάσεις δεδομένων τρισδιάστατων βιολογικών δομών
- Βάσεις δεδομένων γονιδιακής έκφρασης
- Βάσεις δεδομένων γενετικής ποικιλομορφίας
- Βάσεις δεδομένων βιβλιογραφίας
- Βάσεις δεδομένων οικογενειών (κυρίως πρωτεϊνών)
- Εξειδικευμένες βάσεις δεδομένων

Το τελευταίο κριτήριο για την ταξινόμηση των βάσεων δεδομένων είναι η πολιτική συγγραφής και επιμέλειας δεδομένων τους, δηλαδή με βάση το τρόπο συλλογής των πληροφοριών, τα επίπεδα λεπτομέρειας και την περιγραφή, σχολιασμό και ταξινόμηση αυτών των πληροφοριών. Η απόκτηση δεδομένων μπορεί να γίνει με μη αυτόματο τρόπο (δηλαδή, από επιμελητές ή από την επιστημονική κοινότητα), αυτοματοποιημένα (εκτελούνται με χρήση υπολογιστικών μεθόδων), ή

συνδυασμός των δύο προαναφερθέντων κατηγοριών [1], [2]. Στην συνέχεια περιγράφονται περιληπτικά ένα πλήθος σημαντικών βάσεων δεδομένων.

1.1.1 Reactome

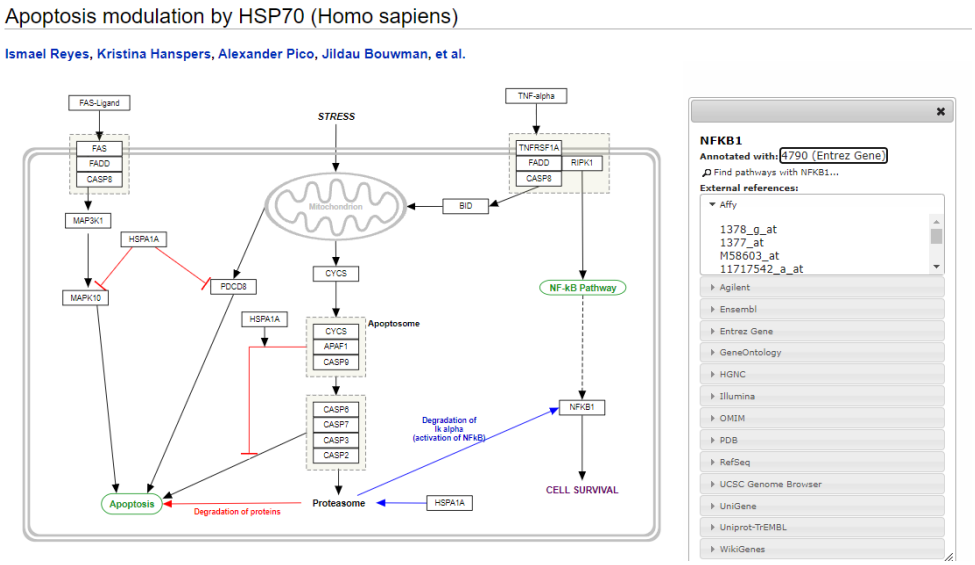
Η Reactome [4] είναι μια ανοιχτή βάση δεδομένων αντιδράσεων, βιολογικών μονοπατιών και διεργασιών, η οποία χρησιμοποιεί ένα απλουστευτικό μοντέλο για την αναπαράσταση της βιολογικής πληροφορίας. Στοχεύει στη παροχή εργαλείων βιοπληροφορικής για την απεικόνιση, ερμηνεία και ανάλυση της γνώσης των βιολογικών μονοπατιών, για τη στήριξη της βασικής και της κλινικής έρευνας, της ανάλυσης του γονιδιώματος, της μοντελοποίησης, της βιολογίας συστημάτων και της εκπαίδευσης. Περιέχει πληροφορίες από 33.453 βιβλιογραφικές αναφορές και κατά κύριο λόγο αποτελεί έναν εκτεταμένο μεταβολικό χάρτη του *H. sapiens*. Μονοπάτια και δεδομένα για άλλους οργανισμούς προέρχονται κυρίως από ορθόλογες μοριακές αντιδράσεις, όπου εφαρμόζονται. Το σκεπτικό πίσω από τη βάση είναι να παρουσιάζει τις βιολογικές διαδρομές για τις κυτταρικές διεργασίες σε μοριακό επίπεδο, οπτικοποιώντας τις σε μοντέλα δεδομένων (Εικόνα 1.1). Η βάση δεδομένων (έκδοση 76) περιέχει 10.867 ανθρώπινα γονίδια, 415 φάρμακα, 1856 μικρά μόρια που χρησιμεύουν ως φυσικά υποστρώματα, καταλύτες ή ρυθμιστές, 11.073 διακριτές πρωτεΐνες και 13.732 αντιδράσεις που ενσωματώνονται σε 2516 ανθρώπινες οδούς ομαδοποιημένες σε 26 υπερ-οδούς (δηλαδή ανοσοποιητικό σύστημα, μεταβολισμός, ασθένειες κλπ.) Οι πληροφορίες στη βάση δεδομένων συντάσσονται από εμπειρογνώμονες βιολόγους. Το περιεχόμενο συχνά παραπέμπει σε άλλες βάσεις, π.χ. NCBI, Ensembl, UniProt, KEGG, ChEBI [5], PubMed και GO, γεγονός που επιτρέπει τη διαδραστική αναζήτηση μονοπατιών-γονιδίων, ενώ συνδυάζει και εξωτερικά δεδομένα όπως δεδομένα έκφρασης. Το περιεχόμενο της βάσης είναι προσβάσιμο, είτε μέσω προγράμματος περιήγησης, είτε μέσω API, καθώς και μέσω εφαρμογής Cytoscape (ReactomeFIViz) [6].



Εικόνα 1.1: Παράδειγμα διαγραμματικής απεικόνισης του μονοπατιού Wnt στη Reactome (στιγμιότυπο οθόνης).

1.1.2 WikiPathways

Η βάση δεδομένων Wikipathways [7] αποτελεί ένα αποθετήριο βιολογικής γνώσης οργανωμένης σε βιολογικές οδούς, όπως σηματοδοτικά, μεταβολικά, ρυθμιστικά μονοπάτια κλπ μέσω διαγραμματικής απεικόνισής τους. Παρουσιάζει αλληλεπιδράσεις τόσο μεταξύ των βιομορίων του κάθε μονοπατιού, όσο και μεταξύ διαφορετικών μονοπατιων. Δημιουργήθηκε το 2007 και ανανεώνεται σχεδόν σε καθημερινή βάση από την επιστημονική κοινότητα, η οποία συμμετέχει σταθερά στην κατασκευή και αναθεώρηση των μονοπατιών που συμπεριλαμβάνονται στη βάση καθώς παρέχει ενσωματωμένα γραφικά εργαλεία για την επεξεργασία και τη διευκόλυνση της αναπαράστασης των βιολογικών οδών και διαδικασιών. Υποστηρίζει 30 είδη οργανισμών μεταξύ των οποίων είναι ο *H. sapiens, M.musculus, D. melanogaster, C. elegans, A. thaliana* και πολλά είδη φυτών και βακτηρίων. Η βάση δεδομένων WikiPathways περιλαμβάνει ένα σύνολο 2958 μονοπατιών (Απρίλιος 2021) που αποτελούνται από πρωτεΐνες, γονίδια, μεταβολίτες, φάρμακα κλπ και περιλαμβάνει 46.105 αλληλεπιδράσεις μεταξύ των βιολογικών οντοτήτων. Κάθε μονοπάτι συνοδεύεται από πληροφορίες για έναν συγκεκριμένο βιολογικό μηχανισμό, συμπεριλαμβανομένου του διαγράμματος, της περιγραφής, των υπερσυνδέσεων με λεπτομερείς πληροφορίες για τα γονίδια, τις πρωτεΐνες και τους μεταβολίτες και τις σχετικές βιβλιογραφικές αναφορές (Εικόνα 1.2). Τέλος, να σημειωθεί ότι το περιεχόμενο της βάσης δεδομένων είναι προσβάσιμο μέσω προγράμματος περιήγησης, API ή μέσω Cytoscape [8] και ο χρήστης μπορεί να το κατεβάσει σε διάφορες μορφές για περαιτέρω ανάλυση των μονοπατιών με διάφορα εργαλεία, όπως το PathVisio [9] και το Cytoscape [10]. Παρέχονται υπερσύνδεσμοι σε άλλες βάσεις δεδομένων για συστατικά των μονοπατιών όπως NCBI, GO, Ensembl [11], UCSC Genome Browser [12], UniProt-TrEMBL, WIKIGENES [13], PDB [14].



Εικόνα 1.2: Διαγραμματική απεικόνιση (στιγμιότυπο οθόνης) μέσω της WikiPathways του μονοπατιού απόπτωσης ρυθμιζόμενο απο την HSP70. Το Μονοπάτι συνοδεύεται από πληροφορίες για τον συγκεκριμένο βιολογικό μηχανισμό, συμπεριλαμβανομένου του διαγράμματος, της περιγραφής, των υπερσυνδέσεων με λεπτομερείς πληροφορίες για τα γονίδια, τις πρωτεΐνες και τους μεταβολίτες και τις σχετικές βιβλιογραφικές αναφορές.

1.1.3 CORUM

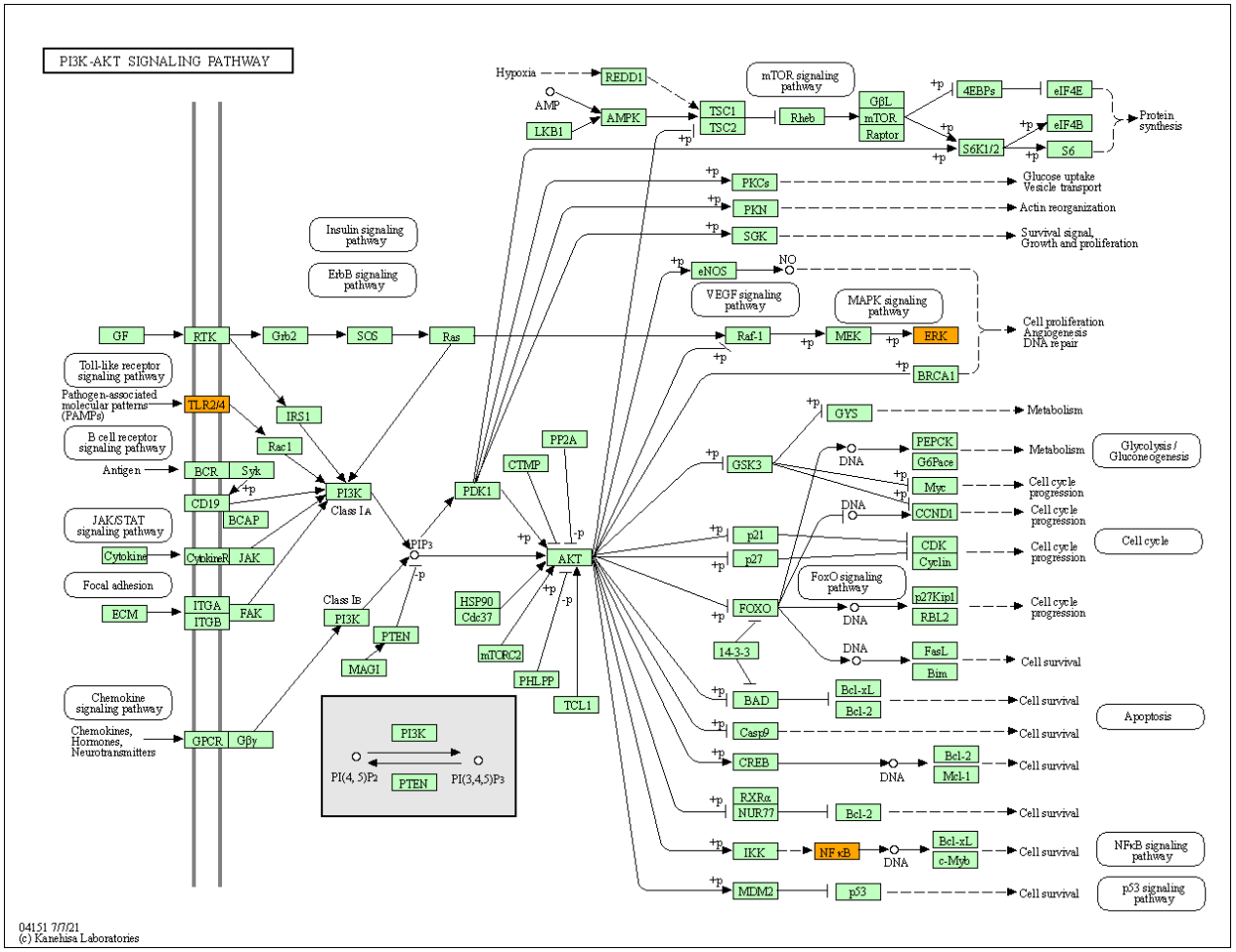
CORUM (Comprehensive Resource of Mammalian protein complexes) [15] αποτελεί μια συλλογή εγγραφών πρωτεϊνικών συμπλεγμάτων των θηλαστικών που συνοδεύονται από αναλυτικό σχολιασμό, ο οποίος περιλαμβάνει πληροφορίες για τη λειτουργία και το ρόλο των πρωτεϊνών, τον εντοπισμό, την σύνθεση των υπομονάδων, τις βιβλιογραφικές αναφορές, λειτουργικό εμπλουτισμό με όρους GO καθώς και συσχετίσεις με ασθένειες. Όλες οι πληροφορίες προέρχονται από πειράματα που δημοσιεύονται σε επιστημονικά άρθρα, ενώ αποκλείονται δεδομένα από πειράματα υψηλής απόδοσης. Για το λόγο αυτό, ο συνολικός αριθμός αλληλεπιδράσεων στη CORUM είναι σχετικά μικρός σε σύγκριση με άλλα αποθετήρια. Ωστόσο, η επιμέλεια των δεδομένων για κάθε καταχώρηση είναι λεπτομερής και προσεγμένη. Οι οργανισμοί που υποστηρίζει είναι 3 *(H. sapiens M. musculus, R. norvegicus)* και περιλαμβάνει 4274 σύμπλοκα πρωτεϊνών θηλαστικών τα οποία προέρχονται από 4473 διαφορετικά γονίδια [1], [15].

1.1.4 Kegg Pathways

Η KEGG (Εγκυκλοπαίδεια του Κιότο για τα γονίδια και τα γονιδιώματα-Kyoto Encyclopedia of Genes and Genomes) [16] αποτελείται από ένα σύνολο υπο-βάσεων δεδομένων που πραγματεύεται τα γονιδιώματα, τις βιολογικές οδούς, τις ασθένειες, τα φάρμακα και τις χημικές ουσίες. Συνίσταται από 15 βάσεις δεδομένων που είναι "με το χέρι" επιμελημένες (manually curated) και μια πρόσθετη υπολογιστικά παραγόμενη. Μεταξύ αυτών, η KEGG PATHWAY [17] που αποτελεί τον πυρήνας της KEGG και περιέχει βιολογικά μονοπάτια που παρουσιάζονται γραφικά με χάρτες οδών, παρόμοιους με της Reactome. Οι καταχωρημένες οντότητες που παρουσιάζονται σε κάθε βιολογική οδό περιλαμβάνουν μόρια, RNAs, γονίδια, πρωτεΐνες και μονοπάτια, καθώς και γονίδια ασθενειών και στόχους φαρμάκων τα οποία αποθηκεύονται ως μεμονωμένες καταχωρίσεις στις άλλες βάσεις δεδομένων της KEGG. Οι χάρτες οδών ταξινομούνται στις εξής ενότητες: Μεταβολισμός, Επεξεργασία γενετικών πληροφοριών (μεταγραφή, μετάφραση, αντιγραφή και επιδιόρθωση DNA κλπ.), Επεξεργασία περιβαλλοντικών πληροφοριών (μεμβρανική μεταφορά (membrane transport), μεταγωγή σήματος (signal transduction), κλπ.), Κυτταρικές διεργασίες (κυτταρική αύξηση, κυτταρικός θάνατος, λειτουργίες κυτταρικής μεμβράνης, κλπ.), Συστήματα οργάνων (ανοσοποιητικό σύστημα, ενδοκρινικό σύστημα, νευρικό σύστημα, κλπ.), Ανθρώπινες ασθένειες και Ανάπτυξη φαρμάκων.

Τα γονίδια που περιέχονται σε κάποιο μονοπάτι συνδέονται με λειτουργίες υψηλότερης τάξης στο πλαίσιο μεμονωμένων κυττάρων ή ολόκληρων οργανισμών. Τέτοιες λειτουργίες απεικονίζονται από έναν ιστό αλληλεπιδράσεων και χημικών αντιδράσεων, σχεδιασμένων με τη μορφή μονοπατιών της KEGG (Εικόνα 1.3), ιεραρχιών BRITE και Module KEGG. Η KEGG περιέχει 34.042.792 γονίδια, 781.759 μονοπάτια και 11.505 αντιδράσεις που σχετίζονται με 545 ευκαρυώτες, 6234 βακτήρια και 343 Αρχαία (Απρίλιος 2021). Παρέχονται σύνδεσμοι σε άλλες βάσεις δεδομένων για τα βιομόρια που συμμετέχουν στα διάφορα μονοπάτια, όπως GO, UniProt, άλλες βάσεις δεδομένων KEGG, Rhea [18], NCBI, PubChem, CheMBL, PDB (Chemical Components), ενώ ενσωματώνονται και οι αναφορές PubMed. Η βάση δεδομένων παρέχει

πρόσβαση μέσω API, ενώ το περιεχόμενο μπορεί να μεταφορτωθεί σε πολλές μορφές, όπως PNG, RDF και KGML. Επιπλέον, έχουν αναπτυχθεί πολλές εφαρμογές Cytoscape, τόσο από τους επιμελητές της βάσης δεδομένων όσο και από τρίτους χρήστες, που ενσωματώνουν την οπτικοποίηση και την ανάλυση δεδομένων KEGG με το Cytoscape [19], [20].



Εικόνα 1.3: Διαγραμματική απεικόνιση του μονοπατιού PI3K-AKT μέσω της KEGG Pathway (στιγμιότυπο οθόνης). Το διάγραμμα είναι διαδραστικό και η επιλογή κάθε όρου αναδρομολογεί το χρήστη σε νέα σελίδα με πληροφορίες για τον εκάστοτε επιλεγμένο όρο.

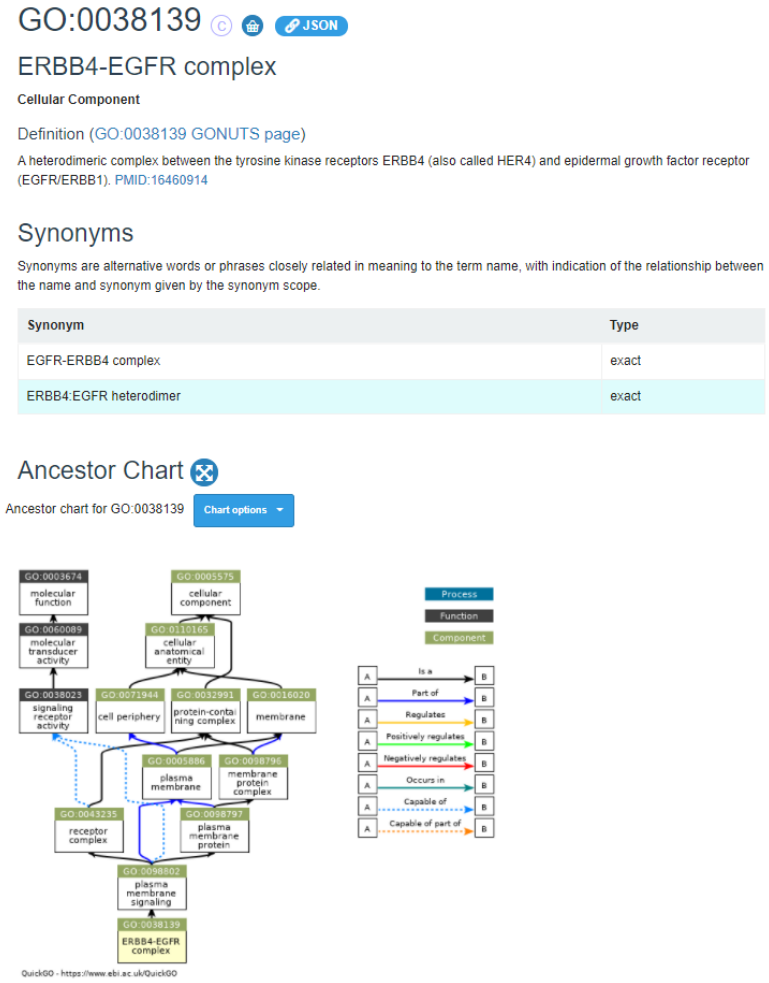
1.1.5 Gene Ontology

Η Gene Ontology (GO) [21], [22] είναι μια αυτοτελής βιοπληροφορική πλατφόρμα που περιέχει πληροφορίες σχετικά με γονίδια, τις λειτουργίες και τις σχέσεις μεταξύ τους με ένα δομημένο τρόπο και συνιστά την πιο ολοκληρωμένη πηγή πληροφοριών σχετικά με τις λειτουργίες των γονιδίων και των γονιδιακών προϊόντων μεταξύ όλων των βάσεων δεδομένων. Αποτελεί ένα ελεγχόμενο λεξιλόγιο βιολογικών όρων που είναι δομημένο και περιέχονται όροι, οι οποίοι είναι γνωστοί ως GO όροι (GO-terms). Σκοπός της αντίστοιχης Κοινοπραξίας (Gene Ontology Consortium) είναι η δημιουργία μιας συνεκτικής, παγκόσμιας ονοματολογίας γονιδίων για όλους τους οργανισμούς. Η Γονιδιακή Οντολογία έχει ενσωματώσει πολλές βάσεις δεδομένων και

αποθετήρια και η ενημέρωσή της είναι διαρκής καθώς η βιολογική γνώση ακόμη αποδελτιώνεται. Στην τελευταία έκδοση (2021) είναι καταχωρημένοι 43.850 όροι (GO terms), 7.928.834 GO annotations, ενώ αφορούν 1.568.828 γονιδιακά προϊόντα απο 5.086 είδη. Η Gene Ontology διαιρείται σε τρεις επιμέρους οντολογίες οι οποίες παρέχουν κοινές πληροφορίες για όλα τα είδη οργανισμών:

- Κυτταρική συνιστώσα (Cellular Component ή CC), περιλαμβάνει όρους οι οποίοι περιγράφουν τα μέρη του κυττάρου (υποκυτταρικές δομές) στα οποία εντοπίζεται ένα γονιδιακό προϊόν.
- Βιολογική διεργασία (Biological Process ή BP), περιλαμβάνει όρους οι οποίοι περιγράφουν μια ακολουθία διακριτών μοριακών λειτουργιών με καθορισμένη αρχή και τέλος για την επίτευξη ενός βιολογικού αποτελέσματος. Ωστόσο δεν περιλαμβάνονται η δυναμική και οι εξαρτήσεις μεταξύ των λειτουργιών αυτών, συνεπώς μια βιολογική διεργασία δεν ισοδυναμεί με ένα βιολογικό μονοπάτι.
- Μοριακή λειτουργία (Molecular Function ή MF), οι όροι που περιλαμβάνονται περιγράφουν λειτουργίες μεμονωμένων ή συμπλόκων γονιδιακών προϊόντων οι οποίες εκτελούνται σε μοριακό επίπεδο, χωρίς περεταίρω λεπτομέρειες για το πλαίσιο στο οποίο λαμβάνουν χώρα.

Οι σχέσεις μεταξύ των όρων της Gene Ontology οργανώνονται σε κατευθυνόμενους ακυκλικούς γράφους (directed acyclic graphs) (Εικόνα 1.4), όπου κάθε όρος έχει καθορισμένες σχέσεις με άλλους όρους του ίδιου ή διαφορετικού τομέα. Κάθε γονίδιο επομένως, αντιστοιχεί σε έναν ακυκλικό γράφο, ο οποίος περιλαμβάνει όλες τις λειτουργίες που είναι γνωστές για αυτό. Το πιο σημαντικό στοιχείο αυτών των οντολογιών-απόψεων είναι η ορθογωνιότητα, η διασφάλιση δηλαδή ύπαρξης κάθε όρου μόνο σε μία από τις τρεις απόψεις. Η ιδιότητα αυτή εγγυάται τη μοναδικότητα ενός γνωρίσματος που αποδίδεται σε ένα γονίδιο. Με απλά λόγια, όταν ένας όρος αποδίδεται σε ένα γονίδιο αυτόματα αποκλείεται η ύπαρξη ενός αντίστοιχου όρου από άλλη οντολογία-άποψη που να φέρει πανομοιότυπες ιδιότητες.



Εικόνα 1.4: Σχηματική αναπαράσταση (στιγμιότυπο οθόνης) ενός κατευθυνόμενου ακυκλικού γράφου που περιγράφει σχέσεις γονιδιακής οντολογίας σχετικά με τον όρο (term) ERBB4-EGFR complex (Cellular Components).

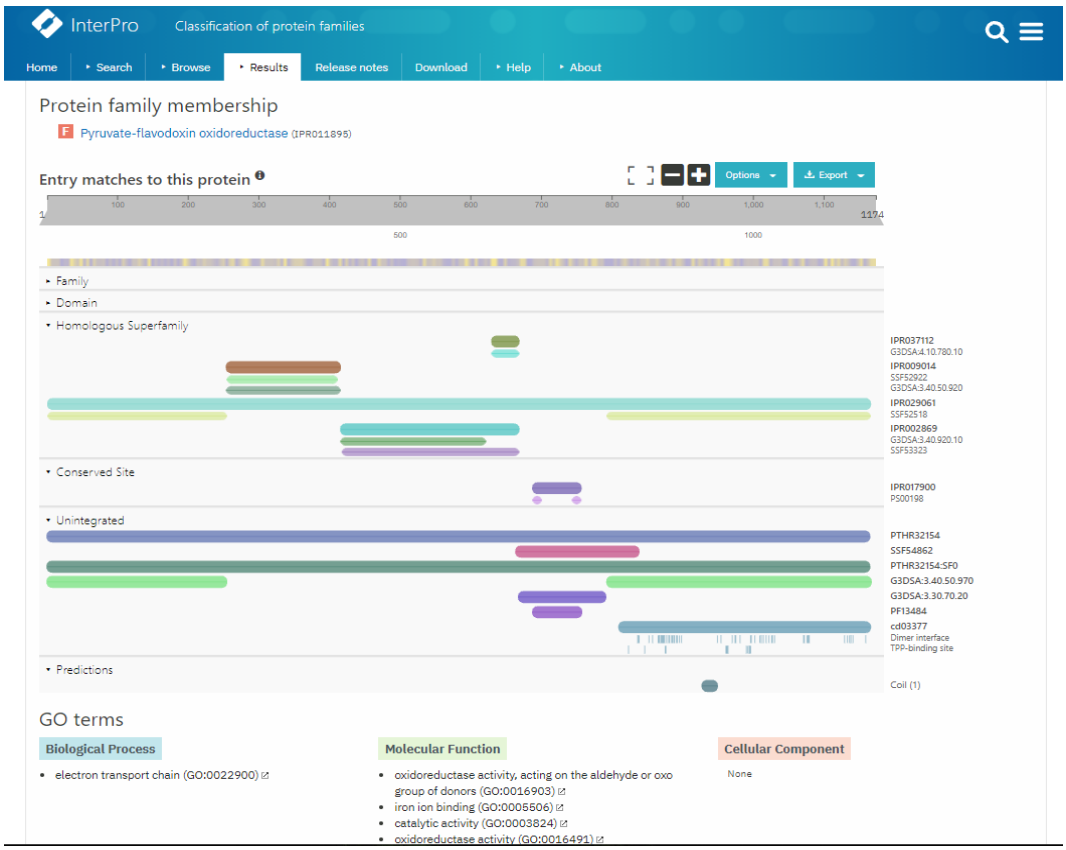
1.1.6 Pfam

Η βάση Pfam [23] αποτελεί μια μεγάλη συλλογή πρωτεϊνικών οικογενειών. Κάθε οικογένεια αντιπροσωπεύεται από πολλαπλές ευθυγραμμίσεις αλληλουχιών και ένα κρυφό hidden Markov model (HMM). Η μέθοδος αυτή συνιστά μια πιο ευαίσθητη μέθοδο στην εύρεση μακρινών ομόλογων με μεγάλη ταχύτητα ταχύτητα και αποτελεσματικότητα. Στην τρέχουσα έκδοση (34.0, Μάρτιος 2021), η βάση περιέχει δεδομένα για 19179 οικογένειες. Οι καταχωρίσεις Pfam ταξινομούνται με έναν από τους έξι τρόπους: οικογένεια, τομέας, επανάληψη, μοτίβα, σπειροειδής πηνίο, ανακατεμένη. Οι σχετικές καταχωρήσεις της Pfam ομαδοποιούνται σε ομάδες. Η σχέση μπορεί να οριστεί από την ομοιότητα της ακολουθίας, της δομής ή του προφίλ HMM. Η PFAM αποτελείται από δύο υποσύνολα, την PFAM-A, και την PFAM-B. Η PFAM-A περιέχει καταχωρήσεις (οικογένειες) υψηλής «ποιότητας», καθώς έχουν όλες υποστεί σχολιασμό από ειδικούς, ενώ υπάρχουν αναφορές σε άλλες βάσεις δεδομένων και κυρίως σε βιβλιογραφία. Η PFAM-B συνιστά ένα υποσύνολο, το οποίο προκύπτει με αυτοματοποιημένο τρόπο εντοπίζοντας

τις ομοιότητες ανάμεσα στις πρωτεϊνικές περιοχές που απομένουν όταν αφαιρεθούν οι περιοχές που αντιστοιχούν στις καταχωρήσεις της PFAM-A.

1.1.7 INTERPRO

Η βάση δεδομένων InterPro [24] ιδρύθηκε το 1999 και παρέχει μια λειτουργική και ολοκληρωμένη κατηγοριοποίηση πρωτεϊνικών αλληλουχιών σε οικογένειες προσδιορίζοντας λειτουργικά αυτοτελείς περιοχές (Domains) και σημαντικά συντηρημένες περιοχές (Εικόνα 1.5). Η InterPro συνιστά μία πηγή που εγκολπώνει πολλές διαφορετικές βάσεις δεδομένων, αποτελεί δηλαδή μια κοινοπραξία και βάσει προγνωστικών μοντέλων (υπογραφές) που παρέχονται από αυτές τις βάσεις πραγματοποιεί την ταξινόμηση των πρωτεϊνών. Η InterPro ενσωματώνει υπογραφές από τις ακόλουθες 13 βάσεις δεδομένων: CATH [25], CDD [26], HAMAP [27], MobiDB Lite [28], Panther [29], Pfam, PIRSF [30], PRINTS [31], Prosite [32], SFLD [33], SMART [34], SUPERFAMILY [35] και TIGRfams [36]. Έτσι, αξιοποιεί τα πλεονεκτήματα κάθε επιμέρους βάσης, με αποτέλεσμα τη δημιουργία ενός ισχυρού εργαλείου διάγνωσης και μια ολοκληρωμένη πηγή πληροφορίας. Το InterProScan είναι το υποκείμενο λογισμικό που επιτρέπει την αναζήτηση αλληλουχιών πρωτεϊνών και νουκλεϊκών οξέων έναντι των υπογραφών της InterPro. Η InterPro ενημερώνεται περίπου κάθε 8 εβδομάδες. Η πρόσβαση είναι ελεύθερη μέσω προγράμματος περιήγησης, API InterPro και λήψη δεδομένων για τοπικές αναλύσεις.



Εικόνα 1.5: Παράδειγμα αναζήτησης πρωτεϊνικής αλληλουχίας στην InterPro (στιγμιότυπο οθόνης).

1.1.8 PubMed

Η βιβλιογραφική βάση PubMed [37] περιλαμβάνει περισσότερες από 23 εκατομμύρια αναφορές σε βιοϊατρική πληροφορία προερχόμενη από τη MEDLINE, περιοδικά και ηλεκτρονικά βιβλία επιστημών υγείας. Ο χρήστης έχει τη δυνατότητα σύνθετων αναζητήσεων με βάση τον τίτλο του άρθρου, ή/και τα ονόματα των συγγραφέων, καθώς και λέξεις κλειδιά. Τα αποτελέσματα των αναζητήσεων παρέχουν τους συγγραφείς του άρθρου με τα στοιχεία του περιοδικούς και της δημοσίευσης, την περίληψη και συνδέσμους προς το πλήρες κείμενο από την PubMed Central ή στον ιστοχώρο του περιοδικού.

1.1.9 TRANSFAC

Η TRANSFAC (TRANScription FACtor database) [38], [39] συνιστά την παλαιότερη ενεργά διατηρούμενη βάση δεδομένων στον τομέα μελέτης των μεταγραφικών παραγόντων και πιο συγκεκριμένα δεδομένα ευκαρυωτικών μεταγραφικών παραγόντων, των γονιδιωματικές θέσεων δέσμευσής τους, καθώς και τα προφίλ δέσμευσης στο DNA. Τα περιεχόμενά της βάσης είναι βασικά δομημένα σε πίνακες που παρέχουν πληροφορίες σχετικά με τους μεταγραφικούς παράγοντες (πίνακας FACTOR) ή τις θέσεις γονιδιωματικής σύνδεσής τους (SITE). Όλες οι πληροφορίες σε αυτούς τους πίνακες έχουν εξαχθεί με μη αυτόματο τρόπο από τις αρχικές δημοσιεύσεις. Ωστόσο, μπορούν να ανακτηθούν δεδομένα ημι-αυτόματα από πειράματα υψηλής απόδοσης, όπως απο ChIPseq. Η TRANSFAC προσπαθεί να καλύψει ολόκληρη την περιοχή των ευκαρυωτικών μεταγραφικών παραγόντων, αν και η κάλυψη διαφορετικών ταξινομημάτων μπορεί να είναι σημαντικά διαφορετική. Στη τρέχουσα έκδοση 2021.2), η βάση περιέχει ενδεικτικά δεδομένα για 48.094 μεταγραφικούς παράγοντες, 50.903 DNA Sites, 102.860 γονίδια, 40.648 βιβλιογραφικές αναφορές και 199.183 ενισχυτές. Η χρήση μιας παλαιότερης έκδοσης της TRANSFAC είναι δωρεάν. Η πρόσβαση στην πιο ενημερωμένη έκδοση απαιτεί άδεια.

1.1.10 miRTarBase

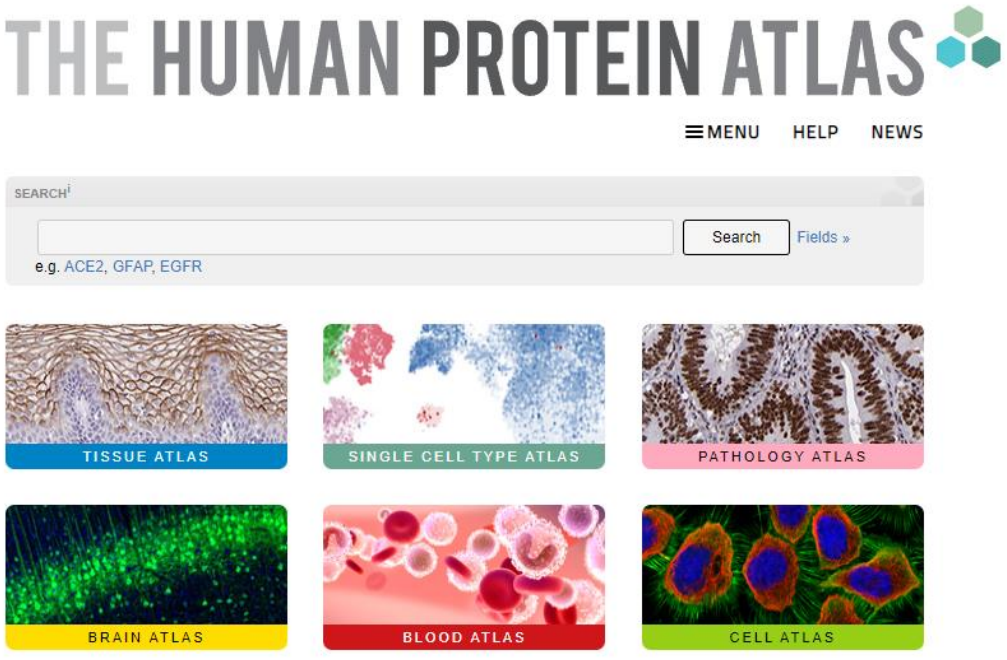
Η miRBase [40] είναι το κύριο διαδικτυακό δωρεάν αποθετήριο για τις microRNA αλληλουχίες και τον σχολιασμό τους και περιλαμβάνει αλληλεπιδράσεις microRNA και γονιδίων, οι οποίες έχουν επιβεβαιωθεί πειραματικά. Η τρέχουσα έκδοση (miRBase 22.1) περιέχει 38.589 πρόδρομες φουρκέτες και πάνω από 48.860 ξεχωριστές ώριμες αλληλουχίες microRNA από 271 είδη. Οι τεχνολογίες αλληλούχισης έχουν προκαλέσει μια απότομη αύξηση του ρυθμού ανακάλυψης νέων μορίων microRNA. Ο χρήστης μπορεί να δει όλα τα δεδομένα που σχετίζονται με ένα δεδομένο microRNA, να φιλτράρει ανά πείραμα και να μετρήσει και να αναζητήσει microRNAs με έκφραση που αφορά τον ιστό και το στάδιο ανάπτυξης. Επίσης, παρέχει σχολιασμένους όρους για κάθε microRNA απο την Gene Ontology καθώς και επιλογή εξόρυξης κειμένου για την αναζήτηση ονομάτων γονιδίων microRNA στο πλήρες κείμενο των άρθρων ανοικτής πρόσβασης με πάνω από 500.000 προτάσεις από 18.542 έγγραφα να περιέχουν ονόματα microRNA. Έτσι κάθε εγγραφή συνοδεύεται από ένα σύνολο βιβλιογραφικών σχολιασμών.

1.1.11 Human Protein Atlas

Το Human Protein Atlas [41] είναι ένα πρόγραμμα που ξεκίνησε το 2003 με στόχο τη χαρτογράφηση όλων των ανθρώπινων πρωτεϊνών στα κύτταρα, τους ιστούς και τα όργανα του ανθρώπου χρησιμοποιώντας τις λεγόμενες ομικές τεχνολογίες και να δημιουργηθεί μια ανοιχτού περιεχομένου βάση για τη διερεύνηση του ανθρώπινου πρωτεώματος που παρέχει γνώση σε όποιον τη χρειάζεται. Αποτελείται από έξι υποκατηγορίες (Εικόνα 1.6), καθεμία εκ των οποίων εστιάζει σε μια συγκεκριμένη πτυχή της ανάλυσης των ανθρώπινων πρωτεϊνών:

- Tissue Atlas, που δείχνει την κατανομή των πρωτεϊνών σε όλους τους κύριους ιστούς και στα όργανα του ανθρώπινου σώματος
- Single Cell Type Atlas, που δείχνει έκφραση γονιδίων που κωδικοποιούν πρωτεΐνες σε μεμονωμένους τύπους ανθρώπινων κυττάρων
- Pathology Atlas, που δείχνει τον αντίκτυπο των επιπέδων πρωτεΐνης στην επιβίωση ασθενών με καρκίνο
- Blood Atlas, που περιγράφει τις πρωτεΐνες που ανιχνεύονται στους τύπους των κυττάρων του αίματος και τις πρωτεΐνες που εκκρίνονται από τους ανθρώπινους ιστούς
- Brain Atlas, για την κατανομή των πρωτεϊνών σε διάφορες περιοχές του εγκεφάλου των θηλαστικών
- Cell Atlas, που δείχνει τον υποκυτταρικό εντοπισμό των πρωτεϊνών σε μεμονωμένα κύτταρα

Στην τρέχουσα έκδοση (20.1, 2021-02-24) η ανάλυση του πρωτεόματος έχει βασιστεί σε 26941 αντισώματα (antibodies targeting) και 17165 πρωτεΐνες.



Εικόνα 1.6: Διεπαφή χρήστη (στιγμιότυπο οθόνης), όπου φαίνονται οι 6 κατηγορίες που συνιστούν την βάση Human Protein Atlas.

1.1.12 Human Phenotype Ontology

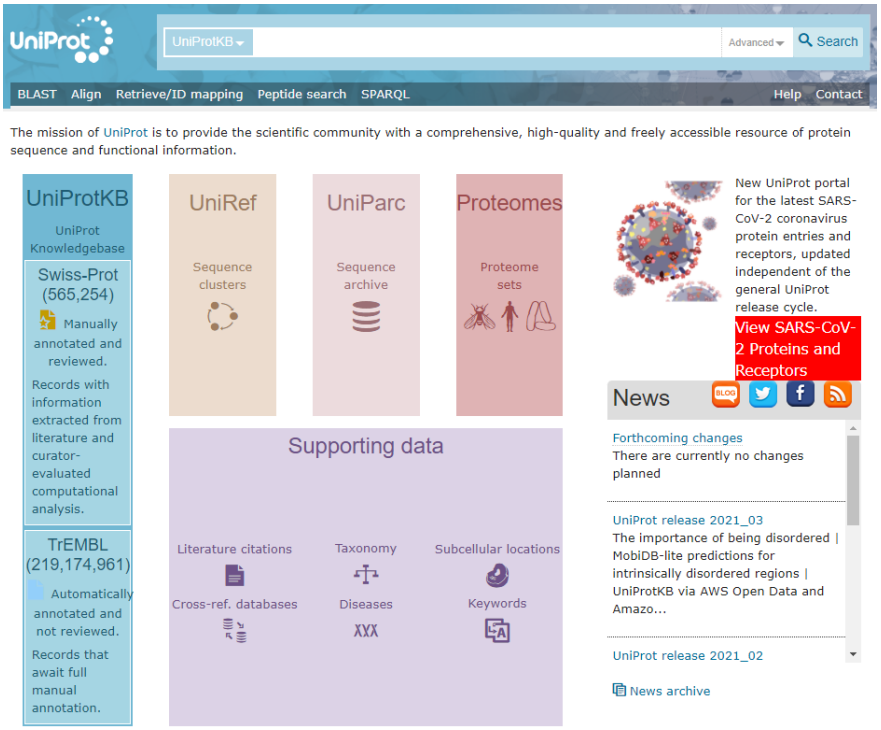
Η Human Phenotype Ontology (HPO) [42] δημιουργήθηκε το 2008 και περιέχει πληροφορίες σχετικά με ασθένειες, φαινοτυπικές ανωμαλίες που μπορεί να σχετίζονται με αυτές και ταυτόχρονα συσχέτιση αυτών με γονίδια και μπορεί να χρησιμοποιηθεί για περιγραφή και υπολογιστική ανάλυση των φαινοτυπικών ανωμαλιών. Η τρέχουσα έκδοση (Αύγουστος 2021) περιέχει πάνω από 13.000 όρους που συνοδεύονται από πάνω 156.000 βιβλιογραφικούς σχολιασμούς για κληρονομήσιμες ασθένειες και οργανώνονται σε κατευθυνόμενους ακυκλικούς γράφους (directed acyclic graphs), όπου κάθε όρος έχει καθορισμένες σχέσεις με άλλους όρους του ίδιου ή διαφορετικού τομέα. Η HPO βασίζεται σε ιατρική βιβλιογραφία, Orphanet, DECIPHER και OMIM. Η HPO έχει αναπτύξει το λογισμικό της με στόχο τη φαινοτυπική διαφορική διάγνωση, τη γονιδιωματική διάγνωση, τη μεταφραστική έρευνα και μιας σειράς εφαρμογών στην υπολογιστική βιολογία παρέχοντας τα μέσα για τον υπολογισμό του κλινικού φαινοτύπου. Επίσης, εστιάζει όλο και περισσότερο στην λήψη δεδομένων φαινοτυπικές ανωμαλιών και ασθενειών από διάφορες ομάδες, όπως διεθνείς οργανισμούς σπάνιων ασθενειών, μητρώα, κλινικά εργαστήρια, βιοϊατρικά και κλινικά εργαλεία λογισμικού με σκοπό τη παγκόσμια ανταλλαγή δεδομένων για τον εντοπισμό αιτίων ανάπτυξης των ασθενειών.

1.1.13 UniProt

Η βάση δεδομένων UniProt (Universal Protein Resource) [43], αποτελεί μια ολοκληρωμένη πηγή που περιέχει πληροφορίες για την αλληλουχία πρωτεϊνών. Η UniProt αποτελείται από τη βάση UniProt Knowledgebase (UniProtKB), τη βάση UniProt Reference Clusters (UniRef) και τη βάση UniProt Archive (UniParc). Η UniProt δημιουργήθηκε με κύριο στόχο να αποτελέσει την πηγή αναφοράς για την περιγραφή των πρωτεϊνών. Δημιουργήθηκε το 2002 από την ένωση των βάσεων Swiss-Prot, TrEMBL και PIRPSD (Translated EMBL Nucleotide Sequence Data Library and Protein Information Resource Protein Sequence Database) (Εικόνα 1.7). Ένα από το πιο σημαντικά της χαρακτηριστικά είναι ότι διασταυρώνει τις πληροφορίες που καταγράφει με άλλες βάσεις βιολογικών δεδομένων και έτσι μπορεί και προσφέρει πρακτικά όλη την επίσημη γνώση που έχουμε πάνω στις πρωτεΐνες.

Τα δεδομένα της Uniprot χωρίζονται γενικά σε δύο κατηγορίες, τα reviewed (επιβεβαιωμένα) – αυτά που προέρχονται από την βάση Swiss-Prot και περιέχουν περισσότερη πληροφορία καθώς έχουν καταγραφεί με ανθρώπινη επιμέλεια και τα unreviewed (μη επιβεβαιωμένα) – αυτά που δίνονται από την TrEMBL που συλλέγει πληροφορίες αυτόματα. Ειδικότερα, η UniprotKB/SwissProt περιέχει 565.254 αλληλουχίες (Έκδοση 2021_03- 2 Ιουνίου 2021), οι οποίες έχουν περάσει από κάποιου είδους έλεγχο και συνοδεύονται από συμπληρωματικά σχόλια όπως βιβλιογραφικές αναφορές, γενικά στοιχεία δευτεροταγούς δομής, συνδέσμους σε άλλες βάσεις δεδομένων σχετικές με κάθε εγγραφή, καθώς και σημειώσεις για τη βιολογική λειτουργία (αν είναι γνωστές), καθώς και άλλες χρήσιμες πληροφορίες. Η Uniprot/TrEMBL περιέχει σήμερα (Έκδοση 2021_03- 2 Ιουνίου 2021) 219.174.961 αλληλουχίες οι οποίες όμως δεν έχουν υποστεί ανθρώπινο σχολιασμό. Η σχέση ανάμεσα στις δύο βάσεις είναι δυναμική και πληροφορίες που αρχικά περιέχονται στην TrEMBL, αφού περάσουν από τον έλεγχο ενός

επιμελητή, μεταφέρονται στην Swiss-Prot ή και το αντίστροφο καθώς κάποια εγγραφή μπορεί να χρειάζεται να ξαναπεράσει από έλεγχο. Η UniProt ενημερώνεται κάθε τέσσερις εβδομάδες. Υπάρχει δυνατότητα λήψης μικρών συνόλων δεδομένων και υποσυνόλων απευθείας από τον ιστότοπο ακολουθώντας τη σύνδεση λήψης σε οποιαδήποτε σελίδα αποτελεσμάτων αναζήτησης.



Εικόνα 1.7: Η διεπαφή ιστού της UniProt (στιγμιότυπο οθόνης) προσφέρει στο χρήστη πρόσβαση στις UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef) και UniProt Archive (UniParc).

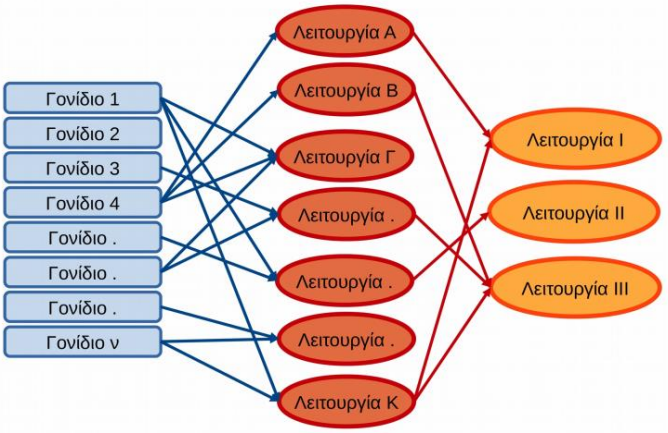
1.1.14 Disease Ontology

Η Disease Ontology [44] αναπτύχθηκε με σκοπό την πλήρη περιγραφή διαφόρων ανθρωπίνων ασθενειών και των χαρακτηριστικών τους. Η οντολογία ορίζει 9069 ασθένειες και 15 σχέσεις. Ουσιαστικά αποτελεί μια ταξινομία και περιέχει αποκλειστικά ιεραρχικές σχέσεις («is-a») και σχέσεις συνωνύμων όρων («synonym»). Οι σχέσεις συνωνύμων χρησιμοποιούνται επιπλέον για να συνδέσουν κάθε τάξη με τις «συνώνυμες» τάξεις της όπως αυτές εμφανίζονται σε άλλες οντολογίες, όπως η Gene Ontology. Η δομή αυτή της βάσης και οι εξωτερικές αναφορές σε άλλες βάσεις θα επιτρέψουν την ενσωμάτωση διαφορετικών συνόλων δεδομένων αναφορικά με την υπό μελέτη νόσο. Προσπαθεί να συσχετίσει ασθένειες σε διάφορα επίπεδα, όπως με τον φαινότυπο, το περιβάλλον και τη γενετική παρέχοντας υπολογιστικές μεθόδους για τις συσχετίσεις αυτές. Η βάση ενσωματώνει και συνδέει πάνω από 46.000 όρους ασθένειας και ιατρικού λεξιλογίου μέσω εκτεταμένων αναφορών (MeSH, ICD, thesaurus NCI, SNOMED και OMIM). Οι όροι οργανώνονται σε κατευθυνόμενους ακυκλικούς γράφους (directed acyclic graphs), όπου κάθε όρος έχει καθορισμένες σχέσεις με άλλους όρους του ίδιου ή διαφορετικού τομέα και αναπαριστά αιτιολογικές κλάσεις ασθενειών

1.2 Ανάλυση λειτουργικού εμπλουτισμού (Functional Enrichment Analysis)

Σε πολλές περιστάσεις στη βιολογική έρευνα, μετά την εκτέλεση πειραμάτων, προκύπτουν λίστες γονιδίων ή πρωτεινών. Ένα βασικό ερώτημα που τίθεται, είναι "σε ποιες βιολογικές διαδικασίες ή σηματοδοτικά μονοπάτια ανήκουν στο σύνολο τους αυτές οι ομάδες". Χάρη στην πληθώρα των δεδομένων που εμπεριέχονται στις παραπάνω βάσεις (Ενότητα 1.1), μπορεί κάποιος να ελέγξει ποια βιολογική διεργασία αντιπροσωπεύουν καλύτερα τα δεδομένα του. Για αυτό το λόγο χρησιμοποιούνται διάφορες στατιστικές μετρικές και εκλεπτισμένα μοντέλα, έτσι ώστε τα αποτελέσματα να είναι όσο το δυνατότερο σαφέστερα. Οι αναλύσεις ομαδοποίησης γονιδίων συμβάλλουν στην ερμηνεία των αποτελεσμάτων από βιολογικής σκοπιάς [1], [45], [46].

Η διαδικασία ομαδοποίησης γονιδίων/πρωτεϊνών αναφέρεται συχνά ως εμπλουτισμός ή λειτουργικός εμπλουτισμός (functional enrichment) και συνίσταται στην κατηγοριοποίηση των γονιδίων/πρωτεϊνών μιας λίστας σε βιολογικες κατηγορίες (π.χ. σηματοδοτικά μονοπάτια) με βάση γνωστές λειτουργίες των γονιδίων που υπάρχουν αποθηκευμένες σε βάσεις δεδομένων (Εικόνα 1.8). Κατηγοριοποιήσεις που δε βασίζονται σε μια απλή αντιστοίχηση, αλλά εμπεριέχουν και ιεραρχικές σχέσεις μεταξύ των ιδιοτήτων με τρόπο τέτοιο που να οδηγούν σε σύνθετες ταξινομήσεις, ονομάζονται οντολογίες. Καθώς οι οντότητες αυτές είναι τα γονίδια και τα χαρακτηριστικά με τα οποία οργανώνονται είναι οι λειτουργικές ονομάζονται, εύλογα, γονιδιακές οντολογίες [45], [47]. Είναι λογικό ότι δεν παρατηρείται αυστηρή αντιστοιχία ενός γονιδίου/πρωτείνης αποκλειστικά σε μία κατηγορία, αλλά οι αντιστοιχίσεις είναι ελεύθερες, δηλαδή ένα γονίδιο/πρωτεΐνη μπορεί να ενταχθεί σε παραπάνω από μία κατηγορίες [48], [49].



Εικόνα 1.8: Διαγραμματική απεικόνιση κατηγοριοποίησης μιας λίστας ν γονιδίων. Δεν παρατηρείται αυστηρή αντιστοιχία ενός γονιδίου/πρωτείνης αποκλειστικά σε μία κατηγορία, αλλά οι αντιστοιχίσεις είναι ελεύθερες, δηλαδή ένα γονίδιο/πρωτεΐνη μπορεί να ενταχθεί σε παραπάνω από μία κατηγορίες. Το γονίδιο 2 δεν αντιστοιχίζεται σε καμία λειτουργία, ενώ τα 1 και 4 σε τρεις διαφορετικές λειτουργίες. Οι ίδιες οι λειτουργίες οργανώνονται ιεραρχικά από ειδικότερες (Α, Β, Γ κλπ) σε γενικότερες (Ι, ΙΙ, ΙΙΙ). (Πηγή Εικόνας [49]).

Έχουν προταθεί πολλοί αλγόριθμοι και εργαλεία, τα οποία επιτελούν αυτοματοποιημένα την ανάλυση λειτουργικού εμπλουτισμού, βοηθώντας έτσι στην ανάκτηση πληροφοριών σε οργανωμενο επίπεδο που άπτεται κάθε λίστας γονιδίων και μπορεί να συνεισφέρει στην κατανόηση και ερμηνεία των αποτελεσμάτων των αναλύσεων γονιδιακής έκφρασης. Ειδικότερα για την ανάλυση λειτουργικού εμπλουτισμού με βάση τους αλγορίθμους που χρησιμοποιούνται

έχουν προταθεί τρεις κατηγορίες-κλάσεις [48] βάσει του σχήματος που χρησιμοποιείται για την ανάλυση των δεδομένων εισόδου:

1. Απλή Ανάλυση Εμπλουτισμού (Singular Enrichment Analysis-SEA)
2. Ανάλυση Εμπλουτισμού σε Σύνολα Γονιδίων (Gene Set Enrichment Analysis-GSEA)
3. Σπονδυλωτή Ανάλυση Εμπλουτισμού (Modular enrichment analysis-MEA)

Όλοι οι αλγόριθμοι γενικότερα βασίζονται σε τρεις βασικούς πυλώνες:

- τα δεδομένα εισόδου αποτελούν μια λίστα γονιδίων ή πρωτεϊνών,
- υπάρχει μια πηγή δεδομένων δηλαδή, τυπικές περιγραφές των λειτουργιών που βασίζονται σε μια βιολογική οντολογία, π.χ., Gene Ontology,
- ο τρόπος αξιολόγησης των κατηγοριοποιήσεων με βάση στατιστικές μεθόδους [50].

Συνεπώς, ο εμπλουτισμός μπορεί να αξιολογηθεί ποσοτικά με γνωστές στατιστικές μεθόδους, όπως το χ-τετράγωνο τεστ, Fisher's exact test, η διωνυμική κατανομή και η υπεργεωμετρική κατανομή κλπ. Έτσι, μπορεί να εξαχθούν συμπεράσματα για την στατιστική σημαντικότητα των εμπλουτισμένων όρων και να αξιολογηθούν ποια αποτελέσματα μπορούν να θεωρηθούν πιο σημαντικά σημαντικά για την εκαστοτε μελετη [45].

1.2.1 Απλή Ανάλυση Εμπλουτισμού (SEA)

Αποτελεί την πρώτη μέθοδο υλοποίησης της ανάλυσης εμπλουτισμού. Ο συγκεκριμένος τύπος ανάλυσης είναι ο απλούστερος αλλά αποτελεί τη βάση για όλες τις πιο σύνθετες μεθοδολογίες. Όλοι οι αλγόριθμοι που ακολουθούν αυτη τη μεθοδολογία δέχονται ως αρχείο εισόδου τα προεπιλεγμένα γονίδια του χρήστη (π.χ. διαφορικά εκφρασμένα γονίδια που επιλέγονται μεταξύ των αποτελεσμάτων της πειραματικής και της ομάδας ελέγχου με βάση τις τιμές p-value ≤ 0.05 και fold change ≥ 1.5). Στην συνέχεια ελέγχουν επαναληπτικά τον εμπλουτισμό κάθε όρου (annotation term) έναν προς έναν με γραμμική συσχέτιση (linear mode). Έτσι κάθε όρος χαρακτηρίζεται από μια πιθανότητα εμπλουτισμού (p-value εμπλουτισμού) και αν κάποια εγγραφή της λίστας περάσει την κατωφλική τιμή αναφέρεται στο χρήστη. Ο υπολογισμός του p-value πραγματοποιείται χρησιμοποιώντας γνωστές στατιστικές μεθόδους, όπως χ τετράγωνο, Fisher's exact test, η διωνυμική πιθανότητα και η υπεργεωμετρική κατανομή [45].

1.2.2.Ανάλυση Εμπλουτισμού σε Σύνολα Γονιδίων (GSEA)

Η Ανάλυση Εμπλουτισμού σε Σύνολα Γονιδίων (Gene Set Enrichment Analysis, GSEA) υπερτερεί σε δύο βασικά σημεία συγκριτικά με την απλούστερη SEA. Αρχικά, ως όρισμα δεν δέχεται το επιλεγμένο με αυθαίρετα κριτήρια από το χρήστη υποσύνολο των διαφορικά εκφραζόμενων γονιδίων. Δεύτερον, λόγω της απουσίας τιμών-κατωφλίων, χρησιμοποιεί το σύνολο των δεδομένων αντί για ένα περιορισμένο μέρος τους. Τα δεδομένα εισόδου στην GSEA είναι οι τιμές έκφρασης από ολόκληρο το πείραμα. Το τελικό αποτέλεσμα είναι κι εδώ μια σειρά από τιμές p-value που αξιολογούν το βαθμό εμπλουτισμού μιας δεδομένης λειτουργίας, όμως ο τρόπος που υπολογίζεται τόσο η κάθε τιμή p-value αλλά και ο εμπλουτισμός διαφέρουν από την SEA [45].

Έστω λοιπόν, μια λίστα L γονιδίων ή πρωτεϊνών ως αποτελεσμα πειραμάτων διαφορικής έκφρασης. Η ανάλυση εμπλουτισμού συνόλου αξιολογεί, την κατανομή/αντιστοιχία των όρων της λίστας L σε μια βιολογική κατηγορία S με τη χρήση στατιστικών εργαλείων.Τα γονίδια που ανήκουν σε μια συγκεκριμένη βιολογική κατηγορία μπορεί να είναι τυχαία διεσπαρμένα εντός της λίστας ή να συσσωρεύονται στην κορυφή ή στη βάση της. Οι βιολογικές κατηγορίες, στο σύνολό τους ονομάζονται συλλογές γονιδιακών συνόλων (gene set collections) [51], [48].
Για την ανάλυση εμπλουτισμού επιτελούνται 3 βήματα [48]:

- Σε πρώτο στάδιο πραγματοποιείται υπολογισμός του σκορ εμπλουτισμού (enrichment score-ES), με χρήση του Kolmogorov-Smirnov (KS) [48], [52] στατιστικού τυχαίων βημάτων (random walk statistic). Το σκορ εμπλουτισμού είναι η μέγιστη απόκλιση από το μηδέν κατά τα τυχαία βήματα: καθώς η λίστα L διατρέχεται από πάνω προς τα κάτω, το συνολικό σκορ προσαυξάνεται όταν ένα γονίδιο της λίστας ανήκει στην κατηγορία S, και μειώνεται όταν ένα γονίδιο της λίστας δεν ανήκει στην ίδια κατηγορία [48].
- Έπειτα, αξιολογείται αν το αποτέλεσμα του σκορ εμπλουτισμού μπορεί να θεωρηθεί στατιστικά σημαντικό με χρήση της τιμής p-value σε σχέση με τη μηδενική κατανομή για το ES [48].
- Τέλος, γίνεται κανονικοποίηση του ES για το σύνολο των γονιδίων και υπολογίζεται το στατιστικό FDR (False Discovery Rate) το οποίο εκτιμά το αναμενόμενο ποσοστό των σφαλμάτων τύπου I, δηλαδή το ποσοστό των κανονικοποιημένων ES τα οποία αντιπροσωπεύουν μια απορριφθείσα θετική έκφραση [48].

Μερικές φορές χρησιμοποιούνται παραμετρικές στατιστικές προσεγγίσεις, όπως z-score, t-test, ανάλυση μετάθεσης κ.λπ.

1.2.3 Σπονδυλωτή (modular) Ανάλυση Εμπλουτισμού (MEA)

Η Σπονδυλωτή Ανάλυση Εμπλουτισμού (Modular Enrichment Analysis, MEA) [45] συμπεριλαμβάνει στην ανάλυση επιπρόσθετους αλγορίθμους που αποσκοπούν στην ανάδειξη ιδιοτήτων δικτύων που λαμβάνουν υπόψη τις σχέσεις μεταξύ λειτουργικών όρων. Αυτό σημαίνει ότι στην περίπτωση που δύο όροι π.χ. γονιδιακής οντολογίας βρεθούν να είναι σημαντικά εμπλουτισμένοι αλλά ταυτόχρονα βρίσκονται και σε γειτονικές θέσεις στο γράφο της ιεραρχίας των όρων, θα θεωρηθούν ακόμα μεγαλύτερης σημασίας. Η μέθοδος αυτή υπερτερεί στο γεγονός ότι δύναται η δυνατότητα για εξόρυξη πληροφορίας συσχετιζόμενη με βαθύτερες βιολογικές σχέσεις, όπως η ιεραρχική οργάνωση κυτταρικών διεργασιών ή οι αλληλεπιδράσεις μεταξύ μονοπατιών. Αντιθέτως, καθώς για την εφαρμογή αυτής της ανάλυσης είναι προαπαιτούμενη η ύπαρξη ιεραρχίας στην οργάνωση των λειτουργικών κατηγοριών (όρων) μπορεί να εφαρμοστεί κυρίως στην περίπτωση των γονιδιακών οντολογιών. Οι μέθοδοι που ενσωματώνουν τη MEA ανήκουν στην τελευταία γενιά εργαλείων λειτουργικής ανάλυσης.

1.3 Στατιστικοί έλεγχοι υποθέσεων

Η ανάλυση των βιολογικών πειραματικών δεδομένων εξαιτίας της πολυπλοκότητας τους και της αναγκαιότητας απο αυτά να προκύψουν νέα στοιχεία, ιδιότητες κλπ απαιτούν την ύπαρξη

ενός μαθηματικού υποβάθρου, το οποίο καλύπτουν οι τομείς της Στατιστικής και των Πιθανοτήτων. Στόχος της στατιστικής ανάλυσης είναι ο έλεγχος της επαναληψιμότητας στα πειράματα μεγάλης κλίμακας ή ακόμα και η εύρεση πιθανών σχέσεων που εμφανίζουν τα πειραματικά δεδομένα με στατιστική και βιολογική σημαντικότητα.

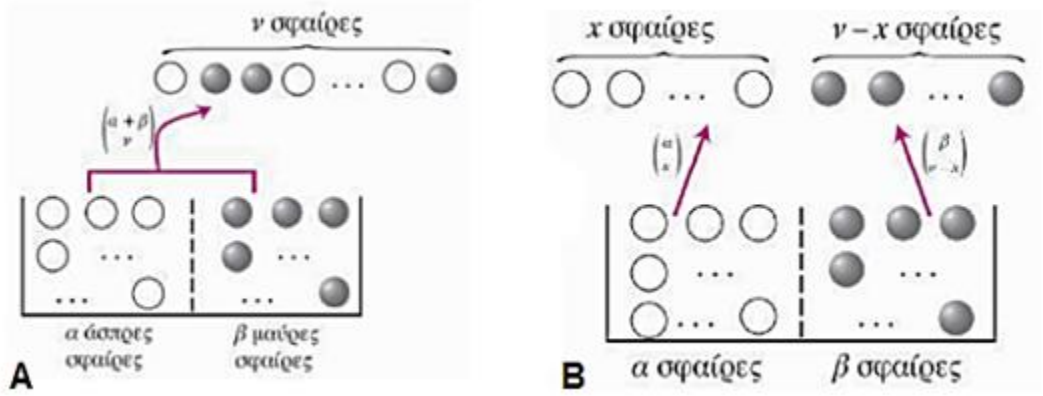
Οι στατιστικοί έλεγχοι υποθέσεων έχουν ως στόχο να εξάγουν συμπεράσματα που αφορά έναν πληθυσμό με βάση δεδομένα από ένα τυχαίο δείγμα του ίδιου πληθυσμού. Σε πειράματα έκφρασης, συγκρίνεται μια συνθήκη μελέτης με μια συνθήκη ελέγχου ως προς μια συγκεκριμένη «υπόθεση» για τη συμπεριφορά των τιμών. Η υπόθεση αυτή ονομάζεται μηδενική υπόθεση (null hypothesis), συμβολίζεται με $H0$ και είναι μια δήλωση για τον πληθυσμό. Επίσης διατυπώνεται μια εναλλακτική υπόθεση για τον πληθυσμό (alternative hypothesis), η οποία συμβολίζεται με H1, και συνήθως είναι το «αντίθετο» της μηδενικής υπόθεσης. Τέλος επιλέγεται και μια στάθμη σημαντικότητας (significance level, α) η οποία εκφράζει την πιθανότητα απόρριψης της μηδενικής υπόθεσης ενώ αυτή είναι αληθής. Στη συνέχεια υπολογίζεται ένα στατιστικό ελέγχου το οποίο «συμπυκνώνει» τα δεδομένα του δείγματος σε έναν και μοναδικό αριθμό. Το στατιστικό που χρησιμοποιείται εξαρτάται από τον τύπο του ελέγχου που πραγματοποιείται (t-test, ANOVA, κ.α.). Στο τελικό στάδιο του ελέγχου υποθέσεων η υπολογιζόμενη τιμή του στατιστικού ελέγχου συγκρίνεται με μια κρίσιμη τιμή, η οποία εξαρτάται από το είδος του στατιστικού και το επίπεδο σημαντικότητας. Αν η τιμή του στατιστικού είναι μεγαλύτερη από την κρίσιμη τιμή τότε η μηδενική υπόθεση απορρίπτεται. Ταυτόχρονα, όσο μεγαλύτερη είναι η τιμή του στατιστικού, τόσο μικρότερη είναι η πιθανότητα η μηδενική υπόθεση να είναι αληθής. Αυτή η τιμή πιθανότητας (probability value ή p-value), συνοψίζει σε έναν αριθμό, το πόσο μεγάλη ή μικρή συμφωνία υπάρχει μεταξύ των δεδομένων και της μηδενικής υπόθεσης. Κατ' αυτόν τον τρόπο σε πειράματα έκφρασης στην υπολογιστική βιολογία για παράδειγμα, μικρές τιμές p-value αποτελούν ισχυρή ένδειξη ότι οι διαφορές που παρατηρούνται στα επίπεδα έκφρασης ενός γονιδίου είναι βιολογικά σημαντικές [53].

1.3.1 Υπεργεωμετρική Κατανομή

Το υπεργεωμετρικό τεστ (Hypergeometric test), το οποίο βασίζεται στην υπεργεωμετρική κατανομή χρησιμοποιείται συχνά στην ανάλυση εμπλουτισμού. Περιγράφει ένα τυχαίο πείραμα με δύο πιθανά αποτελέσματα (επιτυχία - αποτυχία) σε πεπερασμένο πληθυσμό, που επαναλαμβάνεται n φορές χωρίς επανατοποθέτηση. Για να πραγματοποίησει κάποιος μια ανάλυση που βασίζεται στην υπεργεωμετρική κατανομή, χρειάζεται να ορίσει ένα σύμπαν-σύνολο γονιδίων και μια λίστα με τα γονίδια που τον ενδιαφέρουν. Στη συνέχεια εξετάζεται εάν τα γονίδια της λίστας λαμβάνουν μέρος σε υποσύνολα του επιλεγμένου «σύμπαντος». Είναι λογικό λοιπόν ότι η επιλεγμένη λίστα γονιδίων καθορίζει σε μεγάλο βαθμό τα αποτελέσματα της ανάλυσης, αλλά και το σύμπαν έχει μεγάλη επίδραση στα συμπεράσματα. [54].

Για την καλύτερη κατανόηση παρατίθεται το εξής παράδειγμα [55]: Σε ένα κουτί περιέχονται α άσπρες σφαίρες και β μαύρες σφαίρες και εξάγουμε διαδοχικά, τη μια μετά την άλλη, τυχαία ν σφαίρες και χωρίς επανάθεση (Εικόνα 1.9Α). Έστω Χ ο αριθμός των άσπρων σφαιρών που επιλέχθηκαν στο τυχαίο δείγμα μας (Εικόνα 1.9Β). Η πιθανότητα να είναι λευκή η σφαίρα

είναι α/ν. Αν η σφαίρα είναι λευκή, τότε το κουτί περιέχει πλέον ν-1 σφαίρες εκ των οποίων οι α-1 είναι άσπρες και οι β δεν είναι. Η πιθανότητα επιλογής μιας επόμενης λευκής σφαίρας είναι τώρα α-1/ν-1 , έχει δηλαδή αλλάξει από πριν καθώς κάθε σφαίρα που επιλέγεται δεν επανατοποθετείται στο κουτί πριν επιλεχθεί η επόμενη. Στη θεωρία των πιθανοτήτων μια τέτοια διαδικασία ονομάζεται "δειγματοληψία χωρίς επανάθεση". Η κατανομή της τυχαίας μεταβλητής Χ ονομάζεται Υπεργεωμετρική κατανομή με παραμέτρους α, β και ν και συμβολίζεται με h(ν, α, β).

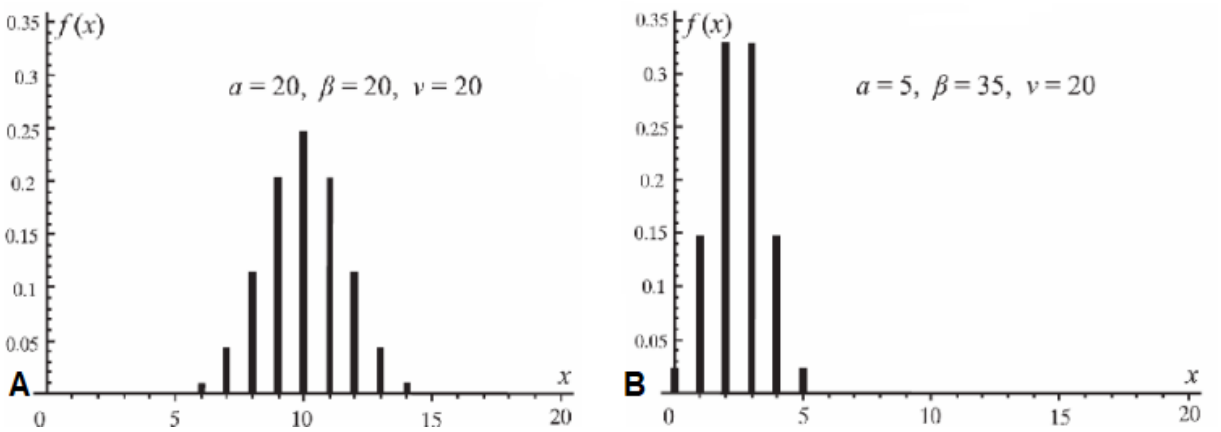


Εικόνα 1.9: Παράδειγμα εφαρμογής υπεργεωμετρικής κατανομής. Α) Κουτί με α άσπρες σφαίρες και β μαύρες σφαίρες και εξάγουμε διαδοχικά, τη μια μετά την άλλη, τυχαία ν σφαίρες και χωρίς επανάθεση. Β) Χ ο αριθμός των άσπρων σφαιρών που επιλέχθηκαν στο τυχαίο δείγμα μας. (Πηγή Εικόνας [55])

'Ετσι, η συνάρτηση της πιθανότητας της υπεργεωμετρικής κατανομής με παραμέτρους α, β και ν που προκύπτει είναι:

$$f(x) = P(X = x) = \frac{\binom{\alpha}{x}\binom{\beta}{\nu - x}}{\binom{a + \beta}{\nu}}, \quad x = \max(0, \nu - \beta), \ldots, \min(\nu, \alpha).$$

Γραφικά, η υπεργεωμετρική κατανομή για συγκεκριμένες τιμές των παραμέτρων α, β και ν παρουσιάζεται στην Εικόνα 1.10.



Εικόνα 1.10: Γραφική απεικόνιση υπεργεωμετρικής κατανομής για συγκεκριμένες τιμές των h(α,β,ν). (Πηγή Εικόνας [55])

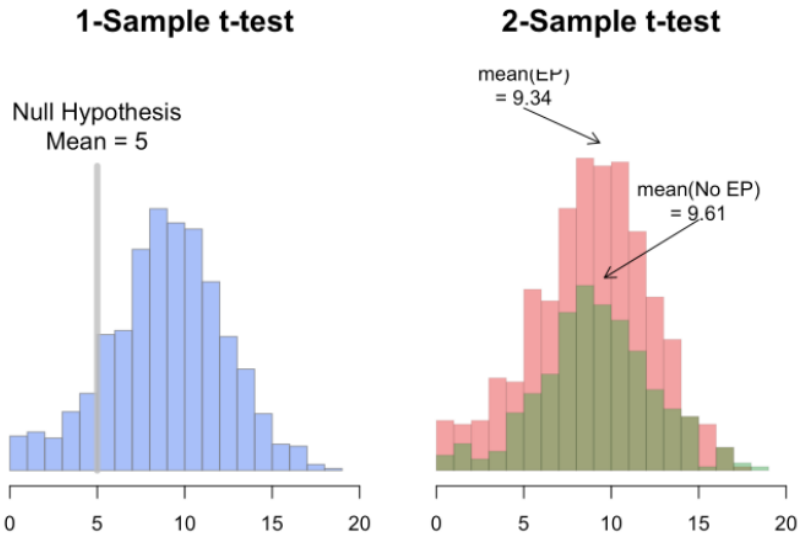
1.3.2 Kolmogorov-Smirnov

Ο έλεγχος Kolmogorov-Smirnov [52] είναι ένα στατιστικό τεστ «καλής προσαρμογής» (goodness of fit test) το οποίο εξετάζει εάν τα πειραματικά δεδομένα είναι συνεπή σε μια αθροιστική συνάρτηση κατανομής (cumulative distribution function). Η εκάστοτε συνάρτηση κατανομής που εξετάζεται τίθεται ως μηδενική υπόθεση. Ο έλεγχος Kolmogorov-Smirnov χρησιμοποιείται επίσης και για τη σύγκριση μεταξύ δύο δειγμάτων. Στην περίπτωση αυτή η μηδενική υπόθεση είναι ότι τα δεδομένα ακολουθούν την ίδια κατανομή.

1.3.3 t-test

Ο t-έλεγχος (t-test ή Student's test) χρησιμοποιείται για τη σύγκριση της μέσης τιμής μεταξύ δύο πληθυσμών και κατ' επέκταση για τον έλεγχο της ομοιότητάς τους. Πιο αναλυτικά, δεδομένων δύο δειγμάτων τιμών $X1$ και X2 , ο t-έλεγχος υπολογίζει ένα μέγεθος t το οποίο είναι μικρό αν οι αντίστοιχοι μέσοι όροι $\mu 1$ και $\mu 2$ των δυο πληθυσμών είναι παραπλήσιοι. Όσο μεγαλύτερο είναι το t τόσο μικρότερη είναι η πιθανότητα οι δύο μέσοι όροι να ταυτίζονται, δηλαδή έχουν στατιστικά σημαντική διαφορά. Υπάρχουν αρκετοί μαθηματικοί τύποι για τον υπολογισμό του στατιστικού t, οι οποίοι εξαρτώνται από το είδος του ελέγχου που πραγματοποιείται (μονοδειγματικός έλεγχος ή έλεγχος μεταξύ δύο δειγμάτων, κανονικοί ή μη πληθυσμοί, γνωστές ή άγνωστες διακυμάνσεις κ.λ.π). Ειδικότερα, μπορούμε να εφαρμόσουμε το t-test [56]:

- **Ανεξάρτητα δείγματα:** Τα δύο δείγματα παίρνονται ανεξάρτητα και τυχαία από τους δύο πληθυσμούς Α και Β. Προϋποθέτει ίσες διακυμάνσεις στους δύο πληθυσμούς
- **Συσχετισμένα (ανά ζεύγη) δείγματα:** Δημιουργούμε ζευγαρωτές παρατηρήσεις. Δηλαδή σε κάθε πειρατική μονάδα (φυτό π.χ.) έχουμε δύο μετρήσεις, την Α και την Β.



Εικόνα 1.11: T κατανομή για ένα δείγμα (Αριστερά) και δύο δείγματα (δεξιά)

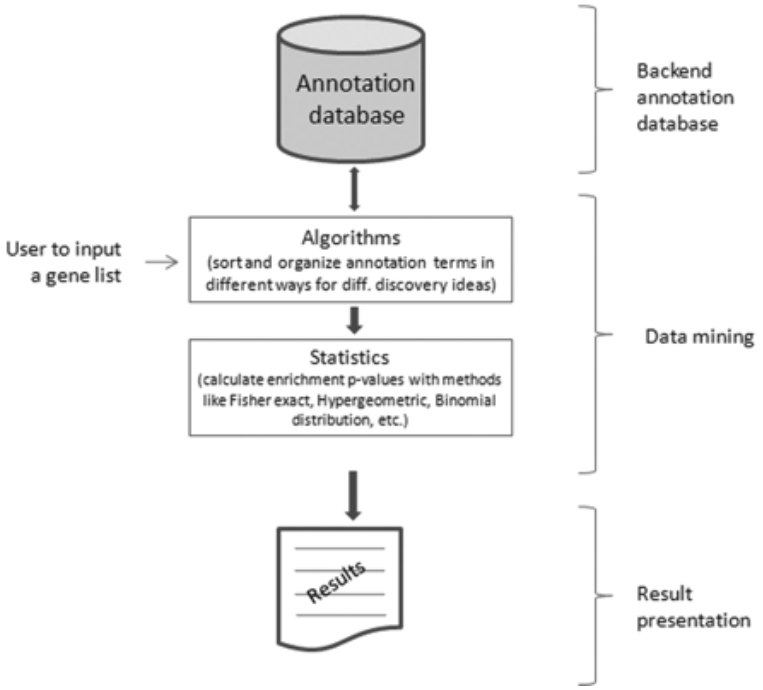
1.4 Εργαλεία λειτουργικού εμπλουτισμού

Οι ταχείς ρυθμοί των νέων ανακαλύψεων στους τομείς της γονιδιωματικής, μεταγραφομικής, πρωτεωμικής κτλ. έχουν οδηγήσει στην προβολή της πολυπλοκότητας των βιολογικών αλληλεπιδράσεων εστιάζοντας έτσι στην έννοια του συστήματος. Όπως έχει ήδη αναφερθεί υπάρχει μια τεράστια πληθώρα βάσεων δεδομένων με αστρονομικά ποσά πληροφορίας στον τομέα της βιολογίας, η οποία συνεχίζει να αυξάνεται με ραγδαίους ρυθμούς. Μέσα σε αυτή τό χαοτικό μέγεθος πληροφοριών η βιοπληροφορική προσπαθεί με τη δημιουργία εργαλείων στη εξόρυξη στοχευμένης και ουσιαστικής πληροφορίας. Τα εργαλεία τα οποία αναπτύσσονται συνεχώς πρέπει να είναι κατασκευασμένα κατ'αυτόν τον τρόπο, ώστε να επιτρέπουν τη σωστή διαχείριση και ανάκτηση των πληροφοριών από τις βάσεις δεδομένων για την καλύτερη ερμηνεία των πειραματικών αποτελεσμάτων.

Ανεξάρτητα από τα ιδιαίτερα χαρακτηριστικά κάθε εργαλείου που επιτελεί κατά βάση ανάλυση εμπλουτισμού, η γενική διαδικασία των εργαλείων αυτών μπορεί να περιγραφεί ότι έχει τρία κύρια επίπεδα (Εικόνα 1.12):

- βάσεις δεδομένων σχολιασμού ως πηγή πληροφοριών,
- εξόρυξη δεδομένων (αλγόριθμος και στατιστικά) και
- παρουσίαση αποτελεσμάτων (διεπαφή και εξερεύνηση).

Επιπλέον, οι βάσεις δεδομένων και άρα οι πληροφορίες που περιέχουν μπορεί να διαφέρουν πολύ από εργαλείο σε εργαλείο. Δεν είναι ασυνήθιστο για τους χρήστες να δοκιμάζουν πολλαπλά εργαλεία με παρόμοια αναλυτική ικανότητα για το ίδιο σύνολο δεδομένων για να πετύχουν τα μέγιστα ικανοποιητικά αναλυτικά αποτελέσματα [45], [57].



Εικόνα 1.12: Η βασική δομή των εργαλείων εμπλουτισμού. Παρόλο που τα εργαλεία ανάλυσης εμπλουτισμού έχουν διαφορετικά χαρακτηριστικά, μπορούν γενικά να περιγραφούν με βάση τρία κύρια επίπεδα: βάσεις δεδομένων σχολιασμού ως πηγή πληροφοριών (backend annotation database), εξόρυξη δεδομένων (data mining), και παρουσίαση αποτελεσμάτων (results presentation). Κάθε ένα από τα επίπεδα, και όχι μόνο οι στατιστικές μέθοδοι, επηρεάζει σε μεγάλο βαθμό τα αναλυτικά αποτελέσματα. (Πηγή Εικόνας: [45])

1.4.1 gProfiler

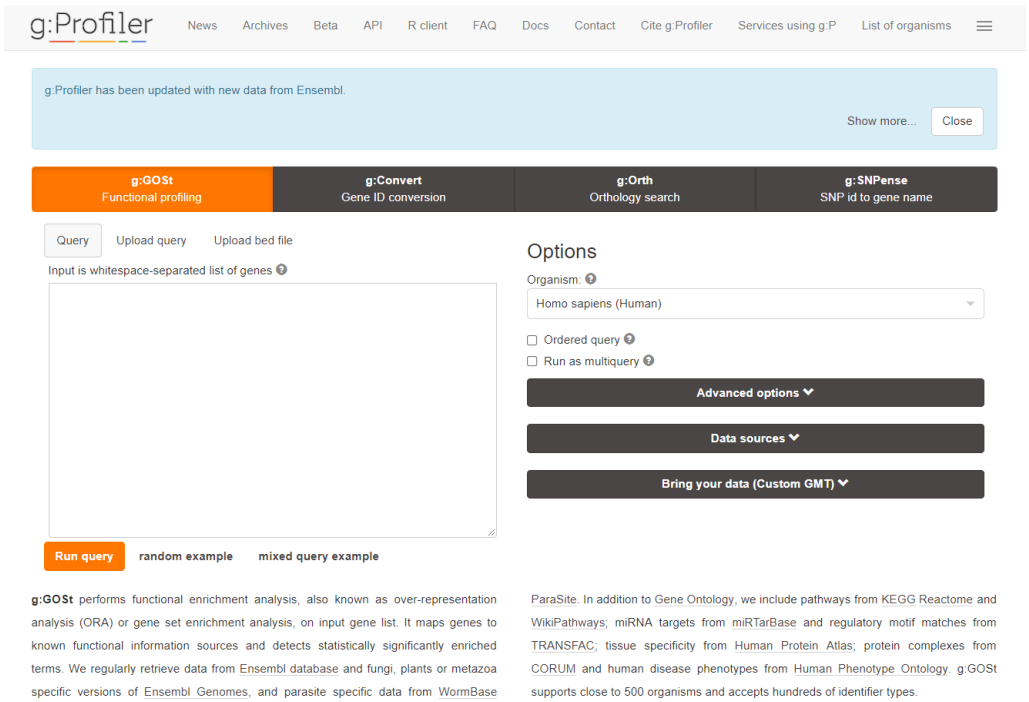
Το g:Profiler [58] είναι ένα σύνολο εργαλείων με δωρεάν πρόσβαση για την ανάλυση λειτουργικού εμπλουτισμού σε βιολογικές κατηγορίες εύρεση βιολογικών κατηγοριών εμπλουτισμένων σε λίστες γονιδίων (g:GOSt), τη μετατροπή μεταξύ αναγνωριστικών IDs (g:Convert), την ορθόλογη χαρτογράφηση γονιδίων μεταξύ οργανισμών (g:Orth) καθώς και τη χαρτογραφηση ανθρώπινων SNP rs-codes (π.χ. rs7961894) σε ονόματα γονιδίων (g:SNPense) .

Ειδικότερα, μέσω του g:GOSt σε μια ή πολλαπλές λίστες γονιδίων εισόδου μπορεί να πραγματοποιηθεί ανάλυση λειτουργικού εμπλουτισμού, γνωστή και ως ανάλυση εμπλουτισμού γονιδίων (GSEA). Τα γονίδια χαρτογραφούνται με βάση γνωστές βάσεις δεδομένων και προκύπτουν κατηγορίες, μονοπάτια κλπ με στατιστικά σημαντικά εμπλουτισμένους όρους. Τα δεδομένα ανανεώνονται από τη βάση δεδομένων Ensembl Genomes και από ειδικά δεδομένα για παράσιτα από το WormBase ParaSite. Εκτός από την οντολογία των γονιδίων για τη λειτουργική ενίσχυση χρησιμοποιεί μεταβολικά μονοπάτια από τις βάσεις KEGG, Reactome και WikiPathways, miRNA στόχους από την miRTarBase, ρυθμιστικά μοτίβα από το TRANSFAC, εξειδίκευση ιστού από το Human Protein Atlas, πρωτεϊνικά σύμπλοκα από το CORUM και φαινότυπους ανθρώπινης νόσου από το Human Phenotype Oncology. Το g:GOSt υποστηρίζει

περίπου 714 οργανισμούς (g:Profiler version e104_eg51_p15_3922dba) και δέχεται εκατοντάδες τύπους αναγνωριστικών.

Η ανάλυση λειτουργικού εμπλουτισμού συνοδεύεται και από άλλα εργαλεια. Πιο συγκεκριμένα, μέσω του g:Convert μπορεί να επιτευχθεί μετατροπή μεταξύ διαφόρων γονιδίων, πρωτεϊνών, ανιχνευτών μικροσυστοιχιών και πολλών άλλων τύπων παρέχοντας περίπου 100 τύπους αναγνωριστικών που συμπεριλαμβάνουν τα αναγνωριστικά Ensembl, Refseq, Illumina, Entrezgene και Uniprot. Το g:Orth μπορεί να χρησιμοποιηθεί για τη μετάφραση αναγνωριστικών γονιδίων μεταξύ οργανισμών. Παρέχει ορθόλογες χαρτογραφήσεις γονιδίων με βάση τις πληροφορίες που ανακτώνται από τη βάση δεδομένων Ensembl. Τέλος, το g:SNPense χαρτογραφεί έναν κατάλογο ανθρώπινων SNP rs-codes (π.χ. rs7961894) σε ονόματα γονιδίων και λαμβάνει χρωμοσωμικές συντεταγμένες και προβλεπόμενα αποτελέσματα παραλλαγής. Η χαρτογράφηση είναι ενεργοποιημένη μόνο για παραλλαγές που επικαλύπτονται με τουλάχιστον ένα γονίδιο Ensembl που κωδικοποιεί πρωτεΐνη. Όλα τα υποκείμενα δεδομένα ανακτώνται από τα δεδομένα της βάσης Ensembl.

Το g:Profiler είναι προσβάσιμο, είτε μέσω προγράμματος περιήγησης, είτε μέσω API, καθώς και μέσω της βιβλιοθήκης gprofiler2 (R package) και σε Python. Επιπλέον, το πακέτο R gprofiler2 παρέχει τις ίδιες διαδραστικές απεικονίσεις και αναλύσεις με αυτές που είναι διαθέσιμες στη διαδικτυακή έκδοση. Ταυτόχρονα, υποστηρίζει προσαρμοσμένα αρχεία GMT για ανάλυση. Τα αποτελέσματα (Εικόνα 1.14) παρουσιάζονται με μορφή Πίνακα που ενσωματώνουν ραβδόγραμμα με βάση το $-\log_{10}(p_{adj})$ και θερμικό χάρτη για την αντιστοιχία γονιδίων στις κατηγορίες που προέκυψαν, καθώς και διαγράμματα Manhattan.



Εικόνα 1.13: Διεπαφή χρήστη g:Profiler.

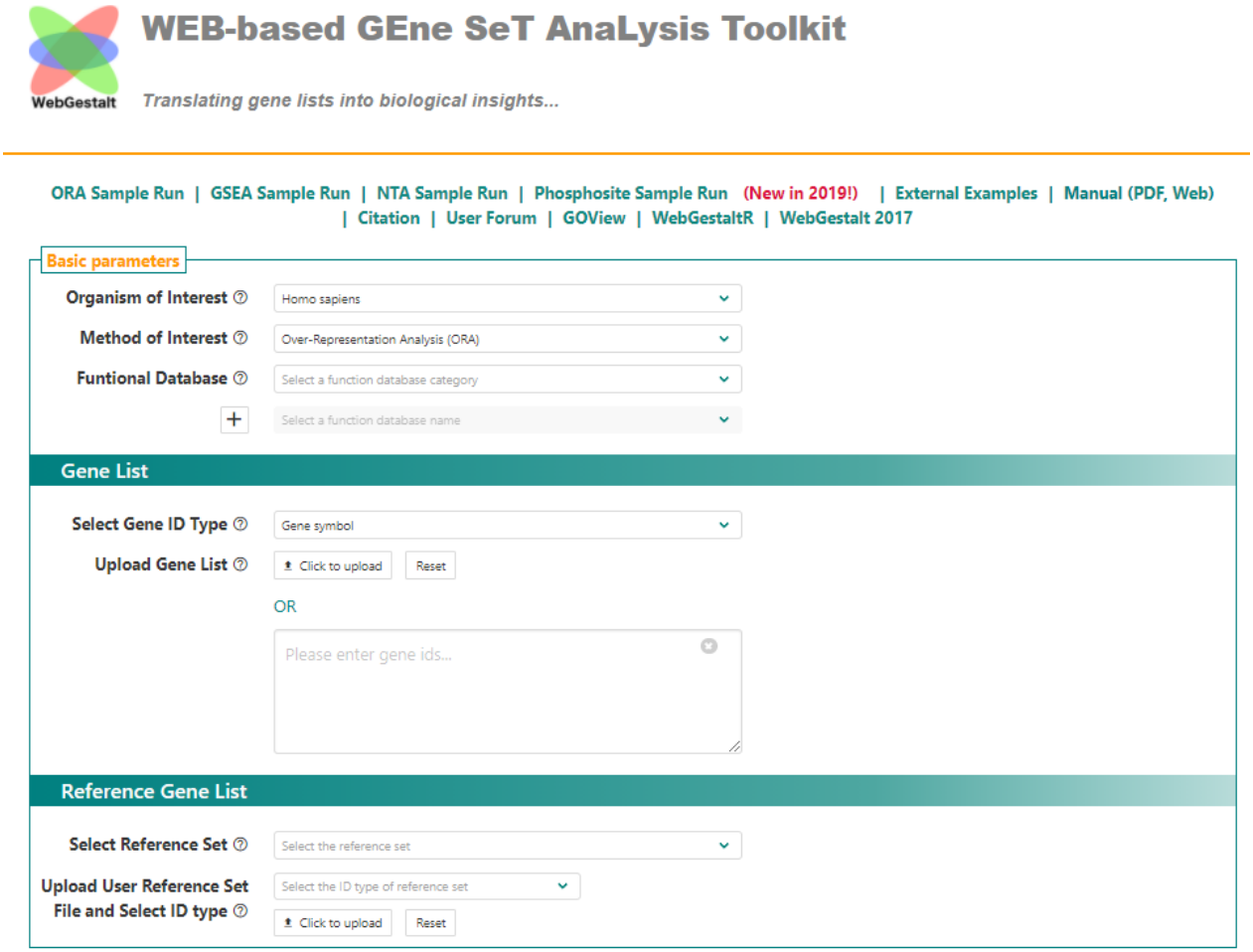
Εικόνα 1.14 : Πίνακας αποτελεσμάτων λειτουργικής ανάλυσης εμπλουτισμού και οπτικοποίηση αποτελεσμάτων με τη χρήση Manhattan Plot.

1.4.2 WebGestalt

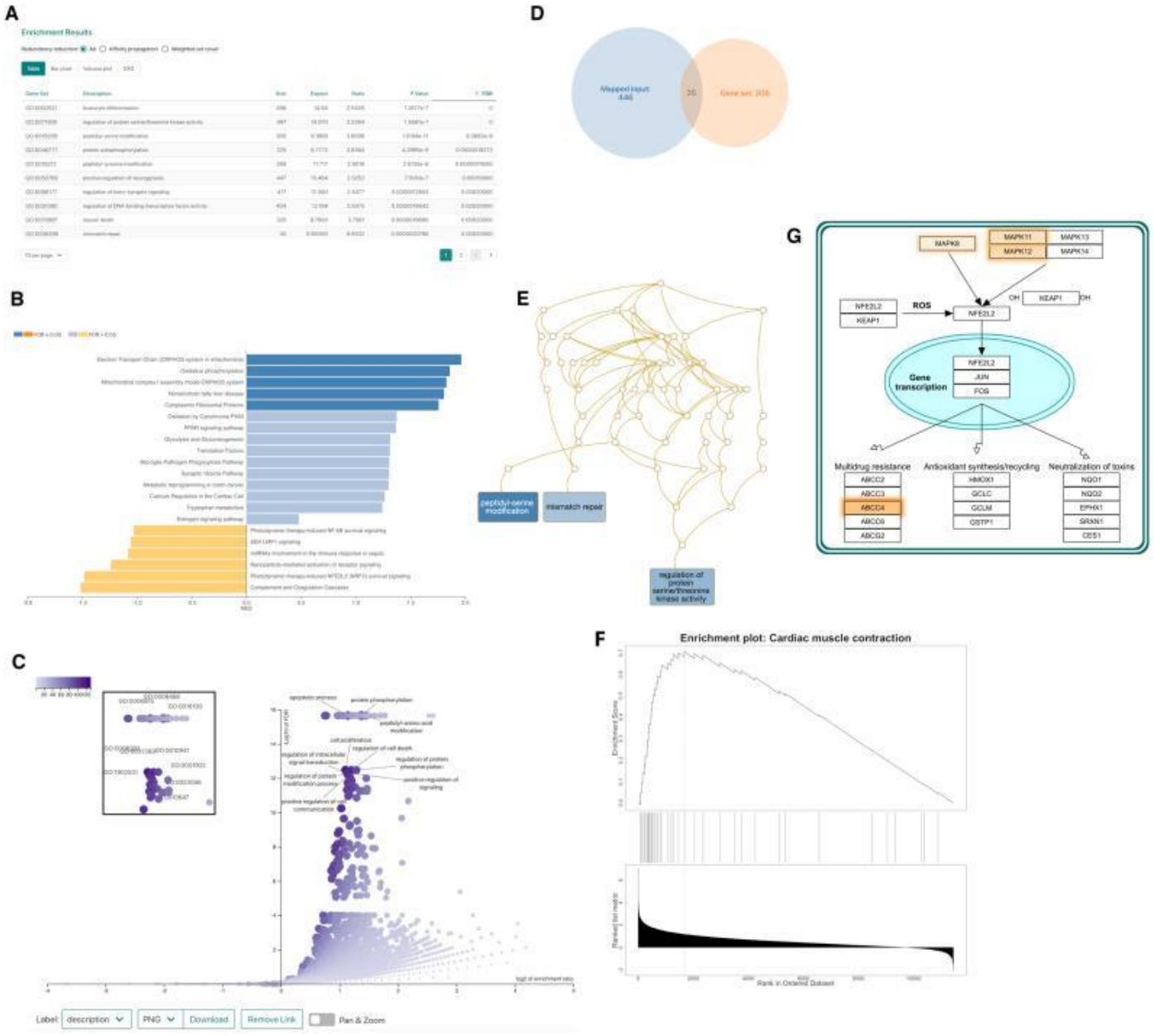
Το WebGestalt [59] αποτελεί ένα web-based εργαλείο αναφορικά με τη λειτουργική ανάλυση εμπλουτισμού σε σύνολα γονιδίων προσφέροντας 3 τύπους ανάλυσεων. Ειδικότερα, παρέχει Over-Representation Analysis (ORA), Gene Set Enrichment Analysis (GSEA),καθώς και Network Topology-based Analysis (NTA) (Εικόνα 1.15). Πρόσφατα έχουν συμπεριληφθεί δεδομένα φωσφατασών για ανάλυση εμπλουτισμού στόχου κινάσης.

Η τρέχουσα έκδοση του WebGestalt (WebGestalt 2019) υποστηρίζει 12 οργανισμούς και 354 αναγνωριστικά γονιδίων από διάφορες βάσεις δεδομένων και τεχνολογικές πλατφόρμες. Με

αυτό τον τρόπο καλύπτονται λειτουργικές κατηγορίες σε διάφορες βιολογικές περιοχές όπως γονιδιακές οντολογίες, μονοπάτια, δίκτυα, συσχέτιση γονιδίων με φαινοτύπους, ασθένειες και φάρμακα καθώς και θέσεις χρωμοσωμάτων, οδηγώντας σε 321,251 στο σύνολο λειτουργικές κατηγορίες. Επιπλέον χαρακτηριστικά του συστήματος αποτελούν οι χάρτες μονοπατιών και η ιεραρχική οπτικοποίηση δικτύων και οντολογιών φαινοτύπων. Για την ανάλυση στην παρούσα έκδοση οι πληροφορίες ανακτώνται από διάφορες βάσεις δεδομένων: Gene Ontology, KEGG, WikiPathways, Reactome, BioGRID, MSigDB, CORUM, Human Phenotype Ontology, DisgeNET, GLAD4U, DrugBank, Chromosomal location, BioGRID, RegPhos, PTMsigDB, Mammalian Phenotype Ontology, NCBI Gene παρέχοντας 8 γενικότερες κατηγορίες επιλογής (geneontology, pathway, network, disease, drug, phenotype, chromosomal Location, community-contributed). Τα αποτελέσματα παρουσιάζονται με μορφή διαδραστικών πινακων, ραβδογραμμάτων, volcano plot, networks, διαγραμμάτων Venn κ.α (Εικόνα 1.16). Είναι προσβάσιμο, είτε μέσω προγράμματος περιήγησης, είτε μέσω API, καθώς και μέσω της βιβλιοθήκης στην R (R package).



Εικόνα 1.15: Διεπαφή χρήστη Webgestalt. Ο χρήστης μπορεί να διαλέξει μεταξύ των: Over-Representation Analysis (ORA), Gene Set Enrichment Analysis (GSEA),καθώς και Network Topology-based Analysis (NTA) και να πραγματοποιήσει την επιθυμητή ανάλυση σε 12 οργανισμούς.

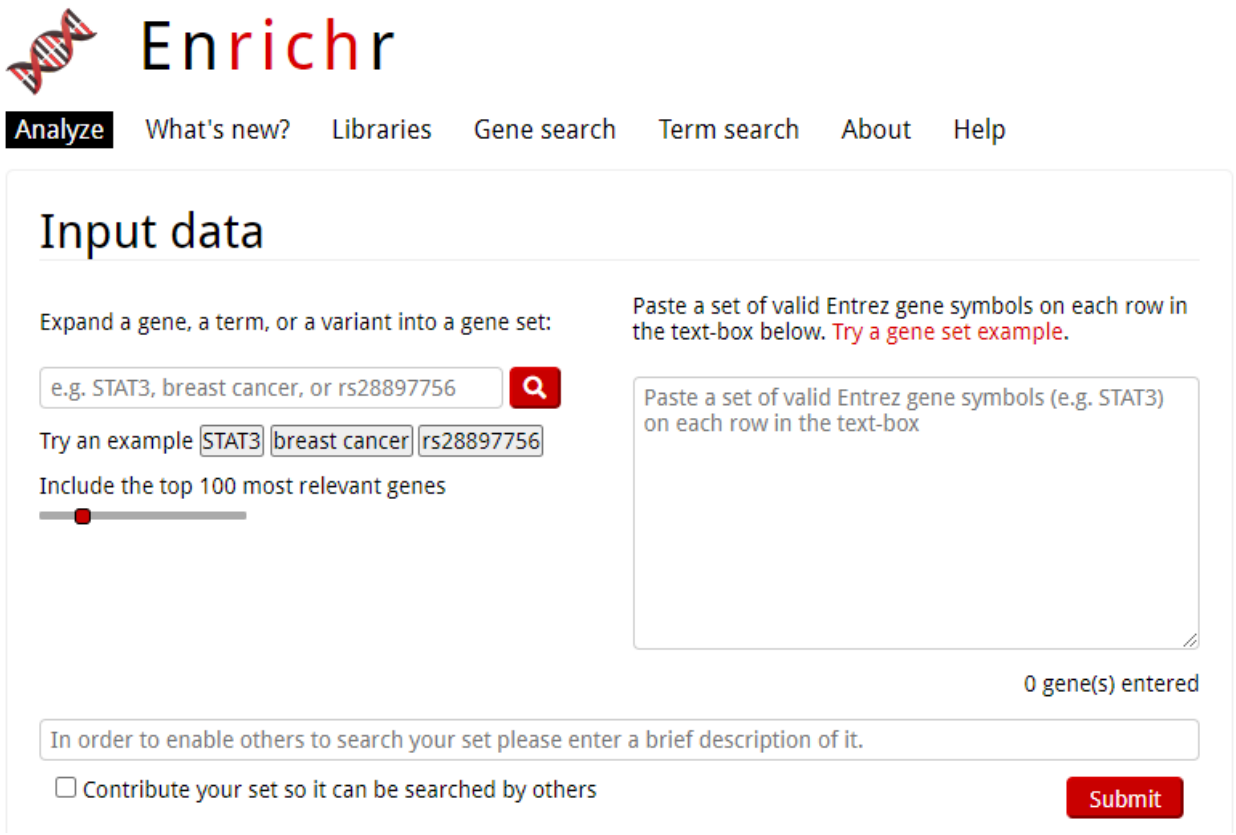


Εικόνα 1.16: Αποτελέσματα λειτουργικής ανάλυσης εμπλουτισμού μέσω WebGestalt 2019. (A) Περιληπτικός πίνακας στατιστικά σημαντικών αποτελεσμάτων με δυνατότητες επιλογή φίλτρων και υποσυνόλων. (B) Ραβδόγραμμα με βάση τις τιμές του σκορ λειτουργικού εμπλουτισμού. (C) Διάγραμμα ηφαιστείου (volcano Plot). (D) Απεικόνιση δικτύου. (E) Το διάγραμμα Venn για την επικάλυψη μεταξύ γονιδίων στη λίστα είσοδο και στα δεδομένα αναφοράς. (F) διάγραμμα εμπλουτισμού GSEA. (G) Προβολή διαδρομής του WikiPathways με επισημασμένα γονίδια. (Πηγή εικόνας [59])

1.4.3 Enrichr

Το βιοπληροφορικό εργαλείο EnrichR [60], [61] χρησιμοποιείται για την ανάλυση λειτουργικού εμπλουτισμού γονιδίων και περιέχει μια μεγάλη συλλογή γονιδιακών συνόλων και βιβλιοθηκών γονιδιακών συνόλων με σκοπό την πραγματοποίηση τέτοιων αναλύσεων. Το Enrichr

(Εικόνα 1.17) στην παρούσα έκδοση υποστηρίζει 377,065 σχολιασμένα σύνολα γονιδίων οργανωμένα σε 190 βιβλιοθήκες. Η μεγάλη συλλογή σχολιασμένων γονιδίων στο Enrichr διευκολύνει τον εμπλουτισμό γονιδιακών συνόλων σε κατηγορίες όπως φάρμακα, ασθένειες, παρενέργειες και άλλους φαινότυπους και βιολογικές διεργασίες. Βάζοντας μία λίστα γονιδίων, επιστρέφει επιγενωμικά, microRNA, μοριακά μονοπάτια και δεδομένα οντολογιών από πολλές βάσεις δεδομένων όπως: GO, KEGG, ChIP-x Enrichment Analysis (ChEA), ENCODE, Connectivity Map (CMAP) κλπ. Παρέχεται η πρόσβαση και μέσω API. Το Enrichr παρέχει τα αποτελέσματα της ανάλυσης εμπλουτισμού σε διάφορες μορφές, με διαδραστικές απεικονίσεις. Οι οπτικοποιήσεις Enrichr υλοποιούνται με τη βιβλιοθήκη JavaScript Document-Driven Documents (D3) για τη δημιουργία διαδραστικών γραφημάτων και ειδικότερα χρησιμοποιούνται γραφήματα ράβδων και οι θερμικοί χάρτες χρησιμοποιώντας το Clustergrammer.

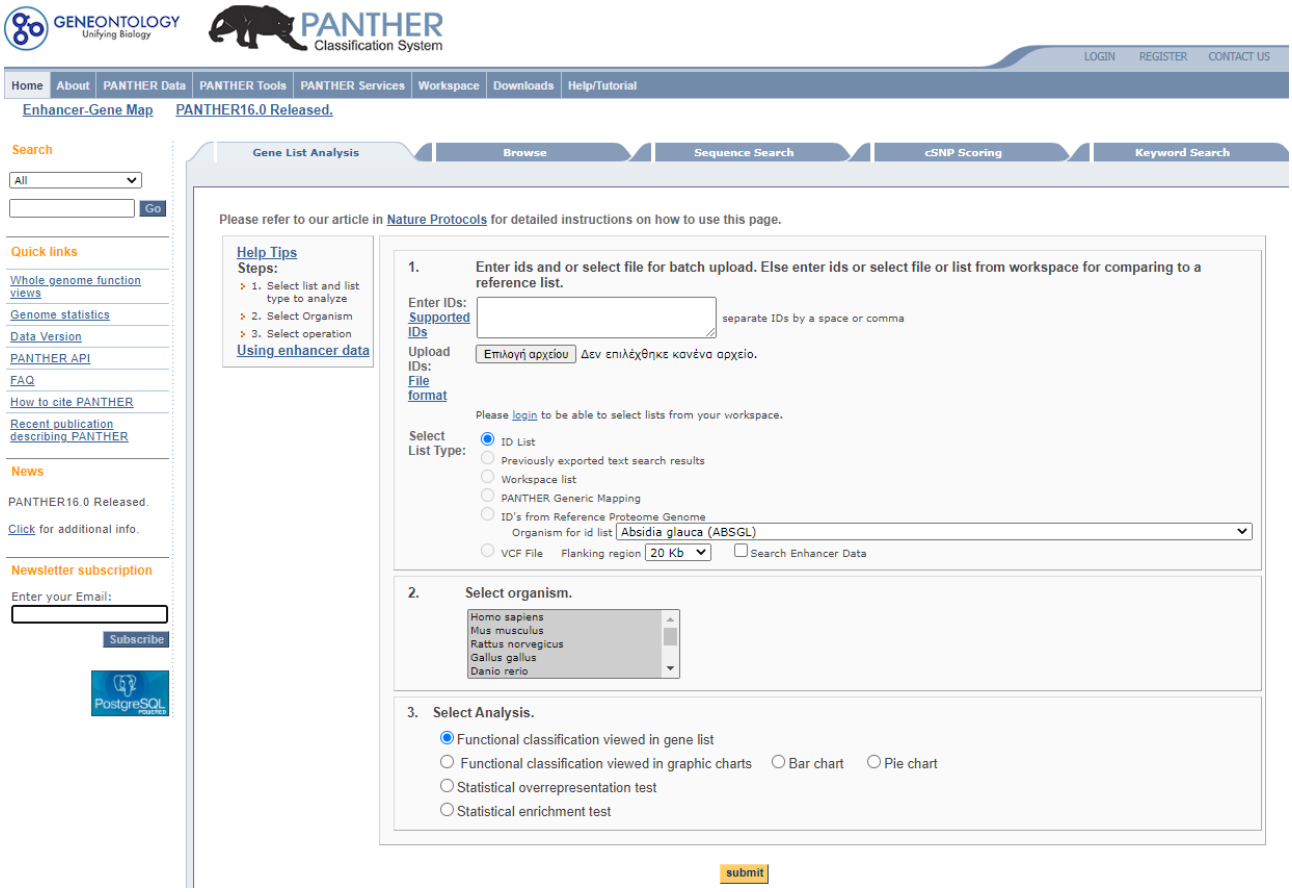


Εικόνα 1.17: Διεπαφή χρήστη στο εργαλείο Enrichr.

1.4.4 PANTHER

Panther (**P**rotein **AN**alysis **TH**rough **E**volutionary **R**elationships) [29]για την ανάλυση εμπλουτισμού με όρους γονιδιακής οντολογίας. Το PANTHER είναι μια πλατφόρμα που αποτελεί ένα σύστημα κατηγοριοποίησης και ταξινόμησης δεδομένων από οικογένειες πρωτεϊνών και γονιδίων καθώς και των λειτουργικά συνδεδεμένων υπο-οικογενειών τους, που χρησιμοποιούνται για την ομαδοποιήση και την αναγνώριση της λειτουργίας των πρωτεϊνών. Οι πρωτεΐνες μπορούν να ομαδοποιηθούν βάσει οικογένειας, μοριακής λειτουργίας τους, βιολογικών διεργασιών στις

οποίες συμμετέχουν, την περιοχή του κυττάρου ή εξωκυττάριου χώρου στην οποία βρίσκονται και τα μονοπάτια στα οποία συμμετέχουν. Οι εμπλουτισμένοι όροι μπορεί να απεικονιστούν με ιεραρχική σειρά βάσει των σχέσεων της GO. Έτσι, συγγενικοί όροι απεικονίζονται πιο κοντά, διευκολύνοντας έτσι την ερμηνεία των βιολογικών αποτελεσμάτων που προκύπτουν από την ανάλυση εμπλουτισμού. Τα GO annotations στο PANTHER ανανεώνονται κάθε μήνα. Υποστηρίζει μια πληθώρα αναγνωριστικών (Ensemble gene, Ensemble protein, Ensemble transcript. Entrez gene id, gene symbol,NCBI, HGNC, International protein index, NCBI UniGene, UniProt, UniProt). Η τρέχουσα έκδοση (Εικόνα 1.18) περιέχει 15635 πρωτεινικές οικογένειες, 142 γονιδιώματα, 2620819 γονίδια και 177 μονοπάτια με 3092 συστατικά και 5996 βιβλιογραφικές αναφορές μονοπατιών. Τέλος, περιέχει 47212 όρους GO. Η πρόσβαση εκτός από την διαδικτυακή διεπαφή μπορεί να γίνει και μέσω API.

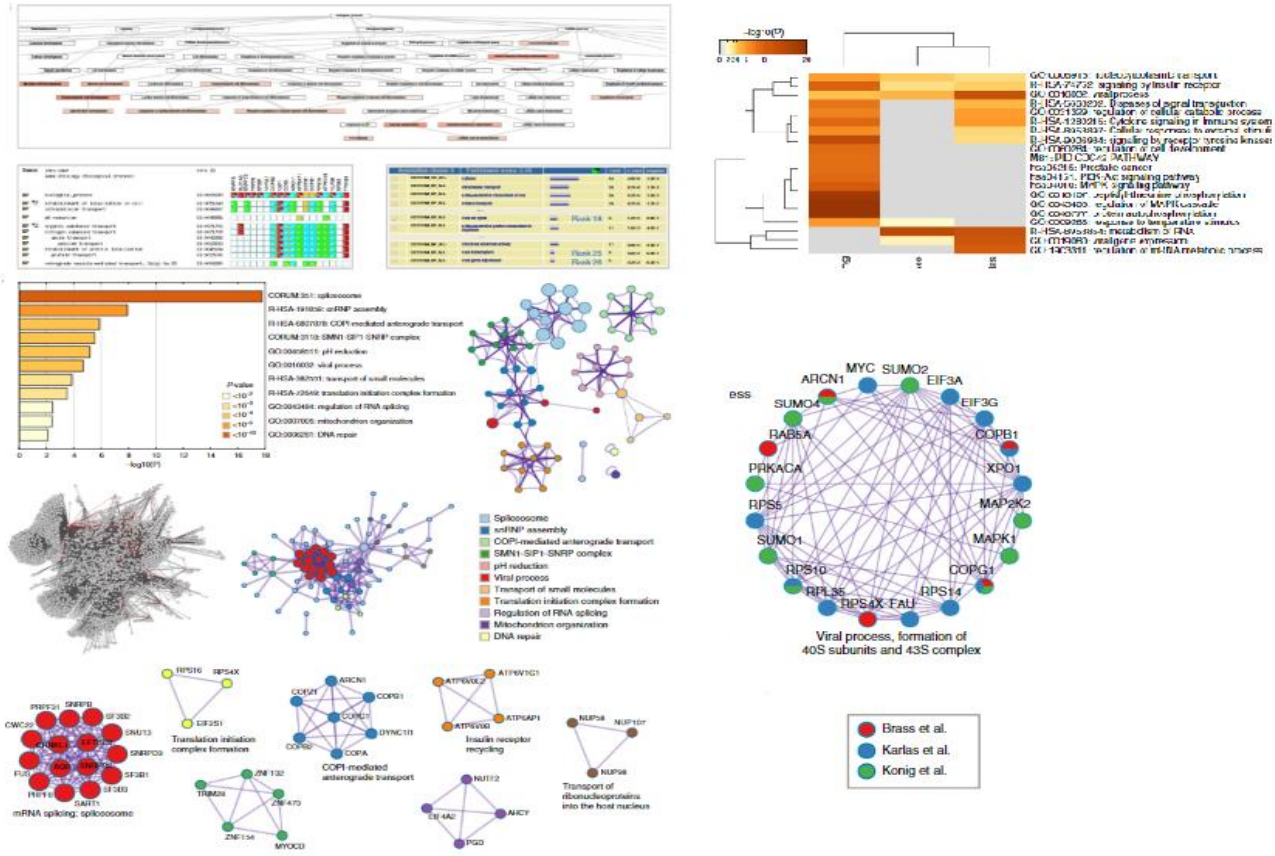


Εικόνα 1.18: Διεπαφή χρήστη για την ανάλυση εμπλουτισμού στο Panther.

1.4.5 Metascape

Το Metascape [62] ενσωματώνει την ανάλυση εμπλουτισμού, την ανάλυση πρωτεϊνικών συμπλεγμάτων καθώς και τη μετα-ανάλυση πολλαπλών λιστών. Υποστηρίζει 10 είδη οργανισμών και ειδικότερα: *H. sapiens, M. musculus, R. norvegicus, D. rerio, D. melanogaster, C. elegans, S. cerevisiae, A. thaliana, S. pombe* και *P. falciparum*. Η προεπιλογή για όλα τα αναγνωριστικά γονιδίων εισόδου είναι να μετατρέπονται σε ανθρώπινα ορθολόγια για ανάλυση σχολιασμού και εμπλουτισμού. Ο χρήστης όμως μπορεί να αλλάξει αυτή την επιλογή. Οι βάσεις από τις οποίες

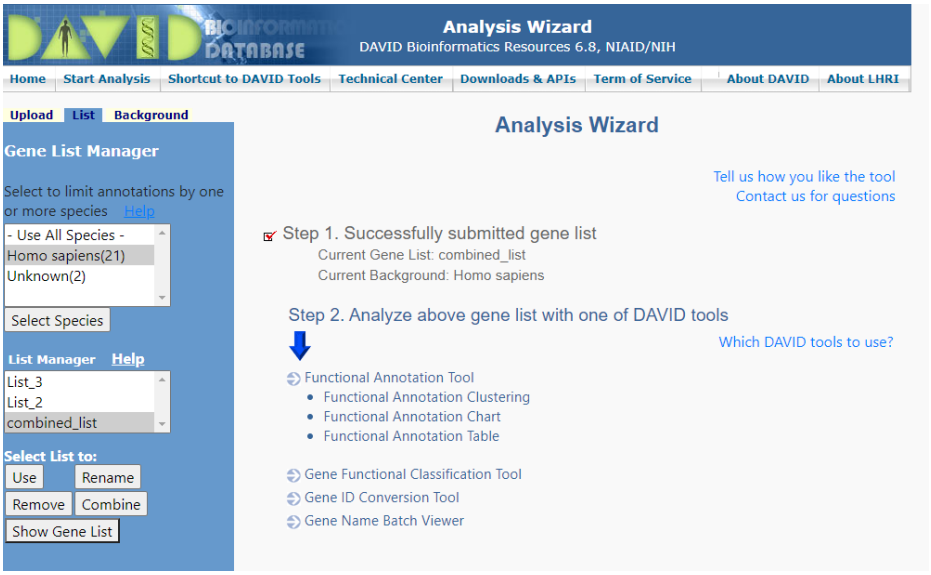
γίνεται εξόρυξη της πληροφορίας είναι οι KEGG Pathway, GO Biological Processes, Reactome Gene Sets, Canonical Pathways, CORUM, TRRUST, DisGeNET, PaGenBase, Transcription Factor Targets, WikiPathways και COVID. Ανανεώνεται σε μηνιαία βάση και χρησιμοποιεί το πιό πρόσφατο περιεχόμενο των βάσεων. Τα αναγνωριστικά των γονιδίων που υποστηρίζει είναι τα εξής: Entrez Gene ID, RefSeq (RNA και Proteins), Gene Symbol, Ensembl (Gene, Transcript, Protein), UCSC, UniProt. Εμφανίζει έναν περιορισμό στο πλήθος των επιτρεπτών γονιδίων ανά ανάλυση στα 3000. Τα αποτελέσματα (Εικόνα 1.19) απεικονίζονται με την μορφή ραβδογραμμάτων, δικτύων εμπλουτισμού και πρωτεϊνών κατασκευασμένων με Cytoscape, θερμικών χαρτών και συνοδεύονται από αναλυτικές περιγραφές που μπορούν να ληφθούν σε μορφη συμπιεσμένου φακέλου που συνοδεύεται από παρουσίαση και τα αποτελέσματα σε μορφή πίνακα. Επίσης, παρεχονται Circos plots με τα δεδομένα μετα-ανάλυσης και απεικονίζουν τα γονιδια που αλληλεπικαλύπτονται στους πολλαπλούς γονιδιακούς καταλόγους σε επίπεδο βιολογικών λειτουργιών. Δεν παρέχει πρόσβαση μέσω API.



Εικόνα 1.19: Διάφορες μορφές απεικόνισης αποτελεσμάτων που παρέχονται από το Metascape. Τα αποτελέσματα απεικονίζονται με την μορφή ραβδογραμμάτων, δικτύων εμπλουτισμού και πρωτεϊνών κατασκευασμένων με Cytoscape, θερμικών χαρτών και συνοδεύονται από αναλυτικές περιγραφές που μπορούν να ληφθούν σε μορφη συμπιεσμένου φακέλου που συνοδεύεται από παρουσίαση και τα αποτελέσματα σε μορφή πίνακα. Επίσης, παρεχονται Circos plots με τα δεδομένα μετα-ανάλυσης και απεικονίζουν τα γονιδια που αλληλεπικαλύπτονται στους πολλαπλούς γονιδιακούς καταλόγους σε επίπεδο βιολογικών λειτουργιών. Πηγή Εικόνας [62]

1.4.6 DAVID

Το DAVID [45], [63] είναι ένα εργαλείο για την εύρεση εμπλουτισμένων βιολογικών όρων και των σχετιζόμενων ομάδων γονιδίων. Ταυτόχρονα, ελέγχει για πιθανή ύπαρξη επιπλέον σχετιζόμενων γονιδίων εκτός των γονιδίων του αρχείου εισόδου. Για την απεικόνιση των αποτελεσμάτων χρησιμοποιεί δισδιάστατα διαγράμματα γονιδίων και αντίστοιχων βιολογικών όρων, χάρτες μονοπατιών της βάσης δεδομένων KEGG και αναδρομολογεί το χρήστη σε βιβλιογραφικά δεδομένα. Καθώς δεν μπορεί να δεχθεί σαν όρισμα μια λίστα με πληθώρα διαφορετικών αναγνωριστικών γονιδίων, ή κάποιου αναγνωριστικού που δεν υποστηρίζει, παρέχει τη δυνατότητα εργαλείου μετατροπής της λίστας του χρήστη σε ένα επιθυμητό τύπου αναγνωριστικού. Επίσης, δίνει την δυνατότητα ο χρήστης να κάνει συγχώνευση των αρχείων εισόδου και να δει το περιεχόμενο της νέας λίστας. Η εξόρυξη της πληροφορίας γίνεται από περισσότερες από 40 κατηγορίες αναγνωρισμένων βιολογικών βάσεων (OMIM_DISEASE, TRANSFAC_ID, BIOCARTA, KEGG, PFAM, PROSITE, Gene Ontology, GENBANK κλπ.) όπως όροι οντολογίας γονιδίων, αλληλεπιδράσεις πρωτεϊνών, λειτουργικές επικράτειες, μονοπάτια, βιβλιογραφικές πηγές, ασθένειες, επικράτειες πρωτεϊνών, αλληλεπιδράσεις κλπ. Η παρούσα έκδοση είναι η DAVID 6.8 (Οκτ 2016). Η εφαρμογή είναι διαθέσιμη online, καθώς και μέσω API, αλλά με κάποιους περιορισμούς στο μέγεθος της λίστας (<500 γονίδια) , πλήθος δοκιμών/ημέρα, περιορισμός στο μέγεθος του URL (2048 χαρακτήρες), ενώ για χρήση της εφαρμογής API από άλλα βιοπληροφορικά εργαλεία υπάρχει περιορισμός 400 γονιδίων.

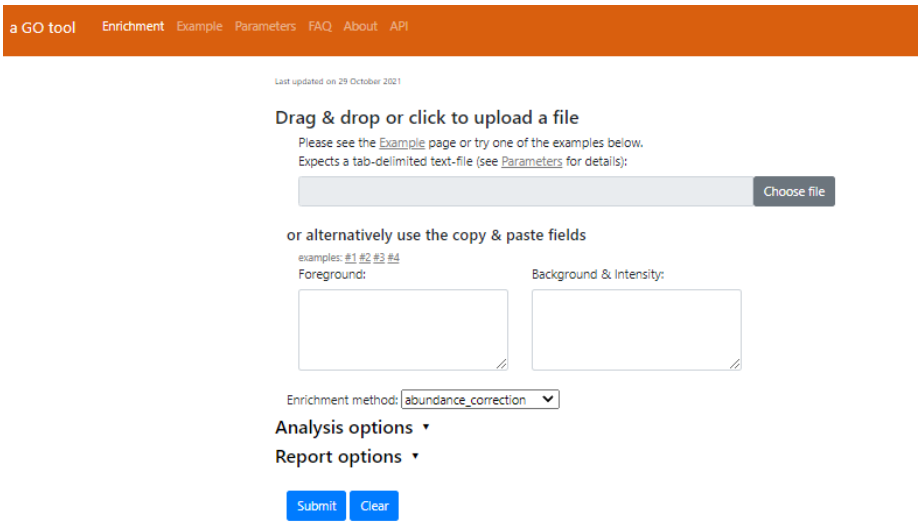


Εικόνα 1.20: Διεπαφή χρήστη για την ανάλυση εμπλουτισμού στο DAVID.

1.4.7 aGOtool

Το aGOtool [64] αποτελεί ένα εργαλείο ανάλυσης λειτουργικού εμπλουτισμού εστιάζοντας στην ανάλυση πρωτεϊνικών δεδομένων (Εικόνα 1.21). Περιέχει διάφορες μεθόδους εμπλουτισμού, μία από αυτές ονομάζεται "abundance_correction", η οποία απευθύνεται ειδικά στις Post Translationally Modified πρωτεΐνες (Μετα-μεταφραστικές τροποποιήσεις πρωτεϊνών). Για την

πραγματοποιήσει των αναλύσεων βασίζεται σε δεδομένα από διάφορες βάσεις δεδομένων και ο χρήστης μπορεί να επιλέξει μεταξύ των εξής βάσεων: GO (molecular function, biological process, cellular component), UniProt, KEGG, PubMed, Reactome, Wiki Pathways, InterPro, PFAM, Brenda Tissues και Diseases. Ανανεώνεται σε μηνιαία βάση. Ταυτόχρονα, εκτός από την ανάλυση λειτουργικού εμπλουτισμού γονιδίων, υποστηρίζει και ανάλυση εμπλουτισμού βιβλιογραφίας. Για να επιτευχθεί αυτό χρησιμοποιεί ένα σύνολο κειμένων όλων των περιλήψεων της PubMed και των άρθρων ανοικτής πρόσβασης του πλήρους κειμένου από το PubMed Central. Οι επιστημονικές δημοσιεύσεις, καθώς και οι περιλήψεις που χρησιμοποιούνται, επεξεργάζονται μέσω των εφαρμογών OnTheFly's [65] ή EXTRACT's [66] ή Named Entity Recognition (NER) [67, p.] σε εβδομαδιαία βάση για τον εντοπισμό βιολογικών οντοτήτων/όρων (γονίδια/πρωτεΐνες, χημικές ενώσεις, οργανισμούς, ιστούς, περιβάλλοντα, ασθένειες, φαινότυπους και όρους GO). Ως αποτέλεσμα, σε όλα τα έγγραφα σχολιάζονται αυτόματα τα γονίδια που αναφέρονται μετατρέποντας κάθε έγγραφο σε «σύνολο γονιδίων». Για την ανάλυση τα αναγνωριστικά πρωτεΐνης μπορούν να παρέχονται ως UniProt Accession (π.χ. P31946), ονόματα καταχώρισης UniProt (UniProt ID) (π.χ. 1433B_HUMAN) ή αναγνωριστικά STRING (π.χ. 9606.ENSP00000361930). Παρέχει δύο τύπους διεπαφών API (ένα προϋπάρχον και ένα για ενσωμάτωση του CytoScape) που επιτρέπουν την πραγματοποίηση αναλύσεων χωρις να χρησιμοποιεί ο χρήστης το περιβάλλον της ιστοσελίδας. Τα αποτελέσματα (Εικόνα 1.22) απεικονίζεται με τη μορφή διαδραστικού Πίνακα και διαγράμματος διασποράς.

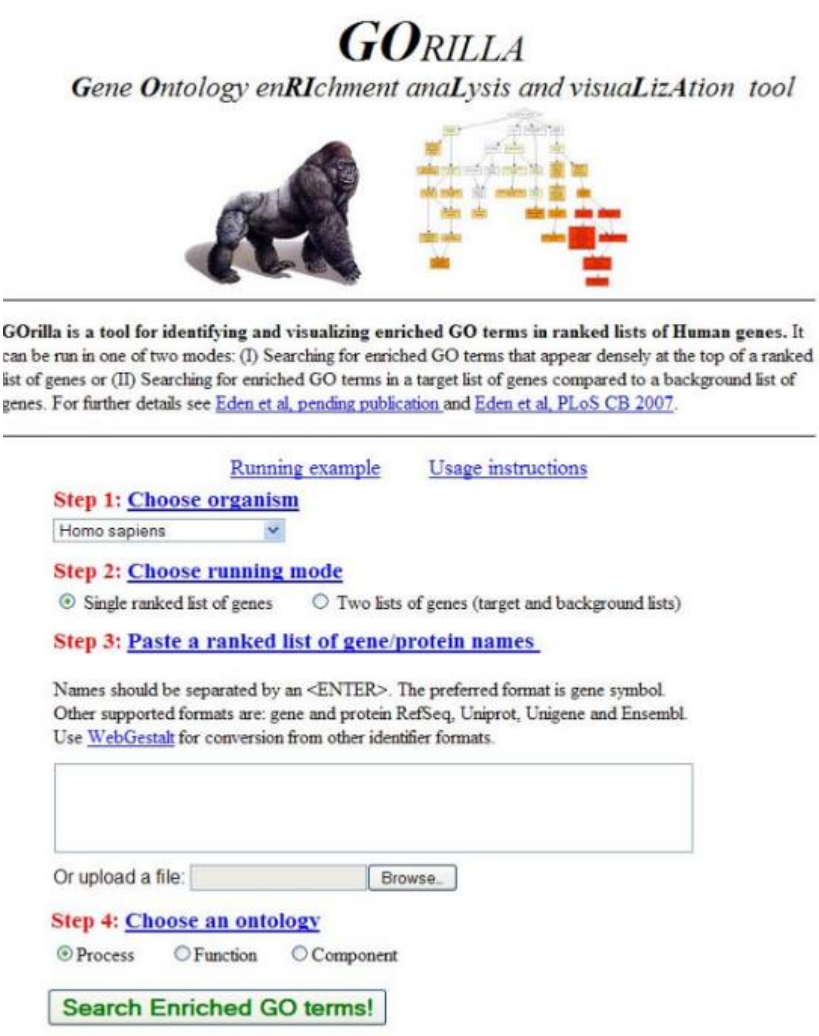


Εικόνα 1.21: Διεπαφή χρήστη aGOtool.

Εικόνα 1.22: Αποτελέσματα λειτουργικής ανάλυσης εμπλουτισμού μέσω aGOtool.

1.4.8 GOrilla

Με στόχο την αποδοτική ανάλυση και επεξεργασία των δεδομένων της GO, έχουν αναπτυχθεί πολλά εργαλεία, από διάφορες επιστημονικές ομάδες και από το GO Consortium. Ένα από αυτά είναι το **GOrilla (*Gene Ontology enRIchment anaLysis and visuaLizAtion tool)*** (Εικόνα 1.23) [68]. Υποστηρίζει 8 οργανισμούς (*H. sapiens, M. musculus,, R. norvegicus, D. rerio, D. melanogaster, C. elegans, S. cerevisiae, A. thaliana).* Αναφορικά με τα αναγνωριστικά εισόδου, εκτός απο τα gene Symbol υποστηρίζει και RefSeq, Uniprot, Unigene και Ensembl.



Εικόνα 1.23: Διεπαφή χρήστη GOrilla.

1.4.9 AmiGO 2

Η πρόσβαση στην GO Database γίνεται εφικτή μέσω του "AmiGO browser and search engine", [69] το οποίο χρησιμοποιείται ευρέως για αναζήτηση πληροφοριών στην οντολογία και ψάχνει στην GO με βάση κάποια GO terms, γονίδια, πρωτεΐνες ή ακριβείς λέξεις κλειδιά. Επιπλέον, το AmiGO δίνει τη δυνατότητα στο χρήστη να κατεβάσει ορολογίες και annotations, παρέχοντας ταυτόχρονα εργαλεία ανάλυσης και επεξεργασίας αυτών των δεδομένων, όπως είναι το GOOSE (online σύστημα εκτέλεσης SQL ερωτημάτων στην GO database. Τέλος, το AmiGO προσφέρει μια γρήγορη μηχανή αναζήτησης (Term Enrichment Service), που επιτρέπει την ανάλυση εμπλουτισμού ειδικών ειδών έναντι της πλήρους βάσης δεδομένων και αναδρομολογεί στην βάση PANTHER.

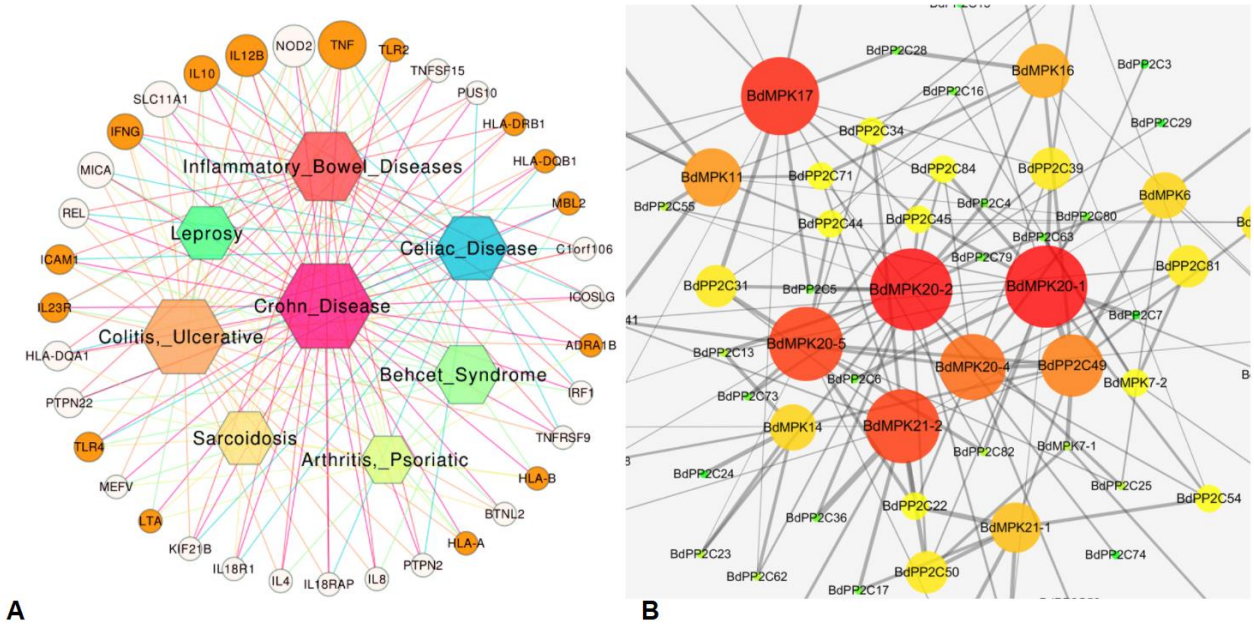
1.5 Βιολογικά Δίκτυα

Τα δίκτυα ή αλλιώς γράφοι, αποτελούν έναν διαγραμματικό τρόπο οπτικοποίησης των σχέσεων μεταξύ διαφόρων οντοτήτων συμβάλλοντας έτσι στην οργάνωση μεγάλου όγκου πληροφοριών. Ένα δίκτυο συνίσταται από τους κόμβους και τις ακμές οι οποίες συνδέουν το πλήθος των κόμβων μεταξύ τους. Υπάρχουν διάφορες κατηγορίες γραφημάτων. Τα πιο γνωστά είναι τα μη κατευθυνόμενα, κατευθυνόμενα, σταθμισμένα, διμερή, πολλαπλά, υπεργραφικά και δέντρα.

Η μελέτη της πολυπλοκότητας των βιολογικών συστημάτων περνάει έτσι, υποχρεωτικά από την επισκόπηση ενός τεράστιου αριθμού διεργασιών που αλληλοδιαπλέκονται δημιουργώντας πυκνά, σύνθετα δίκτυα. Η ανασύσταση, αρχικά, και η ανάλυση στη συνέχεια αυτών των δικτύων σε ό,τι αφορά το μέγεθος, τις επιμέρους υπομονάδες τους και τις στατιστικές τους ιδιότητες μας επιτρέπει να προσεγγίσουμε σε βάθος πολύ σημαντικά ερωτήματα σχετικά με τη ρύθμιση βιολογικών διεργασιών, την οργάνωση των κυττάρων σε επίπεδο συστημάτων αλλά και την ανάδυση γενικότερων ιδιοτήτων που αντανακλούν τον τρόπο με τον οποίο τα πολύπλοκα βιολογικά συστήματα εξελίσσονται σε διαρκώς μεταβαλλόμενα περιβάλλοντα διατηρώντας ταυτόχρονα μια αξιοθαύμαστη σταθερότητα [70], [71].

Μερικά παραδείγματα βιολογικών δικτύων (Εικόνα 1.24) που θα μπορούσαν να διακριθούν είναι [71]:

- **Δίκτυα Πρωτεϊνικών Αλληλεπιδράσεων:** Οι πρωτεΐνες απεικονίζονται με κόμβους και οι αλληλεπιδράσεις με τις ακμές.
- **Σηματοδοτικά Δίκτυα:** Αναπαριστούν την πορεία μετάδοσης σήματος απο τον εξωκυττάριο χώρο στον ενδοκυττάριο και αντίστροφα μέσω όλων των βιομορίων που συμμετέχουν στην μετάδοση του σήματος.
- **Μεταβολικά Δίκτυα:** Περιλαμβάνει όλες τις χημικές αντιδράσεις μέσα στο κύτταρο που αφορούν μια συγκεκριμένη διεργασία π.χ. Γλυκόλυση.
- **Ρυθμιστικά Δίκτυα:** Μοντελοποιείται ο τρόπος που οι πρωτεΐνες και άλλα βιομόρια εμπλέκονται στην διαδικασία της έκφρασης των γονιδίων
- **Νευρωνικά Δίκτυα:** Πληροφορίες για τον τρόπο μετάδοσης σημάτων στο νευρικό σύστημα.
- **Οικολογικά Δίκτυα:** Αναπαριστώνται οι βιοτικές αλληλεπιδράσεις σε ένα οικοσύστημα. Τα είδη των οργανισμών που βρίσκονται σε ένα οικοσύστημα συνδέονται με αλληλεπιδράσεις κατά ζεύγη και μπορεί να είναι είτε τροφικές είτε συμβιωτικές.
- **Δίκτυα ασθενειών:** Παρέχουν πληροφορίες για την προέλευση και την συσχέτιση ασθενειών με βιομόρια.
- **Φυλογενετικά Δίκτυα:** Απεικονίζουν τις εξελικτικές σχέσεις μεταξύ των οργανισμών στο χρόνο.



Εικόνα 1.24: Παραδείγματα βιολογικών δικτύων. Α) Δίκτυο συσχέτισης ασθενειών με γονίδια [72]. Β) Δικτυο απεικόνισης πρωτεϊνικών αλληλεπιδράσεων [73].

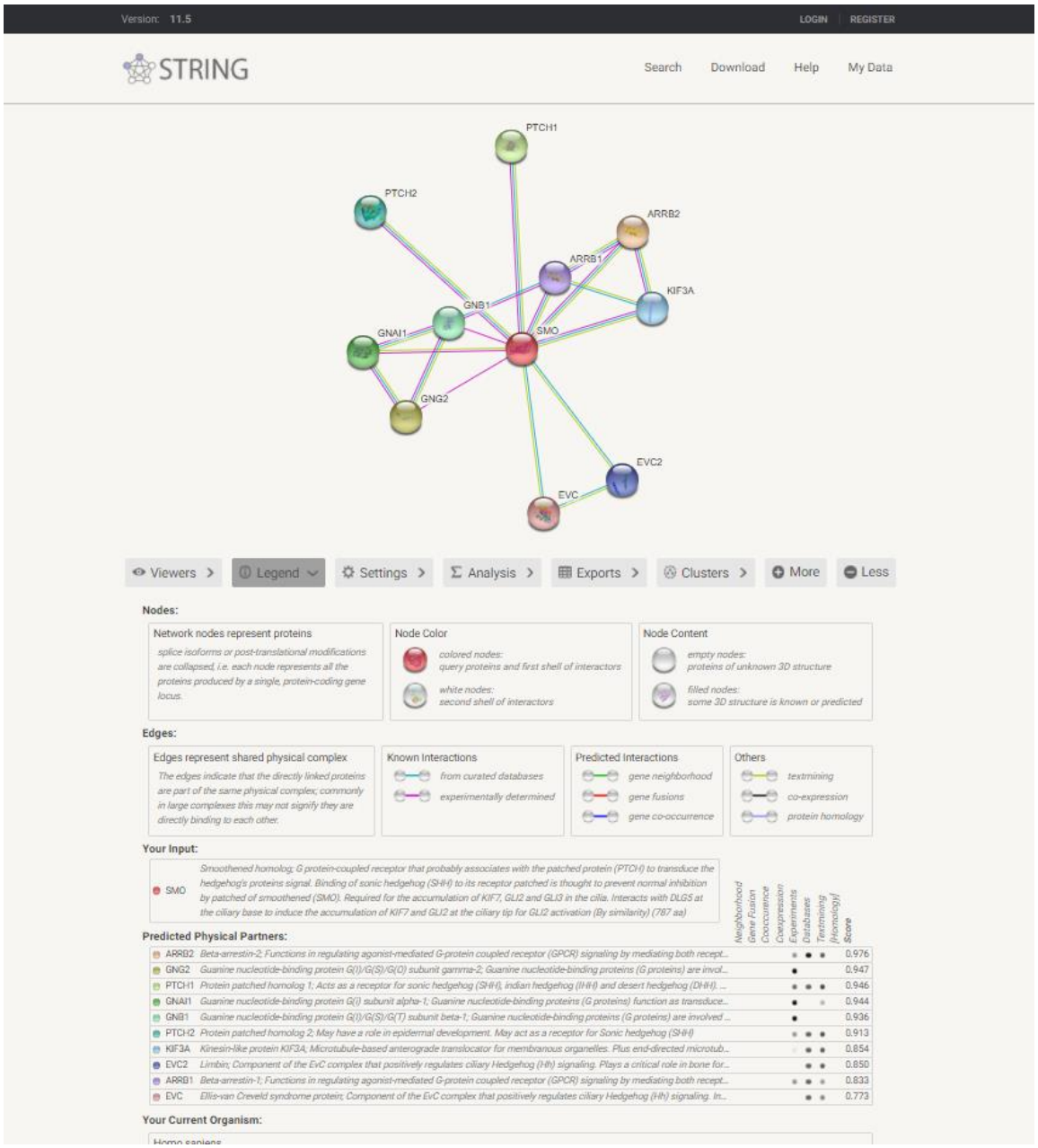
1.6 STRING

Η ανάπτυξη των τεχνολογιών υψηλής απόδοσης σε συνδυασμό με την δυνατότητα των ηλεκτρονικών υπολογιστών να αναλύουν μεγάλο όγκο δεδομένων έδωσε ώθηση στους επιστήμονες να καταχωρήσουν και να εντάξουν μεγάλο όγκο βιολογικών δεδομένων σε εξειδικευμένες βάσεις. Η βάση δεδομένων STRING [74] (The Search Tool for the Retrieval of Interacting Genes) είναι μια διαδικτυακή βάση που παρέχει πληροφορίες σχετικά με αλληλεπιδράσεις πρωτεϊνών (Εικόνα 1.25): είτε άμεσες (φυσικές) είτε έμμεσες (λειτουργικές), συνοψίζοντας τόσο πειραματικά δεδομένα όσο και υπολογιστικές προβλέψεις καθώς και δημόσια αναζήτηση κειμένου. Οι φυσικές αλληλεπιδράσεις αναφέρονται σε πρωτεΐνες που αποτελούν μέρος του ίδιου βιομοριακού συμπλέγματος, ενώ οι λειτουργικές αλληλεπιδράσεις αναφέρονται σε πρωτεΐνες οι οποίες εμπλέκονται στο ίδιο μονοπάτι ή βιολογική διαδικασία.

Η τελευταία έκδοση της STRING 10.0 περιέχει πληροφορίες για πάνω από 2.000 οργανισμούς ταξινομημένες σε οικογένειες με βάση τα διαφορετικά επίπεδα φυλογενετικής ανάπτυξης και μέσα από ένα σύστημα ανάκτησης δεδομένων υπολογίζει με τη χρήση ενός αλγόριθμου ένα σκορ εμπιστοσύνης για τις πιθανές αλληλεπιδράσεις και τις σχέσεις που διέπουν τις πρωτεΐνες.

Η περιγραφή των σχέσεων γίνεται συνήθως με δίκτυο πρωτεϊνικών αλληλεπιδράσεων, το οποίο αναπαρίσταται σαν ένας μη κατευθυνόμενος, αβαρής γράφος G(V,E), με τις πρωτεΐνες σαν σύνολο κόμβων V και τις αλληλεπιδράσεις μεταξύ τους σαν σύνολο ακμών E, προσεγγίζοντας έτσι ολιστικά το σύστημα και δίνοντας στο χρήστη τη δυνατότητα ανάκτησης όλων των πληροφοριών

που αφορούν τις μελετώμενες σχέσεις. Σημαντικό είναι να αναφερθεί ότι ο χαρακτηρισμός των πρωτεϊνών που βρίσκεται στη βάση και η ταξινόμηση των σχημάτων γίνεται βάση της σημαντικότητας των σχέσεων που διέπουν τις πρωτεΐνες, όπως σε φυσικά σύμπλοκα, είτε σηματοδοτικά μονοπάτια, είτε σε αρθρώματα. Επιπλέον, η βάση αυτή χρησιμοποιείται για την εμφάνιση λειτουργικών εμπλουτισμών στις πρωτεϊνικές λίστες, χρησιμοποιώντας πληθώρα συστημάτων λειτουργικής ταξινόμησης, όπως GO, Pfam και KEGG. Η STRING έχει αναπτυχθεί από μια κοινοπραξία των ακαδημαϊκών ιδρυμάτων, συμπεριλαμβανομένων CPR, EMBL, KU, SIB, TUD και UZH.



Εικόνα 1.25: Δίκτυο πρωτεϊνικών αλληλεπιδράσεων STRING.

1.7 Ανάπτυξη Λογισμικού

1.7.1 Γλώσσα προγραμματισμού R

Η R αποτελεί ένα ολοκληρωμένο περιβάλλον λογισμικού για τη διαχείριση δεδομένων, τους υπολογισμούς μεταξύ αυτών, τη στατιστική τους ανάλυση και τη δημιουργία γραφικών απεικονίσεων. Αναπτύχθηκε από τους Ross Ihaka (University of Auckland) και Robert Gentleman και βασίζεται σε μεγάλο βαθμό από τις γλώσσες προγραμματισμού S και Scheme. Επιπλέον, παρέχονται πακέτων (packages), στα οποία μάλιστα κάθε χρήστης μπορεί να συνεισφέρει ελεύθερα. Τα πακέτα περιέχουν συναρτήσεις (functions) και σειρές δεδομένων (datasets). Παρέχονται στο χρήστη περίπου 25 πακέτα, και στη περίπτωση που απαιτούνται πολύπλοκοι στατιστικοί υπολογισμοί ή εξειδικευμένα δεδομένα (π.χ. βιολογικά δεδομένα) ή άλλες βιβλιοθήκες, υπάρχουν διαθέσιμα ελεύθερα πακέτα στα επίσημα διαδικτυακά αποθετήρια CRAN, Bioconductor κλπ [75].

Ο κώδικας της παρούσας εργασίας αναπτύχθηκε στο περιβάλλον του RStudio. Το RStudio είναι ένα ελεύθερο λογισμικό ανοιχτού κώδικα και αποτελεί ένα ολοκληρωμένο περιβάλλον ανάπτυξης (integrated development environment, IDE) για την R.

1.7.2 Ανάπτυξη διαδραστικής εφαρμογής με Shiny

Το Shiny [76], [77] αποτελεί ένα πακέτο του RStudio, το οποίο χρησιμοποιείται για την ανάπτυξη διαδραστικών διαδικτυακών εφαρμογών. Κάθε εφαρμογή που αναπτύσσεται με το Shiny αποτελείται από δύο μέρη: μια ιστοσελίδα η οποία προορίζεται για τον χρήστη και ένα αρχείο που περιέχει τον βασικό κώδικα, ο οποίος «τροφοδοτεί» την εφαρμογή. Η διεπαφή ουσιαστικά αποτελείται από κώδικα HTML ο οποίος όμως γράφεται με τις συναρτήσεις του Shiny. Με τη διεπαφή διαμορφώνεται η διάταξη της εφαρμογής, δεδομένου ότι ορίζονται επακριβώς το γραφικό περιβάλλον, οι θέσεις των διαφόρων λειτουργικών παραμέτρων εισόδου (inputs) καθώς και ο τρόπος εμφάνισης των δεδομένων εξόδου (outputs). Οι παράμετροι εισόδου μπορούν να είναι μεταξύ άλλων της μορφής κουμπιών, πλαισίων ελέγχου, μενού, ενώ στην έξοδο μπορούν να εμφανίζονται κείμενο, γραφήματα, πίνακες κ.α. ανάλογα με τις ανάγκες της εκάστοτε εφαρμογής. Το κομμάτι του server είναι υπεύθυνο για τη λειτουργία της εφαρμογής και αποτελείται από τον ουσιαστικό κώδικα της που επεξεργάζεται τα δεδομένα εισόδου και υπολογίζει τα δεδομένα εξόδου, όταν ο χρήστης αλληλεπιδρά με την εφαρμογή. Τα δεδομένα εξόδου δύνανται να αλλάξουν οποιαδήποτε στιγμή λόγω της αλληλεπίδρασης αυτής, ωστόσο τα επίπεδα της διαδραστικότητας μπορούν να ρυθμιστούν από τον προγραμματιστή για να αποφεύγονται περιττές επανεκτελέσεις τμημάτων του κώδικα και τελικά να επιτυγχάνεται εξοικονόμηση πόρων. Τέλος μια εφαρμογή του Shiny μπορεί να δημοσιευτεί ώστε να είναι προσβάσιμη από όλους στο διαδίκτυο. Η διαδικασία αυτή επιτυγχάνεται είτε ανεβάζοντας την εφαρμογή στο shinyapps.io, το οποίο παρέχεται από το ίδιο το RStudio και είναι δωρεάν, είτε ανεβάζοντας σε άλλο διακομιστή μέσω του Shiny Server που επίσης είναι συμβατό με το RStudio και παρέχει περισσότερη ευελιξία στον προγραμματιστή.

1.8 Διαγράμματα Λογικών Σχέσεων

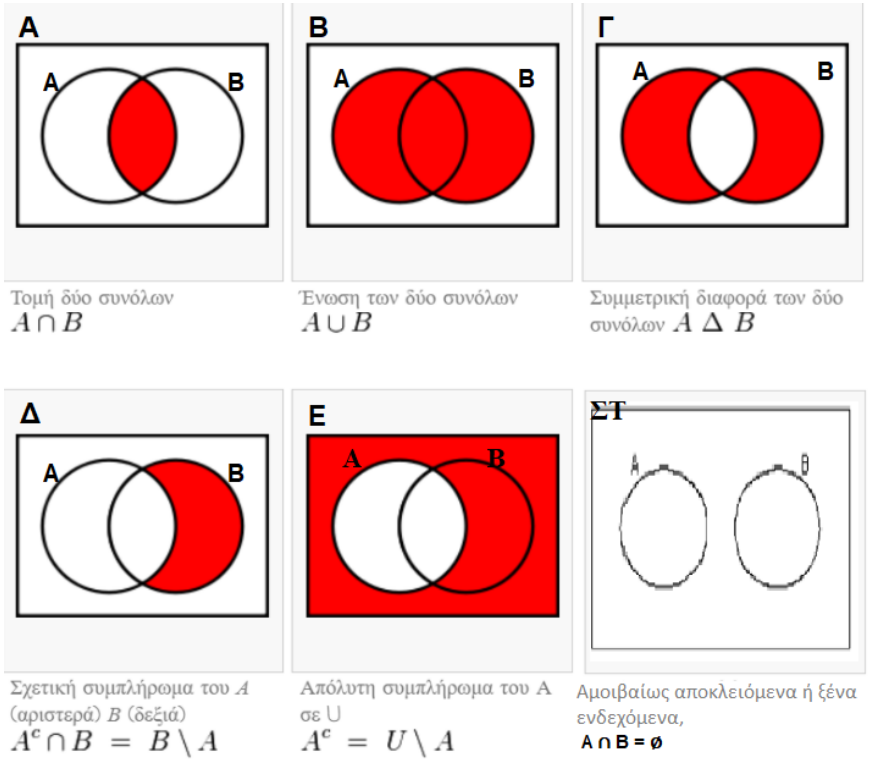
1.8.1 Διάγραμμα Venn (Venn diagram)

Το διάγραμμα Venn [78] είναι ένα διάγραμμα που απεικονίζει όλες τις πιθανές λογικές σχέσεις ανάμεσα σε μια πεπερασμένη συλλογή από σύνολα χρησιμοποιώντας τεμνόμενους κύκλους για να αναπαραστήσει τις ομοιότητες, τις διαφορές και τις σχέσεις μεταξύ εννοιών, ιδεών, κατηγοριών ή ομάδων. Δημιουργήθηκε γύρω στο 1880 από τον Τζον Βενν.

Το διάγραμμα χρησιμοποιείται για τη μελέτη στοιχειωδών στοιχείων της θεωρίας συνόλων και βρίσκει πληθώρα εφαρμογών, όπως στις πιθανότητες, τη λογική, τα στατιστικά στοιχεία, τη γλωσσολογία και την επιστήμη των υπολογιστών, στη βιολογία κλπ.

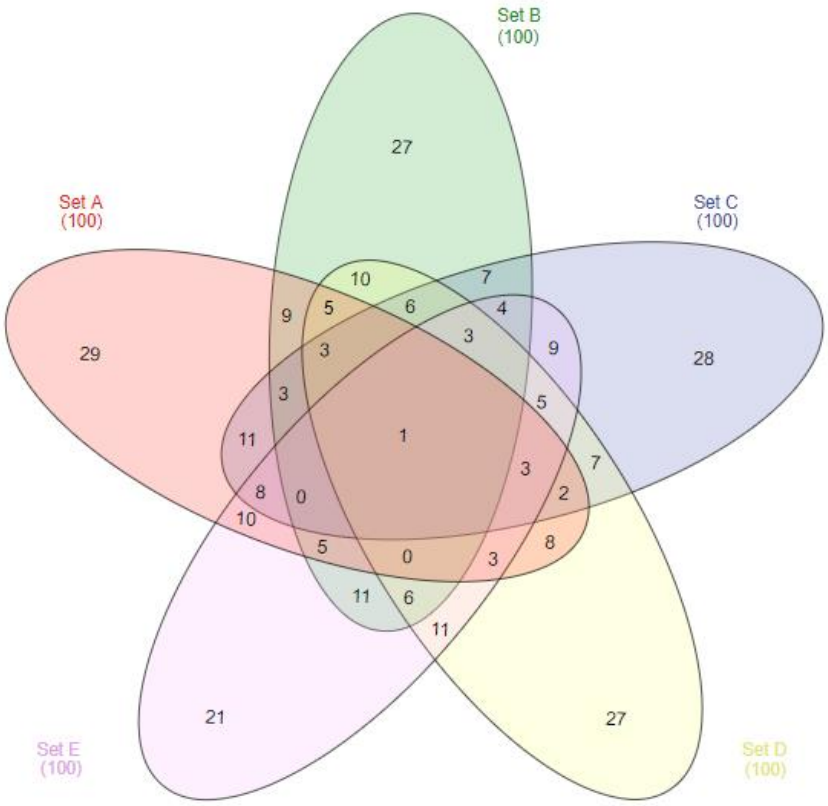
Κάθε διάγραμμα Venn αποτελείται από: ένα ορθογώνιο που συμβολίζει το μεγαλύτερο δυνατό σύνολο που μπορούμε να θεωρήσουμε, ανάλογα με το τί θέλουμε να δείξουμε και συμβολίζεται συνήθως με Ω ή U και από κλειστές γραμμές, συνήθως καμπύλες και κύκλοι, όπου η επιφάνεια που περικλείουν συμβολίζει το ίδιο το σύνολο. Οι πράξεις στην Άλγεβρα των Συνόλων είναι δυο: η ένωση και η τομή. Για δύο σύνολα Α και Β ορίζονται οι εξής πράξεις (Εικόνα 1.26).

1. Η τομή συνόλων Α και Β
2. Η Ένωση συνόλων Α και Β
3. Η συμμετρική διαφορά δύο συνόλων Α και Β
4. Η διαφορά συνόλων (Απόλυτο και Σχετικό συμπλήρωμα)
5. Αμοιβαίως αποκλειόμενα ή ξένα ενδεχόμενα.



Εικόνα 1.26 : Πράξεις με Σύνολα. Α)Το ενδεχόμενο να συμβούν και τα δυο ενδεχόμενα «Α και Β» λέγεται τομή των Α και Β και συμβολίζεται με Α ∩ Β. Β) Αν Α, Β είναι δύο ενδεχόμενα του ίδιου δειγματικού χώρου, τότε το ενδεχόμενο να συμβεί τουλάχιστον ένα από τα δυο λέγεται ένωση των Α και Β και συμβολίζεται με Α ∪ Β. Γ) Να συμβεί ή μόνο το Α ή μόνο το Β ακριβώς 1 από τα 2 Α ∩ Β ′ ∪ Α ′ ∩ Β = (Α − Β) ∪ (Β − Α). Δ) Συμβαίνει μόνο το Β, Α ′ ∩ Β = Β − Α. Ε) Το ενδεχόμενο να μην συμβεί το ενδεχόμενο Α λέγεται συμπληρωματικό ενδεχόμενο και συμβολίζεται με Α'. ΣΤ) Τα 2 ενδεχόμενα δεν μπορούν να συμβούν ταυτόχρονα.

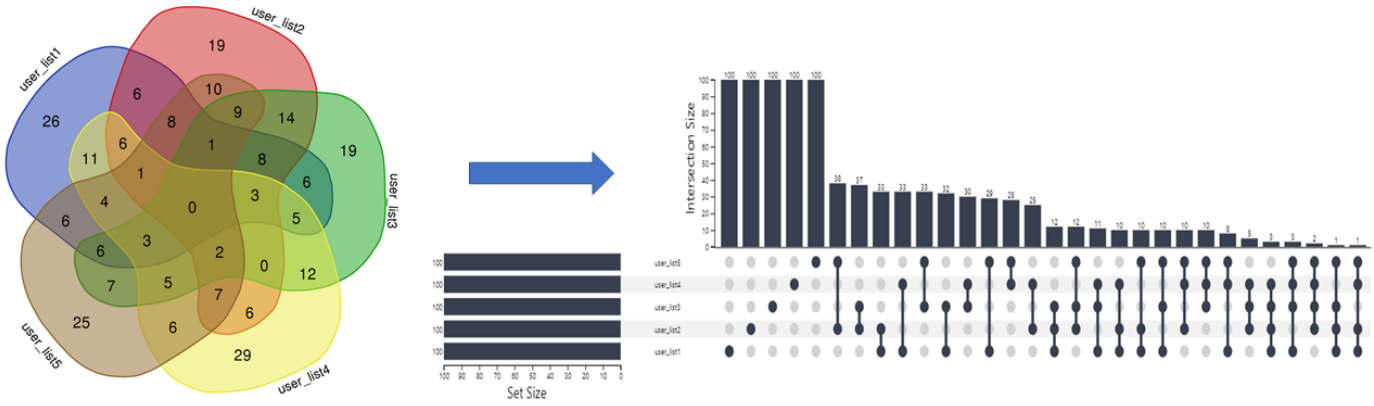
Οι πράξεις μεταξύ περισσότερων συνόλων επιτρέπουν την εποπτική διαχείρισή τους με τη χρήση των διαγραμμάτων Venn. Παρόλα αυτά δεν είναι όμως το ίδιο εύκολο να εκφραστεί το διάγραμμα Venn για μεγάλο αριθμό συνόλων. Στην Εικόνα 1.27 φαίνεται η περίπτωση πέντε συνόλων (Set A-C) που το καθένα αποτελείται από 100 γονίδια που προέκυψαν πειράματα γονιδιακής έκφρασης. Αυξανόμενου του αριθμού των συνόλων είναι δύσκολο να εκτιμηθούν με ευκολία οπτικά οι σχέσεις -κοινά και μη κοινά στοιχεία σε κάθε συνδυασμό- μεταξύ των συνόλων. Η σχεδίαση του διαγράμματος έχει καταστεί εξαιρετικά δύσκολη.Ταυτόχρονα σε περιπτώσεις που τα σύνολα αποτελούνται από μεγάλο αριθμό στοιχείων μπορεί να είναι δύσκολη η απεικόνιση των στοιχείων αυτών σε κάθε συνδυασμό συνόλων [79].



Εικόνα 1.27: Διάγραμμα Venn της τομής 5 συνόλων.

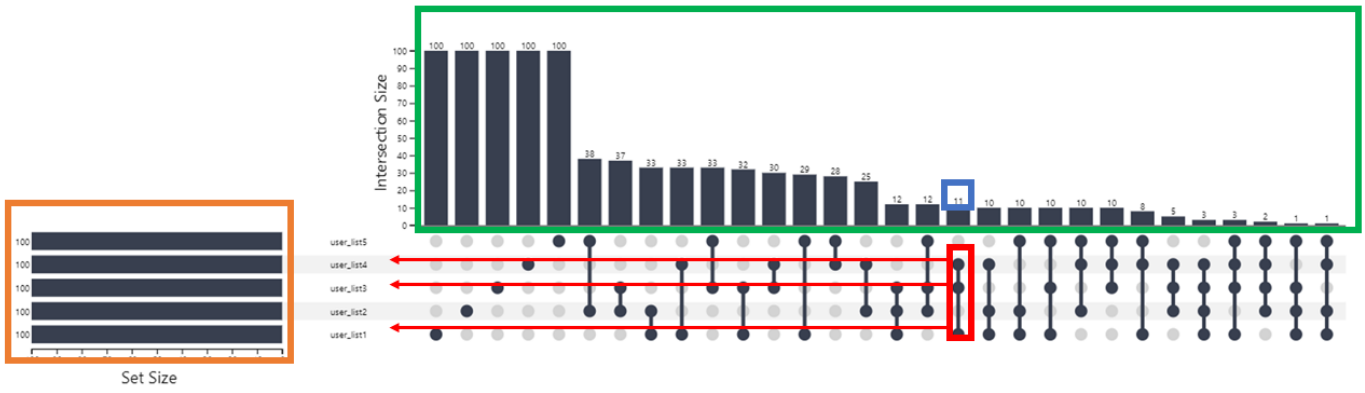
1.8.2 Διάγραμμα UpSet

Η κατανόηση των σχέσεων μεταξύ συνόλων και ειδικότερα βιολογικής προέλευσης, όπως για παράδειγμα από διαφορικά πειράματα γονιδιακής έκφρασης αποτελεί ένα πολύ σημαντικό ζήτημα. Με τη διαγραμματική απεικόνιση Venn είναι εμφανές ότι για πολλαπλό αριθμό συνόλων δεν είναι εύκολη η απεικόνιση, αλλά και η εξαγωγή όλων των πιθανών σχέσεων μεταξύ των υπό μελέτη συνόλων. Τα διαγράμματα UpSet παρέχουν έναν αποτελεσματικότερο και πιο κατανοητό τρόπο οπτικοποίησης όλων των πιθανών σχέσεων μεταξύ μεγάλου αριθμού συνόλων σε σύγκριση με τις παραδοσιακές μεθόδους, όπως το διάγραμμα (Εικόνα 1.28).



Εικόνα 1.28 : Διάγραμμα Venn της τομής 5 συνόλων (Αριστερά) και UpSet plot για τα ίδια 5 σύνολα.

Αποτελείται από δύο άξονες και έναν πίνακα συνδεδεμένων σημείων. Τα κάθετα ορθογώνια (ράβδοι) αντιπροσωπεύουν τον αριθμό των στοιχείων που συμμετέχουν σε κάθε συνδυασμό λιστών. Οι συνδεδεμένες κουκκίδες στον πίνακα υποδεικνύουν ποιος συνδυασμός λιστών αντιστοιχεί σε κάθε κατακόρυφο ορθογώνιο. Τέλος, οι οριζόντιες ράβδοι (Set Size) υποδηλώνουν τις λίστες που συμμετέχουν καθώς και το πλήθος των αρχικών λιστών (Εικόνα 1.29).



Εικόνα 1.29: Διάγραμμα UpSet plot για 5 σύνολα. Αποτελείται από δύο άξονες και έναν πίνακα συνδεδεμένων σημείων. Οι κάθετοι ράβδοι (πράσινο περίγραμμα) αντιπροσωπεύουν τον αριθμό των στοιχείων που συμμετέχουν σε κάθε συνδυασμό λιστών. Οι συνδεδεμένες κουκκίδες στον πίνακα υποδεικνύουν ποιος συνδυασμός λιστών αντιστοιχεί σε κάθε κατακόρυφο ορθογώνιο. Για παράδειγμα, οι επισημασμένες (κόκκινο) κουκίδες αντιστοιχούν στα σύνολα 2, 3 και 5 και η ράβδος που αντιστοιχεί

περιγράφει το σύνολο των κοινών στοιχείων, 11 στη προκειμένη. Τέλος, οι οριζόντιες ράβδοι (Set Size) υποδηλώνουν τις λίστες που συμμετέχουν καθώς και το πλήθος των αρχικών λιστών.

1.9 Σκοπός

Σήμερα ένας μεγάλος αριθμός εργαλείων εμπλουτισμού έχουν αναπτυχθεί για την λειτουργική ανάλυση γονιδιακών λιστών. Η ανάλυση εμπλουτισμού είναι μια πολλά υποσχόμενη στρατηγική υψηλής απόδοσης που μπορεί να αυξήσει την πιθανότητα για τους ερευνητές να εντοπίσουν βιολογικές διεργασίες που είναι πιο σχετικές με τη μελέτη τους. Παρ' όλα αυτά, τα ήδη υπάρχοντα εργαλεία (*Παράγραφος 1.4 Εργαλεία λειτουργικού εμπλουτισμού*), (α) συχνά διαφέρουν ως προς τους οργανισμούς, τα αναγνωριστικά και τις βάσεις δεδομένων που υποστηρίζουν, (β) αναφέρουν τα αποτελέσματα συνήθως με στατικές λίστες και αναπαραστάσεις, και (γ) συχνά δεν είναι σε θέση να χειριστούν και να συγκρίνουν πολλαπλές λίστες για μια πιο συνδυαστική ανάλυση.

Σκοπός της παρούσας εργασίας είναι η δημιουργία ενός διαδικτυακού εργαλείου λειτουργικού και βιβλιογραφικού εμπλουτισμού που διαφοροποιείται από τα υπάρχοντα, λύνοντας τα παραπάνω προβλήματα. Το Flame αποτελεί ένα εύχρηστο εργαλείο που εκτός από την κλασική ανάλυση λειτουργικού αλλά και βιβλιογραφικού εμπλουτισμού, επιτρέπει επίσης το χειρισμό πολλαπλών γονιδιακών λιστών. Αυτό γίνεται με εύκολο και διαδραστικό τρόπο επιτρέποντας την κατασκευή λιστών μέσα από τομές και ενώσεις πολλαπλών λιστών δεδομένων, σε διαδραστικά UpSet διαγράμματα, για περαιτέρω ανάλυση εμπλουτισμού. Ταυτόχρονα, στόχος είναι η παρουσίαση των αποτελεσμάτων με πληθώρα διαγραμμάτων για καλύτερη οπτικοποίηση της βιολογικής πληροφορίας με πολλαπλές επιλογές. Το Flame επιτρέπει επίσης τη δικτυακή ανάλυση των λιστών δεδομένων ώστε να διερευνηθούν πρωτεϊνικές αλληλεπιδράσεις.

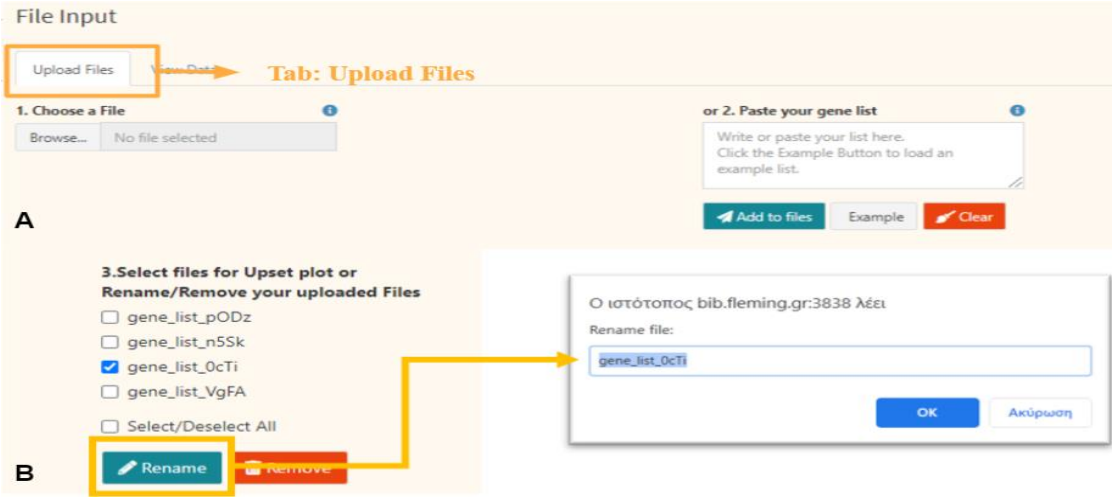
Τέλος, ένα ακόμα σημαντικό πλεονέκτημα του FLAME είναι ότι ακολουθεί μια προσέγγιση οπτικής ανάλυσης που επιτρέπει στους χρήστες να προσαρμόζουν και να παραμετροποιούν τα αναφερόμενα αποτελέσματα μέσω διαδραστικών θερμικών χαρτών (heatmaps), ραβδογραμμάτων, γραφημάτων Μανχάταν, δικτύων και πινάκων διευκολύνοντας έτσι την απεικόνιση, κατανόηση και ερμηνεία των αναλύσεων.

2. ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

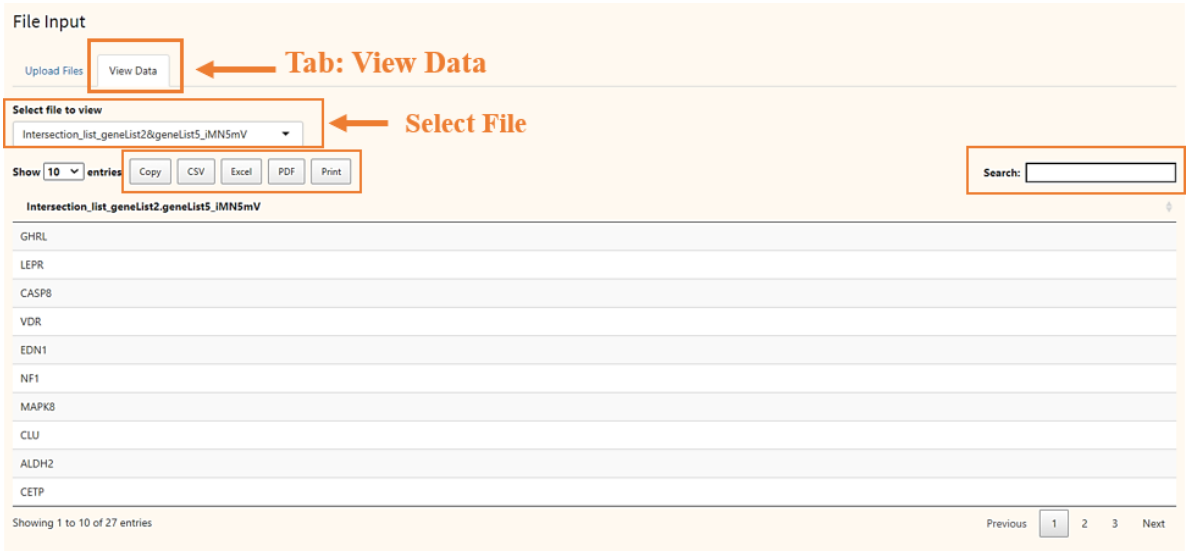
2.1 Αρχεία Εισόδου (Input)

Το Flame επιτρέπει στον χρήστη τόσο την μεταμόρφωση πολλαπλών ξεχωριστών αρχείων όσο και τη δυνατότητα επικόλλησης των γονιδιακών/πρωτεϊνικών κλπ λιστών του σε πλαισίου κειμένου. Για την εισαγωγή αρχείων με οποιονδήποτε από τους δύο τρόπους ο χρήστης πρέπει να μεταβεί απο το Menu στην καρτέλα File Input → Uploads Files (Εικόνα 2.1.A). Ο τύπος αρχείων που υποστηρίζονται είναι αρχεία Flat text (.txt, .tsv, .csv). Μετά την επικόλληση των δεδομένων του στο πλαίσιο κειμένου ο χρήστης για να αποθηκεύσει την λίστα του πρέπει να πατήσει την επιλογή "Add to files". Δίνεται η δυνατότητας χρήσης τυχαίων παραδειγμάτων γονιδιακών λιστών πατώντας το κουμπί "Example". Στην ηλεκτρονική (online) έκδοση το FLAME μπορεί να υποστηρίξει το μέγιστο 10 αρχεία λιστών. Το μέγεθος κάθε αρχείου δεν μπορεί να υπερβαίνει το 1 MB. Ο περιορισμός, τόσο του μεγέθους των αρχείων, όσο του αριθμού τους μπορεί να αρθεί, καθώς δίνεται η δυνατότητα στον χρήστη να κατεβάσει την εφαρμογή μέσω GitHub και να τροποποιήσει τις αντίστοιχες μεταβλητές. Συγκεκριμένα, μπορεί να τροποποιήσει τις μεταβλητες FILE_LIMIT στο αρχείο global.R για τον αριθμό των αρχείων και την shiny.maxRequestSize στο ui.R αρχείο για το μέγεθος (MB) και έπειτα να εκτελέσει τοπικά την εφαρμογή.

Ο κάθε όρος της λίστας, είτε από αρχείο μέσω μεταμόρφωσης, είτε μέσω εισαγωγής από το πλαίσιο κειμένου μπορεί να εισαχθεί χωρισμένος με κόμμα, κενό, πλήκτρο tab ή αλλαγή γραμμής. Στην παρούσα έκδοση το FLAME υποστηρίζει 197 οργανισμούς, καθώς και αρκετούς αναγνωριστικούς κωδικούς γονιδίων (Gene Identifiers, ID) όπως για παράδειγμα IDs πρωτεϊνών, μικροσυστοιχιών κλπ, αναγνωριστικά SNPs, χρωμοσωμικών περιοχών και term IDs, τα οποία χρησιμοποιούνται στην συνάρτηση gConvert της βιβλιοθήκης του gProfiler. Οι διαφορετικές λίστες (αρχεία) που επιθυμεί να διαχειριστεί ο χρήστης δεν μπορούν να έχουν το ίδιο όνομα. Γι αυτόν τον λόγο δίνεται η επιλογή για μετονομασία (Rename) αλλά και διαγραφή (Remove) τόσο μεμονομένων όσο και πολλαπλών αρχείων (Εικόνα 2.1.B). Αφού ο χρήστης εισάγει στην εφαρμογή τα αρχεία του απο την καρτέλα File Input → View Data, έχει τη δυνατότητα να δει το περιεχόμενο από οποιαδήποτε λίστα σε διαδραστικό περιβάλλον με επιλογές αναζήτησης και αποθήκευσης σε άλλες μορφές (Εικόνα 2.3).



Εικόνα 2.1: Εισαγωγή και Διαχείριση αρχείων εισόδου. (A) Εισαγωγή αρχείων εισόδου με μεταμόρφωση πολλαπλών αρχείων ή επικόλληση δεδομένων στο πλαίσιο κειμένου. Μετά την επικόλληση ο χρήστης για να αποθηκεύσει την λίστα του πρέπει να πατήσει την επιλογή "Add to files". Δίνεται η δυνατότητας χρήσης τυχαίων παραδειγμάτων γονιδιακών λιστών πατώντας το κουμπί "Example" (B) Δυνατότητα μετονομασίας ή διαγραφής πολλαπλών ή μεμονωμένων αρχείων.



Εικόνα 2.2: Προβολή περιεχομένου αρχείων εισόδου. Μέσω της καρτέλας View Data ο χρήστης μπορεί να δει όλα τα αρχεία που έχει αποθηκεύσει έχοντας της επιλογές για Αναζήτηση, Αποθήκευση σε διάφορες μορφές καθώς και Εκτύπωση.

2.2 UpSet Plot

Μετά την εισαγωγή των επιθυμητών λιστών προς ανάλυση από τον χρήστη δίνεται η δυνατότητα δημιουργίας γραφήματος UpSet. Για την δημιουργία των διαγραμμάτων UpSet χρησιμοποιήθηκε το πακέτο R/upsetjs. Όπως προαναφέρθηκε στην Ενότητα 1.8.2 το διάγραμμα UpSet παρέχει έναν αποτελεσματικότερο τρόπο οπτικοποίησης των κοινών και μη κοινών στοιχείων πολλαπλών λιστών σε οποιονδήποτε συνδυασμό σε σχέση με το διάγραμμα Venn. Το

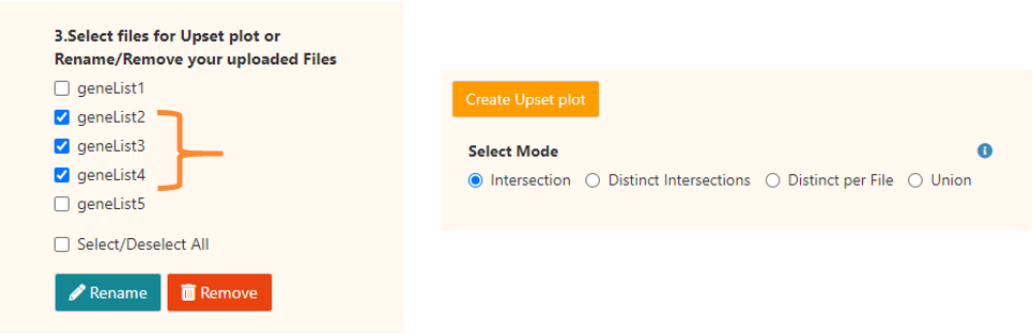
διάγραμμα Venn μπορεί να δώσει μια ευκρινή και κατανοητή εικόνα έχοντας ως όρισμα το πολύ 5 αρχεία εισόδου, σε αντίθεση με το UpSet που προσπερνά αυτό το όριο και δίνει μια εξελιγμένη και κατανοητή εικόνα των αποτελεσμάτων. Στην Εικόνα 2.4A-D φαίνεται μια συγκριτική απεικόνιση με τις δύο μεθόδους χρησιμοποιώντας 3 αρχεία εισόδου με 100 τυχαία γονίδια η κάθε μια. Αντίστοιχα, στην Εικόνα 2.4E απεικονίζεται διάγραμμα σε λειτουργία Distinct Intersection μεταξύ επτά λιστών, κάτι το οποίο δεν θα μπορούσε να γίνει με την χρήση διαγράμματος Venn.

Για να δημιουργηθεί το διάγραμμα UpSet, ο χρήστης πρέπει να επιλέξει το ελάχιστο 2 αρχεία από το πεδίο διαχείρισης αρχείων ("3.Select files for Upset plot or Rename/Remove your uploaded Files") και στην συνέχεια έχοντας επιλέξει την επιθυμητή λειτουργία να πατήσει το κουμπί "Create UpSet Plot" (Εικόνα 2.3).

Το γράφημα UpSet που προκύπτει αποτελείται από δύο άξονες και έναν πίνακα συνδεδεμένων κουκίδων. Τα κάθετα ορθογώνια (ράβδοι) αντιπροσωπεύουν τον αριθμό των στοιχείων που συμμετέχουν σε κάθε συνδυασμό λιστών. Οι συνδεδεμένες κουκίδων στον πίνακα υποδεικνύουν ποιος συνδυασμός λιστών αντιστοιχεί σε κάθε κατακόρυφο ορθογώνιο. Τέλος, οι οριζόντιες ράβδοι (Set Size) υποδηλώνουν το αριθμητικό περιεχόμενο των αρχείων που συμμετέχουν, καθώς και το πλήθος των αρχικών λιστών.

Η παρούσα έκδοση του FLAME υποστηρίζει 4 λειτουργίες (Modes) για την κατασκευή UpSet Plot:

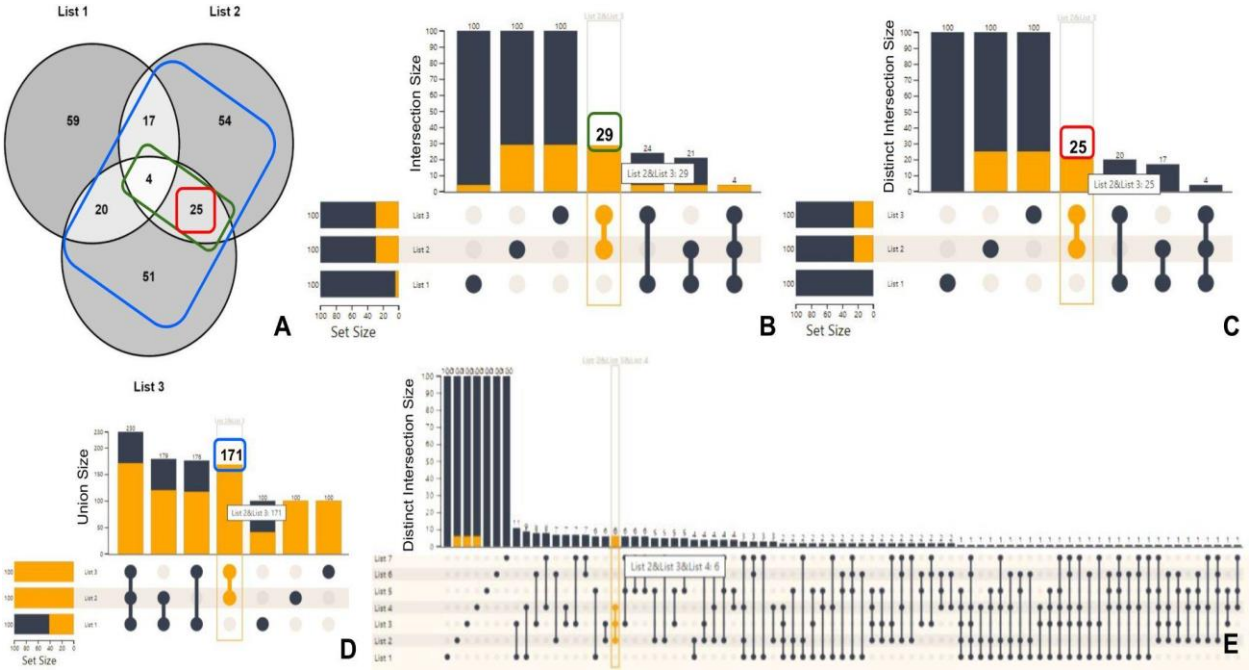
- Intersection,
- Distinct intersection,
- Distinct elements per file και
- Union



Εικόνα 2.3: Επιλογή λιστών για δημιουργία UpSet Plot. Ο χρήστης πρέπει να επιλέξει τουλάχιστον δύο λίστες για τη δημιουργία του διαγράμματος, καθώς και την επιθυμητή επιλογή λειτουργίας.

Το Intersection Mode παρουσιάζει όλα τα κοινά στοιχεία των επιθυμητών συνδυασμών των υπό μελέτη λιστών αν φυσικά οι συνδυασμοί έχουν τουλάχιστον ένα κοινό στοιχείο. Για παράδειγμα αν ο χρήστης έχει να διαχειριστεί 3 λίστες A1, B1, Γ1 θα έχει στην διάθεση του τα κοινά στοιχεία των A1B1, A1Γ1, B1Γ1 και A1B1Γ1. Το Distinct intersection Mode εμφανίζει τα κοινά στοιχεία μεταξύ των συνδυασμών των επιθυμητών λιστών που υπάρχουν σε αυτές αλλά δεν συμμετέχουν σε καμία άλλη λίστα. Η λειτουργία Distinct elements per file παρουσιάζει τα στοιχεία που είναι μοναδικά σε κάθε λίστα και δεν συμμετέχουν σε καμία άλλη. Τέλος, η λειτουργία Union

κατασκευάζει την ένωση των στοιχείων από όλους τους δυνατούς συνδυασμούς των λιστών εισόδου, δείχνοντας τα μοναδικά στοιχεία, έχοντας δηλαδή αφαιρέσει τυχόν πολλαπλά στοιχεία που προέκυψαν από την ένωση N αριθμού λιστών. Σε κάθε mode το διάγραμμα που προκύπτει είναι διαδραστικό και ο χρήστης έχει την δυνατότητα μεταφέροντας τον κέρσορα πάνω σε κάποιο επιθυμητό συνδυασμό λιστών, δηλαδή στα κάθετα ορθογώνια (ράβδους ή στις συνδεδεμένες κουκίδες) να βλέπει τα αντίστοιχα στοιχεία (γονίδα, πρωτεΐνες κλπ) που περιέχει αυτος ο συνδυασμός. Επίσης, κλικάροντας πάνω στον επιθυμητο συνδυασμό, μέσω μηνύματος που εμφανίζεται στην οθόνη μπορεί να αποθηκεύσει την επιλεγμένη λίστα στα αρχεία του και να τη διαχειριστεί για μετέπειτα ανάλυση, όπως τα αρχικά αρχεία εισόδου. Τέλος, ο χρήστης μπορεί να αποθηκεύσει σε μορφή εικόνας (.png, .svg), να μοιραστεί (share via link) με άλλους χρήστες το διάγραμμα που δημιούργησε ή να το κατεβάσει σε JSON αρχείο.



Εικόνα 2.4: Διάγραμμα UpSet Vs διάγραμμα Venn. (A) Τομή τριών γονιδιακών λιστών (100 γονίδια/λίστα) με τη χρήση διαγράμματος Venn. (B) Η επιλογή Intersection του UpSet plot απεικονίζει τον συνολικό αριθμό των κοινών στοιχείων μεταξύ των επιλεγμένων συνόλων, παρόλο που τα στοιχεία αυτά ενδέχεται να συμμετέχουν και σε άλλα σύνολα. Για παράδειγμα, οι λίστες 2 και 3 περιέχουν 29 κοινά γονίδια (πράσινο ορθογώνιο), με 25 από αυτά να είναι μοναδικά κοινά μεταξύ αυτών των λιστών, δηλαδή δεν συμμετέχουν σε κάποιο άλλο σύνολο, και 4 γονίδια κοινά με τη λίστα 1, όπως φαίνεται στο (A). (C) Η επιλογή Distinct intersection απεικονίζει τον κοινό αριθμό γονιδίων μεταξύ των επιλεγμένων συνόλων, τα οποία δεν υπάρχουν σε κανένα άλλο σύνολο. Αυτή η επιλογή είναι η πλησιέστερη στο διάγραμμα Venn. Για παράδειγμα, οι λίστες 2 και 3 έχουν 25 κοινά γονίδια, που δεν περιέχονται σε καμία άλλη λίστα (κόκκινο ορθογώνιο). (D) Η λειτουργία Union κατασκευάζει την ένωση των στοιχείων από όλους τους δυνατούς συνδυασμούς των λιστών εισόδου, δείχνοντας τα μοναδικά στοιχεία, έχοντας δηλαδή αφαιρέσει τυχόν πολλαπλά στοιχεία που προέκυψαν από την ένωση N αριθμού λιστών. Για παράδειγμα, ο συνδυασμός των λιστών 2 και 3 έχει ως αποτέλεσμα 171 συνολικά μοναδικά γονίδια (μπλε ορθογώνιο). (E) Ένα παράδειγμα διαγράμματος UpSet «distinct intersections» με 7 λίστες, απεικόνιση η οποία δεν είναι δυνατή με την χρήση διαγράμματος Venn.

2.3 Functional enrichment analysis

2.3.1 Ontologies και Pathways (gProfiler)

Για τη λειτουργική ανάλυση εμπλουτισμού γονιδίων/πρωτεϊνών σε οντολογίες (Ontologies) και μονοπάτια (Pathways) χρησιμοποιήθηκε η βιβλιοθήκη R/gprofiler2 [58] και συγκεκριμένα το εργαλείο g:GOSt, το οποίο πραγματοποιεί ανάλυση εμπλουτισμού για να εντοπίσει ποιες βιολογικές λειτουργίες και μονοπάτια υπερεκπροσωπούνται απο τα γονίδια/πρωτεΐνες του αρχείου εισόδου με στατιστική σημαντικότητα βάσει των επιλεγμένων βάσεων δεδομένων.

Πλέον ο χρήστης μπορεί ως αρχείο εισόδου, εκτός από τα αρχεία που μεταμορφώνει εξ αρχής, να χρησιμοποιήσει και τα αρχεία που δημιουργησε μέσω του διαγράμματος UpSet. Για την ανάλυση λειτουργικού εμπλουτισμού ο χρήστης καλείται να επιλέξει μια σειρά παραμέτρων (Εικόνα 2.5), για τις οποίες επιθυμεί να γίνει η ανάλυσή. Ειδικότερα, πρέπει να επιλέξει:

1. **Βάσεις Δεδομένων (Select datasources):** Μπορεί να επιλέξει μία η περισσότερες βάσεις δεδομένων από τις παρακάτω.
   - **Gene Ontology (GO)**: Διαχωρίζεται σε 3 υποκατηγορίες:
     α) Molecular Functions (MF, Μοριακή λειτουργία),
     β) Biological Process (BP, Βιολογική διαδικασία) και
     γ) Cellular Component (CC, Κυτταρικός Εντοπισμός)
   - **Βάσεις Δεδομένων Βιολογικών μονοπατιών (Biological Pathway databases)**:
     α) Reactome,
     β) KEGG και
     γ) WikiPathways
   - **Πρωτεϊνικές Βάσεις Δεδομένων (Protein databases):**
     α) CORUM
     β) Human Protein Atlas
   - **Human Phenotype Ontology**
   - **Λειτουργικά μοτίβα στο DNA (Regulatory motifs in DNA):**
     α) TRANSFAC και
     β) mirTarBase

2. **Οργανισμός (Select organism):** Μπορεί να επιλέξει μεταξύ 197 οργανισμών.

3. **Στατιστική μέθοδος διόρθωσης - Έλεγχος Πολλαπλών υποθέσεων (Significance threshold):** Η ανάλυση των αποτελεσμάτων της λειτουργικής ανάλυσης γονιδίων περιλαμβάνει έναν μεγάλο αριθμό ελέγχων πολλαπλών υποθέσεων ταυτόχρονα, καθώς κάθε αρχείο εισόδου συγκρίνεται με χιλιάδες όρους, μονοπάτια, ασθένειες, φαινοτύπους, λειτουργικά μοτίβα κλπ. Με τον έλεγχο των πολλαπλών υποθέσεων μειώνεται η πιθανότητα εμφάνισης ψευδών θετικών αποτελεσμάτων.

Για τη λειτουργική ανάλυση των δεδομένων δίνεται η δυνατότητα στο χρήστη να επιλέξει μέθοδο για τον έλεγχο πολλαπλών υποθέσεων (Significance threshold) μεταξύ των g:SCS threshold, Bonferroni correction και Benjamini-Hochberg FDR (False Discovery Rate).

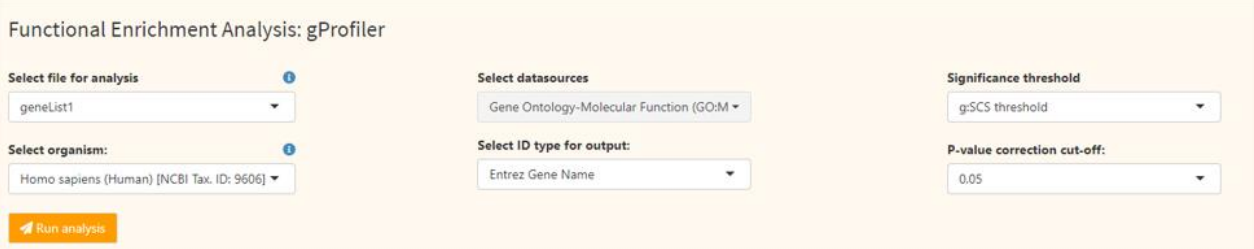
- **g:SCS algorithm:** Η μέθοδος SCS είναι η προεπιλεγμένη μέθοδος για τον υπολογισμό διόρθωσης των πολλαπλών τιμών p-value που προκύπτουν από κάθε ανάλυση εμπλουτισμού. Η μέθοδος αυτή επιλέχθηκε, καθώς οι άλλες δύο έχουν σχεδιαστεί για πολλαπλές δοκιμές ανεξάρτητες μεταξύ τους. Αυτό όμως δεν ανταποκρίνεται στον τύπο των δεδομένων και των αποτελεσμάτων της ανάλυσης εμπλουτισμού, καθώς όπως στην περίπτωση της βάσης δεδομένων GO όλοι οι όροι είναι ιεραρχικά συνδεδεμένοι μεταξύ τους σε επίπεδα. Το αρχικό κατώφλι αντιστοιχεί στην τιμή α = 0.05, δηλαδή τουλάχιστον το 95% των αποτελεσμάτων-αντιστοιχιών πάνω από το κατώφλι θεωρούνται στατιστικά σημαντικά. Το κατώφλι για τη μέθοδο αυτή είναι μια τιμή που έχει υπολογιστεί εκ των προτέρων για λίστες έως 1000 γονιδίων. Δεδομένου ενός σταθερού μεγέθους του αρχείου εισόδου, ο αλγόριθμος εκτιμάει ένα κατώφλι t. Οι τιμές p-value που προκύπτουν από το ερώτημα μετατρέπονται σε διορθωμένες τιμές p-value πολλαπλασιάζοντας αυτές με το λόγο του κατά προσέγγιση κατωφλίου t προς του αρχικού ορίου σε όλο το πείραμα a = 0,05. Ο αλγόριθμος εξετάζει τη δομή του συνόλου των υποκείμενων γονιδιακών συνόλων που αντιστοιχούν σε κάθε όρο για τον κάθε οργανισμό και, ως εκ τούτου, θα πρέπει να δώσει ένα αυστηρότερο κατώφλι για τα στατιστικά σημαντικά αποτελέσματα.
- **Bonferroni correction**: Η "Διόρθωση κατά Bonferroni" είναι μια απλή στατιστική μέθοδος διόρθωσης πολλαπλών υποθέσεων σε περιπτώσεις εμφάνισης στατιστικά σημαντικών συσχετίσεων μεταξύ πολλών μεταβλητών και χειρίζεται την πιθανότητα να εμφανιστεί τουλάχιστον ένα σφάλμα τύπου I (Family Wise Error Rate p-value, FWER). Ως σφάλμα τύπου I ορίζεται η απόρριψη της μηδενικής υπόθεσης, ενώ είναι σωστή. Σύμφωνα με αυτή τη μέθοδο, το αρχικό επίπεδο σημαντικότητας α διαιρείται δια του αριθμού n στατιστικών ελέγχων που πρόκειται να πραγματοποιηθούν, δηλαδή a* = a/n και κάθε αντιστοίχιση με p-value μικρότερο του a* θεωρείται μη στατιστικά σημαντική. Στην ανάλυση με τη χρήση g:GOSt, το επίπεδο σημαντικότητας είναι a = 0.05, ενώ η μεταβλητή n υποδηλώνει τον αριθμό των ανεξάρτητων δοκιμών, δηλαδή τον αριθμό των όρων GO, KEGG κλπ. που προκύπτουν για το δεδομένο δείγμα εισόδου. Από τη βιβλιογραφία προτείνονται δύο τρόποι υπολογισμού της μεταβλητής n. Η πρώτη προσέγγιση συμπεριλαμβάνει μόνο τους όρους στους οποίους συμπεριλαμβάνονται έστω κάποια από τα γονίδια/πρωτεΐνες του αρχείου εισόδου, αντίθετα η δεύτερη προτείνει ως n όλους τους σχολιασμένους όρους για το δεδομένο γονιδίωμα. Κατ επέκταση η λειτουργία g:GOSt, ακολουθεί την πρώτη προσέγγιση καθώς αυτό σημαίνει οτι η μεταβλητή a είναι της τάξης των εκατοντάδων και όχι των χιλιάδων, που θα ήταν στη δεύτερη

περίπτωση. Με τη διόρθωση Bonferroni οι στατιστικοί έλεγχοι γίνονται πιο συντηρητικοί, με αποτέλεσμα να ανιχνεύονται λιγότερα στατιστικά αποτελέσματα.

- **Benjamini-Hochberg FDR:** Η μέθοδος διόρθωσης FDR, είναι λιγότερο συντηρητική διαδικασία από τη διόρθωση Bonferroni, η οποία εκτιμά το αναμενόμενο ποσοστό των ψευδός θετικών αποτελεσμάτων (Σφάλμα τύπου I). Η μέθοδος αυτή λαμβάνει υπόψη όλες τις τιμές p-value που προκύπτουν οι οποίες στην συνέχεια ταξινομούνται κατά αύξουσα σειρά και και επιλέγει το a* να είναι η μεγαλύτερη τιμή p-value. Υπολογίζεται το γινόμενο i·p-value/n και της προσωρινής τιμής, όπου i ο αριθμός των πολλαπλών ερωτημάτων και n η σειρά ως προσωρινή τιμή ορίζεται το 1. Αρχίζοντας από το a*, κάθε διορθωμένη τιμή είναι η μικρότερη μεταξύ του γινομένου i·p-value/n και της προσωρινής τιμής. Κάθε αντιστοίχιση με τιμή p πάνω από το επίπεδο διόρθωσης Benjamini-Hochberg a* απορρίπτεται ως μη σημαντική.

4. **Επίπεδο στατιστικής σημαντικότητας απο το χρήστη (P-value correction cut-off:):** Μέσω της επιλογής αυτής παρέχεται η δυνατότητα επιπλέον φιλτραρίσματος των αποτελεσμάτων αναφορικά με την στατιστική σημαντικότητά τους. Το προεπιλεγμένο κατώφλι είναι p = 0.05, δηλαδή θα εμφανίζονται όλα τα στατιστικά σημαντικά αποτελέσματα. Στην περίπτωση που ο χρήστης ορίσει ως τιμή p = 0.01, θα εμφανιστούν μόνο τα αποτελέσματα που αντιστοιχούν σε τιμή p-value μικρότερη του 0.01. Το όριο που καθορίζεται από το χρήστη διαφέρει από significance threshold, καθώς δεν εμπλέκεται στην στατιστική σημασία των αποτελεσμάτων, αλλά είναι απλά ένα φίλτρο εμφάνισης των αποτελεσμάτων με βάση το p-value.

5. **Τύπος εξόδου δεδομένων:** Σε αντίθεση με τα περισσότερα εργαλεία, όπου ο χρήστης πρέπει να έχει καθορίσει τον τύπο αναγνωριστικού (ID) δεδομένων του, το FLAME μπορεί να υποστηρίξει και να αναλύσει μεικτές λίστες. Ταυτόχρονα όπως δίνεται η επιλογή στο χρήστη, μέσω του εργαλείου g:Convert (Ενότητα 2.7.1) να επιλέξει συγκεκριμένο τύπο αναγνωριστικού για την μορφή των αποτελεσμάτων. Έτσι, ασχέτως του τύπου εισόδου, μέσω της επιλογής Select ID type for output και IDs που παρέχονται για μετατροπή είναι τα εξείς: ChEMBL, Entrez Gene Name, Entrez Gene Accession, Entrez Gene Transcript Name, UniProt Accession, UniProt Gene Name, EMBL Accession, ENSEMBL Protein ID, ENSEMBL Gene ID, ENSEMBL Transcript ID, UniProt Archive, WIKIGENE ID, RefSeq mRNA, RefSeq mRNA Accession, RefSeq Protein Accession, RefSeq Non-coding RNA Accession. Ως προεπιλογή δίνεται ο αρχικός τύπος δεδομένων του χρήστη.



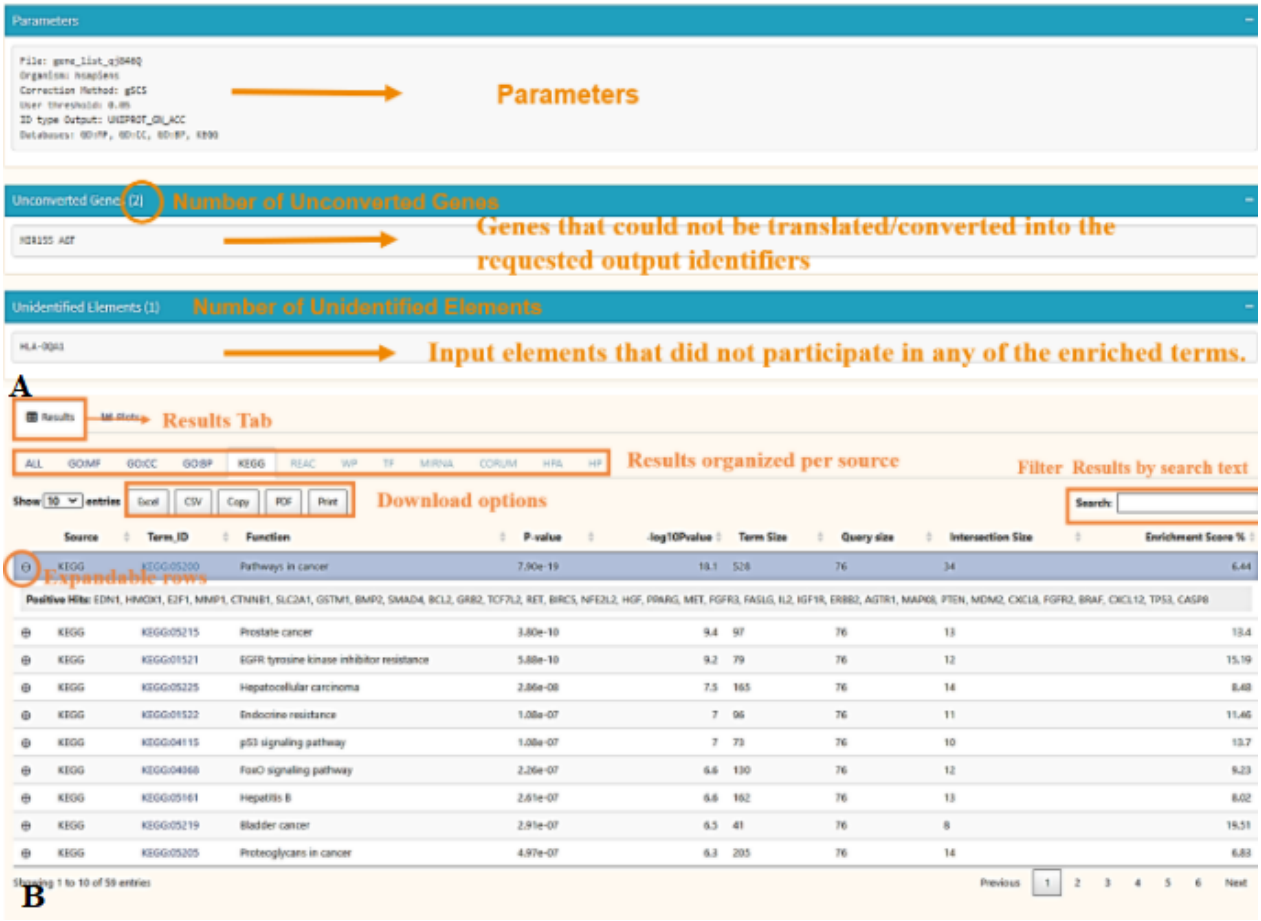
Εικόνα 2.5: Επιλογή παραμέτρων για τη λειτουργική ανάλυση εμπλουτισμού σε οντολογίες και μονοπάτια. Ο χρήστης πρέπει να επιλέξει σε ποιό από τα αρχεία θέλει να κάνει την ανάλυση και παραμέτρους αναφορικά με τον οργανισμό (επιλογή από 197 είδη), Στατιστική μέθοδο διόρθωσης (Significance threshold), επίπεδο στατιστικής σημαντικότητας απο το χρήστη (P-value correction cut-off:) καθώς και τον τύπο εξόδου έχοντας ώς επιλογές ένα μεγάλο εύρος αναγνωριστικών, πχ Entrez, UniProt, EMBL, ENSEMBL, ChEMBL, WikiGene και RefSeq.

Ο Πίνακας των αποτελεσμάτων (Εικόνα 2.6B) περιέχει τις εξής στήλες:

- **Source**: Η βάση δεδομένων από την οποία προέκυψε το αποτέλεσμα.
- **Term ID**: Το μοναδικό αναγνωριστικό ID του όρου/βιολογικής λειτουργίας στην αντίστοιχη βάση δεδομένων. Σχέδον σε όλες τις περιπτώσεις είναι ένας υπερσύνδεσμος που αναδρομολογεί το χρήστη στην σελίδα του όρου (term) της αντίστοιχής βάσης.
- **Function**: το όνομα ή σύντομη περιγραφή του όρου/λειτουργίας/μονοπατιού/φαινοτύπου κλπ που προέκυψαν από την ανάλυση εμπλουτισμού.
- **P-value**: η τιμή p-value που προέκυψε μετά από την επιλεγμένη μέθοδο διόρθωσης πολλαπλών υποθέσεων.
- **-log10p-value**: ο αρνητικός δεκαδικός λογάριθμος του p-value.
- **Term size**: ο αριθμός των γονιδίων που έχουν καταχωρηθεί στον συγκεκριμένο όρο
- **Query size**: ο αριθμός γονιδίων που δόθηκαν προς ανάλυση.
- **Intersection size**: ο αριθμός των γονιδίων της λίστας εισόδου που βρέθηκε να συμμετέχει στην αντίστοιχη Function.
- **Enrichment Score %**: είναι ο λόγος Intersection size/Term size*100%.
- **Positive Hits**: Τα γονίδια χωρισμένα με κόμμα που βρέθηκαν να συμμετέχουν στην αντίστοιχη Function. Η στήλη αυτή είναι “κρυμμένη” και για να την εμφανίσει ο χρήστης πρέπει να πατήσει τον σταυρό επέκτασης του πίνακα στα αριστερά κάθε σειράς.

Τα αποτελέσματα για κάθε βάση δεδομένων παρουσιάζονται σε ξεχωριστό Πίνακα στην καρτέλα Results (Αποτελέσματα) και συλλογικά σε έναν ενιαίο Πίνακα (ALL). Η κάθε βάση δεδομένων, όπως θα δούμε παρακάτω διατηρεί χρωματική κωδικοποίηση. Ο πίνακας μπορεί να αποθηκευτεί σε διάφορες μορφές για περαιτέρω χρήση. Εκτός από τα αποτελέσματα, προσφέρεται η επιλογή (Εικόνα 2.6A) ο χρήστης να βλέπει συγκεντρωμένες σε ένα πλαίσιο όλες τις επιλεγμένες παραμέτρους (Parameters) της ανάλυσης. Τέλος, για κάθε ανάλυση μπορούν να προκύψουν οι εξής πληροφορίες:

- **Unmatched genes**: Γονίδια που δεν μπορούσαν να μετατραπούν στο επιλεγμένο αναγνωριστικό εξόδου και παρέμειναν στην αρχική τους μορφή.
- **Unidentified Elements**: Γονίδια του αρχείου εισόδου, τα οποία δεν βρέθηκαν να συμμετέχουν σε κανένα αποτέλεσμα της ανάλυσης εμπλουτισμού.



Εικόνα 2.6: Παράδειγμα μορφής αποτελεσμάτων λειτουργικής ανάλυσης εμπλουτισμού σε οντολογίες και μονοπάτια. (A) Συγκεντρωτικός πίνακας επιλεγμένων παραμέτρων, καθώς και πίνακες επιπλέον πληροφοριών αναφορικά με: Unmatched genes, δηλαδή τα γονίδια που δεν μπορούσαν να μετατραπούν στο επιλεγμένο αναγνωριστικό εξόδου και παρέμειναν στην αρχική τους μορφή και Unidentified Elements, δηλαδή γονίδια του αρχείου εισόδου, τα οποία δεν βρέθηκαν να συμμετέχουν σε κανένα αποτέλεσμα της ανάλυσης εμπλουτισμού. (B) Διαδραστικός πίνακας αποτελεσμάτων για κάθε βάση δεδομένων που έχει επιλεχθεί ή συλλογικός. Ο πίνακας μπορεί να αποθηκευτεί σε διάφορες μορφές.

2.3.2 Domain και Diseases (aGOtool)

Αντίστοιχα με την ανάλυση εμπλουτισμού σε οντολογίες και μονοπάτια μέσω της βιβλιοθήκης R/gprofiler2, το FLAME παρέχει επίσης την ανάλυση εμπλουτισμού σε Domains και Diseases με τη χρήση της εφαρμογής aGotool μέσω API.

Ο χρήστης πρέπει όπως και στην περίπτωση του gprofiler2 να επιλέξει μια σειρά παραμέτρων (Εικόνα 2.7). Συγκεκριμένα:

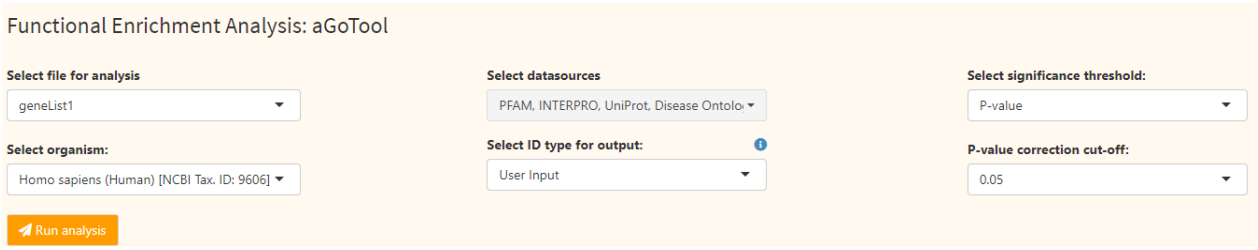
1. **Βάσεις Δεδομένων (Select datasources):** Μπορεί να επιλέξει μία η περισσότερες βάσεις δεδομένων από τις παρακάτω.
    - PFAM
    - INTERPRO
    - UniProt
    - Disease Ontology

2. **Οργανισμός (Select organism):** Μπορεί να επιλέξει μεταξύ 197 οργανισμών.

3. **Στατιστική μέθοδος διόρθωσης - Έλεγχος Πολλαπλών υποθέσεων (Significance threshold):** Η ανάλυση των αποτελεσμάτων της λειτουργικής ανάλυσης γονιδίων περιλαμβάνει ένα μεγάλο αριθμό ελέγχων πολλαπλών υποθέσεων ταυτόχρονα, καθώς κάθε αρχείο εισόδου συγκρίνεται με χιλιάδες όρους, πρωτεΐνες, μονοπάτια, ασθένειες, φαινοτύπους, λειτουργικά μοτίβα κλπ. Με τον έλεγχο των πολλαπλών υποθέσεων μειώνεται η πιθανότητα εμφάνισης ψευδών θετικών αποτελεσμάτων. Για τη λειτουργική ανάλυση των δεδομένων δίνεται η δυνατότητα στο χρήστη να επιλέξει μεταξύ p-value και corrected p-value (FDR).

4. **Τύπος εξόδου δεδομένων (Select ID type for output) :** Σε αντίθεση με τα περισσότερα εργαλεία, όπου ο χρήστης πρέπει να έχει καθορίσει τον τύπο αναγνωριστικού (ID) δεδομένων του, το FLAME μπορεί να υποστηρίξει και να αναλύσει μεικτές λίστες. Ταυτόχρονα όπως δίνεται η επιλογή στο χρήστη, μέσω του εργαλείου g:Convert (Ενότητα 2.7.1) να επιλέξει συγκεκριμένο τύπο αναγνωριστικού για την μορφή των αποτελεσμάτων. Έτσι, ασχέτως του τύπου εισόδου, μέσω της επιλογής Select ID type for output και IDs που παρέχονται για μετατροπή είναι τα εξείς: ChEMBL, Entrez Gene Name, Entrez Gene Accession, Entrez Gene Transcript Name, UniProt Accession, UniProt Gene Name, EMBL Accession, ENSEMBL Protein ID, ENSEMBL Gene ID, ENSEMBL Transcript ID, UniProt Archive, WIKIGENE ID, RefSeq mRNA, RefSeq mRNA Accession, RefSeq Protein Accession, RefSeq Non-coding RNA Accession. Ως προεπιλογή δίνεται ο αρχικός τύπος δεδομένων του χρήστη.

5. **Επίπεδο στατιστικής σημαντικότητας απο το χρήστη (P-value correction cut-off:):** Μέσω της επιλογής αυτής παρέχεται η δυνατότητα επιπλέον φιλτραρίσματος των αποτελεσμάτων αναφορικά με την στατιστική σημαντικότητά τους. Το προεπιλεγμένο κατώφλι είναι p = 0.05, δηλαδή θα εμφανίζονται όλα τα στατιστικά σημαντικά αποτελέσματα. Στην περίπτωση που ο χρήστης, για παράδειγμα ορίσει ως τιμή p = 0.01, θα εμφανιστούν μόνο τα αποτελέσματα που αντιστοιχούν σε τιμή p-value μικρότερη του 0.01. Το όριο που καθορίζεται από το χρήστη διαφέρει από significance threshold, καθώς δεν εμπλέκεται στην στατιστική σημασία των αποτελεσμάτων, αλλά είναι απλά ένα φίλτρο εμφάνισης των αποτελεσμάτων με βάση το p-value. Οι επιλογές που δίνονται είναι ένα εύρος τιμών από 0.1-0.01.



Εικονες 2.7: Επιλογή παραμέτρων για τη λειτουργική ανάλυση εμπλουτισμού σε domains και diseases. Ο χρήστης πρέπει να επιλέξει σε ποιό από τα αρχεία θέλει να κάνει την ανάλυση και παραμέτρους αναφορικά με τον οργανισμό (επιλογή από 197 είδη), Στατιστική μέθοδο διόρθωσης (Significance threshold), επίπεδο στατιστικής σημαντικότητας απο το χρήστη (P-value correction cut-off:) καθώς και τον τύπο εξόδου έχοντας ώς επιλογές ένα μεγάλο εύρος αναγνωριστικών, πχ Entrez, UniProt, EMBL, ENSEMBL, ChEMBL, WikiGene και RefSeq.

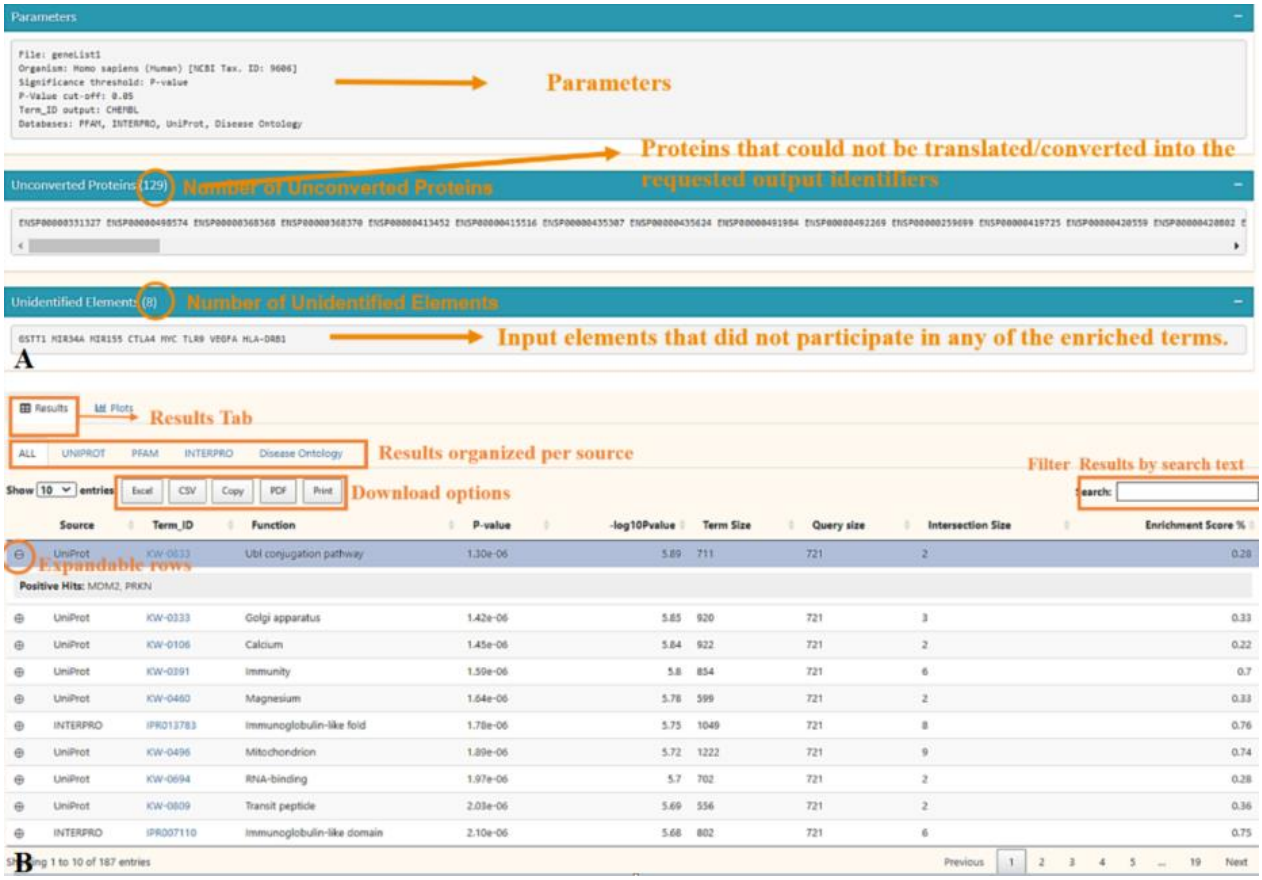
Ο Πίνακας των αποτελεσμάτων (Εικόνα 2.8B) περιέχει τις εξής στήλες:

- **Source**: Η βάση δεδομένων από την οποία προέκυψε το αποτέλεσμα.
- **Term ID**: Το μοναδικό αναγνωριστικό ID του όρου/βιολογικής λειτουργίας στην αντίστοιχη βάση δεδομένων. Σχέδον σε όλες τις περιπτώσεις είναι ένας υπερσύνδεσμος που αναδρομολογεί το χρήστη στην σελίδα του όρου (term) της αντίστοιχής βάσης.
- **Function**: το όνομα ή σύντομη περιγραφή του όρου/λειτουργίας/μονοπατιού/φαινοτύπου κλπ που προέκυψαν από την ανάλυση εμπλουτισμού.
- **P-value**: η τιμή p-value που προέκυψε μετά από την επιλεγμένη μέθοδο διόρθωσης πολλαπλών υποθέσεων.
- **-log10p-value**: ο αρνητικός δεκαδικός λογάριθμος του p-value.
- **Term size**: ο αριθμός των γονιδίων που έχουν καταχωρηθεί στον συγκεκριμένο όρο
- **Query size**: ο αριθμός γονιδίων που δόθηκαν προς ανάλυση.
- **Intersection size**: ο αριθμός των γονιδίων της λίστας εισόδου που βρέθηκε να συμμετέχει στην αντίστοιχη Function.
- **Enrichment Score %**: είναι ο λόγος Intersection size/Term size*100%.
- **Positive Hits**: Τα γονίδια χωρισμένα με κόμμα που βρέθηκαν να συμμετέχουν στην αντίστοιχη Function. Η στήλη αυτή είναι "κρυμμένη" και για να την εμφανίσει ο χρήστης πρέπει να πατήσει τον σταυρό επέκτασης του πίνακα στα αριστερά κάθε σειράς.

Τα αποτελέσματα για κάθε βάση δεδομένων παρουσιάζονται σε ξεχωριστό Πίνακα στην καρτέλα Results (Αποτελέσματα) και συλλογικά σε έναν ενιαίο Πίνακα (ALL). Η κάθε βάση δεδομένων, όπως θα δούμε παρακάτω διατηρεί χρωματική κωδικοποίηση. Ο πίνακας μπορεί να αποθηκευτεί σε διάφορες μορφές για περαιτέρω χρήση. Εκτός από τα αποτελέσματα, προσφέρεται η επιλογή (Εικόνα 2.8A) ο χρήστης να βλέπει συγκεντρωμένες σε ένα πλαίσιο όλες τις επιλεγμένες παραμέτρους (Parameters) της ανάλυσης. Τέλος, για κάθε ανάλυση μπορούν να προκύψουν οι εξής πληροφορίες:

- **Unmatched Proteins**: Πρωτεΐνες που δεν μπορούσαν να μετατραπούν στο επιλεγμένο αναγνωριστικό εξόδου και παρέμειναν στην αρχική τους μορφή.

- **Unidentified Elements**: Πρωτεΐνες του αρχείου εισόδου, τα οποία δεν βρέθηκαν να συμμετέχουν σε κανένα αποτέλεσμα της ανάλυσης εμπλουτισμού.



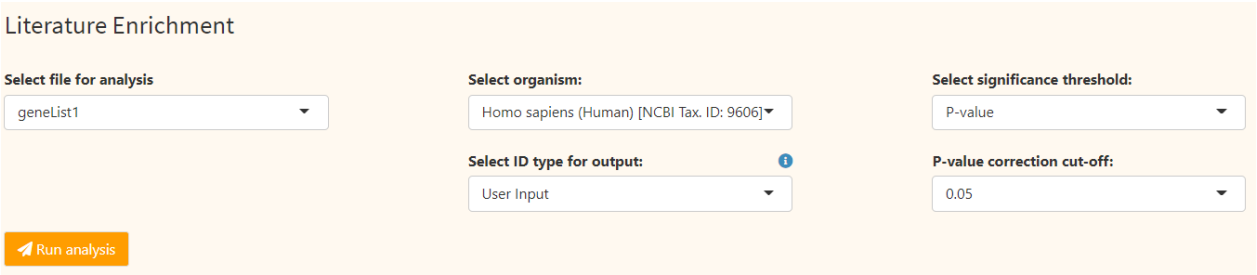
Εικόνα 2.8: Παράδειγμα μορφής αποτελεσμάτων λειτουργικής ανάλυσης εμπλουτισμού σε domains και diseases. (A) Συγκεντρωτικός πίνακας επιλεγμένων παραμέτρων, καθώς και πίνακες επιπλέον πληροφοριών αναφορικά με: Unmatched proteins, δηλαδή πρωτεΐνες που δεν μπορούσαν να μετατραπούν στο επιλεγμένο αναγνωριστικό εξόδου και παρέμειναν στην αρχική τους μορφή και Unidentified Elements, δηλαδή γονίδια/πρωτεΐνες του αρχείου εισόδου, τα οποία δεν βρέθηκαν να συμμετέχουν σε κανένα αποτέλεσμα της ανάλυσης εμπλουτισμού. (B) Διαδραστικός πίνακας αποτελεσμάτων για κάθε βάση δεδομένων που έχει επιλεχθεί ή συλλογικός. Ο πίνακας μπορεί να αποθηκευτεί σε διάφορες μορφές.

2.4 Ανάλυση εμπλουτισμού βιβλιογραφίας (Literature enrichment analysis, aGOtool)

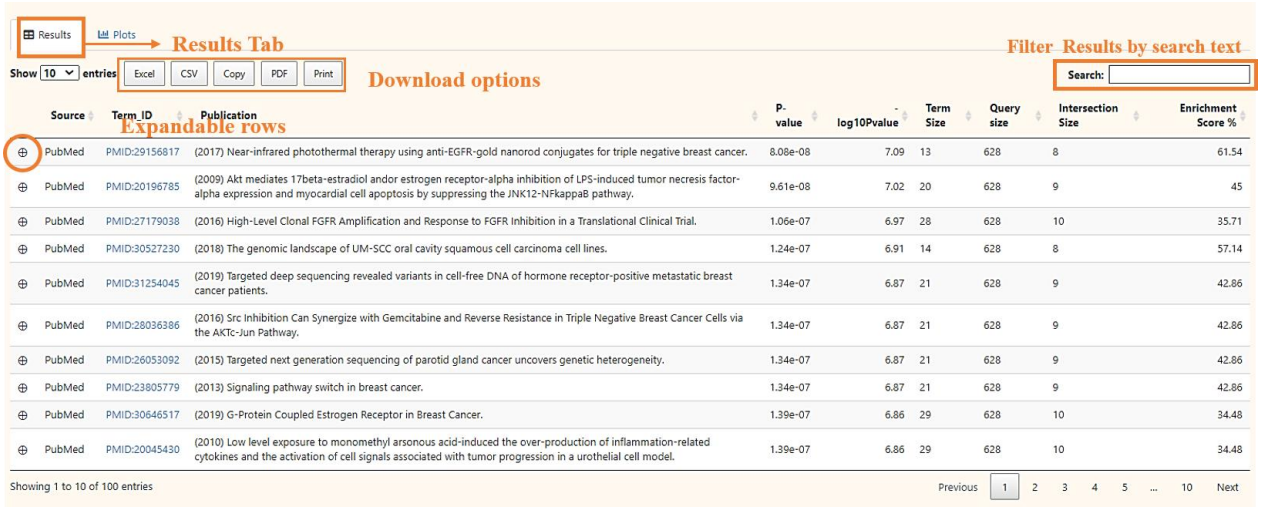
Το FLAME εκτός από την δυνατότητα της λειτουργικής ανάλυσης εμπλουτισμού της γονιδιακής έκφρασης (Functional Enrichment Analysis), παρέχει την επιλογή ανάλυσης εμπλουτισμού βιβλιογραφίας/δημοσιεύσεων (Literature enrichment analysis). Η έννοια της ανάλυσης εμπλουτισμού βιβλιογραφίας είναι παρόμοια με αυτή της λειτουργικής ανάλυσης εμπλουτισμού της γονιδιακής έκφρασης. Το FLAME επιτρέπει στους χρήστες να ανακτήσουν επιστημονικές δημοσιεύσεις που συνδέονται με σημαντικά στατιστική συσχέτιση με τις λίστες γονιδίων/πρωτεϊνών που έχουν εισαχθεί ως αρχεία εισόδου και μελετάει ο χρήστης. Η ανάλυση στην

προκειμένη βασίζεται στην εφαρμογή aGotool μέσω API, η οποία χρησιμοποιεί ένα σύνολο κειμένων όλων των περιλήψεων της PubMed και των άρθρων ανοικτής πρόσβασης του πλήρους κειμένου από το PubMed Central. Οι επιστημονικές δημοσιεύσεις, καθώς και οι περιλήψεις που χρησιμοποιούνται, επεξεργάζονται μέσω των εφαρμογών OnTheFly's ή EXTRACT's ή Named Entity Recognition (NER) σε εβδομαδιαία βάση για τον εντοπισμό βιολογικών οντοτήτων/όρων (γονίδια/πρωτεΐνες, χημικές ενώσεις, οργανισμούς, ιστούς, περιβάλλοντα, ασθένειες, φαινότυπους και όρους GO). Ως αποτέλεσμα, σε όλα τα έγγραφα σχολιάζονται αυτόματα τα γονίδια που αναφέρονται μετατρέποντας κάθε έγγραφο σε «σύνολο γονιδίων».

Ο χρήστης, όπως στην περίπτωση λειτουργικής ανάλυσης εμπλουτισμού (gProfiler, aGOtool) πρέπει να επιλέξει σε ποιό από τα αρχεία θέλει να κάνει την ανάλυση και παραμέτρους αναφορικά με τον οργανισμό (επιλογή από 197 είδη), τιμές στατιστικής σημαντικότητας (Significance threshold: p-value ή FDR), κατώφλι (P-value correction cut-off) καθώς και τον τύπο εξόδου έχοντας ώς επιλογές ένα μεγάλο εύρος αναγνωριστικών, πχ Entrez, UniProt, EMBL, ENSEMBL, ChEMBL, WikiGene και RefSeq (Εικόνα 2.9). Τα αποτελέσματα, όπως φαίνονται στην Εικόνα 2.10 παρουσιάζονται με την μορφή αντίστοιχου διαδραστικού πίνακα, όπως στην περίπτωση της λειτουργικής ανάλυσης.



Εικόνα 2.9: Επιλογή παραμέτρων για την ανάλυση εμπλουτισμού βιβλιογραφίας. Ο χρήστης, όπως στην περίπτωση λειτουργικής ανάλυσης εμπλουτισμού (gProfiler, aGOtool) πρέπει να επιλέξει σε ποιό από τα αρχεία θέλει να κάνει την ανάλυση και παραμέτρους αναφορικά με τον οργανισμό (επιλογή από 197 είδη), τιμές στατιστικής σημαντικότητας (Significance threshold: p-value ή FDR), κατώφλι (P-value correction cut-off) καθώς και τον τύπο εξόδου έχοντας ώς επιλογές ένα μεγάλο εύρος αναγνωριστικών, πχ Entrez, UniProt, EMBL, ENSEMBL, ChEMBL, WikiGene και RefSeq.



Εικόνα 2.10: Παράδειγμα μορφής αποτελεσμάτων της ανάλυση εμπλουτισμού βιβλιογραφίας.

2.5 Δίκτυα Πρωτεϊνικών Αλληλεπιδράσεων - STRING

Το FLAME προσφέρει τη δυνατότητα δημιουργίας διαδραστικών δικτύων πρωτεϊνικών αλληλεπιδράσεων (Protein-Protein Interaction Networks, PPI) για το σύνολο των 197 οργανισμών χρησιμοποιώντας το STRING API. Οι χρήστες μπορούν χρησιμοποιώντας το αρχείο εισόδου τους να απεικονίσουν τα αποτελέσματα ως δίκτυα, στα οποία τα στοιχεία/οντότητες (entities) που αλληλεπιδρούν παρουσιάζονται ως κόμβοι (nodes) και οι αλληλεπιδράσεις μεταξύ τους ως ακμές (edges). Στην online έκδοση, το FLAME επιτρέπει μέγιστο αριθμό πρωτεϊνών στο αρχείο εισόδου τις 500 πρωτεΐνες. Ο περιορισμός αυτός μπορεί να παρακαμφθεί, καθώς δίνεται η δυνατότητα στο χρήστη να κατεβάσει την εφαρμογή μέσω GitHub και να τροποποιήσει την αντίστοιχη μεταβλητή. Συγκεκριμένα, πρέπει να τροποποιήσει τη μεταβλητή STRING_LIMIT στο αρχείο global.R και έπειτα να εκτελέσει τοπικά την εφαρμογή.

Παρά το γεγονός ότι η STRING μπορεί να δεχτεί ως όρισμα διάφορους τύπους αναγνωριστικών για να αποφευχθούν οι χρονικές καθυστερήσεις ή τυχόν σφάλματα αδυναμίας αντιστοίχισης τους απο την STRING συνίσταται η μετατροπή τους σε εναν identifier με συγκεκριμένη μορφή, για παράδειγμα 9606.ENSP00000269305 για την ανθρώπινη TP53. Για το λόγο αυτό, έχει ενσωματωθεί στο υπόβαθρο το εργαλείο g:Convert (βλ. Ενότητα 2.7.1), όπου μετατρέπει αυτόματα το περιεχόμενο της λίστας εισόδου του χρήστη σε Ensembl Protein ID. Να σημειωθεί ότι από τη μετατροπή μπορεί ένα γονίδιο/πρωτεΐνη να αντιστοιχεί σε παραπάνω από ένα ID, γι'αυτό διατηρήθηκε μόνο ένα μετατρεπόμενο αναγνωριστικό Ensembl για κάθε πρωτεινή.

Η STRING υποστηρίζει τόσο τις άμεσες (φυσικές), όσο και τις έμμεσες (λειτουργικές) αλληλεπιδράσεις. Οι φυσικές αλληλεπιδράσεις αναφέρονται σε πρωτεΐνες που αποτελούν μέρος του ίδιου βιομοριακού συμπλέγματος, ενώ οι λειτουργικές αλληλεπιδράσεις αναφέρονται σε πρωτεΐνες οι οποίες εμπλέκονται στο ίδιο μονοπάτι ή βιολογική διαδικασία. Μέσω του FLAME, οι χρήστες μπορούν να θέσουν τις επιθυμητές παραμέτρους (Εικόνα 2.11). Συγκεκριμένα μπορούν να επιλέξουν (select interaction type) εάν θα οπτικοποιήσουν το πλήρες σύνολο (full network) αλληλεπιδράσεων (φυσικών και λειτουργικών) ή απλώς το φυσικό υποδίκτυο (physical sub-network). Επιπλέον, δίνεται η επιλογή να ορίσουν το σκορ της ελάχιστης απαιτούμενης αλληλεπίδρασης (Select interaction score cut-off), δηλαδή να θέσουν ένα κατώφλι στο σκορ εμπιστοσύνης (confidence score), έτσι ώστε τιμές κάτω από αυτό να αποκλείονται από την πρόβλεψη. Ένα χαμηλό σκορ μπορεί να σημαίνει περισσότερες αλληλεπιδράσεις, αλλά ταυτόχρονα περισσότερα ψευδώς θετικά αποτελέσματα. Το σκορ εμπιστοσύνης είναι η κατά προσέγγιση πιθανότητα της ύπαρξης της προβλεπόμενης σύνδεσης μεταξύ δύο ενζύμων στο ίδιο μεταβολικό χάρτη στη βάση δεδομένων KEGG. Οι επιλογές που δίνονται στο χρήστη είναι οι εξής:

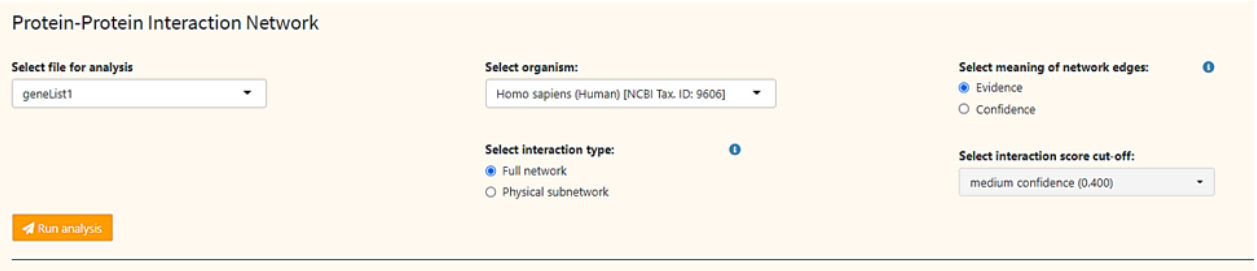
- χαμηλής εμπιστοσύνης - 0.15
- μεσαίας εμπιστοσύνης - 0.4
- υψηλής εμπιστοσύνης - 0.7
- υψηλότερης εμπιστοσύνης - 0.9

Επίσης, δίνεται η δυνατότητα να επιλέξουν τι θα καθορίζουν οι ακμές στην απεικόνισης του δικτύου (Select meaning of network edges). Συγκεκριμένα, οι επιλογές αφορούν είτε τον τύπο των αλληλεπιδράσεων (Evidence mode), είτε τον βαθμό εμπιστοσύνης (Confidence mode). Στην πρώτη περίπτωση, προκύπτει ένα δίκτυο πολλαπλών ακμών (multi-edged graph), όπου το χρώμα των ακμών καθορίζει τον τύπο των αλληλεπιδράσεων (evidence). Συγκεκριμένα, οι ακμές μπορούν να χρωματιστούν με 7 χρώματα:

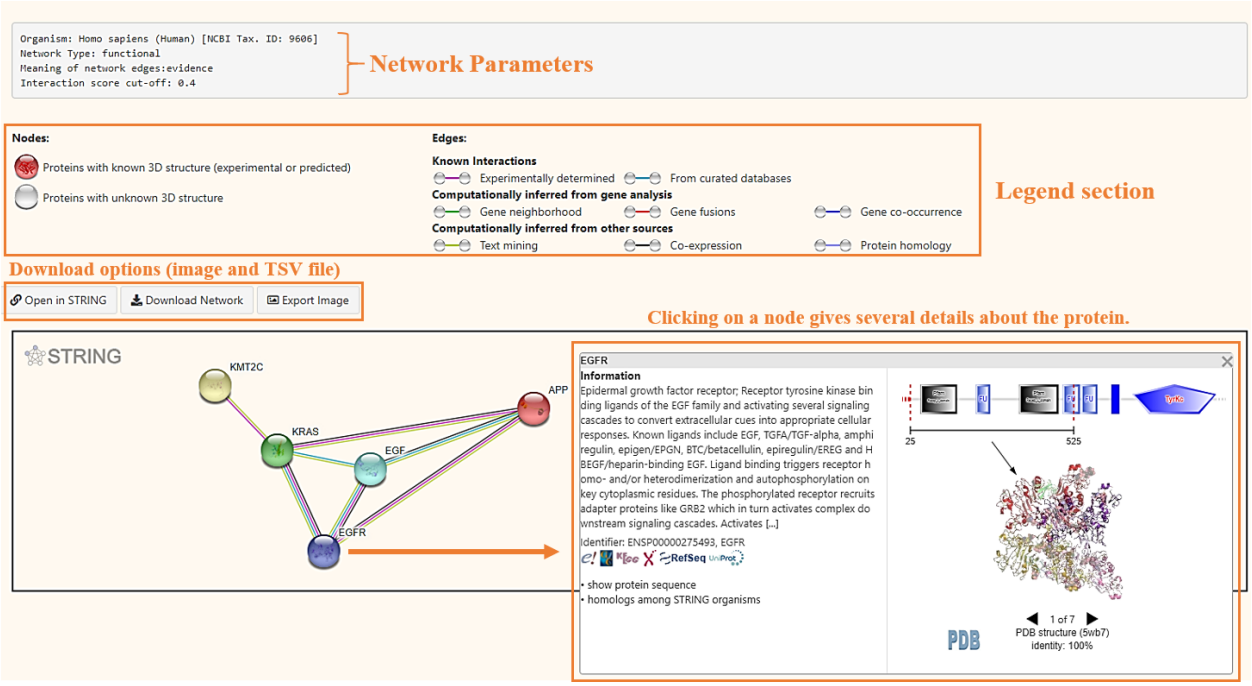
- Κόκκινο - Προβλέψεις γονιδιωματικού περιεχομένου για περιπτώσεις ένωσης/σύντηξης (indicates the presence of fusion evidence)
- Πράσινο - Προβλέψεις γονιδιωματικού περιεχομένου για περιπτώσεις γειτνίασης (neighborhood evidence)
- Μπλε - Προβλέψεις γονιδιωματικού περιεχομένου για περιπτώσεις συνεμφάνισης (co-occurrence evidence)
- Μωβ - Εργαστηριακά πειράματα μεγάλης κλίμακας (experimental evidence)
- Κίτρινο - Αυτοματοποιημένη εξόρυξη κειμένου (text mining evidence)
- Γαλάζιο - από άλλες βάσεις δεδομένων (database evidence)
- Μαύρο - Συντηρημένη συν-έκφραση (coexpression evidence).

Στη δεύτερη περίπτωση, το πάχος των ακμών αντανακλά τη βαθμολογία αλληλεπίδρασης.

Το δίκτυο που προκύπτει εμφανίζεται σε πλαίσιο (Εικόνα 2.12) κάτω από τις παραμέτρους. Είναι διαδραστικό με δυνατότητα μετακίνησης των κόμβων και πατώντας σε κάποιο επιθυμητό κόμβο προβάλλει ένα πλαίσιο με πληροφορίες σχετικά με την πρωτεΐνη που αντιστοιχεί στον κόμβο, καθώς και υπόμνημα ερμηνείας των χρωμάτων των ακμών και των κόμβων. Τέλος, δίνεται η επιλογή αποθήκευσης σε μορφή εικόνας (.png), αρχείου tab-delimited (.tsv) και αναδρομολόγησης στην επίσημη σελίδα της STRING.



Εικόνα 2.11: Επιλογή παραμέτρων για τη δημιουργία δικτύου πρωτεϊνικών αλληλεπιδράσεων (PPI Network) μέσω STRING API. Ο χρήστης έχει δυνατότητα επιλογής του οργανισμού, του τύπο των αλληλεπιδράσεων (select interaction type), δηλαδή εάν θα οπτικοποιήσουν το πλήρες σύνολο (full network) αλληλεπιδράσεων (φυσικών και λειτουργικών) ή απλώς το φυσικό υποδίκτυο (physical sub-network), το τι αντιπροσωπεύουν οι ακμές (Select meaning of network edges) δηλαδή είτε τον τύπο των αλληλεπιδράσεων (Evidence mode), είτε τον βαθμό εμπιστοσύνης (Confidence mode), καθώς και το σκορ της ελάχιστης απαιτούμενης αλληλεπίδρασης (Select interaction score cut-off), δηλαδή να θέσουν ένα κατώφλι στο σκορ εμπιστοσύνης (confidence score), έτσι ώστε τιμές κάτω από αυτό να αποκλείονται από την πρόβλεψη.



Εικόνα 2.12: Διαδραστικό δίκτυο πρωτεϊνικών αλληλεπιδράσεων (Protein-Protein Interaction Networks, PPI) μέσω STRING API βάσει των επιλεγμένων παραμέτρων. Διαδραστικό δίκτυο με δυνατότητα μετακίνησης των κόμβων και πατώντας σε κάποιο επιθυμητό κόμβο προβάλλει ένα πλαίσιο με πληροφορίες σχετικά με την πρωτεΐνη που αντιστοιχεί στον κόμβο, καθώς και υπόμνημα ερμηνείας των χρωμάτων των ακμών και των κόμβων. Τέλος, δίνεται η επιλογή αποθήκευσης σε μορφή εικόνας (.png), αρχείου tab-delimited (.tsv) και αναδρομολόγησης στην επίσημη σελίδα της STRING.

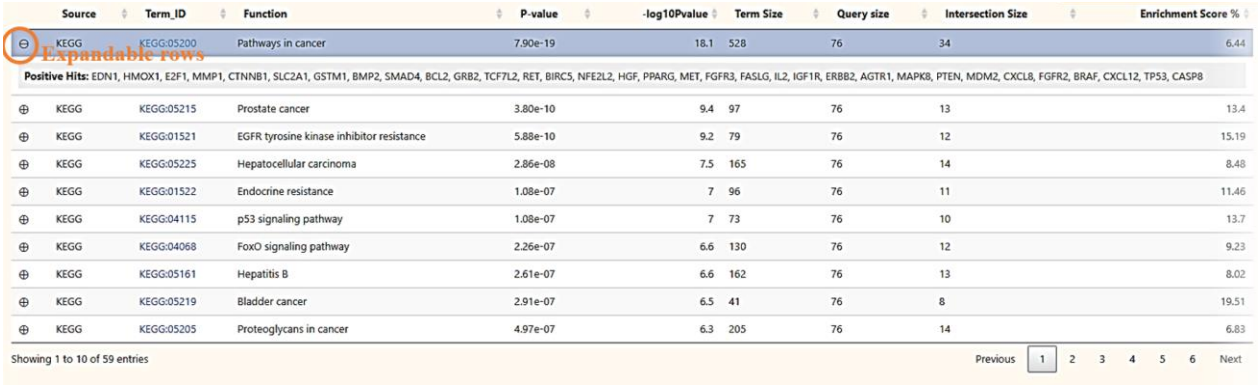
2.6 Γραφικές Απεικονίσεις Αποτελεσμάτων

Το FLAME παρέχει μια σειρά διαδραστικών και παραμετροποιήσιμων διαγραμμάτων για την καλύτερη απεικόνιση των αποτελεσμάτων που προκύπτουν από τις αναλύσεις του λειτουργικού (g:Profiler και aGOtool) και του βιβλιογραφικού εμπλουτισμού (aGOtool). Συγκεκριμένα τα αποτελέσματα μπορούν να απεικονιστούν με τη χρήση πινάκων, διαγραμμάτων διασποράς (scatter Plots), ραβδογραμμάτων (barcharts), θερμικών χαρτών (heatmaps) και δικτύων (networks).

2.6.1 Πίνακες αποτελεσμάτων

Τα αποτελέσματα από τις αναλύσεις εμπλουτισμού που προκύπτουν σε κάθε περίπτωση παρουσιάζονται με την μορφή διαδραστικών πινάκων (Εικόνα 2.13), οι οποίοι περιλαμβάνουν ενδελεχείς πληροφορίες για κάθε όρο που προέκυψε, με δυνατότητα αναζήτησεις με βάση όρους,γονίδια κλπ. Οι πίνακες μέσω του συμβόλου (+) μπορούν να επεκταθούν δίνοντας τη δυνατότητα στο χρήστη να δει τα γονίδια/πρωτεΐνες (positive hits) που βρέθηκαν στον συγκεκριμένο όρο. Για τη δημιουργία των πινάκων χρησιμοποιήθηκε η βιβλιοθήκη DT.
Για παράδειγμα, στην περίπτωση των αποτελεσμάτων της KEGG , μπορεί κανείς να δει πόσες πρωτεΐνες ή γονίδια βρέθηκαν ότι σχετίζονται με κάποιο συγκεκριμένο μονοπάτι και να ανακατευθυνθεί στην επίσημη ιστοσελίδα της KEGG για να δει το σχήμα του μονοπατιού σε

στατική μορφή με όλα τα γονίδια/πρωτεΐνες που εντοπίστηκαν να είναι επισημασμένα με άλλο χρώμα. Τέλος, όλοι οι πίνακες μπορούν να αποθηκευτούν σε διάφορες μορφές (.pdf, .xlsx, .csv).



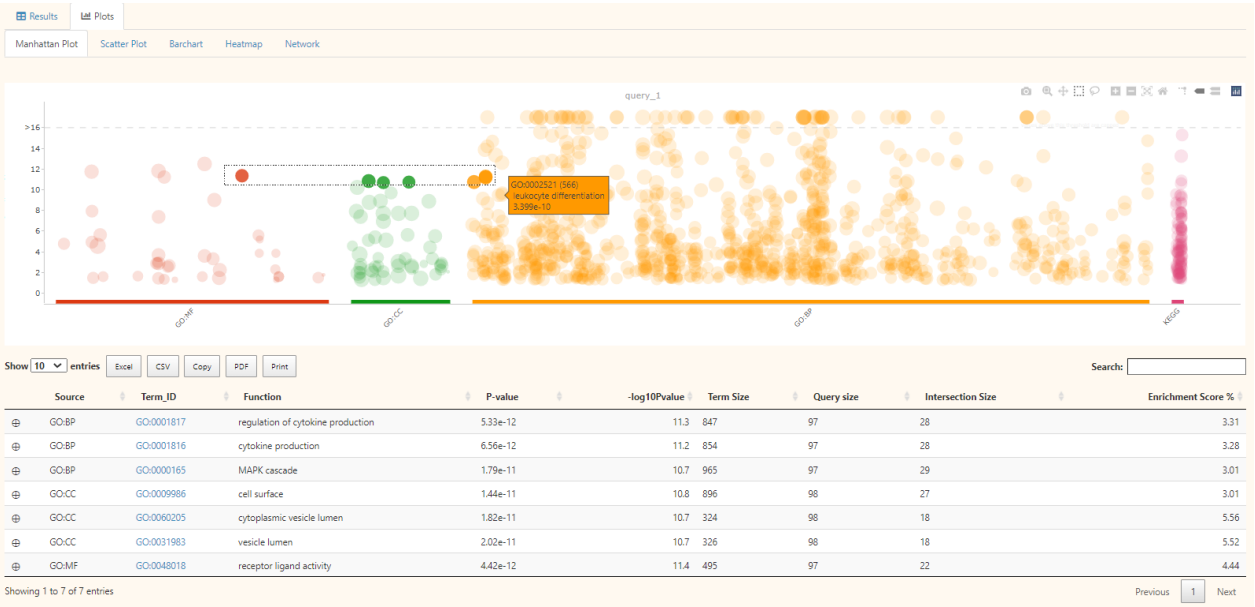
	Source	Term_ID	Function	P-value	-log10Pvalue	Term Size	Query size	Intersection Size	Enrichment Score %
⊖	KEGG	KEGG:05200	Pathways in cancer	7.90e-19	18.1	528	76	34	6.44
	Positive Hits: EDN1, HMOX1, E2F1, MMP1, CTNNB1, SLC2A1, GSTM1, BMP2, SMAD4, BCL2, GRB2, TCF7L2, RET, BIRC5, NFE2L2, HGF, PPARG, MET, FGFR3, FASLG, IL2, IGF1R, ERBB2, AGTR1, MAPK8, PTEN, MDM2, CXCL8, FGFR2, BRAF, CXCL12, TP53, CASP8								
⊕	KEGG	KEGG:05215	Prostate cancer	3.80e-10	9.4	97	76	13	13.4
⊕	KEGG	KEGG:01521	EGFR tyrosine kinase inhibitor resistance	5.88e-10	9.2	79	76	12	15.19
⊕	KEGG	KEGG:05225	Hepatocellular carcinoma	2.86e-08	7.5	165	76	14	8.48
⊕	KEGG	KEGG:01522	Endocrine resistance	1.08e-07	7	96	76	11	11.46
⊕	KEGG	KEGG:04115	p53 signaling pathway	1.08e-07	7	73	76	10	13.7
⊕	KEGG	KEGG:04068	FoxO signaling pathway	2.26e-07	6.6	130	76	12	9.23
⊕	KEGG	KEGG:05161	Hepatitis B	2.61e-07	6.6	162	76	13	8.02
⊕	KEGG	KEGG:05219	Bladder cancer	2.91e-07	6.5	41	76	8	19.51
⊕	KEGG	KEGG:05205	Proteoglycans in cancer	4.97e-07	6.3	205	76	14	6.83

Showing 1 to 10 of 59 entries — Previous 1 2 3 4 5 6 Next

Εικόνα 2.13: Πίνακας αποτελεσμάτων ανάλυσης λειτουργικού εμπλουτισμού.

2.6.2 Διάγραμμα Manhattan

Μόνο στην περίπτωση της λειτουργικής ανάλυσης εμπλουτισμού (g:Profiler), παρέχεται η δυνατότητα απεικόνισης μέσω διαδραστικού διαγράμματος Manhattan βασισμένο στα αποτελέσματα των επιλεγμένων βάσεων δεδομένων. Το συγκεκριμένο διάγραμμα κατασκευάστηκε με τη χρήση των βιβλιοθηκών Plotly και R/gprofiler2. Συγκεκριμένα, πρόκειται ουσιαστικά για έναν τύπο γραφικής παράστασης διασποράς, που χρησιμοποιείται συνήθως για την προβολή δεδομένων με μεγάλο αριθμό σημείων δεδομένων, πολλά μη μηδενικού πλάτους και με κατανομή τιμών υψηλότερου μεγέθους. Στο εν λόγω διάγραμμα (Εικόνα 2.14), στο x αξονα απεικονίζονται οι λειτουργικοί όροι (functional terms) χρωματισμένοι ανάλογα με τη βάση δεδομένων στην οποία αντιστοιχούν. Ο κάθετος y άξονας αντιστοιχεί στον αρνητικό δεκαδικό λογάριθμο των διορθωμένων τιμών p-value (-log10(p-value)). Σημεία που εμφανίζονται ψηλά στο διάγραμμα αντιστοιχούν έτσι σε λειτουργικούς όρους με μικρή τιμή p-value που είναι κατά συνέπεια στατιστικά σημαντικά για τη μελετώμενη γονιδιακή λίστα του χρήστη. Με την τοποθέτηση του κέρσορα πάνω από κάποιο σημείο ανοίγει ένα πλαίσιο με βασικές πληροφορίες αναφορικά με το λειτουργικο όρο. Επίσης, επιλέγοντας ένα ή περισσότερα σημεία με τον λάσσο ή το ορθογώνιο επιλογής (οι επιλογες αυτές υπάρχουν στο πάνω δεξιά τμήμα του διαγράμματος) εμφανίζεται ένας πίνακας κάτω από το διάγραμμα με τις πληροφορίες για τα εν λόγω επιλεγμένα σημεία. Κάθε φορά που νέα σημεία επιλέγονται ο πίνακας ανανεώνεται αυτόματα. Ταυτόχρονα, δίνονται επιλογές στον χρήστη για λήψη του διαγράμματος ως εικόνα (.png), εστίασης (zoom) σε συγκεκριμένα τμήματα του, επαναναδιάταξης αξόνων κλπ.

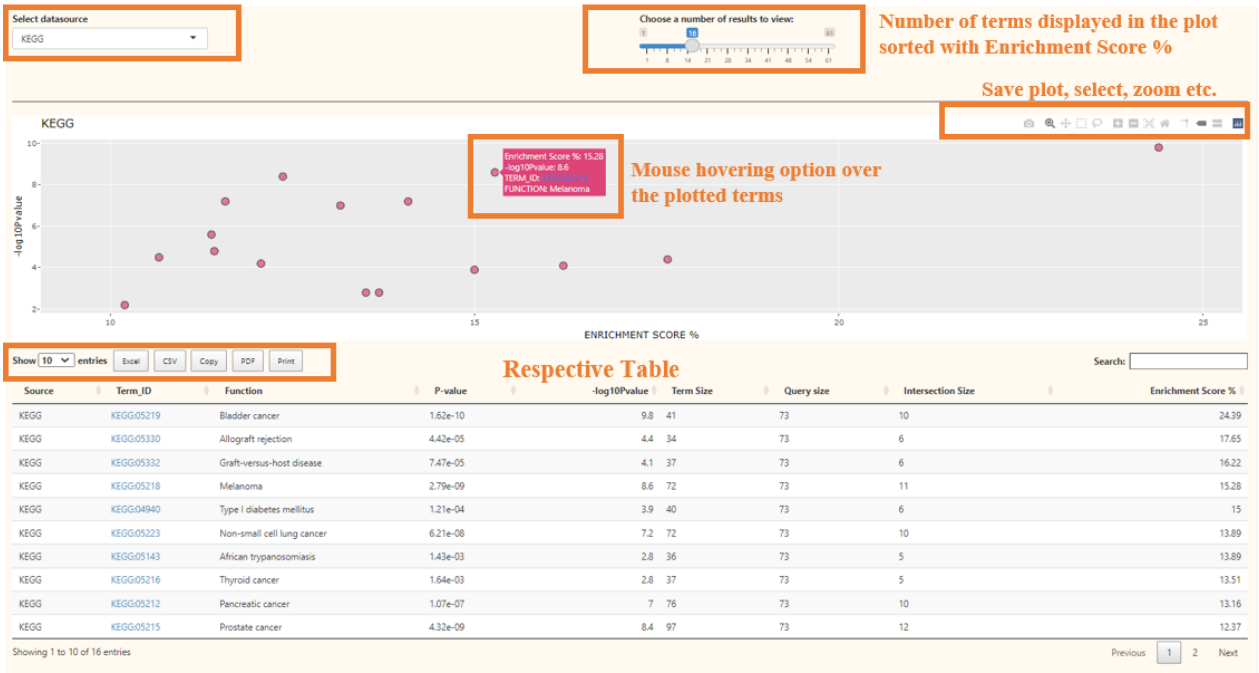


Εικόνα 2.14: Διάγραμμα Manhattan απεικόνισης των αποτελεσμάτων της λειτουργικής ανάλυσης εμπλουτισμου. Στο x αξονα απεικονίζονται οι λειτουργικοί όροι (functional terms) χρωματισμένοι ανάλογα με τη βάση δεδομένων στην οποία αντιστοιχούν. Ο κάθετος y άξονας αντιστοιχεί στον αρνητικό δεκαδικό λογάριθμο των διορθωμένων τιμών p-value (-log10(p-value)).Το διάγραμμα είναι διαδραστικό με δυνατότητα επιλογής συγκεκριμένων σημείων και εμφάνισης αντίστοιχου πίνακα με τις πληροφορίες των επιλεγμένων σημείων.

2.6.3 Διάγραμμα Διασποράς (Scatter Plot)

Η απεικόνιση αποτελεσμάτων με διάγραμμα διασποράς παρέχεται τόσο στην περίπτωση ανάλυσης λειτουργικού εμπλουτισμού της γονιδιακής έκφρασης (g:Profiler και aGOtool), όσο και στην ανάλυση εμπλουτισμού βιβλιογραφίας (aGOtool). Το διάγραμμα κατασκευάστηκε μέσω της βιβλιοθήκης Plotly. Πρόκειται για ένα διάγραμμα (Εικόνα 2.15), όπου ο x άξονας αντιστοιχεί στο Enrichment Score % (βλ. Ενότητα 2.3), ενώ ο y άξονας αντιστοιχεί στον αρνητικό δεκαδικό λογάριθμο των τιμών p-value (-log10(p-value)). Αυτό σημαίνει ότι σημεία που εμφανίζονται στο διάγραμμα πιο ψηλά στον y άξονα και πιο δεξιά στο x άξονα αντιστοιχούν σε όρους με μικρή τιμή p-value και μεγάλη τιμή Enrichment Score, αντίστοιχα, που κατά συνέπεια καθιστά αυτούς τους όρους πιο σημαντικούς. Ο χρήστης μπορεί να επιλέξει για ποιά από τις βάσεις δεδομένων επιθυμεί να δημιουργήσει το διάγραμμα καθώς και να προσαρμόσει τον αριθμό των αποτελεσμάτων στον αριθμό των όρων που επιθυμεί να απεικονισθεί πάνω στο διάγραμμα μέσω του φίλτρου (Choose a number of results to view). Τα αποτελέσματα έχουν καταταχθεί με βάση το Enrichment Score σε φθίνουσα σειρά, δηλαδή αν ο χρήστης επιλέξει να απεικονίσει ένα σημείο, ο όρος που αντιστοιχεί στο σημείο αυτό θα έχει το μεγαλύτερο Enrichment Score. Ανάλογα με το N αριθμό των αποτελεσμάτων που επιλέγει, εμφανίζεται ο αντίστοιχος πίνακας με πληροφορίες για όλους τους N όρους. Ο πίνακας σε κάθε αλλαγή του αριθμού των αποτελεσμάτων ανανεώνεται αυτόματα. Με την τοποθέτηση του κέρσορα πάνω από κάποιο σημείο ανοίγει ένα πλαίσιο με βασικές πληροφορίες αναφορικά με το λειτουργικο όρο. Ταυτόχρονα, δίνονται επιλογές στον χρήστη για

λήψη του διαγράμματος ως εικόνα (.png), εστίασης (zoom) σε συγκεκριμένα τμήματα του, επαναναδιάταξης αξόνων κλπ., καθώς και επιλογές αποθήκευσης του πίνακα.

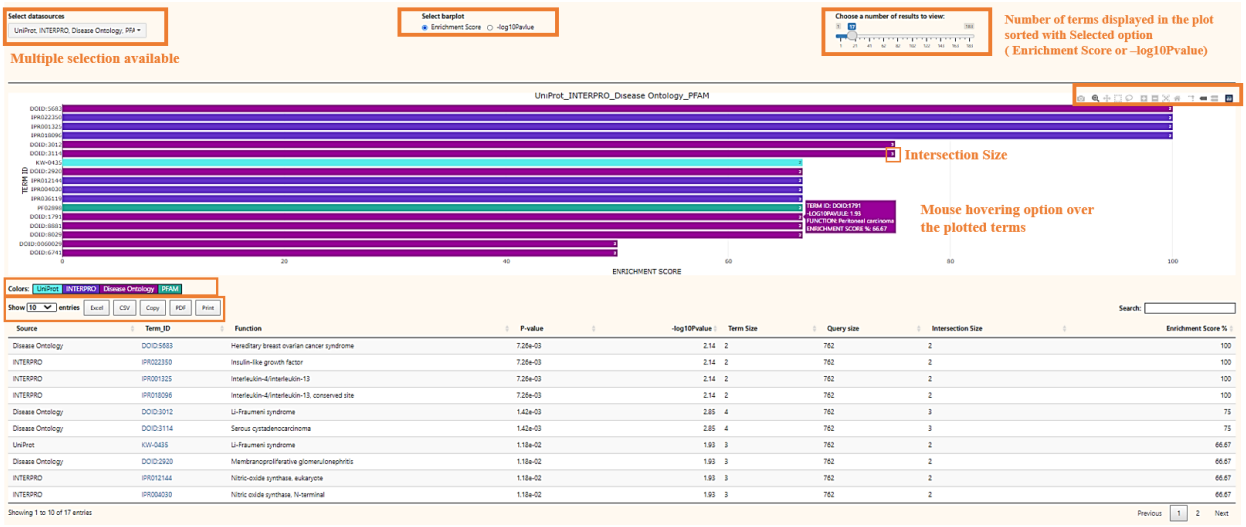


Εικόνα 2.15: Διάγραμμα διασποράς απεικόνισης των αποτελεσμάτων της λειτουργικής και βιβλιογραφικής ανάλυσης εμπλουτισμου. Στο x άξονας αντιστοιχεί στο Enrichment Score %, ενώ ο y άξονας αντιστοιχεί στον αρνητικό δεκαδικό λογάριθμο των τιμών p-value (-log10(p-value)). Το διάγραμμα είναι διαδραστικό και παραμετροποιήσιμο, δηλαδή χρήστης μπορεί να επιλέξει για ποιά από τις βάσεις δεδομένων επιθυμεί να δημιουργήσει το διάγραμμα καθώς και να προσαρμόσει τον αριθμό των αποτελεσμάτων στον αριθμό των όρων που επιθυμεί να απεικονισθεί πάνω στο διάγραμμα μέσω του φίλτρου (Choose a number of results to view). Ανάλογα με το N αριθμό των αποτελεσμάτων που επιλέγει, εμφανίζεται ο αντίστοιχος πίνακας με πληροφορίες για όλους τους N όρους. Με την τοποθέτηση του κέρσορα πάνω από κάποιο σημείο ανοίγει ένα πλαίσιο με βασικές πληροφορίες αναφορικά με το λειτουργικο όρο.

2.6.4 Ραβδόγραμμα (Barchart)

Αντίστοιχα με το διάγραμμα διασποράς, η απεικόνιση αποτελεσμάτων με ραβδόγραμμα δίνεται τόσο στην περίπτωση ανάλυσης λειτουργικού εμπλουτισμού της γονιδιακής έκφρασης (g:Profiler και aGOtool), όσο και στην ανάλυση εμπλουτισμού βιβλιογραφίας (aGOtool). Το διάγραμμα κατασκευάστηκε μέσω της βιβλιοθήκης Plotly. Μια λειτουργία που προστίθεται εδώ (Εικόνα 2.16) είναι ότι ο χρήστης μπορεί να επιλέξει τι θα αντιστοιχεί στο x άξονα του διαγράμματος μέσω της επιλογής Select barplot, δηλαδή Enrichment Score ή -log10(p-value). Ανάλογα με τη μεταβλητή που επιλέγει τα αποτελέσματα κατατάσσονται σε φθίνουσα σειρά. Ο y άξονας αντιστοιχεί στις κωδικές ονομασίες των όρων (Term ID). Αντίθετα, με το διάγραμμα διασποράς, στο οποίο μπορεί να επιλέξει μια βάση για την απεικόνιση των αποτελεσμάτων, εδώ μπορεί να επιλέξει πολλαπλές και οι ράβδοι θα είναι χρωματισμένοι ανάλογα με τις επιλεγμένες βάσεις δεδομένων. Το υπόμνημα του χρωματικού κώδικα εμφανίζεται στο κάτω μέρος του διαγράμματος. Οι επιλογές επιθυμητού αριθμού των αποτελεσμάτων, αυτόματης εμφάνισης

πίνακα, εμφάνισης πληροφοριών με την τοποθέτηση του κέρσορα αλλά και όλες οι ιδιότητες του διαγράμματος είναι ίδιες με αυτές στην περίπτωση του διαγράμματος διασποράς που αναφέρονται στην Ενότητα 2.6.3. Τέλος, στην άκρη κάθε ράβδου εμφανίζεται ένας αριθμός που αντιστοιχεί στον αριθμό των γονιδίων/πρωτεϊνών που περιέχονται στον όρο που αντιστοιχεί στη ράβδο.



Εικόνα 2.16: Ραβδόγραμμα απεικόνισης των αποτελεσμάτων της λειτουργικής και βιβλιογραφικής ανάλυσης εμπλουτισμου. Πολλαπλο ραβδόγραμμα που απεικονίζει τα 17 πρώτα αποτελέσματα των βάσεων UniProt, INTERPRO, Disease Ontology και PFAM, τα οποία έχουν καταταχθεί συνολικά σύμφωνα σε φθίνουσα σειρά βάσει του Enrichment Score. Με την τοποθέτηση του κέρσορα πάνω από κάποια ράβδο προβάλλει ένα πλαίσιο με βασικές πληροφορίες αναφορικά με τον όρο.

2.6.5 Θερμικοί χάρτες (Heatmaps)

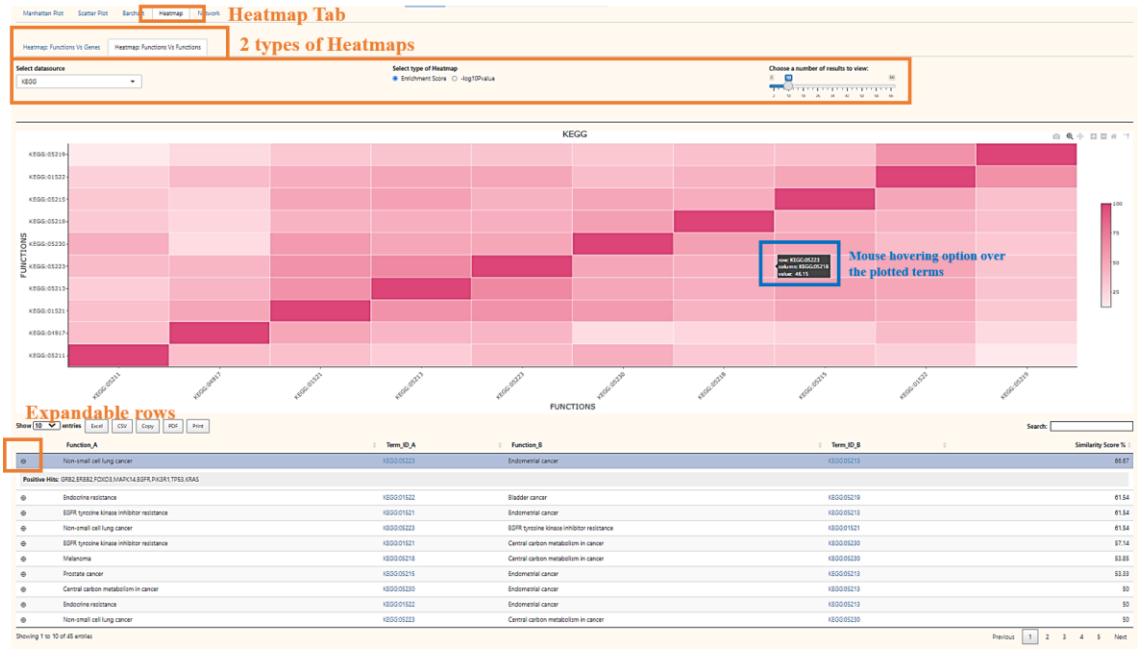
Παράλληλα, για την απεικόνιση των αποτελεσμάτων που προέκυψαν από την ανάλυση λειτουργικού εμπλουτισμού της γονιδιακής έκφρασης (g:Profiler και aGOtool) και την ανάλυση εμπλουτισμού βιβλιογραφίας (aGOtool) μέσω του FLAME έχουν δημιουργηθεί διαγράμματα θερμικών χαρτών (heatmaps). Για τη δημιουργία τους χρησιμοποιήθηκε η βιβλιοθήκη heatmaply.

Ο θερμικός χάρτης (Εικόνα 2.17) που προκύπτει συνοδεύεται επίσης απο επιλογές παραμετροποίησης. Ειδικότερα, ο χρήστης μπορεί να επιλέξει βάσει ποιού μεγέθους θα έχουν ταξινομηθεί τα αποτελέσματά της ανάλυσης μέσω της επιλογής Select type of heatmap, δηλαδή ή Enrichment Score ή -log10(p-value). Ανάλογα με τη μεταβλητή που επιλέγει τα αποτελέσματα κατατάσσονται σε φθίνουσα σειρά και με βάση αυτή την κατάταξη δημιουργούνται τα εκάστοτε heatmaps. Το εύρος τιμών του μεγέθους που επιλέγεται κάθε φορά έχει αντιστοιχιστεί στη χρωματική κωδικοποίηση του χάρτη. Πιο συγκεκριμένα, απουσία οποιασδήποτε σχέσης συμβολίζεται με λευκό χρώμα, ενώ η ένταση του χρώματος σε κάθε πλαίσιο-κελί (cell) αντιστοιχεί στην τιμή του μεγέθους που έχει επιλεγεί (Enrichment Score ή -log10(p-value)), δηλαδή το κελί με τον πιο έντονο χρωματικό τόνο αντιπροσωπεύει την μέγιστη τιμή του μεγέθους. Το χρωματικό υπόμνημα βρίσκεται δεξιά του διαγράμματος. Επίσης, ο χρήστης μπορεί να επιλέξει για ποιά από τις βάσεις δεδομένων επιθυμεί να δημιουργήσει το διάγραμμα και ανάλογα με τη βάση μεταβάλλεται το χρώμα του διαγράμματος, καθώς και να προσαρμόσει τον αριθμό των

αποτελεσμάτων στον αριθμό των όρων που επιθυμεί να απεικονισθεί πάνω στο διάγραμμα μέσω του φίλτρου (Choose a number of results to view).

Παρέχονται δύο τύποι heatmap. Ο πρώτος τύπος πρόκειται για ένα διάγραμμα γονιδίων-όρων εμπλουτισμού (functional ή publication terms), στο οποίο απεικονίζονται ποιά γονίδια περιέχονται σε ποιους όρους. Σε αυτή την περίπτωση υπάρχει μια επιπλέον επιλογή, αυτής της αντιστροφής των αξόνων (Reverse Axis: Functions/Publications-Genes ή Genes-Functions/Publications) για πιο εύληπτη παρουσίαση του χάρτη σε περίπτωση μεγάλου Ν γονιδίων.) Ο δεύτερος τύπος πρόκειται για ένα διάγραμμα λειτουργικών όρων-λειτουργικών όρων ή δημοσίευσης-δημοσίευσης στην περίπτωση εμπλουτισμού βιβλιογραφίας. Για το εν λόγω διάγραμμα υπολογίστηκε ένα νέο μέγεθος, το σκορ ομοιότητας (Similarity Score), το οποίο είναι ο λόγος του αριθμού των κοινών γονιδίων που περιλαμβανονται σε 2 κατηγορίες, προς το αθροιστικό σύνολο των μοναδικών γονιδίων που περιλαμβάνονται σε αυτές τις 2 κατηγορίες (μείον των Νγονιδίων που επαναλαμβάνονται). Έτσι σε αυτόν τον τύπο για κάθε συνδυασμό όρων, ο χρήστης μπορεί να δει ποιά είναι τα κοινά γονίδια και το σκορ ομοιότητας. Οι επιλογές παραμετροποίησης δίνονται και στους δύο τύπους θερμικού χάρτη.

Σε κάθε περίπτωση, το κάθε διάγραμμα χάρτη που προκύπτει είναι διαδραστικό, δηλαδή ο χρήστης μπορεί να το αποθηκεύσει σε μορφή εικόνας (.png), να το μεγεθύνει, να απομονώσει μια επιθυμητή περιοχή και να το μετακινήσει. Επίσης, με την τοποθέτηση του κέρσορα πάνω από κάποιο κελί (cell) ανοίγει ένα πλαίσιο με βασικές πληροφορίες. Ταυτόχρονα, όπως σε όλες τις περιπτώσεις, εμφανίζεται αυτόματα σε κάθε αλλαγή παραμέτρων ο αντίστοιχος πίνακας με όλες τις πληροφορίες αναφορικά με τον εμπλουτισμό, ο οποίος μπορεί να αποθηκευτεί σε διάφορες μορφές για περαιτέρω χρήση.



Εικόνα 2.17: Θερμικός χάρτης απεικόνισης των αποτελεσμάτων της λειτουργικής και βιβλιογραφικής ανάλυσης εμπλουτισμου. Το διάγραμμα που προκύπτει είναι διαδραστικό, δηλαδή ο χρήστης μπορεί να το αποθηκεύσει σε μορφή εικόνας (.png), να το μεγεθύνει, να απομονώσει μια επιθυμητή

περιοχή και να το μετακινήσει. Επίσης, με την τοποθέτηση του κέρσορα πάνω από κάποιο κελί (cell) ανοίγει ένα πλαίσιο με βασικές πληροφορίες. Ταυτόχρονα, εμφανίζεται αυτόματα σε κάθε αλλαγή παραμέτρων ο αντίστοιχος πίνακας με όλες τις πληροφορίες αναφορικά με τον εμπλουτισμό, ο οποίος μπορεί να αποθηκευτεί σε διάφορες μορφές για περαιτέρω χρήση.

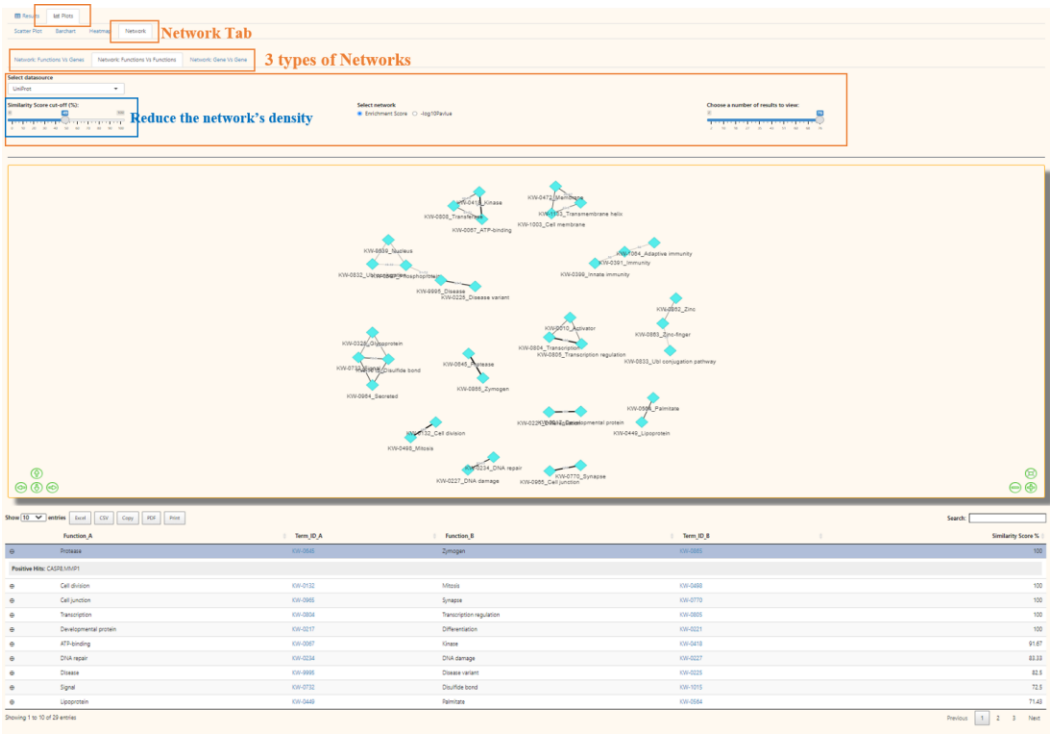
2.6.6 Δίκτυα-Γράφοι (Networks)

Συμπληρωματικά με τους θερμικούς χάρτες, λειτουργούν τα διαγράμματα τύπου δικτύου (network). Τα δίκτυα κατασκευάστηκαν μέσω της βιβλιοθήκης visNetwork και αποθηκεύονται ώς igraph objects (46). Αντίστοιχα, με τους δύο τύπους θερμικών χαρτών, που προαναφέρθηκαν στην Ενότητα 2.6.5, στην περίπτωση των δικτύων προσφέρονται 3 επιλογές. Ειδικότερα, στην πρώτη περίπτωση οι κόμβοι (nodes) αντιπροσωπεύουν γονίδια/πρωτεΐνες με σχήμα κύκλου και λειτουργικούς όρους/δημοσιεύσεις (terms) με ρόμβους, απεικονίζοντας έτσι ποιά γονίδια εμπλέκονται σε ποιες λειτουργίες/δημοσιεύσεις, βλέποντας έτσι ταυτόχρονα όλες τις πιθανές συνδέσεις. Τα χρώματα γονιδίων όρων διαφέρουν. Η δεύτερη επιλογή αφορά ένα δίκτυο, όπου οι κόμβοι αντιστοιχούν σε λειτουργικούς όρους/δημοσιεύσεις και οι ακμές που συνδέουν τους κόμβους αντιπροσωπεύουν το σκορ ομοιότητας (Ενότητα 2.6.5) που σχετίζεται με τον κοινό αριθμό γονιδίων που συμμετέχουν στους δύο όρους που είναι συνδεδεμένοι (σταθμισμένο δίκτυο). Τα κοινά γονίδια αναγράφονται στον πίνακα που συνοδεύει το δίκτυο. Αυτές οι δύο επιλογές παρέχονται και για τους θερμικούς χάρτες. Προστέθηκε στην περίπτωση των δικτύων μια επιπλέον επιλογή, όπου οι κόμβοι αντιπροσωπεύουν γονίδια και οι ακμές που συνδέουν τους κόμβους τον αριθμό των λειτουργικών όρων/δημοσιεύσεων που περιέχονται αυτά τα γονίδια.

Ο χρήστης μπορεί να επιλέξει για ποιά από τις βάσεις δεδομένων επιθυμεί να δημιουργήσει το διάγραμμα και ανάλογα με τη βάση μεταβάλλεται το χρώμα του διαγράμματος, καθώς και να προσαρμόσει τον αριθμό των αποτελεσμάτων στον αριθμό των όρων που επιθυμεί να απεικονισθεί πάνω στο διάγραμμα μέσω του φίλτρου (Choose a number of results to view).
Επίσης, μπορεί να επιλέξει βάσει ποιού μεγέθους θα έχουν ταξινομηθεί τα αποτελέσματά της ανάλυσης μέσω της επιλογής Select network, δηλαδή ή Enrichment Score ή -log10(p-value). Ανάλογα με τη μεταβλητή που επιλέγει τα αποτελέσματα κατατάσσονται σε φθίνουσα σειρά και με βάση αυτή την κατάταξη δημιουργούνται τα εκάστοτε δίκτυα. Στην περίπτωση του δικτύου ομοιότητας των όρων, παρέχεται η δυνατότητα να επιλέξουν την κατώτατη τιμή (κατώφλι) του σκορ ομοιότητας μέσω της επιλογής Similarity Score cut-off (%), έτσι ώστε να εμφανίζονται μόνο οι κόμβοι οι οποίοι συνδέονται με σκορ μεγαλύτερο από την κατωφλική τιμή. Αντίστοιχα, στον τρίτο τύπο δικτύου (συνδέσεις γονιδίων) ο χρήστης μπορεί να ορίσει το ελάχιστο όριο των κοινών όρων που εμφανίζονται δύο γονίδια μέσω της επιλογής Number of common Functions. Με αυτές τις δύο επιπλέον επιλογές απομακρύνονται από το διάγραμμα συνδέσεις με μικρό σκορ ομοιότητας ή μικρό αριθμό κοινών όρων αντίστοιχα, κάνοντας το δίκτυο λιγότερο πυκνό και άρα πιο κατανοητό και πληροφοριακό.

Τα δίκτυα είναι πλήρως διαδραστικά και ο χρήστης μπορεί να κάνει μεγέθυνση/σμίκρυνση, να προσαρμόσει τη διάταξη κόμβων-ακμών και να εξάγει το δίκτυο σε μορφή εικόνας. Τέλος σε κάθε περίπτωση δικτύου εμφανίζεται ο αντίστοιχος πίνακας που ανανεώνεται αυτόματα σε

οποιαδήποτε ρύθμιση των παραμέτρων και έχει επιλογές αποθήκευσης σε διάφορες μορφές για περαιτέρω ανάλυση.



Εικόνα 2.18: Δίκτυο απεικόνισης των αποτελεσμάτων της λειτουργικής και βιβλιογραφικής ανάλυσης εμπλουτισμου. Το δίκτυο είναι πλήρως διαδραστικό και ο χρήστης μπορεί να κάνει μεγέθυνση/σμίκρυνση, να προσαρμόσει τη διάταξη κόμβων-ακμών και να εξάγει το δίκτυο σε μορφή εικόνας. Τέλος σε κάθε περίπτωση δικτύου εμφανίζεται ο αντίστοιχος πίνακας που ανανεώνεται αυτόματα σε οποιαδήποτε ρύθμιση των παραμέτρων και έχει επιλογές αποθήκευσης σε διάφορες μορφές για περαιτέρω ανάλυση.

2.7 Μετατροπή αναγνωριστικών IDs γονιδίων και Ορθόλογη αναζήτηση

Οι βάσεις δεδομένων και τα βιοπληροφορικά εργαλεία διαφόρων αναλύσεων είναι πιθανόν να δέχονται ως τύπο είσοδου τους διαφορετικά αναγνωριστικά (identifiers, ID) γονιδίου/πρωτεΐνης. Αυτή η αναγκαία χρήση κάποιου συγκεκριμένου τύπου ID δημιουργεί έναν φραγμό για τους μη ειδικούς χρήστες. Μέσω του FLAME το πρόβλημα αυτό μπορεί να επιλυθεί, αφού η εφαρμογή χρησιμοποιεί μετατροπείς (βιβλιοθήκη R/gProfiler2), οι οποίοι επιτρέπουν:
i) μετατροπές ID μεταξύ βάσεων δεδομένων, όπως Entrez Gene, Uniprot, ChEMBL, ENSEMBL, RefSeq, και
ii) ορθόλογη μετατροπή γονιδίων από τον οργανισμό προέλευσης στον οργανισμό στόχο μεταξύ των 197 διαφορετικών οργανισμών που υποστηρίζονται από το FLAME.

2.7.1 Μετατροπή αναγνωριστικών γονιδίων (g:Convert)

Το g:Convert [58] είναι ένα εργαλείο για την αυτόματη μετατροπή των γονιδιακών IDs, δηλαδή επιτρέπει τη μετατροπή IDs των γονιδίων, πρωτεϊνών, ανιχνευτών, μικροσυστοιχιών,

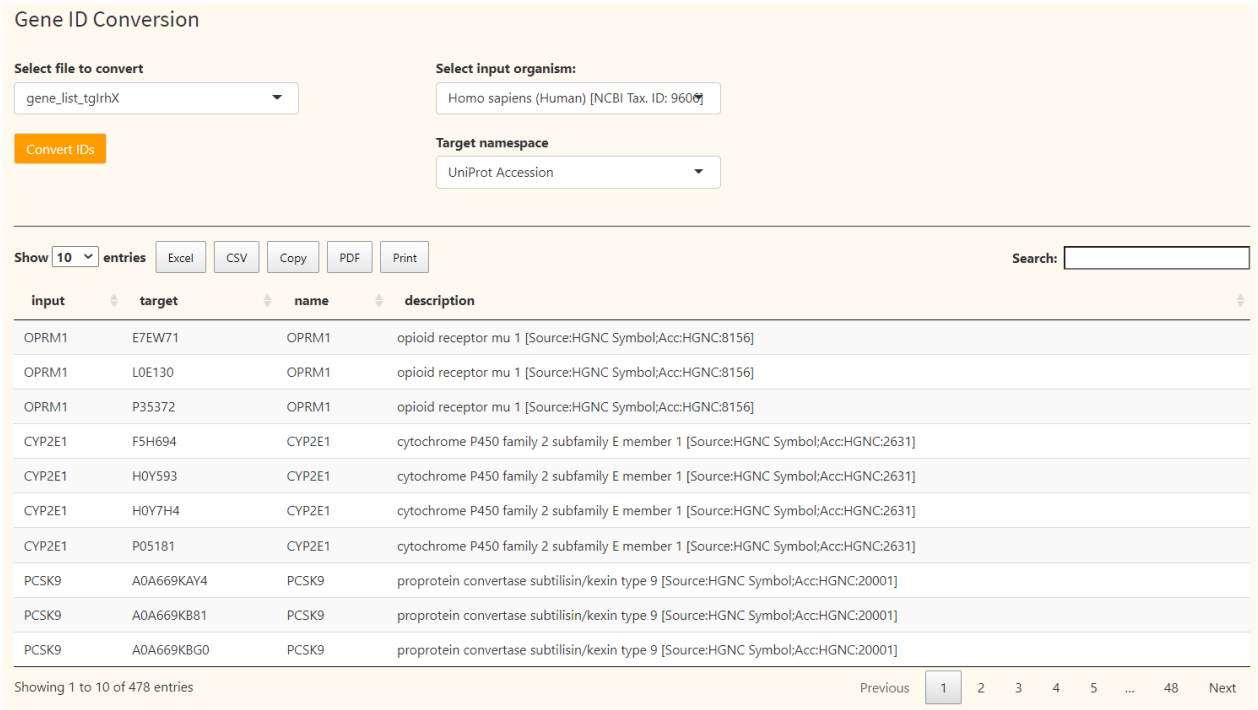
κοινών ονομάτων από μία βάση δεδομένων σε κάποια άλλη. Όλα τα IDs λαμβάνονται μέσω αντιστοίχισης τους των αναγνωριστικών της Ensembl (ENSG).

Μέσω του FLAME, το g:Convert υποστηρίζει σημαντικές βάσεις δεδομένων και τα IDs που παρέχονται για μετατροπή είναι τα εξείς: ChEMBL, Entrez Gene Name, Entrez Gene Accession, Entrez Gene Transcript Name, UniProt Accession, UniProt Gene Name, EMBL Accession, ENSEMBL Protein ID, ENSEMBL Gene ID, ENSEMBL Transcript ID, UniProt Archive, WIKIGENE ID, RefSeq mRNA, RefSeq mRNA Accession, RefSeq Protein Accession, RefSeq Non-coding RNA Accession.

Ο χρήστης μεταβαίνοντας από το Menu→Conversion→Gene ID Conversion μπορεί επιλέγοντας τον οργανισμό προέλευσης των δεδομένων (input organism) του υπό μελέτη αρχείου και το επιθυμητό τύπο αναγνωριστικού (target namespace) σύμφωνα με τις βάσεις δεδομένων που προσφέρονται από το FLAME να ανακτήσει αυτόματα τα αποτελέσματα, αν υπάρχει αντιστοιχία στην επιθυμητή βάση. Ως αρχείο εισόδου μπορεί να χρησιμοποιηθεί μια μικτή λίστα με IDs γονιδίων, SNPs, χρωμοσωμικές περιοχές, term IDs, μικροσυστοιχίες κλπ. Στην περίπτωση λίστας που περιέχει term (όρο) IDs το g:Convert ανακτά όλα τα γονίδια του δεδομένου οργανισμού που σχετίζονται με το δεδομένο term. Για παράδειγμα, η χαρτογράφηση του term GO:0007507 (ανάπτυξη καρδιάς) στον οργανισμό *Homo sapiens* έχει ως αποτέλεσμα το g: Convert να ανακτά περίπου εκατό ανθρώπινα γονίδια που σχετίζονται με την ανάπτυξη της καρδιάς και να χρησιμοποιεί αυτά ως δεδομένα εισόδου πλέον. Στην περίπτωση χρωμοσωμικών περιοχών, όλα τα γονίδια από αυτές τις περιοχές ανακτώνται αυτόματα. Η μορφή για την χρωμοσωμική περιοχή που αναγνωρίζει το εργαλείο είναι "χρωμόσωμα: έναρξη: τέλος" (πχ X: 1: 2000000).

Το αποτέλεσμα είναι ένας διαδραστικός πίνακας με επιλογές αναζήτησης και αποθήκευσης σε διάφορες μορφές για περαιτέρω αναλύσεις. Ο πίνακας των αποτελεσμάτων περιέχει τα ονόματα των γονιδίων της λίστας εισόδου (name), το όνομα του γονιδίου (name), τα IDs μετά τη μετατροπή (target), καθώς και μια σύντομη περιγραφή (description) του ρόλου κάθε γονιδίου (Εικόνα 2.19). Τα γονίδια για τα οποία δεν βρέθηκαν αντιστοιχίες στις επιθυμητή βάση δεδομένων έχουν αφαιρεθεί από τον πίνακα. Τέλος, υπάρχει περίπτωση ένα γονίδιο να αντιστοιχεί σε περισσότερα από ένα IDs, άρα ο συνολικός αριθμός αποτελεσμάτων να είναι μεγαλύτερος από το αριθμητικό περιεχόμενο του αρχείου εισόδου.

Σε αυτό το σημείο αξίζει να σημειωθεί ότι το g:Convert έχει ενσωματωθεί στο υπόβαθρο λειτουργίας των αναλύσεων εμπλουτισμού με αποτέλεσμα ο χρήστης να μην χρειάζεται να επιλέξει το ID των στοιχείων της λίστας εισόδου πριν την ανάλυση, προσπερνώντας έτσι ένα σημαντικό πρόβλημα που υπάρχει σε άλλα εργαλεία, όπου είναι προϋπόθεση η συμπλήρωση του αρχικού αναγνωριστικού της λίστας εισόδου. Ταυτόχρονα, όπως προαναφέρθηκε στην Ενότητα 2.3 δίνεται η δυνατότητα στον χρήστη να επιλέξει και τύπο εξόδου των αποτελεσμάτων του ασχέτως τους τύπου εξόδου.



Εικόνα 2.19: Παράδειγμα μορφής παραμέτρων και αποτελεσμάτων της μετατροπής των γονιδιακών IDs. Στο εν λόγω παράδειγμα έγινε μετατροπή της αρχικής λίστας με τα ανθρώπινα γονίδια (*H. sapiens*) (select input organism) με επιθυμητό ID εξόδου (target namespace) την UniProt Accession και προέκυψαν 478 αποτελέσματα, ενώ το αρχικό αρχείο εισόδου είχε 100 γονίδια. Το αποτέλεσμα είναι ένας διαδραστικός πίνακας με επιλογές αναζήτησης και αποθήκευσης σε διάφορες μορφές για περαιτέρω αναλύσεις. Ο πίνακας των αποτελεσμάτων περιέχει τα ονόματα των γονιδίων της λίστας εισόδου (name), το όνομα του γονιδίου (name), τα IDs μετά τη μετατροπή (target), καθώς και μια σύντομη περιγραφή (description) του ρόλου κάθε γονιδίου

2.7.2 Ορθόλογη μετατροπή γονιδίων (g:Orth)

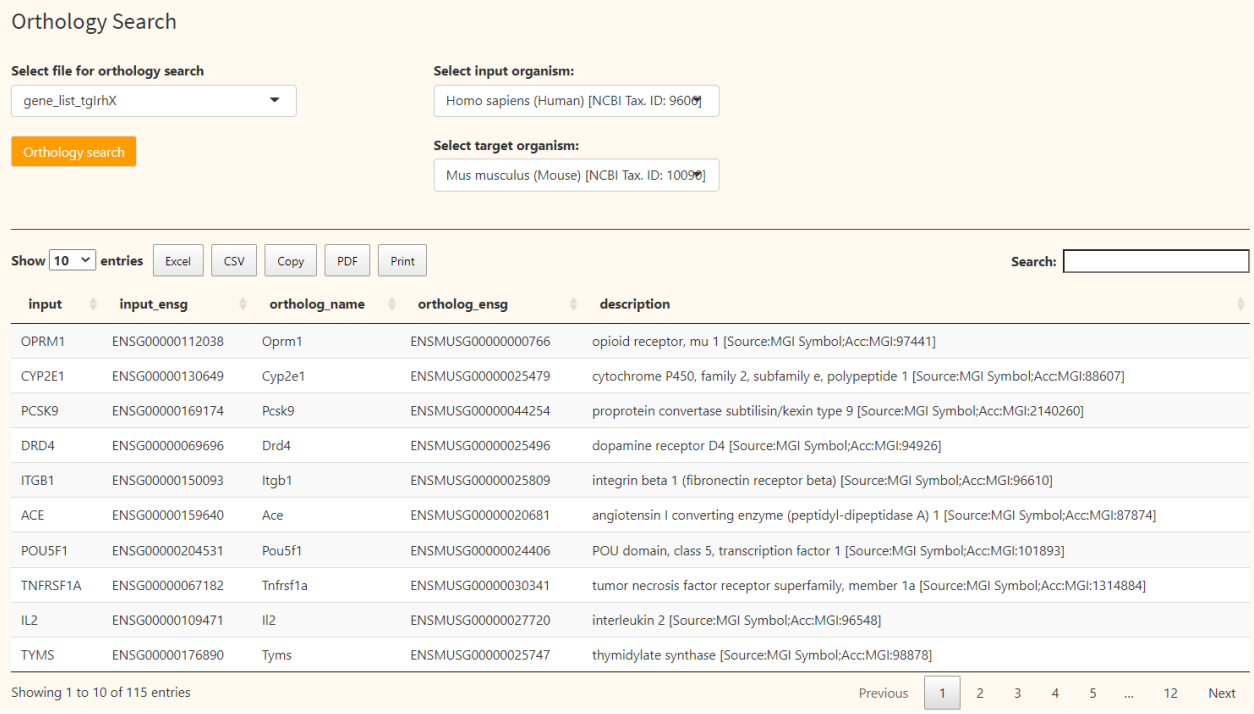
Το g:Orth [58] είναι ένα εργαλείο που χρησιμοποιείται για τη χαρτογράφηση (mapping) ορθόλογων γονιδίων μεταξύ συγγενικών οργανισμών με βάση τα δεδομένα που συλλέγονται στη βάση δεδομένων Ensembl. Τα ορθόλογα γονίδια αποτελούνται από παρόμοιες συντηρημένες αλληλουχίες (ομόλογες αλληλουχίες), καθώς έχουν προκύψει από έναν κοινό πρόγονο απο γεγονότα ειδογένεσης. Τα ορθόλογα γονίδια είναι πιθανόν να έχουν παρόμοιους λειτουργικούς ρόλους και ως εκ τούτου είναι σημαντικά στη λειτουργική ανάλυση.

Ενσωματώνοντας αυτό το εργαλείο στο Flame, ο χρήστης μπορεί να το χρησιμοποιήσει για τη μεταφορά της γνώσης που συλλέγεται για καλά μελετημένους οργανισμούς μοντέλα σε λιγότερο μελετημένα είδη. Η ανάλυση εμπλουτισμού μετά από ορθόλογη χαρτογράφηση μπορεί να οδηγήσει σε νέα αποτελέσματα και καλύτερα ευρήματα σε σχέση με αυτά από τα αρχικά είδη.

Ο χρήστης μεταβαίνοντας από το Menu→Conversion→Orthology Search μπορεί επιλέγοντας τον οργανισμό προέλευσης των δεδομένων (input organism) του υπό μελέτη αρχείου και τον οργανισμό στόχο (target organism) να ανακτήσει αυτόματα τα ορθόλογα γονίδια, αν υπάρχει αντιστοιχία μεταξύ των οργανισμών. Ως αρχείο εισόδου μπορεί να χρησιμοποιηθεί μια μικτή λίστα με IDs γονιδίων, SNPs, χρωμοσωμικές περιοχές, term IDs, μικροσυστοιχίες κλπ. Στην

περίπτωση λίστας που περιέχει term (όρο) IDs το g:Orth ανακτά όλα τα γονίδια του δεδομένου οργανισμού που σχετίζονται με το δεδομένο term. Για παράδειγμα, η χαρτογράφηση του term GO:0007507 (ανάπτυξη καρδιάς) στον οργανισμό *Homo sapiens* έχει ως αποτέλεσμα το g: Orth ανακτά περίπου εκατό ανθρώπινα γονίδια που σχετίζονται με την ανάπτυξη της καρδιάς και να χρησιμοποιεί αυτά ως δεδομένα εισόδου πλέον. Στην περίπτωση χρωμοσωμικών περιοχών, όλα τα γονίδια από αυτές τις περιοχές ανακτώνται αυτόματα. Η μορφή για την χρωμοσωμική περιοχή που αναγνωρίζει το εργαλείο είναι "χρωμόσωμα: έναρξη: τέλος" (πχ X: 1: 2000000).

Το αποτέλεσμα είναι ένας διαδραστικός πίνακας με επιλογές αναζήτησης και αποθήκευσης σε διάφορες μορφές για περαιτέρω αναλύσεις. Εκτός από την ονομαστική μετατροπή των ονομάτων αλλά και των IDs των γονιδίων δίνεται και μια σύντομη περιγραφή του ρόλου κάθε γονιδίου (Εικόνα 2.20). Τα γονίδια για τα οποία δεν βρέθηκαν ορθόλογα έχουν αφαιρεθεί από τον πίνακα.



Εικόνα 2.20: Παράδειγμα μορφής παραμέτρων και αποτελεσμάτων της ορθόλογης μετατροπής γονιδίων. Στο εν λόγω παράδειγμα έγινε χαρτογράφηση της αρχικής λίστας με τα ανθρώπινα γονίδια (*H. sapiens)* στα ορθόλογα γονίδια του ποντικού (*Mus musculus*) και προέκυψαν 115 αποτελέσματα.

2.8 Ανάλυση δεδομένων από εξωτερικές βάσεις δεδομένων (Integration with other applications)

Το FLAME μπορεί να κληθεί από άλλες εφαρμογές μέσω ενός απλού αίτηματος λήψης (get request). Τα ονόματα των γονιδίων/πρωτεϊνών πρέπει να συμπεριληφθούν στη διεύθυνση URL (url_genes variable) και να είναι διαχωρισμένα με κόμμα (,), ενώ στην περίπτωση πολλαπλών λιστών, αυτές πρέπει να έχουν διαχωριστεί με το σύμβολο ερωτηματικών (;). Οι λίστες θα

εμφανίζονται ως μεταφορτωμένα αρχεία. Ένα απλό παράδειγμα URL που κωδικοποιεί τρεις λίστες είναι:
flame.pavlopouloslab.info/?url_genes=MCL1,TTR;APOE,ACE2;TLR4,HMOX1,TP73

3. ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1 Case Study

Το FLAME χρησιμοποιήθηκε για την ανάλυση εκ νέου δεδομένων γονιδιακής έκφρασης που σχετίζονται με τον «καρκίνο σε έδαφος κολίτιδας» (CAC) στα ποντίκια [80]. Ο καρκίνος του παχέος εντέρου (Colorectal Cancer, CRC) διαχωρίζεται στον σποραδικό τύπο (sporadic CRC) και στον καρκίνο σε έδαφος κολίτιδας (Colitis Associated Cancer, CAC), ο οποίος έχει άμεση συσχέτιση με τη φλεγμονώδη νόσο του εντέρου (Intestinal Bowel's Disease, IBD) και έχει υψηλό κίνδυνο θνησιμότητας [81].

Στον ποντικό, η ογκογένεση επάγεται μετά από χρήση αζοξυμεθανίου (azoxymethane AOM) σε συνδυασμό με 4 κύκλους χορήγησης θειικού άλατος δεξτρανών (Dextran Sodium Sulfate, DSS) που προκαλεί χρόνια κολίτιδα. Ωστόσο, η φλεγμονή, η καταστροφή των ιστών και η εμφάνιση καρκίνου εκδηλώθηκαν στο περιφερικό τμήμα (distal) και όχι στο εγγύς (proximal) τμήμα του παχέος εντέρου. Η περιφερική μοίρα (αριστερή) είναι το τελευταίο κομμάτι του παχέος εντέρου και αποτελείται από το κατιόν (descending) κόλον και το σιγμοειδές (sigmoid), ενώ η εγγύς μοίρα (δεξιά) είναι το πρώτο τμήμα του παχέος εντέρου και αποτελείται από το τυφλό (cecum), το ανιόν (ascending) και το εγκάρσιο (transverse) κόλον [80].

Για την ενδελεχή μελέτη και κατανόηση αυτού του φαινομένου πραγματοποιήθηκε διαφορική μελέτη της γονιδιακής έκφρασης μεταξύ του περιφερικού και του εγγύς τμήματος του παχέος εντέρου κατα την διάρκει του CAC. Από τις μελέτες που διεξήχθησαν προέκυψε επίσης ότι τα μεταβολικά μονοπάτια των λιπιδίων στο εγγύς τμήμα συμβάλουν στην αντίσταση της πειραματικά επαγόμενης CAC και κολίτιδας [80].

Το FLAME, χρησιμοποιήθηκε για την εύρεση μονοπατιών και γονιδιακών οντολογιών που μπορεί να σχετίζονται με την ανάπτυξη CAC. Τα γονίδια που βρέθηκαν υπερεκφρασμένα τόσο στα αρχικά στάδια, δηλαδή στους 2 κύκλους χορήγησης DSS όσο και στα τελευταία στάδια (4 DSS κύκλους) της AOM/DSS επαγόμενης καρκινογένεσης στο περιφερικό τμήμα του παχέος εντέρου και όχι στην εγγύς μοίρα θεωρήθηκαν ότι διαδραματίζουν σημαντικό ρόλο στον CAC. Σημειώθηκαν τα 165 γονίδια από το σύνολο τους που υπερεκφράστηκαν μόνο στο περιφερικό τμήμα στους 2 DSS κύκλους και 4 DSS κύκλους με την χρήση του "UpSet Plot" (Εικόνα 3.1.A) του FLAME και το σύνολο αυτό των γονιδίων χαρακτηρίστηκε ως 'the susceptibility-associated gene signature', SAS (Πίνακας 3.1).

Πίνακας 3.1: Τα SAS γονίδια που βρέθηκαν με την χρήση UpSet Plot.

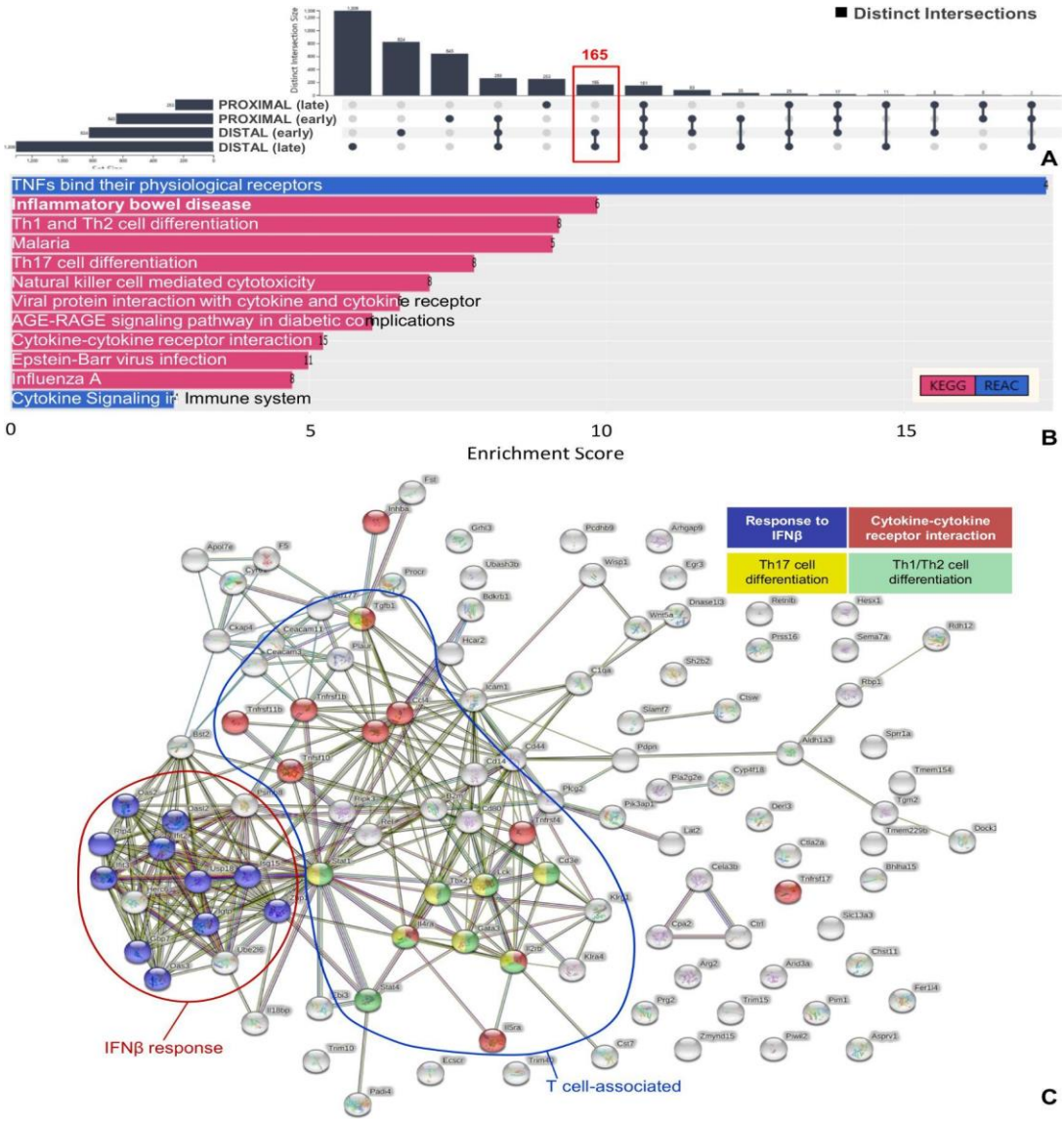
165 Susceptibility-associated gene signature (SAS)
ALDH1A3, Apol7e, Apol9a/Apol9b, ARG2, ARHGAP9, ARID3A, ASPRV1, B2M, BC005685, BDKRB1, BHLHA15, BST2, C1orf187, C1QA, C20orf123, C21orf91, CAPG, CCDC109B, CCL2, CCL20, CCL4, CD14, CD177, CD3E, CD44, CD80, Ceacam11, CEACAM3, CELA3B, Ces2f, CHST11, CKAP4, CPA2, Crem, CST7, Ctla2a, CTRL, CTSW, CYP4F3, CYR61, DDX60, DERL3, DLK1, DNASE1L3, DOCK11, DUOXA1, EBI3, ECSCR, EGR3, Expi, F5, Fer1l4, FST, GATA3, GBP7, GGTA1, GPR110, GRHL3, H2-Q8, HCAR2, HERC6, HESX1, ICAM1, Ifi47, IFIT1B, IFIT2, IFIT3, Igtp, IL18BP, IL2RB, IL4R, IL5RA, INHBA, IRF9, IRGM, Irgm2, ISG15, Klra4, Klrb1c, KLRC1, KLRG1, LAT2, LCK, LGALS9B, LRP8, Ly6a, Lyz1/Lyz2, MS4A7, MT2A, MUC1, NLRP3, OAS2, OAS3, Oasl2, PADI4, Pcdhb9, PDPN, PIK3AP1, PIM1, PIWIL2, PLA2G2E, PLAUR, PLCG2, PPM1M, PRDM1, PREX1, PRG2, PROCR, PRSS16, PSMB8, PSMB9,

RBP1, RDH12, REG1A, REG1B, REL, Retnlb, RIPK3, RNASE3, RNF183, RTP4, SEMA7A, SERPINA1, Serpinb6b, SH2B2, SH3BP2, SH3KBP1, SLAMF7, SLC13A3, SLC14A1, SLC16A6, SLFN12, SMOX, SPP1, Sprr1a, SPTSSB, SRPX2, STAT1, STAT4, TAP1, TBX21, TC2N, TGFB1, TGM1, TGM2, Tlr12, TMEM154, TMEM229B, TNFRSF11B, TNFRSF17, TNFRSF18, TNFRSF1B, TNFRSF4, TNFSF10, TRIM10, TRIM15, TRIM40, UBASH3B, UBE2L6, USP18, WISP1, WNT5A, XAF1, ZBP1, ZMYND15

Στην συνέχεια, πραγματοποιήθηκε ανάλυση λειτουργικού εμπλουτισμού στα γονίδια SAS με σκοπό την συσχέτιση τους σε βιολογικές διεργασίες και μονοπάτια με τη χρήση του Functional Enrichment Analysis: gProfiler Tool που παρέχει το FLAME και ειδικότερα η ανάλυση πραγματοποιήθηκε με βάση τις Kegg και REACTOME. Βρέθηκε, όπως ήταν αναμενόμενο, η "Φλεγμονώδης νόσος του εντέρου" να είναι μεταξύ των πιο στατιστικά σημαντικά εμπλουτισμένων μονοπατιών (Εικόνα 3.1.Β και Παράρτημα Πίνακας 2). Επιπλέον μονοπάτια που βρέθηκαν να είναι στατιστικά σημαντικά σχετίζονται με ανοσολογικές διεργασίες που περιλαμβάνουν τη διαφοροποίηση των Τ βοηθητικών κυττάρων (Th1)/Th2, τη διαφοροποίηση των Th17 κυττάρων, τα NK-κύτταρα (Natural Killer (NK) cell-mediated cytotoxicity) καθώς και μονοπάτια σηματοδότησης μέσω κυτοκινών/χημειοκινών. Τα βιολογικά αυτά μονοπάτια και διεργασίες συνδέονται άρρηκτα με καταστάσεις φλεγμονής, όπου αδυναμία αντιμετώπισης της μπορεί να οδηγήσει σε εξέλιξη καρκίνου του εντέρου [82]. Πράγματι έχει βρεθεί ότι τα λεμφοκύτταρα Th17 επάγουν παθογόνα Th1 κύτταρα που εμπλέκονται στην κολίτιδα [83] και όταν ενεργοποιούνται από τον παράγοντα TGFβ1 (SAS γονίδιο/κυτοκίνη), τα κύτταρα Th17 προάγουν την ανάπτυξη CAC [84]. Επιπλέον, η πειραματικα χρόνια κολίτιδα έχει συσχετιστεί με το διττό προφίλ κυτοκινών Th1/Th2 [85] και υπάρχουν στοιχεία που υποδηλώνουν ότι η φλεγμονή του παχέος εντέρου επαγόμενη από Th2 ενισχύει το σχηματισμό όγκων του παχέος εντέρου [86]. Παρατηρήθηκε επίσης στατιστικά σημαντικός εμπλουτισμός των γονιδίων SAS στο προ-φλεγμονώδες μονοπάτι σηματοδότησης μέσω TNF (Tumor necrosis factor), το οποίο οδηγεί σε αύξηση της ανοσολογικής απόκρισης σε καταστάσεις φλεγμονωδών παθήσεων του εντέρου.

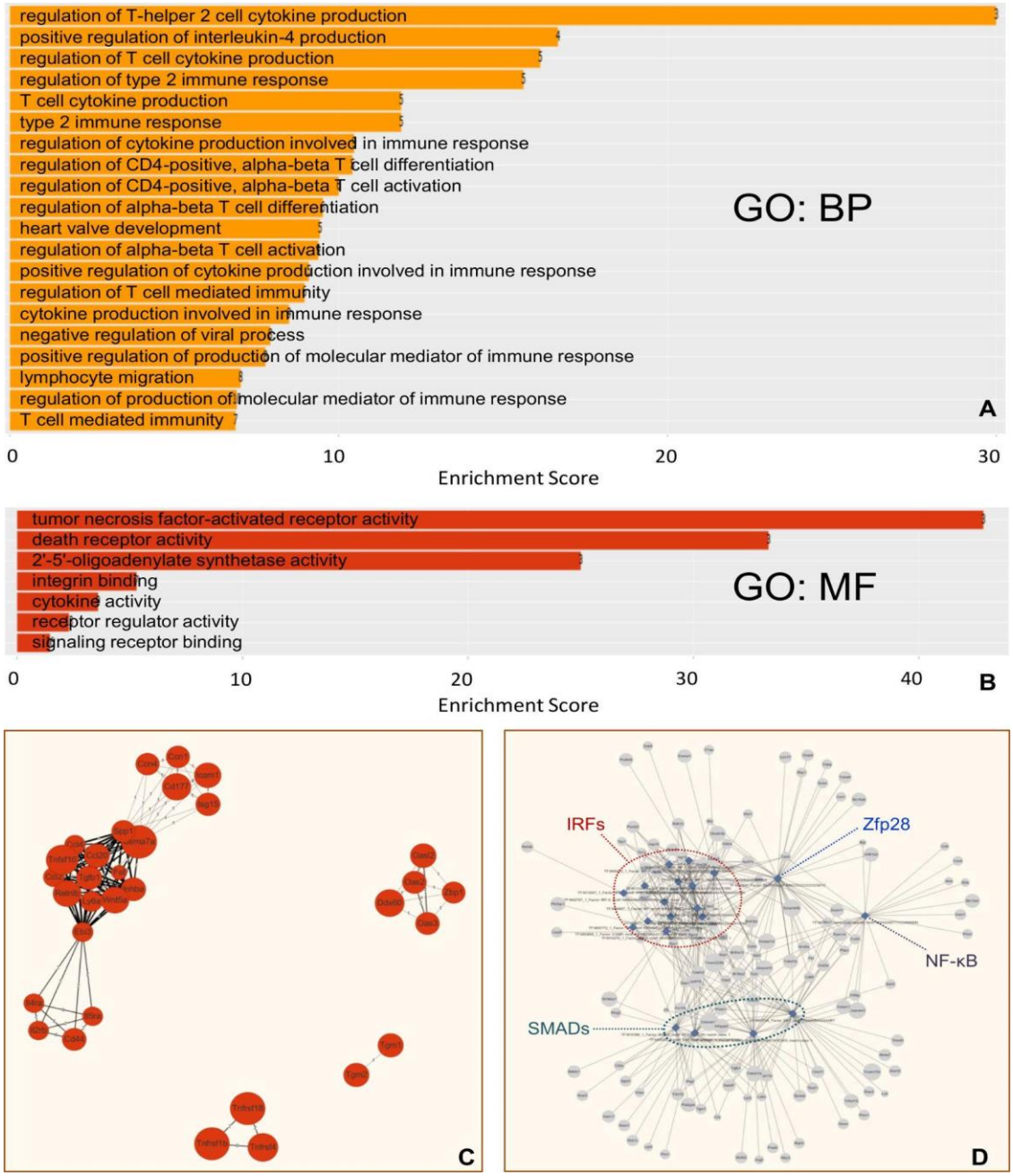
Πραγματοποιήθηκε ταυτόχρονα ανάλυση λειτουργικού εμπλουτισμού σε γονιδιακές οντολογίες (Gene ontology-GO) μέσω της εφαρμογής FLAME. Γονιδία SAS βρέθηκαν ομαδοποιημένα σε βιολογικές διεργασίες που σχετίζονται κυρίως με Τ κύτταρο-επαγόμενη ανοσία και σύνθεση κυτοκίνων (Εικόνα 3.2.Α και Παράρτημα Πίνακας 3). Αναφορικά με τις μοριακές λειτουργίες προέκυψαν διάφορες κατηγορίες (terms) που σχετίζονται με *δραστηριότητες κυτοκινών και χημειοκινών,* -συμπεριλαμβανομένου του παράγοντα TNF- καθώς και με *σηματοδότηση μέσω υποδοχέων θανάτου* (death receptor signaling), οι οποίοι έχουν βρεθεί ότι εμπλέκονται στον CAC [86]. Με τη χρήση των εργαλείων οπτικοποίησης που παρέχονται από το FLAME, κατασκευάστηκε ένα δίκτυο γονιδίου-γονιδίου (gene-gene network) με σκοπό την απεικονιστική εύρεση ομάδων (clusters) αναφορικά με τις κοινές μοριακές λειτουργίες (Εικόνα 3.2.C). Τα αποτελέσματα που προέκυψαν από την χρήση της STRING για τη μελέτη πρωτεϊνικών αλληλεπιδράσεων (φυσικών και λειτουργικών) μέσω του FLAME κατέδειξαν ένα σύνολο αλληλεπιδρώντων συστατικών της ανοσίας που σχετίζεται με τα Τ κύτταρα, καθώς και μια ομάδα πρωτεϊνών ως αποτέλεσμα απόκρισης στον παράγοντα IFNβ (Ιντερφερόνη Β), οι οποίες εμπλέκονται στην επιθηλιακή αναγέννηση βλαβών των ιστών που έχουν προκληθεί από την χορήγηση DSS (Εικόνα 3.1.C) [87].

Τέλος, για περαιτέρω εμβάθυνση σε ρυθμιστικό επίπεδο των βιολογικων διεργασιών και των μονοπατιών, χρησιμοποιήθηκε το FLAME σε επίπεδο ανάλυσης εμπλουτισμού μεταγραφικών παραγόντων. Τα αποτελέσματα έδειξαν μία ομάδα ρυθμιστικών παραγόντων της Ιντερφερόνης (Interferon Regulatory Factor, IRF) (Εικόνα 3.2.D και Παράρτημα Πίνακας 4) οι οποίοι πιθανόν να συσχετίζονται τόσο με τον εμπλουτισμό των γονιδίων SAS σε οντολογίες IFNβ, όσο και με την διαφοροποίηση των T βοηθητικών κυττάρων. Επιπλέον, μια επιπλέον ομάδα μεταγραφικών παραγόντων συσχετίστηκε με τις πρωτεΐνες SMADs, οι οποίες είναι υπεύθυνες για τη μεταγωγή σημάτων από τον υποδοχέα TGFβ1 και φαίνονται να εμπλέκονται στην ανάπτυξη CAC [88], [89].



Εικόνα 3.1: (A) Χρήση του "UpSet Plot" για την εύρεση των κοινών γονιδίων (Distinct intersection Mode) που υπερεκφράστηκαν στους 2 κύκλους χορήγησης DSS (Distal early) και στους 4 DSS κύκλους (Distal late) της AOM/DSS επαγόμενης καρκινογένεσης στο περιφερικό τμήμα του παχέος εντέρου και όχι στην εγγύς μοίρα. Βρέθηκαν 165 γονίδια και το σύνολο αυτό των γονιδίων χαρακτηρίστηκε ως 'the susceptibility-associated gene signature', SAS. Τα γονίδια SAS χρησιμοποιήθηκαν περαιτέρω για ανάλυση

λειτουργικού εμπλουτισμού σε οντολογίες και μονοπάτια. (Β) Αποτελέσματα σε μορφή ραβδογράμματος ανάλυσης λειτουργικού εμπλουτισμού των γονιδίων SAS με βάση τις βάσεις KEGG (φούξια) και REACTOME (μπλε). (C) Μελέτη πρωτεϊνικών αλληλεπιδράσεων (φυσικών και λειτουργικών) μέσω της STRING που παρέχεται από το FLAME. Παρατηρήθηκαν ομαδοποιημένα συνολα (clusters) αλληλεπιδρώντων συστατικών: ως αποτέλεσμα απόκρισης στον παράγοντα IFNβ (μπλέ χρώμα), καθώς και Τ κύτταρα που εμπλέκονται στην κυτταρική ανοσία συμπεριλαμβανομένων της διαφοροποίησης Th1/Th2 (πράσινο χρώμα), της διαφοροποίησης Th17 (κίτρινο χρώμα) και της αλληλεπίδρασης υποδοχέα κυτοκίνης-κυτοκίνης (κόκκινο χρώμα).



Εικόνα 3.2: (A) Αποτελέσματα ανάλυσης λειτουργικού εμπλουτισμού σε μορφή ραβδογράμματος των SAS γονιδίων με βάση την Gene Ontology-Biological Processes (GO:BP). (B) Αποτελέσματα ανάλυσης λειτουργικού εμπλουτισμού σε μορφή ραβδογράμματος των SAS γονιδίων με βάση την Gene Ontology-Molecular Functions (GO:MF). (C) Αναπαράσταση των αποτελεσμάτων της GO:MF με την χρήση δικτύου γονιδίων (gene-gene Network), όπου το πάχος το ακμών είναι ανάλογο με τον αριθμό των κοινών μοριακών λειτουργιών στις οποίες εντοπίζονται τα γονίδια που ενώνουν οι ακμές (δηλαδή τα γονίδια των κόμβων). (D) Μη σταθμισμένο (unweighted) δίκτυο (Functions vs Genes Network) μεταγραφικών παραγόντων που πιθανόν συμμετέχουν στην ρύθμιση της έκφρασης των γονιδίων SAS.

4. ΣΥΖΗΤΗΣΗ-ΣΥΜΠΕΡΑΣΜΑΤΑ

Οι σημαντικές αλλαγές που συντελέστηκαν τις τελευταίες δεκαετίες στο πεδίο της βιολογίας, σε συνδυασμό με την εξέλιξη των επονομαζόμενων -omics τεχνολογιών, αλλά και στη μελέτη της βιοποικιλότητας και της διατήρησης της φύσης, οδήγησαν σε ραγδαία αύξηση των πληροφοριών που παράγονται από τη βιολογική κοινότητα. Αυτό οδήγησε στην ανάγκη για αποτελεσματική διαχείριση, έλεγχο και ανάλυση όλων αυτών των δεδομένων με τελικό σκοπό την αξιοποίησή τους για την εξαγωγή σημαντικών Βιολογικών συμπερασμάτων. Απόρροια αυτών αποτελεί η ανάπτυξη εξειδικευμένων βιοπληροφορικών εργαλείων, αλλά και η προσαρμογή ήδη υπαρχόντων συστημάτων κατάλληλων για την αποθήκευση, οπτικοποίηση και ανάλυση των δεδομένων, δίνοντας το έναυσμα για τη μεγάλη ανάπτυξη, που παρατηρείται στις μέρες μας, στο πεδίο της Βιοπληροφορικής.

Οι βιολογικές βάσεις δεδομένων, αποτελούν βασικό κομμάτι της σύγχρονης βιοπληροφορικής, καθώς αποτελούν τη βασική πηγή δεδομένων από την οποία ένας ερευνητής αντλεί τα δεδομένα στα οποία θα βασίσει την ανάλυση του. Σήμερα, οι βάσεις περιέχουν πολύ μεγάλο όγκο δεδομένων ενώ είναι απαραίτητο να ανανεώνονται καθημερινά. Η συντήρηση μιας βάσης απαιτεί μεγάλο αριθμό εξειδικευμένων επιστημόνων που θα ασχολούνται αποκλειστικά με την επισήμανση ενδεχόμενων λαθών καθώς και με το σχολιασμό (annotation) των νεοεισερχόμενων δεδομένων.

Η εμφάνιση των εργαλείων εμπλουτισμού για τη διαχείριση των αποτελεσμάτων γονιδιακής έκφρασης έχει επιφέρει σημαντικές αλλαγές στην ανάλυση δεδομένων -omics, συμπεριλαμβανομένης της γονιδιωματικής, πρωτεωμικής και μεταβολικής ανάλυσης. Η διαχείριση αυτών των δεδομένων για ερμηνεία συνιστά ένα απαιτητικό και πολυεπίπεδο έργο, καθώς τα βιολογικά συστήματα και οι αλληλεπιδράσεις εντός αυτών και μεταξύ των βιομορίων που τα συνιστούν είναι εκ φύσεως πολύπλοκες και ένα μεγάλο μέρος τους είναι ακόμα και σήμερα αχαρτογράφητο. Ταυτόχρονα ο όγκος της βιολογικής πληροφορίας είναι τόσο μεγάλος που πλέον ο επιστήμονας πρέπει να καταφέρει να προσεγγίσει κάθε πιθανή πηγή που μπορεί να του φανεί χρήσιμη, κάτι το οποίο θα ήταν χρονοβόρο και επιφοβο χωρίς τα εργαλεία εμπλουτισμού.

Ως εκ τούτου, τα εργαλεία που χρησιμοποιούνται για την ανάλυση και τον εντοπισμό των υποκείμενων βιολογικών διεργασιών που εμπλέκονται σε αυτά τα δεδομένα είναι σημαντικά και απαραίτητα για τη λήψη χρήσιμων βιολογικών δεδομένων. Η ανάγκη οργάνωσης όλης της γνωστής βιολογικής πληροφορίας με σαφή τρόπο, ώστε να αποτυπώνονται με ακρίβεια το σύνολο της βιολογικής γνώσης και οι ιεραρχικές σχέσεις μεταξύ των διαφόρων λειτουργιών, οδήγησε στη δημιουργία πολύπλοκων ιεραρχικών ταξινομήσεων με τη χρήση εξειδικευμένων εργαλείων βιοπληροφορικής. Ειδικότερα τα γονίδια οργανώνονται με βάση τις λειτουργικές τους ιδιότητες, όπως τα βιολογικά μονοπάτια ή οι βασικές διεργασίες στις οποίες συμμετέχουν. Οι σχέσεις μεταξύ γονιδίων και λειτουργιών δεν είναι αλληλοαποκλειόμενες, αλλά ελεύθερες, δηλαδή ένα γονίδιο μπορεί να αντιστοιχηθεί σε περισσότερες από μια λειτουργίες.

Σήμερα ένας μεγάλος αριθμός εργαλείων εμπλουτισμού έχουν αναπτυχθεί για την αντιμετώπιση του προβλήματος λειτουργικής ανάλυσης μεγάλων λιστών γονιδίων. Η ανάλυση εμπλουτισμού είναι μια πολλά υποσχόμενη στρατηγική υψηλής απόδοσης που μπορεί να αυξήσει

την πιθανότητα για τους ερευνητές να εντοπίσουν βιολογικές διεργασίες που είναι πιο σχετικές με τη μελέτη τους. Μέχρι τώρα, η ανάπτυξη εργαλείων εξακολουθεί να αναπτύσσεται με ταχείς ρυθμούς και αυτός ο τομέας είναι πολύ παραγωγικός. Ωστόσο, ο υπερβολικός αριθμός εργαλείων εμπλουτισμού μπορεί να προκαλέσει δυσκολίες στους χρήστες να επιλέξουν μεταξύ όλων των υπαρχόντων εργαλείων για την καλύτερη ανάλυση των δεδομένων τους.

Το FLAME, συνιστά ένα διαδικτυακό εργαλείο για συνδυαστικές αναλύσεις δεδομένων, καθώς επιτρέπει τον συνδυασμό πολλών λιστών πριν από την ανάλυση εμπλουτισμού. Οι χρήστες μπορούν ως αρχεία εισόδου να χρησιμοποιήσουν αρκετές λίστες και να τις επεξεργαστούν με τη χρήση διαδραστικών γραφημάτων UpSet, ως αποτελεσματικότερη εναλλακτική λύση έναντι των διαγραμμάτων Venn, με σκοπό την εύρεση τομών/ενώσεων κλπ μεταξύ όλων των επιθυμητών συνδυασμών των αρχείων. Ο εμπλουτισμός λειτουργικότητας και βιβλιογραφίας, καθώς και ενσωματωμένα εργαλεία για μετατροπές αναγνωριστικών των γονιδίων και εύρεση ορθόλογων γονιδίων προσφέρονται από τις εφαρμογές g:Profiler και aGOtool για 197 οργανισμούς.

Στην παρούσα έκδοση, το FLAME μπορεί να αναλύει γονίδια/πρωτεΐνες και να τα εντάσσει σε γονιδιακές οντολογίες, μονοπάτια, ρυθμιστικά μοτίβα, λειτουργικές επικράτειες, ασθένειες, φαινότυπους και βιβλιογραφικές αναφορές, ενώ μπορεί επίσης να δημιουργήσει δίκτυα πρωτεϊνικών αλληλεπιδράσεων που προέρχονται από την STRING. Για τον έλεγχο της λειτουργικότητας του FLAME, πραγματοποιήθηκε μελέτη δεδομένων γονιδιακής έκφρασης που σχετίζονται με την ευαισθησία του περιφερικού τμήματος του παχέος εντέρου στον πειραματικό καρκίνο του παχέος εντέρου (Καρκίνος σε έδαφος κολίτιδας).

Η εφαρμογή FLAME συνιστά μια διαδραστική φιλική προς το χρήστη πλατφόρμα που επιτρέπει τον εύκολο χειρισμό δεδομένων και αποτελεσμάτων, τα οποία μπορούν να απεικονιστούν ως διαδραστικοί και παραμετροποιήσιμοι πίνακες με τα αντίστοιχα διαγράμματα heatmaps, barcharts, διαγράμματα Manhattan και δίκτυα.

Η δυνατότητα που παρέχει στους χρήστες να χειρίζονται πληθώρα γονιδιακών/πρωτεϊνικών λιστών μέσω διαδραστικών UpSet plot του προσδίδει ένα πλεονέκτημα σε σχέση με τα ήδη υπάρχοντα εργαλεία που εκτελούν ανάλυση λειτουργικού εμπλουτισμού, καθώς ο χειρισμός πολλαπλών λιστών γονιδίων που προκύπτουν από πειράματα γονιδιακής έκφρασης μεγάλης κλίμακας κατά τη διεξαγωγή μιας ανάλυσης λειτουργικού εμπλουτισμού είναι συχνά ένα δύσκολο κομμάτι της ανάλυσης. Το FLAME συνδυάζει έναν σημαντικό αριθμό δυνατοτήτων και εργαλείων για βιοπληροφορικές αναλύσεις παρέχοντας στο χρήστη ένα ολοκληρωμένο περιβάλλον για τις αναλύσεις του.

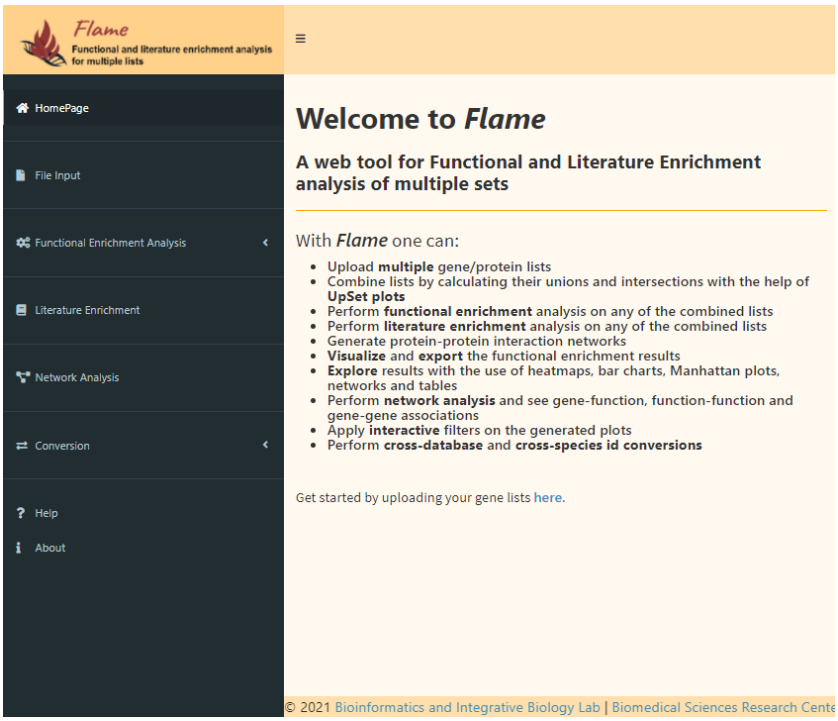
Μελλοντικά, για βελτίωση και αναβάθμιση της παρούσας έκδοσης προτείνεται αύξηση/προσθήκη οργανισμών συμπεριλαμβανομένων οργανισμών όπως τα βακτηρία, data integration (ανάλυση έτοιμων συνόλων δεδομένων από εξωτερικές βάσεις δεδομένων), κλήση του FLAME από τα υπόλοιπα εργαλεία του εργαστηρίου μας μέσω του GET request, ενσωμάτωση περισσότερων τύπων αναγνωριστικών και σύνδεση με επιπλέον βάσεις δεδομένων.

5. ΔΙΑΘΕΣΙΜΟΤΗΤΑ

Από την παρούσα πτυχιακή προέκυψε οι εξής επιστημονικές δημοσιεύσεις:

- **F. Thanati**, E. Karatzas, F. Baltoumas, D. J. Stravopodis, A. G. Eliopoulos, and G. Pavlopoulos, "FLAME: a web tool for functional and literature enrichment analysis of multiple gene lists," Biology, 2021, doi: 10.3390/biology10070665
- Baltoumas FA, Zafeiropoulou S, Karatzas E, Koutrouli M, **Thanati F,** Voutsadaki K, Gkonta M, Hotova J, Kasionis I, Hatzis P, Pavlopoulos GA. Biomolecule and Bioentity Interaction Databases in Systems Biology: A Comprehensive Review. Biomolecules. 2021; 11(8):1245. https://doi.org/10.3390/biom11081245
- Baltoumas, F.A., Zafeiropoulou, S., Karatzas, E., Paragkamian, S., **Thanati, F.**, Iliopoulos, I., Eliopoulos, A.G., Schneider, R., Jensen, L.J., Pafilis, E., Pavlopoulos, G.A. (2021) OnTheFly2.0: a text-mining web application for automated biomedical entity recognition, document annotation, network and functional enrichment analysis. bioRxiv 2021.05.14.444150. http://doi.org/10.1101/2021.05.14.444150

Η διαδικτυακή εφαρμογή του FLAME είναι διαθέσιμη στην εξής ιστοσελίδα: http://flame.pavlopouloslab.info. Στην Εικόνα 5.1 φαίνεται η αρχική σελίδα της εφαρμογής.



Εικόνα 5.1: Αρχική σελίδα της εφαρμογής FLAME.

Ο κώδικας είναι διαθέσιμος και ελεύθερα προσβάσιμος στο αποθετήριο GitHub: https://github.com/PavlopoulosLab/FLAME

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] F. A. Baltoumas *et al.*, "Biomolecule and Bioentity Interaction Databases in Systems Biology: A Comprehensive Review," *Biomolecules*, vol. 11, no. 8, p. 1245, Aug. 2021, doi: 10.3390/biom11081245.

[2] G. C. K. W. Koh, P. Porras, B. Aranda, H. Hermjakob, and S. E. Orchard, "Analyzing protein-protein interaction networks," *J. Proteome Res.*, vol. 11, no. 4, pp. 2014–2031, Apr. 2012, doi: 10.1021/pr201211w.

[3] J. D. L. Rivas and C. Fontanillo, "Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks," *PLOS Comput. Biol.*, vol. 6, no. 6, p. e1000807, 2010, doi: 10.1371/journal.pcbi.1000807.

[4] B. Jassal *et al.*, "The reactome pathway knowledgebase," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D498–D503, Jan. 2020, doi: 10.1093/nar/gkz1031.

[5] J. Hastings *et al.*, "ChEBI in 2016: Improved services and an expanding collection of metabolites," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1214-1219, Jan. 2016, doi: 10.1093/nar/gkv1031.

[6] G. Wu, E. Dawson, A. Duong, R. Haw, and L. Stein, "ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis," *F1000Research*, vol. 3, p. 146, Sep. 2014, doi: 10.12688/f1000research.4431.2.

[7] M. Martens *et al.*, "WikiPathways: connecting communities," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D613–D621, Jan. 2021, doi: 10.1093/nar/gkaa1024.

[8] M. Kutmon, S. Lotia, C. T. Evelo, and A. R. Pico, "WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization." F1000Research, Sep. 11, 2014. doi: 10.12688/f1000research.4254.2.

[9] M. Kutmon *et al.*, "PathVisio 3: An Extendable Pathway Analysis Toolbox," *PLOS Comput. Biol.*, vol. 11, no. 2, p. e1004085, 2015, doi: 10.1371/journal.pcbi.1004085.

[10] P. Shannon *et al.*, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.

[11] K. L. Howe *et al.*, "Ensembl 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D884–D891, Jan. 2021, doi: 10.1093/nar/gkaa942.

[12] M. Haeussler *et al.*, "The UCSC Genome Browser database: 2019 update," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D853–D858, Jan. 2019, doi: 10.1093/nar/gky1095.

[13] R. Hoffmann, "A wiki for the life sciences where authorship matters," *Nat. Genet.*, vol. 40, no. 9, pp. 1047–1051, Sep. 2008, doi: 10.1038/ng.f.217.

[14] H. M. Berman *et al.*, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, Jan. 2000, doi: 10.1093/nar/28.1.235.

[15] M. Giurgiu *et al.*, "CORUM: the comprehensive resource of mammalian protein complexes-2019," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D559–D563, Jan. 2019, doi: 10.1093/nar/gky973.

[16] M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.

[17] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D353–D361, Jan. 2017, doi: 10.1093/nar/gkw1092.

[18] R. Alcántara *et al.*, "Rhea—a manually curated resource of biochemical reactions," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D754–D760, Jan. 2012, doi: 10.1093/nar/gkr1126.

[19] "Cytoscape App Store - CytoKegg." https://apps.cytoscape.org/apps/cytokegg (accessed Sep. 28, 2021).

[20] L. Nersisyan, R. Samsonyan, and A. Arakelyan, "CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows," *F1000Research*, vol. 3, p. 145, 2014, doi: 10.12688/f1000research.4410.2.

[21] Gene Ontology Consortium, "The Gene Ontology resource: enriching a GOld mine," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D325–D334, Jan. 2021, doi: 10.1093/nar/gkaa1113.

[22] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.
[23] J. Mistry *et al.*, "Pfam: The protein families database in 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D412–D419, Jan. 2021, doi: 10.1093/nar/gkaa913.
[24] M. Blum *et al.*, "The InterPro protein families and domains database: 20 years on," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D344–D354, Jan. 2021, doi: 10.1093/nar/gkaa977.
[25] I. Sillitoe *et al.*, "CATH: expanding the horizons of structure-based functional annotations for genome sequences," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D280–D284, Jan. 2019, doi: 10.1093/nar/gky1097.
[26] S. Lu *et al.*, "CDD/SPARCLE: the conserved domain database in 2020," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D265–D268, Jan. 2020, doi: 10.1093/nar/gkz991.
[27] I. Pedruzzi *et al.*, "HAMAP in 2015: updates to the protein family classification and annotation system," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D1064-1070, Jan. 2015, doi: 10.1093/nar/gku1002.
[28] D. Piovesan *et al.*, "MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D471–D476, Jan. 2018, doi: 10.1093/nar/gkx1071.
[29] H. Mi *et al.*, "PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D394–D403, Jan. 2021, doi: 10.1093/nar/gkaa1106.
[30] A. N. Nikolskaya, C. N. Arighi, H. Huang, W. C. Barker, and C. H. Wu, "PIRSF family classification system for protein functional and evolutionary analysis," *Evol. Bioinforma. Online*, vol. 2, pp. 197–209, Feb. 2007.
[31] T. K. Attwood *et al.*, "The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012," *Database J. Biol. Databases Curation*, vol. 2012, p. bas019, 2012, doi: 10.1093/database/bas019.
[32] C. J. A. Sigrist *et al.*, "New and continuing developments at PROSITE," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D344-347, Jan. 2013, doi: 10.1093/nar/gks1067.
[33] E. Akiva *et al.*, "The Structure-Function Linkage Database," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D521-530, Jan. 2014, doi: 10.1093/nar/gkt1130.
[34] I. Letunic and P. Bork, "20 years of the SMART protein domain annotation resource," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D493–D496, Jan. 2018, doi: 10.1093/nar/gkx922.
[35] A. P. Pandurangan, J. Stahlhacke, M. E. Oates, B. Smithers, and J. Gough, "The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D490–D494, Jan. 2019, doi: 10.1093/nar/gky1130.
[36] D. H. Haft, J. D. Selengut, R. A. Richter, D. Harkins, M. K. Basu, and E. Beck, "TIGRFAMs and Genome Properties in 2013," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D387-395, Jan. 2013, doi: 10.1093/nar/gks1234.
[37] D. A. Lindberg, "Internet access to the National Library of Medicine," *Eff. Clin. Pract. ECP*, vol. 3, no. 5, pp. 256–260, Oct. 2000.
[38] E. Wingender, T. Schoeps, and J. Dönitz, "TFClass: an expandable hierarchical classification of human transcription factors," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D165-170, Jan. 2013, doi: 10.1093/nar/gks1123.
[39] V. Matys *et al.*, "TRANSFAC®: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 374–378, Jan. 2003.
[40] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "miRBase: from microRNA sequences to function," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D155–D162, Jan. 2019, doi: 10.1093/nar/gky1141.
[41] M. Uhlén *et al.*, "Proteomics. Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, p. 1260419, Jan. 2015, doi: 10.1126/science.1260419.
[42] S. Köhler *et al.*, "The Human Phenotype Ontology in 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1207–D1217, Jan. 2021, doi: 10.1093/nar/gkaa1043.
[43] The UniProt Consortium, "UniProt: the universal protein knowledgebase in 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D480–D489, Jan. 2021, doi: 10.1093/nar/gkaa1100.

[44] L. M. Schriml *et al.*, "Human Disease Ontology 2018 update: classification, content and workflow expansion," *Nucleic Acids Res.*, vol. 47, no. Database issue, pp. D955–D962, Jan. 2019, doi: 10.1093/nar/gky1032.
[45] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, Jan. 2009, doi: 10.1093/nar/gkn923.
[46] F. Maleki, K. Ovens, D. J. Hogan, and A. J. Kusalik, "Gene Set Analysis: Challenges, Opportunities, and Future Research," *Front. Genet.*, vol. 11, p. 654, 2020, doi: 10.3389/fgene.2020.00654.
[47] P. H. Guzzi, "Functional Enrichment Analysis Methods," in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 896–897. doi: 10.1016/B978-0-12-809633-8.20404-4.
[48] A. Subramanian *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/pnas.0506580102.
[49] C. Nikolaou, P. Chouvardas, Χ. Νικολάου, and Π. Χουβαρδάς, "Ανάλυση της Γονιδιακής Έκφρασης," 2015, Accessed: Sep. 24, 2021. [Online]. Available: https://repository.kallipos.gr/handle/11419/1585
[50] P. H. Guzzi, M. Mina, C. Guerra, and M. Cannataro, "Semantic similarity analysis of protein data: assessment with biological features and issues," *Brief. Bioinform.*, vol. 13, no. 5, pp. 569–585, Sep. 2012, doi: 10.1093/bib/bbr066.
[51] A. Keller, C. Backes, and H.-P. Lenhof, "Computation of significance scores of unweighted Gene Set Enrichment Analyses," *BMC Bioinformatics*, vol. 8, p. 290, Aug. 2007, doi: 10.1186/1471-2105-8-290.
[52] N. R. Clark and A. Ma'ayan, "Introduction to statistical methods for analyzing large data sets: gene-set enrichment analysis," *Sci. Signal.*, vol. 4, no. 190, p. tr4, Sep. 2011, doi: 10.1126/scisignal.2001966.
[53] G. Montana, "Statistical methods in genetics," *Brief. Bioinform.*, vol. 7, no. 3, pp. 297–308, Sep. 2006, doi: 10.1093/bib/bbl028.
[54] S. Falcon and R. Gentleman, "Hypergeometric Testing Used for Gene Set Enrichment Analysis," in *Bioconductor Case Studies*, New York, NY: Springer New York, 2008, pp. 207–220. doi: 10.1007/978-0-387-77240-0_14.
[55] "Εισαγωγή στις Πιθανότητες.Θεωρία και Εφαρμογές.Μέρος Ι.-Β΄ Έκδοση," *stamoulis.gr*. https://www.stamoulis.gr/Εισαγωγή-στις-Πιθανότητες-Θεωρία-και-Εφαρμογές-Μέρος-Ι-Β΄-Έκδοση_p-373872.aspx (accessed Oct. 23, 2021).
[56] T. K. Kim, "T test as a parametric statistic," *Korean J. Anesthesiol.*, vol. 68, no. 6, pp. 540–546, Dec. 2015, doi: 10.4097/kjae.2015.68.6.540.
[57] S. Y. Rhee, V. Wood, K. Dolinski, and S. Draghici, "Use and misuse of the gene ontology annotations," *Nat. Rev. Genet.*, vol. 9, no. 7, pp. 509–515, Jul. 2008, doi: 10.1038/nrg2363.
[58] U. Raudvere *et al.*, "g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W191–W198, Jul. 2019, doi: 10.1093/nar/gkz369.
[59] Y. Liao, J. Wang, E. J. Jaehnig, Z. Shi, and B. Zhang, "WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W199–W205, Jul. 2019, doi: 10.1093/nar/gkz401.
[60] E. Y. Chen *et al.*, "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool," *BMC Bioinformatics*, vol. 14, p. 128, Apr. 2013, doi: 10.1186/1471-2105-14-128.
[61] Z. Xie *et al.*, "Gene Set Knowledge Discovery with Enrichr," *Curr. Protoc.*, vol. 1, no. 3, p. e90, 2021, doi: 10.1002/cpz1.90.
[62] Y. Zhou *et al.*, "Metascape provides a biologist-oriented resource for the analysis of systems-level datasets," *Nat. Commun.*, vol. 10, no. 1, p. 1523, Apr. 2019, doi: 10.1038/s41467-019-09234-6.
[63] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, 2009, doi: 10.1038/nprot.2008.211.
[64] C. Schölz, D. Lyon, J. C. Refsgaard, L. J. Jensen, C. Choudhary, and B. T. Weinert, "Avoiding abundance bias in the functional annotation of post-translationally modified proteins," *Nat. Methods*,

vol. 12, no. 11, pp. 1003–1004, Nov. 2015, doi: 10.1038/nmeth.3621.
[65] F. A. Baltoumas *et al.*, "OnTheFly2.0: a text-mining web application for automated biomedical entity recognition, document annotation, network and functional enrichment analysis," May 2021. doi: 10.1101/2021.05.14.444150.
[66] E. Pafilis *et al.*, "EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation," *Database*, vol. 2016, no. baw005, Jan. 2016, doi: 10.1093/database/baw005.
[67] E. Pafilis and L. J. Jensen, "Real-time tagging of biomedical entities," Sep. 2016. doi: 10.1101/078469.
[68] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists," *BMC Bioinformatics*, vol. 10, p. 48, Feb. 2009, doi: 10.1186/1471-2105-10-48.
[69] S. Carbon *et al.*, "AmiGO: online access to ontology and annotation data," *Bioinforma. Oxf. Engl.*, vol. 25, no. 2, pp. 288–289, Jan. 2009, doi: 10.1093/bioinformatics/btn615.
[70] A. Barve and A. Wagner, "A latent capacity for evolutionary innovation through exaptation in metabolic systems," *Nature*, vol. 500, no. 7461, pp. 203–206, Aug. 2013, doi: 10.1038/nature12301.
[71] M. Koutrouli, E. Karatzas, D. Paez-Espino, and G. A. Pavlopoulos, "A Guide to Conquer the Biological Network Era Using Graph Theory," *Front. Bioeng. Biotechnol.*, vol. 8, p. 34, 2020, doi: 10.3389/fbioe.2020.00034.
[72] K. Sun, N. Buchan, C. Larminie, and N. Pržulj, "The integrated disease network," *Integr. Biol.*, vol. 6, no. 11, pp. 1069–1079, Oct. 2014, doi: 10.1039/C4IB00122B.
[73] M. Jiang, C. Niu, J. Cao, D. Ni, and Z. Chu, "In silico-prediction of protein–protein interactions network about MAPKs and PP2Cs reveals a novel docking site variants in Brachypodium distachyon," *Sci. Rep.*, vol. 8, no. 1, p. 15083, Oct. 2018, doi: 10.1038/s41598-018-33428-5.
[74] L. J. Jensen *et al.*, "STRING 8--a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D412-416, Jan. 2009, doi: 10.1093/nar/gkn760.
[75] F. Morandat, B. Hill, L. Osvald, and J. Vitek, "Evaluating the Design of the R Language," in *ECOOP 2012 – Object-Oriented Programming*, Berlin, Heidelberg, 2012, pp. 104–131. doi: 10.1007/978-3-642-31057-7_6.
[76] D. Attali, "Building Shiny apps - an interactive tutorial," *Dean Attali*, Dec. 07, 2015. https://deanattali.com/blog/building-shiny-apps-tutorial/ (accessed Oct. 04, 2021).
[77] Winston Chang *et al.*, "Web Application Framework for R. R package version 1.6.0." [Online]. Available: https://CRAN.R-project.org/package=shiny
[78] P. Bardou, J. Mariette, F. Escudié, C. Djemiel, and C. Klopp, "jvenn: an interactive Venn diagram viewer," *BMC Bioinformatics*, vol. 15, p. 293, Aug. 2014, doi: 10.1186/1471-2105-15-293.
[79] B. Hur, D. Kang, S. Lee, J. H. Moon, G. Lee, and S. Kim, "Venn-diaNet : venn diagram based network propagation analysis framework for comparing multiple biological experiments," *BMC Bioinformatics*, vol. 20, no. Suppl 23, p. 667, Dec. 2019, doi: 10.1186/s12859-019-3302-7.
[80] K. K. Gkouskou *et al.*, "Apolipoprotein A-I inhibits experimental colitis and colitis-propelled carcinogenesis," *Oncogene*, vol. 35, no. 19, pp. 2496–2505, May 2016, doi: 10.1038/onc.2015.307.
[81] A. K. Rustgi, "The genetics of hereditary colon cancer," *Genes Dev.*, vol. 21, no. 20, pp. 2525–2538, Oct. 2007, doi: 10.1101/gad.1593107.
[82] S. I. Grivennikov, "Inflammation and colorectal cancer: colitis-associated neoplasia," *Semin. Immunopathol.*, vol. 35, no. 2, pp. 229–244, Mar. 2013, doi: 10.1007/s00281-012-0352-6.
[83] S. N. Harbour, C. L. Maynard, C. L. Zindl, T. R. Schoeb, and C. T. Weaver, "Th17 cells give rise to Th1 cells that are required for the pathogenesis of colitis," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 22, pp. 7061–7066, Jun. 2015, doi: 10.1073/pnas.1415675112.
[84] L. G. Perez *et al.*, "TGF-β signaling in Th17 cells promotes IL-22 production and colitis-associated colon cancer," *Nat. Commun.*, vol. 11, no. 1, p. 2608, May 2020, doi: 10.1038/s41467-020-16363-w.
[85] L. A. Dieleman *et al.*, "Chronic experimental colitis induced by dextran sulphate sodium (DSS) is characterized by Th1 and Th2 cytokines," *Clin. Exp. Immunol.*, vol. 114, no. 3, pp. 385–391, Dec. 1998,

doi: 10.1046/j.1365-2249.1998.00728.x.
[86] E. Osawa *et al.*, "Predominant T helper type 2-inflammatory responses promote murine colon cancers," *Int. J. Cancer*, vol. 118, no. 9, pp. 2232–2236, May 2006, doi: 10.1002/ijc.21639.
[87] C. McElrath *et al.*, "Critical role of interferons in gastrointestinal injury repair," *Nat. Commun.*, vol. 12, no. 1, p. 2624, May 2021, doi: 10.1038/s41467-021-22928-0.
[88] A. L. Means *et al.*, "Epithelial Smad4 Deletion Up-Regulates Inflammation and Promotes Inflammation-Associated Cancer," *Cell. Mol. Gastroenterol. Hepatol.*, vol. 6, no. 3, pp. 257–276, 2018, doi: 10.1016/j.jcmgh.2018.05.006.
[89] E. Troncone, I. Marafini, C. Stolfi, and G. Monteleone, "Transforming Growth Factor-β1/Smad7 in Intestinal Immunity, Inflammation, and Cancer," *Front. Immunol.*, vol. 9, p. 1407, 2018, doi: 10.3389/fimmu.2018.01407.