Τα παρακάτω διαγράμματα δημιουργήθηκαν με τη συνάρτηση enrichrVisual, η οποία παίρνει ως είσοδο τον πίνακα δεδομένων [data.visual](#data_visual), τον φάκελο αποθήκευσης των διαγραμμάτων και τον όρο με τον οποίο έγινε εμπλουτισμός (παράδειγμα: “KEGG Pathways”) για την προσθήκη των κατάλληλων τίτλων στα διαγράμματα.

1) Density plot-P values of KEGG Pathways

* κατανομές των τιμών των στηλών P.value, P.adjust.

2) Top KEGG Pathways

* μέτρηση το πόσες φορές (n) εμφανίζεται κάθε μονοπάτι στον πίνακα data.visual
* συνένωση της πληροφορίας με τον πίνακα data.visual
* επιλογή των στηλών: Term, n, P.adjust
* ομαδοποίηση με βάση τις στήλες Term, n και summarise το P.adjust ώστε να είναι ο μέσος όρος των διαφορετικών τιμών που εμφανίζονται στα διαφορετικά TADs
* ταξινόμηση του πίνακα με φθίνουσα σειρά ως προς τη στήλη n
* αν οι γραμμές του πίνακα ξεπερνούν τις 30, τότε επιλογή των 30 πρώτων γραμμών και αναφορά αυτής της επιλογής ως υπότιτλο στο γράφημα
* δημιουργία ιστογράμματος, στο οποίο ο άξονας x αντιπροσωπεύει τη μέση τιμή P.adjust, ο άξονας y τους όρους KEGG Pathways και το χρώμα των ράβδων τη μεταβλητή n.

3) KEGG Pathways in different TADs

* χρήση των στηλών Term, n του πίνακα που δημιουργήθηκε για το προηγούμενο γράφημα
* δημιουργία ιστογράμματος, στο οποίο ο άξονας x αντιπροσωπεύει τον αριθμό των TADs, στα οποία βρέθηκε ο κάθε όρος και ο άξονας y τους όρους KEGG Pathways.

4) Adjusted P values per TAD

* ομαδοποίηση του πίνακα data.visual ανά TAD
* για κάθε TAD:
  + ταξινόμηση των γραμμών κατά φθίνουσα σειρά σύμφωνα με τη στήλη P.adjust
  + αν οι γραμμές του πίνακα ξεπερνούν τις 30, τότε επιλογή των 30 πρώτων γραμμών και αναφορά αυτής της επιλογής ως υπότιτλο στο γράφημα
  + δημιουργία ιστογράμματος για κάθε TAD στο οποίο ο άξονας y αναπαριστά τους KEGG Pathways που βρέθηκαν στο TAD και ο άξονας x την πιθανότητα P.adjust που αντιστοιχεί σε κάθε όρο.

5) Adjusted P values per KEGG Pathways

* ομαδοποίηση του πίνακα data.visual ανά KEGG Pathway
* για κάθε Term:
  + ταξινόμηση των γραμμών κατά φθίνουσα σειρά σύμφωνα με τη στήλη P.adjust
  + αν οι γραμμές του πίνακα ξεπερνούν τις 30, τότε επιλογή των 30 πρώτων γραμμών και αναφορά αυτής της επιλογής ως υπότιτλο στο γράφημα
  + δημιουργία ιστογράμματος για κάθε KEGG Pathway στο οποίο ο άξονας y αναπαριστά τα TADs στα οποία βρέθηκε ο όρος και ο άξονας x αναπαριστά την πιθανότητα εύρεσης P.adjust του όρο σε κάθε TAD.

6) Top 10 KEGG Psthways network graph

* μέτρηση το πόσες φορές (n) εμφανίζεται κάθε όρος (KEGG Pathway) στον πίνακα data.visual
* επιλογή των δέκα όρων με τις μεγαλύτερες τιμές n
* συνένωση με τον πίνακα data.visual
* επιλογή των στηλών Term, TAD
* δημιουργία γράφου, στον οποίο οι κόμβοι αντιστοιχούν στα TADs (κάθε TAD αντιστοιχεί σε έναν κόμβο), και οι ακμές συνδέουν τους κόμβους που ανήκουν στην ίδια ομάδα, δηλαδή KEGG Pathway. Σε κάθε ομάδα αντιστοιχεί διαφορετικό χρώμα ακμών.

7)Pathview

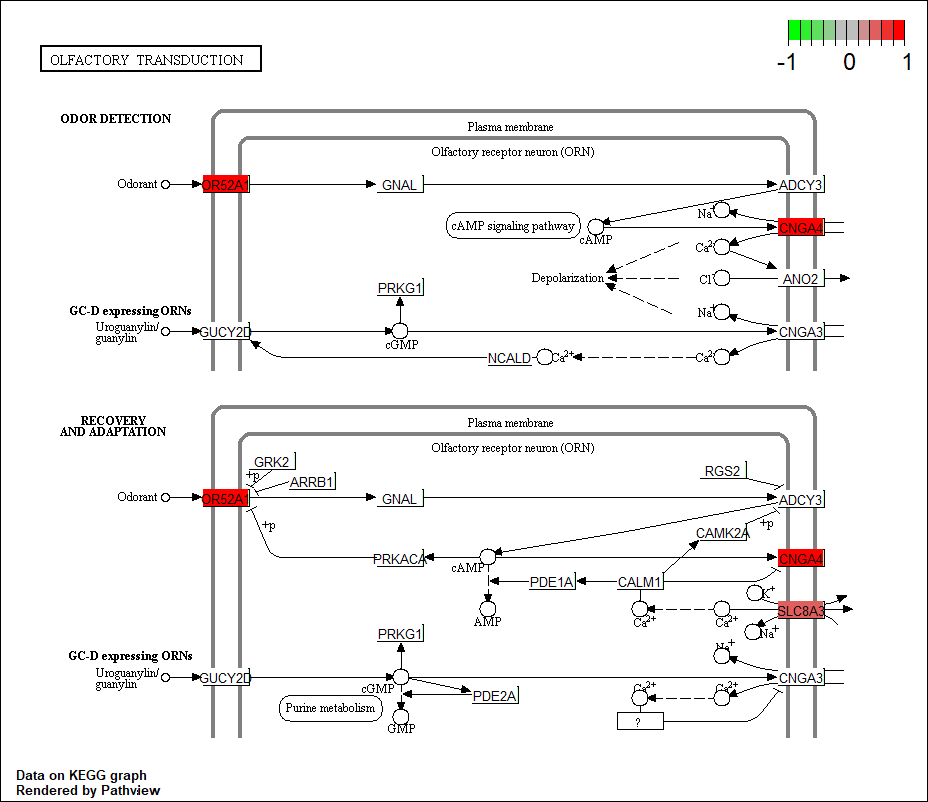
Τα διαγράμματα δημιουργήθηκαν με τη συνάρτηση pathVisual, η οποία παίρνει ως είσοδο τον πίνακα full.data (με τα δεδομένα εισόδου: integrated\_table\_with\_sign\_tads.csv ), τον πίνακα δεδομένων [pathview.input](#pathview_input) και το φάκελο αποθήκευσης των διαγραμμάτων.

* επιλογή των μονοπατιών (κάθε μονοπάτι αντιστοιχεί σε μία γραμμή) από τον πίνακα pathview.input για τα οποία έγινε αντιστοίχιση με τα kegg ids
* επιλογή των Gene\_id, diff από τον πίνακα full.data και αποθήκευση στον πίνακα data.S
* για κάθε μονοπάτι:
  + διάσπαση της στήλης Genes
  + έπειτα συνένωση με τον πίνακα data.S ως προς τις στήλες (Gene\_id == Genes)
  + ομαδοποίηση ανά path, Genes και summarise ώστε για κάθε γονίδιο να αντιστοιχεί diff = mean(diff)
  + τα δεδομένα αυτά αποτελούν είσοδο της συνάρτησης pathview
    - kegg id του μονοπατιού
    - διάνυσμα με ονόματα στοιχείων τα γονίδια του μονοπατιού (Genes) και στοιχεία τη μέση τιμή της διαφοράς (mean(diff)) ανάμεσα στους δύο τύπους δειγμάτων (ss6,ss8).

\*Όσον αφορά το μικρό αριθμό γονιδίων σε ορισμένα μονοπάτια, παρατήρησα ότι ο αριθμός των γονιδίων που εμφανίζονται στη χαρτογράφηση του μονοπατιού μπορεί να είναι μικρότερος από τον αριθμό των γονιδίων που έγιναν αντιστοίχιση ανάμεσα στο dataset και τη βιβλιοθήκη KEGG\_Human\_2019. Παράδειγμα αποτελεί το μονοπάτι “Olfactory transduction”, το οποίο έγινε αντιστοίχιση στο dataset εισόδου σε 67 γονίδια:

OR10J1;OR10J3;OR52J3;OR10J5;OR51D1;OR52N1;OR52N5;OR52N4;OR52N2;OR51L1;OR52W1;OR6Y1;OR51Q1;OR56B1;OR52B4;OR52B2;CNGA4;OR56B4;OR52B6;OR51G2;OR51G1;OR52I2;OR52I1;OR52E4;OR52E2;OR6P1;OR52M1;OR10Z1;OR56A5;OR52R1;OR51T1;OR52E8;OR52E6;OR52A1;OR56A4;OR56A3;OR52A5;OR56A1;OR51F2;OR6K6;OR51F1;OR52H1;OR51B2;OR6K3;OR6K2;OR52D1;OR10T2;OR52L1;OR51S1;OR51B6;OR51B5;OR51B4;OR10K2;OR51E1;OR51I2;OR51E2;OR51M1;OR6N2;OR6N1;OR52K2;OR51I1;OR52K1;SLC8A3;OR51V1;OR10X1;OR51A2;OR51A7

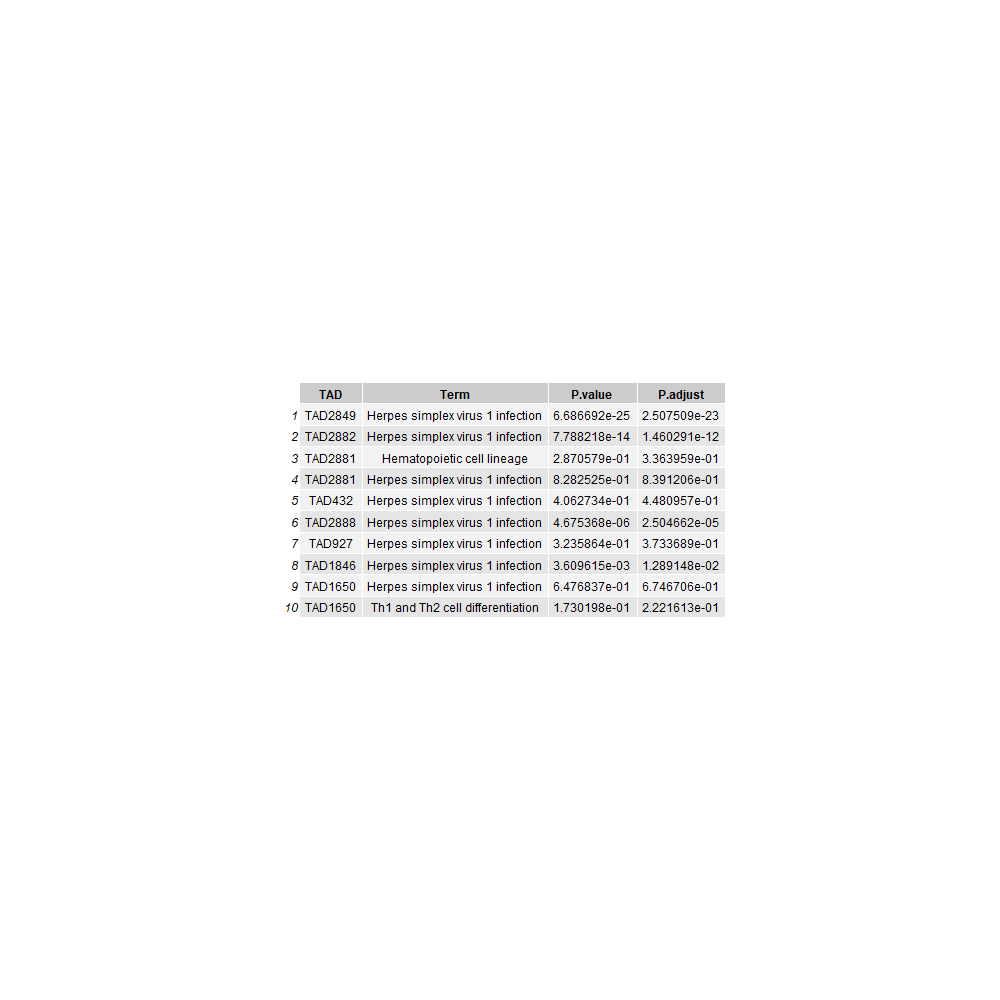
και η έξοδος της συνάρτησης pathview τα αντιστοιχεί σε πολύ λιγότερους κόμβους:



KEGG γράφημα του μονοπατιού Olfactory transduction.

\*\*Ωστόσο και πάλι όταν εφαρμόζω τη μεθοδολογία εμπλουτισμού ανά TAD (και μόνο τότε!), εφαρμόζοντας cut-off: P value<0.05, έχω στο αποτέλεσμα μονοπάτια ακόμη και με ένα γονίδιο, ενώ έχω συνολικά πάνω από 200 στατιστικά σημαντικά μονοπάτια. Το τελευταίο ειδικά με οδηγεί στο συμπέρασμα ότι τελικά ο εμπλουτισμός ανά TAD δεν είναι σωστή στρατηγική στη συγκεκριμένη περίπτωση, καθώς αντιτίθεται του στόχου της εργασίας, ειδικά όταν στο ανθρώπινο γονιδίωμα υπάρχουν συνολικά 537 μονοπάτια και τα 200 από αυτά θεωρούνται σημαντικά.

**Βήματα για την παραγωγή του πίνακα εισόδου data.visual της συνάρτησης enrichrVisual**

****

Στιγμιότυπο του πίνακα data.visual.

**-enrich per TAD (με υπογράμμιση βρίσκονται τα διαφορετικά σημεία σε σχέση με τη μέθοδο enrich all)**

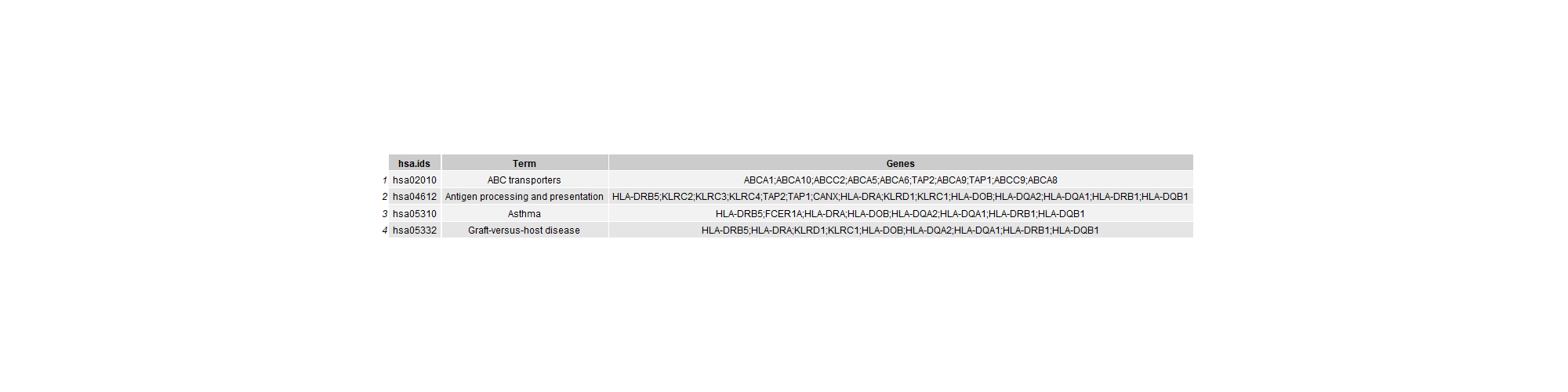
Μεθοδολογία εμπλουτισμού – enrichPerTAD

* χρήση της συνάρτησης enrichR με εισόδους:
  + τη στήλη Gene\_id των δεδομένων εισόδου (integrated\_table\_with\_sign\_tads.csv / έξοδος της TADiff.R), σε αυτή την περίπτωση όχι ολόκληρη τη στήλη, αλλά τα γονίδια που αντιστοιχούν σε κάθε TAD ξεχωριστά
  + τις βάσεις δεδομένων σύμφωνα με τις οποίες θα γίνει ο εμπλουτισμός GO\_Molecular\_Function\_2018 , GO\_Biological\_Process\_2018 , KEGG\_2019\_Human
* η συνάρτηση επιστρέφει μία λίστα με τρεις πίνακες, ο καθένας με 9 στήλες (Term, Overlap, P.value, Adjusted.P.value, Old.P.value, Old.Adjusted.P.value, Odds.Patio, Combined.Score, Genes) και γίνεται επιλογή των αποτελεσμάτων με P.value < 0.05
* τέλος γίνεται συνένωση (bind\_rows) των επιμέρους πινάκων για κάθε βάση δεδομένων που χρησιμοποιήθηκε για εμπλουτισμό

Ανάλυση δεδομένων και υπολογισμός p value – analysisPerTAD, calculatePvalue, produceOutputs

* η ανάλυση γίνεται ξεχωριστά για κάθε πίνακα με τα αποτελέσματα του εμπλουτισμού από τη συνάρτηση enrichPerTAD
* επιλογή των στηλών Term, Overlap,Genes
* η στήλη Overlap είναι τύπου character και έχει τη μορφή: αριθμητής/παρονομαστής, όπου ο αριθμητής αντιπροσωπεύει τον αριθμό γονιδίων από τα δεδομένα εισόδου που αντιστοιχούν σε αυτό τον όρο (αυτά αναγράφονται αναλυτικά στη στήλη Genes) και ο παρονομαστής το συνολικό αριθμό γονιδίων του ανθρώπινου γονιδιώματος που αντιστοιχούν σε αυτό τον όρο.
* γίνεται χωρισμός της στήλης σε δύο επιμέρους στήλες numerator και denominator, ώστε να υπάρχει σε ξεχωριστή στήλη διαθέσιμη η πληροφορία για τον αριθμό των γονιδίων που αντιστοιχούν σε κάθε όρο στη βάση δεδομένων
* ομαδοποίηση των δεδομένων με βάση των όρο (Term) και summarise για τη δημιουργία πίνακα με τρεις στήλες : Term, denominator, Genes. Σε αυτόν, κάθε γραμμή αντιστοιχεί σε έναν όρο, ενώ στη στήλη Genes βρίσκονται όλα τα γονίδια του dataset που αντιστοιχούν σε αυτόν
* χωρισμός της στήλης Genes, ώστε σε κάθε γραμμή να βρίσκεται ένα γονίδιο και συνένωση στις στήλες των γονιδίων (Gene\_id == Genes) με τον πίνακα data.with.genes, ο οποίος έχει παραχθεί από τα δεδομένα εισόδου πριν τον εμπλουτισμό και περιέχει δύο στήλες: Gene\_id, tad\_name. Ως αποτέλεσμα, γίνεται γνωστό σε ποιό/ποιά TADs συναντάται κάθε όρος και σε ποια γονίδια οφείλεται αυτό
* έπειτα εκτελείται hypergeometric test for over-represented terms και υπολογισμός p value για κάθε όρο και σε κάθε TAD με τη χρήση της συνάρτησης phyper(hitInSample-1,hitInPop, failInPop,sampleSize, lower.tail=FALSE)
* hitInSample = αριθμός των γονιδίων του TAD που αντιστοιχούν στον όρο (με χρήση της συνάρτησης count()
* hitInPop = ο αριθμός των γονιδίων που αντιστοιχούν στον όρο στη βάση δεδομένων (η τιμή που αντιστοιχεί στη στήλη denominator και υπολογίστηκε πριν από το Overlap)
* failInPop = Ο συνολικός αριθμός των γονιδίων που καλύπτει η βάση – hitInPop
* sampleSize = Ο αριθμός των γονιδίων από τα δεδομένα εισόδου που αντιστοιχούν σε αυτό το TAD
* και τέλος προσαρμογή p values με τη μέθοδο “fdr”.

**Βήματα για την παραγωγή του πίνακα εισόδου pathview.input της συνάρτησης pathVisual**

****

Στιγμιότυπο του πίνακα pathview.input.

Με χρήση της συνάρτησης getKEGGIDs

* η συνάρτηση λαμβάνει ως είσοδο τον πίνακα enriched\_terms, ο οποίος έχει παραχθεί από τη διαδικασία του enrichment (enrich all ή enrich per TAD) και επιλέγει τις στήλες Term, Genes
* έπειτα γίνεται ομαδοποίηση ανά μονοπάτι
* τέλος, χρησιμοποιεί τη συνάρτηση keggFind() του πακέτου KEGGREST, με την οποία κάνει αναζήτηση στην Kegg.db για τα keg ids του κάθε όρου.