3η εργαστηριακή άσκηση Βιοπληροφορικης ΙΙ

Εκπωνήθηκε από τον: Τζοβαλίν Κιαζίμις με ΑΜ: 02401 6ο εξάμηνο

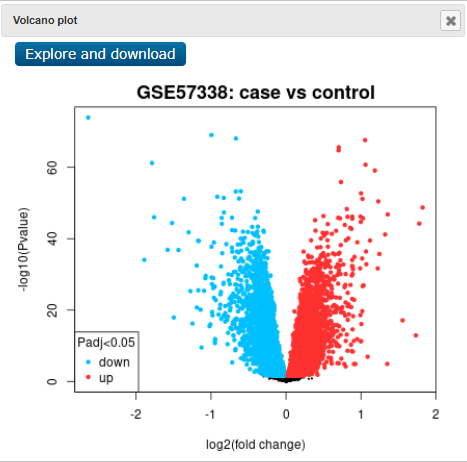
Α) Ανάλυση γονιδιακής έκφρασης με το GEO2R και ανάλυση εμπλουτισμού με το gProfiler

Για την υλοποίηση των συγκεκριμένων εργαλείων g-profiler και GEO2R χρησιμοποίηθηκε η μελέτη case –control με τίτλο:   
  
*RNA-Seq Identifies Novel Myocardial Gene Expression Signatures of Heart Failure [microarray]* και Αccession number: GSE57338 με 313 συνολικά δείγματα

GEO2R tool:Για την ανάλυση των διαφορών στην έκφραση μεταξύ ομάδων ασθενών με και χωρίς καρδιακή ανεπάρκεια για καλύτερη οπτικοποίηση και ανάλυση αποτελεσμάτων.Το εργαλείο επιτρέπει την στατιστική σύγκριση και ταυτοποίηση διαφορετικα εκφραζόμενων γονιδίων.

Συγκεκριμένα έχουμε 177 samples ως cases και τα υπόλοιπα 136 περιλαμβάνονται από υγιείς άτομα.

Αφού χωρίσουμε το Dataset κάνουμε ανάλυση με τις παραμέτρους που υποδείχθηκαν στο εργαστήριο και έχουμε τα εξής

Volcano plot :  


Y –axis (p-value): δείχνει πόσο στατιστικά σημαντική είναι η διαφορά

X-axis (fold hange): δείχνει πόσο αλλάζει η έκφραση ενός γονιδίου

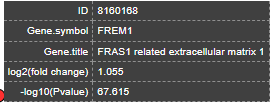
[μπλε με αρνητικές τιμές] : σημαντικά υποεκφρασμένα γονίδια

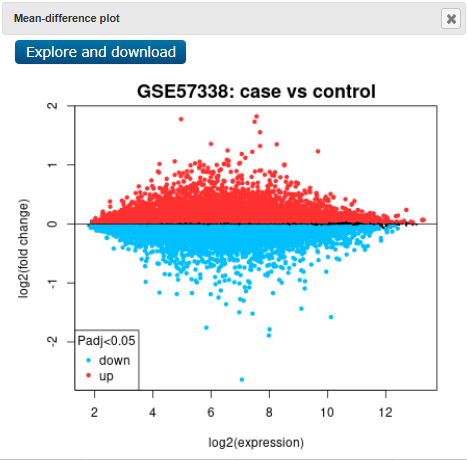
[Κόκκινο με θετικές τιμές]: σημαντικά υπερεκφρασμένα γονίδια   
Όσο πιο ψηλα στο plot άλλο τόσο πιο σημαντικά στατιστικό είναι ένα γονίδιο,άρα και μικρότερα p-value

Στατιστικά πιο σημαντικό υποεκφρασμένο γονίδιο :



Στατιστικά πιο σημαντικό υπερκφρασμένο γονίδιο:

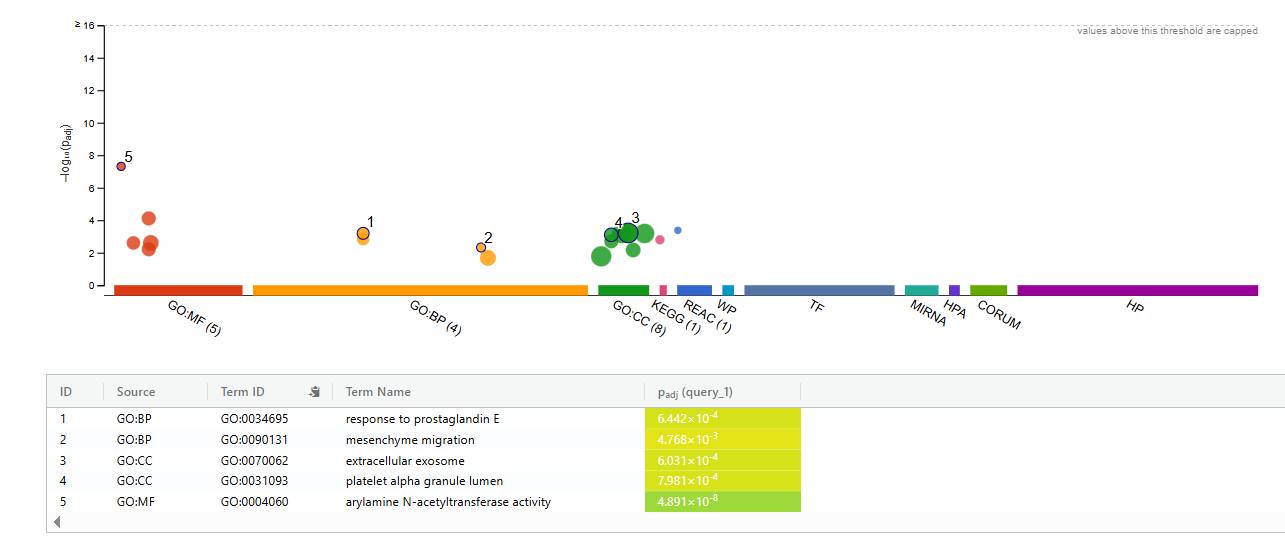


Mean – Difference plot:  


X-axis: δείχνει το μέσο όρο έκφρασης γονιδίου

Υ-axis: δείχνει τη διαφορά στην έκφραση μεταξύ δύο ομάδων

Παίρνουμε τα 100 πιο σημαντικά σταστιστικά γονίδια απο την ανάλυση του GEO2R και κάνουμε ανάλυση εμπλουτισμού χρησιμοποιώντας το εργαλειο g-profiler το οποίο Βοηθά να βρεις τι λειτουργίες, διαδικασίες ή pathways σχετίζονται στατιστικά με τασημαντικά γονίδια

Το πιο σημαντικό στατιστικό γονίδιο είναι αυτό με τα στοιχεια :   
GO:MF | GO:0004060 arylamine N-acetyl transferase activity | p-value:4.891×10-8

Β)GWAS μετα-ανάλυση με το PLINK και ανάλυση εμπλουτισμού με το gProfiler

Κατεβάζουμε τα ετοιμα dataset και εκτελούμε τις παρακάτω εντολές:  
1> #Preprocessing

plink --file toy --make-bed --out toy\_bin

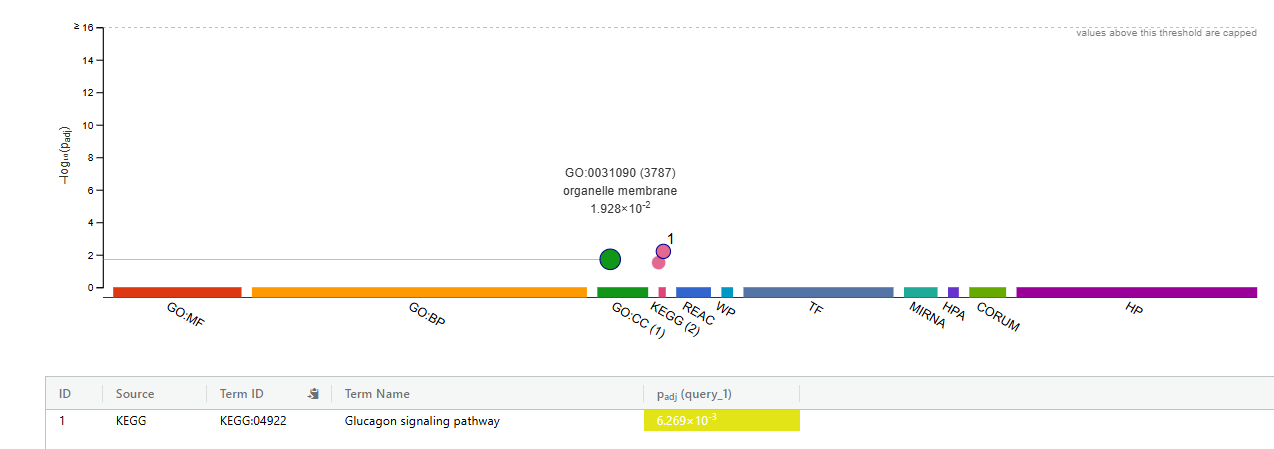
2>#Association Test

plink --bfile toy\_bin --assoc --out assoc\_results

3>#Meta-analysis

plink --meta-analysis DGI\_three\_regions\_PLINK\_adjusted.txt MAGIC\_FUSION\_PLINK\_adjusted.txt

magic\_SARDINIA\_PLINK\_adjusted.txt + logscale no-allele report-al

Εξάγουμε το τελευταίο αρχείο με όνομα plink.meta(περιέχετε και στο github repo) και έχουμε τα εξής :  
με το πιο στατιστικά σημαντικό το:  
KEGG KEGG:04922 Glucagon signaling pathway με p-value 6.269 x 10^(-3)  
  
σημειώσεις γενικές για το τι στοιχεία παίρνουμε από το plink :

| Στήλη | Περιγραφή  
| SNP/ID | Το όνομα του SNP ή marker  
| CHR | Χρωμόσωμα  
| BP | Θέση στο χρωμόσωμα  
| Allele1, Allele2 | Αλληλόμορφα (μπορεί να απουσιάζουν)  
| Effect | Συντελεστής επίδρασης (effect size, π.χ. beta)  
| StdErr | Τυπικό σφάλμα του effect  
| P | P-value για τη συνολική μετα-ανάλυση  
| N | Μέγεθος δείγματος (συνολικά από όλες τις μελέτες)  
| Direction | Κατεύθυνση επίδρασης στα διάφορα datasets (αν υπάρχει)