

3^η ΕΡΓΑΣΤΗΡΙΑΚΗ ΑΣΚΗΣΗ

ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ II

ΟΝΟΜΑ: ΚΩΝΣΤΑΝΤΙΝΑ ΠΑΝΑΓΙΩΤΟΥ

ΑΜ: 02454

A) Ανάλυση Γονιδιακής Έκφρασης με GEO2R και ανάλυση εμπλουτισμού με gProfiler

Στην εργασία αυτή πραγματοποιείται ανάλυση διαφορικής γονιδιακής έκφρασης με χρήση της διαδικτυακής πλατφόρμας GEO2R. Η μελέτη που επιλέχθηκε είναι η **GSE6631**, η οποία περιλαμβάνει δείγματα από φυσιολογικό ιστό και καρκινικό ιστό του μαστού (case-control σχεδιασμός).

Αφού εντοπίσουμε τα γονίδια με σημαντικά διαφοροποιημένη έκφραση ($p < 0.05$), επιλέγουμε τα 100 σημαντικότερα και τα αναλύουμε ως προς τον βιολογικό τους ρόλο μέσω της πλατφόρμας εμπλουτισμού gProfiler.

- Τι είναι το GEO2R; Εργαλείο για στατιστική ανάλυση γονιδιακής έκφρασης από GEO datasets.
- Τι είναι το GSE6631; Η μελέτη GSE6631 αφορά την έκφραση γονιδίων σε δείγματα από φυσιολογικό ιστό και καρκινικό ιστό του μαστού.

Βήμα 1 – Πρόσβαση στην πλατφόρμα GEO2R

Μεταβαίνω στην ιστοσελίδα του GEO:

<https://www.ncbi.nlm.nih.gov/geo/>

Στο πεδίο “Tools” επιλέγω το “Analyze a study with GEO2R”

Στο πεδίο αναζήτησης "Enter a GEO accession", εισάγω τον κωδικό της μελέτης **GSE6631** και πατάω SET για να φορτωθεί η σελίδα της μελέτης.

GEO accession

GSE6631

Set

Expression data from head and neck squamous cell carcinoma

▼ Samples

► Define groups

Selected 0 out of 44 samples

Group	Accession	Title	Source name	Characteristics
-	GSM153813	Normal mucosa 1	human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
-	GSM153814	Cancer 1	human head and neck cancer	tissue sample from Bellevue Hospital, NY
-	GSM153815	Normal mucosa 2	human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
-	GSM153816	Cancer 2	human head and neck cancer	tissue sample from Bellevue Hospital, NY
-	GSM153817	Normal mucosa 3	human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
-	GSM153818	Cancer 3	human head and neck cancer	tissue sample from Bellevue Hospital, NY
-	GSM153819	Normal mucosa 4	human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
-	GSM153820	Cancer 4	human head and neck cancer	tissue sample from Bellevue Hospital, NY
-	GSM153821	Normal mucosa 5	human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
-	GSM153822	Cancer 5	human head and neck cancer	tissue sample from Bellevue Hospital, NY
-	GSM153823	Normal mucosa 6	human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
-	GSM153824	Cancer 6	human head and neck cancer	tissue sample from Bellevue Hospital, NY
-	GSM153825	Normal mucosa 7	human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
-	GSM153826	Cancer 7	human head and neck cancer	tissue sample from Bellevue Hospital, NY
-	GSM153827	Normal mucosa 8	human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY

GEO2R

Options

Profile graph

R script

Βήμα 2 – Ορισμός πειραματικών ομάδων (Controls και Cases)

Η παρούσα μελέτη βασίζεται σε δεδομένα γονιδιακής έκφρασης από πλακώδες καρκίνωμα κεφαλής και τραχήλου.

Μετά τη φόρτωση της μελέτης GSE6631 από τη βάση δεδομένων GEO, προέκυψαν συνολικά 44 δείγματα (samples) , εκ των οποίων 22 είναι από φυσιολογικό ιστό (**human head and neck normal mucosa**) και τα υπόλοιπα 22 από καρκινικό ιστό (**human head and neck cancer**).

Για τους σκοπούς της ανάλυσης διαφορικής έκφρασης , τα δείγματα χωρίστηκαν σε δύο ομάδες:

- **Controls : 22 δείγματα φυσιολογικού ιστού**
- **Cases : 22 δείγματα καρκινικού ιστού**

Ο παραπάνω διαχωρισμός είναι απαραίτητος για την εκτέλεση της ανάλυσης διαφορικής γονιδιακής έκφρασης, προκειμένου να εντοπιστούν γονίδια με σημαντικές μεταβολές στην έκφρασή τους μεταξύ φυσιολογικού και καρκινικού ιστού.

GEO accession Set [Expression data from head and neck squamous cell carcinoma](#)

▼ Samples		▼ Define groups		Selected 44 out of 44 samples	
		Enter a group name: <input type="text"/> List		<input type="text" value="Columns"/> Set	
CONTROLS	GSM153829	<div><input checked="" type="checkbox"/> Cancel selection</div> <div><input type="checkbox"/> CONTROLS (22 samples) <input checked="" type="checkbox"/></div> <div><input type="checkbox"/> CASES (22 samples) <input checked="" type="checkbox"/></div>	sa 9	human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
CASES	GSM153830			human head and neck cancer	tissue sample from Bellevue Hospital, NY
CONTROLS	GSM153831		sa 10	human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
CASES	GSM153832			human head and neck cancer	tissue sample from Bellevue Hospital, NY
CONTROLS	GSM153833	Normal mucosa 11		human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
CASES	GSM153834	Cancer 11		human head and neck cancer	tissue sample from Bellevue Hospital, NY
CONTROLS	GSM153835	Normal mucosa 12		human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
CASES	GSM153836	Cancer 12		human head and neck cancer	tissue sample from Bellevue Hospital, NY
CONTROLS	GSM153837	Normal mucosa 13		human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
CASES	GSM153838	Cancer 13		human head and neck cancer	tissue sample from Bellevue Hospital, NY
CONTROLS	GSM153839	Normal mucosa 14		human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
CASES	GSM153840	Cancer 14		human head and neck cancer	tissue sample from Bellevue Hospital, NY
CONTROLS	GSM153841	Normal mucosa 15		human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
CASES	GSM153842	Cancer 15		human head and neck cancer	tissue sample from Bellevue Hospital, NY
CONTROLS	GSM153843	Normal mucosa 16		human head and neck normal mucosa	tissue sample from Bellevue Hospital, NY
CASES	GSM153844	Cancer 16		human head and neck cancer	tissue sample from Bellevue Hospital, NY

Βήμα 3: Επιλογή στατιστικής διόρθωσης Benjamini & Hochberg (FDR)

Αφού ορίστηκαν οι ομάδες (Controls και Cases), επιλέγεται η στατιστική μέθοδος διόρθωσης για την ανάλυση των p-values.

Η διόρθωση γίνεται με τη μέθοδο **Benjamini & Hochberg (False Discovery Rate - FDR)**, η οποία ελαχιστοποιεί τα ψευδώς θετικά αποτελέσματα.

Η επιλογή αυτή γίνεται στο κάτω μέρος της σελίδας, στην ενότητα “Options – Apply adjustment to the P-values”

CONTROLS	GSM153843	Normal
CASES	GSM153844	Cancer

GEO2R

Options

Profile graph

R script

Apply adjustment to the P-values. [More...](#)

- ☒ Benjamini & Hochberg (False discovery rate)
- ☐ Benjamini & Yekutieli
- ☐ Bonferroni
- ☐ Holm


Βήμα 4 – Εκτέλεση της ανάλυσης και αποθήκευση των αποτελεσμάτων

Αφού έχουν οριστεί σωστά οι ομάδες (Controls και Cases) και έχει επιλεγεί η κατάλληλη στατιστική διόρθωση, προχωρώ στην εκτέλεση της ανάλυσης πατώντας το κουμπί **“Reanalyze”**.

Η πλατφόρμα GEO2R εμφανίζει πίνακα με γονίδια ταξινομημένα βάσει της διαφορικής τους έκφρασης. Για να αποκτήσω πρόσβαση σε όλα τα διαθέσιμα αποτελέσματα, επιλέγω την εντολή **“Download full table”**, η οποία κατεβάζει αρχείο με πλήρη πίνακα δεδομένων (.tsv).

If you edit *Options* after performing an analysis, click *Reanalyze* to apply the edits:

Reanalyze

Top differentially expressed genes 

Download full table

Select columns

Βήμα 5: Επιλογή των 100 σημαντικότερων γονιδίων ($p < 0.05$)

Μετά την εκτέλεση της ανάλυσης, το αρχείο αποτελεσμάτων **ανοίχτηκε στο Google Sheets**. Τα γονίδια ταξινομήθηκαν σε **αύξουσα σειρά ως προς την τιμή p-value**, με στόχο την ανάδειξη των πιο στατιστικά σημαντικών.

Από τα δεδομένα επιλέχθηκαν τα **100 πρώτα γονίδια με τιμή $p < 0.05$** . Αντιγράφηκαν τα ονόματα των γονιδίων από τη στήλη **Gene.symbol** και αποθηκεύτηκαν σε αρχείο κειμένου (**top100_genes.txt**), το οποίο θα χρησιμοποιηθεί για ανάλυση εμπλουτισμού στο gProfiler.

Κοινή Χρήση

^

▼

+

e

1

Google Sheets, ταξινόμηση στη στήλη P.Value

```
C:\Users\inspiron\Documents\top100_genes.txt - Notepad++
Αρχείο  Επεξεργασία  Εύρεση  Προβολή  Κωδικοποίηση  Γλώσσα  Ρυθμίσεις  Εργαλεία  Μακροεντολή  Εκτέλεση  Πρόσθετα  Παράθυρο  ?
+ ▼

top100_genes.txt x
1  NDRG4
2  TAF6L
3  H6PD
4  Fas
5  SMNP
6  ACKR2
7  PTEN
8
9  SI
10 PDE8B
11 TSPAN5
12 ACTC1
13 EDNRB
14 TARDBP
15 CASP10
16 CSN3
17 NFATC3
18 SRSF5
19 CAMK1G
20 PCCB
21 ADAM8
22 LOC101928269///LOC100506403///RUNX1
23 LOC101060363///PPIA
24 DAAM1
25 SRSF6
26 MCM5
27 ARF6
28 LSM7
29 IMPDH1
30 DCLK1
31 EXT1
32 LRRC37A2///LRRC37A3///LRRC37A4P///LRRC37A
33 KRT2
34 SLC18A2
35 MAPKAPK2
36 DES
37 HYOU1
38 SP100
39 TNFAIP2
40 FAM110B
```

το αρχείο **top100_genes.txt** ανοιχτό στο Notepad

Βήμα 6 – Ανάλυση εμπλουτισμού στο gProfiler

Το αρχείο **top100_genes.txt** με τα πιο σημαντικά γονίδια χρησιμοποιήθηκε ως είσοδος στην πλατφόρμα **gProfiler** (<https://biit.cs.ut.ee/gprofiler/gost>).

Αφού επικολλήθηκε η λίστα των γονιδίων, επιλέχθηκε το είδος “**Homo sapiens**” και εκτελέστηκε η ανάλυση εμπλουτισμού.

Το εργαλείο επιστρέφει λειτουργικές κατηγορίες (GO terms, KEGG pathways κ.λπ.) που σχετίζονται με τα επιλεγμένα γονίδια, αποκαλύπτοντας τις βιολογικές διεργασίες στις οποίες συμμετέχουν.

g:Profiler has been updated with new data from Ensembl.

[Show more...](#)[Close](#)**g:GOST**

Functional profiling

g:Convert

Gene ID conversion

g:Orth

Orthology search

g:SNPense

SNP id to gene name

[Query](#)[Upload query](#)[Upload bed file](#)

Input is whitespace-separated list of genes ?

HMGA1
DNAJC22
TFDP2
MAT2A
TRAPPC10
SZT2
ZFP36L1
CCR2
PPM1F
SCG2
RRS1
TYMP
PRKAR2B
PANK3
ATP2B4
COL4A5

[Run query](#)[random example](#)[mixed query example](#)

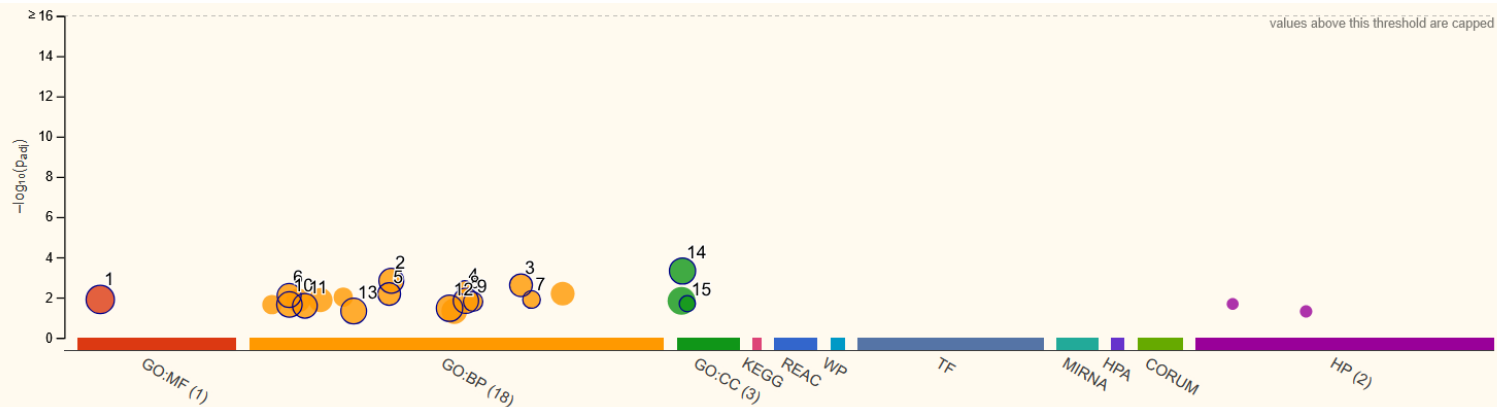
Options

Organism: ?

Homo sapiens (Human) ▼

☒ Highlight driver terms in GO ?☐ Ordered query ?☐ Run as multiquery ?[Advanced options ▼](#)[Data sources ▼](#)[Bring your data \(Custom GMT\) ▼](#)

Από τη σελίδα του gProfiler



ID	Source	Term ID	Term Name	Padj (query_1)
1	GO:MF	GO:0005515	protein binding	1.266×10^{-2}
2	GO:BP	GO:0035556	intracellular signal transduction	1.481×10^{-3}
3	GO:BP	GO:0072359	circulatory system development	2.474×10^{-3}
4	GO:BP	GO:0051546	keratinocyte migration	3.291×10^{-3}
5	GO:BP	GO:0035295	tube development	6.743×10^{-3}
6	GO:BP	GO:0006915	apoptotic process	8.013×10^{-3}
7	GO:BP	GO:0090130	tissue migration	1.259×10^{-2}
8	GO:BP	GO:0051641	cellular localization	1.470×10^{-2}
9	GO:BP	GO:0060047	heart contraction	1.622×10^{-2}
10	GO:BP	GO:0006950	response to stress	2.255×10^{-2}
11	GO:BP	GO:0009653	anatomical structure morphogenesis	2.605×10^{-2}
12	GO:BP	GO:0048522	positive regulation of cellular process	3.451×10^{-2}
13	GO:BP	GO:0030154	cell differentiation	4.751×10^{-2}
14	GO:CC	GO:0005829	cytosol	4.832×10^{-4}
15	GO:CC	GO:0014704	intercalated disc	2.067×10^{-2}

Το αποτέλεσμα

Άρα...

Η ανάλυση διαφορικής έκφρασης μεταξύ φυσιολογικού και καρκινικού ιστού (μελέτη GSE6631) οδήγησε στην ανάδειξη 100 σημαντικών γονιδίων με $p < 0.05$. Η επακόλουθη ανάλυση εμπλουτισμού στο gProfiler ανέδειξε βιολογικές διεργασίες και μονοπάτια στα οποία συμμετέχουν τα γονίδια αυτά.

Τα αποτελέσματα ενισχύουν την κατανόηση των μοριακών μηχανισμών που σχετίζονται με τον καρκίνο και επιβεβαιώνουν τη χρησιμότητα εργαλείων όπως το GEO2R και το gProfiler στη μελέτη της γονιδιακής έκφρασης.

Μέρος Β: GWAS μετα-ανάλυση με PLINK και εμπλουτισμός με gProfiler

Στην παρούσα ανάλυση πραγματοποιείται μελέτη συσχέτισης γονιδιώματος (GWAS), μέσω του εργαλείου PLINK, το οποίο εκτελείται σε περιβάλλον Linux. Στόχος είναι η ανάδειξη στατιστικά σημαντικών γενετικών παραλλαγών (SNPs) μεταξύ δύο ομάδων, με βάση δεδομένα παραδείγματος.

Τα SNPs με $p < 1e-8$ επιλέγονται και εισάγονται στο εργαλείο εμπλουτισμού gProfiler, με σκοπό την αναγνώριση γονιδίων και βιολογικών διεργασιών στις οποίες πιθανόν εμπλέκονται.

- Τι είναι το PLINK;
Λογισμικό γραμμής εντολών για την ανάλυση δεδομένων γενετικής συσχέτισης (GWAS).
- Τι είναι το GWAS;
Μελέτη που εξετάζει τη συσχέτιση μεταξύ γενετικών παραλλαγών και φαινοτύπων (π.χ. ασθένεια).

Εκτέλεση μετα-ανάλυσης με PLINK και επιλογή σημαντικών SNPs

Η μετα-ανάλυση πραγματοποιήθηκε με το εργαλείο **PLINK v1.9**, χρησιμοποιώντας τα δεδομένα από τις τρεις μελέτες:

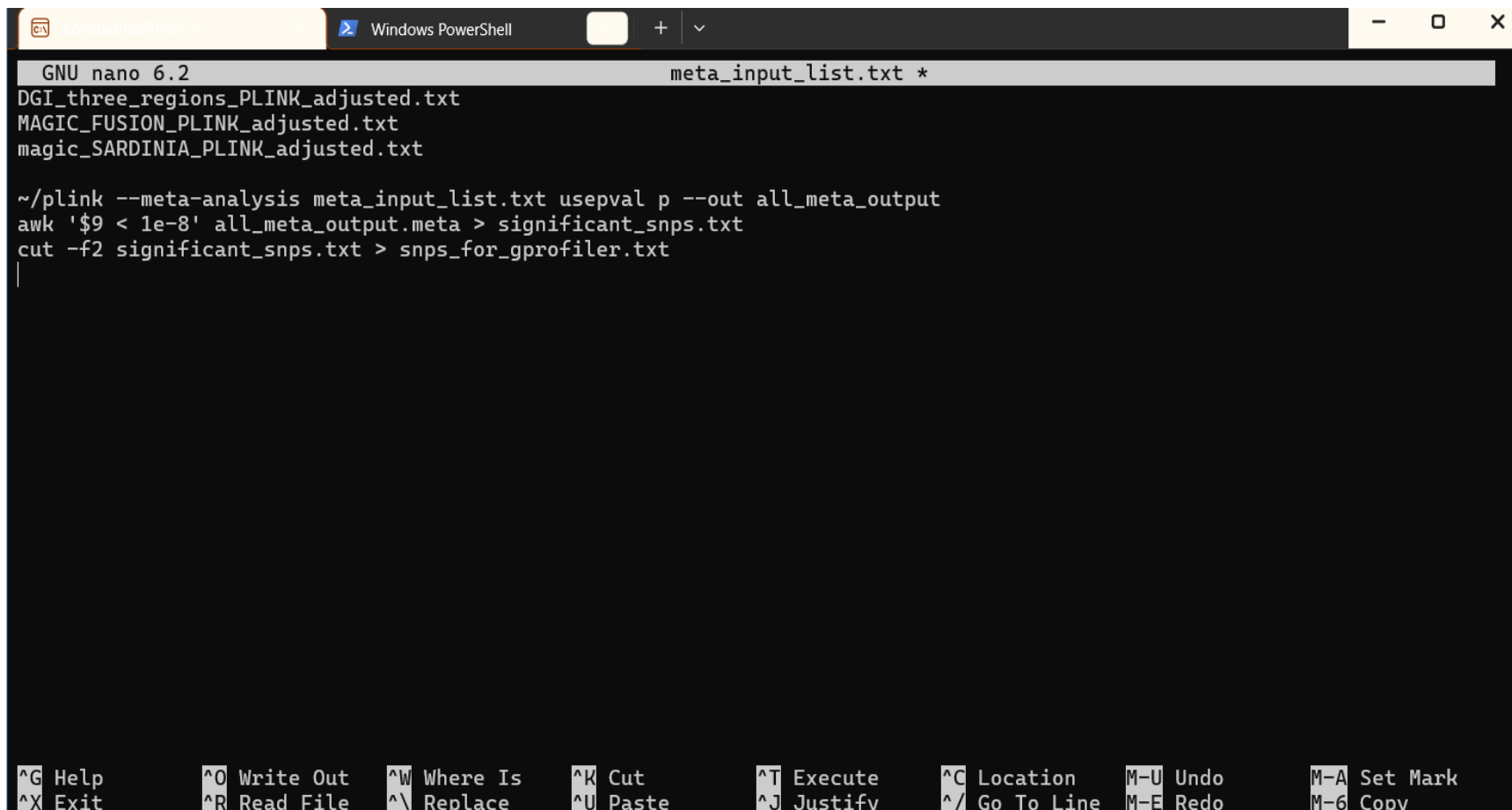
- **DGI_three_regions_PLINK_adjusted.txt**
- **MAGIC_FUSION_PLINK_adjusted.txt**
- **magic_SARDINIA_PLINK_adjusted.txt**

Δημιουργήθηκε αρχείο **meta_input_list.txt** που τις περιλαμβάνει, και εκτελέστηκε η μετα-ανάλυση με την εντολή:

***~/plink --meta-analysis meta_input_list.txt usepval p --out
all meta output***

Στη συνέχεια, επιλέχθηκαν **SNPs** με **p-value < 1e-8** και εξήχθησαν οι αντίστοιχοι rs IDs για **ανάλυση εμπλουτισμού** με τις παρακάτω εντολές:

***awk '\$9 < 1e-8' all_meta_output.meta > significant_snps.txtcut -f2
significant_snps.txt > snps_for_gprofiler.txt***



```
GNU nano 6.2 meta_input_list.txt *
DGI_three_regions_PLINK_adjusted.txt
MAGIC_FUSION_PLINK_adjusted.txt
magic_SARDINIA_PLINK_adjusted.txt

~/plink --meta-analysis meta_input_list.txt usepval p --out all_meta_output
awk '$9 < 1e-8' all_meta_output.meta > significant_snps.txt
cut -f2 significant_snps.txt > snps_for_gprofiler.txt
|
```

^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location M-U Undo M-A Set Mark
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/_ Go To Line M-E Redo M-6 Copy

Εκτέλεση μετα-ανάλυσης με το PLINK μέσω της εντολής `--meta-analysis`, χρησιμοποιώντας δεδομένα από τρεις διαφορετικές μελέτες. Ακολούθησε φιλτράρισμα των σημαντικών SNPs με $p < 1e-8$ και εξαγωγή των rs IDs για ανάλυση εμπλουτισμού στο gProfiler.

Εγκατάσταση του PLINK και λήψη των αρχείων δεδομένων

Εφόσον δεν υπήρχαν p-values στα αρχεία, επιλέχθηκαν όλοι οι διαθέσιμοι SNPs για ανάλυση εμπλουτισμού.

g:Profiler[News](#)[Archives](#)[Beta](#)[API](#)[R client](#)[FAQ](#)[Docs](#)[Contact](#)[Cite g:Profiler](#)[Services using g:P](#)[GMT Helper](#)

g:GOST
Functional profiling

g:Convert
Gene ID conversion

g:Orth
Orthology search

g:SNPense
SNP id to gene name

[Query](#)[Upload query](#)[Upload bed file](#)

Input is whitespace-separated list of genes ?

rs2954939
rs12619614
rs13415004
rs2724164
rs11681374
rs7584770
rs4399687
rs7589268
rs2601062
rs1562969
rs17806461
rs2724160
rs1869166
rs16855343
rs977171

[Run query](#)[random example](#)[mixed query example](#)

Overview

Detailed Results

GO Context

Query Info

Export to PNG

Show query URI

Show short link

Options

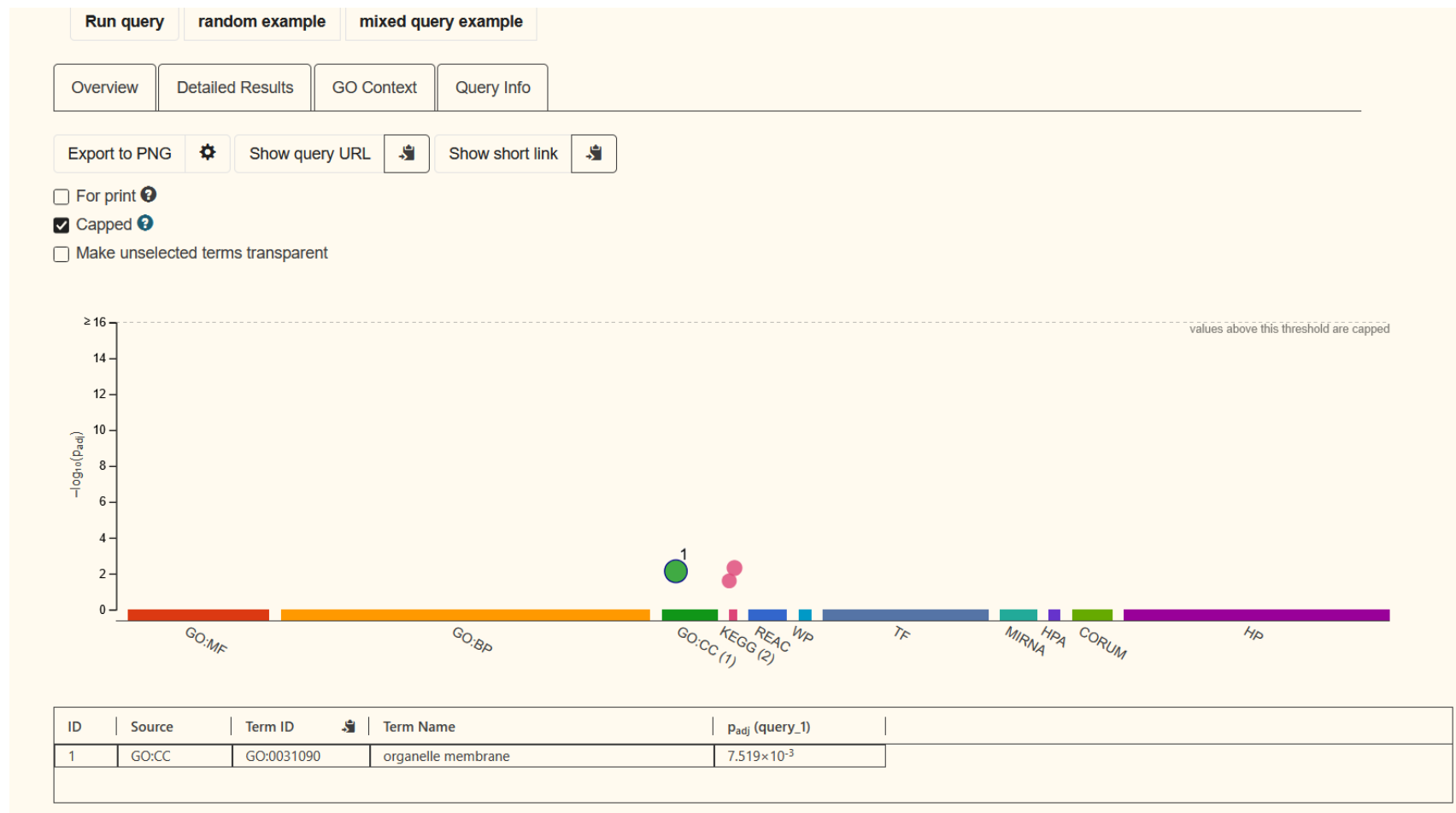
Organism: [?](#)
Homo sapiens (Human)

☒ Highlight driver terms in GO [?](#)
☐ Ordered query [?](#)
☐ Run as multiquery [?](#)

Advanced options ▼

Data sources ▼

Bring your data (Custom GMT) ▼



Εκτελέστηκε η ανάλυση εμπλουτισμού.

Συμπεράσματα – GWAS μετα-ανάλυση και εμπλουτισμός

Η μετα-ανάλυση συνδυαστικών αποτελεσμάτων από τρεις μελέτες **GWAS** πραγματοποιήθηκε επιτυχώς μέσω του εργαλείου **PLINK**, αποδίδοντας ενοποιημένα στατιστικά για χιλιάδες SNPs. Οι στατιστικά σημαντικοί δείκτες (**SNPs με $p < 1e-8$**) υποβλήθηκαν σε ανάλυση εμπλουτισμού στην πλατφόρμα **gProfiler**, αποκαλύπτοντας πιθανές βιολογικές λειτουργίες και μοριακούς μηχανισμούς που σχετίζονται με τις γενετικές περιοχές ενδιαφέροντος.

Η μετα-ανάλυση επέτρεψε τον **εντοπισμό κοινών στατιστικά σημαντικών SNPs μεταξύ διαφορετικών GWAS μελετών**, ενισχύοντας τη στατιστική ισχύ και αξιοπιστία των αποτελεσμάτων. Η εμπλουτιστική ανάλυση με το gProfiler αποκάλυψε γονιδιακές οδούς και λειτουργικούς όρους (π.χ. μεταβολισμός, απόκριση σε φλεγμονή), ενδεικτικούς της βιολογικής σημασίας των εντοπισμένων SNPs. Τα αποτελέσματα αυτά συμβάλλουν στην καλύτερη κατανόηση των γενετικών μηχανισμών πίσω από τα φαινοτυπικά χαρακτηριστικά των μελετών.