ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

Πανεπιστήμιο Κρήτης
Σχολή Θετικών και Τεχνολογικών Επιστημών
Τμήμα Μαθηματικών και Εφαρμοσμένων Μαθηματικών

Μεροληψία φύλου σε κλινικές μελέτες: μια στατιστική προσέγγιση

Φοιτήτρια: Μικέλα Δάφνη Λοΐζου Μάνσκε, tem2206
Επιβλέπων καθηγητής: Παύλος Παυλίδης, Associate Professor, Department of Biology UoC and Affiliated Researcher, ICS-FORTH

Στόχος εργασίας

Εξέταση της **μεροληψίας φύλου** σε κλινικές μελέτες, συγκεκριμένα μέσω της αναζήτησης **διαφορικά εκφραζόμενων γονιδίων** μεταξύ φύλων

Μεροληψία φύλου σε κλινικές μελέτες

Η **μεροληψία φύλου** (gender bias) σε κλινικές μελέτες αναφέρεται στην **υποεκπροσώπηση** ή στον **αποκλεισμό** των γυναικών από κλινικές μελέτες.

Δύο βασικά επιχειρήματα:

1. Το γυναικείο σώμα είναι σχεδόν ίδιο με το ανδρικό
2. Ο εμμηνορρυσιακός κύκλος και οι εγκυμοσύνη αυξάνουν την πολυπλοκότητα της ερευνητικής διαδικασίας

Παραδείγματα υποεκπροσώπησης γυναικών σε έρευνες

Αμερικάνικη έρευνα για τον HIV: Γυναίκες αποτελούν το **19,2%** των συμμετεχόντων σε αντιρετροϊκές μελέτες, το **38,1%** σε μελέτες εμβολιασμού και το **11,1%** σε μελέτες για την εξεύρεση θεραπείας.

Καρδιαγγειακές παθήσεις: Οι πρώτες έρευνες διεξήχθησαν μόνο σε άντρες, ενώ οι γυναίκες αποτελούν το **25%** των συμμετεχόντων σε 31 μελέτες-ορόσημα για τη συμφορητική καρδιακή ανεπάρκεια.

Μελέτες σε ζώα: **22%** των μελετών σε ζώα δεν προσδιορίζουν το φύλο, ενώ όσες το κάνουν, περιλαμβάνουν κατά **80%** μόνο αρσενικά.

Συνέπειες μεροληψίας φύλου

Αγνοούνται σημαντικές βιολογικές διαφορές ανάμεσα στα φύλα, που επηρεάζουν τον **μεταβολισμό φαρμάκων**, την **καρδιακή λειτουργία**, την **ανοσοποιητική απόκριση** και την **εκδήλωση συμπτωμάτων**. Ως συνέπειες έχουμε:

- Αναποτελεσματικές ή επικίνδυνες θεραπείες
- Ανεπιθύμητες ενέργειες φαρμάκων (ADR)
- Καθυστερημένες ή λανθασμένες διαγνώσεις
- Ακατάλληλες ιατρικές οδηγίες ή συσκευές

Ανάλυση διαφορικής γονιδιακής έκφρασης και μεροληψία φύλου

Οι μελέτες **γονιδιακής έκφρασης** παρέχουν σημαντικές πληροφορίες σε μοριακό επίπεδο σχετικά με το πώς οι ασθένειες μπορεί να εκδηλώνονται με **διαφορετικό τρόπο σε άνδρες και γυναίκες.**

Με την εφαρμογή **στατιστικής ανάλυσης** σε σύνολα δεδομένων γονιδιακής έκφρασης, γίνεται δυνατός ο εντοπισμός γονιδίων που εκφράζονται **διαφορετικά** μεταξύ αρσενικών και θηλυκών δειγμάτων.

Γονίδια

Τα **γονίδια** (Genes) είναι αλληλουχίες νουκλεοτιδίων που βρίσκονται στα χρωμοσώματα, τα οποία είναι νηματοειδείς δομές που σχηματίζονται από DNA.

Η **γονιδιακή έκφραση** (Gene expression) είναι η διαδικασία με την οποία οι πληροφορίες ενός γονιδίου χρησιμοποιούνται στη σύνθεση ενός λειτουργικού γονιδιακού προϊόντος, όπως πρωτεΐνες ή μη κωδικοποιητικό RNA.

Τα **επίπεδα γονιδιακής έκφρασης** (Gene expression levels) ποσοτικοποιούν το πόσο εκφράζεται ένα γονίδιο ... μεταξύ δύο πειραματικών συνθηκών, ... γονίδια ... διαφορικά εκφραζόμενα (**DEGs**) ...

Μέτρηση και αναπαράσταση επιπέδων γονιδιακής έκφρασης

- Τεχνολογία μέτρησης γονιδιακής έκφρασης: **Microarrays**

- Αναπαράσταση επιπέδων γονιδιακής έκφρασης: **nxm πίνακας**.

  n γραμμές ~ 10.000-40.000 **γονίδια**

  m στήλες ~ 10–40 **δείγματα**

- Κάθε εγγραφή του πίνακα αντιπροσωπεύει το **επίπεδο έκφρασης** ενός συγκεκριμένου γονιδίου για ένα συγκεκριμένο δείγμα

Σύνολα Δεδομένων

Αναλύθηκαν **30** σύνολα δεδομένων microarray γονιδιακής έκφρασης διαφόρων ασθενειών και παθήσεων από τη βάση δεδομένων Gene Expression Omnibus. (**GEO**)

Συχνά περιλαμβάνουν πολλαπλές πειραματικές συνθήκες που αντιπροσωπεύονται απο υποσύνολα για τα δείγματα (**factors**), για παράδειγμα φύλο, κατάσταση ασθένειας, ηλικία κλπ.

Έλεγχος υποθέσεων

Βασικός στόχος είναι να προσδιοριστεί εάν υπάρχει **στατιστικά σημαντική επίδραση του φύλου** στην διαφορική γονιδιακή έκφραση μεταξύ υγιούς και ασθενούς κατάστασης

Μηδενική υπόθεση H_0 :

Το φύλο **δεν** έχει σημαντική επίδραση στη διαφορική γονιδιακή έκφραση μεταξύ υγιούς και ασθενούς κατάστασης

Εναλλακτική υπόθεση H_1 :

Το φύλο **έχει** σημαντική επίδραση στη διαφορική γονιδιακή έκφραση μεταξύ υγιούς και ασθενούς κατάστασης

Το κριτήριο για τη στατιστική σημαντικότητα ορίζεται σε όριο **p-value < 0.05**

Limma (Linear models for microarray data)

Αρχικά χρησιμοποιήθηκε το **t-test**, όμως επειδή δεν επαρκεί για μεγάλες και σύνθετες μελέτες τελικά χρησιμοποιήθηκε η βιβλιοθήκη της limma

Η Limma είναι μια βιβλιοθήκη της R για την **ανάλυση διαφορικής γονιδιακής έκφρασης**

Χρησιμοποιεί **γραμμικά μοντέλα** και την εμπειρική Bayes προσέγγιση για την εκτίμηση της διακύμανσης, βελτιώνοντας έτσι την ισχύ της στατιστικής ανάλυσης

Empirical Bayes: Η εμπειρική Bayes προσέγγιση εκτιμά την εκ των προτέρων κατανομή από τα παρατηρούμενα δεδομένα

Ορολογία

- **Παράγοντες (factors):** Κατηγορικές **μεταβλητές** που περιγράφουν τις **ομάδες των δειγμάτων**. *πχ. φύλο, κατάσταση ασθένειας*

- **Επίπεδα (levels):** Δυνατές **κατηγορίες** (τιμές) που μπορεί να πάρει ένας παράγοντας. *πχ. για τον παράγοντα φύλο: γυναίκα, άνδρας*

- **Παράμετροι (parameters): Συντελεστές** γραμμικού μοντέλου, οι οποίοι εκφράζουν τη **μέση γονιδιακή έκφραση** για κάθε επίπεδο. *πχ. β1: έκφραση για γυναίκες, β2: έκφραση για άνδρες*

- **Αντίθεση (contrast):** Η μαθηματική **διαφορά παραμέτρων** που θέλουμε να συγκρίνουμε στατιστικά. *πχ. β1 - β2*

Limma (Linear models for microarray data)

Για **κάθε** γονίδιο, η limma εφαρμόζει το γραμμικό μοντέλο:

$$Y = X\beta + \varepsilon$$

όπου:

Y: Διάνυσμα με τις **τιμές έκφρασης** του γονιδίου σε όλα τα δείγματα

X: Design matrix - περιγράφει σε ποιά ομάδα (**επίπεδο/level**) ανήκει κάθε δείγμα

β: Παράμετροι του μοντέλου - εκφράζουν τις εκτιμώμενες μέσες εκφράσεις (ή τις διαφορές μεταξύ ομάδων)

ε: Το **σφάλμα** - η απόκλιση κάθε παρατήρησης από την τιμή που προβλέπει το μοντέλο

Ένα ξεχωριστό γραμμικό μοντέλο εφαρμόζεται σε κάθε γονίδιο, εκτιμώντας τις **παραμέτρους β** και τη διακύμανση της έκφρασης για κάθε **επίπεδο ή συνδυασμό επιπέδων** των **παραγόντων** του μοντέλου. Η διακύμανση **σταθεροποιείται** μέσω της εμπειρικής μεθόδου Bayes (**empirical Bayes**) και εκτελούνται έλεγχοι υποθέσεων για τους παραμέτρους **β**. Τα p-values που προκύπτουν προσαρμόζονται για πολλαπλούς ελέγχους με τη μέθοδο FDR (**False Discovery Rate**).

Means Model

Μοντελοποίηση / εκτίμηση της μέσης γονιδιακής έκφρασης για κάθε **επίπεδο (level)** ενός **παράγοντα (factor)**

$$expression = \beta1 * \texttt{level1} + \beta2 * \texttt{level2}$$

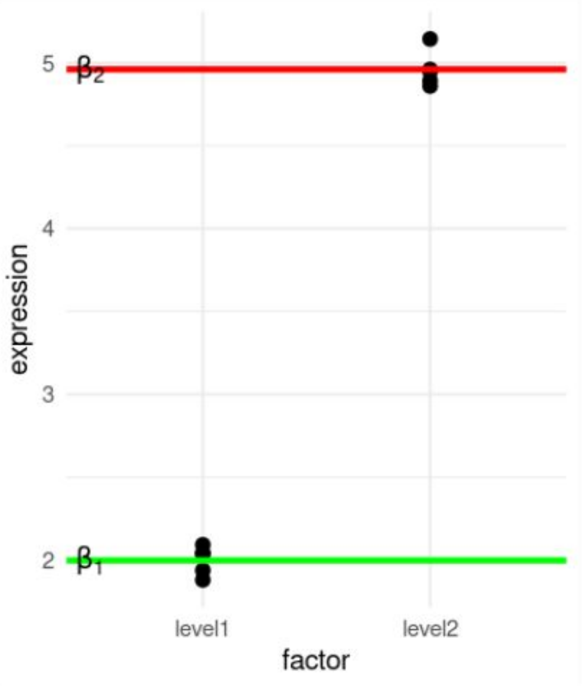
$\beta1$: παρατηρούμενη μέση γονιδιακή έκφραση για `level1`

$\beta2$: παρατηρούμενη μέση γονιδιακή έκφραση για `level2`

Τα levels είναι δυαδικά (0/1) και αμοιβαία αποκλειόμενα

Σε περίπτωση **πολλαπλών παραγόντων**, το μοντέλο εκτιμά ξεχωριστή μέση έκφραση για κάθε **συνδυασμό επιπέδων**

Το **mean-reference** model εκτιμά κατευθείαν την **διαφορά** μεταξύ επιπέδων



Design & Contrast Matrices

Design matrix: Ορίζει τα **επίπεδα** παραγόντων (levels) που συμπεριλαμβάνονται

- Στήλες: **επίπεδο** (level) ή **συνδυασμός επιπέδων**

Contrast matrix: Χρησιμοποιείται για τον προσδιορισμό των **επιπέδων**

- Στήλες: **επίπεδο** (level) ή **συνδυασμός επιπέδων**

	Sample	B1	B2
1	1	1	0
2	2	1	0
3	3	1	0
4	4	0	1
5	5	0	1
6	6	0	1

	Parameter	C1	C2
1	B1	1	-1
2	B2	-1	1

Εφαρμογή: Limma

Η **ανάλυση των δεδομένων** έγινε με την χρήση της Limma στην R, χρησιμοποιώντας τα μοντέλα means-model και mean-reference model για την επαλήθευση της εγκυρότητας της εφαρμογής.

- Δημιουργία **παραγόντων (factors)**
- Δημιουργία **design matrix**
- Δημιουργία **αντιθέσεων (contrasts)**
- Εφαρμογή γραμμικών μοντέλων με χρήση **lmFit** και **contrasts.fit**
- Διόρθωση αποτελεσμάτων με μέθοδο **eBayes**
- Έλεγχος για DEGs με κριτήριο **p-value < 0.05**
- Εξαγωγή **gene IDs** των γονιδίων που πληρούν το παραπάνω κριτήριο

Εφαρμογή: ερμηνεία και διόρθωση αποτελεσμάτων

- Αρχικά, τα DEGs που βρέθηκαν εξέφραζαν κυρίως **έμφυλες διαφορές**, αντί για τις διαφορές μεταξύ ασθενών και υγιών
- Έγινε επανάληψη της διαδικασίας αυτή τη φορά ελέγχοντας για DEGs μεταξύ της **υγιούς** και **φρενώους** κατάστασης
- Η τελική ανάλυση έγινε στην **τομή** των αποτελεσμάτων των δύο εφαρμογών

Αποτελέσματα

- Στο **63,3%** (19/30) των συνόλων δεδομένων δεν βρέθηκαν DEGs μεταξύ φύλων
  Η H_0 δεν μπορεί να απορριφθεί
- Στο **36,7%** (11/30) των συνόλων δεδομένων βρέθηκαν DEGs μεταξύ φύλων
  Η H_0 απορρίπτεται, παρέχοντας αποδείξεις για την H_1

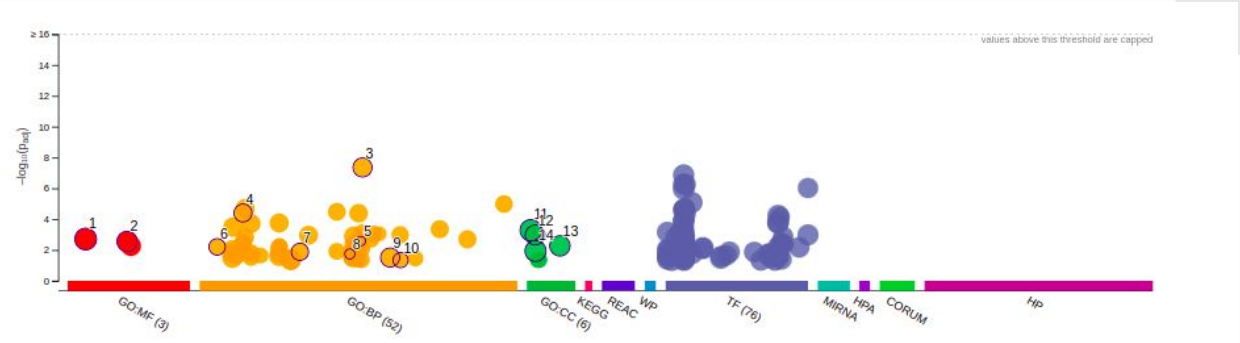
Αποτελέσματα

Συνολικά βρέθηκαν **677 διαφορικά εκφραζόμενα γονίδια** μεταξύ φύλων και μεταξύ της υγιής κατάστασης και της ασθένειας

11 από αυτά (**1,6%**) εμφανίστηκαν σε παραπάνω από ένα σύνολο δεδομένων

Η λίστα των DEGs αναλύθηκε με την βοήθεια **gProfiler** (εργαλείο γονιδιακού προφίλ) και **τεχνητής νοημοσύνης**

Αποτελέσματα



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0005515	protein binding	1.949×10^{-3}
2	GO:MF	GO:0036094	small molecule binding	2.780×10^{-3}
3	GO:BP	GO:0051239	regulation of multicellular organismal process	4.410×10^{-6}
4	GO:BP	GO:0009719	response to endogenous stimulus	3.989×10^{-5}
5	GO:BP	GO:0051045	negative regulation of membrane protein ectod...	2.658×10^{-4}
6	GO:BP	GO:0003013	circulatory system process	6.283×10^{-4}
7	GO:BP	GO:0033993	response to lipid	1.314×10^{-3}
8	GO:BP	GO:0048103	somatic stem cell division	1.809×10^{-3}
9	GO:BP	GO:0065008	regulation of biological quality	3.081×10^{-3}
10	GO:BP	GO:0071559	response to transforming growth factor beta	4.582×10^{-3}
11	GO:CC	GO:0005737	cytoplasm	5.274×10^{-4}
12	GO:CC	GO:0012505	endomembrane system	1.067×10^{-3}
13	GO:CC	GO:0071944	cell periphery	5.408×10^{-3}
14	GO:CC	GO:0016020	membrane	1.155×10^{-2}

version	e112_eg59_p19_25aa4782
date	4/7/2025, 6:34:25 PM
organism	hsapiens

g:Profiler

Στιγμιότυπο οθόνης από την ανάλυση της λίστας των DEGs μέσω της διαδικτυακής πλατφόρμας g:Profiler.

Αποτελέσματα

Ανάλυση με χρήση τεχνητής νοημοσύνης:

Τα σημαντικότερα αποτελέσματα εμπλουτισμού λειτουργιών (GO terms) έδειξαν:

- **Μοριακές λειτουργίες (MF):** σύνδεση με πρωτεΐνες και μικρά μόρια
- **Βιολογικές διεργασίες (BP):** απόκριση σε ενδογενή ερεθίσματα, λιπίδια και TGF-β· διαίρεση σωματικών βλαστικών κυττάρων
- **Κυτταρικά συστατικά (CC):** κυτταρόπλασμα, μεμβράνη και ενδομεμβρανικό σύστημα

Τα γονίδια αυτά φαίνεται να συμμετέχουν σε ορμονικά ρυθμιζόμενες διεργασίες και παθοφυσιολογικούς μηχανισμούς που **διαφέρουν ανά φύλο**, ενισχύοντας την υπόθεση της **φύλο-εξαρτώμενης γονιδιακής ρύθμισης**.

Συμπεράσματα / Συζήτηση

Παρότι τα περισσότερα σύνολα δεδομένων **δεν έδειξαν σημαντικές διαφορές**, εντοπίστηκαν **έμφυλες αποκλίσεις** στη γονιδιακή έκφραση σε αρκετές περιπτώσεις.

Ένδειξη της συστηματικής **απουσίας δεδομένων** για το φύλο αποτέλεσε η αδυναμία εύρεσης συνόλων δεδομένων στο GEO τα οποία έχουν το φύλο ως παράγοντα (**0.241%**, 74 από τα 30702).

Το πρόβλημα της μεροληψίας στην έρευνα παρατηρείται και σε:

- Μη λευκούς πληθυσμούς
- Ηλικιωμένους
- Trans και intersex άτομα

Φαίνεται η αναγκαιότητα **συμπεριληπτικής έρευνας** για μια πιο δίκαιη, αξιόπιστη και αποτελεσματική υγειονομική περίθαλψη για όλους.

Ερωτήσεις;