ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

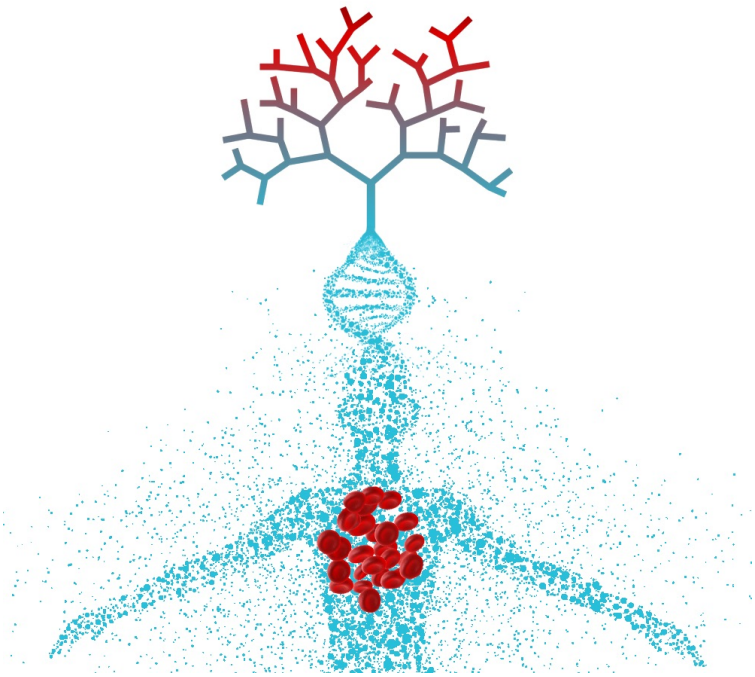
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό
Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

«Ανάλυση αποτελεσμάτων Omics με στόχο τον προσδιορισμό βιοδεικτών»



Γεώργιος Σέντης

Πτυχιούχος Τμήματος Βιολογίας, Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης

ΑΘΗΝΑ 2019

HELLENIC REPUBLIC
National and Kapodistrian
University of Athens
EST. 1837

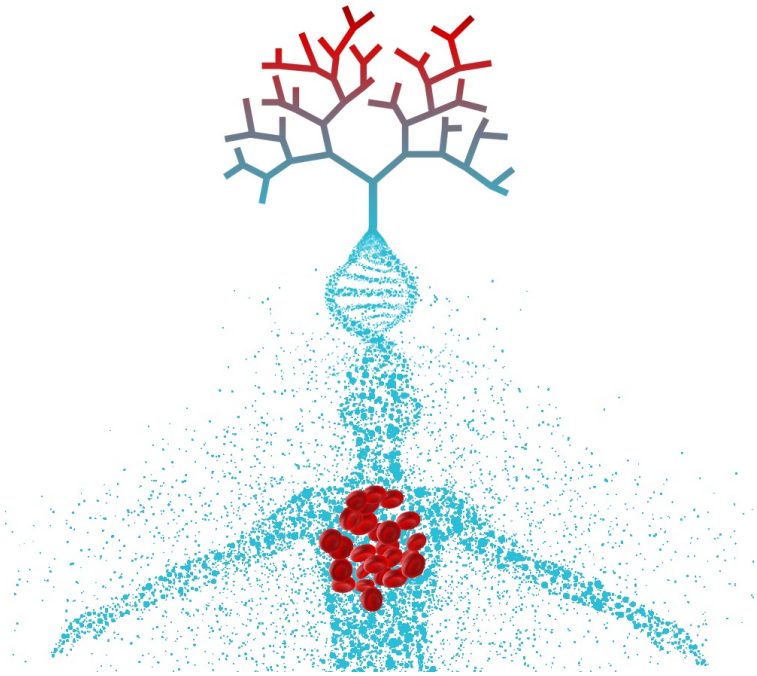
HELLENIC REPUBLIC
National and Kapodistrian
University of Athens

SCHOOL OF SCIENCE
DEPARTMENT OF BIOLOGY

MASTER IN «BIOINFORMATICS»

Master Diploma Thesis

«Omics data analysis for biomarkers identification»



GEORGIOS SENTIS

B.Sc. Biology, Aristotle University of Thessaloniki

A T H E N S 2 0 1 9

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

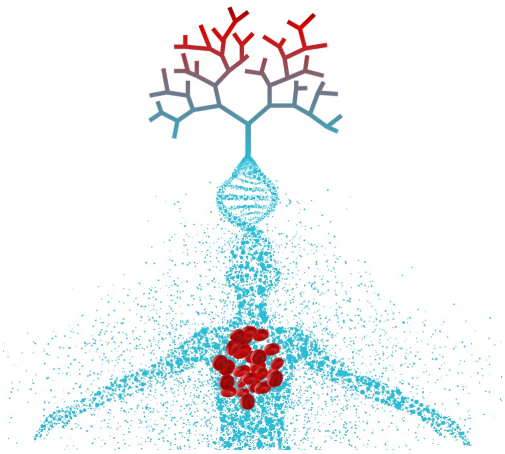
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό
Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

«Ανάλυση αποτελεσμάτων Omics με στόχο τον προσδιορισμό βιοδεικτών»



Τριμελής εξεταστική επιτροπή

Δρ. Βασιλική Α. Οικονομίδου, Επίκουρη Καθηγήτρια (Επιβλέπουσα)
Τομέας Βιολογίας Κυττάρου & Βιοφυσικής,
Τμήμα Βιολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Δρ. Δημήτριος Ι. Στραβοπόδης, Αναπληρωτής Καθηγητής
Τομέας Βιολογίας Κυττάρου & Βιοφυσικής,
Τμήμα Βιολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Δρ. Βασίλειος Κουβέλης, Επίκουρος Καθηγητής
Τομέας Γενετικής & Βιοτεχνολογίας,
Τμήμα Βιολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Επιστημονική υπεύθυνη
Δρ. Ελένη Κατσαντώνη, Ερευνήτρια Γ΄
Κέντρο Βασικής Έρευνας,
Ίδρυμα Ιατροβιολογικών Ερευνών Ακαδημίας Αθηνών

Πρόλογος – Ευχαριστίες

Η παρούσα εργασία εκπονήθηκε στο Εργαστήριο της Κας Κατσαντώνη στο Κέντρο Βασικής Έρευνας του Ιδρύματος Ιατροβιολογικών Ερευνών, Ακαδημίας Αθηνών υπό την επίβλεψη και καθοδήγηση της Κας Κατσαντώνη.

Θα ήθελα να ευχαριστήσω την Κα Κατσαντώνη για την ευκαιρία που μου έδωσε να δουλέψω στο εργαστήριό της και για την καθοδήγησή της καθ' όλη τη διάρκεια της εκπόνησης της εργασίας. Θα ήθελα επίσης να ευχαριστήσω τα μέλη του Εργαστηρίου Μάρκο Φουντουλάκη και Αιμιλία Καφαλίδου για τη βοήθειά τους, όλους τους φοιτητές και μεταδιδακτορικούς ερευνητές που μοιραζόμασταν τον ίδιο χώρο για το ευχάριστο κλίμα εργασίας που δημιούργησαν και τη μεταδιδακτορική ερευνήτρια Αικατερίνη Νάνου για τις πολύτιμες συμβουλές της. Στη συνέχεια, θα ήθελα να ευχαριστήσω την επιβλέπουσα Κα Οικονομίδου για το συντονισμό της εργασίας και, μαζί με τα άλλα δύο μέλη της τριμελούς επιτροπής Κο Κουβέλη και Κο Στραβοπόδη, για την κριτική ανάγνωση της εργασίας μου.

Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου για τη στήριξή της όλα αυτά τα χρόνια και κυρίως κατά τη διάρκεια των μεταπτυχιακών μου σπουδών.

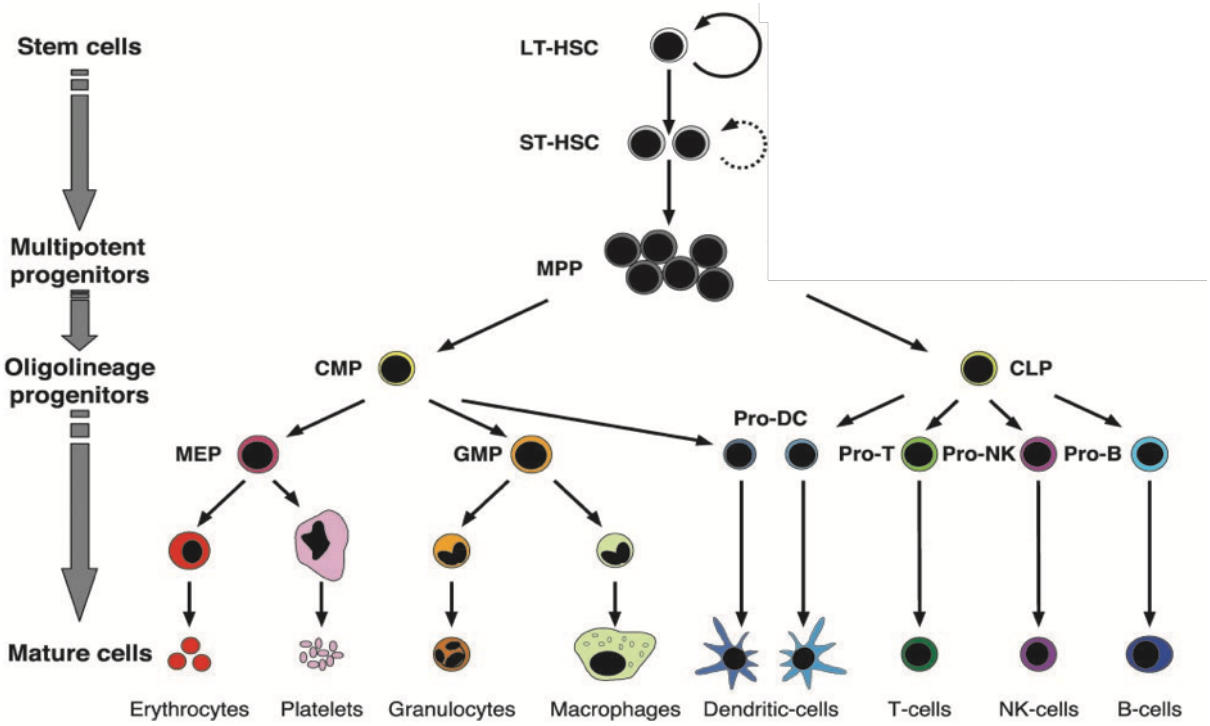
ΠΕΡΙΕΧΟΜΕΝΑ

Α. Εισαγωγή

1. Ερυθροποίηση

Η ερυθροποίηση είναι μια αυστηρά ελεγχόμενη διαδικασία που περιλαμβάνει τη διαφοροποίηση των αιμοποιητικών βλαστικών κυττάρων, τα οποία βρίσκονται στο μυελό των οστών, σε ώριμα ερυθροκύτταρα, τα οποία μεταφέρουν οξυγόνο στους ιστούς, με σκοπό τη διατήρηση της ομοιόστασης των επιπέδων οξυγόνου (O_2) στο σώμα (Elliott et al., 2008; Testa, 2004; Zivot et al., 2018). Μέσω της ερυθροποίησης παράγονται περίπου 200 δισεκατομμύρια ερυθροκύτταρα καθημερινά, ενώ η παραγωγή αυτή μπορεί να αυξηθεί σημαντικά ανάλογα με τις ανάγκες του οργανισμού (Valent et al., 2018).

Η συνεχής παραγωγή κυττάρων του αίματος διασφαλίζεται με την παρουσία ολοδύναμων αιμοποιητικών βλαστικών κυττάρων (HSCs), τα οποία έχουν την ικανότητα αυτοανανέωσης και διαφοροποίησης (Testa, 2004). Αρχικά κατά την ερυθροποίηση, από τα HSCs προκύπτουν δύο μεγάλες κατηγορίες προγονικών κυττάρων, τα κοινά μυελικά προγονικά κύτταρα (CMPs) και τα κοινά λεμφικά προγονικά κύτταρα (CLPs). Από τα CMPs προκύπτουν οι πληθυσμοί των μεγακαρυωτικών/ερυθροκυτταρικών προγονικών κυττάρων

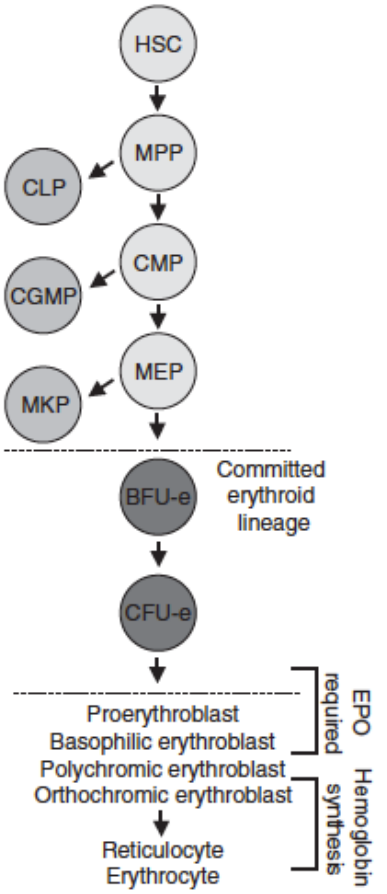


Εικόνα 1. *Αιμοποιητικά κύτταρα και προγονικές κυτταρικές σειρές διαφοροποίησης (από Passegue et al., 2003)*

(MEPs) και των κοινών προγονικών μυελικών κοκκιοκυττάρων (GMPs) (Εικόνα 1). Στη συνέχεια, ο πληθυσμός των MEPs μπορεί να δώσει μετά τη διαφοροποίησή του άλλες δύο

κυτταρικές κατηγορίες, είτε τους μεγακαρυοβλάστες (MKPs) είτε τα ερυθροποιητικά προγονικά κύτταρα. Από τους μεγακαρυοβλάστες προκύπτουν τα αιμοπετάλια, ενώ από τα ερυθροποιητικά προγονικά κύτταρα προκύπτουν, έπειτα από μερικά επιπρόσθετα στάδια διαφοροποίησης, τα ώριμα ερυθροκύτταρα (Singh, 2018, Εικόνα 2).

Στην πρώτη καθορισμένη ερυθροποιητική προγονική κυτταρική σειρά διαφοροποίησης ανήκουν κύτταρα τα οποία μπορούν να αναπτύξουν σε κυτταροκαλλιέργειες μονάδες «εκρηκτικής αύξησης» ερυθροειδών κυττάρων BFU-E (Burst-forming units – erythroid). Όταν ένα κύτταρο BFU-E αναπτυχθεί σε θρεπτικό το οποίο περιέχει ερυθροποιητίνη, μπορεί να δώσει περίπου 500 ώριμα ερυθροκύτταρα σε 6-10 ημέρες. Τα κύτταρα BFU-E διαφοροποιούνται σε έναν άλλο τύπο κυττάρων που ονομάζονται CFU-E (Colony-forming units – erythroid), τα οποία με τη σειρά τους διαφοροποιούνται σε κλασσικούς ερυθροβλάστες (Singh, 2018; Zivot et al., 2018).



Εικόνα 2. *Ερυθροποίηση. Σχηματική απεικόνιση της διαδοχής των σειρών διαφοροποίησης για τον σχηματισμό των ώριμων ερυθροκυττάρων. Επισημαίνονται τα στάδια όπου είναι απαραίτητη η ερυθροποιητίνη καθώς και τα στάδια στα οποία γίνεται η σύνδεση αιμοσφαιρίνης (από Singh, 2018).*

Τα στάδια των ερυθροβλαστών ακολουθούν την εξής σειρά: προερυθροβλάστης, βασεόφιλος ερυθροβλάστης, πολυχρωματικός ερυθροβλάστης και ορθοχρωματικός ερυθροβλάστης. Για τη μετάβαση στο στάδιο του προερυθροβλάστη καθώς και του βασεόφιλου ερυθροβλάστη είναι απαραίτητη η ερυθροποιητίνη (Singh, 2018). Η φάση του ερυθροβλάστη περιλαμβάνει τη σταδιακή συσσώρευση αιμοσφαιρίνης, τη μείωση του κυτταρικού μεγέθους και τη συμπύκνωση του πυρήνα που έχει ως τελικό αποτέλεσμα την αποβολή του και τη διαφοροποίηση των ερυθροβλαστών σε δικτυοερυθροκύτταρα (Zivot et al., 2018).

Η τελευταία φάση της ερυθροποίησης περιλαμβάνει την ωρίμανση των δικτυοερυθροκυττάρων σε ώριμα ερυθροκύτταρα. Αυτή λαμβάνει χώρα στα ερυθροβλαστικά νησίδια του μυελού των οστών, όπου 1 ή 2 μακροφάγα κύτταρα περιβάλλονται από μέχρι και 30 κύτταρα του ερυθροποιητικού συστήματος σε διάφορες φάσης της ερυθροποίησης (από το στάδιο του κυττάρου CFU-E έως και του ερυθροβλάστη). Η σύνδεση αυτή με τα μακροφάγα δημιουργεί ένα περιβάλλον επικοινωνίας στο οποίο λαμβάνουν χώρα όλες οι

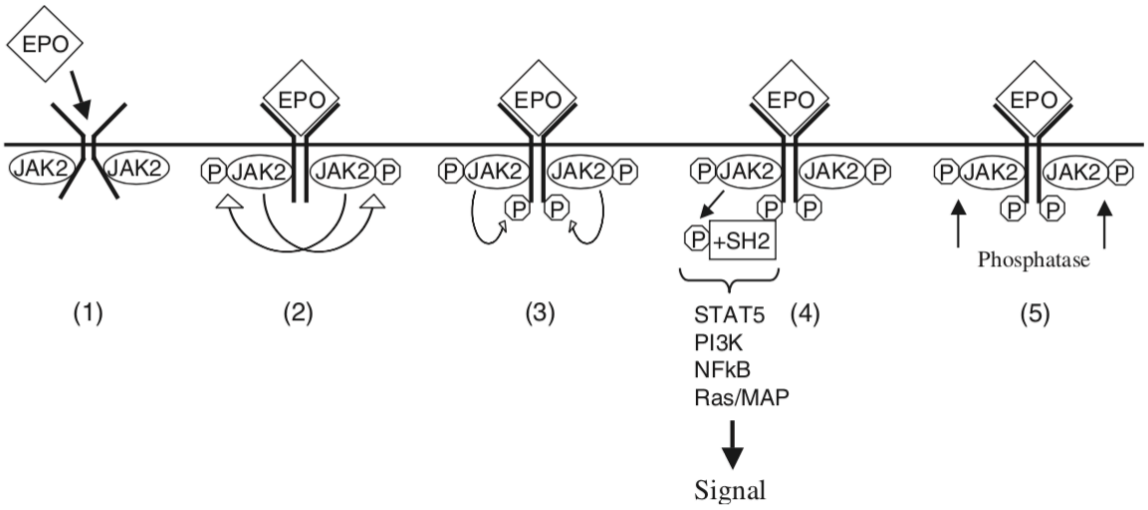
αλληλεπιδράσεις που είναι απαραίτητες για την διαφοροποίηση και τον πολλαπλασιασμό των κυττάρων. Τα διαφοροποιημένα ώριμα ερυθροκύτταρα απελευθερώνονται στο περιφερικό αίμα (Valent et al., 2018; Zivot et al., 2018).

1.1 Μεταγραφικοί παράγοντες στην ερυθροποίηση

Στη διαδικασία της ερυθροποίησης εμπλέκονται πολλοί μεταγραφικοί παράγοντες οι οποίοι ρυθμίζουν την έκφραση των γονιδίων στόχων τους. Ένας από αυτούς είναι ο επαγόμενος από την υποξία παράγοντας HIF, ο οποίος ανήκει στην οικογένεια μεταγραφικών παραγόντων έλικας-στροφής-έλικας (basic helix-loop-helix). Προσδένεται σε ειδικές αλληλουχίες του DNA που αναφέρονται ως στοιχεία απόκρισης στην υποξία (Hypoxia-response elements – HREs) και στον ενισχυτή της 3' περιοχής του γονιδίου της ερυθροποιητίνης και επάγει την έκφρασή του στους νεφρούς και στο ήπαρ (Singh, 2018).

Η ερυθροποιητίνη (EPO) είναι μια κυτοκίνη που συντίθεται κυρίως στους νεφρούς. Στοχεύει κύτταρα που βρίσκονται στο μυελό των οστών και εκφράζουν τον υποδοχέα της ερυθροποιητίνης (EPOR) και έχει ως κύρια λειτουργία τη ρύθμιση της κατανομής οξυγόνου στους περιφερικούς ιστούς (Testa, 2004; Zivot et al., 2018). Η δράση της EPO διεκπεραιώνεται μέσω της πρόσδεσής της στον υποδοχέα της EPOR, ο οποίος βρίσκεται στην επιφάνεια των ερυθροποιητικών προγονικών κυττάρων (Εικόνα 3). Η πρόσδεση της EPO στον υποδοχέα της έχει ως αποτέλεσμα την ενεργοποίηση πολλαπλών μονοπατιών μεταγωγής σήματος, συμπεριλαμβανομένων και των σηματοδοτικών μονοπατιών του μεταγωγέα σήματος και ενεργοποιητή της μεταγραφής 5 (STAT5), της PI3 κινάσης/AKT και SHC/RAS/mitogen-activated protein kinase (MAPK). Η απενεργοποίηση οποιουδήποτε εκ των δύο πρώτων σηματοδοτικών μονοπατιών έχει ως αποτέλεσμα μειωμένη παραγωγή ερυθροκυττάρων, ενώ η απενεργοποίηση του τρίτου μονοπατιού έχει ελαφρές επιδράσεις στην ερυθροποίηση (Hattangadi et al., 2011).

Οι πρωτεΐνες της οικογένειας GATA (GATA-1 και GATA-2) προσδένονται στην περιοχή του υποκινητή του γονιδίου της EPO και το ρυθμίζουν αρνητικά. Εκτός από τη ρύθμιση του γονιδίου της EPO, οι πρωτεΐνες GATA-1 και GATA-2 έχουν σημαντικό ρόλο στη ρύθμιση της έκφρασης ειδικών γονιδίων για την κάθε κυτταρική σειρά διαφοροποίησης κατά την ερυθροποίηση, καθώς η σχετική αναλογία έκφρασης των GATA-1 και GATA-2 προωθεί την έκφραση των γονιδίων στόχων που είναι απαραίτητα για την ωρίμανση των ερυθροκυττάρων και την έκφραση των γονιδίων της β-σφαιρίνης (Zivot et al., 2018). Στα ερυθροκύτταρα, μεταγραφικοί παράγοντες ειδικοί για κάθε κυτταρική σειρά διαφοροποίησης



***Εικόνα 3**. Σχηματική αναπαράσταση του μηχανισμού σηματοδότησης της ερυθροποιητίνης μέσω του υποδοχέα της. (1) Οι υποδοχείς EPO υπάρχουν σαν ανενεργά προσχηματισμένα διμερή στην επιφάνεια του κυττάρου με κινάσες JAK2 συνεχώς προσδεδεμένες. (2) Η σύνδεση της ερυθροποιητίνης προκαλεί αναδιάταξη των υποδοχέων που επιτρέπει την trans-φωσφορυλίωση και ενεργοποίηση των κινασών JAK2. (3) Οι ενεργοποιημένες κινάσες φωσφορυλιώνουν κατάλοιπα τυροσίνης στους υποδοχείς της EPO δημιουργώντας θέσεις πρόσδεσης για πρωτεΐνες που περιέχουν SH2. (4) Σηματοδοτικά μόρια που περιέχουν αυτοτελείς λειτουργικές περιοχές SH2 προσδένονται στους φωσφορυλιωμένους υποδοχείς EPO και φωσφορυλιώνονται από τις κινάσες JAK2. Τα ενεργοποιημένα σηματοδοτικά μόρια επάγουν την ερυθροποίηση. (5) Τερματισμός της μετάδοσης σήματος συμβαίνει όταν φωσφατάσες αποφωσφορυλιώνουν του υποδοχείς της EPO και τα σύμπλοκα που είναι προσδεδεμένα στην ερυθροποιητίνη ενδοκυτταρώνονται (από Singh, 2018)*

όπως ο GATA1, SCL/Tal1, LMO2 και άλλοι, αλληλοεπιδρούν με επαγόμενους από την EPO μεταγωγείς σήματος και ενεργοποιητές τη μεταγραφής, όπως ο STAT5, για να εκφράσουν τα απαιτούμενα για την ερυθροποιητική διαφοροποίηση (Hattangadi et al., 2011).

Ο STAT5 είναι ένας μεταγραφικός παράγοντας της οικογένειας STAT, στην οποία ανήκουν μεταγραφικοί παράγοντες που έχουν ρόλο ως μεταγωγείς σήματος και ενεργοποιητές της μεταγραφής (Signal transducers and activators of transcription). Οι παράγοντες αυτής της οικογένειας βρίσκονται ανενεργοί στο κυτταρόπλασμα και ενεργοποιούνται από κάποιο εξωκυτταρικό σήμα (όταν για παράδειγμα μια κυτοκίνη, ένας αυξητικός παράγοντας ή μια ορμόνη προσδένονται σε συγκεκριμένο επιφανειακό υποδοχέα του κυττάρου). Ειδικότερα για τον STAT5, ένας από τους υποδοχείς κυτοκίνης που προκαλεί την ενεργοποίησή του είναι ο EPOR μετά από πρόσδεση της EPO (Testa, 2004). Μετά την πρόσδεση ενός συνδέτη στον υποδοχέα, προκαλείται η φωσφορυλίωση των κινασών JAK (Janus Kinase) οι οποίες αλληλοφωσφορυλιώνονται. Μετά τη φωσφορυλίωση και ενεργοποίησή τους, οι JAKs φωσφορυλιώνουν τον υποδοχέα δημιουργώντας θέσεις πρόσδεσης για τον STAT5, τον οποίο και φωσφορυλιώνουν. Το μονοπάτι JAK-STAT αλληλοεπιδρά και με άλλα μονοπάτια, όπως αυτά των κινασών ERK-MAPK και της κινάσης PI3K (Katsantoni, 2012). Ο ενεργοποιημένος STAT5 διμερίζεται και μετατοπίζεται από το

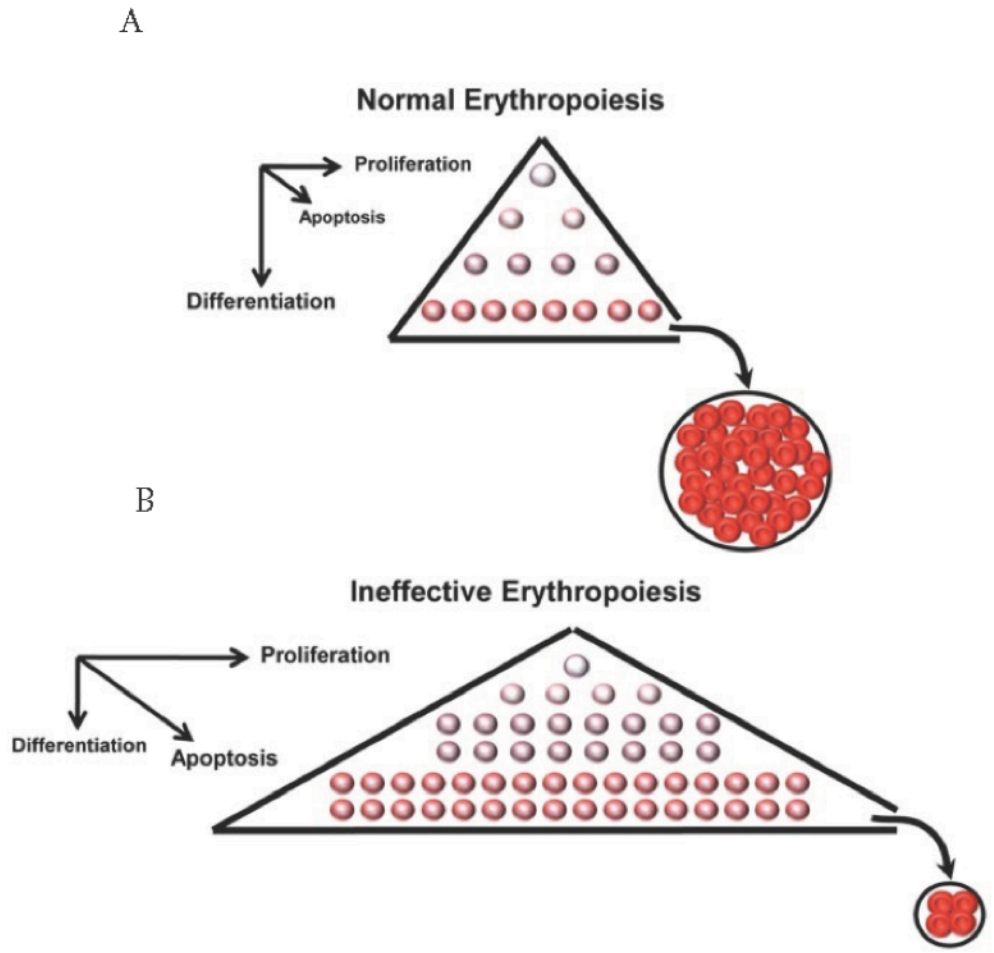
κυτταρόπλασμα στον πυρήνα, όπου και δρα ρυθμίζοντας τη μεταγραφή γονιδίων σχετικών με τον κυτταρικό πολλαπλασιασμό και τη διαφοροποίηση. Στα γονίδια στόχους του STAT5 περιλαμβάνεται και το αντιαποπτωτικό γονίδιο *Bcl-X_L*. Η ρύθμιση του γονιδίου από τον STAT5 εξηγεί την αντιαποπτωτική του δράση στις ερυθροποιητικές κυτταρικές σειρές. Απώλεια έκφρασης και των δύο ισομορφών (STAT5a & STAT5b) του παράγοντα STAT5 σε έμβρυα ποντικών έχει ως αποτέλεσμα σοβαρή αναιμία λόγω μειωμένης επιβίωσης των ερυθροποιητικών προγονικών κυττάρων (Testa, 2004).

2. Αναιμία

Η αναιμία είναι μια κατάσταση στην οποία ο αριθμός των ερυθρών αιμοσφαιρίων στο αίμα ή η ικανότητά τους να μεταφέρουν οξυγόνο στους ιστούς είναι ανεπαρκής και δεν ανταποκρίνεται στις φυσιολογικές ανάγκες του οργανισμού. Οι ανάγκες αυτές ποικίλουν ανάλογα με την ηλικία, το φύλο, το υψόμετρο και άλλους παράγοντες. Διάφοροι παράγοντες μπορούν να προκαλέσουν αναιμία, ενώ μπορεί να υπάρχουν περισσότερες από μια αιτίες πρόκλησης αναιμίας σε ένα άτομο. Στους παράγοντες αυτούς ανήκουν η σιδηροπενία, η οποία είναι υπεύθυνη για το 50% των περιπτώσεων αναιμίας, ενώ η αυξημένη απώλεια αίματος κατά την έμμηνο ρύση και οι οξείες ή χρόνιες μολύνσεις μπορούν επίσης να χαμηλώσουν τη συγκέντρωση της αιμοσφαιρίνης (Hb) του αίματος και να προκαλέσουν αναιμία. Αυξημένο κίνδυνο αναιμίας έχουν άτομα με ελλείψεις βιταμινών όπως οι βιταμίνες A, B12 και ριβοφλαβίνη. Άλλες αιτίες αναιμίας είναι οι αιμοσφαιρινοπάθειες (De Benoist et al., 2008).

2.1 Μεσογειακή αναιμία

Η μεσογειακή αναιμία ή θαλασσαιμία είναι μια αιμοσφαιρινοπάθεια στην οποία παρατηρείται αναποτελεσματική ερυθροποίηση που χαρακτηρίζεται από αυξημένη απόπτωση των ερυθροκυττάρων που ωριμάζουν (Εικόνα 4). Στην θαλασσαιμία έχει εντοπιστεί μεγάλος αριθμός διαφορετικών μεταλλαγών που προκαλούν μη φυσιολογική έκφραση των γονιδίων των σφαιρινών, με αποτέλεσμα την μερική ή πλήρη μείωση της σύνθεση των αλυσίδων των σφαιρινών (Rivella, 2009).



Εικόνα 4*. Σχηματική αναπαράσταση της φυσιολογικής και της αναποτελεσματικής ερυθροποίησης. Α) Σε φυσιολογικές συνθήκες, οι ερυθροβλάστες παράγουν ερυθροκύτταρα και υπάρχει ισορροπία μεταξύ πολλαπλασιασμού, διαφοροποίησης και απόπτωσης των κυττάρων. Β) Στην αναποτελεσματική ερυθροποίηση, προγονικά ερυθροποιητικά κύτταρα αποπίπτουν, η παραγωγή ερυθροκυττάρων ελαττώνεται και διαταρράσεται η ισορροπία μεταξύ πολλαπλασιασμού, διαφοροποίησης και απόπτωσης των κυττάρων. Στη θαλασσαιμία τα προγονικά ερυθροποιητικά κύτταρα πολλαπλασιάζονται με αυξημένο ρυθμό και ταυτόχρονα μειώνεται ο ρυθμός κυτταρικής διαφοροποίησης, οδηγώντας σε αύξηση των προγονικών ερυθροποιητικών κυττάρων παρά τον αυξημένο ρυθμό απόπτωσης (Ginzburg & Rivella, 2011).*

Η φυσιολογική ενήλικη ανθρώπινη αιμοσφαιρίνη Α (HbA) αποτελείται από 2 ζεύγη αλυσίδων σφαιρινών α και β ($\alpha_2\beta_2$), η σύνθεση των οποίων είναι συνήθως συντονισμένη ώστε να υπάρχουν σε ίσες ποσότητες (Ribeil et al., 2013). Ανάλογα με το ποια γονίδια σφαιρινών έχουν επηρεαστεί, οι θαλασσαιμίες διακρίνονται σε α- και β-θαλασσαιμίες. Η α-θαλασσαιμία συχνά οφείλεται σε ελλείψεις μέσα στο σύμπλεγμα γονιδίων των α-σφαιρινών που οδηγούν σε απώλεια λειτουργίας ενός ή και των δύο γονιδίων της α-σφαιρίνης σε κάθε ένα από τους δύο γονιδιακούς τόπους. Σε περίπτωση απενεργοποίησης και των 4 γονιδίων της α-σφαιρίνης παρατηρείται η μείζων α-θαλασσαιμία, οποία μπορεί να προκαλέσει ενδομήτριο θάνατο. Η β-θαλασσαιμία οφείλεται συνήθως σε σημειακές μεταλλαγές του β-

γονιδίου της σφαιρίνης που περιλαμβάνουν μόνο ένα ή περιορισμένο αριθμό νουκλεοτιδίων. Ανάλογα με την περιοχή της μεταλλαγής, η σοβαρότητα της πάθησης ποικίλλει με τις πιο σοβαρές συνέπειες να παρατηρούνται σε μεταλλαγές του υποκινητή και των μεταγραφόμενων περιοχών (Rivella, 2009). Οι επιπτώσεις αυτών των μεταλλαγών έχουν ως αποτέλεσμα τη διαταραχή της ισορροπίας στη σύνθεση των αλυσίδων της α- και β-σφαιρίνης και τη συσσώρευση ελεύθερων αλυσίδων α-σφαιρίνης, οι οποίες σχηματίζουν τοξικά συσσωματώματα. Τα συσσωματώματα τελικά οδηγούν σε μειωμένη ζωή των ερυθροκυττάρων λόγω αιμόλυσης, σε πρόωρο θάνατο των προγονικών κυττάρων στο μυελό των οστών και τελικά σε αναιμία (Ribeil et al., 2013). Η β-θαλασσαιμία χωρίζεται σε τρεις τύπους, την ελάσσονα, την ενδιάμεση και τη μείζονα. Η ελάσσων β-θαλασσαιμία δεν εμφανίζει συμπτώματα και ο φορέας έχει φυσιολογικό προσδόκιμο ζωής, η ενδιάμεση συνδέεται με διάφορα συμπτώματα και εμφανίζεται με ποικίλους βαθμούς σοβαρότητας, ενώ η μείζων είναι η πιο βαριά μορφή. Σε ασθενείς που εξαρτώνται από μεταγγίσεις αίματος παρατηρείται υπερφόρτωση σιδήρου και είναι απαραίτητη η χηλική θεραπεία για την αποσιδήρωσή τους. Μια συχνή αιτία θανάτου ασθενών με μείζων β-θαλασσαιμία είναι οι καρδιολογικές επιπλοκές λόγω υπερφόρτωσης σιδήρου. Οι επιπλοκές από την θεραπεία αλλά και ο σύνθετος φαινότυπος της β-θαλασσαιμίας είναι δύο από τους λόγους για τους οποίους είναι απαραίτητη η ανάπτυξη νέων μοριακών μεθόδων για την κατηγοριοποίηση των ασθενών, αλλά και για τη διαχείριση της θεραπείας τους. Ένα σημαντικό πρόβλημα που υπάρχει στην κλινική πράξη είναι η διάκριση μεταξύ της ενδιάμεσης β-θαλασσαιμίας (Thalassaemia Intermedia, TI) και της μείζονος β-θαλασσαιμίας (Thalassaemia Major, TM), ώστε να αποφευχθούν μεταγγίσεις αίματος, οι οποίες δεν είναι απαραίτητες στους ασθενείς του ενδιάμεσου τύπου, και να ξεκινήσουν εγκαίρως οι μεταγγίσεις σε ασθενείς του μείζονος τύπου. Συνεπώς, είναι σημαντική η ανακάλυψη νέων βιοδεικτών και παραγόντων που συνεισφέρουν στη διάκριση μεταξύ των δύο καταστάσεων της ασθένειας, η οποία σε πολλές περιπτώσεις είναι δύσκολη (Katsantoni, 2019).

3. Ανακάλυψη βιοδεικτών

Ο βιοδείκτης είναι ένα αντικειμενικά μετρήσιμο χαρακτηριστικό το οποίο περιγράφει μια φυσιολογική ή μη φυσιολογική βιολογική κατάσταση σε έναν οργανισμό. Βιομόρια όπως το DNA, το RNA, οι πρωτεΐνες και τα πεπτίδια ή οι διάφορες χημικές μετατροπές των βιομορίων μπορούν να αποτελέσουν χρήσιμους βιοδείκτες. Με την πρόοδο των τεχνολογιών γενομικής ανάλυσης και των στοχευμένων μοριακών θεραπειών, ιδιαίτερη σημασία έχουν

αρχίσει να λαμβάνουν οι βιοδείκτες στην κλινική πράξη (Goossens et al., 2015). Μια κατηγοριοποίηση των βιοδεικτών, όπως παρουσιάζεται από τους McDermott et al 2013, είναι ο διαχωρισμός τους σε βιοδείκτες κινδύνου, διαγνωστικούς και προγνωστικούς βιοδείκτες. Οι βιοδείκτες κινδύνου αναγνωρίζουν ασθενείς οι οποίοι πιθανώς θα αναπτύξουν μια ασθένεια. Οι διαγνωστικοί βιοδείκτες συμβάλλουν στον εντοπισμό μιας ασθένειας σε πρώιμο στάδιο, βοηθούν στην ταξινόμηση σε υποτύπους της ασθένειας και στον χαρακτηρισμό της απόκρισης στη θεραπεία. Τέλος, οι προγνωστικοί βιοδείκτες συμμετέχουν στην πρόβλεψη της προόδου ή της επανεμφάνισης της ασθένειας και αναγνωρίζουν ασθενείς που είναι πιθανότερο να ανταποκριθούν σε κάποια θεραπεία (McDermott et al., 2013). Η ανακάλυψη βιοδεικτών για εφαρμογή σε κλινικό επίπεδο και για βελτίωση της εξατομικευμένης ιατρικής ακριβείας είναι πολύ σημαντική σήμερα.

Η ιατρική ακριβείας (Precision Medicine), γνωστή παλαιότερα και ως εξατομικευμένη ιατρική, είναι μια μορφή ιατρικής που λαμβάνει υπόψη της τα ειδικά χαρακτηριστικά ενός ασθενή ώστε να εξατομικεύσει την πρόληψη, τη διάγνωση και τη θεραπεία. Εκτός από τις κλινικές και επιδημιολογικές πληροφορίες, η ιατρική ακριβείας βασίζεται σε πληροφορίες που παρέχονται από διάφορα πεδία ομικών αναλύσεων που εμφανίστηκαν τα τελευταία χρόνια όπως είναι η γενομική, η μεταγραφομική και η πρωτεομική. Σε καθεμία από αυτές τις ομικές αναλύσεις μελετάται ένα διαφορετικό σύνολο βιομορίων. Στη γενομική πραγματοποιείται μελέτη του γονιδιώματος ή ενός μεγάλου υποσυνόλου του, στη μεταγραφομική μελετάται το σύνολο των μεταγράφων ενός βιολογικού επιπέδου οργάνωσης (κύτταρο, ιστός, οργανισμός), ενώ στην πρωτεομική γίνεται ανάλυση των πρωτεϊνών που απαντώνται σε κάποιο από τα βιολογικά επίπεδα οργάνωσης (Quezada et al., 2017).

Στο παρελθόν, η ανακάλυψη ενός βιοδείκτη ήταν περισσότερο στοχευμένη και βασιζόταν σε μελέτες και στοιχεία σχετικά με την ασθένεια, καθιστώντας τα πρώτα στάδια της ανακάλυψης αρκετά δύσκολα. Πλέον, οι ομικές τεχνολογίες προσφέρουν τη δυνατότητα μελέτης χιλιάδων μορίων χωρίς να υπάρχει κάποια προηγούμενη γνώση για αυτά ή συσχέτισή τους με κάποια ασθένεια, με αποτέλεσμα να είναι δυνατή η διατύπωση μιας ερευνητικής υπόθεσης βασισμένης στα δεδομένα αυτά (Goossens et al., 2015). Οι τεχνολογίες αυτές έχουν επιτρέψει την πρόσβαση σε μεγάλη ποσότητα ποσοτικών δεδομένων διαφορικώς εκφραζόμενων mRNAs, microRNAs και πρωτεϊνών από μελέτες ασθενών και υγιών ατόμων. Για την ανάλυση και εξαγωγή συμπερασμάτων από τέτοιου είδους και μεγέθους δεδομένα είναι απαραίτητη η χρήση βιοπληροφορικών εργαλείων (McDermott et al., 2013).

3.1 Αλληλούχιση RNA

Μια από τις πιο συχνά χρησιμοποιούμενες ομικές τεχνολογίες για ανακάλυψη βιοδεικτών είναι η αλληλούχιση του μεταγραφώματος ή αλλιώς αλληλούχιση RNA (RNA-Sequencing, RNA-Seq). Η αλληλούχιση πραγματοποιείται σε μηχανήματα αλληλούχιση νέας γενιάς (Next Generation Sequencing, NGS) και αποτελεί την πιο ευρέως χρησιμοποιούμενη μέθοδο για τη μελέτη γονιδιακής έκφρασης. Συγκριτικά με την παλαιότερη μέθοδο των μικροσυστοιχιών, η RNA-Seq έχει λιγότερο θόρυβο, μεγαλύτερη αξιοπιστία και εύρος στα δεδομένα που παράγονται.

Εικόνα 5. *Πρωτόκολλο προετοιμασίας βιβλιοθηκών για RNA-Seq (προσαρμογή από Illumina TruSeq RNA Sample Preparation Kit v2)*

Το μεγαλύτερο πλεονέκτημα της μεθόδου είναι η ακριβής αλληλούχιση του μεταγράφου, η οποία μπορεί να οδηγήσει στην ανακάλυψη άγνωστων γονιδίων και νέων μεταγράφων γνωστών γονιδίων (Hrdlickova et al., 2017).

Κατά το RNA-Seq, δεν αλληλουχείται το μόριο RNA αλλά ο συμπληρωματικός κλώνος DNA. Αυτό συμβαίνει λόγω των πλεονεκτημάτων που έχει πειραματικά το μόριο του DNA, για παράδειγμα της ευκολότερης διαχείρισής του με ήδη εδραιωμένες τεχνικές, και της μεγαλύτερης σταθερότητάς του έναντι του RNA. Η διαδικασία της προετοιμασίας της βιβλιοθήκης από το δείγμα προς αλληλούχιση συνοψίζεται στα παρακάτω βήματα και στην Εικόνα 5:

- Απομόνωση των μορίων RNA, μέσω σφαιριδίων που φέρουν ολιγονουκλεοτίδια θυμίνης (T) στα οποία προσδένονται τα μόρια RNA με ουρά πολυ-αδενίνης (poly-A).
- Κατακερματισμός των μορίων σε μικρότερου μήκους αλληλουχίες μέσω ενζυμικής επεξεργασίας.
- Αντίστροφη μεταγραφή των μορίων με τυχαία εξαμερή ως εκκινητές και παραγωγή μορίων συμπληρωματικού DNA (First strand cDNA synthesis).
- Σύνθεση του δεύτερου κλώνου cDNA (Second strand cDNA synthesis).

- Επισκευή των μονόκλωνων άκρων των cDNA με χρήση πολυμεράσης και 3΄-5΄ εξωνουκλεάσης.
- Αδενυλίωση των 3΄ άκρων για την αποφυγή πρόσδεσης των τμημάτων μεταξύ τους
- Πρόσδεση αλληλουχιών προσαρμογέων (Adapters). Οι αλληλουχίες προσαρμογείς δίνουν τη δυνατότητα προσδιορισμού του κλώνου από τον οποίο προήλθε η αλληλουχία (Strand-specificity). Στις αλληλουχίες προσαρμογείς μπορούν να συμπεριληφθούν και μοριακές ετικέτες (molecular labels – barcodes), οι οποίες βοηθάνε στη διάκριση του δείγματος από το οποίο προήλθε η κάθε αλληλουχία, στην περίπτωση που περισσότερα από ένα δείγματα RNA αλληλουχηθούν ταυτόχρονα.
- Ενίσχυση με αλυσιδωτή αντίδραση πολυμεράσης (PCR) με τυχαίους εκκινητές.
- Επιβεβαίωση της ποιότητας των βιβλιοθηκών.
- Κανονικοποίηση της ποσότητας και συνένωσή τους προς αλληλούχιση.

Η αλληλούχιση των δειγμάτων μπορεί να γίνει Single-end, στην οποία το μηχάνημα αλληλούχισης διαβάζει την κάθε αλληλουχία DNA μόνο σε μία κατεύθυνση ή Paired-end, στην οποία η κάθε αλληλουχία DNA διαβάζεται και προς τις δύο κατευθύνσεις (Hrdlickova et al., 2017).

3.1.1 Βιοπληροφορική Ανάλυση Αλληλούχισης RNA (RNA-Seq)

Τα αποτελέσματα της αλληλούχισης RNA νέας γενιάς δημιουργούν μεγάλο όγκο δεδομένων και χρειάζονται επεξεργασία με υπολογιστικές τεχνικές. Ένας από τους λόγους πραγματοποίησης ενός πειράματος αλληλούχισης RNA είναι η αναζήτηση γονιδίων, τα οποία έχουν απορρυθμιστεί σημαντικά από τη φυσιολογική κατάσταση του κυττάρου. Η απάντηση δίνεται με ανάλυση διαφορικής έκφρασης (Differential Expression Analysis), η οποία αποτελείται από πέντε βασικά βήματα όπως περιγράφονται στο άρθρο των Yamalanchili et al (2017). Το μηχάνημα αλληλούχισης νέας γενιάς παράγει εκατομμύρια αλληλουχίες, που αποκαλούνται διαβάσματα (reads), τις οποίες αποθηκεύει σε αρχεία κειμένου της μορφής *fastq,* μαζί με πληροφορίες για την ποιότητα της αλληλούχισης.

Το πρώτο βήμα στην ανάλυση διαφορικής έκφρασης είναι ο έλεγχος της ποιότητας των δεδομένων. Σε αυτό το βήμα ελέγχονται η ποιότητα της αλληλούχισης των βάσεων σε κάθε διάβασμα, το περιεχόμενο των διαβασμάτων σε νουκλεοτίδια γουανίνης-κυτοσίνης (GC content), καθώς και αλληλουχίες οι οποίες μπορεί να απαντώνται πολύ περισσότερο από το μέσο όρο (overrepresented sequences).

Στο επόμενο βήμα αφαιρούνται οι αλληλουχίες ή τμήματα αυτών που δεν έχουν καλή ποιότητα ώστε να χρησιμοποιηθούν μόνο αξιόπιστα δεδομένα στην επακόλουθη ανάλυση. Στα μηχανήματα NGS είναι συχνό το φαινόμενο πτώσης της ποιότητας της αλληλούχισης στα άκρα της παραγόμενης αλληλουχίας. Συνεπώς, αφαιρείται συνήθως ένας μικρός αριθμός βάσεων από την αρχή και το τέλος κάθε διαβάσματος.

Κατά το τρίτο βήμα πραγματοποιείται η στοίχιση των διαβασμάτων στο γονιδίωμα αναφοράς με τη χρήση κάποιου κατάλληλου εργαλείου. Στα ευκαρυωτικά γονιδιώματα, λόγω της παρουσίας ιντρονίων στις αλληλουχίες των γονιδίων, χρησιμοποιούνται αλγόριθμοι στοίχισης που λαμβάνουν υπόψη τους τα σημεία ματίσματος και μπορούν να στοιχίσουν ένα διάβασμα σωστά χωρίζοντας κατάλληλα την αλληλουχία του στο σημείο που μεσολαβεί το ιντρόνιο στο αντίστοιχο γονιδίωμα αναφοράς. Το αποτέλεσμα της στοίχισης είναι ένα αρχείο σε μορφή BAM (δυαδικό αρχείο) ή SAM (το αντίστοιχο μη δυαδικό αρχείο) που περιλαμβάνει όλες της πληροφορίες για τη στοίχιση. Ανάλογα με τον αριθμό των θέσεων στο γονιδίωμα στα οποία έχουν στοιχηθεί, τα διαβάσματα μπορούν να ταξινομηθούν σε μοναδικά στοιχισμένα, αν έχουν στοιχηθεί μόνο σε ένα σημείο, και πολλαπλά στοιχισμένα, αν έχουν στοιχηθεί σε περισσότερα από ένα σημεία.

Τη στοίχιση ακολουθεί η ποσοτικοποίηση της γονιδιακής έκφρασης. Ο υπολογισμός της έκφρασης ενός γονιδίου γίνεται με βάση των αριθμό των διαβασμάτων που στοιχήθηκαν μοναδικά σε κάποιο γονίδιο. Οι αλγόριθμοι που αποδίδουν τα μοναδικά στοιχισμένα διαβάσματα σε κάποιο γονίδιο λειτουργούν αντλώντας παράλληλα τις πληροφορίες για το σημείο της στοίχισης του διαβάσματος από το αρχείο SAM και για το σχολιασμό (Annotation) του γονιδιώματος από κάποιο αρχείο μορφής gtf. Σε περίπτωση που ένα διάβασμα έχει στοιχηθεί μοναδικά σε ένα σημείο του γονιδιώματος αλλά επικαλύπτεται με παραπάνω από ένα γονίδια, είναι στην ευχέρεια του ερευνητή να επιλέξει αν αυτό θα αποδοθεί και στα δύο γονίδια ή σε κανένα από τα δύο γονίδια.

Το τελευταίο βήμα είναι η ανάλυση διαφορικής έκφρασης η οποία πραγματοποιείται με τη χρήση στατιστικών πακέτων. Τα δεδομένα εισάγονται με τη μορφή αρχείων κειμένου που περιέχουν την πληροφορία των τιμών έκφρασης όπως αυτή προέκυψε από το προηγούμενο βήμα. Μαζί με τα δεδομένα εισάγεται και ένα αρχείο κειμένου που περιγράφει τις ιδιότητες των δειγμάτων, όπως την ονομασία τους, τον φαινότυπο ή γονότυπο τον οποίο φέρουν, το φύλο τους ή και άλλα χαρακτηριστικά. Το αποτέλεσμα της ανάλυσης που προκύπτει περιλαμβάνει τιμές μεταβολής έκφρασης μεταξύ των δύο καταστάσεων οι οποίες συγκρίνονται (με τη μορφή log2 Fold Change), καθώς και τιμές στατιστικής σημαντικότητας των τιμών αυτών (Yalamanchili et al., 2017).

3.1.2 Αλληλούχιση RNA και Βιοδείκτες

Η αναγνώριση των στατιστικώς σημαντικών διαφορικώς εκφραζόμενων γονιδίων ανάμεσα στις διάφορες καταστάσεις κάποιας ασθένειας (π.χ. ενδιάμεση β-θαλασσαιμία – μείζων β-θαλασσαιμία), καθώς και υγιών ατόμων, είναι σημαντική σε πολλές μελέτες γονιδιακής έκφρασης. Μια σημαντική πρόκληση είναι η εύρεση του ελάχιστου δυνατού συνδυασμού γονιδίων που μπορούν να δώσουν τη μέγιστη δυνατή προβλεπτική δύναμη για διαγνωστικούς σκοπούς στην ιατρική (Zararsiz et al., 2017). Οι τιμές γονιδιακής έκφρασης που προκύπτουν από την ανάλυση πειραμάτων αλληλούχισης RNA έχουν κινήσει το ερευνητικό ενδιαφέρον με αποτέλεσμα να έχουν προταθεί πολλοί αλγόριθμοι μηχανικής μάθησης για την ταξινόμηση δειγμάτων με χρήση αυτών των δεδομένων. Οι αλγόριθμοι μηχανικής μάθησης είναι αρκετά υποσχόμενοι με εφαρμογές σε πολυδιάστατα σύνολα δεδομένων όπως είναι αυτά που προκύπτουν από ομικά πειράματα. Σε πολλές εφαρμογές στόχος είναι η δημιουργία και χρήση ενός καλού προβλεπτικού μοντέλου, ενώ σε άλλες περιπτώσεις στόχος είναι η ταυτοποίηση των μεταβλητών που επιτρέπουν την καλή προβλεπτική ικανότητα. Ο εντοπισμός των μεταβλητών που αυξάνουν την προβλεπτική ισχύ επιτρέπει την απομάκρυνση των υπολοίπων μεταβλητών που εισάγουν θόρυβο στο μοντέλο πρόβλεψης. Ένας από τους πιο κατάλληλους αλγορίθμους για τις δύο αυτές εφαρμογές είναι ο αλγόριθμος Τυχαίο Δάσος (Random Forest, RF), ο οποίος παρέχει τιμές σημαντικότητας των μεταβλητών σχετικά με την προβλεπτική ισχύ του μοντέλου (Degenhardt et al., 2019).

3.2 Μέθοδοι Μηχανικοί Μάθησης

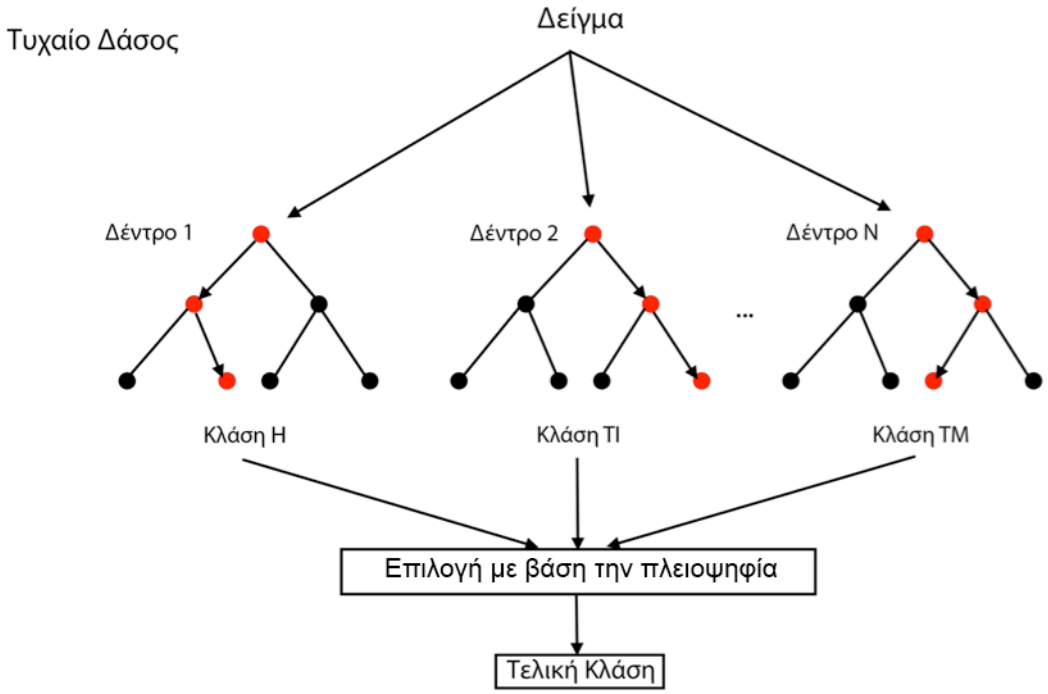
Οι μέθοδοι μηχανικής μάθησης είναι αλγόριθμοι που είναι σε θέση να εκπαιδεύονται από υπάρχοντα δεδομένα και να βελτιώνουν την απόδοσή τους στην εργασία για την οποία προορίζονται. Η διαδικασία εκμάθησης μπορεί να είναι με επίβλεψη (supervised learning) ή χωρίς επίβλεψη (unsupervised learning). Στην επιβλεπόμενη μάθηση, ο αλγόριθμος εκπαιδεύεται σε ένα υποσύνολο των δεδομένων, που ονομάζεται σύνολο εκπαίδευσης, με τα παραδείγματα που περιέχονται σε αυτό να χαρακτηρίζονται από μια κατηγορία (πχ υγιής – ΤΙ – ΤΜ). Αντίθετα, στη μάθηση χωρίς επίβλεψη τα παραδείγματα δεν χαρακτηρίζονται από καμία κατηγορία με αποτέλεσμα ο αλγόριθμος να χρησιμοποιεί μόνο τα χαρακτηριστικά τους. Σε πειράματα κατηγοριοποίησης (classification), η κατηγορία στην οποία ανήκει το κάθε παράδειγμα υποδεικνύει και το επιθυμητό αποτέλεσμα του αλγορίθμου μηχανικής μάθησης.

3.2.1 Αλγόριθμος «Δέντρο Απόφασης»

Το «Δέντρο Απόφασης» είναι ένα αλγόριθμος μηχανικής μάθησης με επίβλεψη που χρησιμοποιείται για ταξινόμηση. Ο αλγόριθμος χρησιμοποιείται με σύνολα δεδομένων στα οποία υπάρχουν δείγματα με ένα σύνολο χαρακτηριστικών (π.χ. τιμές γονιδιακής έκφρασης) και ανήκουν σε κάποια κλάση (π.χ. Υγιής – ΤΙ – ΤΜ). Το «Δέντρο Απόφασης» εκπαιδεύεται με ένα υποσύνολο των συνολικών δεδομένων με γνωστές ετικέτες που ονομάζεται σύνολο εκπαίδευσης και σχηματίζει ένα σύνολο από κανόνες αποφάσεων με σκοπό την ταξινόμηση όσο το δυνατόν περισσότερων δειγμάτων στη σωστή κλάση. Στη συνέχεια αξιολογείται η επίδοσή του στα υπόλοιπα δείγματα που δεν υπάρχουν στο σύνολο εκπαίδευσης και δημιουργούν το σύνολο δοκιμών (test set), το οποίο δεν φέρει γνωστές ετικέτες.

3.2.2 Αλγόριθμος «Τυχαίο Δάσος»

Ο αλγόριθμος «Τυχαίο Δάσος» είναι μια μέθοδος μηχανικής μάθησης που αναπτύχθηκε από τον Leo Breiman (Breiman, 2001) και χρησιμοποιεί ένα συνδυασμό «Δέντρων Απόφασης» (Εικόνα 6). Κάθε ένα από τα δέντρα απόφασης εκπαιδεύεται σε διαφορετικό υποσύνολο των δεδομένων και οι μεταβλητές που επιλέγουν από το σύνολο των χαρακτηριστικών για να σχηματίσουν τους κανόνες αποφάσεων είναι τυχαίες. Σε έναν κανόνα απόφασης μπορούν να χρησιμοποιηθούν συνήθως πάνω από μια μεταβλητές που να δίνουν το ίδιο προβλεπτικό αποτέλεσμα. Το συνολικό αποτέλεσμα του αλγορίθμου είναι το κοινό αποτέλεσμα της πλειοψηφίας των δέντρων απόφασης που συμμετέχουν στο τυχαίο δάσος. Το Τυχαίο Δάσος έχει ορισμένα χαρακτηριστικά που το καθιστούν κατάλληλο για τη χρήση με δεδομένα από πειράματα αλληλούχισης RNA. Μπορεί να χρησιμοποιηθεί όταν τα χαρακτηριστικά-μεταβλητές είναι πολύ περισσότερα από τα δείγματα, έχει καλή προβλεπτική ικανότητα ακόμα και όταν οι περισσότερες μεταβλητές αποτελούν θόρυβο, δεν υπερπροσαρμόζεται στην ταξινόμηση συγκεκριμένων συνόλων δεδομένων, μπορεί να λάβει ως είσοδο συνεχείς και διακριτές μεταβλητές και υπάρχουν πολλές υψηλής ποιότητας και ελεύθερες προς χρήση υλοποιήσεις του (Diaz-Uriarte, 2007; Diaz-Uriarte and de Andres, 2005).



***Εικόνα 6**. Οπτική παρουσίαση της λειτουργίας του Τυχαίου Δάσους. Κάθε δέντρο απόφασης παράγει ένα σύνολο κανόνων ταξινόμησης με βάση τις τιμές των μεταβλητών εισόδου (γονίδια) και ταξινομεί κάθε δείγμα σε μία κλάση. Η τελική κλάση στην οποία ταξινομείται κάθε δείγμα είναι αυτή που έχει επιλεγεί από τα περισσότερα δέντρα απόφασης.*

3.2.3 Εργαλείο GeneSrF

Το GeneSrF (Gene Selection using Random Forest, Diaz-Uriarte, 2007) είναι ένα διαδικτυακό εργαλείο που χρησιμοποιεί τον αλγόριθμο του τυχαίου δάσους, στα πλαίσια της ταξινόμησης ατόμων σε κλάσεις (ασθενείς – υγιείς), ώστε να επιλέξει τα μικρότερα δυνατά σύνολα γονιδίων, τα οποία μπορούν να ταξινομήσουν με μεγάλη ακρίβεια ένα δείγμα. Η εφαρμογή λαμβάνει ως είσοδο δύο αρχεία κειμένου, ένα με τις τιμές γονιδιακής έκφρασης και ένα με τις κλάσεις στις οποίες ανήκουν τα δείγματα. Στη συνέχεια αναλαμβάνει την εκπαίδευση του αλγορίθμου με βάση τα δεδομένα και την παραγωγή κάποιων στατιστικών μέτρων για την αξιοπιστία του προβλεπτικού μοντέλου και τη σημαντικότητα κάθε χαρακτηριστικού. Για την εκπαίδευση και αξιολόγηση το εργαλείο χρησιμοποιεί 200 επαναλήψεις της δειγματοληψίας με τη μέθοδο Bootstrap: Σε κάθε επανάληψη επιλέγονται διαφορετικά υποσύνολα των δεδομένων για την εκπαίδευση και στη συνέχεια υπολογίζεται το ποσοστό σφάλματος (Error rate) του μοντέλου με βάση τα δείγματα που δε συμμετείχαν στην εκπαίδευση του (δείγματα Out-Of-Bag) (Diaz-Uriarte, 2007). Κατά την ανάλυση των αποτελεσμάτων όλων των επαναλήψεων, το εργαλείο υπολογίζει:

- Το ποσοστό σφάλματος του μοντέλου (Bootstrap estimate of prediction error).
- Τον αριθμό των μεταβλητών που επιλέχθηκαν σε κάθε επανάληψη (Number of variables in bootstrapped forests).
- Τη συχνότητα με την οποία ταξινομήθηκε σε κάθε κλάση καθένα από τα δείγματα στις επαναλήψεις που δεν αποτελούσε μέρος του συνόλου εκπαίδευσης (Mean class membership probabilities from out of bag samples).
- Τη συχνότητα εμφάνισης κάθε γονιδίου σε κάθε επανάληψη (Variable frequencies in bootstrapped models).
- Και τη σημαντικότητα κάθε μεταβλητής στα αρχικά δεδομένα (Variable/gene importances from original data).

Η εφαρμογή επιλέγει τελικά ένα σύνολο γονιδίων που παρουσιάζουν την υψηλότερη σημαντικότητα και επανεμφανίζονται συχνά κατά τις επαναλήψεις του αλγορίθμου.

4. Σκοπός

Η παρούσα εργασία είχε ως σκοπό τη χρήση υπολογιστικών εργαλείων σε βιολογικά δεδομένα από πειράματα αλληλούχισης RNA ατόμων με υγιή ή θαλασσαιμικό φαινότυπο, ώστε να ανακαλυφθούν γονίδια που θα μπορούσαν να αποτελέσουν πιθανούς βιοδείκτες για την έγκαιρη και έγκυρη ταξινόμηση των ασθενών με θαλασσαιμία στις κατηγορίες της ενδιάμεσης και μείζονος θαλασσαιμίας με τη χρήση μοριακών μεθόδων στην κλινική πράξη.

Β. Υλικά και Μέθοδοι

1. Δείγματα

Στην εργασία χρησιμοποιήθηκαν δεδομένα από πειράματα RNA-Seq του εργαστηρίου σε ασθενείς με β-θαλασσαιμία και υγιή άτομα. Πιο συγκεκριμένα, χρησιμοποιήθηκαν 54 δείγματα από τα οποία τα 24 συλλέχθηκαν από τα νοσοκομεία της Ferrara και Rovigo στην Ιταλία (8 από υγιή άτομα, 8 από ασθενείς με TI, 8 από ασθενείς με TM) σε συνεργασία με τον Καθηγητή Roberto Gambari και τα υπόλοιπα 30 συλλέχθηκαν από την Κλινική Θαλασσαιμίας στη Λευκωσία και το Ινστιτούτο Νευρολογίας και Γενετικής της Κύπρου (10 από υγιή άτομα, 10 από ασθενείς με TI, 10 από ασθενείς με TM) σε συνεργασία με την Καθηγήτρια Μαρίνα Κλεάνθους. Τα δείγματα οργανώθηκαν σε 18 ομάδες, καθεμία με ένα υγιές, ένα TI και ένα TM δείγμα. Τα δείγματα της ίδιας ομάδας προήλθαν από το ίδιο ερευνητικό κέντρο ή νοσοκομείο, καλλιεργήθηκαν ταυτόχρονα και είχαν ομοιογένεια ως προς την ηλικία και το φύλο.

2. Ετοιμασία Βιβλιοθηκών και αλληλούχιση

Η απομόνωση RNA πραγματοποιήθηκε με τη χρήση Tri Reagent (Sigma) και οι βιβλιοθήκες για την αλληλούχιση κατασκευάστηκαν με τη χρήση του TruSeq RNA Sample Preparation kit v2 (Illumina RS-122-2001) χρησιμοποιώντας 1.5-2 μg του συνολικού RNA. Η αλληλούχιση των βιβλιοθηκών ήταν Single-end, χωρίς Strand-specificity και το μήκος των αλληλουχιών που προέκυψαν ήταν 50 ή 51 νουκλεοτίδια ανάλογα με τη βιβλιοθήκη. Ο έλεγχος της ποιότητας των βιβλιοθηκών έγινε με Agilent Bioanalyzer 2100 (DNA chips 1000, Agilent, 5067-1504) και όλες οι βιβλιοθήκες αλληλουχήθηκαν σε ένα Illumina HiSeq2000. Τα πειράματα πραγματοποιήθηκαν από την Δρ. Χ. Τουμπέκη.

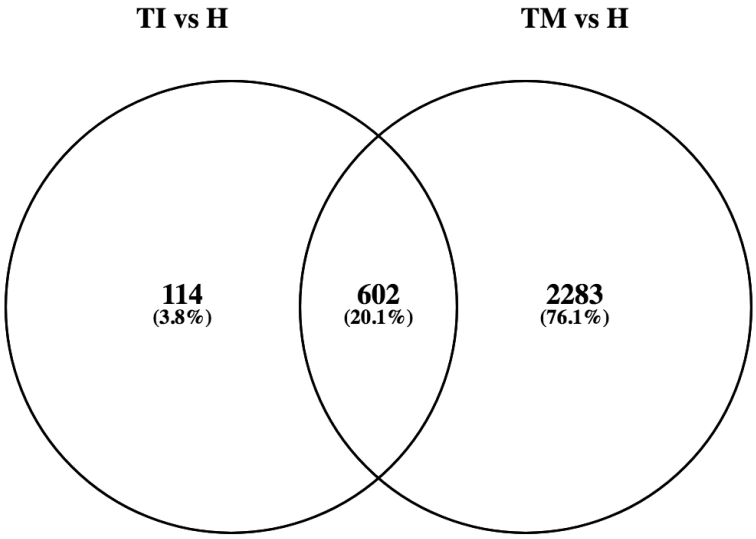
3. Ανάλυση δεδομένων NGS

Μετά την αλληλούχιση ο έλεγχος της ποιότητας των βιβλιοθηκών πραγματοποιήθηκε με το εργαλείο FastQC. Οι βάσεις που βρέθηκαν με χαμηλή ποιότητα (1-3 βάσεις από την αρχή της αλληλουχίας) και οι αλληλουχίες προσαρμογείς αποκόπηκαν με τη χρήση του Trimmomatic (v0.30). Στη συνέχεια ακολούθησε χαρτογράφηση των διαβασμάτων στο ανθρώπινο μεταγράφωμα (hg38) με τη χρήση του TopHat2. Το εργαλείο HTSeq (v0.5.4) χρησιμοποιήθηκε στο επόμενο βήμα για την ποσοτικοποίηση της γονιδιακής έκφρασης. Μετά την ανάλυση, 5 δείγματα της αλληλούχισης RNA (1 Υγιές δείγμα, 3 TI δείγματα και 1

ΤΜ δείγμα) αφαιρέθηκαν από τις παρακάτω αναλύσεις λόγω χαμηλής ποιότητας αλληλούχισης ή χαμηλής απόδοσης της χαρτογράφησής τους. Επιπλέον, ένα ΤΜ δείγμα επανεκτιμήθηκε από το ιατρικό προσωπικό ως ΤΙ μετά την επιλογή των ασθενών και την κατασκευή των βιβλιοθηκών με αποτέλεσμα τα τελικά δείγματα RNA να είναι 17 Υγιή, 16 ΤΙ και 16 ΤΜ δείγματα. Στη συνέχεια πραγματοποιήθηκε ανάλυση διαφορικής έκφρασης γονιδίων στην R με το πακέτο DESeq2 (v1.8.1) με βήματα κανονικοποιήσεων και χρησιμοποιώντας τις ομάδες ως blocking factor ώστε να αποκλειστούν διαφορές που μπορεί να οφείλονται στην ηλικία. Τα διαφορικώς εκφραζόμενα γονίδια θεωρήθηκαν στατιστικά σημαντικά όταν η τιμή του Padjusted για τα συγκεκριμένα γονίδια, όπως αυτή υπολογιζόταν από το DESeq2, ήταν <0.1. Οι παραπάνω αναλύσεις βιοπληροφορικής έγιναν από την Δρ. Α. Νάνου.

4. Επιλογή γονιδίων με το GeneSrF

Για την επιλογή γονιδίων χρησιμοποιήθηκε το εργαλείο GeneSrF (http://genesrf.iib.uam.es). Στο εργαλείο εισάγονται δύο πίνακες, ένας με τα δεδομένα γονιδιακής έκφρασης και ένας με τις ετικέτες των δειγμάτων (Η, ΤΙ, ΤΜ). Για τον πίνακα γονιδιακής έκφρασης χρησιμοποιήθηκαν 2999 γονίδια που βρέθηκαν ως διαφορικώς εκφραζόμενα σε τουλάχιστον μια από τις συγκρίσεις ΤΙ vs Η και ΤΜ vs Η (Εικόνα 7).



Εικόνα 7. *Διάγραμμα Venn των στατιστικώς σημαντικών διαφορικώς εκφραζόμενων γονιδίων των συγκρίσεων ΤΙ vs Η και ΤΜ vs Η δειγμάτων. Στη σύγκριση ΤΙ vs Η βρέθηκαν 716 διαφορικώς εκφραζόμενα γονίδια ενώ στη σύγκριση ΤΜ vs Η 2885 εκ των οποίων τα 602 ήταν κοινά. Συνολικά προέκυψαν 2999 διαφορικώς εκφραζόμενα γονίδια.*

Από το αρχείο με τα δεδομένα έκφρασης κάθε δείγματος απομονώθηκαν οι τιμές των 2999 γονιδίων και δημιουργήθηκε ένας πίνακας 2999 γραμμές x 50 στήλες με κάθε γραμμή να αντιστοιχεί σε ένα γονίδιο και κάθε στήλη σε ένα δείγμα, με εξαίρεση την πρώτη στήλη η

οποία φέρει τις ονομασίες των γονιδίων. Στις τιμές έγινε κανονικοποίηση με τη μέθοδο της μεταμόρφωσης με σταθεροποίηση της διασποράς (Variance stabilizing transformation). Σε ένα δεύτερο πίνακα με διαστάσεις 1 γραμμή x 49 στήλες αποθηκεύτηκαν με αντίστοιχη σειρά οι ετικέτες των δειγμάτων. Οι ετικέτες κωδικοποιήθηκαν με αριθμούς και τα υγιή άτομα αναπαρίστανται ως 0, τα TI ως 1 και τα TM ως 2. Πραγματοποιήθηκαν 15 συγκρίσεις σε καθεμία από τις οποίες οι πίνακες εισόδου αποτελούσαν υποσύνολα των αρχικών δεδομένων. Στις συγκρίσεις TI vs H, TM vs H και TI vs TM αφαιρέθηκαν από τους πίνακες εισόδου αντίστοιχα τα δείγματα TM, TI και H. Στις συγκρίσεις θηλυκών ατόμων αφαιρέθηκαν όλα τα αρσενικά δείγματα, ενώ αντίστοιχα έγινε και στις συγκρίσεις αρσενικών ατόμων. Τέλος, στις συγκρίσεις ανάμεσα στα φύλα, τα θηλυκά άτομα κωδικοποιήθηκαν με τον αριθμό 0 ενώ τα αρσενικά άτομα με τον αριθμό 1. Οι συγκρίσεις που πραγματοποιήθηκαν με το GeneSrF είναι οι ακόλουθες:

- Υγιείς (H) vs Thalassaemia Intermediate (TI) vs Thalassaemia Major (TM)
    - Όλα τα δείγματα
    - Μόνο θηλυκά δείγματα
    - Μόνο αρσενικά δείγματα
- H vs TI
    - Όλα τα δείγματα
    - Μόνο θηλυκά δείγματα
    - Μόνο αρσενικά δείγματα
- H vs TM
    - Όλα τα δείγματα
    - Μόνο θηλυκά δείγματα
    - Μόνο αρσενικά δείγματα
- TI vs TM
    - Όλα τα δείγματα
    - Μόνο θηλυκά δείγματα
    - Μόνο αρσενικά δείγματα
- Θηλυκά δείγματα vs Αρσενικά δείγματα
    - Υγιή δείγματα
    - TI δείγματα
    - TM δείγματα

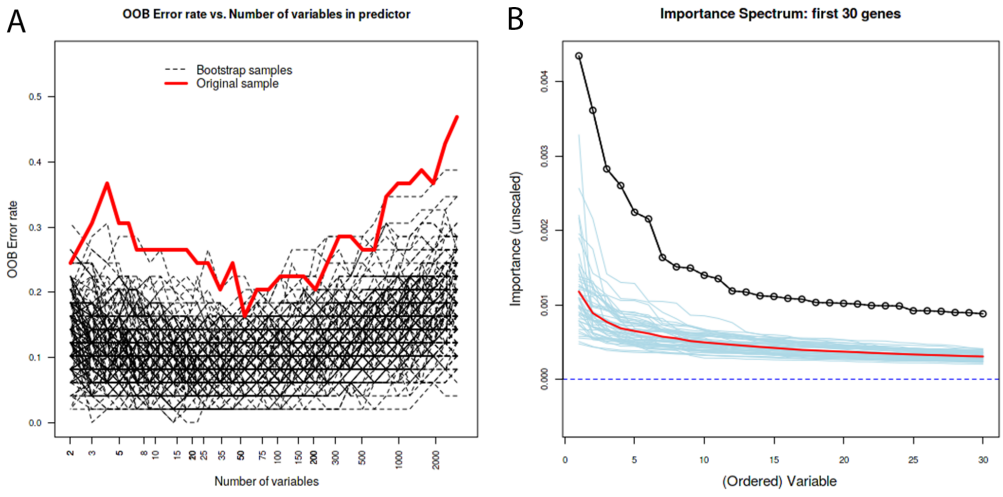
Γ. Αποτελέσματα

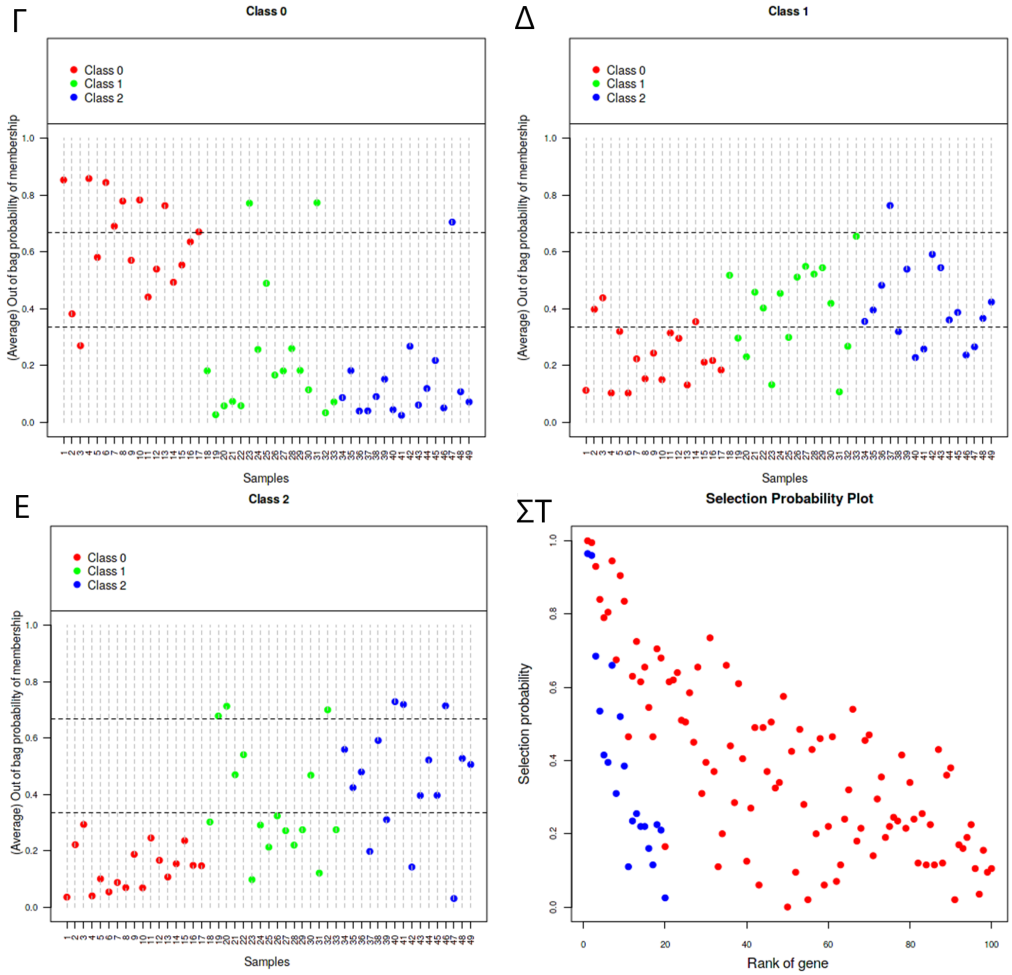
Στα αποτελέσματα του εργαλείου παρουσιάζονται με διαγράμματα οι εξής πληροφορίες:

- Σφάλμα Out-of-Bag σε συνάρτηση με τον αριθμό των γονιδίων που χρησιμοποιήθηκαν (OOB error vs. num of genes).
- Μέση τιμή των προβλέψεων της κλάσης κάθε δείγματος όταν δεν αποτελούσε μέρος του εκπαιδευτικού συνόλου (OOB predictions).
- Σημαντικότητα των γονιδίων στα αρχικά δεδομένα σε σχέση με τη σημαντικότητα των γονιδίων σε επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα (Importance Spectrum Plots).
- Πιθανότητα μιας μεταβλητής να επιλεγεί μέσα στα κορυφαία 20 ή 100 πιο σημαντικά γονίδια (Selection probability plot).

Η εκτέλεση του αλγορίθμου με όλα τα δείγματα είχε ως αποτέλεσμα ταξινόμηση των δειγμάτων με πιθανότητα σφάλματος 36,89% (εκτίμηση με τη μέθοδο Bootstrap) ενώ η πιθανότητα σφάλματος με τυχαία ταξινόμηση είναι 65,31%. Ο αλγόριθμος επέλεξε 34 γονίδια ως σημαντικά, τα ακόλουθα (με φθίνουσα σειρά σημαντικότητας): *TRIB3*, *ADM2*, *ASS1*, *GPT2*, *DEXI*, *PHGDH*, *ATF5*, *TERF2*, *PACSIN2*, *SHMT2*, *MOK*, *MIOX*, *FAM129A*, *TNFRSF11A*, *BRD2*, *C17orf107*, *AGA*, *CCDC169*, *WARS*, *PLIN2*, *SNAP47*, *CLDN7*, *FAM83A*, *ATP6V1E2*, *TUBG2*, *LOC653513*, *OPLAH*, *CYGB*, *FBXO22*, *TMPPE*, *GRTP1*, *ZNF609*, *ID1*, *FDXACB1*. Η πιθανότητα σωστής ταξινόμησης ενός υγιούς δείγματος είναι πολύ μεγαλύτερη από την πιθανότητα σωστής ταξινόμησης ενός δείγματος ΤΙ ή ΤΜ, όπως φαίνεται από τα διαγράμματα Γ, Δ, Ε της Εικόνας 8.

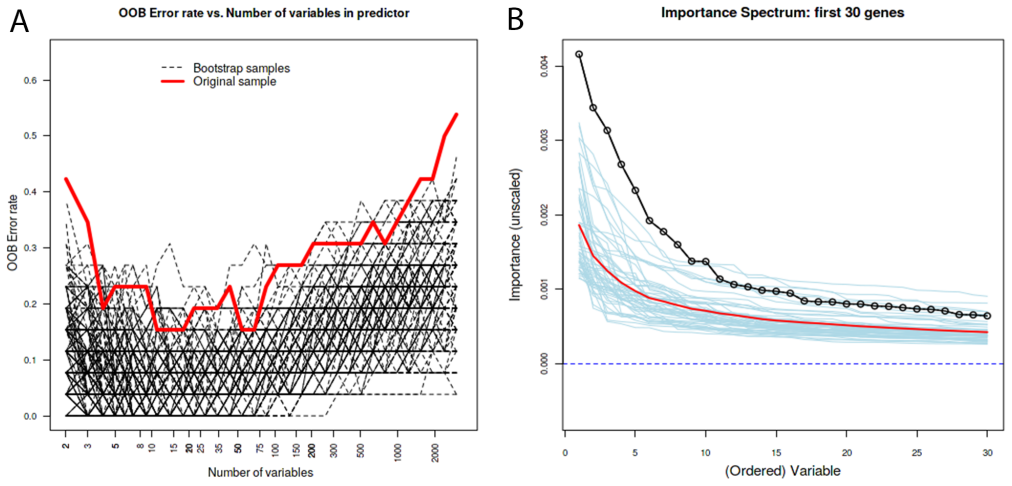
Σύγκριση 1: Υγιή vs ΤΙ vs ΤΜ – Όλα τα δείγματα

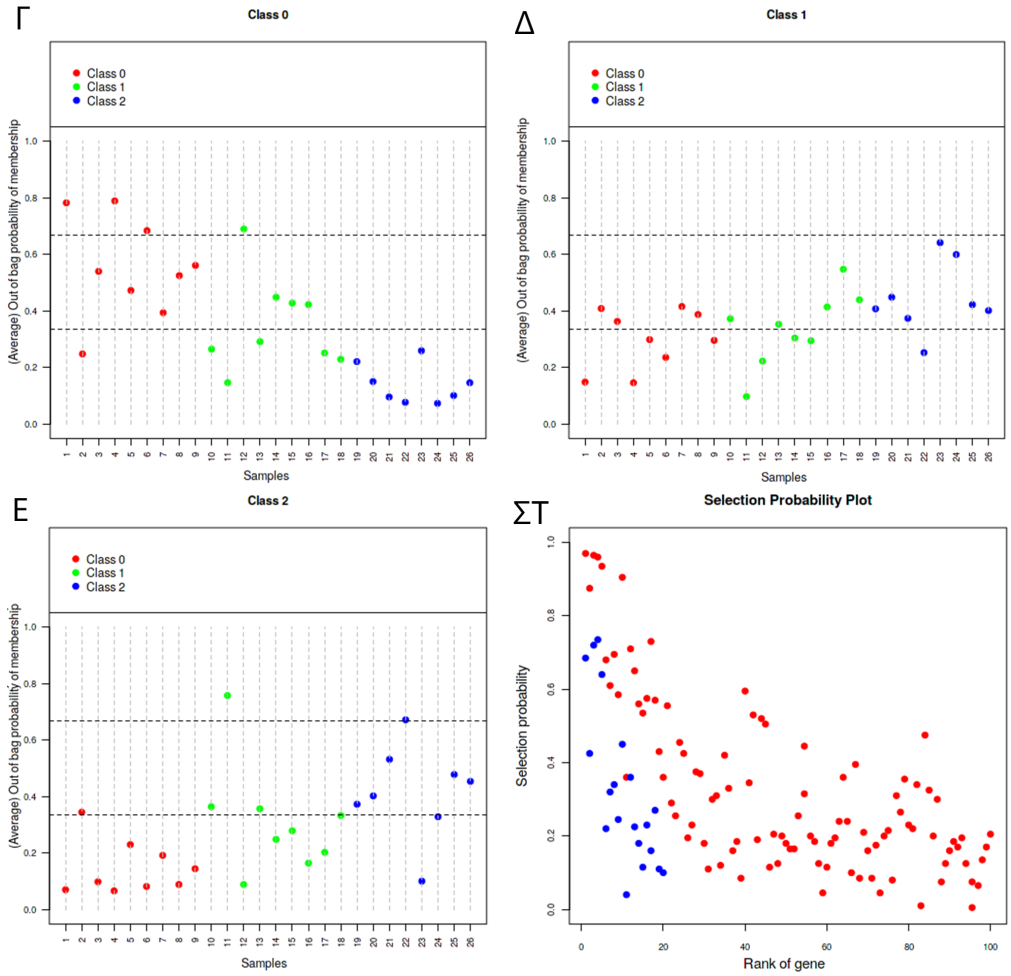




Εικόνα 8. *Σύγκριση δειγμάτων Η vs ΤΙ vs ΤΜ. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση 50 γονιδίων. Β) Η σημαντικότητα των γονιδίων είναι μεγαλύτερη στα δεδομένα με πραγματικές ετικέτες σε σχέση με τη σημαντικότητα των γονιδίων σε επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των υγιών ατόμων. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση ΤΙ. Ε) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση ΤΜ. ΣΤ) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Σύγκριση 2: Υγιή vs TI vs TM – Μόνο θηλυκά δείγματα

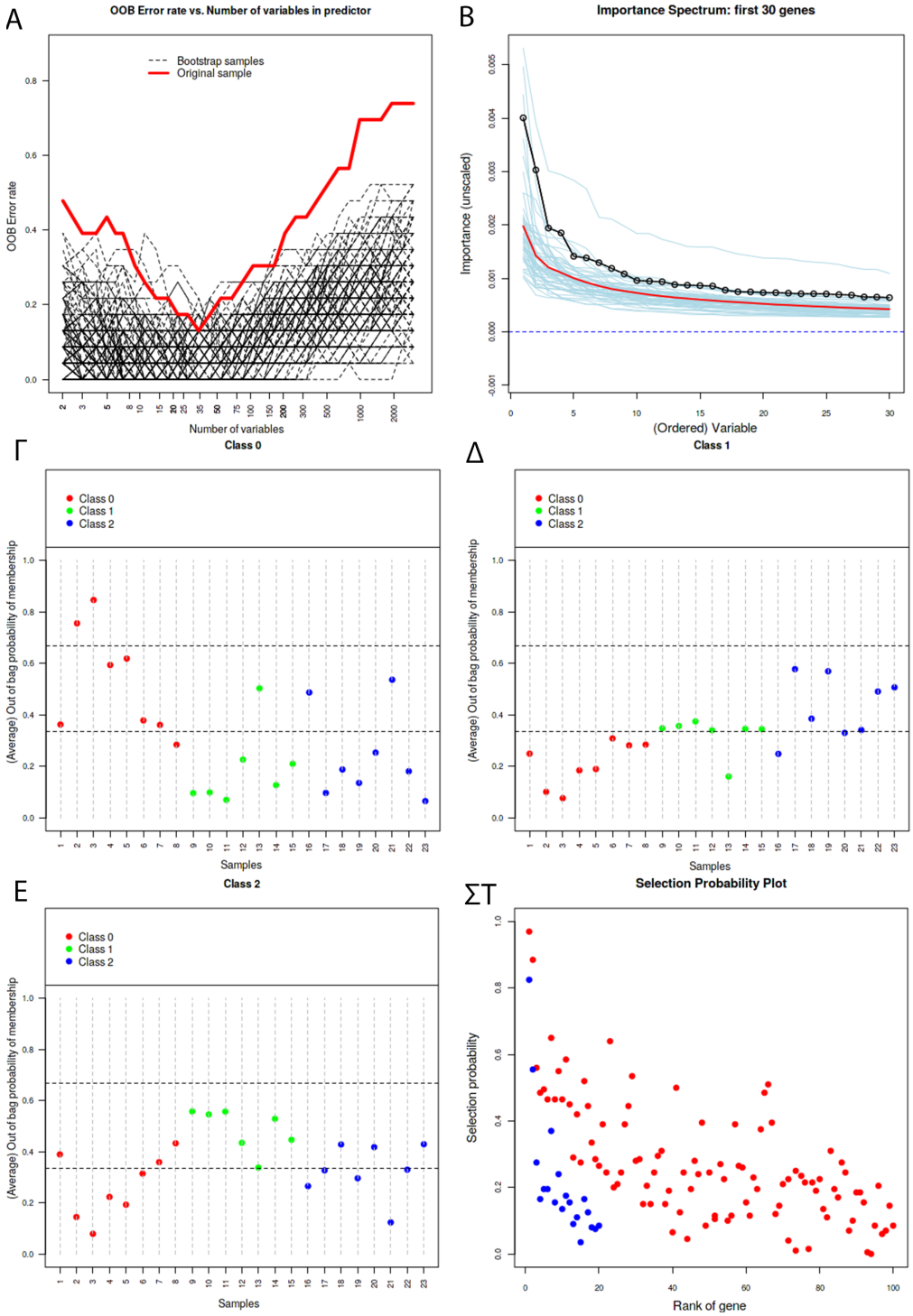




Εικόνα 9. *Σύγκριση θηλυκών δειγμάτων Η vs ΤΙ vs ΤΜ. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση 10 γονιδίων. Β) Η σημαντικότητα των πρώτων 8 γονιδίων είναι αυξημένη στα δεδομένα με πραγματικές ετικέτες σε σχέση με τη σημαντικότητα των γονιδίων σε επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των υγιών ατόμων. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση ΤΙ. Ε) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση ΤΜ. ΣΤ) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Στη σύγκριση όλων των θηλυκών δειγμάτων (Εικόνα 9) το ποσοστό σφάλματος ήταν 46,57%, ενώ η πιθανότητα σφάλματος με τυχαία ταξινόμηση είναι 65,39%. Τα 4 επιλεγμένα γονίδια με φθίνουσα σειρά σημαντικότητας ήταν τα: *FAM83A*, *MIOX*, *ASS1*, *TRIB3*.

Σύγκριση 3: Υγιή vs TI vs TM – Μόνο αρσενικά δείγματα

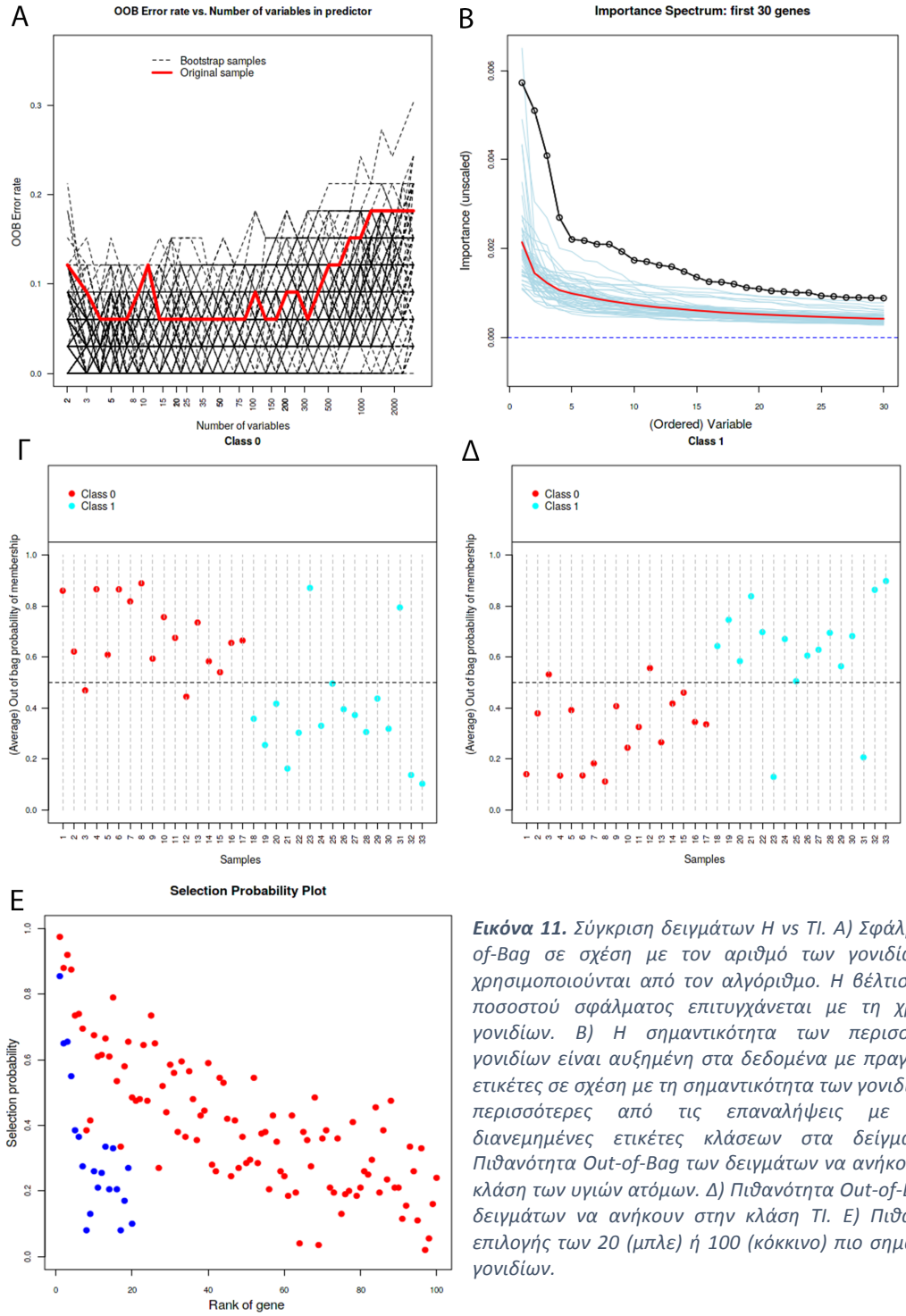


Εικόνα 10. *Σύγκριση αρσενικών δειγμάτων Η vs ΤΙ vs ΤΜ. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση περίπου 32 γονιδίων. Β) Η σημαντικότητα των πρώτων 5 γονιδίων είναι αυξημένη στα δεδομένα με πραγματικές ετικέτες σε σχέση με τη σημαντικότητα των γονιδίων στις περισσότερες από τις επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των υγιών ατόμων. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση ΤΙ. Ε) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση ΤΜ. ΣΤ) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Στη σύγκριση όλων των αρσενικών δειγμάτων (Εικόνα 10) το ποσοστό σφάλματος ήταν 56,08% ενώ το τυχαίο σφάλμα είναι 65,22%. Τα επιλεγμένα γονίδια είναι 22:

LOC100507634, *CYGB*, *IL10RA*, *ATF5*, *TRIB3*, *TERF2*, *MOK*, *ADM2*, *PPP1R14C*, *WARS*, *LINC00936*, *LINC00883*, *TBCB*, *MOCOS*, *GPX4*, *GPT2*, *WDR25*, *FAM129A*, *SFXN1*, *RPL23AP53*, *KMO*, *SERPINE2*.

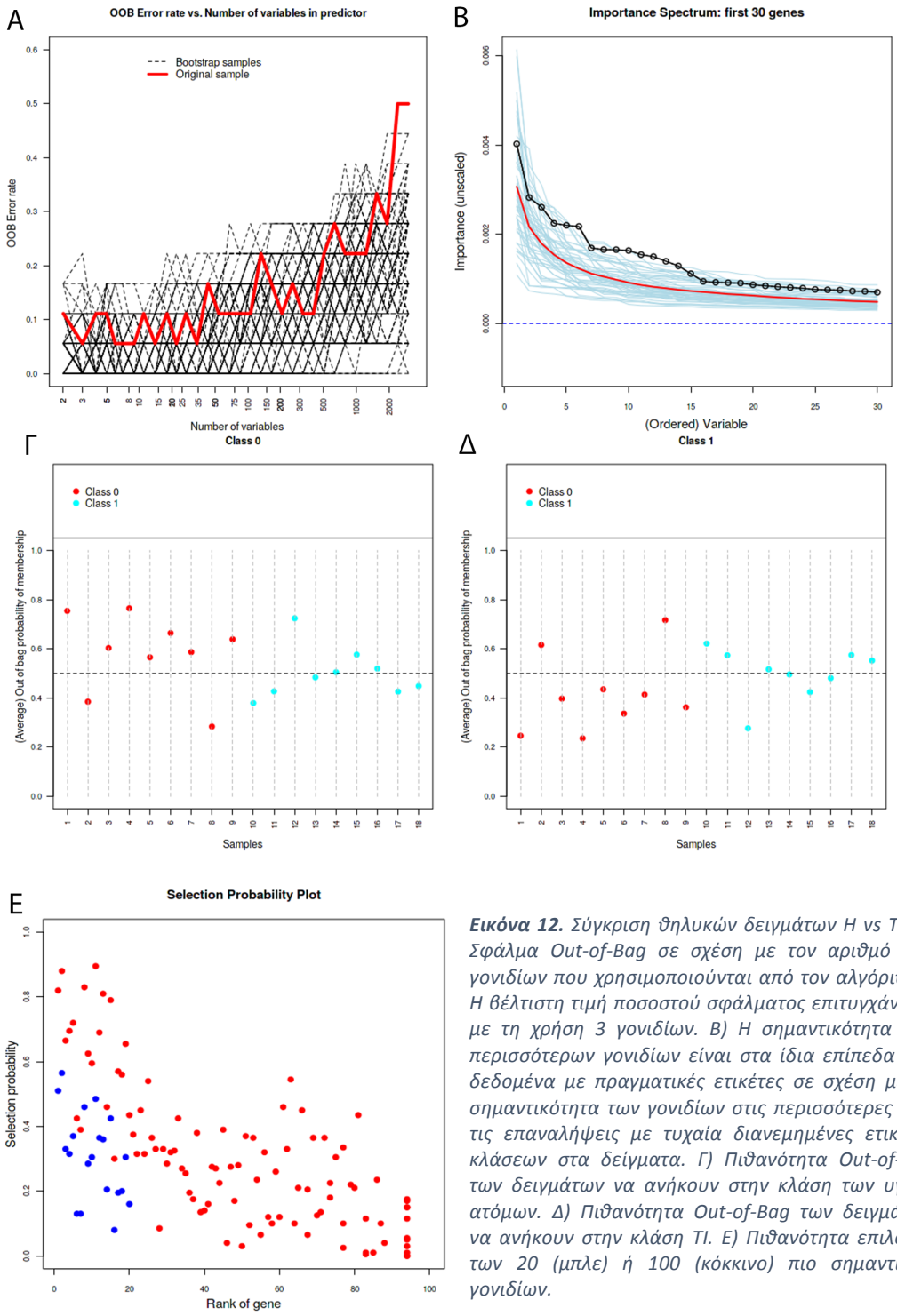
Σύγκριση 4: Υγιή vs TI – Όλα τα δείγματα



Εικόνα 11. *Σύγκριση δειγμάτων Η vs ΤΙ. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση 4 γονιδίων. Β) Η σημαντικότητα των περισσότερων γονιδίων είναι αυξημένη στα δεδομένα με πραγματικές ετικέτες σε σχέση με τη σημαντικότητα των γονιδίων στις περισσότερες από τις επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των υγιών ατόμων. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση ΤΙ. Ε) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Κατά τη δυαδική κατηγοριοποίηση όλων των δειγμάτων Η και ΤΙ (Εικόνα 11), το ποσοστό σφάλματος ήταν 24,00% και η πιθανότητα τυχαίου σφάλματος είναι 48,49%. Τα επιλεγμένα γονίδια είναι 3: *ADM2*, *DEXI*, *ASS1*.

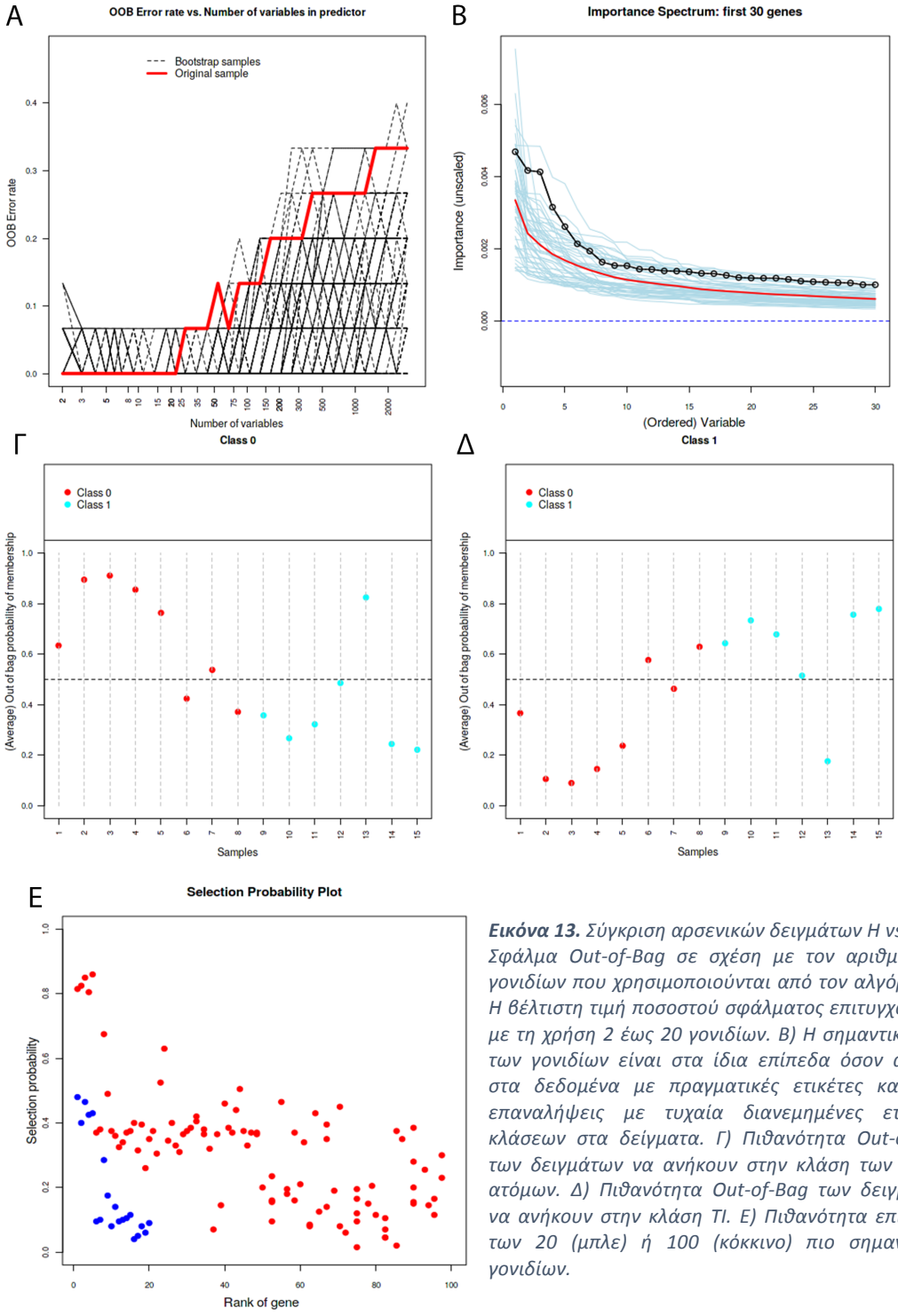
Σύγκριση 5: Υγιή vs TI – Μόνο θηλυκά δείγματα



Εικόνα 12. *Σύγκριση θηλυκών δειγμάτων H vs TI. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση 3 γονιδίων. Β) Η σημαντικότητα των περισσότερων γονιδίων είναι στα ίδια επίπεδα στα δεδομένα με πραγματικές ετικέτες σε σχέση με τη σημαντικότητα των γονιδίων στις περισσότερες από τις επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των υγιών ατόμων. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση TI. Ε) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Στην ταξινόμηση μόνο των θηλυκών H vs TI δειγμάτων (Εικόνα 12), το ποσοστό σφάλματος υπολογίστηκε στο 41,61%, ενώ η πιθανότητα σφάλματος με τυχαία ταξινόμηση των δειγμάτων είναι 50,00%. Τα 3 γονίδια που επιλέχθηκαν από τον αλγόριθμο είναι: *ASS1*, *ADM2*, *TCAF2*.

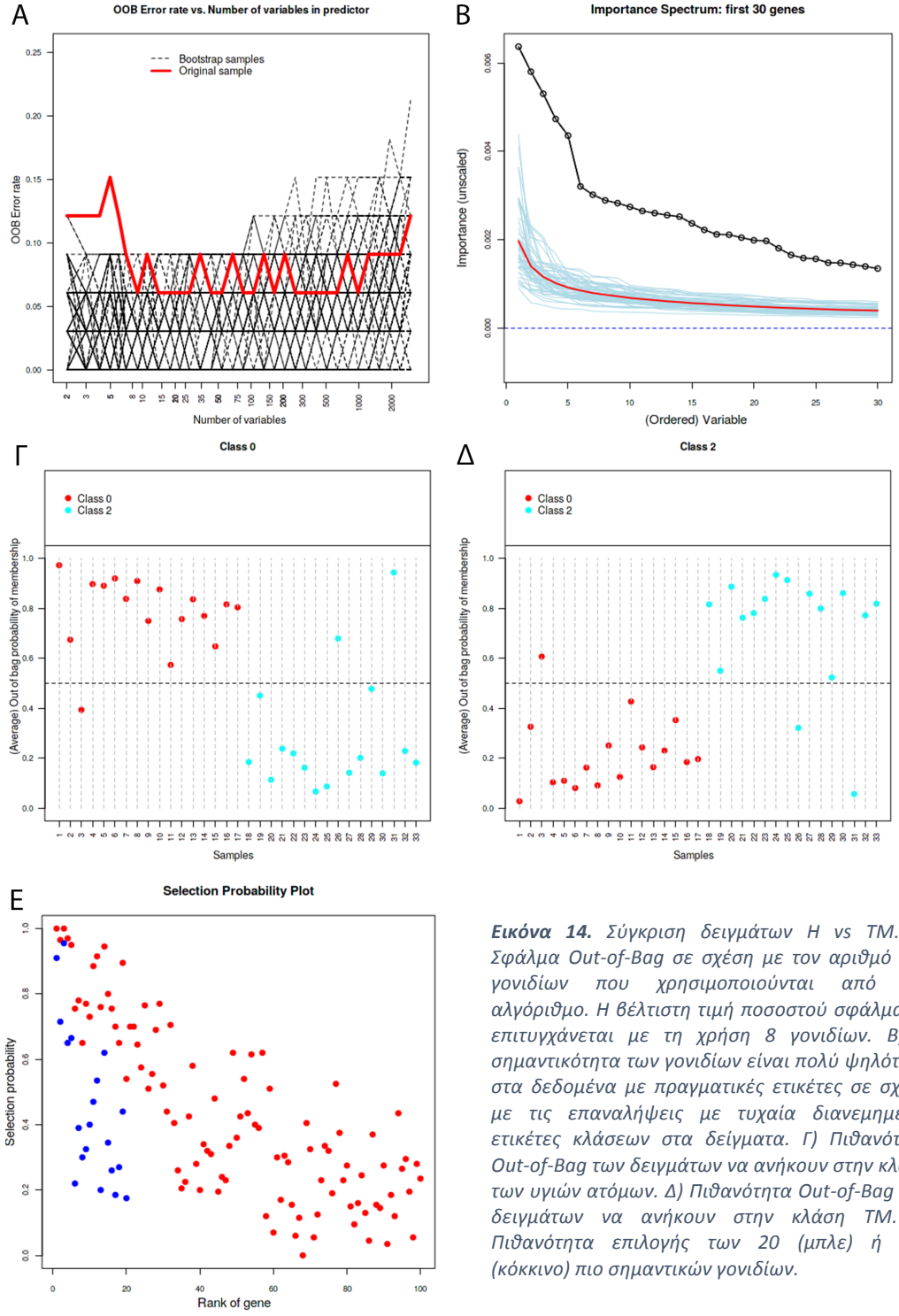
Σύγκριση 6: Υγιή vs TI – Μόνο αρσενικά δείγματα



Εικόνα 13. *Σύγκριση αρσενικών δειγμάτων H vs TI. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση 2 έως 20 γονιδίων. Β) Η σημαντικότητα των γονιδίων είναι στα ίδια επίπεδα όσον αφορά στα δεδομένα με πραγματικές ετικέτες και στις επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των υγιών ατόμων. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση TI. Ε) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Στην σύγκριση μόνο των αρσενικών H vs TI δειγμάτων (Εικόνα 13), το ποσοστό σφάλματος υπολογίστηκε στο 28,11%, ενώ η πιθανότητα σφάλματος με τυχαία ταξινόμηση των δειγμάτων είναι 46,67%. Τα γονίδια που επιλέχθηκαν από τον αλγόριθμο είναι 2: *ZCRB1*, *VIMP*.

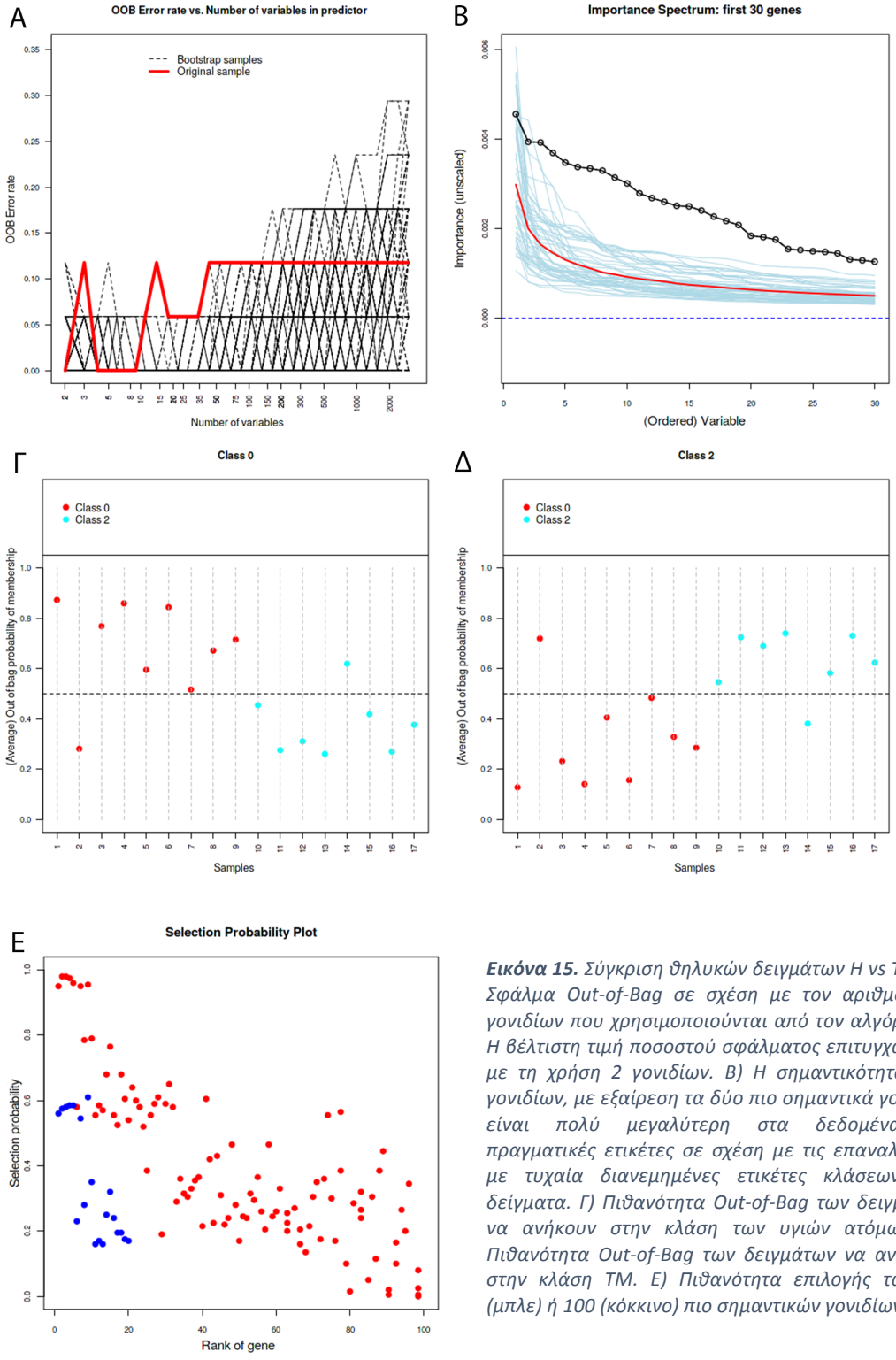
Σύγκριση 7: Υγιή vs ΤΜ – Όλα τα δείγματα



Εικόνα 14. *Σύγκριση δειγμάτων Η vs ΤΜ. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση 8 γονιδίων. Β) Η σημαντικότητα των γονιδίων είναι πολύ ψηλότερα στα δεδομένα με πραγματικές ετικέτες σε σχέση με τις επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των υγιών ατόμων. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση ΤΜ. Ε) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Η ταξινόμηση των δειγμάτων Η vs ΤΜ (Εικόνα 14) είχε πολύ χαμηλό ποσοστό σφάλματος 15,49%. Το σφάλμα με τυχαία ταξινόμηση είναι 48,49%. Ο αλγόριθμος επέλεξε 7 γονίδια, τα παρακάτω: ADM2, *GPT2*, *TRIB3*, *PHGDH*, *PACSIN2*, *WARS*, *GRTP1*.

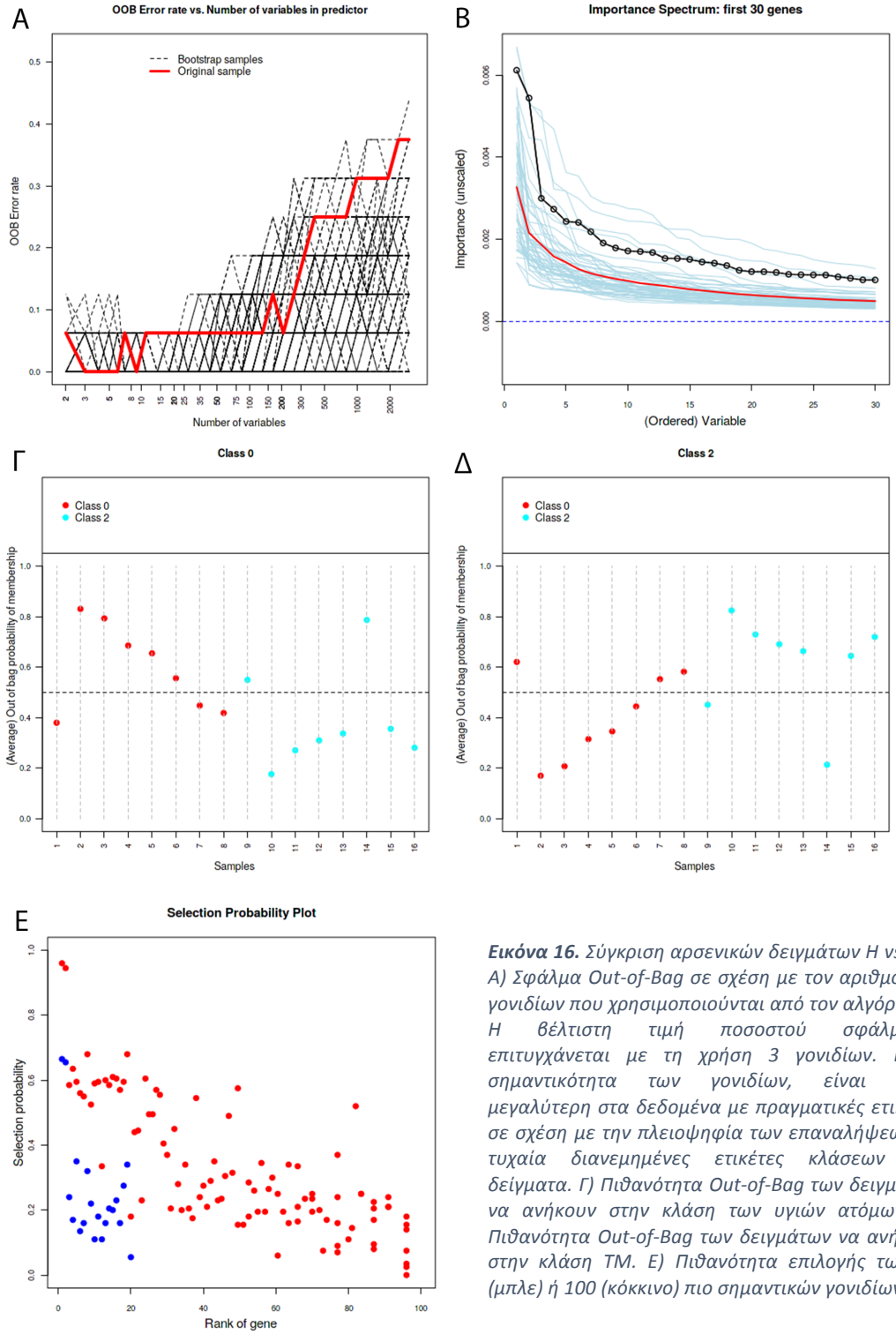
Σύγκριση 8: Υγιή vs ΤΜ – Μόνο θηλυκά δείγματα



Εικόνα 15. *Σύγκριση θηλυκών δειγμάτων Η vs ΤΜ. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση 2 γονιδίων. Β) Η σημαντικότητα των γονιδίων, με εξαίρεση τα δύο πιο σημαντικά γονίδια, είναι πολύ μεγαλύτερη στα δεδομένα με πραγματικές ετικέτες σε σχέση με τις επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των υγιών ατόμων. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση ΤΜ. Ε) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Το ποσοστό σφάλματος στη σύγκριση θηλυκών Η vs ΤΜ δειγμάτων (Εικόνα 15) ήταν 25,77% και το τυχαίο σφάλμα είναι 47,06%. Τα επιλεγμένα γονίδια είναι 2: *THEMIS2*, *FAM83A*.

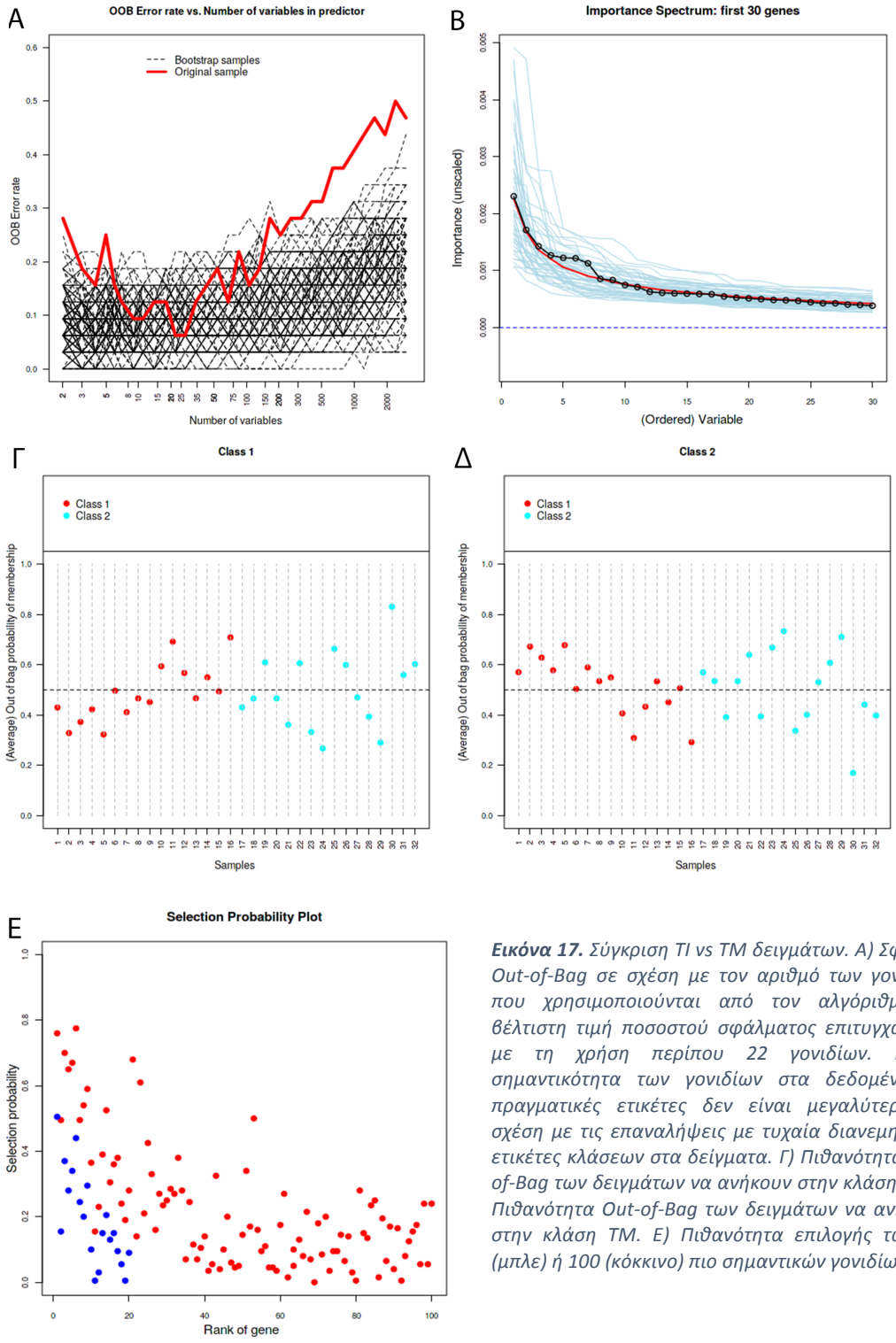
Σύγκριση 9: Υγιή vs ΤΜ – Μόνο αρσενικά δείγματα



Εικόνα 16. *Σύγκριση αρσενικών δειγμάτων Η vs ΤΜ. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση 3 γονιδίων. Β) Η σημαντικότητα των γονιδίων, είναι πολύ μεγαλύτερη στα δεδομένα με πραγματικές ετικέτες σε σχέση με την πλειοψηφία των επαναλήψεων με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των υγιών ατόμων. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση ΤΜ. Ε) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Το ποσοστό σφάλματος της σύγκρισης αρσενικών Η vs ΤΜ δειγμάτων (Εικόνα 16) ήταν 32,16% και το τυχαίο σφάλμα είναι 50%. Τα γονίδια που επιλέχθηκαν από τον αλγόριθμο ήταν 3: *CYGB*, *LOC100507634*, *SNAP47*.

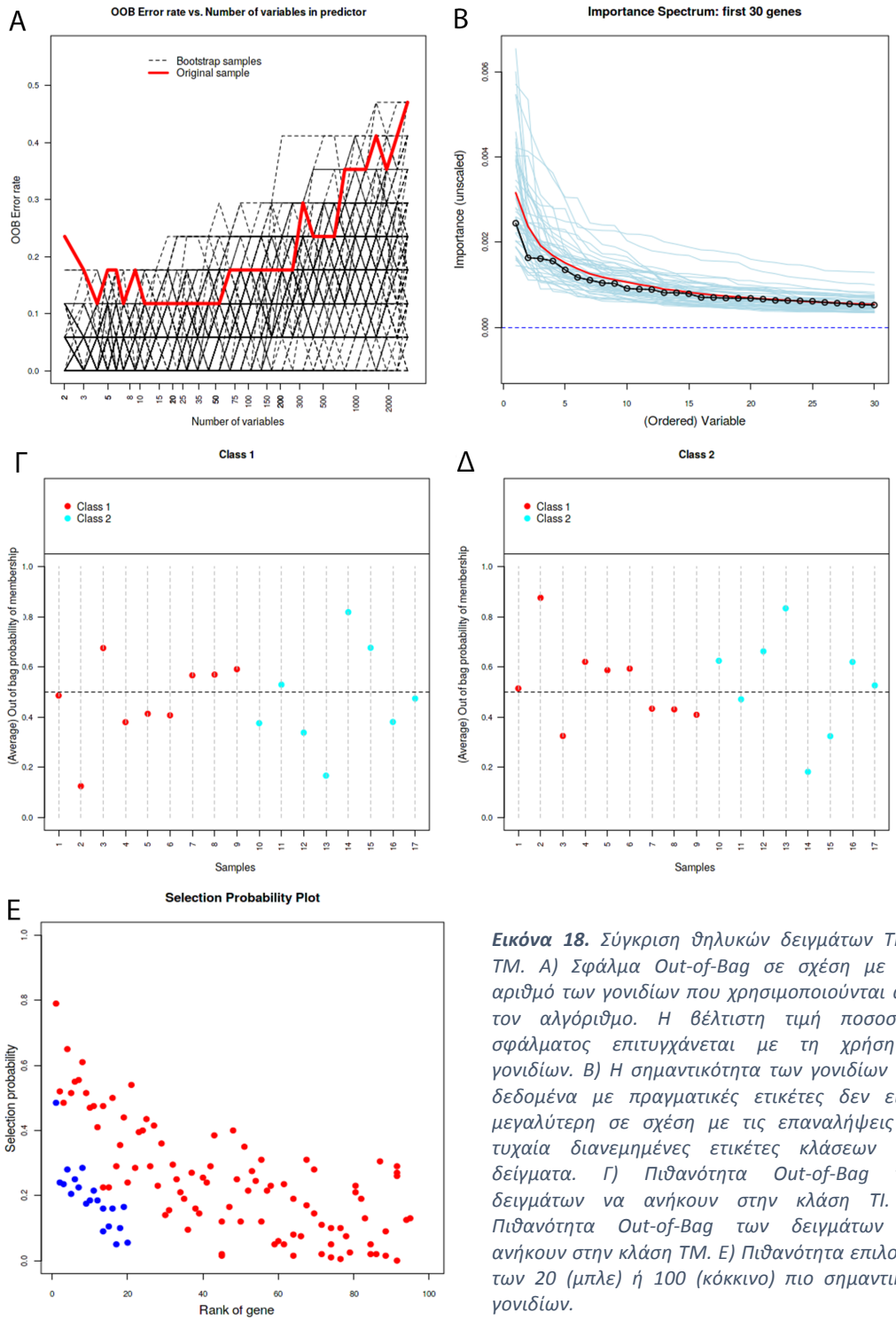
Σύγκριση 10: TI vs TM – Όλα τα δείγματα



Εικόνα 17. *Σύγκριση TI vs TM δειγμάτων. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση περίπου 22 γονιδίων. Β) Η σημαντικότητα των γονιδίων στα δεδομένα με πραγματικές ετικέτες δεν είναι μεγαλύτερη σε σχέση με τις επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση TI. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση TM. Ε) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Στη σύγκριση TI vs TM δειγμάτων (Εικόνα 17) το σφάλμα εκτιμήθηκε στο 50,34%, ενώ η πιθανότητα σφάλματος με τυχαία ταξινόμηση είναι 50%. Τα επιλεγμένα γονίδια είναι 9: *LINC00883*, *FBXO22*, *PHLDA3*, *FASTKD2*, *CCDC169*, *BRD2*, *KMO*, *PACSIN2*, *LOC101928034*.

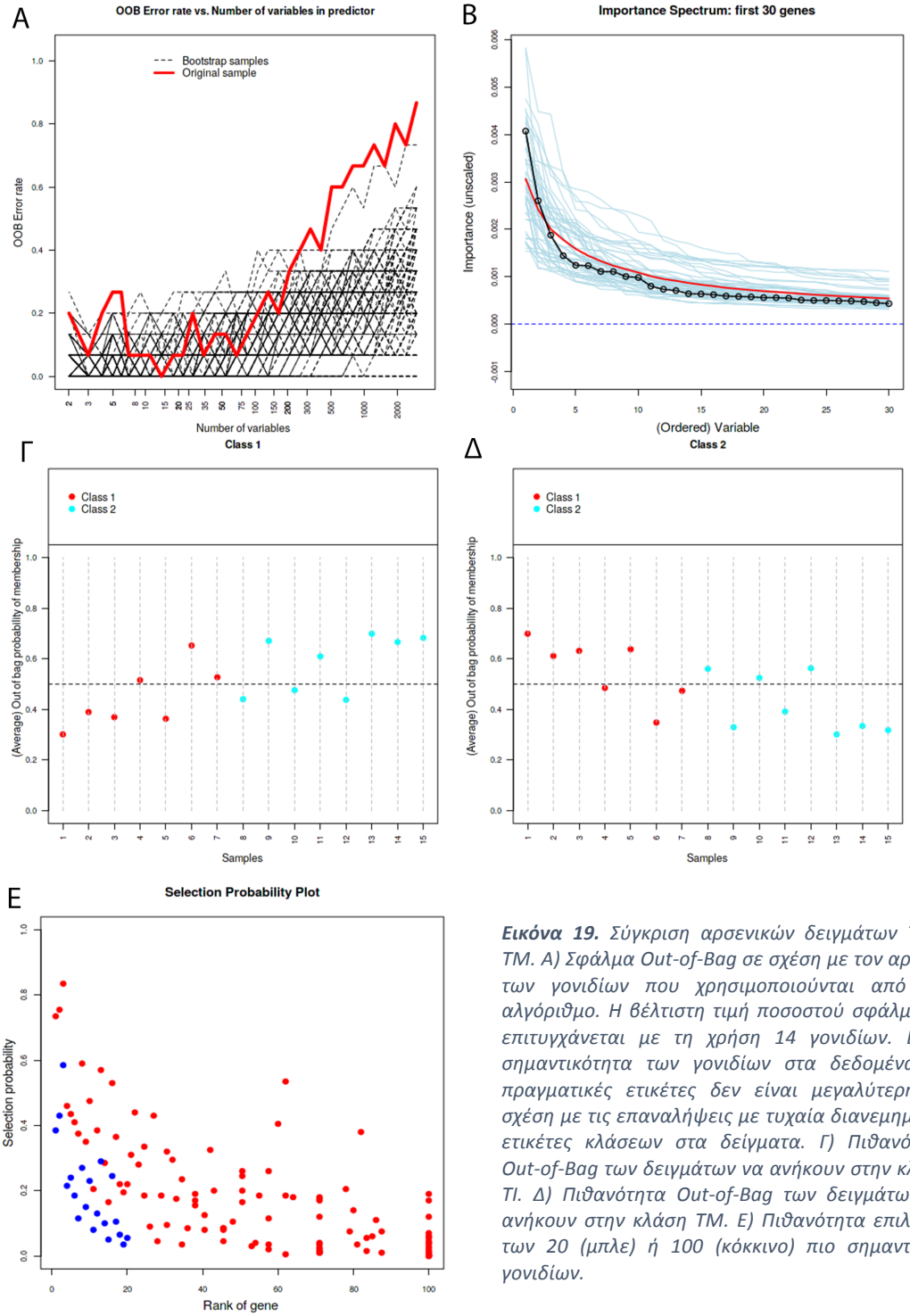
Σύγκριση 11: TI vs TM – Μόνο θηλυκά δείγματα



Εικόνα 18. *Σύγκριση θηλυκών δειγμάτων TI vs TM. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση 4 γονιδίων. Β) Η σημαντικότητα των γονιδίων στα δεδομένα με πραγματικές ετικέτες δεν είναι μεγαλύτερη σε σχέση με τις επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση TI. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση TM. Ε) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Η ταξινόμηση των θηλυκών δειγμάτων TI vs TM (Εικόνα 18) είχε ποσοστό σφάλματος 50,16% και η πιθανότητα σφάλματος τυχαία ταξινόμησης είναι 47,06%. Τα 3 επιλεγμένα γονίδια είναι: *FAM210A*, *SBF2-AS1*, *ZHX2*.

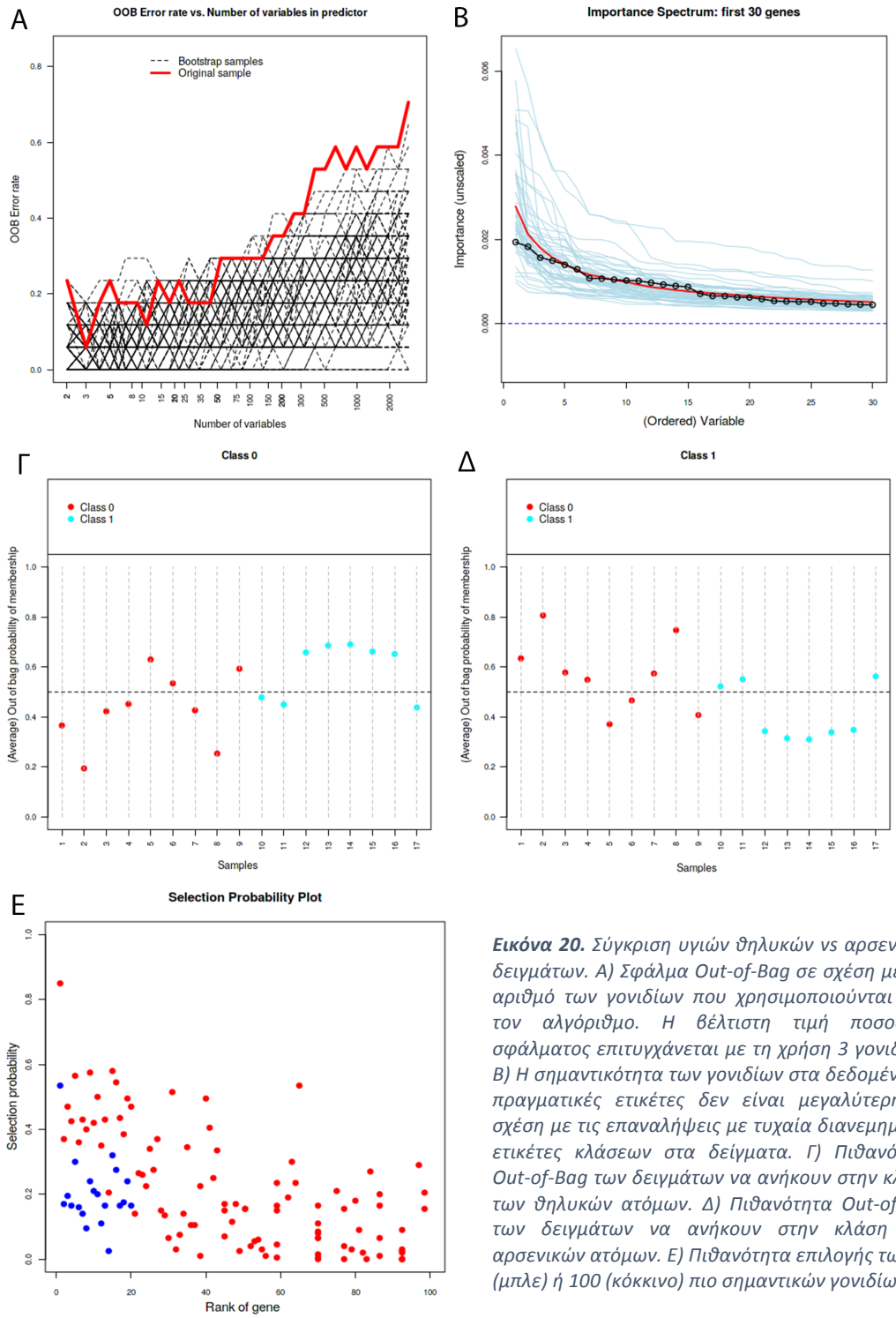
Σύγκριση 12: TI vs TM – Μόνο αρσενικά δείγματα



Εικόνα 19. *Σύγκριση αρσενικών δειγμάτων TI vs TM. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση 14 γονιδίων. Β) Η σημαντικότητα των γονιδίων στα δεδομένα με πραγματικές ετικέτες δεν είναι μεγαλύτερη σε σχέση με τις επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση TI. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση TM. Ε) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Το ποσοστό σφάλματος της ταξινόμησης των αρσενικών TI vs TM δειγμάτων (Εικόνα 19) ήταν 55,63% ενώ το τυχαίο σφάλμα είναι 46,67%. Τα επιλεγμένα γονίδια είναι 14: *PPP1R14C*, *LINC00936*, *MIR7113*, *LINC00883*, *LYNX1*, *MFSD4*, *MCM8*, *TLE3*, *CDC42EP2*, *ANO10*, *CIART*, *ALKBH7*, *KMO*, *FABP4*.

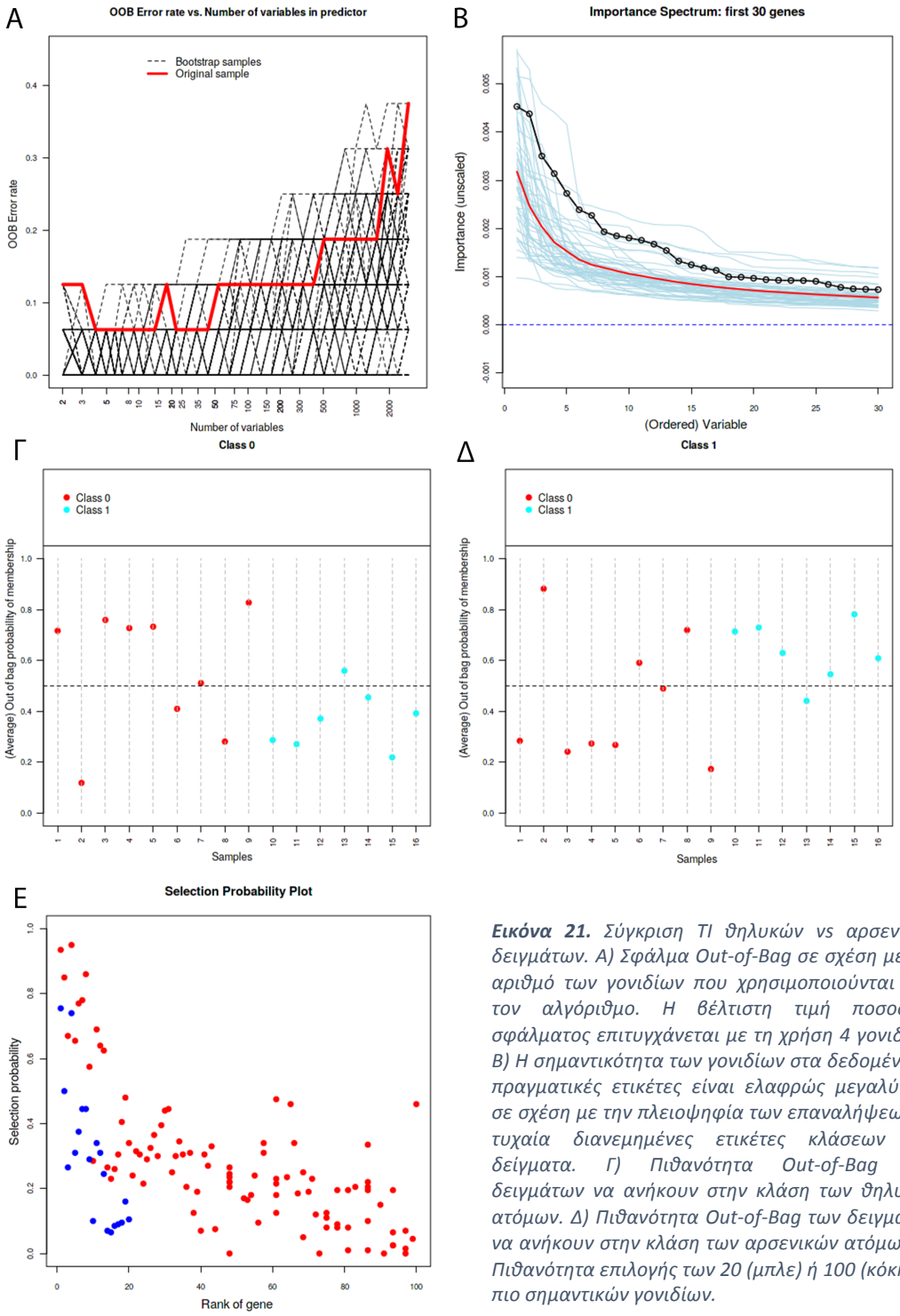
Σύγκριση 13: Θηλυκά vs Αρσενικά – Υγιή Δείγματα



Εικόνα 20. *Σύγκριση υγιών θηλυκών vs αρσενικών δειγμάτων. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση 3 γονιδίων. Β) Η σημαντικότητα των γονιδίων στα δεδομένα με πραγματικές ετικέτες δεν είναι μεγαλύτερη σε σχέση με τις επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των θηλυκών ατόμων. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των αρσενικών ατόμων. Ε) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Στη σύγκριση υγιών θηλυκών vs αρσενικών δειγμάτων (Εικόνα 20) το ποσοστό σφάλματος ήταν 56,57%. Το ποσοστό σφάλματος με τυχαία ταξινόμηση είναι 47,06%. Τα 3 γονίδια που επέλεξε ο αλγόριθμος είναι: *SGMS1-AS1*, *SCML1*, *LOC100507462*.

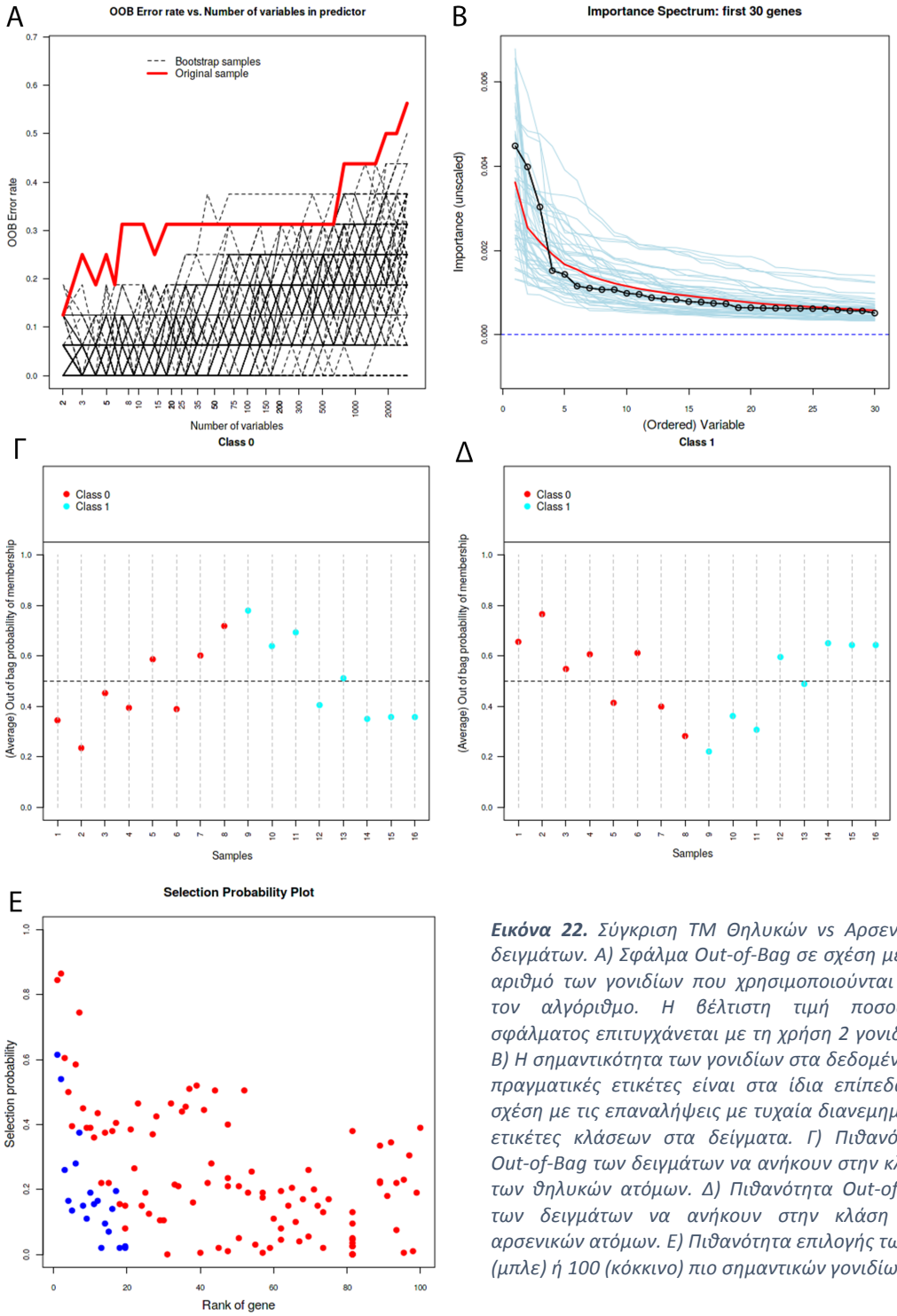
Σύγκριση 14: Θηλυκά vs Αρσενικά – ΤΙ Δείγματα



Εικόνα 21. *Σύγκριση ΤΙ θηλυκών vs αρσενικών δειγμάτων. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση 4 γονιδίων. Β) Η σημαντικότητα των γονιδίων στα δεδομένα με πραγματικές ετικέτες είναι ελαφρώς μεγαλύτερη σε σχέση με την πλειοψηφία των επαναλήψεων με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των θηλυκών ατόμων. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των αρσενικών ατόμων. Ε) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Η ταξινόμηση των δειγμάτων ΤΙ σε αρσενικά και θηλυκά (Εικόνα 21) είχε ποσοστό σφάλματος 33,13% ενώ το ποσοστό σφάλματος με τυχαία ταξινόμηση είναι 43,75%. Τα επιλεγμένα γονίδια είναι 4: *CA5BP1*, *ZCRB1*, *FBXO22*, *PDHX*.

Σύγκριση 15: Θηλυκά vs Αρσενικά – TM Δείγματα



Εικόνα 22. *Σύγκριση TM Θηλυκών vs Αρσενικών δειγμάτων. Α) Σφάλμα Out-of-Bag σε σχέση με τον αριθμό των γονιδίων που χρησιμοποιούνται από τον αλγόριθμο. Η βέλτιστη τιμή ποσοστού σφάλματος επιτυγχάνεται με τη χρήση 2 γονιδίων. Β) Η σημαντικότητα των γονιδίων στα δεδομένα με πραγματικές ετικέτες είναι στα ίδια επίπεδα σε σχέση με τις επαναλήψεις με τυχαία διανεμημένες ετικέτες κλάσεων στα δείγματα. Γ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των θηλυκών ατόμων. Δ) Πιθανότητα Out-of-Bag των δειγμάτων να ανήκουν στην κλάση των αρσενικών ατόμων. Ε) Πιθανότητα επιλογής των 20 (μπλε) ή 100 (κόκκινο) πιο σημαντικών γονιδίων.*

Η σύγκριση TM θηλυκών vs αρσενικών δειγμάτων (Εικόνα 22) είχε ποσοστό σφάλματος 51,74% ενώ το ποσοστό σφάλματος με τυχαία ταξινόμηση είναι 50%. Τα επιλεγμένα γονίδια είναι 2: *EIF1AX*, *TBC1D16*.

Τα αποτελέσματα όλων των συγκρίσεων συνοψίζονται στον Πίνακα 1.

Σύγκριση	**Σφάλμα Bootstrap**	**Τυχαίο Σφάλμα**	**Γονίδια με φθίνουσα σειρά σημαντικότητας**
Η vs ΤΙ vs ΤΜ – Όλα	36,89%	65,31%	34 γονίδια: *TRIB3*, *ADM2*, *ASS1*, *GPT2*, *DEXI*, *PHGDH*, *ATF5*, *TERF2*, *PACSIN2*, *SHMT2*, *MOK*, *MIOX*, *FAM129A*, *TNFRSF11A*, *BRD2*, *C17orf107*, *AGA*, *CCDC169*, *WARS*, *PLIN2*, *SNAP47*, *CLDN7*, *FAM83A*, *ATP6V1E2*, *TUBG2*, *LOC653513*, *OPLAH*, *CYGB*, *FBXO22*, *TMPPE*, *GRTP1*, *ZNF609*, *ID1*, *FDXACB1*
Η vs ΤΙ vs ΤΜ – Θηλυκά	46,57%	65,39%	4 γονίδια: *FAM83A*, *MIOX*, *ASS1*, *TRIB3*
Η vs ΤΙ vs ΤΜ – Αρσενικά	56,08%	65,22%	22 γονίδια: *LOC100507634*, *CYGB*, *IL10RA*, *ATF5*, *TRIB3*, *TERF2*, *MOK*, *ADM2*, *PPP1R14C*, *WARS*, *LINC00936*, *LINC00883*, *TBCB*, *MOCOS*, *GPX4*, *GPT2*, *WDR25*, *FAM129A*, *SFXN1*, *RPL23AP53*, *KMO*, *SERPINE2*
Η vs ΤΙ – Όλα	24,00%	48,49%	3 γονίδια: *ADM2*, *DEXI*, *ASS1*
Η vs ΤΙ – Θηλυκά	41,61%	50,00%	3 γονίδια: *ASS1*, *ADM2*, *TCAF2*
Η vs ΤΙ – Αρσενικά	28,11%	46,67%	2 γονίδια: *ZCRB1*, *VIMP*
Η vs ΤΜ – Όλα	15,49%	48,49%	7 γονίδια: ADM2, *GPT2*, *TRIB3*, *PHGDH*, *PACSIN2*, *WARS*, *GRTP1*
Η vs ΤΜ – Θηλυκά	25,77%	47,06%	2 γονίδια: *THEMIS2*, *FAM83A*
Η vs ΤΜ – Αρσενικά	32,16%	50%	3 γονίδια: *CYGB*, *LOC100507634*, *SNAP47*
ΤΙ vs ΤΜ – Όλα	50,34%	50%	9 γονίδια: *LINC00883*, *FBXO22*, *PHLDA3*, *FASTKD2*, *CCDC169*, *BRD2*, *KMO*, *PACSIN2*, *LOC101928034*
ΤΙ vs ΤΜ – Θηλυκά	50,16%	47,06%	3 γονίδια: *FAM210A*, *SBF2-AS1*, *ZHX2*
ΤΙ vs ΤΜ – Αρσενικά	55,63%	46,67%	14 γονίδια: *PPP1R14C*, *LINC00936*, *MIR7113*, *LINC00883*, *LYNX1*, *MFSD4*, *MCM8*, *TLE3*, *CDC42EP2*, *ANO10*, *CIART*, *ALKBH7*, *KMO*, *FABP4*
Θηλυκά vs Αρσενικά – Η	56,57%	47,06%	3 γονίδια: *SGMS1-AS1*, *SCML1*, *LOC100507462*
Θηλυκά vs Αρσενικά – ΤΙ	33,13%	43,75%	4 γονίδια: *CA5BP1*, *ZCRB1*, *FBXO22*, *PDHX*
Θηλυκά vs Αρσενικά – ΤΜ	51,74%	50%	2 γονίδια: *EIF1AX*, *TBC1D16*

Πίνακας 1. *Σύνοψη των αποτελεσμάτων των συγκρίσεων με το εργαλείο GeneSrF.*

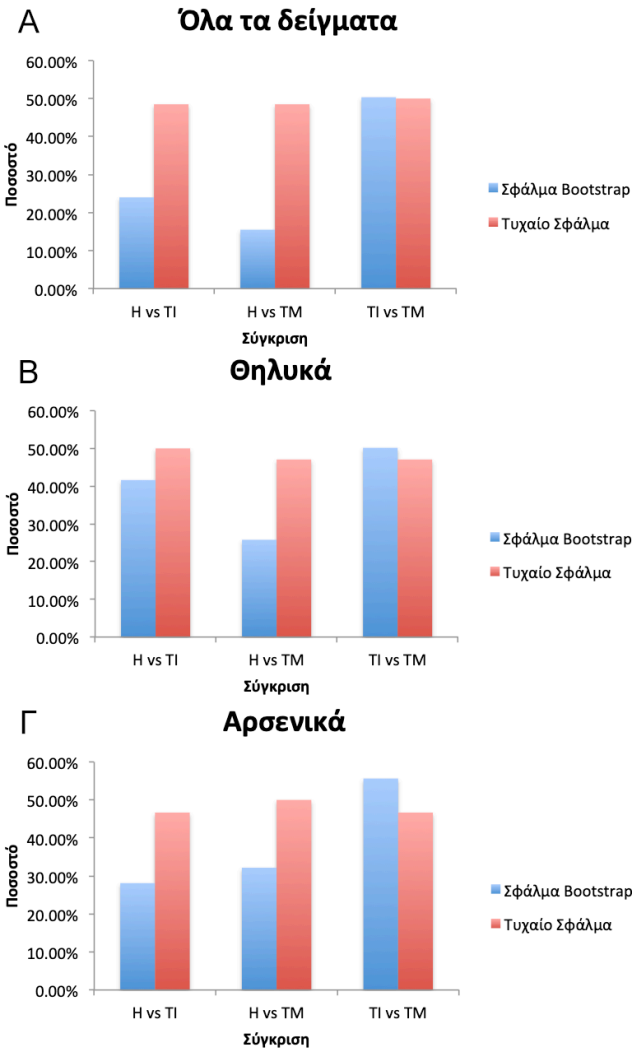
Δ. Συζήτηση

Η β-θαλασσαιμία είναι ένα και σοβαρό κληρονομικό νόσημα παγκοσμίως, αλλά και στη χώρα μας. Είναι μονογονιδιακή νόσος και έχει αυτοσωμική υπολειπόμενη κληρονόμηση. Τα άτομα με θαλασσαιμία έχουν μειωμένη έως μηδενική παραγωγή σφαιρινών και παρουσιάζουν σοβαρές παθολογικές επιπλοκές. Η αντιμετώπισή της γίνεται με μεταγγίσεις και αποσιδήρωση, μεταμόσχευση μυελού και συμπτωματική αντιμετώπιση των επιπλοκών. Η διάκριση μεταξύ της μείζονος και της ενδιάμεσης μορφής με μοριακό τρόπο είναι σημαντική για την αποφυγή των μη απαραίτητων μεταγγίσεων και των επακόλουθων επιπλοκών στα άτομα με ενδιάμεσο φαινότυπο και την έγκαιρη έναρξη των μεταγγίσεων σε άτομα με μείζων θαλασσαιμία.

Για την εύρεση προβλεπτικών βιοδεικτών έχει αποτελέσει δημοφιλή επιλογή η αλληλούχιση του μεταγραφώματος (van Rensburg and Loxton, 2015; Akond et al., 2018; van Rheenen et al., 2018) και στη συνέχεια η επεξεργασία των δεδομένων με μεθόδους μηχανικής μάθησης (Grissa et al., 2016; Mamoshina et al., 2018; Vafaee et al., 2018). Η μέθοδος μηχανικής μάθησης που επιλέχθηκε στην παρούσα Εργασία, «Τυχαίο Δάσος», είναι από τις πλέον χρησιμοποιούμενες για την επιλογή βιοδεικτών και για ταξινόμηση των ασθενών σε φαινοτυπικές κατηγορίες (Valdes et al., 2016; Lee, 2017).

Στις συγκρίσεις της παρούσας εργασίας, δείκτης της αξιοπιστίας πρόβλεψης ενός εκπαιδευμένου μοντέλου είναι το ποσοστό σφάλματος με τη μέθοδο Bootstrap σε σχέση με το ποσοστό τυχαίας πρόβλεψης. Δείκτη της εμπιστοσύνης στην επιλογή των μεταβλητών αποτέλεσε η συχνότητα εμφάνισης μιας μεταβλητής ως σημαντικού χαρακτηριστικού στις επαναλήψεις του αλγορίθμου. Από τα ποσοστά σφάλματος των συγκρίσεων υγιών ατόμων σε σχέση με τους ασθενείς ΤΙ και σε σχέση με τους ασθενείς ΤΜ φαίνεται πως τα δεδομένα διαφορικής έκφρασης αυξάνουν την αξιοπιστία της πρόβλεψης, αφού τα αντίστοιχα ποσοστά σφάλματος με τυχαία πρόβλεψη είναι πολύ μεγαλύτερα σε όλες τις συγκρίσεις (Εικόνα 23). Ειδικά στη σύγκριση Η vs ΤΜ με όλα τα δείγματα παρατηρείται η μεγαλύτερη βελτίωση αξιοπιστίας με το ποσοστό σφάλματος να είναι 15,49% ενώ το ποσοστό σφάλματος με τυχαία ταξινόμηση είναι 48,49%. Στις συγκρίσεις μεταξύ των ασθενών ΤΙ και ΤΜ, τα ποσοστά σφάλματος (50.34% - όλα τα δείγματα, 50.16% - θηλυκά, 55.63% - αρσενικά) ήταν σε λίγο μεγαλύτερα επίπεδα από ότι το σφάλμα της τυχαίας πρόβλεψης (50%, 47.06%, 46.67%, αντίστοιχα). Άρα ο αλγόριθμος δε μπόρεσε να διακρίνει αν ένα δείγμα είναι ΤΙ ή ΤΜ χρησιμοποιώντας τα δεδομένα που του παρείχαμε. Συνεπώς, τα επιλεγμένα γονίδια στις συγκρίσεις αυτές δεν είναι αξιόπιστα για χρήση ως βιοδείκτες κατηγοριοποίησης μεταξύ των

ΤΙ και ΤΜ ασθενών. Τα αποτελέσματα συμφωνούν με τα δεδομένα από τις αναλύσεις διαφορικής έκφρασης, καθώς στην ανάλυση μεταξύ των δειγμάτων ΤΜ και ΤΙ δεν είχε προσδιοριστεί κανένα στατιστικώς σημαντικό αποτέλεσμα.



Εικόνα 23. *Διαγράμματα ποσοστών σφάλματος. Στα διαγράμματα συγκρίνονται το ποσοστό σφάλματος πρόβλεψης της κλάσης των δειγμάτων (υπολογισμένο με τη μέθοδο Bootstrap) έναντι του ποσοστού σφάλματος με τυχαία πρόβλεψη. Ποσοστά σφάλματος σε συγκρίσεις με: Α) Όλα τα δείγματα, Β) Θηλυκά δείγματα, Γ) Αρσενικά δείγματα.*

Συνοψίζοντας, η προσπάθεια ταξινόμησης των δειγμάτων είχε επιτυχία στις συγκρίσεις μεταξύ των δειγμάτων ΤΜ και Η, ενώ βελτιωμένη αξιοπιστία πρόβλεψης είχαν και οι συγκρίσεις ΤΙ vs Η δειγμάτων. Αναγνωρίστηκαν επίσης γονίδια, όπως τα *ADM2*, *ASS1*, *PHGDH* κ.α. τα οποία θα μελετηθούν περαιτέρω πειραματικά. Στις συγκρίσεις μεταξύ ΤΙ και ΤΜ δειγμάτων η διάκριση των δειγμάτων δεν ήταν εφικτή με τη χρήση των συγκεκριμένων δεδομένων γονιδιακής έκφρασης. Ένας πιθανός λόγος είναι ο μικρός αριθμός δείγματος (49) που χρησιμοποιήθηκε στην Εργασία. Επίσης, οι καταστάσεις ΤΙ και ΤΜ ίσως αποτελούν μέρος ενός ευρέους φαινοτυπικού φάσματος και να μην είναι διακριτές. Σε μελλοντικές μελέτες με αυξημένο αριθμό δειγμάτων και με την ενσωμάτωση πρωτεομικών δεδομένων, τα οποία σχετίζονται πιο άμεσα με τον φαινότυπο, η ταξινόμηση των δειγμάτων μέσω του αλγορίθμου και η επιλογή υποψήφιων γονιδίων-βιοδεικτών θα έχει μεγαλύτερη αξιοπιστία.

Ε. ΠΕΡΙΛΗΨΗ – ABSTRACT

Η β-θαλασσαιμία είναι μια αιμοσφαιρινοπάθεια στην οποία παρατηρείται αναποτελεσματική ερυθροποίηση. Χαρακτηρίζεται από μεγάλο αριθμό μεταλλάξεων του γονιδίου της β-σφαιρίνης, οι οποίες μειώνουν τη σύνθεσή της. Ο φαινότυπος που παρουσιάζει η ασθένεια μπορεί να είναι ασυμπτωματικός, ενδιάμεσος (Thalassaemia Intermedia, TI) ή μείζων (Thalassaemia Major, TM) με διαφορετική θεραπευτική αντιμετώπιση (μεταγγίσεις αίματος, αποσιδήρωση και μεταμόσχευση μυελού). Η έγκαιρη κατηγοριοποίηση των ασθενών ως TI ή TM είναι σημαντική στην κλινική πράξη για την αποφυγή των μη απαραίτητων μεταγγίσεων και των επακόλουθων επιπλοκών τους στα άτομα με ενδιάμεσο φαινότυπο και την έγκαιρη έναρξη των μεταγγίσεων σε άτομα με μείζων θαλασσαιμία. Για το σκοπό αυτό χρησιμοποιήθηκαν δεδομένα γονιδιακής έκφρασης (RNA-Seq) 2999 στατιστικώς σημαντικών διαφορικώς εκφραζόμενων γονιδίων από 49 δείγματα (υγιών ατόμων και ασθενών με ενδιάμεση και μείζων θαλασσαιμία), τα οποία αναλύθηκαν με το εργαλείο GeneSrF που χρησιμοποιεί τον αλγόριθμο μηχανικής μάθησης "Τυχαίο Δάσος". Το GeneSrF χρησιμοποιήθηκε για την ανακάλυψη γονιδίων που μπορούν να χρησιμοποιηθούν ως βιοδείκτες. Τα αποτελέσματα έδειξαν ότι η εκπαίδευση του αλγορίθμου για τη διάκριση μεταξύ ασθενών και υγιών (Η) ατόμων οδηγεί σε μεγάλη αξιοπιστία στη διάκριση μεταξύ Η & TI και μεγαλύτερη αξιοπιστία στη διάκριση μεταξύ Η & TM ατόμων. Ωστόσο, η διάκριση μεταξύ των δειγμάτων TI και TM δεν ήταν εφικτή. Τέλος, η παρούσα εργασία εντόπισε γονίδια που θα ελέγχουν πειραματικά και πιθανώς να συνεισφέρει μελλοντικά στον προσδιορισμό βιοδεικτών για την αποτελεσματική κατηγοριοποίηση των θαλασσαιμικών ασθενών.

β-Thalassaemia is a hemoglobinopathy characterized by ineffective erythropoiesis. A large number of mutations has been identified in the β-globin gene, resulting in reduced production of β-globin. The phenotype of a person with β-Thalassaemia can be asymptomatic, intermediate (Thalassaemia Intermedia – TI) or severe (Thalassaemia Major – TM). Each phenotype requires different therapeutic management (such as blood transfusion, iron chelation or bone marrow transplantation). Identification of molecular biomarkers for patient stratification is necessary to guide therapeutic decisions in clinical practice. In this study, gene expression data (RNA-Seq) from 2999 statistically significant differentially expressed genes from 49 individuals (Healthy, TI and TM) have been used to train a Random Forest machine learning tool named GeneSrf, to identify genes that distinguish different groups and

might be used as biomarkers for patient stratification in thalassaemia. Our results show that RNA-Seq data does increase the success rate of the predictive model generated by the Random Forest algorithm to stratify TM vs H and TI vs H patients. However, stratification of TM vs TI patients was not possible. This study has identified genes that will be further experimentally validated and might be useful as biomarkers for efficient β-thalassaemia patient stratification.

Βιβλιογραφία

Akond, Z., Alam, M., Mollah, M.N.H., 2018. Biomarker Identification from RNA-Seq Data using a Robust Statistical Approach. Bioinformation 14, 153–163. https://doi.org/10.6026/97320630014153

Breiman, L., 2001. Random Forests. Machine Learning 45, 5-32. https://doi.org/10.1023/A:1010933404324

De Benoist, B., World Health Organization, Centers for Disease Control and Prevention (U.S.), 2008. Worldwide prevalence of anaemia 1993-2005 of: WHO Global Database of anaemia. World Health Organization, Geneva.

Degenhardt, F., Seifert, S., Szymczak, S., 2019. Evaluation of variable selection methods for random forests and omics data sets. Briefings in Bioinformatics 20, 492–503. https://doi.org/10.1093/bib/bbx124

Diaz-Uriarte, R., 2007. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC Bioinformatics 8. https://doi.org/10.1186/1471-2105-8-328

Diaz-Uriarte, R., de Andres, S.A., 2005. Variable selection from random forests: application to gene expression data. arXiv:q-bio/0503025.

Elliott, S., Pham, E., Macdougall, I.C., 2008. Erythropoietins: A common mechanism of action. Experimental Hematology 36, 1573–1584. https://doi.org/10.1016/j.exphem.2008.08.003

Goossens, N., Nakagawa, S., Sun, X., Hoshida, Y., 2015. Cancer biomarker discovery and validation 21.

Grissa, D., Pétéra, M., Brandolini, M., Napoli, A., Comte, B., Pujos-Guillot, E., 2016. Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data. Frontiers in Molecular Biosciences 3. https://doi.org/10.3389/fmolb.2016.00030

Hattangadi, S.M., Wong, P., Zhang, L., Flygare, J., Lodish, H.F., 2011. From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. Blood 118, 6258–6268. https://doi.org/10.1182/blood-2011-07-356006

Hrdlickova, R., Toloue, M., Tian, B., 2017. RNA-Seq methods for transcriptome analysis: RNA-Seq. Wiley Interdisciplinary Reviews: RNA 8, e1364. https://doi.org/10.1002/wrna.1364

Katsantoni, E., 2019. Omics Studies in Hemoglobinopathies. Molecular Diagnosis & Therapy 23, 223–234. https://doi.org/10.1007/s40291-019-00386-1

Katsantoni, E., 2012. Protein Complexes and Target Genes Identification by in Vivo Biotinylation: The STAT5 Paradigm. Science Signaling 5, pt13–pt13. https://doi.org/10.1126/scisignal.2003622

Lee, J., 2017. Patient-Specific Predictive Modeling Using Random Forests: An Observational Study for the Critically Ill. JMIR Medical Informatics 5, e3. https://doi.org/10.2196/medinform.6690

Mamoshina, P., Volosnikova, M., Ozerov, I.V., Putin, E., Skibina, E., Cortese, F., Zhavoronkov, A., 2018. Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification. Frontiers in Genetics 9. https://doi.org/10.3389/fgene.2018.00242

McDermott, J.E., Wang, J., Mitchell, H., Webb-Robertson, B.-J., Hafen, R., Ramey, J., Rodland, K.D., 2013. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. Expert Opinion on Medical Diagnostics 7, 37–51. https://doi.org/10.1517/17530059.2012.718329

Passegue, E., Jamieson, C.H.M., Ailles, L.E., Weissman, I.L., 2003. Normal and leukemic hematopoiesis: Are leukemias a stem cell disorder or a reacquisition of stem cell characteristics? Proceedings of the National Academy of Sciences 100, 11842–11849. https://doi.org/10.1073/pnas.2034201100

Quezada, H., Guzmán-Ortiz, A.L., Díaz-Sánchez, H., Valle-Rios, R., Aguirre-Hernández, J., 2017. Omics-based biomarkers: current status and potential use in the clinic. Boletín Médico del Hospital Infantil de México 74, 219–226. https://doi.org/10.1016/j.bmhimx.2017.03.003

Ribeil, J.-A., Arlet, J.-B., Dussiot, M., Cruz Moura, I., Courtois, G., Hermine, O., 2013. Ineffective Erythropoiesis in β -Thalassemia. The Scientific World Journal 2013, 1–11. https://doi.org/10.1155/2013/394295

Rivella, S., 2009. Ineffective erythropoiesis and thalassemias: Current Opinion in Hematology 16, 187–194. https://doi.org/10.1097/MOH.0b013e32832990a4

Singh, A.K., 2018. Erythropoiesis, in: Textbook of Nephro-Endocrinology. Elsevier, pp. 207–215. https://doi.org/10.1016/B978-0-12-803247-3.00012-X

Testa, U., 2004. Apoptotic mechanisms in the control of erythropoiesis. Leukemia 18, 1176–1199. https://doi.org/10.1038/sj.leu.2403383

Vafaee, F., Diakos, C., Kirschner, M.B., Reid, G., Michael, M.Z., Horvath, L.G., Alinejad-Rokny, H., Cheng, Z.J., Kuncic, Z., Clarke, S., 2018. A data-driven, knowledge-based approach to biomarker discovery: application to circulating microRNA markers of colorectal cancer prognosis. npj Systems Biology and Applications 4. https://doi.org/10.1038/s41540-018-0056-1

Valdes, G., Solberg, T.D., Heskel, M., Ungar, L., Simone, C.B., 2016. Using machine learning to predict radiation pneumonitis in patients with stage I non-small cell lung cancer treated with stereotactic body radiation therapy. Physics in Medicine and Biology 61, 6105–6120. https://doi.org/10.1088/0031-9155/61/16/6105

Valent, P., Büsche, G., Theurl, I., Uras, I.Z., Germing, U., Stauder, R., Sotlar, K., Füreder, W., Bettelheim, P., Pfeilstöcker, M., Oberbauer, R., Sperr, W.R., Geissler, K., Schwaller, J., Moriggl, R., Béné, M.C., Jäger, U., Horny, H.-P., Hermine, O., 2018. Normal and pathological erythropoiesis in adults: from gene regulation to targeted treatment concepts. Haematologica 103, 1593–1603. https://doi.org/10.3324/haematol.2018.192518

van Rensburg, I.C., Loxton, A.G., 2015. Transcriptomics: the key to biomarker discovery during tuberculosis? Biomarkers in Medicine 9, 483–495. https://doi.org/10.2217/bmm.15.16

van Rheenen, W., Diekstra, F.P., Harschnitz, O., Westeneng, H.-J., van Eijk, K.R., Saris, C.G.J., Groen, E.J.N., van Es, M.A., Blauw, H.M., van Vught, P.W.J., Veldink, J.H., van den Berg, L.H., 2018. Whole blood transcriptome analysis in amyotrophic lateral sclerosis: A biomarker study. PLOS ONE 13, e0198874. https://doi.org/10.1371/journal.pone.0198874

Yalamanchili, H.K., Wan, Y.-W., Liu, Z., 2017. Data Analysis Pipeline for RNA-seq Experiments: From Differential Expression to Cryptic Splicing: Data Analysis Pipeline for RNA-seq Experiments, in: Bateman, A., Pearson, W.R., Stein, L.D., Stormo, G.D., Yates, J.R. (Eds.), Current Protocols in Bioinformatics. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 11.15.1-11.15.21. https://doi.org/10.1002/cpbi.33

Zararsiz, G., Goksuluk, D., Klaus, B., Korkmaz, S., Eldem, V., Karabulut, E., Ozturk, A., 2017. voomDDA: discovery of diagnostic biomarkers and classification of RNA-seq data. PeerJ 5, e3890. https://doi.org/10.7717/peerj.3890

Zivot, A., Lipton, J.M., Narla, A., Blanc, L., 2018. Erythropoiesis: insights into pathophysiology and treatments in 2017. Molecular Medicine 24. https://doi.org/10.1186/s10020-018-0011-z