ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Υπολογιστική μέθοδος για την πρόβλεψη και χαρτογράφηση αλληλεπιδράσεων των μικρών RNAs του ξενιστή με γονίδια του μικροβιώματος του

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Γιώργος Σκούφος

Επιβλέπουσα: Άρτεμις Γ. Χατζηγεωργίου,

Καθηγήτρια Βιοπληροφορικής, ΠΘ

Δεκέμβριος, 2022

Το έργο συγχρηματοδοτείται από την Ελλάδα και την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) μέσω του Επιχειρησιακού Προγράμματος «Ανάπτυξη Ανθρώπινου Δυναμικού, Εκπαίδευση και Διά Βίου Μάθηση», στο πλαίσιο της Πράξης «Ενίσχυση του ανθρώπινου ερευνητικού δυναμικού μέσω της υλοποίησης διδακτορικής έρευνας» (MIS-5000432), που υλοποιεί το Ίδρυμα Κρατικών Υποτροφιών (ΙΚΥ)

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

Επιχειρησιακό Πρόγραμμα
Ανάπτυξη Ανθρώπινου Δυναμικού,
Εκπαίδευση και Διά Βίου Μάθηση
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

An in-silico approach for predicting and characterizing interactomes among bacterial genes and host microRNAs

PhD DISSERTATION

Giorgos Skoufos

Supervisor: Artemis G. Hatzigeorgiou,

Bioinformatics Professor

December, 2022

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Strengthening Human Resources Research Potential via Doctorate Research" (MIS-5000432), implemented by the State Scholarships Foundation (IKY)

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Υπολογιστική μέθοδος για την πρόβλεψη και χαρτογράφηση αλληλεπιδράσεων των μικρών RNAs του ξενιστή με γονίδια του μικροβιώματος του

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Γιώργος Σκούφος

Συμβουλευτική Επιτροπή

Άρτεμις Γ. Χατζηγεωργίου, Καθηγήτρια, Πανεπιστήμιο Θεσσαλίας (Επιβλέπουσα)

Γεώργιος Σταμούλης, Καθηγητής, Πανεπιστήμιο Θεσσαλίας

Παντελής Μπάγκος, Καθηγητής, Πανεπιστήμιο Θεσσαλίας

Επταμελής εξεταστική επιτροπή

Άρτεμις Γ. Χατζηγεωργίου, Καθηγήτρια, Πανεπιστήμιο Θεσσαλίας (Επιβλέπουσα)

Γεώργιος Σταμούλης, Καθηγητής, Πανεπιστήμιο Θεσσαλίας

Παντελής Μπάγκος, Καθηγητής, Πανεπιστήμιο Θεσσαλίας

Γεωργία Γ. Μπράλιου, Επίκουρη Καθηγήτρια, Πανεπιστήμιο Θεσσαλίας

Διονύσιος Σγούρας, Διευθυντής Ερευνών, Ελληνικό Ινστιτούτο Παστέρ

Γεράσιμος Ποταμιάνος, Αναπληρωτής Καθηγητής, Πανεπιστήμιο Θεσσαλίας

Παναγιώτα Τσομπανοπούλου, Καθηγήτρια, Πανεπιστήμιο Θεσσαλίας

Δεκέμβριος, 2022

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διδακτορική διατριβή, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της διατριβής, αποτελούν αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλουν οποιασδήποτε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχουν έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνω επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο Δηλών

Γιώργος Σκούφος

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

An in-silico approach for predicting and characterizing interactomes among bacterial genes and host microRNAs

PhD DISSERTATION

Giorgos Skoufos

Advisory Committee

Artemis G. Hatzigeorgiou, Professor, University of Thessaly

Georgios Stamoulis, Professor, University of Thessaly

Pantelis Bagos, Professor, University of Thessaly

Examination Committee

Artemis G. Hatzigeorgiou, Professor, University of Thessaly

Georgios Stamoulis, Professor, University of Thessaly

Pantelis Bagos, Professor, University of Thessaly

Georgia G. Braliou, Assistant Professor, University of Thessaly

Dionyssios Sgouras, Research Director, Hellenic Pasteur Institute

Gerasimos Potamianos, Associate Professor, University of Thessaly

Panagiota Tsompanopoulou, Professor, University of Thessaly

December, 2022

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this Ph.D. dissertation, as well as the electronic files and source codes developed or modified in the course of this dissertation, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this dissertation or part of it does not belong to me because it is a product of plagiarism.

The Declarant

Giorgos Skoufos

Ευχαριστίες

Αρχικά, ευχαριστώ ιδιαίτερα την Καθηγήτρια Άρτεμη Χατζηγεωργίου που το 2017, με υποδέχθηκε στο DIANA-Lab και μου έδωσε την ευκαιρία να πραγματοποιήσω τη Διδακτορική μου Διατριβή, δίνοντας προτεραιότητα στις προσωπικές μου ερευνητικές επιθυμίες. Την ευχαριστώ για την επίβλεψη, την υποστήριξη και την εμπιστοσύνη που μου έδειξε και συνεχίζει να μου δείχνει όλα αυτά τα χρόνια. Με την Άρτεμη ξενυχτήσαμε πολλές φορές, κυνηγώντας προθεσμίες για χρηματοδοτικά και δημοσιεύσεις και κόντρα στις πιθανότητες, βγαίναμε πάντα κερδισμένοι.

Συνεχίζοντας, θα ήθελα να ευχαριστήσω τους Καθηγητές Γεώργιο Σταμούλη και Παντελή Μπάγκο, μέλη της Τριμελούς Συμβουλευτικής Επιτροπής και της Επταμελούς Εξεταστικής Επιτροπής για τις συμβουλές που μου έδωσαν κατά τη διάρκεια συγγραφής της παρούσας Διατριβής. Επίσης, τους ευχαριστώ θερμά για την ανταπόκριση και τη βοήθειά τους ως προς τις υποχρεώσεις μου απέναντι στο Ίδρυμα Κρατικών Υποτροφιών. Ευχαριστώ θερμά και τις Καθηγήτριες Γεωργία Μπράλιου και Παναγιώτα Τσομπανοπούλου, τον Καθηγητή Γεράσιμο Ποταμιάνο και τον Ερευνητή Διονύσιο Σγούρα που δέχθηκαν να συμμετάσχουν ως μέλη της Επταμελούς Εξεταστικής Επιτροπής αλλά και για τις υποδείξεις τους ως προς την παρουσίαση των αποτελεσμάτων της παρούσας Διατριβής.

Ευχαριστώ θερμά τον Επίκουρο Καθηγητή Ιωάννη Βλάχο που πίστεψε σε μένα από πολύ νωρίς, με καθοδήγησε με τον καλύτερο δυνατό τρόπο και κατ' επέκταση, διαμόρφωσε στο μέγιστο βαθμό την ερευνητική μου σκέψη. Ευχαριστώ επίσης τη συνάδελφο και φίλη Δρ. Δήμητρα Καραγκούνη που με συμπεριέλαβε σε δύο ερευνητικές της εργασίες αλλά και για την παρέα και στήριξη τα πρώτα χρόνια εκπόνησης της Διδακτορικής μου Διατριβής.

Έχω την ανάγκη να εκφράσω τις θερμές μου ευχαριστίες στον φίλο μου και στενό συνεργάτη, Δρ. Σπύρο Ταστσόγλου, ο οποίος εδώ και έξι χρόνια αποτελεί το ερευνητικό μου αποκούμπι. Με το Σπύρο πορευόμαστε μαζί από την αρχή, ανταλλάσσοντας καλές (και κακές) ερευνητικές ιδέες, σκαρώνοντας όμορφες εργασίες και στιγμές. Ελπίζω πως θα συνεχίσουμε μαζί για πολλά ακόμα χρόνια.

Ευχαριστώ την υποψήφια διδάκτορα Βασιλική Κώτσιρα για την παρέα και την ερευνητική μας συνεργασία, τους υποψήφιους διδάκτορες Θάνο Αλεξίου και Φίλιππο Καρδαρά για την άψογη συνεργασία σε τρεις ερευνητικές εργασίες και τον Δρ. Νίκο Περδικοπάνη για την άριστη συνεργασία και την συμπόρευση από το μεταπτυχιακό Βιοπληροφορικής, που ξεκίνησε το 2014 έως και σήμερα. Επίσης, θέλω να ευχαριστήσω τα υπόλοιπα (πρώην και νυν) μέλη του DIANA-Lab, Ελίζα Ζαχαροπούλου, Μάριο Μιλιώτη, Δημήτρη Γρηγοριάδη,

Άννα Καραβαγγέλη και Δρ. Ιωάννη Καβακιώτη για τη διάθεση για κουβέντα, την συνεργασία και το καλό κλίμα εντός και εκτός εργαστηρίου. Ευχαριστώ και τους συνεργάτες μας από το Πανεπιστήμιο της Πενσυλβάνια, Καθηγητή Κώστα Κουμένη, Δρ. Ιωάννη Βεργινάδη, Δρ. Αναστασία Βελαλοπούλου καθώς επίσης και τους συνεργάτες μας από το Ελληνικό Ινστιτούτο Παστέρ, Ερευνήτρια Ευδοκία Καραγκούνη, Δρ. Μαρία Αγάλλου και Δρ. Μαρίτσα Μαργαρώνη για την εμπιστοσύνη που μου έδειξαν και την άριστη συνεργασία, μέσα από την οποία καταφέραμε να δημοσιεύσουμε τρεις ερευνητικές εργασίες. Θέλω επίσης να ευχαριστήσω το Κρατικό Ίδρυμα Υποτροφιών που χρηματοδότησε την ερευνητική μου δραστηριότητα.

Για το τέλος, άφησα πολλά ονόματα και ακόμα πιο πολλά συναισθήματα. Ευχαριστώ την Ελένη (μαμά), το Νίκο (μπαμπάς) και την Αντιγόνη (γιαγιά) που ήταν και είναι πάντα δίπλα μου και δεν αφήνουν τίποτα - και ποτέ - να μπαίνει εμπόδιο σε όσα θέλω να κάνω. Ευχαριστώ τον φίλο μου Πάνο, για την παρέα, τα ταξίδια, τα μπαρ, τα ποτά, τις συναυλίες και όλα όσα κάνουν τη ζωή ομορφότερη εδώ και οχτώ περίπου χρόνια. Ευχαριστώ το φίλο μου Δαμιανό που εδώ και τριάντα χρόνια (not a typo) είναι δίπλα μου και πρακτικά έχουμε περάσει μαζί όλα τα όμορφα και όχι τόσα όμορφα και που βγαίνουμε νικητές. Ευχαριστώ τον αδερφό μου Θανάση και τις αδερφές μου Καλλιρρόη κι Αντιγόνη για όλα όσα έχουν κάνει και όλα όσα θα κάνουν για μένα στο μέλλον. Ευχαριστώ και τους φίλους μου Στέφανο, Αντρέα, Μάριο, Δημήτρη και Ναταλία για την παρέα, τη στήριξη και τις όμορφες στιγμές.

Κλείνοντας, θέλω να ευχαριστήσω τη Μαρία που αποτελεί το κίνητρο για όλα όσα θα διαβάσετε παρακάτω. Που γεμίζει τη ζωή μου με χαρούμενες στιγμές και που όταν είμαι κοντά της, μπορώ να είμαι ο εαυτός μου!

Computers aren't the thing.
They're the thing that gets us to the thing.

Joe MacMillan, HACF

Διδακτορική Διατριβή

Υπολογιστική μέθοδος για την πρόβλεψη και χαρτογράφηση αλληλεπιδράσεων των μικρών RNAs του ξενιστή με γονίδια του μικροβιώματος του

Γιώργος Σκούφος

Περίληψη

Η επιστήμη της Μεταγονιδιωματικής έχει προσφέρει τα τελευταία χρόνια σημαντικές δυνατότητες προς τη μελέτη των μικροοργανισμών (Βακτήρια, Αρχαία, Μύκητες, Ιοί) σε περίπλοκα περιβαλλοντικά δείγματα όπως είναι το νερό ωκεανών, το χώμα αλλά και το ανθρώπινο σώμα. Με τον όρο Μικροβίωμα χαρακτηρίζεται το σύνολο των μικροοργανισμών που εποικίζουν ένα περιβάλλον, καθώς επίσης και οι λειτουργίες που επιτελούν σε αυτό. Οι τεχνικές Αλληλούχησης Επόμενης Γενιάς (π.χ., Shotgun metagenomics) έχουν παίξει καθοριστικό ρόλο στη μελέτη των περίπλοκων αλληλεπιδράσεων μεταξύ Ξενιστή-Μικροβιώματος στον άνθρωπο και σε άλλους οργανισμούς. Το Human Microbiome Project και παρόμοιες μελέτες μεγάλης κλίμακας των τελευταίων ετών, αποκάλυψαν τη σημασία του Μικροβιώματος και τη συμμετοχή του σε παθολογικές καταστάσεις συμπεριλαμβανομένων των φλεγμονωδών νόσων του εντέρου, νεοπλασματικών παθήσεων, μεταβολικών διαταραχών, νευροεκφυλιστικών ασθενειών και άλλων παθήσεων.

Τα microRNAs (miRNAs) αποτελούν μη-κωδικά μετάγραφα μήκους ~22 νουκλεοτιδίων. Είναι κύριοι μετα-μεταγραφικοί ρυθμιστές της γονιδιακής έκφρασης με καθοριστικό ρόλο σε πολυάριθμες φυσιολογικές αλλά και παθολογικές καταστάσεις. Τα miRNAs προσδένουν σε μικρές ακολουθίες των στόχων τους που ονομάζονται Στοιχεία Αναγνώρισης από miRNAs (miRNA Recognition Elements, MREs). Κατά τη στόχευση, τα miRNAs προσδένονται στα MREs, με βάση σύνθετους κανόνες τέλειας ή ατελούς συμπληρωματικότητας του RNA, επάγοντας την καταστολή της μετάφρασης ή και την αποικοδόμησή του. Η πλειοψηφία των βιολογικών διεργασιών στα ανώτερα θηλαστικά ρυθμίζεται από τη δράση των miRNAs σε μετα-μεταγραφικό επίπεδο.

Το Μικροβίωμα διαμορφώνει τη γονιδιακή έκφραση του ξενιστή, αλλά οι υποκείμενοι μοριακοί μηχανισμοί της αλληλεπίδρασης αυτής παραμένουν ασαφείς. Προσφάτως έχει αναδειχθεί η ρύθμιση της έκφρασης των miRNAs αφενός ως κρίσιμο συστατικό της

απόκρισης του ξενιστή σε μολύνσεις, αφετέρου δε ως στρατηγική που εκμεταλλεύονται οι μικροοργανισμοί για να χειραγωγήσουν μοριακά μονοπάτια του ξενιστή. Πρόσφατες μελέτες δείχνουν πως miRNAs του ξενιστή δύνανται να εισχωρήσουν σε μικροβιακούς οργανισμούς, επηρεάζοντας τη γονιδιακή τους έκφραση και κρίσιμες διεργασίες τους όπως το ρυθμό αύξησης.

Στα πλαίσια της παρούσας Διατριβής, αναπτύχθηκε υπολογιστική μεθοδολογία για την ανάλυση δεδομένων Μεταγονιδιωματικής από πειράματα Αλληλούχησης Επόμενης Γενιάς. Η εν λόγω μεθοδολογία χρησιμοποιεί μία σειρά από αλγοριθμικές τεχνικές και δομές δεδομένων με σκοπό την ποσοτικοποίηση μικροοργανισμών σε δείγματα Shotgun metagenomics με δραματικά μειωμένες απαιτήσεις σε μνήμη RAM, εξαιρετικά υψηλή ταχύτητα και ακρίβεια που ξεπερνάει το state-of-the-art. Ταυτόχρονα, αναπτύχθηκε γραφικό περιβάλλον μέσω του οποίου προσφέρονται μία σειρά από στατιστικές αναλύσεις και επιλογές οπτικοποίησης για τα δεδομένα ποσοτικοποίησης που προέκυψαν από τα αναλυμένα δείγματα. Επίσης, σχεδιάστηκε και αναπτύχθηκε βάση δεδομένων που περιέχει πειραματικά υποστηριζόμενες συσχετίσεις μεταξύ βακτηρίων και ασθενειών. Η βάση δεδομένων περιέχει περισσότερες από 7900 συσχετίσεις, που αφορούν 43 ασθένειες και 1396 μικροοργανισμούς. Τέλος, πραγματοποιήθηκε ανάλυση δειγμάτων NGS (Shotgun metagenomics, small RNA-Seq) και αναπτύχθηκε υπολογιστική μεθοδολογία με σκοπό την εξαγωγή πιθανών στόχων των ανθρώπινων miRNAs σε γονίδια του Μικροβιώματός του. Για τις ανάγκες της μελέτης, έγινε χρήση βιοφυσικών και βιοχημικών κανόνων γενικής φύσεως που ισχύουν για τις RNA-RNA αλληλεπιδράσεις, όπως η προσβασιμότητα ακολουθίας-στόχου, η ενέργεια πρόσδεσης για τη δημιουργία του διμερούς και συνολικός αριθμός ταιριασμάτων και αναντιστοιχιών στην περιοχή πρόσδεσης.

Κατά την εκπόνηση της Διδακτορικής μου Διατριβής, συμμετείχα σε 8 δημοσιευμένες εργασίες σε επιστημονικά περιοδικά υψηλού κύρους (μία παρουσίαση μεθοδολογίας ποσοτικοποίησης μικροοργανισμών, δύο εργασίες σχετικές με τη μελέτη της στόχευσης mRNAs/lncRNAs από miRNAs, μία παρουσίαση Βάσης Δεδομένων με συσχετίσεις της αφθονίας μικροβίων με ασθένειες, μία μεταγραφωματική μελέτη για τη διερεύνηση του μικροπεριβάλλοντος του καρκίνου με τεχνικές single-cell RNA-Seq, μία μεταγραφωματική μελέτη για τη διερεύνηση της τεχνικής ακτινοβόλησης όγκων FLASH, μία μεταγραφωματική μελέτη για τη διερεύνηση αλληλεπιδράσεων μεταξύ Ξενιστή-Λεϊσμάνιας και μία παρουσίαση Βάσης Δεδομένων με miRNAs διαγνωστικής/προγνωστικής αξίας). Έχω παρουσιάσει ερευνητικά αποτελέσματα σε 6 επιστημονικά συνέδρια και ημερίδες (4 εθνικά, 2 διεθνή), και έχω συμβάλει στη διοργάνωση του Πανευρωπαϊκού Συνεδρίου Βιοπληροφορικής ECCB 2018. Σύμφωνα με το Google Scholar οι εργασίες στις οποίες μετέχω έχουν λάβει μέχρι σήμερα 841 αναφορές.

Το Νοέμβριο του 2018, έγινε δεκτή η αίτηση μου στο Ίδρυμα Κρατικών Υποτροφιών (ΙΚΥ) για χρηματοδότηση Διατριβής με διάρκεια 36 μηνών.

Λέξεις-κλειδιά:

Μικρά RNAs, Μεταγονιδιωματική, Μικροβίωμα, Βιολογία των RNAs, Πειράματα Αλληλούχησης Επόμενης Γενιάς, Βακτήρια, Αλληλεπιδράσεις RNA μορίων, Υπολογιστική Βιολογία

Ph.D. Dissertation

An in-silico approach for predicting and characterizing interactomes among bacterial genes and host microRNAs

Giorgos Skoufos

Abstract

The study of metagenomics has offered us novel aspects of the world around us and within, while Shotgun sequencing revolutionized the field, offering higher throughput and resolution. With the term Microbiome we refer to the entirety of microorganisms (i.e., Bacteria, Yeast, Archaea, Viruses) present in a given environment, as well as to their associated function(s). Next-Generation Sequencing (NGS) techniques have shed light to the complex host-microbiome relationships in humans, mice and other organisms, enabling detailed, and even population-scale studies. The Human Microbiome Project and other similar scientific endeavors revealed the importance of the microbiome and its implications in pathological conditions, including GI tract inflammatory diseases, neoplastic conditions, metabolic disorders and neurodegenerative diseases.

microRNAs (miRNAs) are short RNA molecules (~22 nucleotides long) which exert their post-transcriptional regulatory function by targeting miRNA Recognition Elements (MREs) on the sequence of coding and non-coding RNAs (e.g. messenger RNAs, mRNAs, long non-coding RNAs, lncRNAs). miRNA-MRE binding follows complex rules of perfect or imperfect RNA complementarity and results in translational suppression and/or degradation of the targeted transcripts. The vast majority of biological processes in higher mammals are under miRNA regulation at the post-transcriptional level.

The Microbiome regulates host gene expression but the underlying molecular mechanisms of this interplay remain unclear. Recently, the regulation by miRNAs has emerged as both a critical component of the host's response to infections and as a strategy exploited by microorganisms to manipulate host molecular pathways. Recent studies show that host miRNAs can enter microbial cells and regulate their gene expression and critical processes such as growth rates.

During the course of my PhD thesis, a computational method for the analysis of metagenomics datasets derived from NGS techniques was designed and implemented. The

methodology utilizes a series of algorithmic techniques and data structures to quantify microbial abundances in Shotgun metagenomics datasets with dramatically reduced memory footprints, lightning-fast speed and accuracy that goes beyond the state-of-the-art. On top of the quantification results, a user-friendly graphical interface offers numerous downstream analyses modules, enabling users to explore and visualize microbial abundances but also conduct analyses of differential abundance, diversity indices and correlation. Moreover, a database hosting experimentally supported microbe-disease associations was developed. The database is comprised of more than7900 associations, linking 43 diseases and 1396 microorganisms. Finally, NGS-derived samples (i.e., Shotgun metagenomics, small RNA-Seq) were analyzed and a computational methodology was developed in order to extract possible targets of human miRNAs in genes of the host's Microbiome. For the needs of the study, general biochemical and biophysical rules of the RNA-RNA interaction space were utilized, including target sequence accessibility, dimer binding energy and sequence complementarity (i.e., matches and mis-matches in the miRNA binding region).

During my doctoral dissertation, I participated in 8 published studies (i.e., one presentation of a microbial quantification application in Shotgun metagenomics samples, two studies related to miRNA targeting, the showcase of a database of experimentally supported microbe-disease associations, one transcriptomic analyses for the study of the tumor microenvironment using single-cell RNA Sequencing, one transcriptomic study to investigate FLASH radiation for the treatment of tumors, one transcriptomic study to investigate Host-Leishmania interactions and the showcase of a database providing circulating miRNAs of diagnostic/prognostic value). I presented research results in 6 scientific conferences (4 national and 2 international) and contributed to the organization of the European Conference of Computational Biology (ECCB 2018). As of today, my publications have been cited 841 times, according to Google Scholar.

On November 2018, the State Scholarship Foundation (I.K.Y.) accepted to fund my PhD proposal in the form of a 36-month scholarship.

Keywords:

small RNAs, Metagenomics, Microbiome, RNA biology, Next-Generation Sequencing, Bacteria, RNA-RNA interactions, Computational Biology

Table of contents

Περίληψη ..... *xvi*
Abstract ..... *xix*
Table of contents *1*
List of figures ..... *xxiii*
List of tables ..... *xxvii*
Abbreviations *29*
INTRODUCTION *31*
CHAPTER 1 – RNA biology *31*
1.1 Fundamentals of RNA **31**
1.2 The magnificent world of RNA species **34**
1.2.1 Messenger RNA (mRNA) 34
1.2.2 Transfer RNA (tRNA) 37
1.2.3 Ribosomal RNA (rRNA) 39
1.2.4 microRNA (miRNA) 40
1.2.4.1 Biogenesis 41
1.2.4.2 Primary function 42
1.2.4.3 Basic targeting rules 43
1.2.4.4 Extracellular miRNA 45
1.2.4.5 Atypical miRNA function 46
1.2.5 Long non-coding RNA (lncRNA) 48
1.2.6 Other RNA species 49
1.3 RNA in pathology **50**
CHAPTER 2 - Microbiome *51*
2.1 Microbiome and metagenomics **51**
2.2 Bacteria **52**
2.3 Human microbiome **55**
2.3.1 Development 55
2.3.2 Architecture 56
2.3.3 Pathophysiology 58
METHODS AND RESULTS *60*
CHAPTER 3 – Metagenomics analysis suite *60*
3.1 Overview of AGAMEMNON quantification suite **61**
3.2 Benchmarking AGAMEMNON against state-of-the-art methodologies **64**
3.3 Microbial quantification in host RNA/DNA-Seq samples **68**
3.4 Downstream analysis of quantification results with R-Shiny **69**
CHAPTER 4 – An expanded microbe-disease association compendium *70*
4.1 Peryton's content and statistics **70**

4.2 Interface, modules and implementation of Peryton 71

CHAPTER 5 – An extended catalogue of bacterial small RNA-RNA interactions..... 73

5.1 Agnodice's content and statistics 75

5.2 Agnodice's data collection and curation process..... 76

5.3 Interface, modules and implementation of Agnodice 76

5.4 Comparison of Agnodice with existing resources..... 78

CHAPTER 6 – miRNA targets on non-coding transcripts and expression of lncRNAs at sub-cellular resolution..... 79

6.1 A database of experimentally supported miRNA targets on non-coding transcripts 79

6.2 Expression levels of lncRNAs at cellular and sub-cellular resolution..... 79

CHAPTER 7 – Host miRNA-bacterial mRNA interactions on the spotlight 82

7.1 Quantification of miRNAs in the gut extracellular space..... 83

7.2 Quantification of microbial abundances in the human gut 86

7.3 Discovery of potential host miRNA-bacterial mRNA interactions 88

7.4 Functional interpretation and validation of the human miRNA-bacterial gene candidate pairs..... 95

CHAPTER 8 – Discussion and conclusions 105

CHAPTER 9 – Publications 107

CHAPTER 10 – Posters.....109

List of figures

Figure 1.1: (A) Molecular and chemical structure of RNA. (B) Comparison of ribose (RNA sugar) and deoxyribose (DNA sugar). (C) Examples of a secondary and a tertiary structure of an RNA molecule (Figure was modified for the purpose of this thesis. Original figures: National Human Genome Research Institute, study.com and Qi Zhao et al. 2021). 31

Figure 1.2: A simple classification of some of the known RNA types based on their coding capacity and size (Figure was created for the purpose of this thesis). 32

Figure 1.3: Schematic representation of the process of transcription (DNA to RNA) utilizing RNA polymerase II and the process of translation (mRNA to protein) from the ribosome (Figure adopted from: Clancy, S. & Brown, W. (2008) Translation: DNA to mRNA to Protein. Nature Education 1(1):101). 34

Figure 1.4: Illustration of a precursor mRNA (pre-mRNA), a mature mRNA and their components (Original figure (source: Wikipedia) was modified for the purpose of this thesis). 35

Figure 1.5: Overview of some of the most common types of alternative splicing (Figure adopted from: Jonathan Dornell, Alternative Splicing: Importance and Definition, Genomics research). 36

Figure 1.6: Presentation of the genetic code blueprint. Each codon (combination of three nucleotides) translates in one of the 20 available amino acids (Figure adopted from: National Human Genome Research Institute). 37

Figure 1.7: Illustration of the process of protein synthesis in the presence of an mRNA, tRNAs and the ribosome (Figure adopted from: Wikipedia). 38

Figure 1.8: The secondary and territory structure of a typical transfer RNA (Figure was adopted from: Wikipedia). 39

Figure 1.9: Illustration of the SSU and LSU of the ribosome, an mRNA under translation and tRNAs that carry amino acids to a newly synthesized amino acid chain (Figure was adopted from: Biology dictionary). 40

Figure 1.10: miRNA canonical biogenesis pathway (Figure was created for the purpose of this thesis). 42

Figure 1.11: (A) Example of a perfect miRNA-mRNA seed match. (B) Example of high and low energy regions. High energy indicates weak force and vice versa (Figure was created for the purpose of this thesis). 44

Figure 1.12: Routes of extracellular miRNAs secretion and uptake (Figure was adopted from Yuko Ito et al., 2021) 45

Figure 1.13: Mechanism of target-directed miRNA degradation (Figure was adopted from Charlie Y. Shi et. al., 2020). 47

Figure 1.14: Classification of lncRNAs based on their genomic position in respect to the neighboring coding genes (Figure was adopted from Juliane C. R. Fernandes et al., 2019). 48

Figure 2.1: Types of bacterial shapes and multicellularity (Figure was adopted from RI Krasner et al., 2014). 53

Figure 2.2: (A) Gram+ and Gram- bacterial cell wall. (B) Bacterial growth phases. (C) Bacterial Cell Structure. (D) Horizontal gene transfer processes (Figure was adopted from Wikipedia and NCBI) 54

Figure 2.3: Dominant phyla at different anatomical sites (Figure was adopted from Ilseung Cho et al. 2012). 57

Figure 2.4: The established human microbiome-disease associations (Figure was adopted from Kaijian Hou et al. 2022) 58

Figure 3.1: Overview of AGAMEMNON's workflow. (Figure was adopted from Skoufos G. et al., 2022) 62

Figure 3.2: Overview of AGAMEMNON's quantification algorithm. (Figure was adopted from Skoufos G. et al., 2022) 63

Figure 3.3: A–F The mean squared log error (MSLE) and the number of false positive taxa (FP) between true and estimated read counts at the levels of genus, species, and strain using the Illumina 400 dataset and reference 1. (Figure was adopted from Skoufos G. et al., 2022) 65

Figure 3.4: The mean squared log error (MSLE) and the number of false positive taxa (FP) between true and estimated read counts at the levels of genus, species, and strain using reference 3. (Figure was adopted from Skoufos G. et al., 2022) 66

Figure 3.5: The pairwise Spearman correlation of each method in three human fecal samples at the levels of genus and species. (Figure was adopted from Skoufos G. et al., 2022) 67

Figure 3.6: The mean squared log error (MSLE) and the number of false positive taxa (FP) between true and estimated read counts at the levels of genus, species, and strain using mixed datasets one and two and the human-subset reference. (Figure was adopted from Skoufos G. et al., 2022) 68

Figure 3.7: Screenshots of AGAMEMNON's Shiny application. (Figure was adopted from Skoufos G. et al., 2022) 69

Figure 4.1: Peryton's main user interface. (Figure was adopted from Skoufos G. et al., 2021) 71

Figure 4.2: Peryton visualizations. (Figure was adopted from Skoufos G. et al., 2021) 72

Figure 5.1: Schematic representation of the development and information flow, from the researcher bench to the creation of Agnodice resource. sRNA interactome data, generated by low- or high-yield experimental methods, are contained in hundreds of articles, supplementary materials, or they exist as publicly available raw sequencing libraries. After querying for candidate sources, meticulous curation and analysis procedures are applied. sRNA-RNA interactions are detected either via from scratch analysis of raw NGS data, or by means of manual curation. The resulting set of entries is harmonized, accompanied with rich experiment- or study-specific metainformation and broken down into an efficient database schema. Inter-connection with external resources, including Peryton database of bacterial-disease associations, PubMed and NCBI Taxonomy, is performed. The resulting content is provided in the form of an open, user-friendly online database, supporting numerous querying, filtering, visualization and download functionalities (Figure was created for the purpose of this thesis) 74

Figure 5.2: (A) Top 12 most frequent sRNA regulators and (B) top 12 most frequent regulated genes catered in Agnodice. (C) Top 8 species in terms of total interactions in the database. (D) Number of total interactions per experimental method (interactions by low-yield methods summed together). Interaction sums are transformed in log10 space (Figure was created for the purpose of this thesis) 75

Figure 5.3: Main query/result interface offered in Agnodice. 77

Figure 6.1: Screenshot of DIANA-LncBase v3.0 expression module. (Figure was adopted from Karagkouni G. et al., 2020) 80

Figure 6.2: DIANA-LncBase v3.0 visualizations. (Figure was adopted from Karagkouni G. et al., 2020) 81

Figure 7.1: Bar plots presenting the most abundant miRNAs in terms of mean read counts from the first dataset (Figure was created for the purpose of this thesis) 84

Figure 7.2: Box plots presenting the most abundant miRNAs in terms of median read counts from the first dataset (Figure was created for the purpose of this thesis) 84

Figure 7.3: Bar plot presenting the most abundant miRNAs in terms of mean read counts from the second dataset (Figure was created for the purpose of this thesis) 85

Figure 7.4: Box plot presenting the most abundant miRNAs in terms of median read counts from the second dataset (Figure was created for the purpose of this thesis) 85

Figure 7.5: Bar plot presenting the most abundant bacterial species in the human gut using datasets from the Human Microbiome Project (Figure was created for the purpose of this thesis) 87

Figure 7.6: Box plot presenting the most abundant bacterial species in the human gut using datasets from the Human Microbiome Project (Figure was created for the purpose of this thesis) 88

Figure 7.7: Example of a predicted RNA secondary structure. Hairpin, Internal, Bulge and Multibranch loops are all single-stranded RNA regions and thus more easily accessible by miRNAs (Figure was created for the purpose of this thesis) 89

Figure 7.8: The distribution of duplex structure energy, GU wobbles in miRNA seed, consecutive miRNA seed matches and number of accessible nucleotides in the gene sequence for all predicted interacting miRNAs-genes for the bacterial strains Enterococcus faecium ATCC 8459 = NRRL B-2354 (NC_020207.1), Streptococcus suis SC84 (NC_012924.1), Buchnera aphidicola str. Bp (NC_004545.1) and Escherichia coli K12 (NC_000913.1) (Figure was created for the purpose of this thesis) 91

Figure 7.9: The distribution of duplex structure energy, GU wobbles in miRNA seed, consecutive miRNA seed matches and number of accessible nucleotides in the gene sequence for all predicted interacting miRNAs-genes for the bacterial strains Odoribacter splanchnicus strain NCTC10825 (NZ_LT906459.1), Aeromonas veronii strain FDAARGOS_632 (NZ_CP044060.1), Clostridium baratii strain CDC51267 (NZ_CP014203.1) and Blautia argi strain KCTC 15426 (NZ_CP030280.1) (Figure was created for the purpose of this thesis) 92

Figure 7.10: The distribution of human miRNA-bacterial gene interaction normalized scores for the genes originating from the bacterial strains Butyrivibrio proteoclasticus B316 (NC_014387.1), Bacillus subtilis subsp. subtilis str. 168 (NC_000964.3), Chlamydia trachomatis D/UW-3/CX (NC_000117.1) and Escherichia coli K12 (NC_000913.1) (Figure was created for the purpose of this thesis) 93

Figure 7.11: The distribution of human miRNA-bacterial gene interaction normalized scores for bacterial genes originating from all bacterial genomes together (n = 115) (Figure was created for the purpose of this thesis) 94

Figure 7.12: The distribution of the top human miRNA-bacterial gene interactions in terms of normalized scores (i.e., score >= 0.6) (Figure was created for the purpose of this thesis) 95

Figure 7.13: Frequency of essential and non-essential genes per bacterial genome (subset 1) (Figure was created for the purpose of this thesis) 96

Figure 7.14: Frequency of essential and non-essential genes per bacterial genome (subset 2) (Figure was created for the purpose of this thesis) 96

Figure 7.15: Frequency of essential and non-essential genes per bacterial genome (subset 3) (Figure was created for the purpose of this thesis) 97

Figure 7.16: Frequency of essential and non-essential genes per bacterial genome (subset 4) (Figure was created for the purpose of this thesis) 97

Figure 7.17: Frequency of essential and non-essential genes per bacterial genome (subset 5) (Figure was created for the purpose of this thesis) 98

Figure 7.18: Top 30 bacterial genes in terms of total number of interactions (Figure was created for the purpose of this thesis) 103

Figure 7.19: Examples of human miRNA-bacterial gene interactions (Figure was created for the purpose of this thesis) 104

List of tables

Table 7.1: *Reference bacterial strains from NCBI RefSeq*…… 99

Table 7.2: *Results of goodness-of-fit test (chi-squared test with one dimensional contingency table) to assess the significance of imbalance between essential and non-essential genes. 172 out of the 174 bacterial genomes showed a significant enrichment in miRNA targeting events (interactions) towards essential genes implying a functional role of the predicted interactions (e.g., control of bacterial growth-rate by host miRNAs)* …… 110

Abbreviations

Ribonucleic acid	RNA
RNA Polymerase	RNA Pol I
Messenger RNA	mRNA
Precursor mRNA	pre-mRNA
5' untranslated region	5' UTR
3' untranslated region	3' UTR
Ribonucleoprotein	RNP
Transfer RNA	tRNA
Precursor transfer RNA	pre-tRNAs
Ribosomal RNA	rRNA
Small ribosome subunit	SSU
Large ribosome subunit	LSU
Nucleotides	nt
microRNAs	miRNAs
RNA binding proteins	RBPs
Argonaute RISC Catalytic Component 2	AGO2
DiGeorge Syndrome Critical Region 8	DGCR8
Exportin 5	XPO5
RNA-induced silencing complex	RISC
Mature miRNAs function within an AGO protein	miRISC
Coding	CDS
Minimum free energy	MFE
Extracellular miRNA	ExmiRNA
Microvesicles	MVs
Nanometers	nm
miRNA-encoded peptides	miPEPs
Long non-coding RNAs	lncRNAs
Next-Generation Sequencing	NGS
Small nuclear RNA	snRNA
Small nucleolar RNA	snoRNA
Small interfering RNA	siRNA
Piwi-interacting RNA	piRNA
Circular RNA	circRNA
Vascular endothelial growth factor	VEGF
Glutamate receptor 2	GluA2
Amyotrophic lateral sclerosis	ALS
Human Microbiome Project	HMP
Millimeters	mm
Centimeter	cm
Inclusion bodies	IBs
Horizontal gene transfer	HGT
Full-term	FT
Preterm infants	PT
Short chain fatty acids	SCFAs
Lowest Common Ancestor	LCA

The Cancer Genome Atlas	TCGA
Genome Analysis Toolkit	GATK
Expectation maximization	EM
Paired-end	PE
Single-end	SE
Potentially Removable	PR
Mean Squared Log Error	MSLE
False positive	FP
Principal component analysis	PCA
Multidimensional scaling	MDS
Small non-coding RNAs	sRNAs
RNA interaction by ligation and sequencing	RIL-Seq
Crosslinking, ligation and sequencing of hybrids	CLASH
Global sRNA target identification by ligation and sequencing	GRIL-Seq
Crosslinking and immunoprecipitation followed by sequencing	CLIP-Seq
Ligation of interacting RNA followed by high-throughput sequencing	LIGR-Seq
MS2-affinity purification coupled with RNA sequencing	MAPS
Relative Concentration Index	RCI

INTRODUCTION

CHAPTER 1 – RNA biology

1.1 Fundamentals of RNA

Ribonucleic acid (RNA) is a polymeric molecule essential in almost all known biological processes. The chemical structure of RNA is composed of nucleotides attached by 5’ - 3’ phosphodiester bonds between ribose sugars (*Figure 1.1*). The molecular formula of ribose is different compared to deoxyribose (i.e., DNA sugar) in that it contains a hydroxyl group attached to the pentose ring in the 2' position (*Figure 1.1*). RNA primary structure consists of adenine, guanine (purines), cytosine and uracil (pyrimidines) (*Figure 1.1*). Adenine and uracil form two while cytosine and guanine three hydrogen bonds respectively. Hydrogen bond-based base pairing allows for complex RNA secondary structures[1] (*Figure 1.1*). The RNA tertiary structure[2] is the result of RNA folding and consists of helical duplexes, triple-stranded structures and more (*Figure 1.1*).

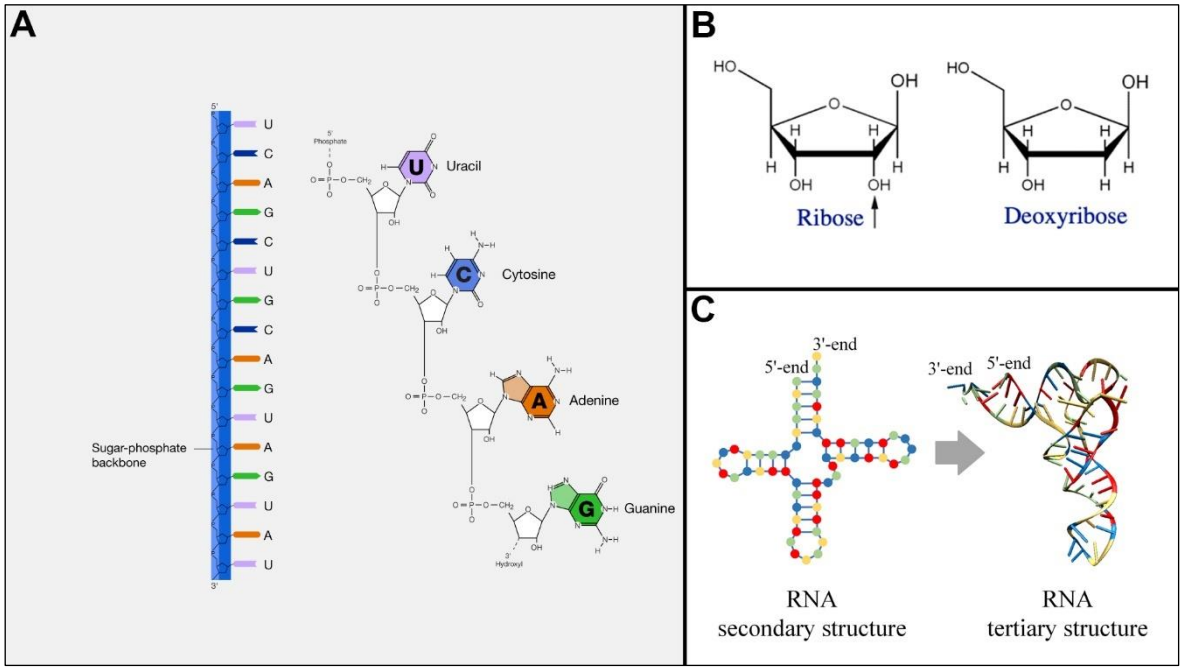

Figure 1.1: (A) Molecular and chemical structure of RNA. (B) Comparison of ribose (RNA sugar) and deoxyribose (DNA sugar). (C) Examples of a secondary and a tertiary structure of an RNA molecule (Figure was modified for the purpose of this thesis. Original figures: National Human Genome Research Institute, study.com and Qi Zhao et al. 2021).

RNA like DNA are nucleic acids which constitute one of the four (the other three being proteins, lipids and carbohydrates) macromolecules vital for all known forms of life. Unlike

DNA which has a double-stranded helix structure, RNA is primarily single-stranded though it can also be found in double-stranded forms. Moreover, with a process known as “thymine ↔ uracil exchange”, thymine found in the DNA chain are converted into its unmethylated form, uracil, to give birth to RNA. Synthesis of RNA is almost always catalyzed by RNA polymerase[3], a special enzyme that uses DNA as a roadmap to produce RNA molecules through the process of transcription. In prokaryotes like bacteria, a single RNA Polymerase (RNA Pol I) synthesizes all types of RNA. On the contrary, transcription in Eukarya occurs by at least three different RNA polymerases (i.e., RNA Pol I, II and III) each of which is responsible for the transcription of certain RNA types.

RNA can be classified (*Figure 1.2*) using many different schemes; some of the most common separation strategies are based on their (i) coding capacity, (ii) cellular localization and (iii) function. In the first case, they are divided into coding (i.e., RNAs that code for proteins) and non-coding RNAs. In the second, they are classified as nuclear or cytoplasmic RNAs though in some cases sub-cellular compartments are also utilized to get higher classification resolution. Finally, classification based on the function of RNA is also employed. For instance, regulatory RNA is a typical umbrella term used to describe a superset of RNA species (e.g., microRNAs, long non-coding RNAs, etc.) with regulatory mechanisms. Other commonly used RNA properties based on which classification is conducted are their structural landscape, biogenesis routes or size.

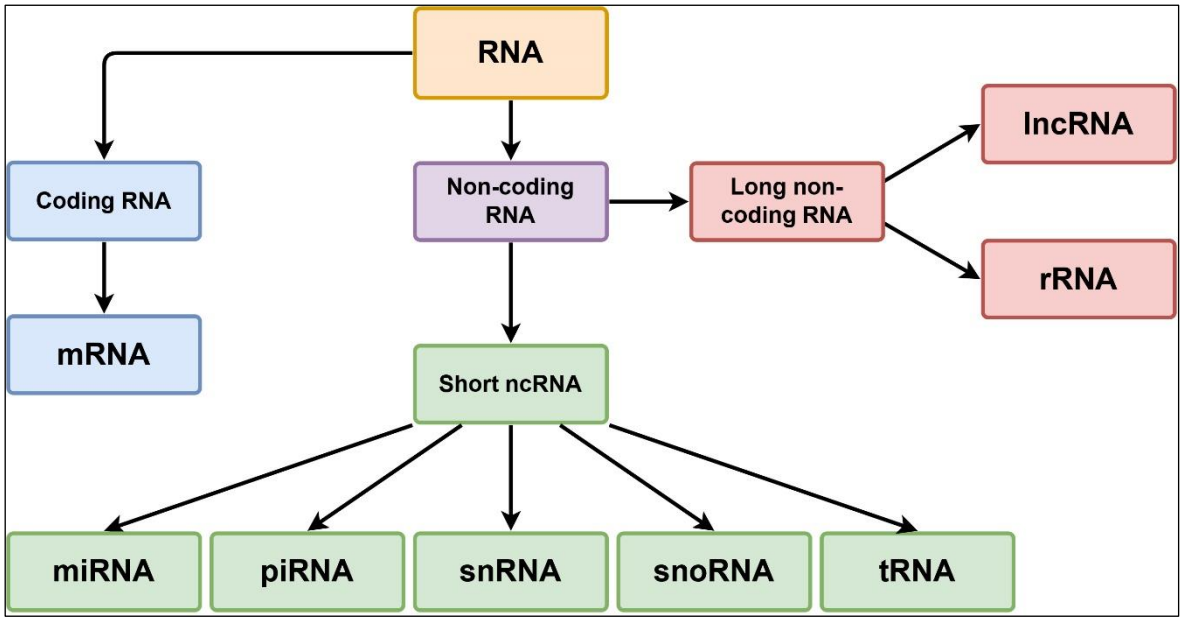

Figure 1.2: A simple classification of some of the known RNA types based on their coding capacity and size (Figure was created for the purpose of this thesis).

Landmark discoveries from the last 80 years[4-10] shaped the field of RNA biology and paved the way for RNA based diagnostics, therapeutics, advances in basic research and biotechnological breakthroughs.

1.2 The magnificent world of RNA species

1.2.1 Messenger RNA (mRNA)

Messenger RNA (mRNA) belongs to a family of single stranded RNA molecules that carries the genetic information from DNA to be translated into proteins[11]. The process by which a DNA sequence is copied to make an RNA molecule is called transcription while the process by which RNA molecules are used to synthesize proteins is called translation (*Figure 1.3*). In eukaryotes, mRNA molecules are synthesized in the cell nucleus using DNA nucleotides as a blueprint; the process of mRNA transcription[11] requires nucleotide triphosphates as substrates and is catalyzed by the enzyme RNA polymerase II. RNA polymerases[12-14] are fundamental to maintain normal cell function and thus, they can be found in all known living organisms[15, 16].

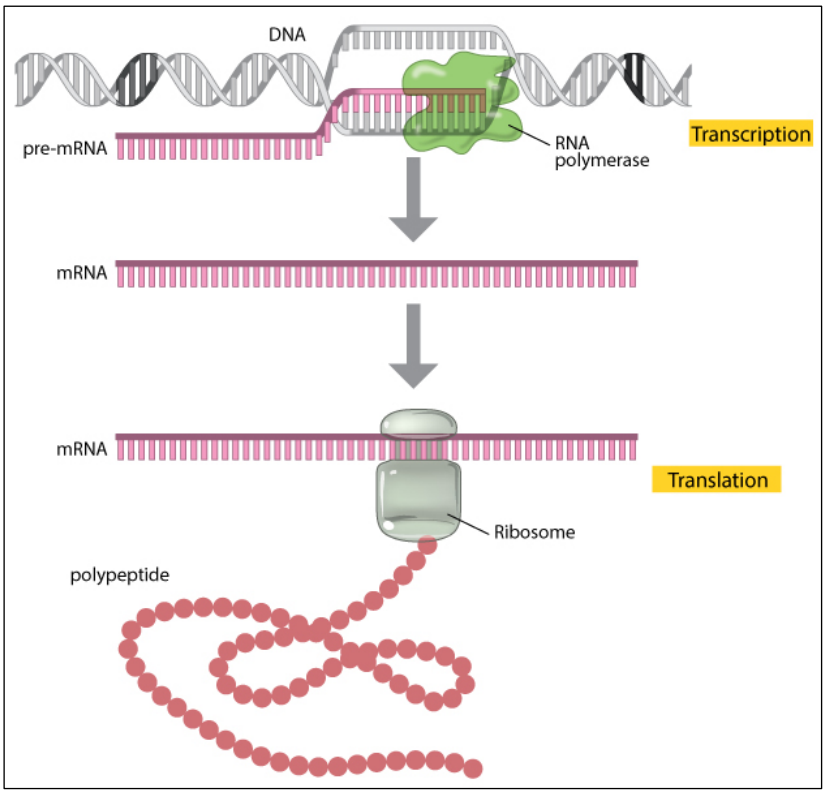

Figure 1.3: Schematic representation of the process of transcription (DNA to RNA) utilizing RNA polymerase II and the process of translation (mRNA to protein) from the ribosome (Figure adopted from: Clancy, S. & Brown, W. (2008) Translation: DNA to mRNA to Protein. Nature Education 1(1):101).

The initial form of an unprocessed mRNA is termed precursor mRNA (pre-mRNA)[11]. The pre-mRNA form includes exons, introns and the 5' and 3' untranslated regions (5' and 3' UTRs). Through the process of mRNA processing, introns are removed by a mechanism known as splicing and exons are joined together (*Figure 1.4*)[17]. Splicing is carried out by

spliceosomes[18], a ribonucleoprotein (RNP) complex found primarily within the nucleus of eukaryotic cells. Furthermore, the 5' cap addition is occurred shortly after the start of transcription[19]. That is, a modified guanine nucleotide that is added in the front of the 5' end of the mRNA molecule. This addition is extremely crucial for (i) the recognition of mRNA by the ribosome and (ii) the protection of mRNA molecules from ribonucleases, special enzymes responsible for the degradation/cleavage of RNA[20]. Finally, multiple adenosine monophosphates are added to the 3' end of mRNA (polyA tail) by a molecular process called polyadenylation[21, 22]. This process occurs at the same time or immediately after transcription. PolyA tails play catalytic role in mRNA protection, stability and export from the nucleus to the cell cytoplasm. Of note, a recently published study showcases a strong correlation between polyA tail length and translational efficiency of mRNA transcripts[23].

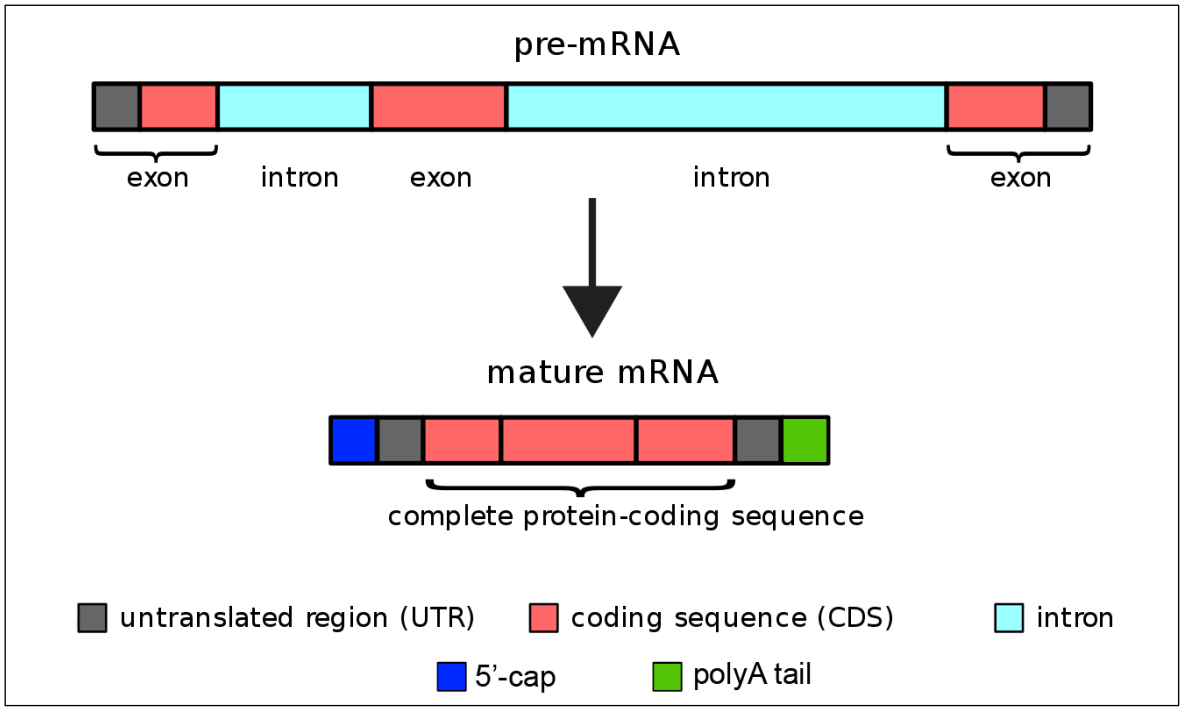

Figure 1.4: Illustration of a precursor mRNA (pre-mRNA), a mature mRNA and their components (Original figure (source: Wikipedia) was modified for the purpose of this thesis).

Maturation of pre-mRNA has multiple different paths. The most well-studied is alternative splicing (Nobel prize in Physiology and Medicine, 1993, Richard J. Roberts και Phillip A. Sharp)[8, 24]. During alternative splicing (*Figure 1.5*), multiple different mRNA transcripts can be produced from the same DNA sequence of a gene. Alternative mRNA molecules are also termed gene isoforms or just isoforms and emanate from the arrangement of exons

and/or introns in different combinations[25]. Different gene isoforms can produce proteins with different amino acid sequences and ultimately different function.

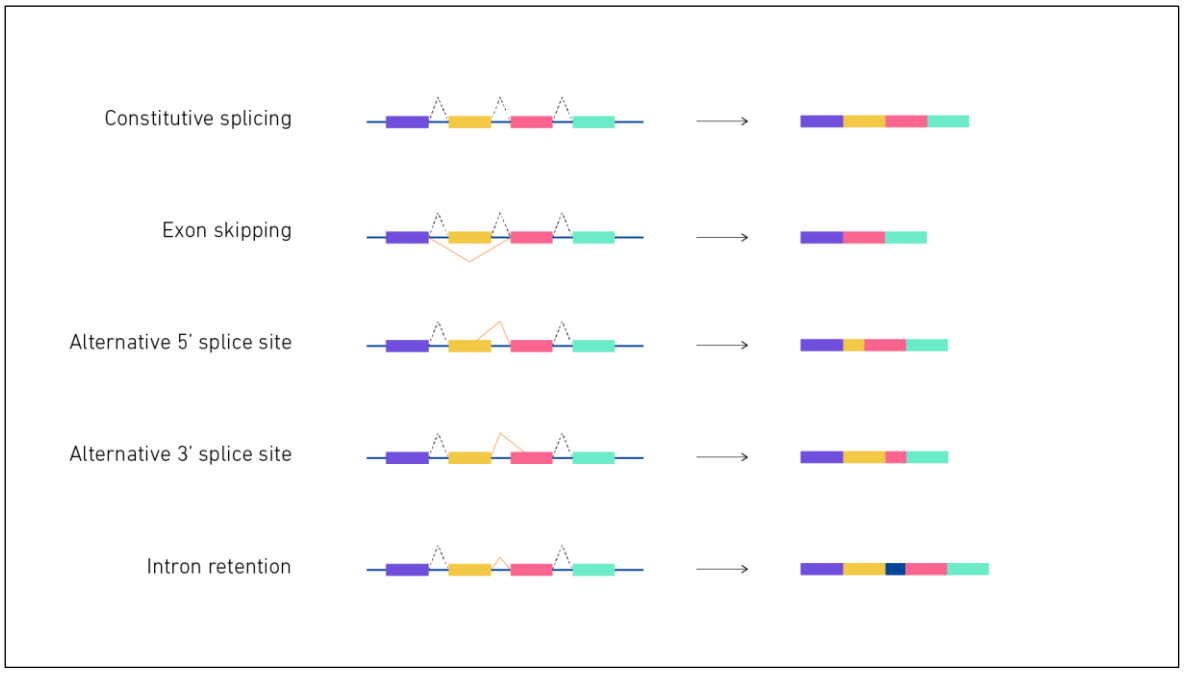

Figure 1.5: Overview of some of the most common types of alternative splicing (Figure adopted from: Jonathan Dornell, Alternative Splicing: Importance and Definition, Genomics research).

Alternative splicing constitutes an important mechanism that extends the molecular arsenal of eukaryotic cells. For instance, while *Homo sapiens* comprise ~20,000 protein-coding genes, the total number of unique products that can be produced by these genes exceeds 100,000 proteins.

Despite the fact that eukaryotic and prokaryotic mRNA molecules have many structural and functional differences, they also share many similarities. In messenger RNA, genetic information is encoded using an alphabet of four nucleotides. These nucleotides comprise adenine (A), cytosine (C), guanine (G) and uracil (U). Different combinations of threes of these nucleotides form the codons. Each codon codes for a specific amino acid (*Figure 1.6*) with the exception of stop codons which specify the termination of protein synthesis from the ribosome. The start codon is usually the nucleotide triplet AUG which signals for the initiation of protein synthesis in both eukaryotic and prokaryotic species. Often, alternative start and stop codons are also encountered.

In eukaryotes, most of the times, a single mRNA molecule codes for a single protein (monocistronic mRNA)[26]. On the contrary, mRNA in prokaryotes usually codes for a series of different proteins (polycistronic mRNA)[26]. For instance, trp[27] operon is a well-studied polycistronic mRNA that constitutes a DNA region which codes for six polypeptides that catalyze the tryptophan biosynthesis. Bacterial mRNAs has a much shorter half-life

compared to their eukaryotic counterparts[28] which also function as a mechanism of gene regulation. Finally, even though eukaryotic mRNAs require extensive processing and export from the cell nucleus, bacterial mRNAs follow a different path of maturation since they do not acquire a 5' cap structure and they rarely contain a polyA tail.

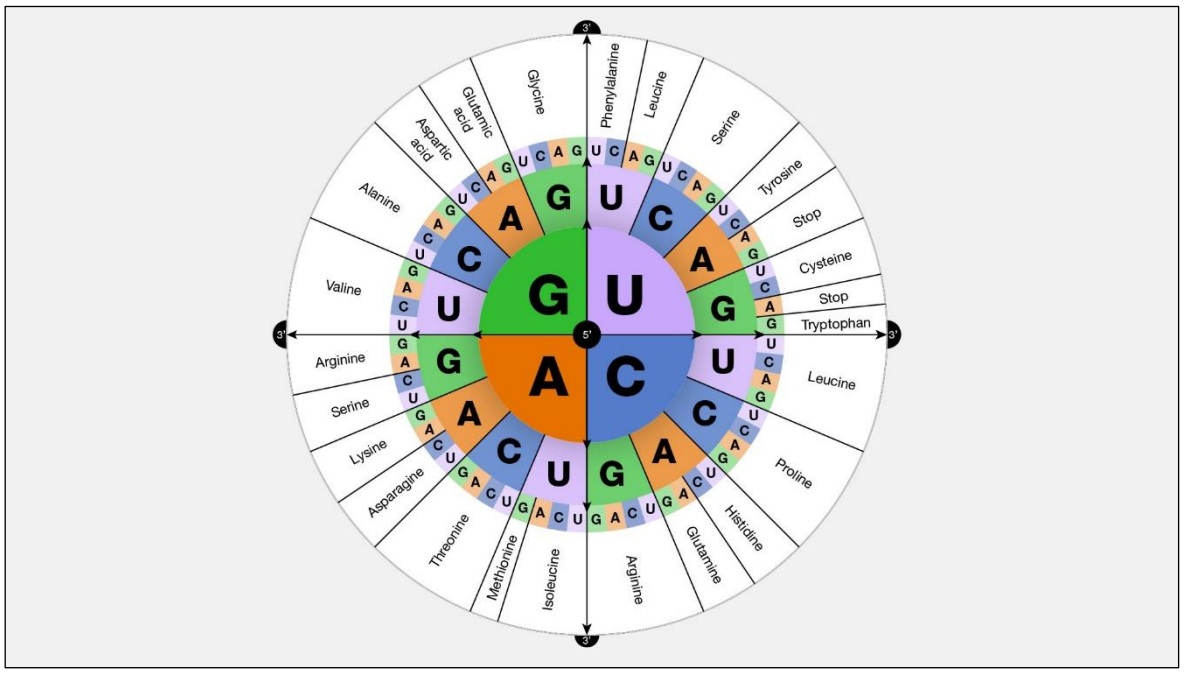

Figure 1.6: Presentation of the genetic code blueprint. Each codon (combination of three nucleotides) translates in one of the 20 available amino acids (Figure adopted from: National Human Genome Research Institute).

1.2.2 Transfer RNA (tRNA)

Transfer RNA (tRNA)[29] constitutes a molecule typically ranging between 76 – 90 nucleotides that acts as the physical link between the mRNA molecule and the amino acid sequence of a protein[9]. This process (*Figure 1.7*) is carried out with the transfer of amino acids from tRNAs to the ribosome. tRNAs can be found in all three kingdoms of life (Bacteria, Archaea, Eukarya). The complementarity between mRNA codons and tRNA anticodons results in the accurate protein synthesis based on the genetic information that is carried out by DNA through mRNAs. Therefore, tRNAs, mRNAs and the ribosome are the three essential components of protein synthesis.

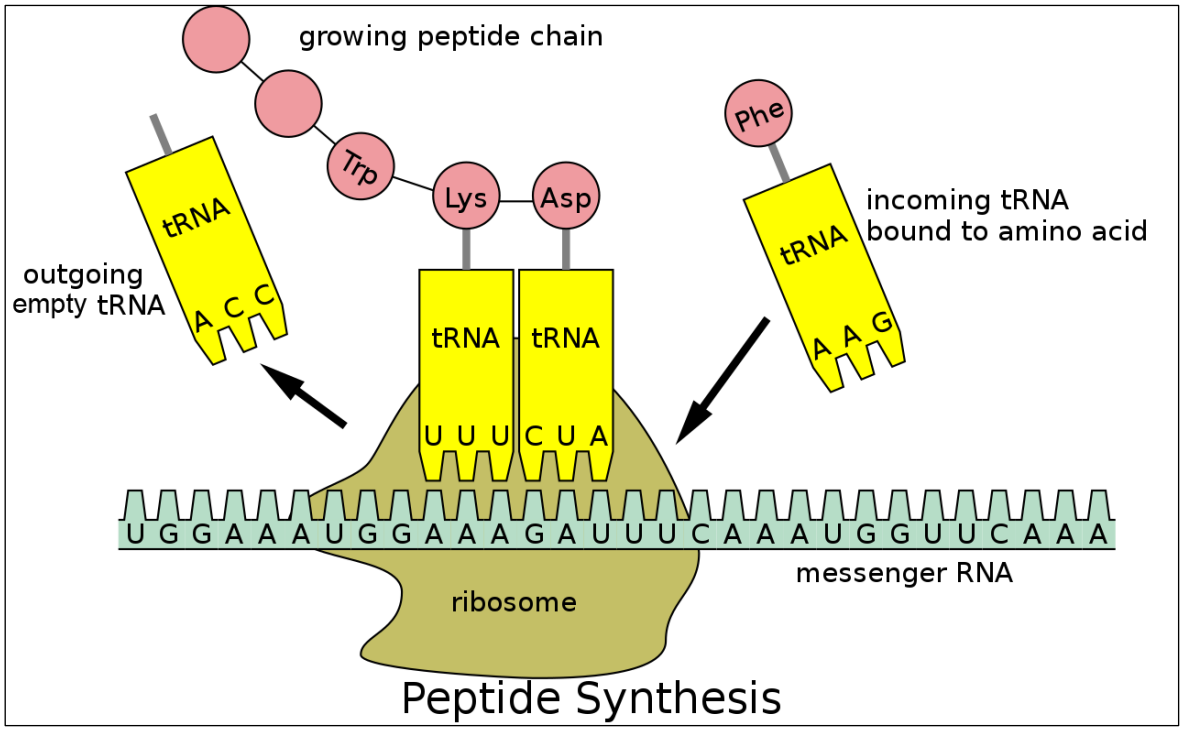

Figure 1.7: Illustration of the process of protein synthesis in the presence of an mRNA, tRNAs and the ribosome (Figure adopted from: Wikipedia).

tRNAs have a characteristic secondary structure that resembles the structure of a cloverleaf. During transcription, in its initial form, a tRNA recommends a single-stranded RNA molecule. However, soon after transcription ends, complementary nucleotides assemble base pairs in different regions of the tRNA forming double-stranded regions and loops. Thus, in their final form, tRNAs constitute double-stranded molecules with an *L*-shaped territory structure (*Figure 1.8*). In eukaryotes, transcription of tRNAs occurs by RNA polymerase III or Pol III. In their precursor form (pre-tRNAs), tRNAs carry extra nucleotides in their 5' and 3' ends. These sequence parts are termed "5' leader" and "3' trailer" regions. The 5' leader and 3' trailer regions are removed from RNase P and RNase Z respectively. Furthermore, tRNAs also include introns which are subjected to splicing by utilizing the TSEN nuclear endonuclease complex.

The human genome encodes for more than 500 tRNAs. Nevertheless, a recently published study[30] suggests that almost half of them are either lowly expressed or completely silent. Generally, there is great diversity in the number of tRNAs among eukaryotes ranging between 170 to 570 tRNAs[31]. On the contrary, bacteria, in proportion to their genome size, usually comprise a lower number of tRNAs. For instance, the model species *Escherichia coli* employs ~46 tRNAs[32]. Expectantly, in prokaryotes and other unicellular organisms, the abundance of each tRNA positively correlates with the codon usage of highly expressed genes[33].

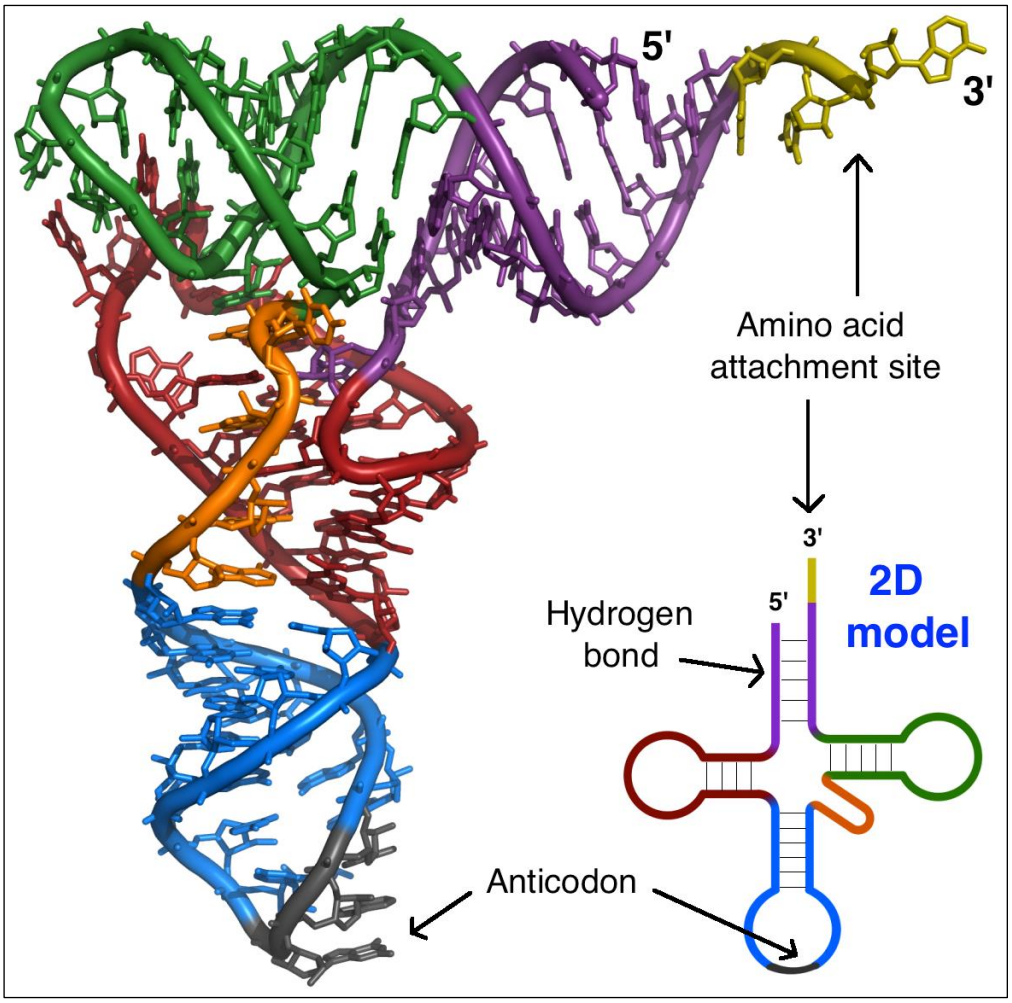

Figure 1.8: The secondary and territory structure of a typical transfer RNA (Figure was adopted from: Wikipedia).

1.2.3 Ribosomal RNA (rRNA)

Ribosomal RNA (rRNA)[34, 35] is a class of non-coding RNA species which is one of the primary building blocks of the ribosome[36]. As other RNA species, their transcription occurs in the cell nucleus and subsequently exported to the cytoplasm to bound to ribosomal proteins and form the small (SSU) and large (LSU) ribosome subunits. Usually, a ribosome consists of ~60% rRNAs and ~40% ribosomal proteins[37]. During mRNA translation, rRNA is responsible for the binding of mRNA and tRNA in order to aid with the process of translating codons into amino acids. rRNA catalyzes protein synthesis when tRNA intercedes between the SSU and LSU. Interactions of mRNA's codons with tRNA's

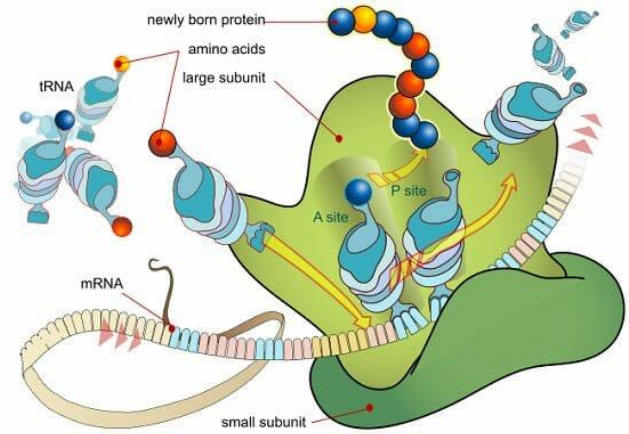

Figure 1.9: Illustration of the SSU and LSU of the ribosome, an mRNA under translation and tRNAs that carry amino acids to a newly synthesized amino acid chain (Figure was adopted from: Biology dictionary).

anticodons takes place in the small ribosome subunit. In the large subunit, tRNA amino acid acceptor interacts with the LSU rRNA which has a characteristic three-dimensional shape with internal loops and helices that form specific sites (A, P, and E sites) within the ribosome[38]. The P-site is responsible for binding a growing polypeptide, while the A-site withholds the incoming tRNA loaded with amino acids. After peptide bond formation, tRNA briefly binds to the E-site before exiting the ribosome (*Figure 1.9*).

Bacterial ribosomes incorporate three types of rRNAs, 23S and 5S rRNAs in the large subunit[39] and 16S rRNA in the small unit[40]. 23S length is close to 3,000 nucleotides (nt) while 5S and 16S close to 1,500nt. These three types of rRNAs are combined with ~50 ribosomal proteins to form the two ribosomal units. In eukaryotes[41], the SSU contains a single small rRNA (~1,800nt) while the LSU consists of two small (~1,800nt) and one large (~5,000nt) ribosomal RNA. The combination of these rRNAs with ~70 proteins assembles the small and large ribosomal subunits. In general, both the small and large ribosomal subunits of eukaryotes are larger compared to their prokaryotic equivalents.

rRNA sequences can be found in all known forms of life and are conserved across species. Their utilization is extremely common to study *inter*- and *intra*-species evolutionary relationships[42, 43] and to conduct phylogenetic analyses especially in the setting of bacterial organisms where the 16S gene helps delineate between species with similar genomic architectures[44, 45]. In eukaryotes such us *Homo sapiens*, rRNA is synthesized by RNA polymerase I using the genetic information that can be found repeatedly in the genome (rDNA genes)[46, 47].

1.2.4 microRNA (miRNA)

microRNAs (miRNAs) are small (~21 nt in length) non-coding RNAs that serve as super regulators of gene expression at the post-transcriptional level[48, 49]. They play important roles in many aspects of molecular biology in both physiological and pathological phenotypes including cardiovascular[50] and neural[51] development, stem cell

differentiation[52, 53], metabolism[54], apoptosis[55], neurodegenerative diseases[56] and tumour[57, 58]. miRNAs were discovered simultaneously in 1993 by two independent labs (V. Ambros and G. Ruvkun) working with the model species *C. elegans[10, 59]*. To this day, miRNAs have been identified and characterized in animals, plants and some viruses. They primarily function through base-pairing with complementary sequences within other classes of coding and non-coding RNA molecules. In the setting of miRNA-mRNA pairs, inhibition of gene expression occurs through one or more of the following mechanisms: (a) mRNA degradation, (b) mRNA destabilization, (c) inhibition of translation[49].

A large body of evidence from recently published studies suggest that miRNAs also have alternative functions; in some cases miRNAs have been identified to (i) upregulate protein expression[60], (ii) activate transcription by direct interactions with DNA[61], (iii) target non-coding RNAs in the cell nucleus[62], (iv) be present in the extracellular space[63], (v) participate in *transkingdom* RNA-RNA interactions[64] and more molecular and cellular processes[65, 66]. Their presence in the extracellular space (e.g., blood plasma), inside exosomes, exosome-like extracellular micro-vesicles or bounded in RNA binding proteins (RBPs) like Argonaute RISC Catalytic Component 2 (AGO2), has been extensively studied under the contexts of cell-cell signaling[67], biomarker potential[68, 69] and more.

1.2.4.1 Biogenesis

The biogenesis of miRNAs starts in the cell nucleus and specifically from DNA regions (i.e., miRNA genes) or clusters of miRNA genes. Additionally, miRNAs can also be transcribed from intronic and/or UTR regions of protein-coding genes. Initially, primary miRNAs (pri-miRNAs) are transcribed by RNA Polymerase II/III and then processed into pre-miRNAs by the microprocessor complex[70] which minimally consists of two proteins, DiGeorge Syndrome Critical Region 8 (DGCR8) which is an RBP and Drosha, a ribonuclease III enzyme. The most common motif recognized by DGCR8 within pri-miRNAs is the N6-methyladenylated four-nucleotide sequence GGAC[71]. Subsequently, Drosha is recruited and the cleavage of the RNA duplex at the base of the hairpin structure takes place yielding the pre-miRNA product. Next, the exportin 5 (XPO5)/RanGTP complex[72] exports the pre-miRNA to the cytoplasm which is further processed by Dicer, an RNase III endonuclease. Dicer removes the terminal loop of the pre-miRNA[73] resulting in the mature miRNA duplex (*Figure 1.10*). The miRNA duplex is then loaded into the RNA-induced silencing complex (RISC), a multiprotein complex that includes the RISC catalytic component AGO2, SND1, AEG-1, FMR1, VIG, R2D2 and Armitage-RNA helicase1[74-76]. Finally, AGO2 is responsible for the removal of one of the two strands of the miRNA duplex. The directionality of the selected strand gives the name to the mature miRNA. If the mature miRNA originates from the 5’ end of the pre-miRNA hairpin it is termed miRNA-5p. On the contrary, if it originates from the 3’ end, it is termed miRNA-3p.

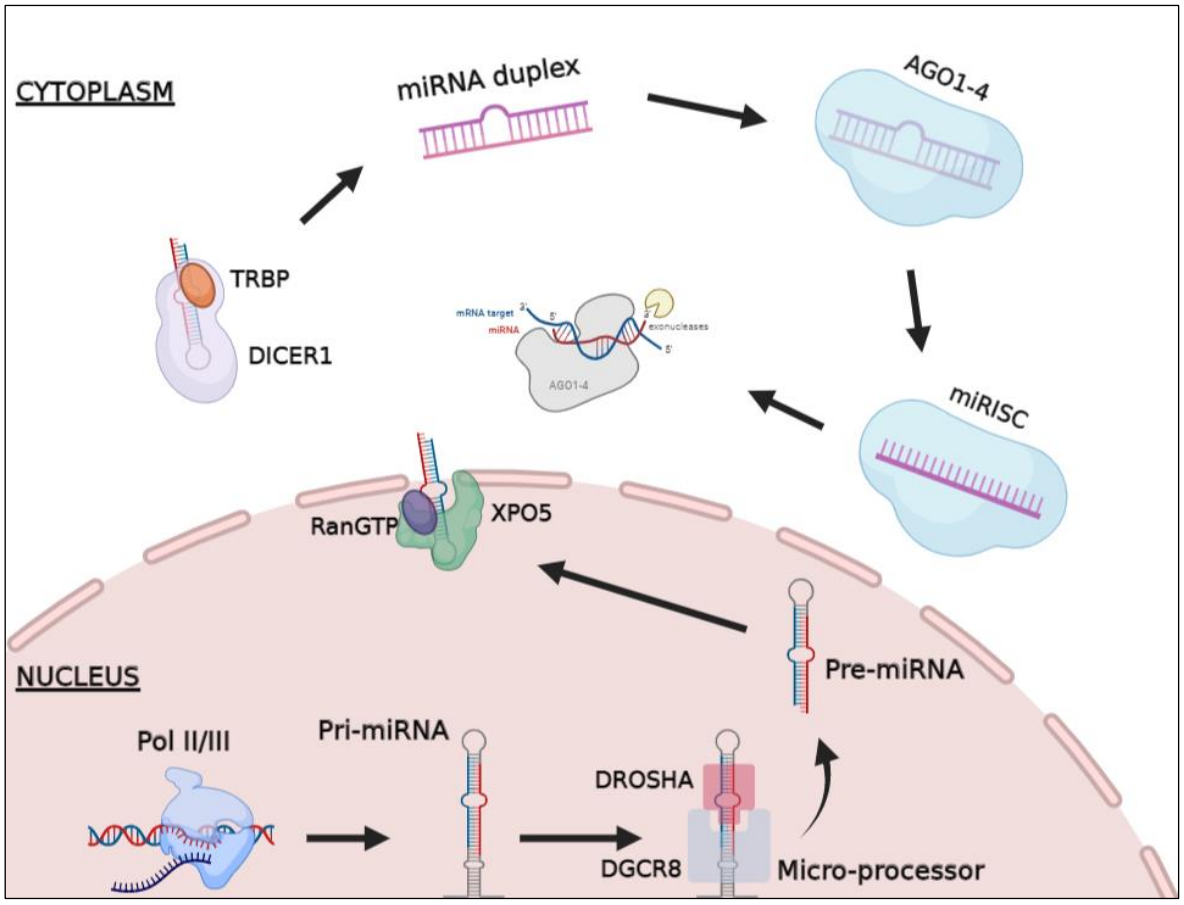

Figure 1.10: miRNA canonical biogenesis pathway (Figure was created for the purpose of this thesis).

Non-canonical miRNA biogenesis pathways also exist; Such pathways can be grouped under two main categories, the Drosha/DGCR8-independent and Dicer-independent pathways. A well-studied example of Drosha/DGCR8-independent pre-miRNAs is MiRtrons[77] which are derived from intronic regions of mRNAs during splicing. Instead of the Microprocessor complex, they are processed by splicing. Moreover, the Dicer-independent derived miRNAs require AGO2 to complete their maturation[78]. This occurs mainly due to their short stem-loop structure which is insufficient for recognition and processing by Dicer. Pre-miR-451 is an example of such miRNA. Finally, another route of non-canonical miRNA biogenesis is the XPO5-independent pathway[79].

1.2.4.2 Primary function

Mature miRNAs function within an AGO protein (miRISC). Mammals encode for four AGO proteins (AGO1-4) yet AGO2 is the most highly expressed; it is also the only protein of the AGO family that can catalyze the cleavage of an RNA target with full complementarity with the guide strand of the miRNA[76, 80]. Their primary and most studied mode of action is the regulation of gene expression at the post-transcriptional level (*Figure 1.11*). Negative

regulation by miRNAs is accomplished through base-pairing with complementary sequences within their mRNA targets; the two basic mechanisms of gene regulation include: (a) mRNA cleavage and (b) translational repression[48]. The mode of regulation (either (a) or (b)) is determined by the degree of complementarity between the miRNA-mRNA pair. mRNA cleavage occurs under perfect or near-perfect complementarity scenarios whereas translational repression takes place in scenarios where miRNA-mRNA complementarity is limited[49]. Additionally, limited complementarity also permits the miRNA-induced deadenylation of their targets, accelerating mRNA degradation[81]. Imperfect base-pairing allows for a single miRNA to regulate many mRNA targets and a single mRNA to be regulated by many different miRNAs. This phenomenon is observed mostly in animals where it adds extra complexity in the endeavors to completely characterize their miRNA targetomes.

Most of the experimentally-supported and/or validated animal miRNA interactions have been shown to have a binding preference towards the 3' UTR of their target genes[82]. miRNA binding sites have also been characterized in other regions of the gene body including coding (CDS) and intronic regions[83], and less frequently the 5' UTR.

1.2.4.3 Basic targeting rules

The characterization of animal miRNA targeting repertoires through experimental (e.g., CLIP-Seq, CLASH, Luciferase Reporter Assay)[84, 85] and computational (e.g., microCLIP, microT-CDS, TargetScan) methods[86-89], led to a core set of miRNA targeting rules/features. Over the years, these rules have been refined and updated many times. Nevertheless, some of them are well-established and universally accepted.

Seed sequence of miRNAs[49, 89] is defined as the seven nucleotides at position 2-8 starting from the 5' end towards the 3' end of the mature miRNA. The rule termed "Seed match" refers to the match (in terms of Watson-Crick base-pairs) between the miRNA seed sequence and a segment of the target sequence (*Figure 1.11*). Such matches occur when adenosine pairs with uracil and guanine pairs with cytosine. Perfect seed match refers to the binding of the miRNA seed with a target without any mismatches or gaps. Distinct seed match types have been identified over the years; each one with different efficiency and characteristics. Four established seed match types include the following: **(1)** 6mer - a perfect match between the miRNA seed and mRNA for six nucleotides (either at positions 2-7 or 3-8 of the seed), **(2)** 7mer-A1 - a perfect match from nucleotides at positions 2–7 of the miRNA seed in addition to an adenine across from the miRNA nucleotide at position 1 of the seed, **(3)** 7mer-m8 - a perfect match from nucleotides at positions 2–8 of the miRNA seed and **(4)** 8mer - a perfect match from nucleotides 2–8 of the miRNA seed in addition to an adenine across from the miRNA nucleotide at position 1 of the seed.

Free energy in the context of miRNA targeting, estimates the stability between a miRNA-RNA pair[90]. According to the second law of thermodynamics, the internal energy of a closed system decreases and will approach a minimum value at equilibrium. The minimum free energy (MFE) between two RNA molecules upon interaction indicates how strong they are bind to each other (*Figure 1.11*). The lower the free energy, the higher the stability and interaction strength between the RNA molecules under investigation. The computational calculation of the optimal MFE in miRNA-mRNA pairs, usually results in more accurate results.

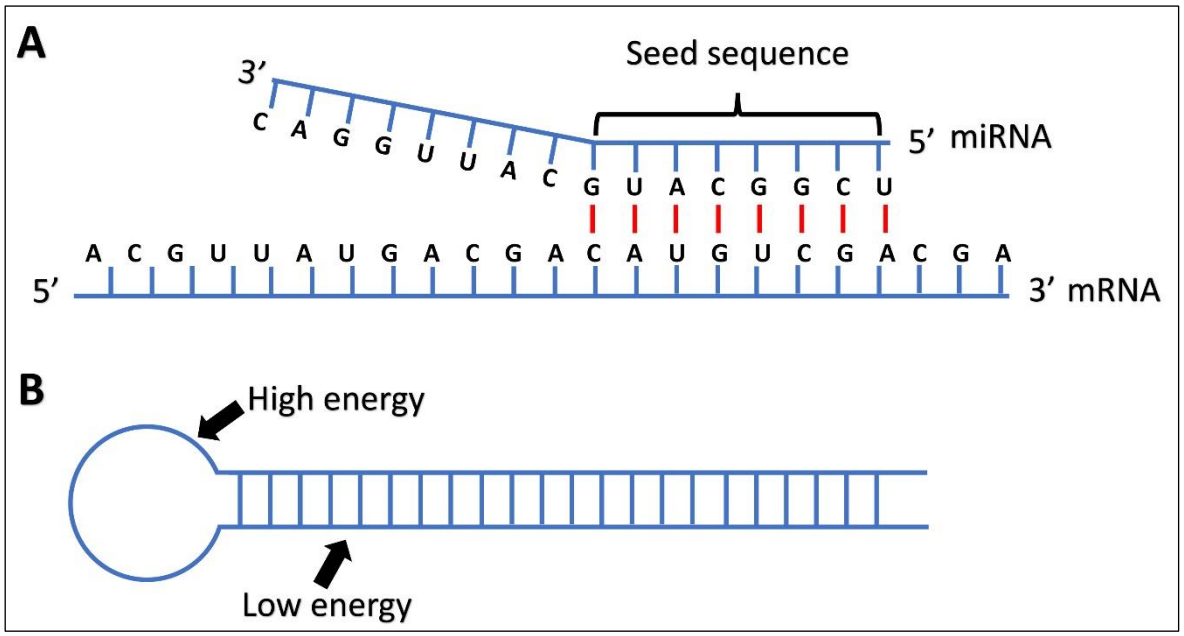

Figure 1.11: (A) Example of a perfect miRNA-mRNA seed match. (B) Example of high and low energy regions. High energy indicates weak force and vice versa (Figure was created for the purpose of this thesis).

Accessibility[91] is a structural characteristic of the RNA targets of miRNAs. After transcription and processing, mRNAs and other RNA species also conceive double-stranded regions; these regions can interfere with the miRNAs ability to bind to the desired target location. Usually, a miRNA initially binds to a short (accessible) region of its target RNA. Subsequently, the target starts to unfold and the miRNA gain ground to ultimately fully bind to the selected region. Expectedly, for an mRNA to unfold, energy needs to be consumed; thus, the likelihood of a miRNA to have the capacity to bind to an mRNA is conversely proportional to the amount of energy required to make a site accessible.

Conservation is another important feature utilized to (a) design target prediction methodologies[88] and (b) distinguish between functional and non-functional miRNA targeting events[92]. In general, conservation refers to the maintenance of a DNA region (e.g., a gene) across different species. In the context of miRNA targeting, the spotlight of

conservation is either on the side of mRNAs, most commonly in their 5’ and 3’ UTRs, the side of the miRNAs, usually in their seed region which is significantly more conserved compared to the rest of the miRNA body or both. When a predicted miRNA target is conserved, it is used as an evidence that the targeting event is functional because it is evolutionary selected/conserved.

1.2.4.4 Extracellular miRNA

In a breakthrough publication in 2007, Valadi *et al.*[93] demonstrated for the first time that a special type of nano-sized biovesicles termed exosomes contain RNA. Analysis of total RNA derived from isolated exosomes revealed an enrichment for mRNAs and miRNAs. Additionally, they showed that exosomal mRNA can be translated after delivered to another cell and function appropriately. Since then, many studies focused on extracellular RNAs established the presence of miRNAs in a variety of body fluids[94] including plasma, urine, bronchial lavage and saliva. Even though the mechanisms of extracellular miRNA (ExmiRNA) transport and their mode of action were under close investigation for the last fifteen years, the field of extracellular RNAs in still under active research.

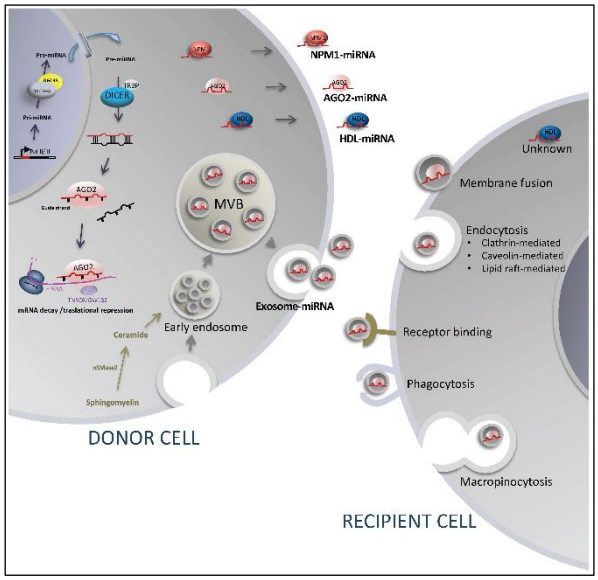

Figure 1.12: Routes of extracellular miRNAs secretion and uptake (Figure was adopted from Yuko Ito et al., 2021)

Endogenous miRNAs are released to the extracellular space (*Figure 1.12*) through the following three routes: (i) passive leakage mainly due to cell necrosis or apoptosis, (ii) active secretion via microvesicles (MVs)[93, 95] which constitute cell-derived membrane particles ranging from 30–100 (exosomes)[96] to 100 to 1000 (shedding vesicles) nanometers (nm) in diameter and (iii) active secretion of miRNAs bound to RBPs (i.e., AGO2, NPM1, HDL)[97-99]. The latter two, protect miRNAs from degradation by RNases of the extracellular space and constitute the models through which ExmiRNAs play regulatory roles in cellular processes[93]. The sorting of miRNAs into EVs, the selection of recipient cells and the uptake of ExmiRNAs by distal cells need further research. Nevertheless, a recently published study[100] demonstrates that the determinant of secretion of miRNAs in EVs is a set of sorting sequences that miRNAs possess. Furthermore, uptake of MV-encapsulated ExmiRNAs by other cells takes place through the recognition of specific surface molecules;

next, internalization of MVs occurs by phagocytosis, endocytosis, or direct fusion with the plasma membrane.

ExmiRNAs thought to function in many different ways; studies from the last years demonstrate that EV-contained miRNAs act as gene expression regulators in the recipient cells[101], play roles in cell-cell communication[67], they orchestrate *transkingdom* RNA-RNA interactions[102, 103] and most importantly, serve as diagnostic and/or prognostic biomarkers of a wide spectrum of pathologies[104-106]. Surprisingly, most of the miRNAs encapsulated in EVs are either in a protein-free form (i.e., naked miRNAs) who's not involve any proteins from the AGO family (AGO1-4) or are bounded to different RBPs (e.g., NPM1) both of which contradict with the current understanding of the miRNA biogenesis pathways and function[107].

1.2.4.5 Atypical miRNA function

Even though miRNAs constitute one of the most studied classes of RNAs, they are primarily known for their ability to post-transcriptionally regulate gene expression. Since their discovery, research efforts have been focused on the identification/prediction and cataloguing of miRNA targets in coding[86, 87, 108], and non-coding[109] transcripts, their capacity to shape a plethora of biological processes[110, 111] and recently their potential as minimally invasive diagnostic/prognostic disease biomarkers[106, 112] and their utilization for therapeutic interventions[113]. Nonetheless, a relatively small body of evidence suggests that miRNAs repertoire goes beyond canonical gene regulation paradigms.

Pri-miRNAs encode regulatory peptides [66] in *Medicago truncatula* and *Arabidopsis thaliana* plant species. These miRNA-encoded peptides (miPEPs) are generated through open reading frame sequences contained in primary miRNA transcripts. Interestingly, the mechanism of action of these miPEPs is to increase the transcription of the pri-miRNAs they originate from ultimately leading to enhanced abundance of their mature form and subsequent downregulation of their corresponding protein-coding gene targets.

miRNAs are found bound to proteins outside the AGO1-4 set. A large body of evidence from the last decade suggest that mature microRNAs interact with additional RBPs including NPM1[99], HDL proteins[114] and hnRNP E2[115]. As said, these findings somehow contradict with the established miRNA knowledge which dictates that AGO proteins are absolutely essential for miRNAs biogenesis and function. Even though NPM1 and HDL proteins are thought to play roles in the transfer of miRNAs into the extracellular space and not in their function *per se*, interactions of miRNAs with hnRNP E2 (specifically miR-328), prevents CEBPA mRNA binding to the same RBP (i.e., hnRNP E2) and in turn rescues CEBPA translation both *in vitro* and *in vivo*[115].

Target-directed miRNA degradation is another uncommon or understudied function of microRNAs. It was first observed in fruit fly by the utilization of synthetic miRNAs

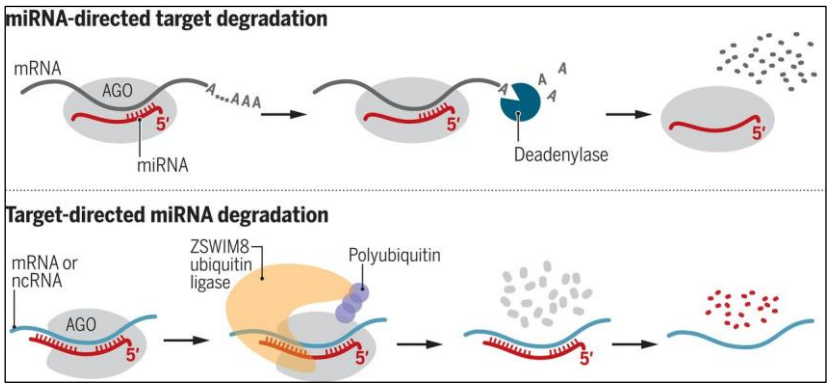

Figure 1.13: Mechanism of target-directed miRNA degradation (Figure was adopted from Charlie Y. Shi et. al., 2020).

specifically engineered to share a high degree of complementarity (beyond the seed region) with their targets[116]. In two recent publications[117, 118], the mechanism of miRNA degradation was elucidated (*Figure 1.13*), revealing that when paired to specific mRNA targets, a ubiquitin ligase (ZSWIM8) is recruited to the miRISC complex catalyzing the degradation of AGO and thus, exposing the miRNA into cellular nucleases ultimately leading to miRNA degradation.

miRNAs targeting repertoire includes mitochondrial transcripts. Mitochondria are membrane-bounded cell organelles with their own genome; they can be found in most known eukaryotic organisms and they generate most of the chemical energy needed to power the cell. In *Homo sapiens*, they encode for 37 genes (13 mRNAs, 22 tRNAs and 2 rRNAs)[119]. A study published in 2012[120] established a miRNA-mitochondrion interplay in which miR-181c translocate into the mitochondrion and regulates the expression of MT-COX1 (downregulation) and MT-COX2 (upregulation) following the canonical targeting pathway. microRNAs with the ability to directly regulate MT-genes are termed mitomiRs and their role in pathophysiology is only recently started to be appreciated[121, 122].

***Transkingdom* RNA-RNA interactions** with the participation of microRNAs have been reported in many contexts and pairs of different organisms including (i) animal miRNA-bacterial mRNA[102, 103], (ii) plant miRNA-parasite mRNA[123-125] and (iii) viral miRNA- host mRNA[126, 127] (and *vice versa*). Especially in the context of animal miRNA-bacterial mRNA interplay, the phenomenon considers to have a global effect in the microbiome (i.e., the totality of the microorganisms in a specific environment) of the host. This could be the result of many host miRNAs regulating the expression of many different mRNAs from many different microorganisms or more likely due to the symbiotic relationships that usually bacterial species share with each other; that is, a microRNA may control the growth rate of a specific bacterium by directly regulating a set of its genes and subsequently a domino effect follows for the rest of the local microbiota. Nevertheless, in a study published in 2016[103], researchers demonstrated for the first time that extracellular miRNAs present in the feaces of a mouse model, could control and shape the gut microbiota in a miRNA-mediated inter-species gene regulation manner. Specifically, these miRNAs can enter bacteria such us *F. nucleatum* and *E. coli* and regulate bacterial gene transcripts which in turn affects bacterial growth rates and ultimately the rest of the gut microbiome.

Furthermore, in a Dicer1$^{\Delta IEC}$ mouse model, a colitis phenotype was observed (paired with uncontrolled growth of the gut microbiota) deeming mature miRNAs necessary for host gut homeostasis. Finally, restoration of the healthy gut microbiome was achieved through fecal miRNA transplantation.

1.2.5 Long non-coding RNA (lncRNA)

Long non-coding RNAs (lncRNAs) are a diverse class of non-coding RNA molecules exceeding 200nt in length. The last decade, advancements in Next-Generation Sequencing (NGS) methodologies, multi-modal data integration techniques and community efforts have increased the number of annotated lncRNAs tremendously. For instance, the species *Homo sapiens* comprise more than 270,000 annotated lncRNAs[128]. Even though the number of annotated human lncRNAs is impressive, less than 2,000 of them (i.e., < 1%) are studied in depth.

Most of the times, lncRNAs follow the biogenesis route of mRNAs; they are transcribed in the cell nucleus by RNA Polymerase II, acquire a 5′ methyl-cytosine cap and a 3’ polyA tail though they can also exist without polyadenylation[129]. They are also subject to canonical and alternative splicing leading to different isoforms from the same genomic loci[130]. Their biogenesis can also be regulated by several epigenetic modifications and can be processed by additional non-canonical mechanisms. For example, RNase P-induced cleavage of some lncRNAs produces mature 3’ ends[131]. Based on the genomic architecture (*Figure 1.14*) of their neighborhood, lncRNAs are usually classified into four wide categories as follows: (1) intergenic lincRNAs, (2) intronic lncRNAs, (3) bidirectional and (4) antisense lncRNAs. Additionally, they can also be categorized as nuclear or cytoplasmic lncRNAs since they are present in both cellular compartments[109] having a selective tendency over the one or the other.

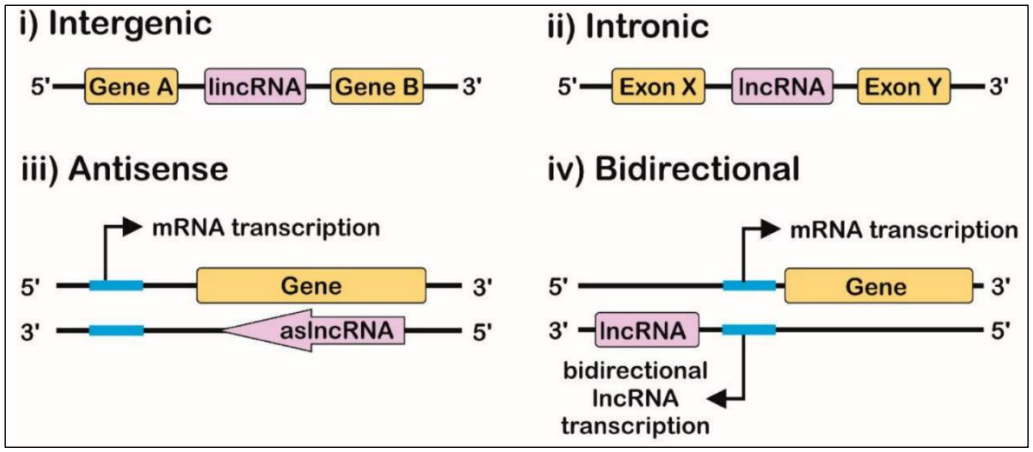

Figure 1.14: Classification of lncRNAs based on their genomic position in respect to the neighboring coding genes (Figure was adopted from Juliane C. R. Fernandes et al., 2019).

Nuclear lncRNAs can regulate gene transcription by recruiting chromatin modifying complexes and regulating transcription factors[132]. Moreover, they can regulate gene expression by controlling chromosomal conformations or pre-mRNA splicing[133]. Cytoplasmic lncRNAs regulate mRNA expression by controlling mRNA stability[134], mRNA translation[135], or by competing for microRNA binding[136, 137]. Finally, production of small active peptides is also feasible for a subset of lncRNAs[138, 139].

1.2.6 Other RNA species

The RNA universe is constantly expanding with the addition of new RNA types with different biogenesis avenues and diverse function.

Small nuclear RNA (snRNA) is a class of small RNA molecules that are found in the cell nucleus of Eukarya; their median length is ~150nt. They are transcribed by RNA Pol II/III[140] and function as processing machines of pre-mRNAs in the nucleus. They have also been shown to aid in the regulation of transcription factors, RNA polymerase II and the maintenance of the telomeres[141].

Small nucleolar RNA (snoRNA) is a type of functional small non-coding RNA species[142]. They are transcribed by RNA Pol II and their primary mode of action is to guide chemical modifications of other RNA types (e.g., rRNAs, tRNAs). snoRNAs can be divided in the following two[143] broad categories: (i) C/D box snoRNAs which guide methylation and (ii) H/ACA box snoRNAs which guide pseudouridylation. Their length ranges between 60 to 300nt. Interestingly, is some rare cases, snoRNAs can also act as miRNAs[144].

Small interfering RNA (siRNA) is a class of small non-coding double-stranded RNA molecules (~21nt in length) which act within the RNA interference (RNAi) pathway. siRNAs impede the expression of protein-coding genes that share sequence complementarity causing post-transcriptional mRNA degradation and thus, averting translation[145].

Piwi-interacting RNA (piRNA) is the most abundant class of small non-coding RNA in Eukarya with their length ranging between 26 to 31 nucleotides[146]. Even though their role has not yet been fully elucidated, it is thought that piRNAs primary function is to guide PIWI proteins to induce cleavage of RNA targets[147] and/or methylation of DNA[148]. They possess unique challenges in terms of annotation, experimental verification and abundance estimation and thus, piRNAs constitute one of the most challenging research topics in RNA biology[149].

Circular RNA (circRNA) is a special type of single-stranded eukaryotic RNA molecule that unlike linear RNAs, it forms a covalently closed continuous loop (i.e., joined 5’ and 3’ ends). They originate mostly from pre-mRNA transcripts through a non-canonical splicing event called back-splicing[150]. Back-splicing is a splicing mechanism in which the 5’

terminus of a pre-mRNA upstream exon is non-colinearly spliced with the 3' terminus of a downstream exon; their unique structural properties (lack of 5' and 3' ends) provide them with resistance to exonuclease-mediated degradation. While the function of most circRNAs remains a mystery, a small subset of them are shown to encode proteins[151] or act as gene regulators[152].

1.3 RNA in pathology

While disease-causing mutations at the DNA level have been known and extensively studied for decades[153], RNA implications in human pathologies have relatively recently started to be appreciated[154]. Even though mutations at the RNA level are inherited from DNA, post-transcriptional RNA modifications allow for the differentiation of RNA at the sequence level. Thus, additional implications can arise even if the DNA region from which an RNA transcribed contains no mutations at all. This is one of the many molecular mechanisms by which RNA can induce pathogenesis. Additional RNA-based mechanisms that contribute to disease phenotypes include: misfolded RNA structures[155], over- and/or under-expressed coding or non-coding RNAs[156], extensive regulation of specific protein-coding genes by miRNAs, RNA-protein interactions[157] and more.

Due to the computational and experimental obstacles governing the study of a subset of the aforementioned mechanisms (e.g., RNA folding), the knowledge is limited; on the other hand, for some mechanisms (e.g., dysregulation of RNA), there are numerous studies contributing to the overall knowledge of RNA pathology. For instance, up-regulation of miR-9 has been found to be activated by c-Myc, resulting in cancer metastasis. Mechanistically, miR-9 suppresses E-cadherin at the mRNA level which in turn promotes b-catenin signaling leading to the overexpression of vascular endothelial growth factor (VEGF), ultimately inducing epithelial-to-mesenchymal transition, cell motility, angiogenesis and metastasis[158].
In a different axis, the glutamate receptor 2 (GluA2), which functions as a ligand-gated ion channel in the central nervous system and plays a major role in excitatory synaptic transmission, is edited in many of its nucleotides under physiological conditions. However, in individuals with sporadic amyotrophic lateral sclerosis (ALS), motor neuron GluA2 mRNA has been found unedited[159]. Furthermore, ADAR2, the protein that catalyse widespread A-to-I editing within RNA sequences, found to be decreased in expression in ALS patients[160]. This and other similar studies, indicate a vital role of A-to-I editing in proper neuronal functioning in mammals.

CHAPTER 2 - Microbiome

2.1 Microbiome and metagenomics

For more than a decade, the word microbiome was defined as the totality of microorganisms present in a specific environment. Moreover, microbiota was practically used as a synonymous word and they both could be used under the same context. In 2020, a panel of experts separated the two words[161]; the meanings they introduced are the following:

"*The microbiome is defined as a characteristic microbial community occupying a reasonable well-defined habitat which has distinct physio-chemical properties. The microbiome not only refers to the microorganisms involved but also encompass their theatre of activity, which results in the formation of specific ecological niches. The microbiome, which forms a dynamic and interactive micro-ecosystem prone to change in time and scale, is integrated in macro-ecosystems including eukaryotic hosts, and here crucial for their functioning and health.*"

"*The microbiota consists of the assembly of microorganisms belonging to different kingdoms (prokaryotes (bacteria, archaea), eukaryotes (algae, protozoa, fungi etc.), while "their theatre of activity" includes microbial structures, metabolites, mobile genetic elements (such as transposons, phages, and viruses), and relic DNA embedded in the environmental conditions of the habitat.*"

In other words, their proposal updated the term "microbiome" to also include the function of microorganisms at the individual and/or collective level. Even though this is an interesting definition, in reality, most scientists of relevant fields keep using both words interchangeably. On the other hand, metagenomics is the study of genetic material directly collected from environmental sources[162, 163]. Additionally, in metagenomics based sequencing techniques (i.e., 16S rRNA-Seq and Shotgun metagenomics), there is no need for isolation of individual species and lab cultivation. Finally, the term "metatranscriptomics" has the exact same meaning though it refers to the study of RNA material while in metagenomics DNA is in the spotlight. The study of metagenomics has offered us novel views of the world around us and within, while shotgun sequencing has revolutionized the field offering higher resolution and throughput. Next-Generation Sequencing techniques have shed light to the complex host-microbiome relationships in humans and other organisms, enabling detailed or even population-scale studies. The Human Microbiome Project (HMP)[164] and other similar studies revealed the importance of the microbiome and its implications in pathological conditions, including gastrointestinal tract

inflammatory diseases[165], neoplastic conditions[166-168], metabolic disorders[169], neurodegenerative diseases[170], and adverse outcomes in pregnancy[171].

2.2 Bacteria

Bacteria are prokaryotic microorganisms present in most habitats on earth. They are mostly unicellular organisms though there are some classes of multicellular bacteria like *Actinomycetes*, *Chloroflexi* and *Magnetomorum rongchengroseum*. In general prokaryotes consist of two different domains sharing an ancient common ancestor, Bacteria and Archaea. The first ever observation of a *bacterium* was made by Antonie van Leeuwenhoek in 1676 using a self-made microscope with a single lens[172].

Three main characteristics assemble the morphology of a *bacterium*; size, shape and multicellularity. Their average size (diameter) ranges between 0.2 – 2 micrometers (μm). The smallest known bacteria range between 200 – 500 nanometers; for instance, the species *Mycoplasma gallicepticum* has an average size between 200 – 300 nm. On the contrary, one of the largest bacterial species ever discovered is *Thiomargarita namibiensis* which may grow to be as large as 0.75 millimeters (mm). To date, the largest known bacterial species is with an unprecedented size of one centimeter (cm)[173]. In terms of shape (*Figure 2.1*), most bacteria are either rod shaped (*bacilli*) or spherical (*cocci*); additionally, there are comma-shaped (*vibrio*) and spiral-shaped (*spirilla*) bacteria. The shape of a *bacterium* is primarily being formed based on its cytoskeleton and cellular wall characteristics; unusual bacterial shapes (e.g., star-shaped bacteria) have also been characterized. Shape is extremely important for their (i) movement in liquids, (ii) attachment to surfaces and (iii) acquisition of nutrients[174]. Finally, multicellularity (*Figure 2.1*) is another important factor of bacterial morphology. Most known bacteria exist as unicellular microorganisms (i.e., single, independent cells); moreover, they can be found as diploids (*neiserria*), chains (*streptococci*) and in grape like structures (*staphylococci*). Sometimes, multicellularity is condition-dependent (i.e., same bacteria can be observed in different structures under different environmental stimuli)[175].

Their cellular structure (*Figure 2.2*) comprise a plasma membrane which mainly consists of phospholipids; the plasma membrane encapsulates the cytoplasm, ribosomes, inclusion bodies, plasmids and the nucleoid (i.e., DNA). The genetic material of bacteria is usually a single circular chromosome but a few exceptions exist. Plasmids are small extrachromosomal circular DNA molecules with the ability to replicate independently. Like other prokaryotic and eukaryotic organisms, ribosomes are the protein production machines of bacteria but their structure divergence from that of eukaryotic cells and Archaea[176].

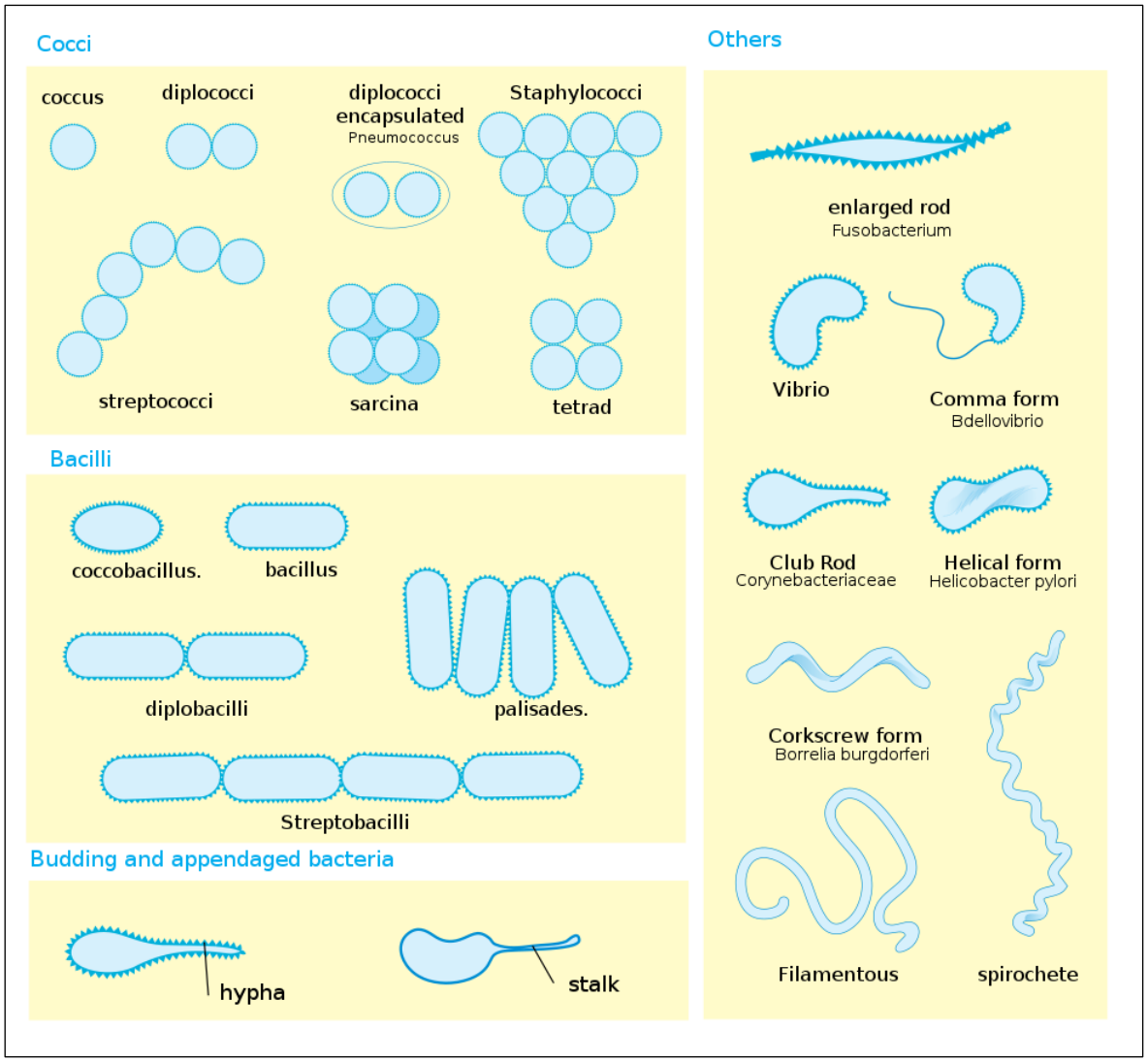

Figure 2.1: Types of bacterial shapes and multicellularity (Figure was adopted from RI Krasner et al., 2014).

Finally, inclusion bodies (IBs)[177] are functional aggregates of proteins, usually assembled under environmental stress conditions. Even though the formation mechanism of IBs is not completely understood, it is thought that they are part of the quality control system of bacterial cells. Bacterial extracellular structure (i.e., outside of the cell membrane), consists of the cell wall and the cell capsule. Additionally, many bacteria comprise *pili*, *flagella* and *fimbriae* (*Figure 2.2*). The cell wall is made of peptidoglycan which is crucial for the survival of many bacterial species. For instance, penicillin, the first and most widely used antibiotic, interferes with the synthesis of peptidoglycan[178] to kill bacteria.

In bacteria, one of the most widely used classification strategies utilizes characteristics of their cell wall to divide them into two broad categories; Gram-positive and Gram-negative bacteria (*Figure 2.2*). The technique conducted to classify bacteria in one of the aforementioned groups is called "[179]" and it was developed by Hans Christian Gram in

1884. In Gram-positive bacteria, there is a thick peptidoglycan layer and they lack an outer lipid membrane. On the contrary, the peptidoglycan layer of Gram-negative bacteria is thin but they contain an additional lipid membrane in the outer part of the cell.

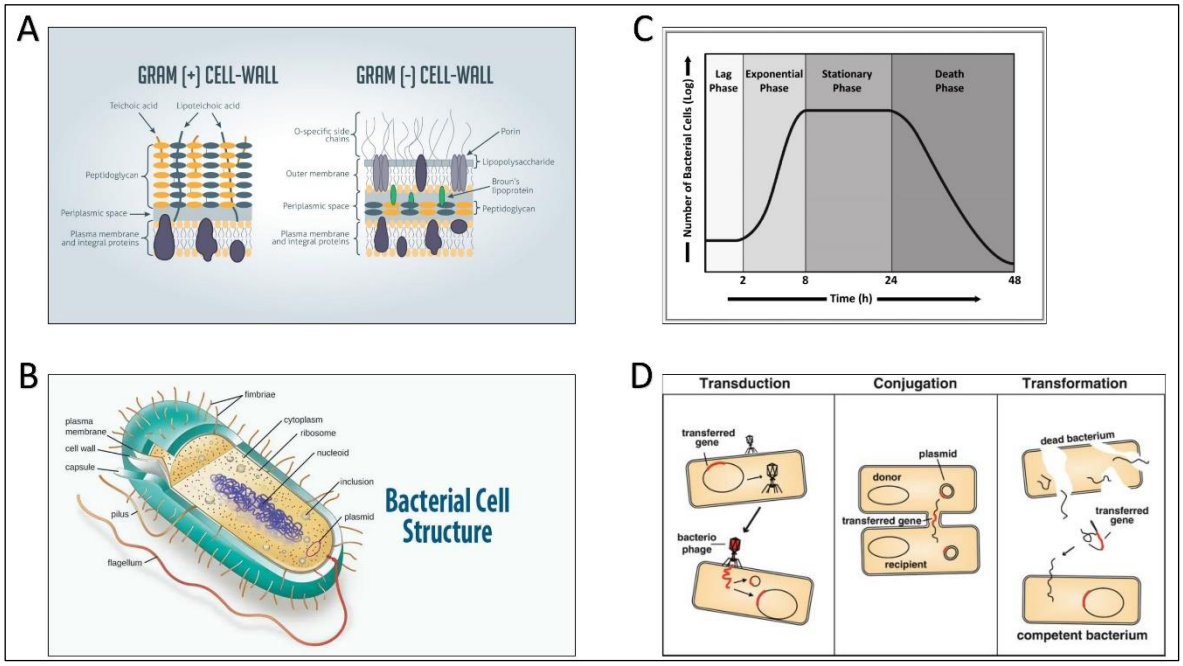

Figure 2.2: (A) Gram+ and Gram- bacterial cell wall. (B) Bacterial growth phases. (C) Bacterial Cell Structure. (D) Horizontal gene transfer processes (Figure was adopted from Wikipedia and NCBI)

Bacterial growth[180] is divided into four phases (*Figure 2.2*); lag phase, exponential phase, stationary phase and death phase. During lag phase, bacteria adapt to the environment's growth conditions (e.g., high-nutrient). At lag phase, bacteria are incapable to divide, instead through the process of maturation, the synthesis of RNA and enzymes takes place. Lag phase duration ranges from one hour to several days. In exponential phase, bacterial cells are dividing regularly by binary fission. At this phase, cells are growing by geometric progression. The cells divide at a constant rate depending upon the composition of the growth medium and the conditions of incubation. Significant reduction of nutrients is the cause of transitioning from the exponential to the stationary phase. In stationary phase, the rate of cell growth and death are in equilibrium. As a result, in stationary phase, the number of bacterial cells remains approximately the same. Finally, the death phase is practically the opposite of exponential phase. When nutrients are depleted or due to other conditions (e.g., nonoptimal temperature), bacterial cells die and their number declines geometrically.

Horizontal gene transfer (HGT) is another important molecular mechanism utilized by many multicellular and unicellular organisms including bacteria (*Figure 2.2*). Transduction, conjugation and transformation are the three processes by which bacterial HGT is achieved.

In bacterial transduction, a specific type of virus termed “bacteriophage” transfers genetic material from one *bacterium* to another. Bacteriophages are viruses that infect bacterial cells and use them as a host. Conjugation is the process by which physically attached/close bacterial cells are exchanging genetic material either by direct cell-to-cell contact or by bridge-like connections. Bacterial transformation refers to the process by which bacteria take up (i.e., absorb) naked DNA from their surrounding environment[181, 182]. Plasmids are usually central players in horizontal gene transfer. HGT is an extremely important mechanism that plays pivotal roles in the spread of antibiotic resistance[183], in the evolution of species and more.
Even though HGT between bacteria is a well-studied and established field of research, additional types of HGT exist too. An extremely interesting field of active research is the *transkingdom* horizontal gene transfer. Known cases of *transkingdom* HGT include: (i) bacteria to plants[184], (ii) viruses to plants[185], (iii) plants to animal[186], (iv) bacteria to fungi[187] and (v) bacteria to animals[188].

Bacteria form either symbiotic or parasitic relationships with plants and animals. In the latter case, bacterial pathogens can lead to many human pathologies. Even though this is established knowledge, it hasn’t always been like that; it took the pioneering work of Girolamo Fracastoro, Marcus von Plenciz, Louis Pasteur and Robert Koch to establish the germ theory of disease[189, 190].

2.3 Human microbiome

2.3.1 Development

The human microbiome is a complex ecosystem comprising bacteria, archaea, fungi, viruses and bacteriophages. It has long been believed that the average human body (~70 kg) consist of ~10^{14} bacterial cells[191] resulting in a bacterial to human cells (B/H) ratio of 10:1. A recently published study[192] revisited this estimate to 3.8×10^{13} bacterial cells leading to an updated B/H ratio of 1:1.

Microbial colonization begins at birth though there are studies that both challenge[193] and support[194] the sterile womb hypothesis (i.e., the amniotic fluid is sterile). The first two factors that shape the microbiota of a neonate are the mode of delivery and duration of pregnancy. In the first case, multiple studies indicate a significant difference in the abundance of specific bacterial taxa but also in the overall bacterial composition between spontaneous vaginally- and Caesarean section-delivered newborns[195, 196]. In the latter case, the abundance of specific phyla residing in the gut microbiota of full-term (FT) infants are significantly reduced compared to preterm (PT) infants[196]. For instance, Cian J. Hill

et al.[196] demonstrated that *Proteobacteria*, a Gram-negative phylum with numerous pathogenic bacteria (e.g., *Salmonella, Escherichia, Yersinia*) is significantly more abundant in PT compared to FT infants. A plethora of other factors influence the development of the gut microbiota in the first days after birth but also the first months and years of a human's life. These factors include (i) breastfeeding, (ii) antibiotic exposure, (iii) environment, and (iv) post-breastfeeding diet. Generally, species from the anaerobic genera *Bacteroides*, *Clostridium*, and *Bifidobacterium spp.* are some of the very first colonizers of the human gut. At the phylum level, *Proteobacteria* and *Actinobacteria* are the first and most abundant phyla that establish their presence in the human body. The first months of a newborn are characterized by low microbial diversity; *Firmicutes* and *Bacteroidetes* are usually the next players in the developing human gut, quickly becoming the dominant phyla while at the same time shaping a more diverse ecosystem. After a year, each infant possesses a unique microbial footprint while at age ~3, their microbiome is almost identical both in terms of diversity and composition with that of an adult[197, 198].

Immune-specific homeostasis of the human body largely depends on the diversity and composition of the microbiome and thus, great effort has been made towards the full characterization of the microbiome of adult humans. These endeavors possess unique challenges due to technical obstacles, high variation among individuals and limitations in the samples under study. Nevertheless, the general consensus is that adults have highly personalized microbiomes with some core elements being common between individuals[199]. The composition of such microbiomes is extremely stable overtime though opportunistic infections and other factors have the potential to perturb the composition of bacteria and other microorganisms that assemble the microbiome of adults. Some factors that can significantly change the adult microbiome include (i) major changes in diet habits, (ii) smoking, (iii) alcohol consumption, (iv) genetics and (v) pathogenicity[200].

Elderly people (age > 65) usually face a sudden reduction in bacterial diversity while the dominant species present in their gut change. Beneficial bacterial species generally decline and at the same time, anaerobic bacteria increase. Additionally, the overall abundance and circulation of short chain fatty acids (SCFAs) is significantly decreased compared to younger adults[201].

The microbial composition and diversity are both unambiguous indicators of human age; machine learning models have been trained and employed, predicting age with high accuracy (±3 years off)[202].

2.3.2 Architecture

Defining the composition of the human microbiome is an extremely difficult task. To date, a plethora of small- and large-scale studies are focused in either health- or disease-specific phenotypes[164]. In healthy individuals, *Firmicutes, Bacteroidetes* and *Actinobacteria* are

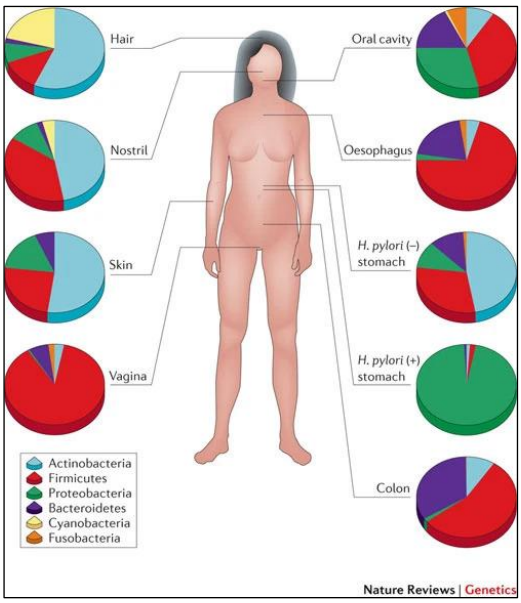

Figure 2.3: Dominant phyla at different anatomical sites (Figure was adopted from Ilseung Cho et al. 2012).

the most abundant phyla in the human colon making up for > 90% of the composition of the gut (*Figure 2.3*)[203]. *Proteobacteria* are also present most of the times. At the genus level, *Alistipes*, *Prevotella*, *Paraprevotella*, *Parabacteroides* and *Odoribacter* are some of the most abundant taxa. At the species level, the diversity of the human gut microbiota differs significantly among individuals harboring 200 (low *alpha* diversity) to more than 1,000 (high *alpha* diversity) unique bacterial species.

Additionally, the *beta* diversity (i.e., difference in terms of bacterial composition between two different habitats) between any two body parts of the same person is extremely high. Notably, microbial composition is more similar within (different persons, same body part) than between body habitats (same person, different body parts)[204]. The oral cavity is the second most complex and rich microbial environment of the human body highly enriched by the species *Streptococcus spp.* At the phylum level, the most dominant bacteria include *Firmicutes*, *Proteobacteria*, *Bacteroidetes* and *Actinobacteria*. Moreover, *Fusobacteria* and *Cyanobacteria* are also abundant in the oral cavity[205].

The healthy human stomach is enriched in acid-resistant bacterial strains assembling a core microbiome enriched for *Prevotella*, *Streptococcus*, *Veillonella*, *Rothia* and *Haemophilus*[206]. This relatively stable ecosystem can change upon long-term *Helicobacter pylori* infection. In such case, *H. pylori* becomes the dominant species lowering the overall diversity of the human stomach while also introducing quantitative changes in the ecosystem[207]. For instance, *H. pylori* infected patients have a relative lack of *Proteobacteria* and *Bacteroidetes*, and a relative abundance of *Streptococcus* and *Prevotella*[208]. Furthermore, the human skin is colonized by millions of beneficial microorganisms that usually act as a barrier to prevent the invasion of pathogens. The composition of microbial communities is primarily dependent on the physiology of the different skin sites (i.e., moist, dry and sebaceous). Sebaceous sites are highly enriched in *Propionibacterium* which is a lipophilic bacterial genus whereas *Staphylococcus* and *Corynebacterium* are the dominant genera of moist skin sites[209, 210]. On the contrary, fungal communities are independent of the skin physiology since they are highly similar in all body sites. Some of the most abundant fungi throughout the human body include *Malassezia spp.*, *Aspergillus spp.*, *Cryptococcus spp.*, *Rhodotorula spp.* and *Epicoccum spp.* Finally, skin viruses are host- rather than body site-dependent[211]. Additionally, many

body sites that long-believed to be sterile have recently being re-evaluated; many studies have been focused in such body sites to identify local and highly functional microbial communities. Some examples are the circulating microorganisms in the blood and the human breast microbiome[168, 212].

2.3.3 Pathophysiology

Over the last decade, the field of microbiome research has experienced an exponential growth. Currently, numerous research projects worldwide are exploring the possibilities of microbiome in health, disease, nutrition and reproduction. Even though the human body comprise multiple local microbiota (e.g., oral, skin, gut), most of these endeavors are focused in the gut microbiome which is the most complex microbial ecosystem of the human body and is also considered critical in terms of health and disease.

Typically, the healthy gut microbiota is characterized by high taxonomic diversity, high gene richness and is also relatively stable over time. A plethora of studies with a focus in the health benefits of the microbiome have shown that the gut microbiota, among others, is closely involved in nutrient extraction, metabolism, and immunity. Energy and nutrient extraction from food (e.g., *Bacteroides* in the large intestine are responsible for sugar harvesting[213]), takes place by utilizing a versatile set of microbial genes, specialized enzymes and molecular pathways that the microbiota adds in the biochemical arsenal of its host. Additionally, crucial molecules for the human health (e.g., vitamins, amino acids and lipids), are synthesized and provided by the gut microbiota[214]. In the immune system axis, the human microbiota contributes in the protection of the host from pathogens (e.g., opportunistic infections) by producing antimicrobial substances that fight bacterial/viral invaders. Finally, the healthy microbiota also provides significant support in the development of intestinal mucosa and the immune system[215].

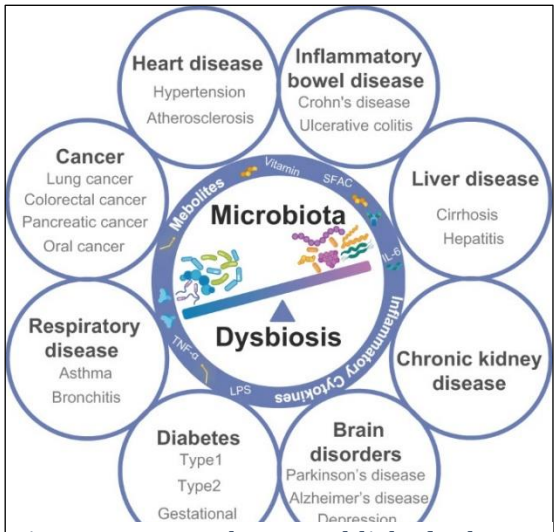

Figure 2.4: The established human microbiome-disease associations (Figure was adopted from Kaijian Hou et al. 2022)

On the contrary, a large body of evidence have established the connection of the human microbiome with the development and progression of numerous diseases such as cancer, metabolic and brain disorders, liver and cardiovascular diseases (*Figure 2.4*). Most of these studies provide quantitative results (i.e. high and/or low abundance of specific bacterial taxa are linked to specific disease phenotypes) suggesting that host-microbiome symbiosis and host homeostasis are extremely crucial to maintain health. Mechanistically, there are many pathways through which

microbes can disrupt a healthy phenotype. For instance, in cancer development and progression, it has been shown that bacteria may influence cell proliferation, perturb immune responses and affect host metabolism. Furthermore, the bacterial-induced disruption of the epithelial barrier is another major source of pathogenesis. The aforementioned and additional processes can promote human pathology and play significant roles not only in pathogenesis but also severity of symptoms and disease progression. Some of the molecular causing agents of bacterial-induced dysbiosis/disease include: (i) Short-chain fatty acids (SFCAs), (ii) endotoxins and (iii) bacterial metabolites.

To date, thousands of "*bacterium*-disease" pairs have been identified. Some well-studied examples comprise *Helicobacter pylori*-Stomach cancer, *Fusobacterium nucleatum*-Colorectal cancer, *Campylobacter concisus*-Pediatric Chron's disease, Adherent-Invasive *Escherichia coli*-Chron's disease and *Aggregatibacter actinomycetemcomitans*-Coronary artery disease.

METHODS AND RESULTS

CHAPTER 3 – Metagenomics analysis suite

The increased resolution of shotgun metagenomics samples comes along with numerous technical challenges, mostly derived from the huge size of the generated FASTQ files and the extended compendium of fully sequenced microbial genomes that are utilized during the alignment step. Numerous methodologies have offered significant advances in many aspects of the shotgun metagenomics pipeline, including indexing, taxonomic assignment and differential abundance analyses.

Kraken[216], an algorithm for taxonomic assignment of microbial sequencing reads, achieved a significant improvement in analysis speed by utilizing the concept of exact alignment of k-mers to the Lowest Common Ancestor (LCA) containing the query k-mer. MetaPhlAn 3[217] is a tool for profiling of microbial communities and utilizes a collection of clade-specific gene markers that uniquely characterize specific phyla, genera and/or microbial species. It assigns sequenced fragments by aligning them against the gene markers database using Bowtie 2[218]. Kaiju[219] translates DNA into protein sequences (amino acid sequences) and searches for maximum exact matches in a pre-computed compendium of proteins from microbial genomes. Finally, Schaeffer *et al.*[220] demonstrated the ability to achieve fast and accurate read assignment transferring technology from RNA-Seq to metagenomics by applying the concept of pseudoalignment implemented in Kallisto[221], and the subsequent use of an expectation maximization (EM) algorithm for the quantification of microbial abundances in shotgun metagenomics samples.

Recent studies[222, 223] have shown that bacterial, archaeal and/or viral footprint can be present in host tissue and bulk RNA and DNA sequencing libraries. These could be the due to sample contamination or local microbiota present within a tissue of biofluid sample. The analysis of such samples could prove invaluable since they could be used for the quantification of bacterial/viral species infiltrating tissue samples. These studies have mostly focused on the viral content of these samples due to the lower complexity of the task. Notably, a recent reanalysis of The Cancer Genome Atlas (TCGA)[224] revealed microbial content in both tissue and blood samples and across different cancer types, highlighting the importance of the human microbiome for oncology-related studies.

AGAMEMNON[225] is an accurate metagenomics and metatranscriptomics quantification analysis suite that addresses open challenges in the field but also provides an end-tp-end methodology. It caters every step of the analysis pipeline, from alignment to statistical

analysis and data visualization by utilizing a series of advancements that enable minimum RAM memory requirements compared to other alignment-based algorithms, enabling the use of larger collections of microbial genomes for hypothesis-free investigations.

3.1 Overview of AGAMEMNON quantification suite

AGAMEMNON (*Figure 3.1*)[225] follows multiple well-known existing metagenomic profiling tools[226-229], in that it divides the task of profiling into two independent steps of read alignment and subsequent abundance estimation. It utilizes the Pufferfish[230] data structure for space and time-efficient representation and indexing of a collection of microbial genomes, coupled with the concept of selective alignment which allows for fast alignment of sequencing reads against a collection of genomes and then feeds a novel quantification algorithm for metagenomic samples, in order to quantify the abundance of the microbial genomes.

The main approach in the abundance estimation step is based on the expectation maximization (EM) algorithm and targets maximizing the likelihood of the observed reads by gradually altering the abundance value associated to different taxa. However, the EM approach is modified and adapted based on specific properties of metagenomic data; mainly, (i) high similarity among the strain sequences, (ii) taxonomic tree and strain relationships through the tree hierarchy, and (iii) high number of unknown species. Additional modules enable concurrent deconvolution and quantification of host and microbial RNAs from the same samples or microbial abundance from host DNA samples, contaminant detection, differential abundance analysis between samples, and visual investigations using a dedicated R-Shiny[231] application. Finally, AGAMEMNON supports single-cell techniques right out of the box, for all analyses modules.

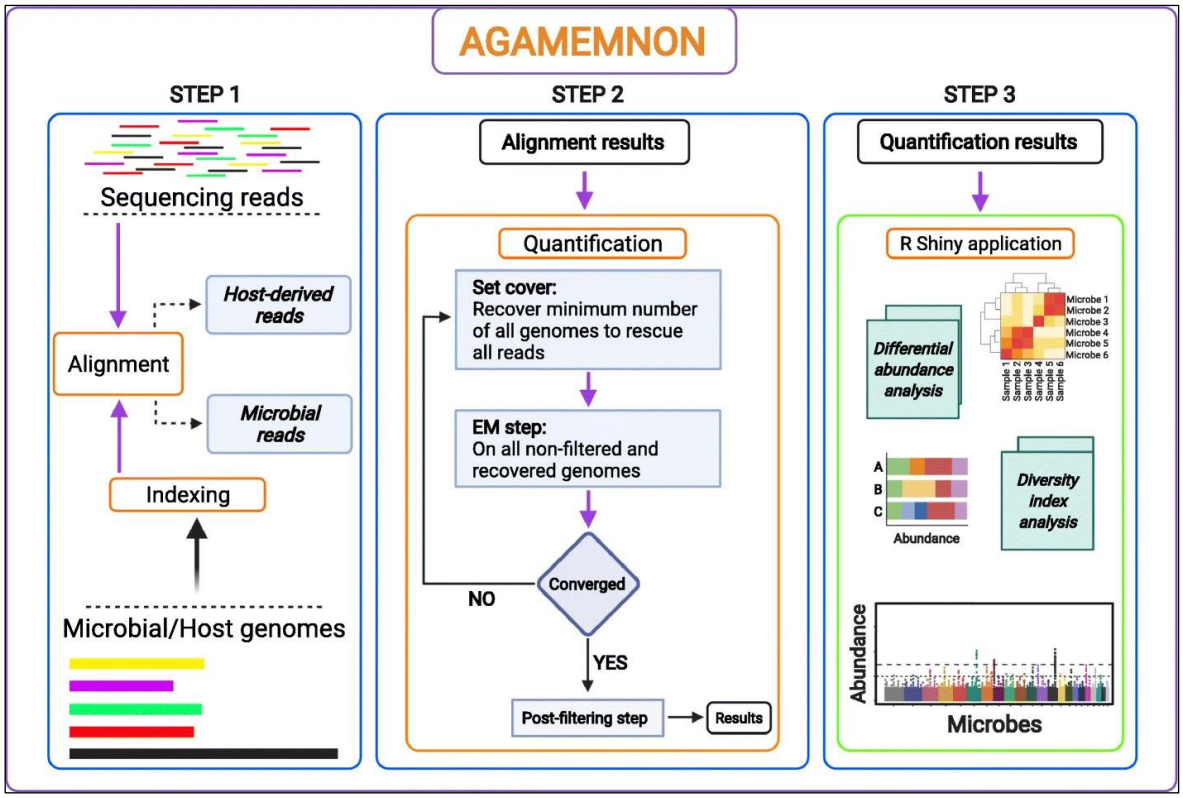

Figure 3.1: Overview of AGAMEMNON's workflow. (Figure was adopted from Skoufos G. et al., 2022)

AGAMEMNON uses an EM algorithm to probabilistically resolve the origin of reads to individual references in the second step of the pipeline (*Figure 3.2*). This step contributes to its enhanced quantification accuracy at the species and strain levels. Unlike methods such as Kraken[216] and Kraken 2[232] that propagate reads that have multiple best assignments to a higher taxonomic rank, AGAMEMNON, via the inference performed in its quantification engine, makes use of other reads and their probabilistic allocations to determine the probability that the ambiguous read arises from the different references to which it aligns well. Similar to other EM algorithms, iterations over these two steps of Expectation and Maximization until the convergence are performed. In each iteration, the model calculates the read probability distribution in the Expectation step and assigns the reads across strains to maximize the probability of observed reads in the Maximization step.

The EM procedure is specifically modified according to fundamental properties and challenges of metagenomic quantification. For instance, in metagenomic indexes, there is often high similarity among the strain sequences belonging to the same species[228], which increases the complexity of disentangling reads at lower levels of the taxonomy. Additionally, reads coming from unknown species or unknown strains can be falsely assigned to entries existing in the index, resulting in false positive non-zero values. As part of the quantification pipeline, these challenges are addressed through iterative, mass-preserving filtering. We look for groups of references that share the same class of reads and

are fully ambiguous without any preference towards a reference over the others to detangle reads in the Maximization step. We call such a group of references, a "multi-mapping island." We reduce the problem of multi-mapping islands to a "set-cover" problem and solve it by adopting an existing approximate set-cover solution. Essentially, we select the minimum number of strains that explain all multi-mapping reads distributed across the strains of each multi-mapping island. The remaining strains in each island are removed prior to the next EM step, significantly improving the accuracy of the proposed quantification model. Through this approach, we tackle the problem by sparsifying the solution (i.e., the set of species that may be assigned a non-zero abundance) in a manner that still retains all mapped reads. The "set-cover" step is called after every k iterations of EM until there are no multi-mapping islands left. At this point, EM continues until termination, which happens either if (a) it reaches the maximum number of iterations (default = 1000 iterations) or (b) the genomes abundance change between the two iterations is adequately small. The quantification procedure is completed by the final step of removing genomes with abundance values lower than a cutoff threshold (default = 2).

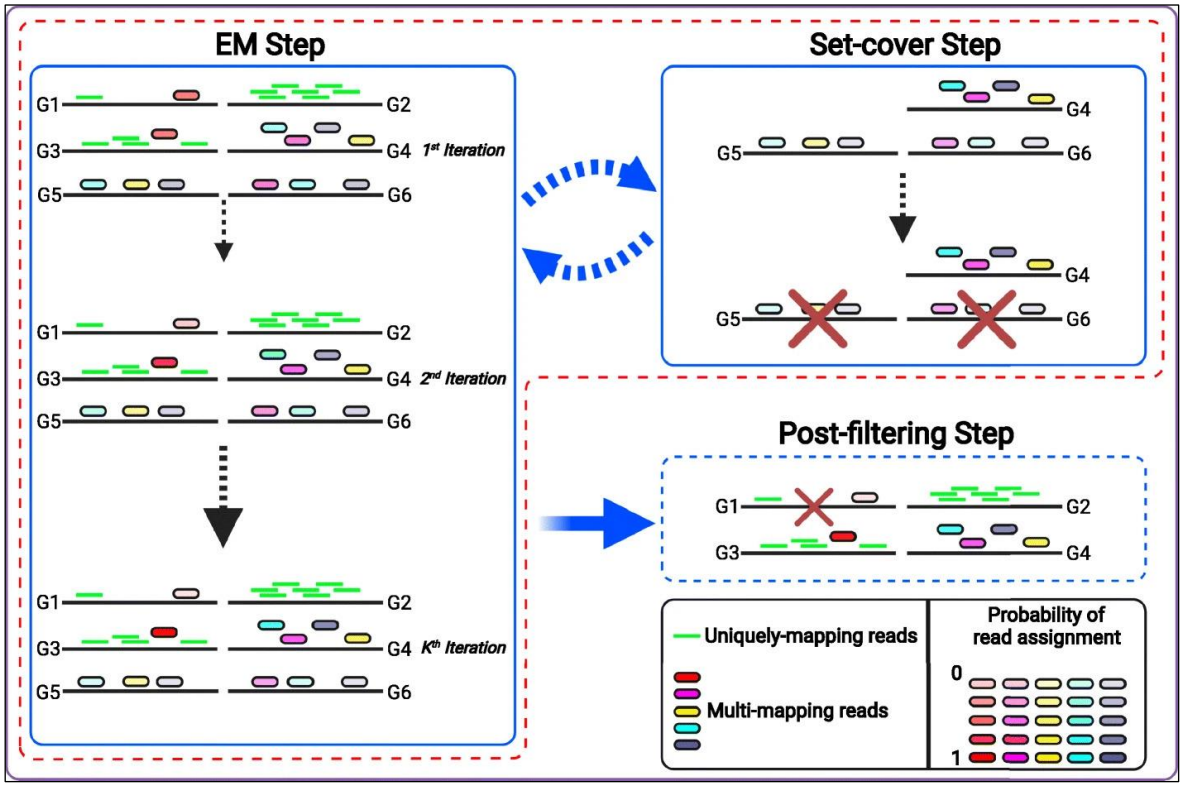

Figure 3.2: Overview of AGAMEMNON's quantification algorithm. (Figure was adopted from Skoufos G. et al., 2022)

3.2 Benchmarking AGAMEMNON against state-of-the-art methodologies

We benchmarked AGAMEMNON using simulated[233], synthetic[234], and real datasets against Kaiju[219], Kraken 2[232], Bracken[228], MetaPhlAn 3[217], and meta-Kallisto[220].

To assess the accuracy of the methods, we used the Mean Squared Log Error (MSLE) and the total number of reported false positive (FP) taxa in different read thresholds. Briefly, MSLE is defined by the following formula:

$$L(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N}(\log(yi\ +1) - \log(\hat{y}i\ +1))^2$$

where *y* and *yˆ* are numeric vectors comprising the ground truth and estimated read counts respectively. *N* is the total number of reported microbes by each method.

The Illumina 400 dataset[233] that was used in our benchmarks incorporates 400 different microbial genomes. A less complex version of the dataset (Illumina 100) has been commonly used as a test set in the field. Furthermore, the synthetic dataset[234] used is a product of a real shotgun metagenomics sequencing experiment of a predefined mock microbial community. The mock community comprises 12 bacterial strains spreading over 2 phyla.

The choice of a reference compendium for index creation is an important aspect, since it can affect the complexity of the task at hand. To simulate real-world scenarios but also to enable us to assess the effects of reference choice on algorithm outcomes, we incorporated two different microbial references to our benchmarks. The first reference (REF-1) comprises all complete and latest bacterial and archaeal genomes from NCBI RefSeq database (n = 1,840). In this reference, ~ 36% of the genomes present in the Illumina 400 dataset are missing from the reference and that allows us to mimic the common scenario of unknown microbial sequences in metagenomics samples. Such cases can lead to false positives, by assigning reads of unindexed microbes to the closest match in the index. The second reference (REF-2) comprises 44,694 sequences (> 8,500 genomes). Importantly, we removed 63% of all genomes (i.e., 252 entire genomes) from REF-2 that are part of the Illumina 400 dataset. After the aforementioned removal, reference 2 contains only 148 out of the 400 genomes that are part of the Illumina 400 dataset. Strain level results for that particular scenario are calculated by considering (a) how accurate the quantification of the abundances is for the 148 strains that are both part of the reference and present in the dataset and (b) how many reads are mis-classified into different strains for the 252 strains that are missing from the reference but are part of the dataset.

Moreover, in the relevant test for the synthetic community, all (100%) of the strains and species present in the synthetic dataset are not included in reference 3. In that scenario, where all the strains present in the dataset are missing from the reference, we believe it is still

informative to calculate metrics of accuracy. In that case, the number of the false positive strains represents the number of falsely reported taxa (in terms of presence/absence) and MSLE shows the total error in terms of mis-assigned counts (i.e., the degree of overestimated abundances for each of the falsely reported taxa). For example, a method that does not assign reads of missing strains to those present in the index will outperform a method assigning falsely the majority of those reads (while keeping all other assignments equal).

Even though AGAMEMNON is a complete analysis suite and not just a quantification or alignment method, its engine shows robust top-of-the-line performance, across all test sets. Specifically, AGAMEMNON exhibited top performance accuracy in most of the tests. (*Figure 3.3*, *Figure 3.4*).

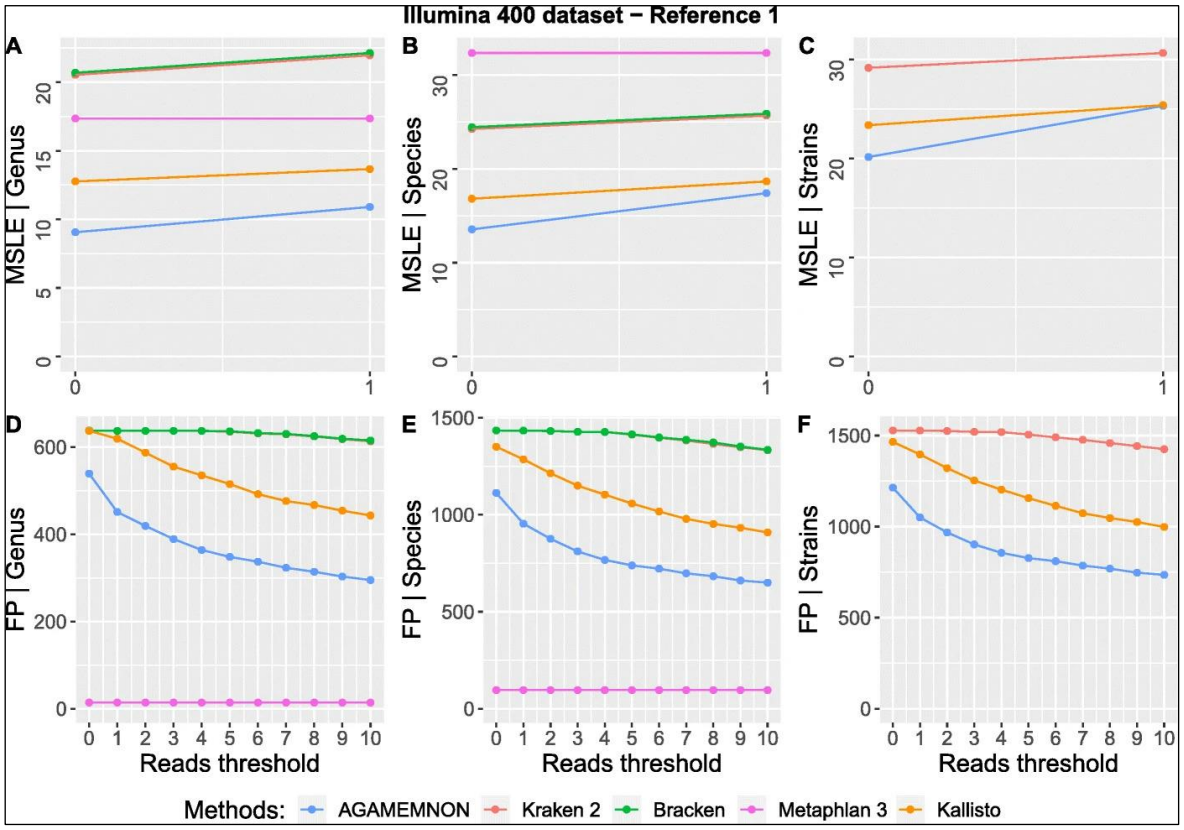

Figure 3.3: A–F The mean squared log error (MSLE) and the number of false positive taxa (FP) between true and estimated read counts at the levels of genus, species, and strain using the Illumina 400 dataset and reference 1. (Figure was adopted from Skoufos G. et al., 2022)

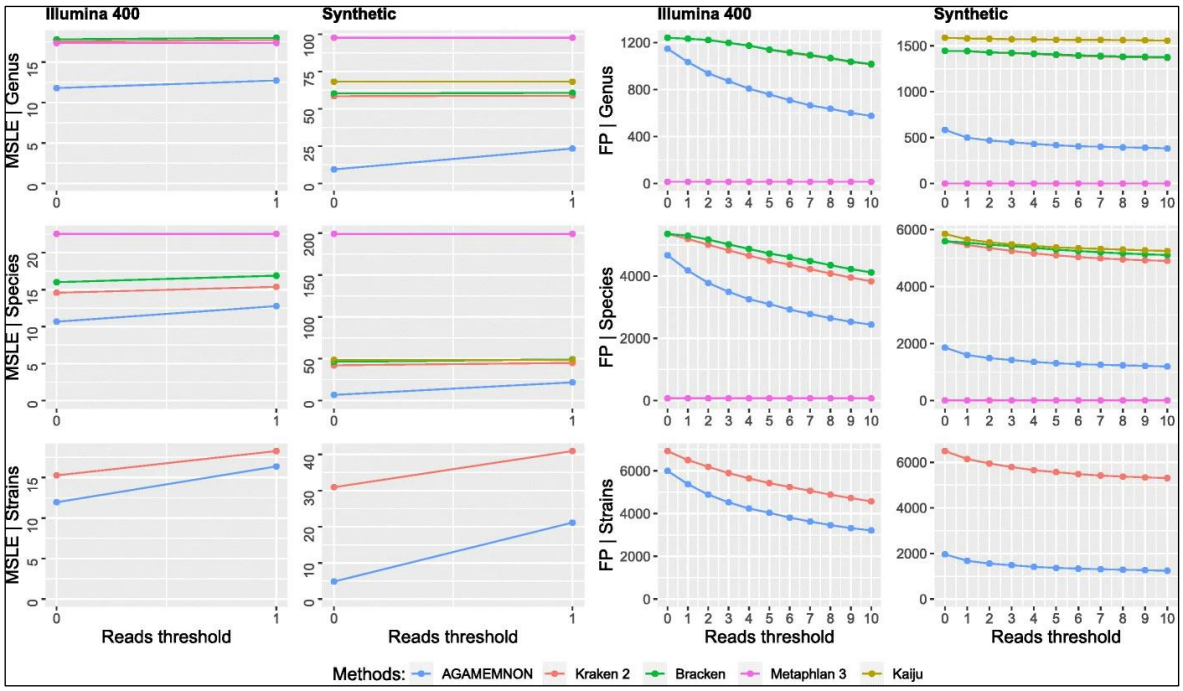

Figure 3.4: The mean squared log error (MSLE) and the number of false positive taxa (FP) between true and estimated read counts at the levels of genus, species, and strain using reference 3. (Figure was adopted from Skoufos G. et al., 2022)

In *Figure 3.3*, we present the results using the Illumina 400 dataset and REF-1. As shown, AGAMEMNON displayed better performance in terms of mean squared log error (MSLE) in both genus and species resolution (panels A, B) while at the strain level, AGAMEMNON and Kallisto had the lowest MSLE (panel C). MetaPhlAn 3 had the smallest number of false positives (FP) in all tested taxonomic ranks with AGAMEMNON following (panels D, E, F). Metaphlan's small number of false positives despite the low observed accuracy is expected since it uses a predefined clade-specific marker database with significantly reduced query space compared to the reference used by Kraken 2, Bracken, Kallisto, and AGAMEMNON. Kaiju was not included in the analyses presented in *Figure 3.3*, since it only supports analyses using the complete RefSeq as reference (presented in *Figure 3.4*) and not custom annotations.

Next, we compared the methods using both the Illumina 400 and the synthetic datasets against REF-2 index. In terms of MSLE, AGAMEMNON performed better in all tested cases and taxonomic ranks (*Figure 3.4*, left panel). MetaPhlAn 3 had the smallest number of false positives (FP) with AGAMEMNON following (*Figure 3.4*, right panel). Kaiju is included only in the synthetic tests, since its index (which cannot be altered) already includes the omitted species and strains. We were also not able to run meta-Kallisto using REF-2, since the indexing step of REF-2 required more RAM than what was available in our largest server instance (512 GB).

To assess the concordance of the methods when analyzing shotgun metagenomics sequencing experiments, we also quantified the microbial abundances of three human stool

samples originating from the Human Microbiome Project[205] using REF-2. In this comparison, we included Kaiju, Bracken, Kraken 2, MetaPhlAn 3, and AGAMEMNON (*Figure 3.5*). As shown, all methods (excepting MetaPhlAn 3) exhibit a positive Spearman's rho > 0.5 in almost all samples and both taxonomic ranks. As expected, the highest correlation is between Kraken 2 and Bracken, since Bracken utilizes Kraken 2 output as the foundation of its abundance estimation calls. AGAMEMNON has a strong positive correlation with both Bracken and Kraken 2 at both the genus (> 0.7) and species (> 0.5) levels. These results demonstrate that most of the methods have a relative agreement in three experimentally derived human datasets.

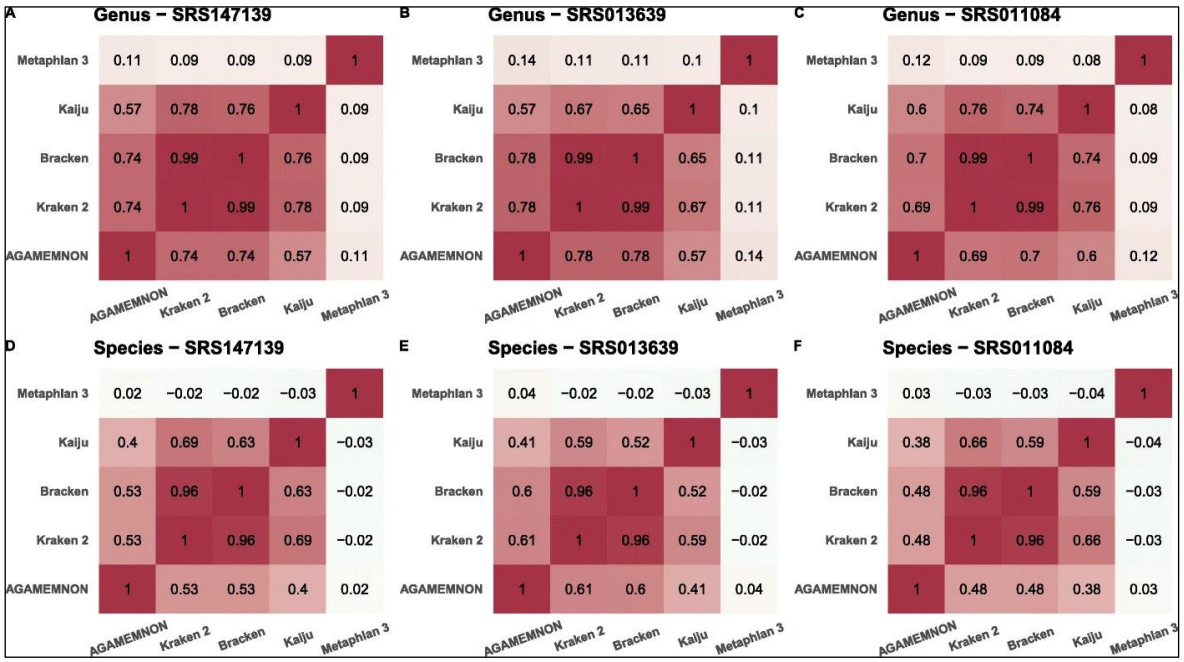

Figure 3.5: The pairwise Spearman correlation of each method in three human fecal samples at the levels of genus and species. (Figure was adopted from Skoufos G. et al., 2022)

In terms of execution speed and memory footprint, MetaPhlAn 3 and Kraken 2 proved to be the most efficient algorithms. It is worth noting though that AGAMEMNON is the only method (tested in this study) that performs actual alignments against a full reference. This information (i.e., SAM files) can be stored locally and used downstream to the quantification results. The incorporation of the pufferfish data structure in the quantification engine enables AGAMEMNON to require ~ 6.5-fold less RAM than Kallisto, a pseudoalignment-based approach. The differences are also evident during indexing, an important bottleneck for this class of implementations, since medium to large-size microbial compendia could require more than 0.5 TB RAM for indexing, which is not always available.

3.3 Microbial quantification in host RNA/DNA-Seq samples

AGAMEMNON also supports the quantification of microbial fragments in host tissue and/or biofluid RNA/DNA samples. To this end, AGAMEMNON separates the host reads by utilizing HISAT2[235, 236] and subsequently employs aligns the reads failing to align to the host genome and/or transcriptome against a user-defined collection of microbial genomes. Finally, it uses its quantification engine to derive the microbial abundances of the microbial in the sample under investigation. We evaluated AGAMEMNON's host sample analysis capabilities against GATK PathSeq[237] and the HUMAnN3[217] pipeline in host tissue analysis scenarios. Two different simulated datasets were created using ART[238], which included a high (Dataset ONE, 7.53%) and a low (Dataset TWO, 3.77%) microbial read content in human. In both of the datasets, AGAMEMNON outperformed both HUMAnN3 and PathSeq (*Figure 3.6*). Importantly, the percentage of mis-classified microbial reads to the host genome made by AGAMEMNON has no impact on the overall accuracy (< 0.001%).

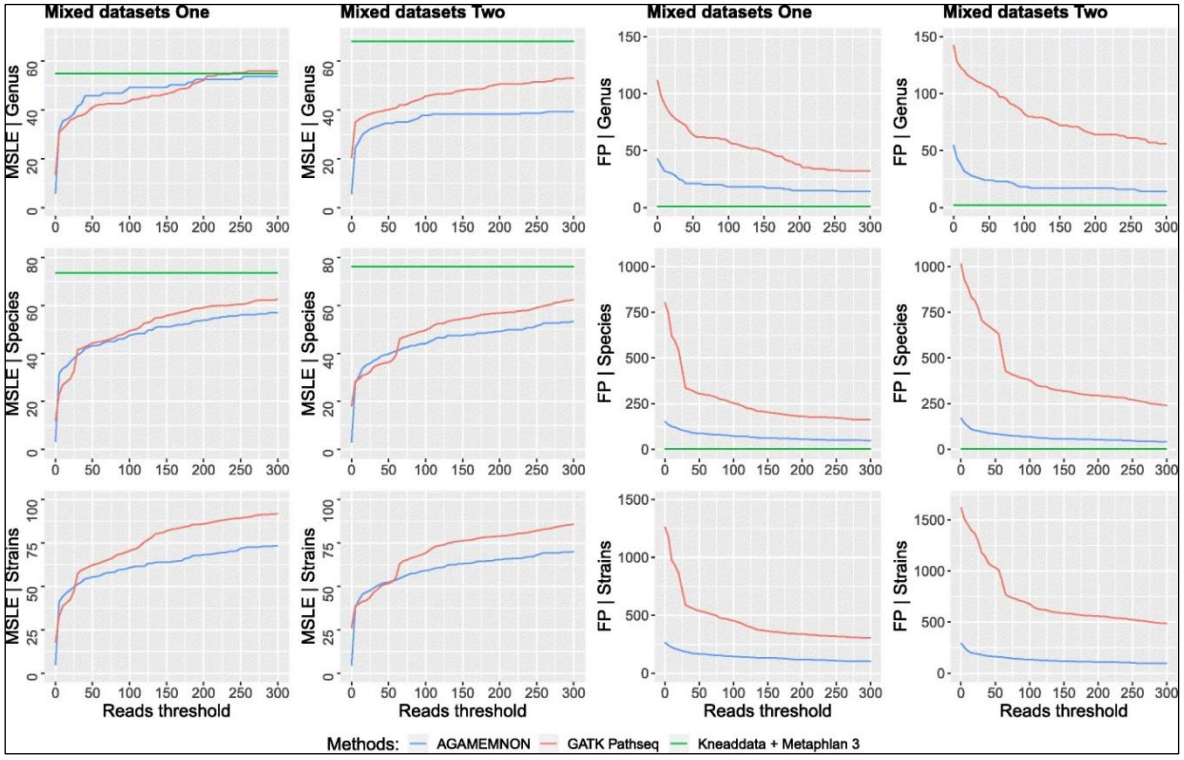

Figure 3.6: The mean squared log error (MSLE) and the number of false positive taxa (FP) between true and estimated read counts at the levels of genus, species, and strain using mixed datasets one and two and the human-subset reference. (Figure was adopted from Skoufos G. et al., 2022)

3.4 Downstream analysis of quantification results with R-Shiny

AGAMEMNON offers a powerful R-Shiny analysis suite (*Figure 3.7*), where users can explore and visualize quantification results as well as perform differential abundance or diversity index analyses. This module supports simple and sophisticated exploratory visualizations including Manhattan plots, diversity indices (Bray-Curtis, Euclidean, Canberra) heat maps, boxplots and dimensionality reduction methods (principal component analysis (PCA), multidimensional scaling (MDS)) and others.

Figure 3.7: Screenshots of AGAMEMNON's Shiny application. (Figure was adopted from Skoufos G. et al., 2022)

CHAPTER 4 – An expanded microbe-disease association compendium

Over the past decade, numerous research projects and the development of novel Next-Generation Sequencing and *in-silico* techniques have expanded our view of the implication of microorganisms to human pathologies[164, 239, 240]. A plethora of studies identify associative relationships between the bacterial abundance and the existence, progress and outcomes of several diseases starting from disorders of the gastrointestinal tract[241], but surprisingly extending to other pathologies, including cancer[242, 243], neurodegenerative disorders[170] and cardiovascular diseases[244]. A number of studies blaze a trail in microbiota research, going beyond the qualitative model and focusing on the causative effects of the microbiome. For instance, *Fusobacterium nucleatum* has been found to selectively stimulate the growth of colorectal tumor cells in a colorectal cancer progression model comprising cell lines from human colonic adenoma[245]. The stimulation is achieved mainly using the Adhesin A protein expressed by FadA gene. Contrarily, microbiome-based diagnostic and therapeutic interventions for treatment of gastrointestinal and neurodegenerative diseases and other types of pathologies are being actively pursued[246].

The systematic cataloging of the rapidly expanding volume of microbe-disease associations is indispensable to basic and applied microbiome research. To this end, we developed Peryton, a novel database comprising experimentally supported microbe-disease associations.

4.1 Peryton's content and statistics

Peryton[247] constitutes a novel resource of experimentally supported microbe-disease associations, currently hosting more than 7,900 entries linking 43 diseases and 1,396 microorganisms. Peryton's content is exclusively sustained by manual curation of biomedical articles. Importantly, diseases and microorganisms are provided in a systematic, standardized manner using reference resources (e.g., NCBI Taxonomy and MeSH terms) to create database dictionaries. Information about the experimental design and techniques, the study cohort, microorganisms and diseases are annotated and catered to users.

The manual curation of ~320 publications yielded more than 7,900 microbe-disease associations. Diseases comprise 10 gastrointestinal disorders, 7 cardiovascular diseases, 23 cancer types and 3 neurodegenerative disorders. Peryton's entries span over all known taxonomic ranks, with the genus-related associations having the highest frequency (3,680, 46%). Importantly, 21.54% of the associations provide information at species level or below (i.e., strain and sub-strain level). For each entry, information including bacterial abundance, the groups under study, experimental design, study cohorts, sample type, the applied high- or low-throughput techniques, Next-Generation Sequencing (NGS) sample accession

numbers and article metadata are annotated and catered to users. Interestingly, in more than 50% of the associations, the sample size (i.e. the number of individuals participated in the study) is >50. Finally, diseases and microorganisms are provided in a systematic, standardized manner by using the vocabularies provided by MeSH and the NCBI Taxonomy database[248], respectively.

4.2 Interface, modules and implementation of Peryton

We designed a user-friendly interface (*Figure 4.1*) that provides a number of functionalities to enhance user experience and enable ingenious use of Peryton. One or more microorganisms and/or diseases can be queried at the same time. Advanced filtering options to refine search results can be applied for experimental methodology, disease type, sample type, taxonomic rank, sample size etc. Direct text-based filtering of results enables refinement of returned information and the conducting of tailored queries suitable to different research questions. Peryton is interconnected with NCBI Taxonomy database (22), PubMed database and MeSH terms. Additionally, we compiled a list of common contaminants following Eisenhofer et al. (23) and integrated it in the database. Therefore, users can see whether or not a microorganism participating in an association has been deemed a common contaminant.

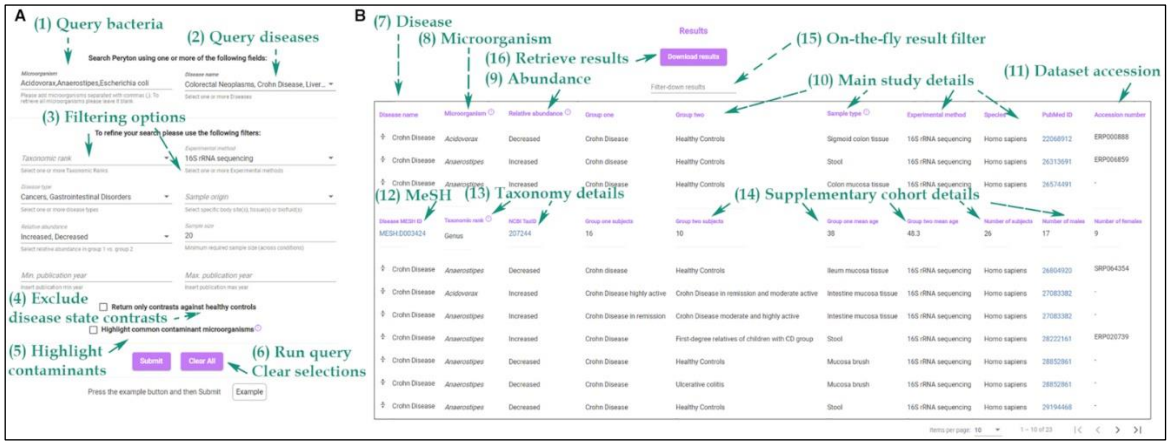

Figure 4.1: Peryton's main user interface. (Figure was adopted from Skoufos G. et al., 2021)

Peryton also provides interactive visualizations (*Figure 4.2*) to effectively capture different aspects of its content. Via Network graphs, Chord diagrams and Hierarchy diagrams, users can browse into the available content and perform observations about microbe-disease relationships using information from the latest relevant literature.

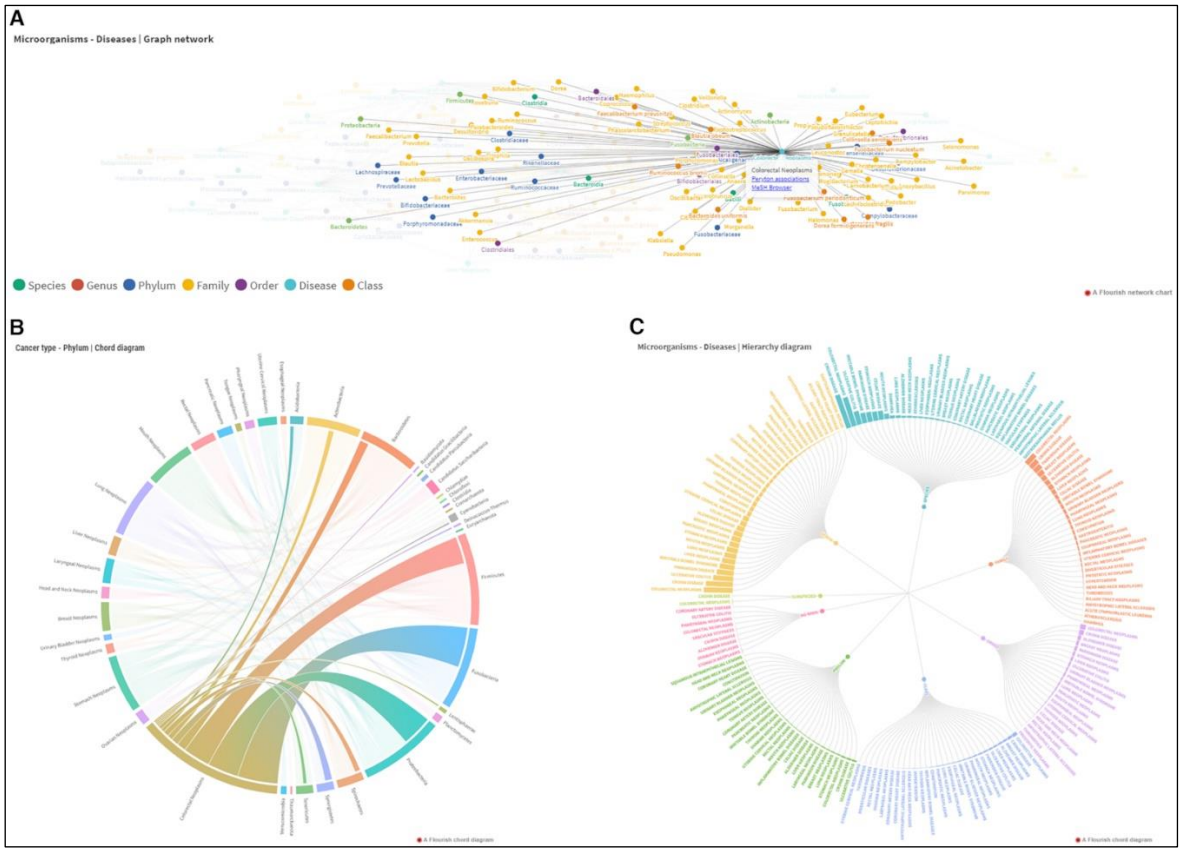

Figure 4.2: Peryton visualizations. (Figure was adopted from Skoufos G. et al., 2021)

CHAPTER 5 – An extended catalogue of bacterial small RNA-RNA interactions

Small non-coding RNAs (sRNAs) have been detected in most known eukaryotic and prokaryotic organisms and have been found to be involved in a variety of biological functions, prominently including the post-transcriptional regulation of gene expression. In the bacterial kingdom, sRNAs typically range between 50 and 500 nucleotides (nt)[249]. Their main function is to post-transcriptionally regulate gene expression in a negative and/or positive manner. Usually, gene regulation by sRNAs occurs via imperfect base-pairing with their RNA targets, in response to stress, metabolism and environmental stimuli[250, 251]. Bacterial sRNAs play major roles during bacterial infections[252, 253], in the control of antibiotic resistance genes and virulence factors[254, 255], during intra- and inter-species communication[67] and in other crucial molecular and cellular processes, deeming them extremely important for (i) basic research in microbiology, (ii) applied biomedical and clinical research and (iii) the development of novel sRNA-based therapeutic interventions and biomarkers.

Over the past decade, the field of bacterial sRNAs has experienced tremendous growth. A flourishing body of evidence emerges and our understanding of bacterial sRNA implications grows deeper. Biotechnological breakthroughs mainly in the field of Next-Generation Sequencing (NGS) enabled the identification of sRNA-RNA interactions in a high-throughput manner[256-258]. To date, many experimental methods, featuring diverse protocols yet ultimately the same scope, have been developed. In RIL-Seq (RNA interaction by ligation and sequencing), CLASH (crosslinking, ligation and sequencing of hybrids) and GRIL-Seq (global sRNA target identification by ligation and sequencing), RNAs are ligated prior to sequencing resulting in the generation of chimeric fragments and reads. While RIL-Seq and CLASH are RBP-dependent (i.e., they are focused on immunoprecipitation of a specific RNA-binding protein) and sRNA-independent (i.e., many sRNAs and many RNA targets are captured in one experiment), GRIL-Seq is RBP-independent and sRNA-dependent (i.e., one sRNA and many RNA targets are captured in one experiment, irrespective of the RBPs involved). Furthermore, CLIP-Seq (crosslinking and immunoprecipitation followed by sequencing), a technique that has been widely utilized for the study of eukaryotic RBP-bound RNAs, has been successfully applied in the bacterial RNA space as well[259].

Microbiology is currently lacking a carefully curated collection of the rapidly expanding universe of bacterial sRNA-RNA interactions. We developed Agnodice (*Figure 5.1*), the first version of an effort to systematically catalogue and annotate experimentally supported bacterial sRNA-RNA interactions that, for the first time, incorporates thousands of bacterial sRNA-RNA interactions, supports advanced querying/filtering capacity and exploratory

visualizations in a user-friendly manner, and hosts interactions derived from a diverse set of experimental methodologies including state-of-the-art NGS interactome identification techniques. The arsenal of experimental techniques that are currently part of the database is divided in two broad categories; low- and high-throughput methods. These two categories are further dissected into direct (assessing RNA-RNA binding sites) and indirect (lacking information on RNA-RNA binding sites) techniques. Interactions identified via low-yield methods are collected exclusively by manual curation of the available literature, while interactions produced by high-throughput experiments are either collected by manual curation (in the case of CLASH, ligation of interacting RNA followed by high-throughput sequencing (LIGR-Seq), MS2-affinity purification coupled with RNA sequencing (MAPS) and GRIL-Seq datasets) or by de novo analysis of raw datasets (from RIL-Seq and CLIP-Seq methods).

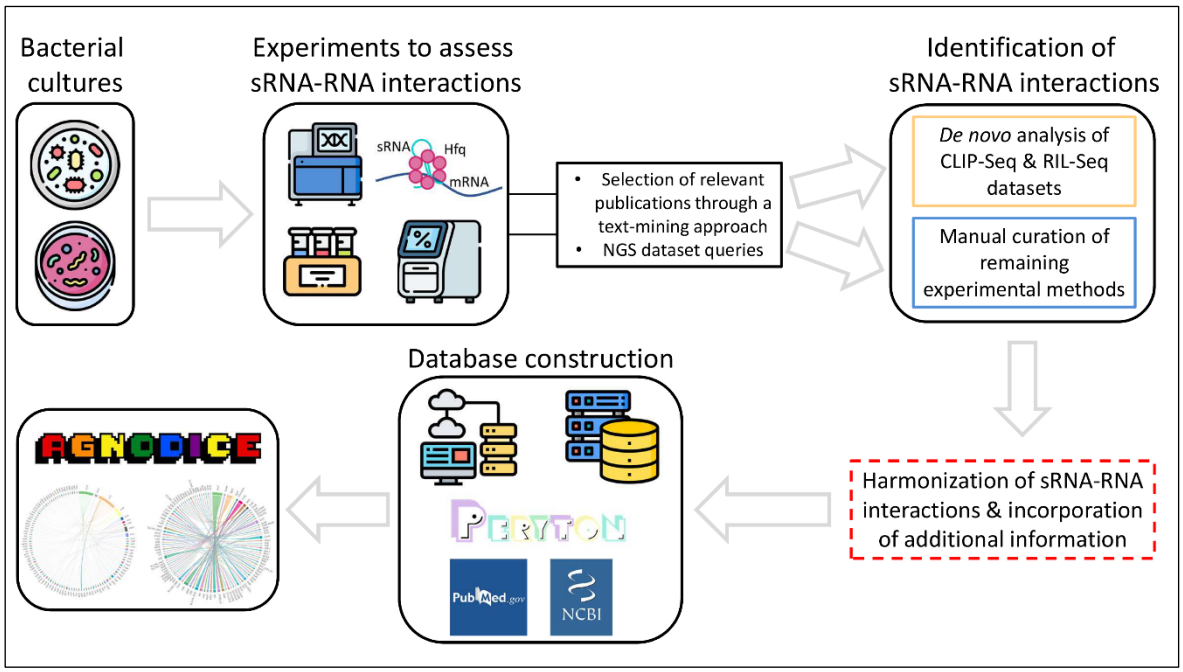

Figure 5.1: Schematic representation of the development and information flow, from the researcher bench to the creation of Agnodice resource. sRNA interactome data, generated by low- or high-yield experimental methods, are contained in hundreds of articles, supplementary materials, or they exist as publicly available raw sequencing libraries. After querying for candidate sources, meticulous curation and analysis procedures are applied. sRNA-RNA interactions are detected either via from scratch analysis of raw NGS data, or by means of manual curation. The resulting set of entries is harmonized, accompanied with rich experiment- or study-specific metainformation and broken down into an efficient database schema. Inter-connection with external resources, including Peryton database of bacterial-disease associations, PubMed and NCBI Taxonomy, is performed. The resulting content is provided in the form of an open, user-friendly online database, supporting numerous querying, filtering, visualization and download functionalities (Figure was created for the purpose of this thesis)

5.1 Agnodice's content and statistics

De novo analyses of high-throughput data and manual curation of more than 100 studies yielded ~22,000 bacterial sRNA-RNA interactions. Agnodice database provides a total of 12,230 unique sRNA-RNA interacting pairs between 390 sRNAs and ~6,630 coding/non-coding RNAs. Out of the total entries, more than 1,000 are derived from high-confidence methodologies directly assessing RNA-RNA binding sites, such as CLIP-Seq. In addition, Agnodice incorporates ~4,550 coding RNAs annotated with 4,100 unique RNA products (i.e., proteins), and ~130 non-coding RNAs. In total, Agnodice features interactions derived from 45 experimental methods (36 low-throughput and 9 high-throughput). Moreover, the database comprises interactions dependent on three different RBPs, the major regulator Hfq, CsrA and ProQ, as well as interactions for which no RBP-related evidence was obtained. Finally, for every annotated interaction, the database integrates additional information including article metadata, microorganism names and TaxIDs (exclusively derived from the NCBI Taxonomy database), bacterial lineages (i.e., genus, species and strain names), information on the specifics of the experimental methods, MFE calculations and more. Basic statistics of Agnodice are presented in *Figure 5.2*.

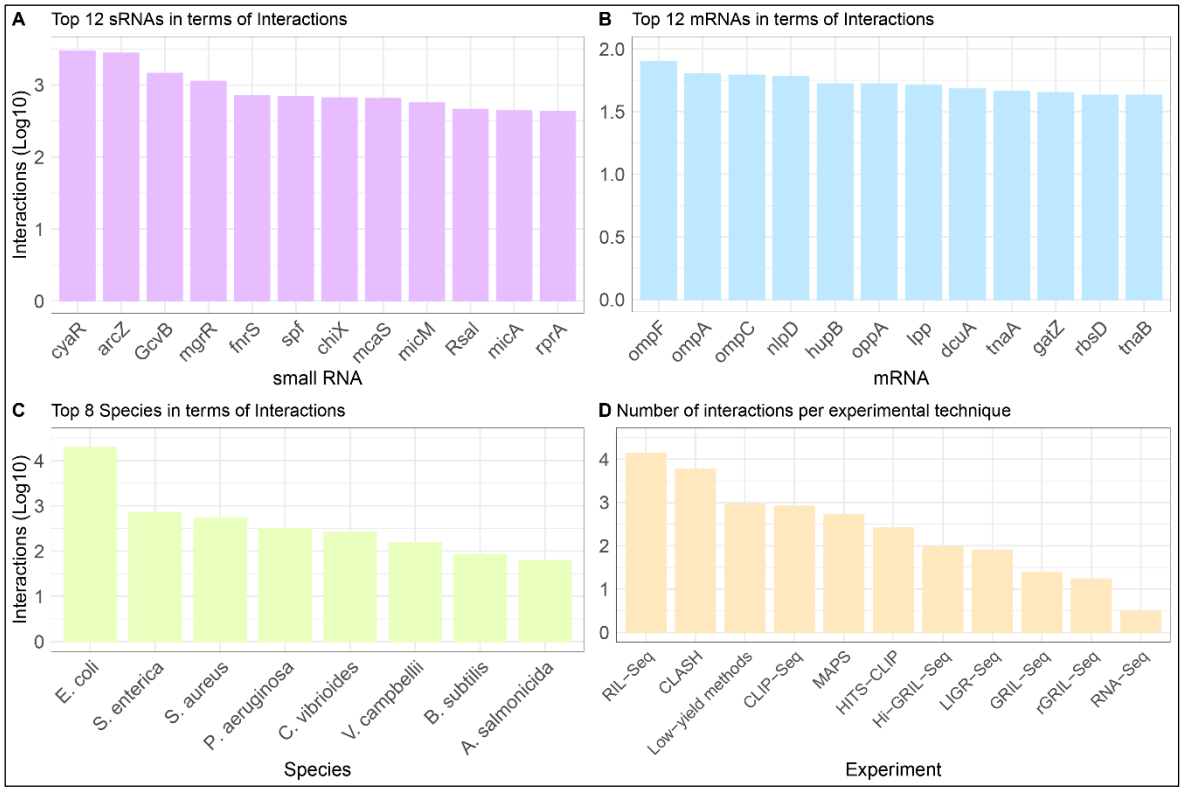

Figure 5.2: (A) Top 12 most frequent sRNA regulators and (B) top 12 most frequent regulated genes catered in Agnodice. (C) Top 8 species in terms of total interactions in the database. (D) Number of total interactions per experimental method (interactions by low-yield methods summed together). Interaction sums are transformed in log10 space (Figure was created for the purpose of this thesis)

5.2 Agnodice's data collection and curation process

In order to collect as many interactomic NGS datasets as possible, PubMed was utilized and queried using a series of keywords (e.g., "RIL-Seq", "CLIP-Seq", "CLASH", "MAPS" and "bacteria" and/or "sRNAs" and/or "interactions"). This resulted in a total of 23 datasets containing the following NGS methodologies: CLIP-Seq, HITS-CLIP, LIGR-Seq, RIL-Seq, MAP-Seq, GRIL-Seq, rGRIL-Seq and CLASH. Furthermore, in order to retrieve studies with experimentally confirmed interactions, we developed a text mining pipeline with full-text capacity dedicated to indexing sRNA-RNA interactions. For pre-processing, we used a set of approximately 200 related scientific publications. Then, we evaluated every related full-text article, regarding the existence of sRNAs and targets, as well as by the importance of individual words on it. Sentences possibly containing the desired associations were retained for manual curation. Curators performed meticulous manual curation of studies potentially containing interactions verified by low-throughput methods, populating a table with 33 pre-defined fields. The final collection was independently cross-checked and post-processed to ensure the uniformity and quality of its content. The criteria required to record an interaction entry were (i) that the interaction should be supported by at least one experimental method and not be solely based on evidence from target prediction algorithms (ii) detailed information on the experimental design, microorganism and interaction components should be provided by the authors and (iii) interactions which were assessed using statistical methods (mainly high-throughput studies) should be deemed statistically significant (i.e., $p < 0.05$). In total, ~1,000 interactions, associated with more than 70 different bacterial strains, were derived from the manual curation process.

5.3 Interface, modules and implementation of Agnodice

Agnodice is built under a mindset of a user-friendly interface that provides researchers with a variety of functionalities and allows them to reach out to an all-in-one resource, perform hypotheses, come up with potentially interacting RNA candidates and unravel complex biological questions. The database supports queries using one or more sRNA(s), gene(s) and/or microorganism names from any of genus, species and strain taxonomic ranks, as well as the application of smart filtering options, including "regulation type" (i.e., activation, repression or unknown), "experimental method name", "publication years", "RNA-RNA interaction energy" etc. Furthermore, filtering options incorporate check boxes for including only interactions derived using NGS-based techniques and/or techniques directly assessing RNA-RNA binding events. Special effort has been placed to offer direct interconnections with a number of useful resources including NCBI Taxonomy[248] reference organism indices, PubMed database to access relevant publications and Peryton, the database of experimentally supported microbe-disease associations. Comprehensive details on the query options and the format of the returned results are available in *Figure 5.3*.

A supplemental free-browse mode has been developed in which users can navigate throughout the entire collection of interactions without the need for a query and/or filtering. Through a dedicated Visualizations page, users can utilize Chord diagrams to focus on the strongest sRNA-RNA interactions (i.e., interactions reproduced independently by different studies) of *E. coli*, *S. enterica* and *P. aeruginosa* species. The DB Statistics page presents the top 12 entries among key components of Agnodice (namely, microorganisms, sRNAs, genes and experimental methods). Unrestricted download options with additional complementary metadata (e.g., publication information, bacterial lineage information etc.) permits the storage of Agnodice results locally, allowing users to utilize them in any meta-analysis scenario. A detailed and informative Help page is available to ensure that users perform tasks across our resource in an effortless way. Finally, interested researchers can submit their own interactions (which will be manually inspected by curators and, until then, remain in a provisional state) by filling the form provided through the web interface. In this manner, the scientific community may contribute to the future establishment of a centralized information hub of bacterial sRNA-RNA interactions which will facilitate the kick-start of new experiments.

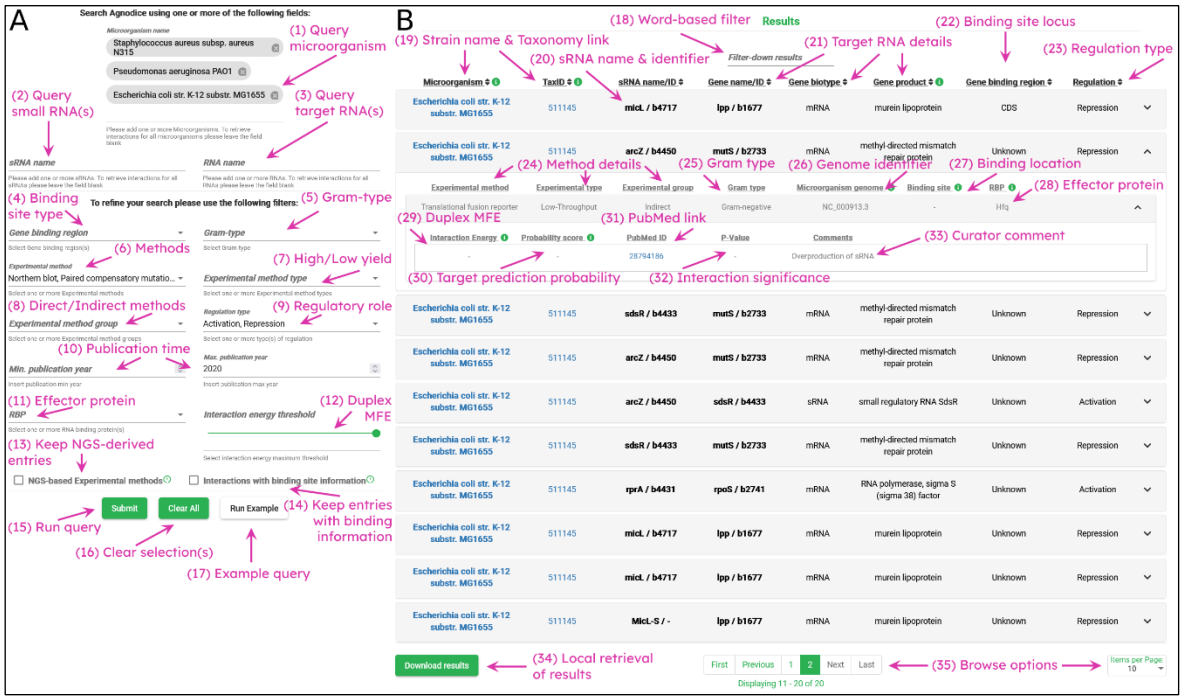

Figure 5.3: Main query/result interface offered in Agnodice.

Finally, Agnodice was built using the MVC architecture as a relational database and is being hosted on Apache HTTP server 2.4. The back-end consists of a PostgreSQL server 11.8 (https://www.postgresql.org/) where Agnodice's data is stored in multiple tables featuring relational connections for optimal storing and querying. The PHP framework Laravel 8

(https://laravel.com/) (PHP 7.2) handles the back-end logic including the connection to the PostgreSQL server for the storing and retrieval of the data. The front-end is designed as a one-page website using Angular 14 (https://angular.io/) and the Angular Material UI library (https://material.angular.io/). Finally, the database statistics are presented using the Chart JS (https://www.chartjs.org/) library, while Flourish (https://flourish.studio/) is utilized for the more complex visualizations provided.

5.4 Comparison of Agnodice with existing resources

Databases addressing similar topics include RegulonDB[260], sRNAMap[261], sRNAdb[262], BSRD[263] and sRNATarBase[264]. RegulonDB, hosting ~230 sRNA-mRNA interactions, is a reference database dedicated to the pathogenic species E. coli. sRNAMap emphasizes sRNA annotation and provides ~60 sRNA interactions. sRNAdb also focuses on bacterial sRNA annotations, lacking information on their targets. BSRD provides ~205 validated sRNA-RNA interactions. Finally, sRNATarBase v3.0, features ~500 experimentally-supported low-throughput sRNA-RNA interactions.

In addition, one recently published study provides an easily accessible and interactive browser where users may navigate through bacterial sRNA interactions, which have exclusively been derived through RIL-Seq datasets. The browser stores built-in RIL-Seq-derived interactions from previously conducted experiments but also offers to users the option to upload and visualize their own RIL-Seq results[265]. Finally, experimentally supported targets can also be found in another online resource, that provides RNA-RNA interactions (https://rilseqdb.cs.huji.ac.il/Interactions) derived exclusively from two RIL-Seq experiments.

Agnodice incorporates orders of magnitude more experimentally supported bacterial sRNA-RNA interactions (~22,000). It is unique in (i) supporting advanced querying/filtering capacity in an intuitive user-friendly scheme, and (ii) hosting interactions derived from a diverse set of low-yield and state-of-the-art NGS interactome identification techniques (e.g., RIL-Seq, CLASH, CLIP-Seq, MAPS).

CHAPTER 6 – miRNA targets on non-coding transcripts and expression of lncRNAs at sub-cellular resolution

Long non-coding RNAs (LncRNAs) and microRNAs (miRNAs) constitute two of the most widely studied non-coding RNA species. They regulate gene expression at transcriptional, RNA processing, translational, and even post-translational levels by interacting with nucleic acids (both DNA and RNA) and proteins.

Apart from interacting with messenger RNAs, miRNAs also include lncRNAs in their targeting repertoire. The miRNA-lncRNA interplay is still a field under active research and development. The two primary modes of action of miRNAs on lncRNAs (or *vice versa*) are (i) the direct targeting of lncRNAs with functional roles by miRNAs (e.g., control of cell proliferation through the interaction of miR-34a with XIST) and (ii) lncRNAs acting as miRNA sponges which in turn reduces the regulatory effect of miRNAs on their mRNA targets. Furthermore, the knowledge of the total expression levels of lncRNAs in the cell or their compartment-specific expression (i.e., nucleus or cytoplasm) can prove invaluable for the study of their function (e.g., direct interaction with DNA) and their potential as indirect mRNA regulators.

6.1 A database of experimentally supported miRNA targets on non-coding transcripts

DIANA-LncBase v3.0 constitutes a reference database with experimentally supported miRNA targets on non-coding (LncRNA) transcripts. It provides ~ half a million interactions, corresponding to ~240,000 unique tissue and/or cell-type specific miRNA–lncRNA pairs. Interactions are derived from (i) the manual curation of the available literature and (ii) the re-analysis of more than 300 Next-Generation Sequencing datasets (e.g., AGO2-CLIP-Seq). The database include miRNA targets on lncRNAs for two species (i.e., *Homo sapiens* and *Mus musculus)* derived from 14 different experimental methodologies. Most of the available interactions are produced from the analysis of AGO-CLIP datasets by utilizing the microCLIP framework.

6.2 Expression levels of lncRNAs at cellular and sub-cellular resolution

In DIANA-LncBase v3.0 we also developed a novel module comprising lncRNA expression profile in numerous cell-lines at cellular and sub-cellular resolution (*Figure 6.1*). Users can retrieve the expression profiles of lncRNAs (i) within the cell (Expression mode) and (ii) comparatively between the nuclear and cytoplasmic subcellular compartments (Localization mode), coupled with a wide range of cell types in *Homo sapiens* and *Mus musculus* species.

Transcript per million (TPM) values describing the expression of lncRNAs are provided to users. In case of more than one biological replicates the median TPM value is specified. Specifically, in the 'Expression' mode the user can also retrieve results by selecting a particular range of TPM values, described as 'Low' (range:1–10), 'Medium' (range: 11–600) and 'High' (range: >600). In 'Localization' mode TPM values, estimated separately in nucleus and cytoplasm, are provided, followed by the Relative Concentration Index (RCI) and the apparent inclination of the sub-localization of lncRNAs, either towards the nucleus or the cytoplasm. The user can also retrieve the targets of the specified lncRNAs via a dedicated inter-connected link with the module for the experimentally supported targets.

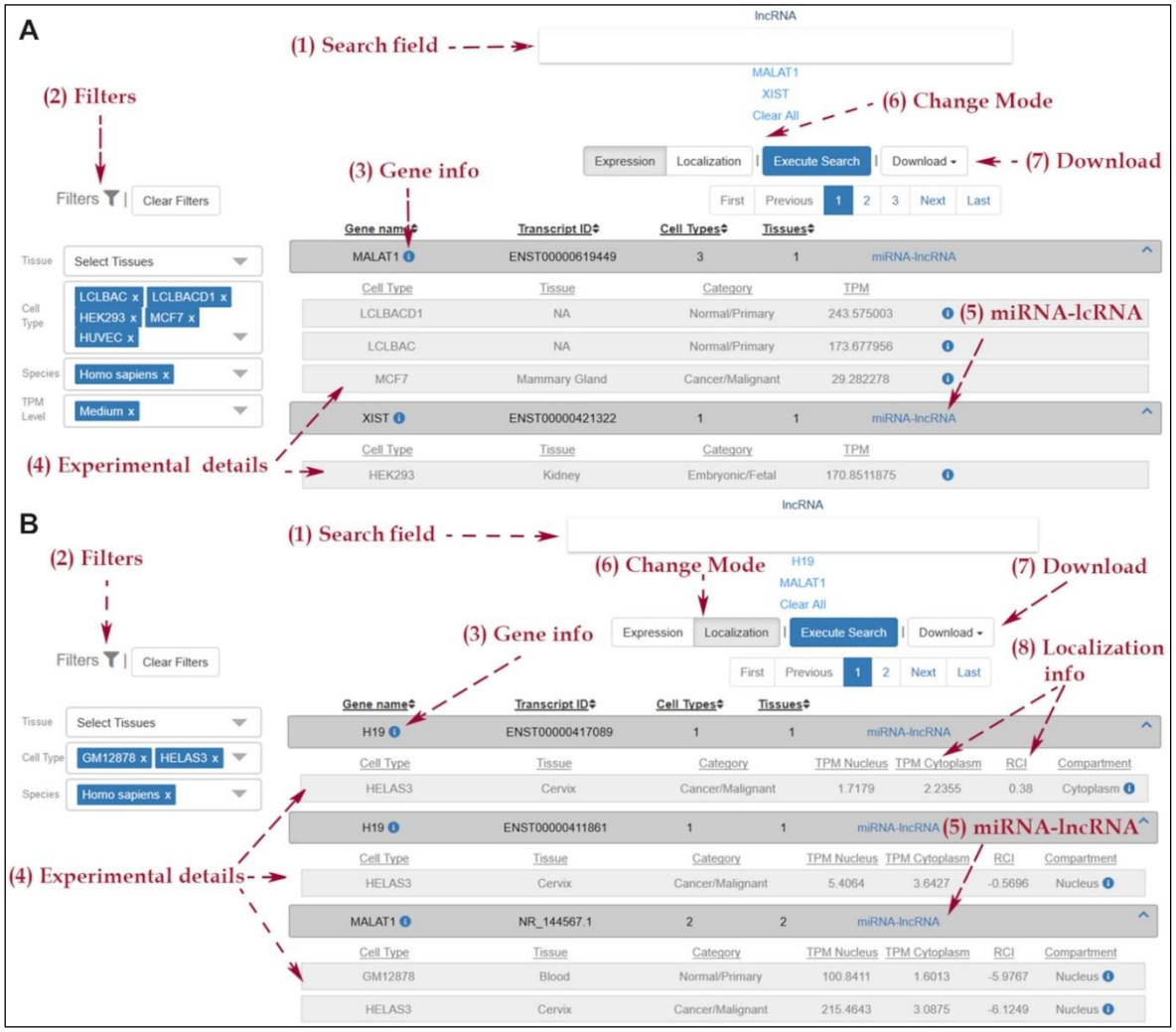

Figure 6.1: Screenshot of DIANA-LncBase v3.0 expression module. (Figure was adopted from Karagkouni G. et al., 2020)

To produce the lncRNA expression profiles, raw RNA-Seq datasets were retrieved from the ENCODE[266, 267] and GEO[268] repositories, corresponding to 34 distinct cell-lines/types and tissues for the *Homo sapiens* and *Mus musculus* species. RNA-Seq datasets corresponding to similar cell types and tissues with AGO-CLIP-Seq samples were

preferentially selected. Raw datasets were quality checked and pre-processed using FastQC and Cutadapt[269]. Quantification was conducted at the transcript level, using Salmon[270] version on quasi-mapping mode and Transcripts Per Million (TPM) values were extracted. 48 whole transcriptome libraries, corresponding to 22 cell types/lines and tissues were analyzed. Transcripts with TPM > 1 were retained, while median TPM values were estimated in case of more than one biological replicates. For the characterization of the subcellular localization of transcripts, 55 libraries from RNA-Seq experiments, conducted separately in RNA isolated from the nucleus and the cytoplasm in 15 distinct cell types/tissues, were pre-processed. Transcripts were filtered out to present TPM > 1 in at least one of the two subcellular compartments. We adopted the Relative Concentration Index (RCI)[271], estimated by transforming the cytoplasmic-to-nuclear TPM fraction into log2 scale, to define the trend of lncRNA transcripts localization towards the two different cellular compartments. Human transcriptomes were compiled from ENSEMBL 96[272], RefSeq 109[273] and Cabili et al.[274], as well as mouse transcriptomes derived from ENSEMBL 96[272] and RefSeq 106[273]. A dedicated visualizations page (*Figure 6.2*) was also designed to help users inspect the expression of lncRNAs using bar plots.

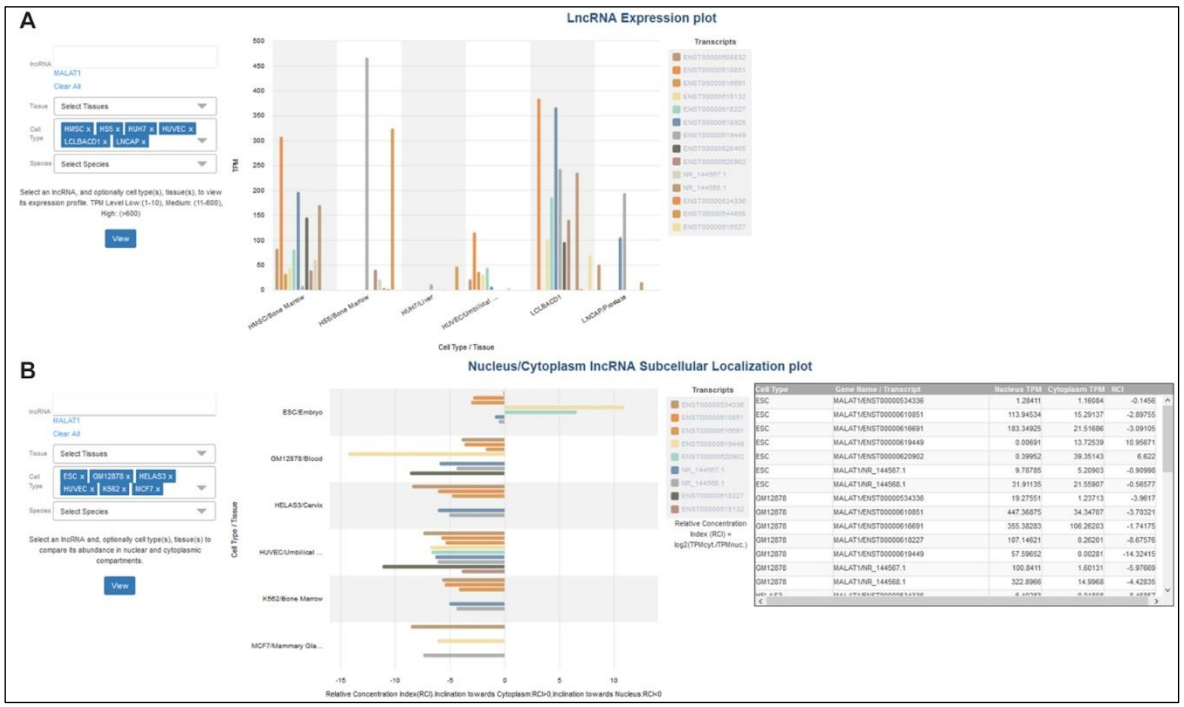

Figure 6.2: DIANA-LncBase v3.0 visualizations. (Figure was adopted from Karagkouni G. et al., 2020)

CHAPTER 7 – Host miRNA-bacterial mRNA interactions on the spotlight

In recent years, advances in NGS technologies and the development of specialized bioinformatics methods have made possible the realization of significant breakthrough in the study of host-microbiome interactions in *Homo sapiens* and other host species. It is now experimentally verified that microorganisms coexisting in human tissues and biofluids play important roles in the proper functional activity of the host by modulating fundamental cellular and molecular processes such as signal transduction, immunity and metabolism[275]. Dysregulation of the host's microbial composition has been associated with the initiation and/or progression of complex diseases such as cancer[168]. The host-microbiome interplay a is being actively investigated both at the genome level (microbial abundances, polymorphic sites etc.) and the transcriptome level (gene expression of specific species and strains) to identify specific microbes, and/or molecular features within them, with therapeutic potential or superior capacity to be used as disease risk predictors[276].

To date, a relatively small body of experimental evidence suggest that there is a high frequency of indirect and direct interactions between host miRNAs and microbial organisms. The inter-molecular communication between the host and its microbiome plays a catalytic role in their symbiosis. miRNAs have a key role in this communication as shown by studies in recent years. It is hypothesized that pathogenic microorganisms modulate the expression of host miRNAs in order to enhance their survival. At the same time, it has been experimentally confirmed that extracellular host miRNAs enriched in stool samples, can enter bacteria of the host's gut microbiota, regulate the expression of bacterial genes, and affect their growth[64]. It is also known that extracellular miRNAs have a long life span, due to their protection by microvesicles, such as exosomes, and/or RNA binding proteins (e.g., AGO2).

Studies assessing the direct or indirect influence of the microbiome on the expression of miRNAs and *vice versa* are limited. The main reasons are the lack of suitable algorithms for the analysis of the relevant datasets and the extremely high computational complexity of the task at hand. Another critical limitation is the absence of the necessary metagenomics samples that could support such a large-scale study. The computational prediction of the interactomes among host microRNAs and bacterial genes is an extremely challenging task and is expected to prove of major importance in a multitude of applications, such as the manipulation of the host microbiota with the utilization of miRNAs.

7.1 Quantification of miRNAs in the gut extracellular space

The latest version of miRBase (v22.1)[277], the reference database of miRNA sequences and annotation, consists of more than 3,500 *Homo sapiens* microRNAs. Moreover, it is estimated that the average human gut microbial super-transcriptome is comprised of more than 24 million bacterial genes. Thus, the possible interacting miRNA-microbial gene pairs amount to more than 90 billion. If we take into consideration that a single miRNA can bind to multiple regions of the same gene, the number of possible binding sites is even higher. These numbers are beyond reach even if we computationally attack the problem of predicting miRNA targets on bacterial genes. To this end, in order to reduce the search space from both sets of players (i.e., miRNAs and bacterial genes), we (i) identified the most abundant extracellular miRNAs in the gut and (ii) the most abundant bacterial species in the human gastrointestinal (GI) tract.

Stool samples are the ideal sample type to identify miRNAs present in the extracellular space of the gut. Therefore, , we downloaded from the GEO repository 119 publicly available small RNA-Seq (sRNA-Seq) datasets generated from stool samples. The samples spread over two datasets (39 and 80 samples respectively) and originated from 63 healthy humans and 53 colorectal adenoma patients. Raw sRNA-Seq datasets were quality checked and pre-processed using FastQC and Cutadapt[269]. Next, we removed 47 out of the 119 samples before quantification since they failed to pass the quality standards that we had defined. Quantification of sRNAs (e.g., miRNAs, tRNAs) with a focus on miRNAs was conducted using Manatee[278], extracting read counts per sample for downstream analyses. The average alignment rate of the stool samples to the human genome was low (< 35%); that was expected since stool samples comprise a complex biological source with a diverse set of RNA molecules (i.e., host RNA, bacterial RNA and viral RNA content). The miRNAs exhibiting the highest abundance in terms of mean and median read counts and with a presence in at least one third of the samples (i.e., 33% of the samples) were retained from both datasets for further analysis.
hsa-miR-1246, hsa-let-7a-5p, hsa-let-7b-5p and *hsa-miR-192-5p* were among the most abundant extracellular miRNAs in the stool both in terms of mean and median read counts (*Figure 7.1*, *Figure 7.2*, *Figure 7.3*, *Figure 7.4*). The final set of microRNAs, was produced based on the union of the highly abundant miRNAs from the two datasets. In total, 79 miRNAs were deemed as highly abundant.

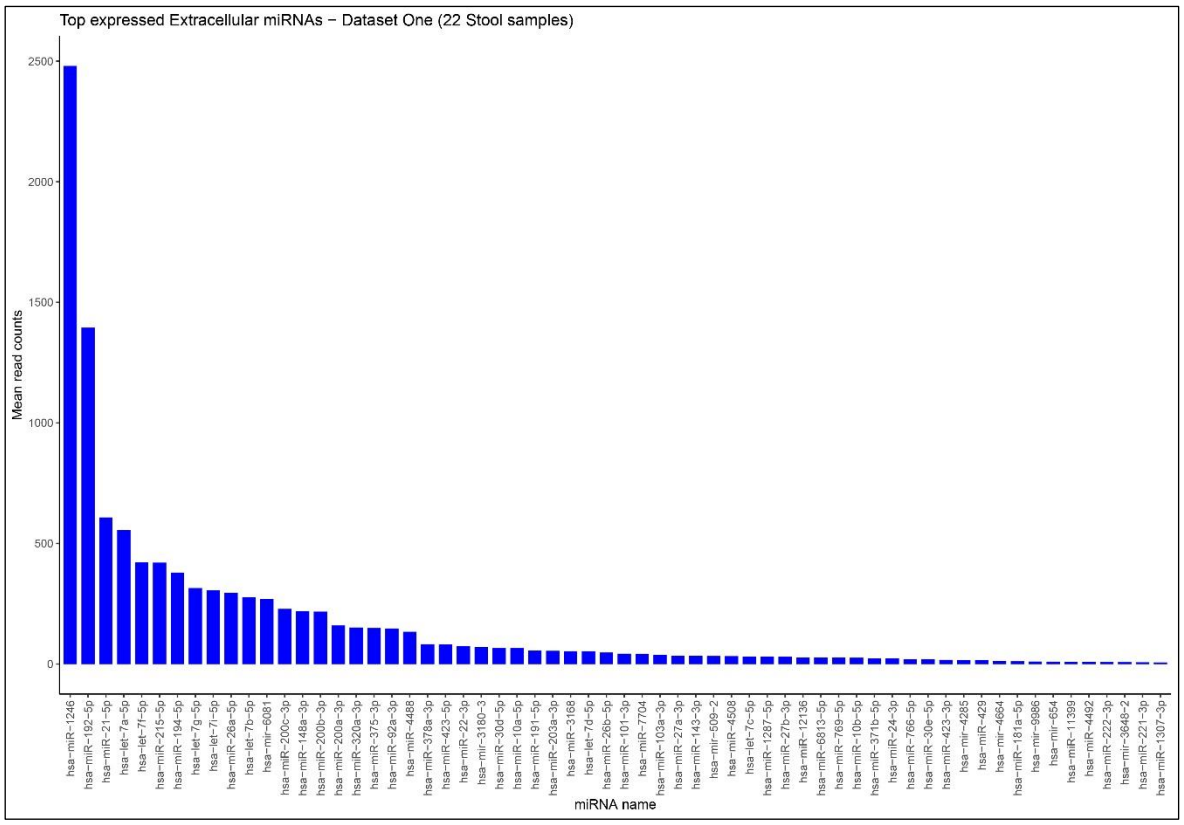

Figure 7.1: Bar plots presenting the most abundant miRNAs in terms of mean read counts from the first dataset (Figure was created for the purpose of this thesis)

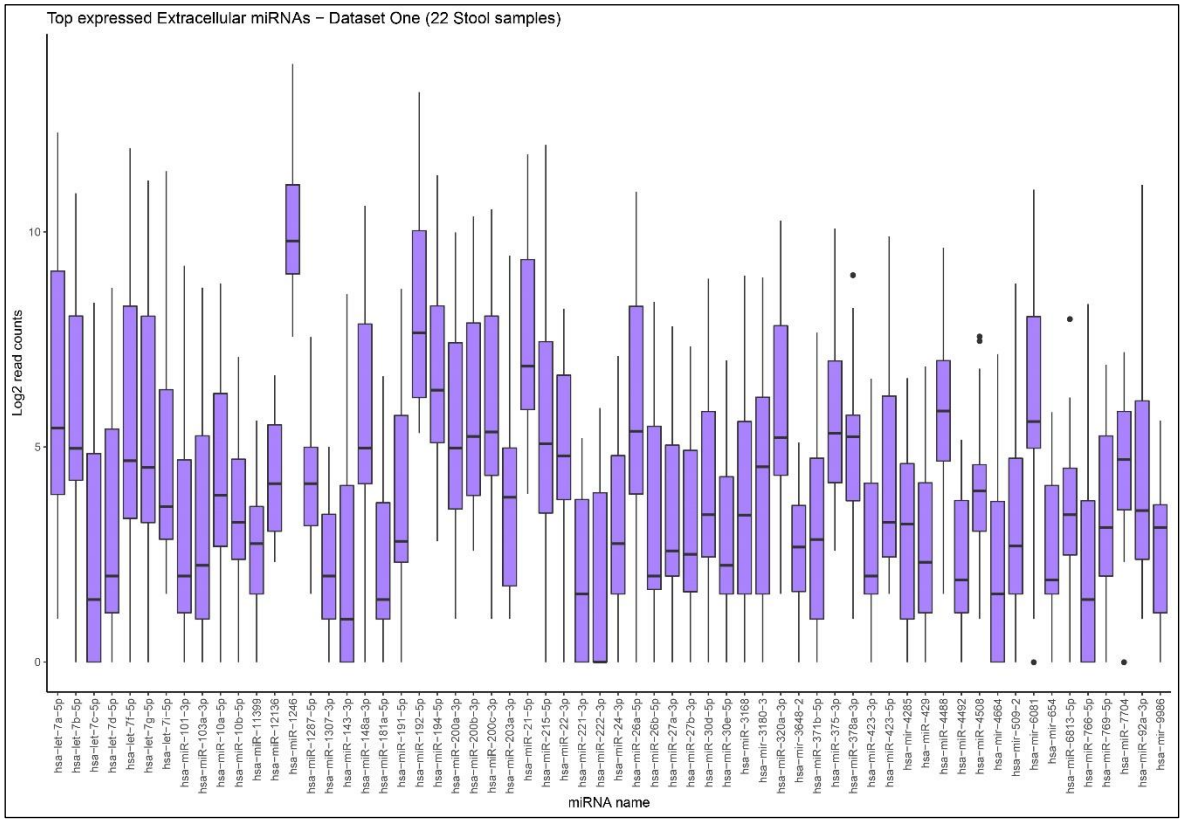

Figure 7.2: Box plots presenting the most abundant miRNAs in terms of median read counts from the first dataset (Figure was created for the purpose of this thesis)

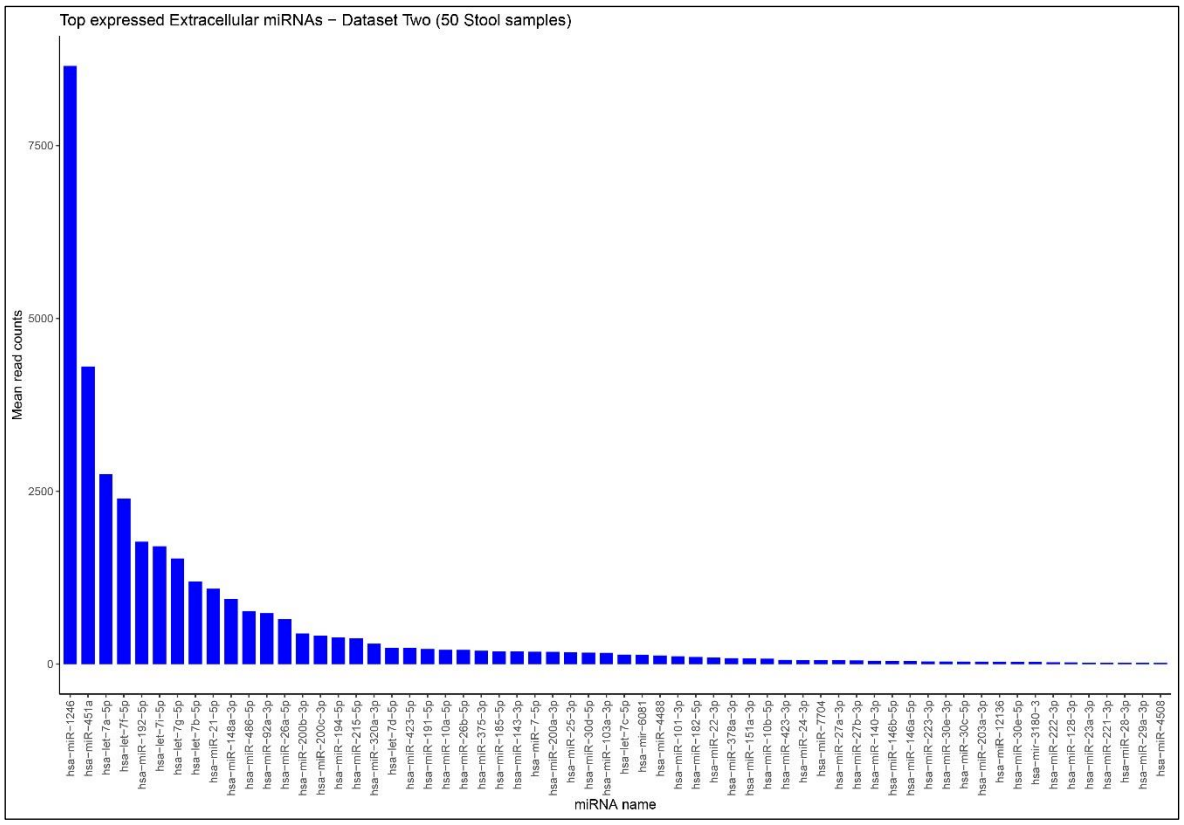

Figure 7.3: Bar plot presenting the most abundant miRNAs in terms of mean read counts from the second dataset (Figure was created for the purpose of this thesis)

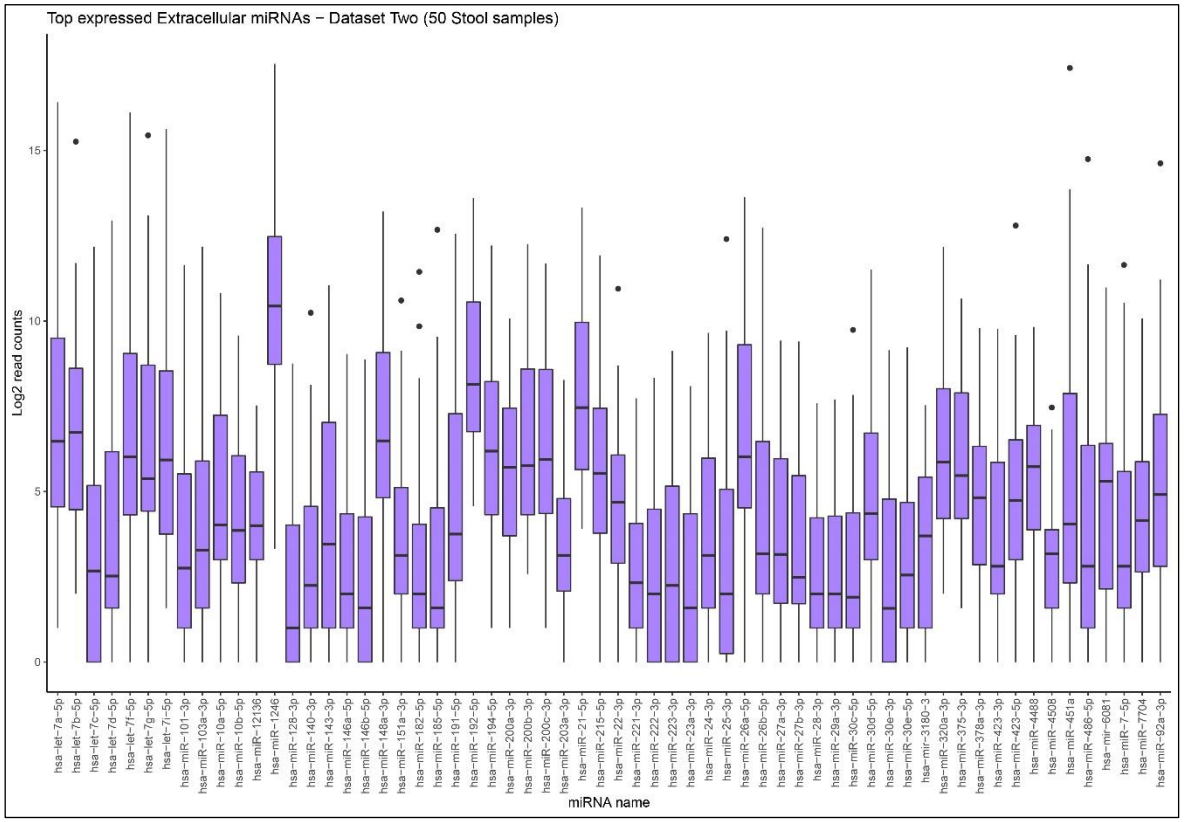

Figure 7.4: Box plot presenting the most abundant miRNAs in terms of median read counts from the second dataset (Figure was created for the purpose of this thesis)

7.2 Quantification of microbial abundances in the human gut

To identify the most abundant bacterial species in the human gut we downloaded 227 Shotgun metagenomics datasets from the Human Microbiome Project (HMP)[240]. The respective samples originated from healthy humans, while the bacterial DNA was isolated from stool. Raw Shotgun metagenomics datasets were quality checked and pre-processed using FastQC and Cutadapt[269]. We built a microbial reference index comprising all complete bacterial genomes available in the NCBI RefSeq database (> 30,000 genomes)[273]. For the quantification of microbial abundances, we utilized AGAMEMNON[225]. The average alignment rate of the samples was relatively high (> 60%) if we consider that a large part of the datasets probably originate from yet unknown bacterial species that are not part of the utilized microbial reference. To identify the bacterial species with the highest abundance, we calculated the median read counts of each *bacterium* and retained the top 100 bacterial species (*Figure 7.5*, *Figure 7.6*).

Phocaeicola vulgatus, *Bacteroides ovatus*, *Phocaeicola dorei*, *Bacteroides fragilis* and *Bacteroides uniformis* were among the most highly abundant bacterial species in the human gut. Each of the identified species comprise dozens, hundreds or even thousands of different strains/sub-species. Since the next step of the analysis needs specific bacterial genomes, for each of the identified species, we retrieved the species representative strain. On top of these bacterial genomes, we also included 15 NCBI reference genomes (Table 1). These reference genomes (e.g., *Escherichia coli str. K-12 substr. MG1655*) have been selected by NCBI based on their wide recognition as being the community standards for basic research. Additionally, other reference genomes were selected based on medical importance, sequence and annotation quality, and the availability of experimental support. In total, 62 miRNAs and 115 bacterial genomes were retained for downstream analysis.

Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819	*Escherichia coli str. K-12 substr. MG1655*	*Bacillus subtilis subsp. subtilis str. 168*
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	*Shigella flexneri 2a str. 301*	*Klebsiella pneumoniae subsp. pneumoniae HS11286*
Staphylococcus aureus subsp. aureus NCTC 8325	*Pseudomonas aeruginosa PAO1*	*Caulobacter vibrioides NA1000*
Listeria monocytogenes EGD-e	*Chlamydia trachomatis D/UW-3/CX*	*Acinetobacter pittii PHEA-2*
Mycobacterium tuberculosis H37Rv	*Coxiella burnetii RSA 493*	*Escherichia coli O157:H7 str. Sakai*

Table 7.1: Reference bacterial strains from NCBI RefSeq

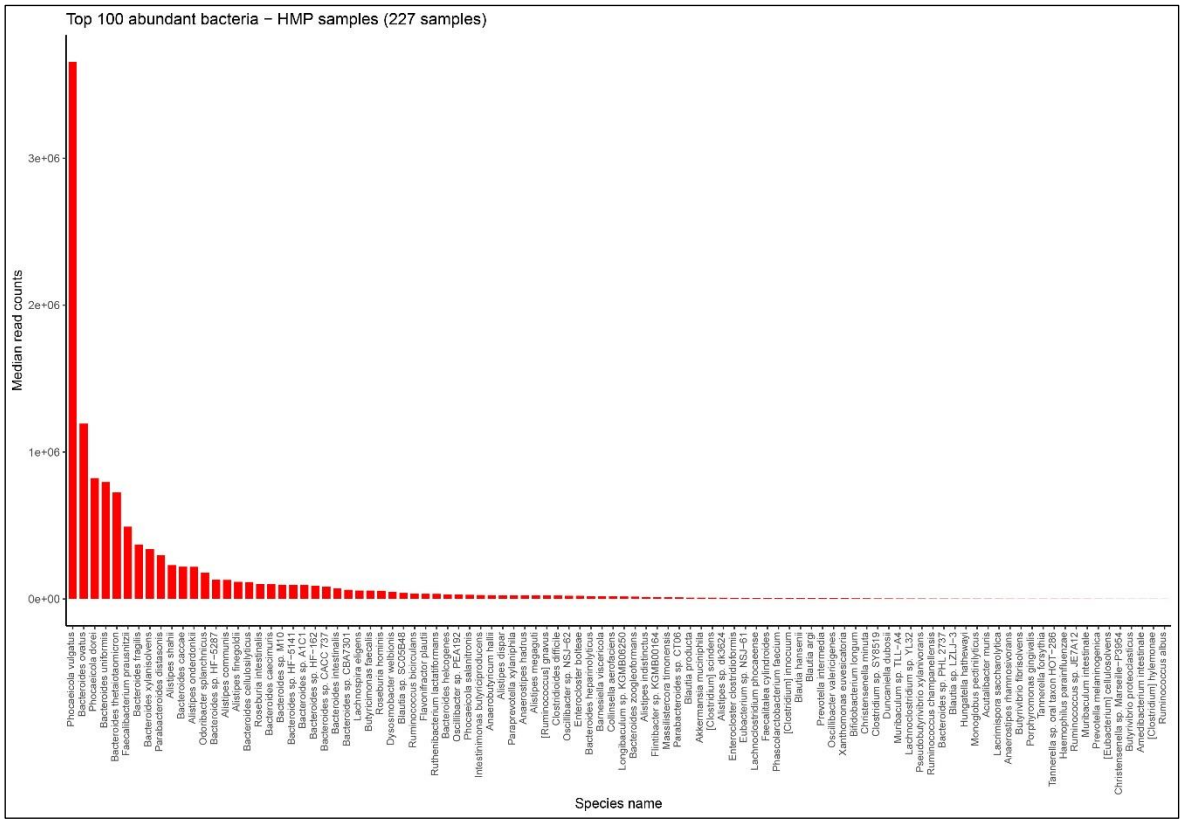

Figure 7.5: Bar plot presenting the most abundant bacterial species in the human gut using datasets from the Human Microbiome Project (Figure was created for the purpose of this thesis)

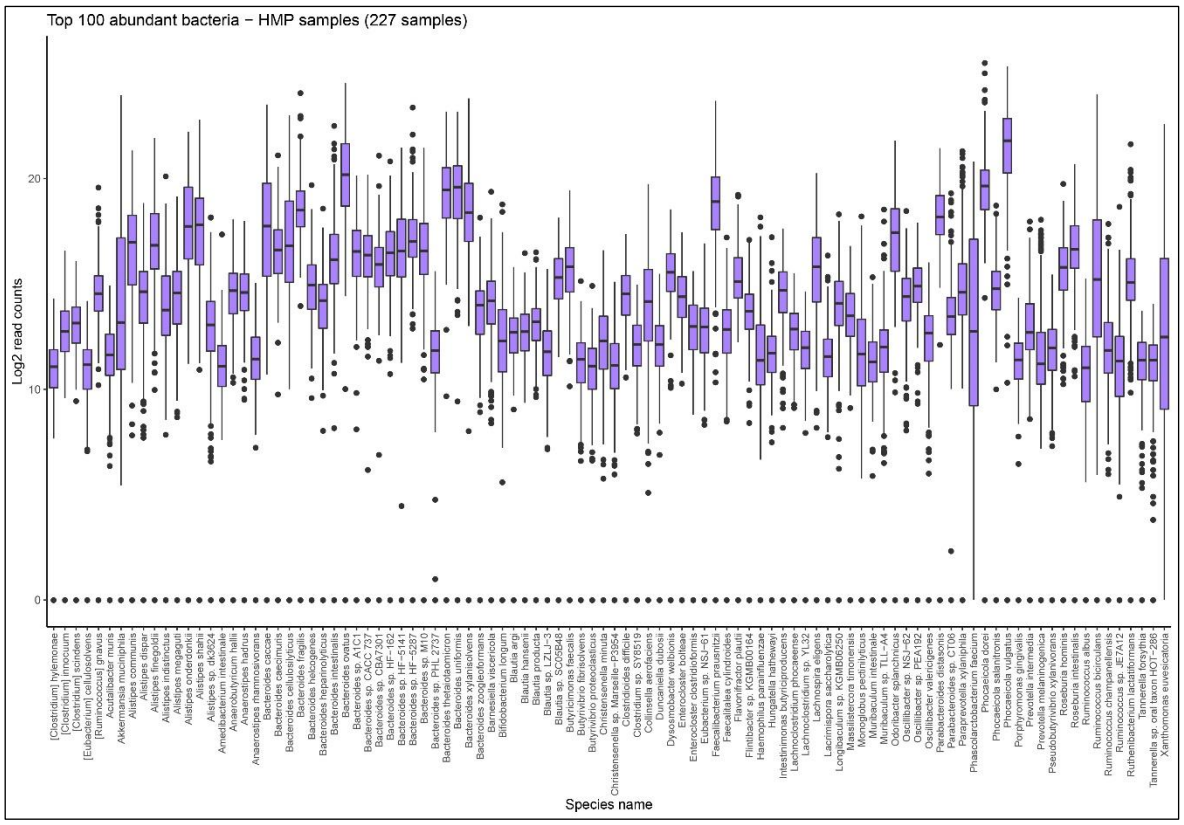

Figure 7.6: Box plot presenting the most abundant bacterial species in the human gut using datasets from the Human Microbiome Project (Figure was created for the purpose of this thesis)

7.3 Discovery of potential host miRNA-bacterial mRNA interactions

To discover potential host miRNA-bacterial mRNA interacting pairs, initially we retrieved (i) from miRBase[277] the mature sequences of the 79 microRNAs deemed highly abundant in the gut extracellular space (chapter 7.1) and (ii) from NCBI RefSeq[273] the transcriptome (i.e., the full range of RNA molecules transcribed by an organism) of each of the 115 highly abundant bacterial genomes (chapter 7.2). By utilizing the host miRNA and bacterial gene sequences, we studied known biochemical and biophysical rules of the RNA-RNA interaction space, including target sequence accessibility, dimer binding energy and sequence complementarity (i.e., matches and mis-matches in the miRNA binding region).

To this end, initially we used an in-house version of BBMap[279], specifically tailored to conduct alignments using the known miRNA binding types against target sequences. BBMap[279] was executed in an “all against all” fashion (i.e., all miRNA sequences against all gene sequences from all bacterial genomes). The output of BBMap consists of pairs of miRNAs-bacterial genes that passed an alignment score threshold (i.e., only adequate alignments were considered for downstream analysis), coupled with additional alignment-

based characteristics such as relative coordinates of the match between the miRNA and bacterial gene sequences, consecutive matches in the miRNA seed, number of GU wobbles[280] in the miRNA seed, binding type (e.g., seed match, imperfect seed match) etc. BBMap[279] results significantly reduced the candidate miRNA-bacterial gene couples, since many alignments did not pass the threshold that had been set.

Next, by considering only the subset of adequate miRNA-bacterial gene candidate pairs denoted by BBMap, we utilized RNAduplex from the ViennaRNA[281] package to calculate minimum free energy (MFE) estimates of the duplex between each miRNA and bacterial gene's RNA, for each miRNA-bacterial gene pair. MFE is an indicator of the stability between two interacting RNAs (i.e., the lower the MFE, the higher the stability between the interacting molecules under investigation).

Furthermore, the accessibility of the bacterial genes participating in the candidate miRNA-bacterial gene couples was calculated with RNAfold from the Vienna RNA suite[281]. In a nutshell, RNAfold predicts the secondary structure (*Figure 7.7*) of single RNA/DNA sequences using the dynamic programming algorithm originally proposed by Zuker and Stiegler[282]. Usually, the secondary structure predictions on RNA molecules characterize their subregions as single-stranded or double-stranded depending on their self-complementarity with other subregions of the same RNA. Single-stranded regions are more easily accessible for targeting by miRNAs, while for double-stranded regions, the RNA target would have to undergo secondary structure changes and/or unfold locally first.

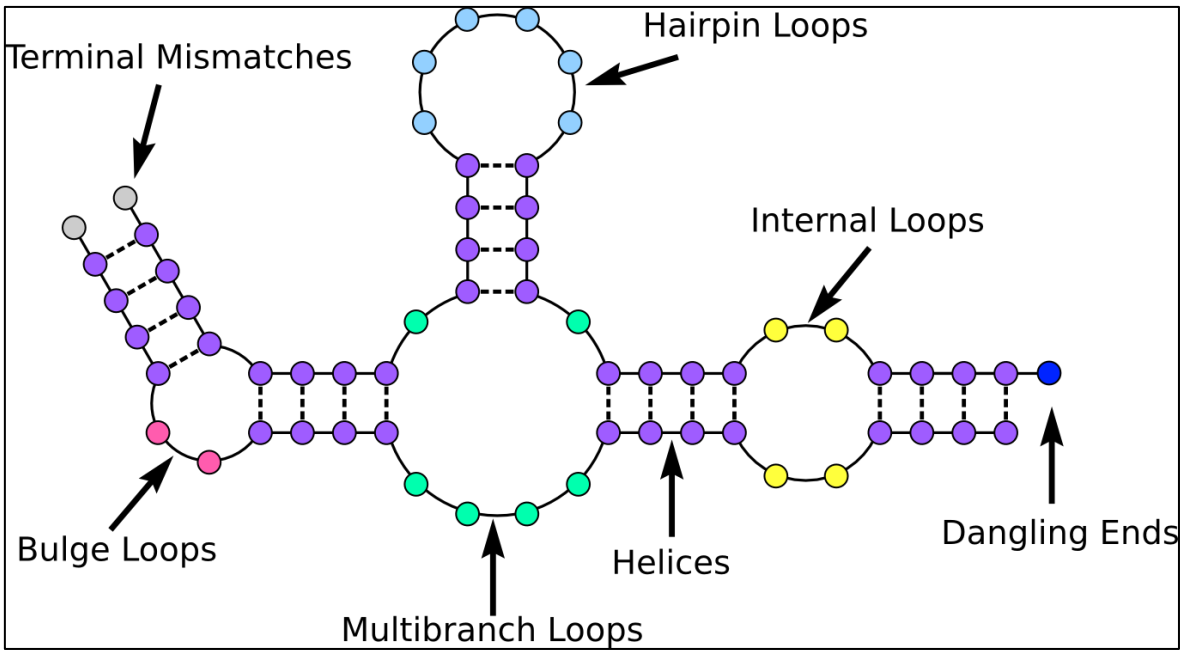

Figure 7.7: Example of a predicted RNA secondary structure. Hairpin, Internal, Bulge and Multibranch loops are all single-stranded RNA regions and thus more easily accessible by miRNAs (Figure was created for the purpose of this thesis)

Collectively, for each candidate human miRNA-bacterial gene we calculated and investigated the following general biochemical and biophysical characteristics:

- Sequence complementarity
- Dimer minimum free binding energy
- Accessibility of the genes using their secondary structure
- Number of GU wobbles in the miRNA seed region
- Number of consecutive matches in the miRNA seed region
- Information of whether the miRNA seed region participates in the miRNA-gene duplex

The minimum free energy values of candidate interacting human miRNAs-bacterial genes, when bacterial genes are grouped by their genome of origin, follow the normal distribution starting from values close to zero and expanding to values lower than -30. The median of the distribution is close to -10. Furthermore, most of the times, there are no GU wobbles in the miRNA seed, with a few exceptions ranging from 1 to 4 GU wobbles. In addition, the median number of consecutive miRNA seed matches usually equals 5 matches but ranges from 1 and up to 8 nucleotides. Finally, the median number of accessible nucleotides in the gene binding region is 10 but ranges from 1 to more than 20 nucleotides. These findings are also shown in *Figure 7.8* and *Figure 7.9* for the miRNA-gene interactions of eight representative bacterial genomes but follow the same distributions for all 115 bacterial genomes that were part of the analysis.

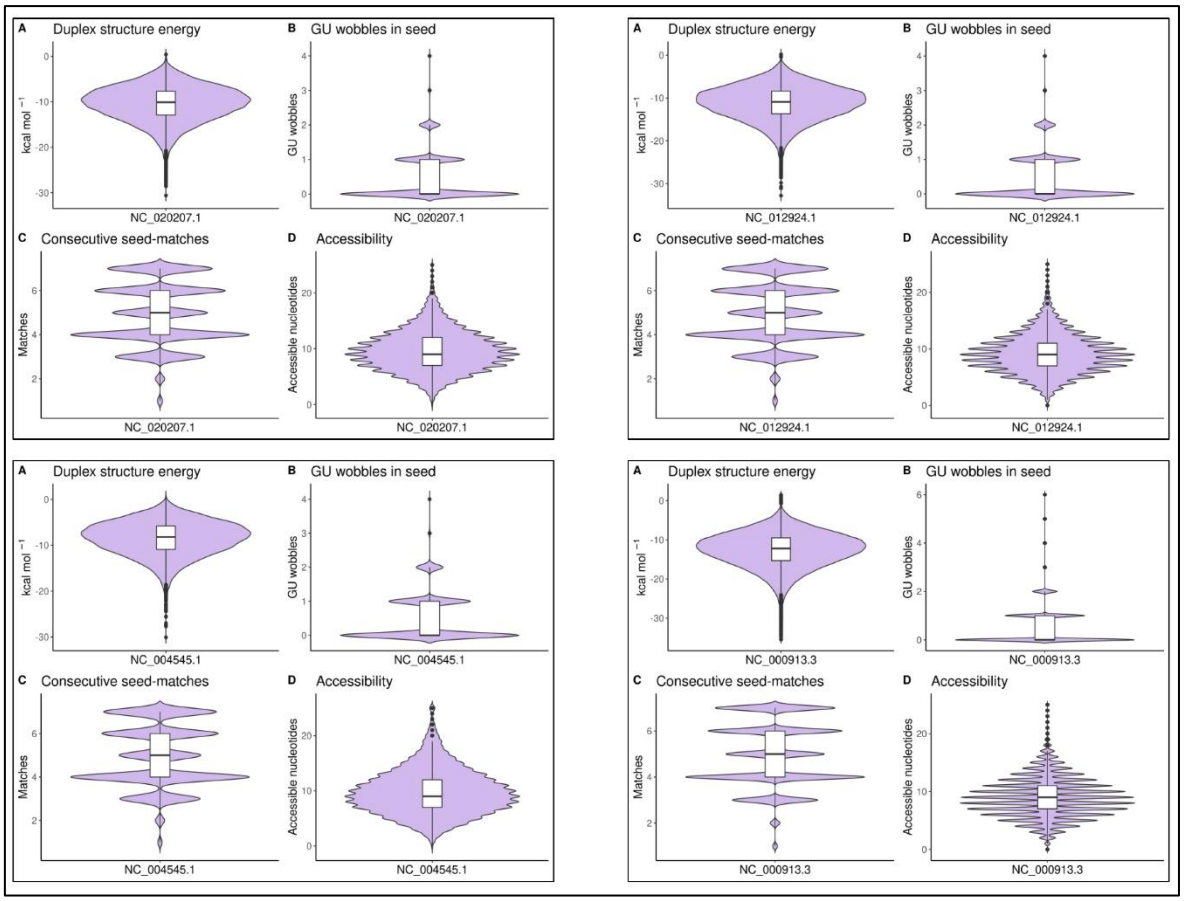

Figure 7.8: The distribution of duplex structure energy, GU wobbles in miRNA seed, consecutive miRNA seed matches and number of accessible nucleotides in the gene sequence for all predicted interacting miRNAs-genes for the bacterial strains Enterococcus faecium ATCC 8459 = NRRL B-2354 (NC_020207.1), Streptococcus suis SC84 (NC_012924.1), Buchnera aphidicola str. Bp (NC_004545.1) and Escherichia coli K12 (NC_000913.1) (Figure was created for the purpose of this thesis)

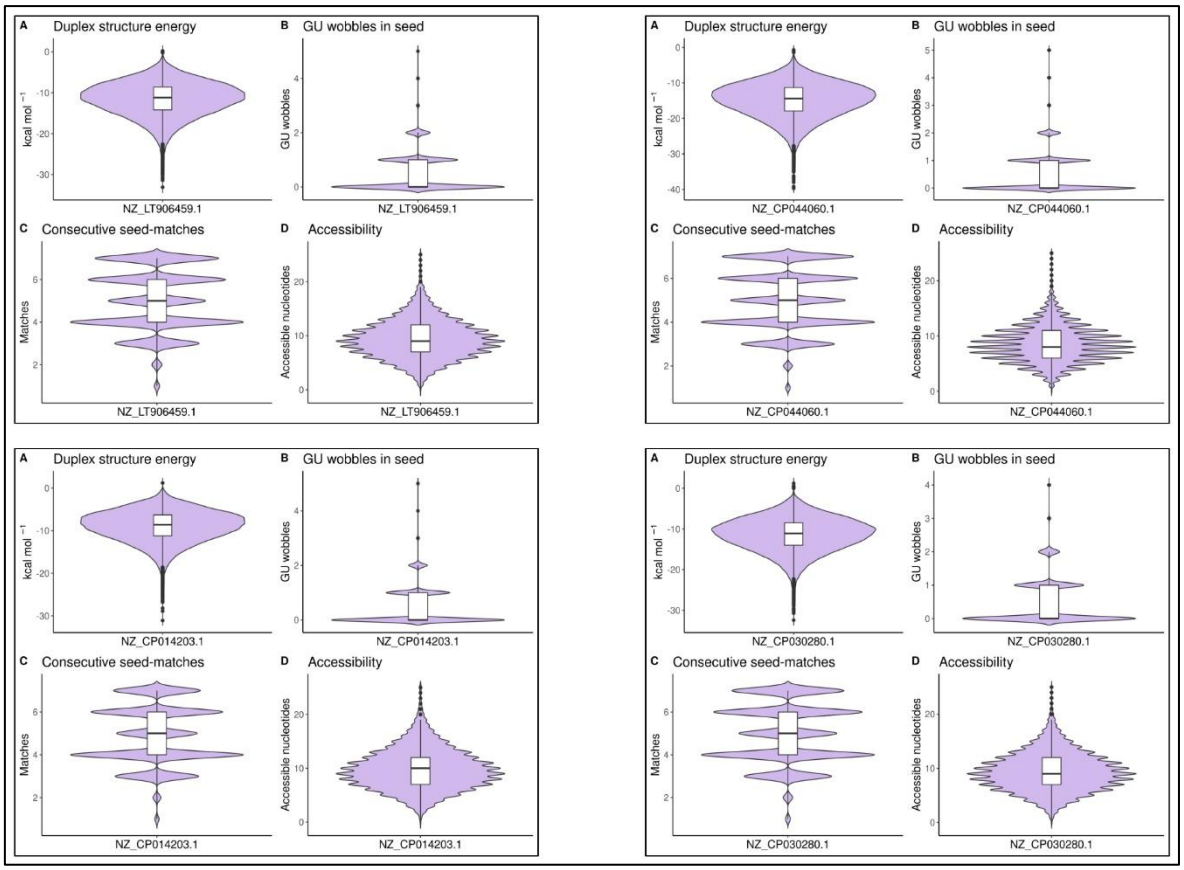

Figure 7.9: The distribution of duplex structure energy, GU wobbles in miRNA seed, consecutive miRNA seed matches and number of accessible nucleotides in the gene sequence for all predicted interacting miRNAs-genes for the bacterial strains Odoribacter splanchnicus strain NCTC10825 (NZ_LT906459.1), Aeromonas veronii strain FDAARGOS_632 (NZ_CP044060.1), Clostridium baratii strain CDC51267 (NZ_CP014203.1) and Blautia argi strain KCTC 15426 (NZ_CP030280.1) (Figure was created for the purpose of this thesis)

To rank the predicted human miRNA-bacterial gene pairs, we applied a function that takes into consideration the biochemical and biophysical properties of the interactions and calculates a unique score for each of the interacting pairs. The scoring function is demonstrated below:

$$S(i) = |\,MFE\,| + C + A + GU + SM$$

where MFE denotes the minimum free energy, C denotes the number of consecutive miRNA seed matches, A denotes the number of accessible nucleotides in the gene binding region, GU denotes the number of GU wobbles in the miRNA seed and SM is a binary value indicating whether the miRNA seed participates in the miRNA-gene duplex (1) or not (0). Downstream of the calculated scores, a normalization is performed to ensure that the scores will be bounded between 0 and 1. A value close to 1 indicates a good interacting score while a value close to 0 indicates the opposite.

The normalization scheme that is applied to the raw score values is demonstrated in the equation below:

$$Ns(i) = \frac{S(i) - min(S)}{max(S) - min(S)}$$

Where S(i) is the score of a miRNA-gene interacting pair and S is the numerical vector comprising all interacting scores.

The distribution of the normalized miRNA-gene interaction scores independently in four different bacterial genomes is shown in *Figure 7.10*. Furthermore, the distribution of the normalized scores for all bacterial genomes together (n = 115) is presented in *Figure 7.11*. It is evident that both the per bacterial genome interaction scores and the scores from all bacterial genomes together follow a Normal distribution with a median close to 0.5.

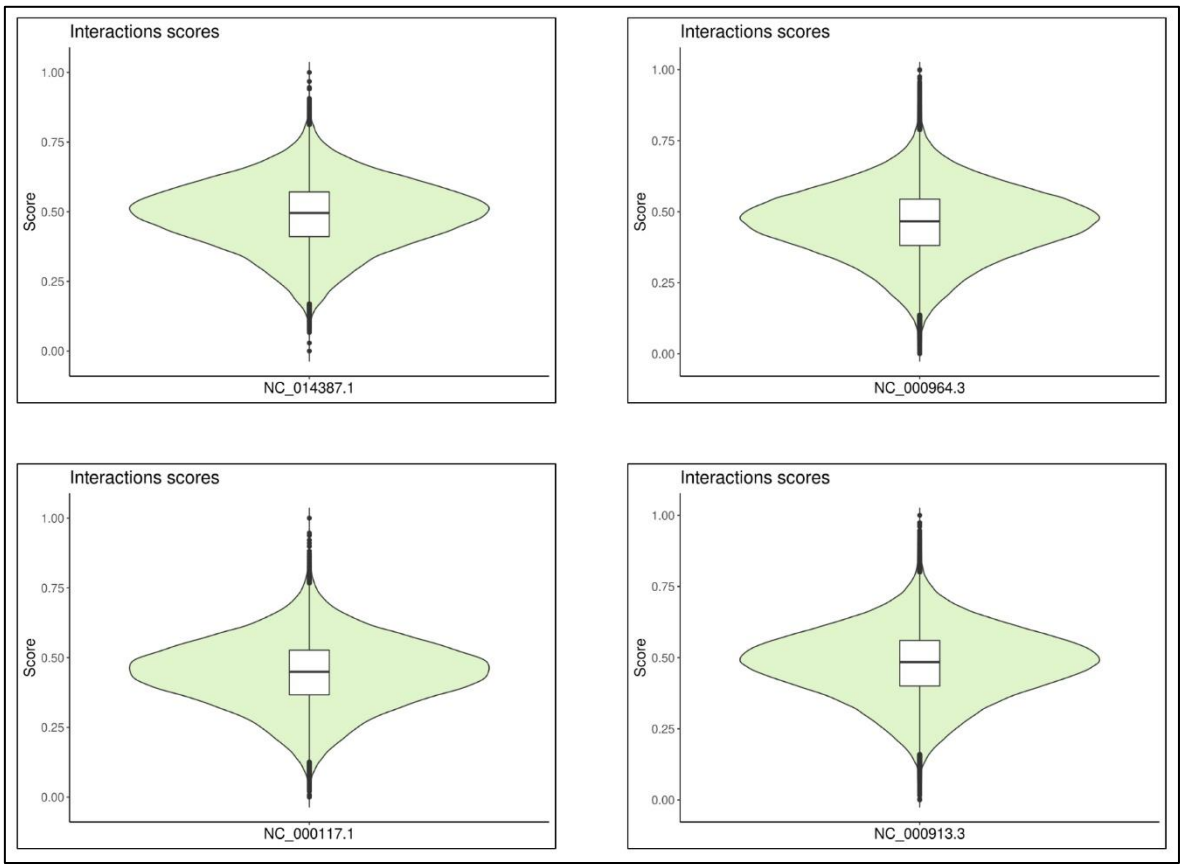

Figure 7.10: The distribution of human miRNA-bacterial gene interaction normalized scores for the genes originating from the bacterial strains Butyrivibrio proteoclasticus B316 (NC_014387.1), Bacillus subtilis subsp. subtilis str. 168 (NC_000964.3), Chlamydia trachomatis D/UW-3/CX (NC_000117.1) and Escherichia coli K12 (NC_000913.1) (Figure was created for the purpose of this thesis)

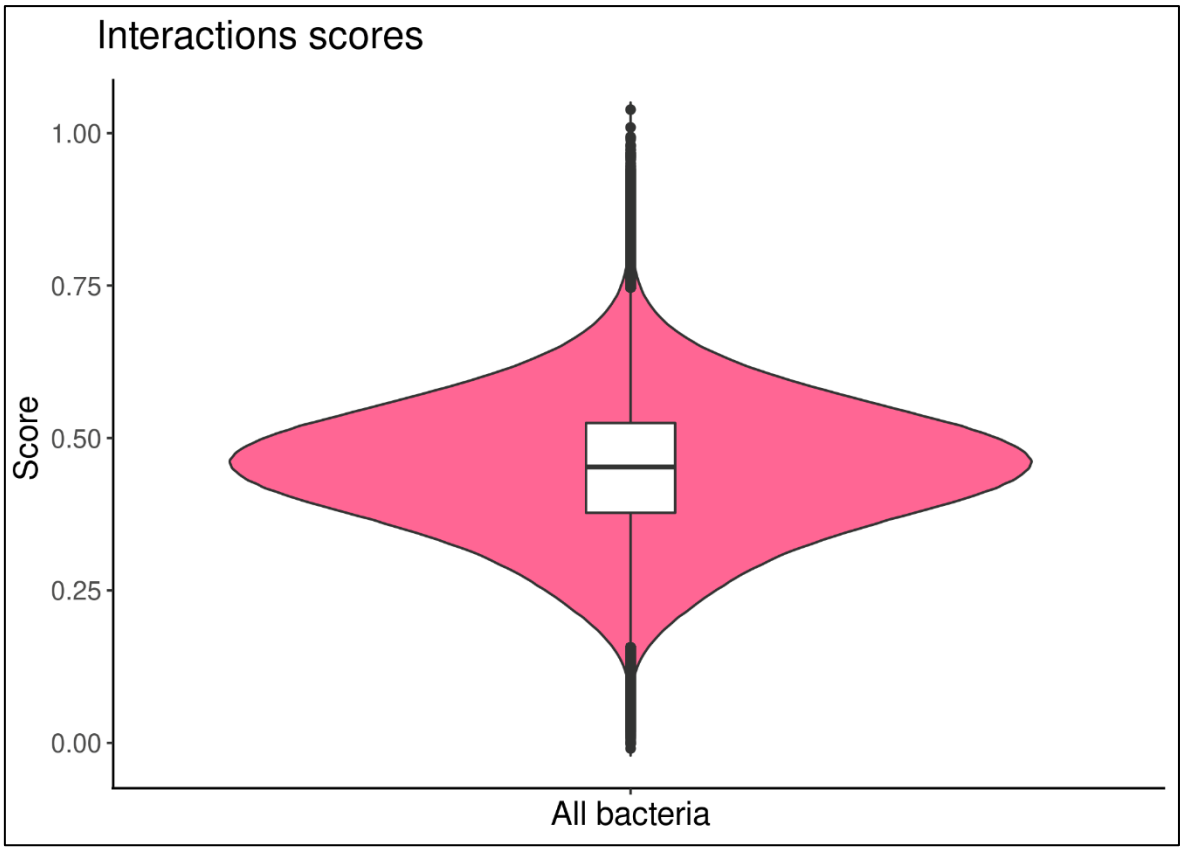

Figure 7.11: The distribution of human miRNA-bacterial gene interaction normalized scores for bacterial genes originating from all bacterial genomes together (n = 115) (Figure was created for the purpose of this thesis)

7.4 Functional interpretation and validation of the human miRNA-bacterial gene candidate pairs

The total number of human miRNA-bacterial gene interactions derived from the aforementioned analysis exceed 11 million pairs. As shown in *Figure 7.11*, these interactions occupy the full range of normalized scores (i.e., 0 - 1) with a median value close to 0.5. To focus on the most promising interactions, we kept the pairs with normalized score >= 0.6 for further analysis. The total number of interactions after filtering comprise ~670,000 entries. Their median score is 0.63 and range between 0.6 and 1. The distribution of the normalized scores for the top interactions is presented in *Figure 7.12*.

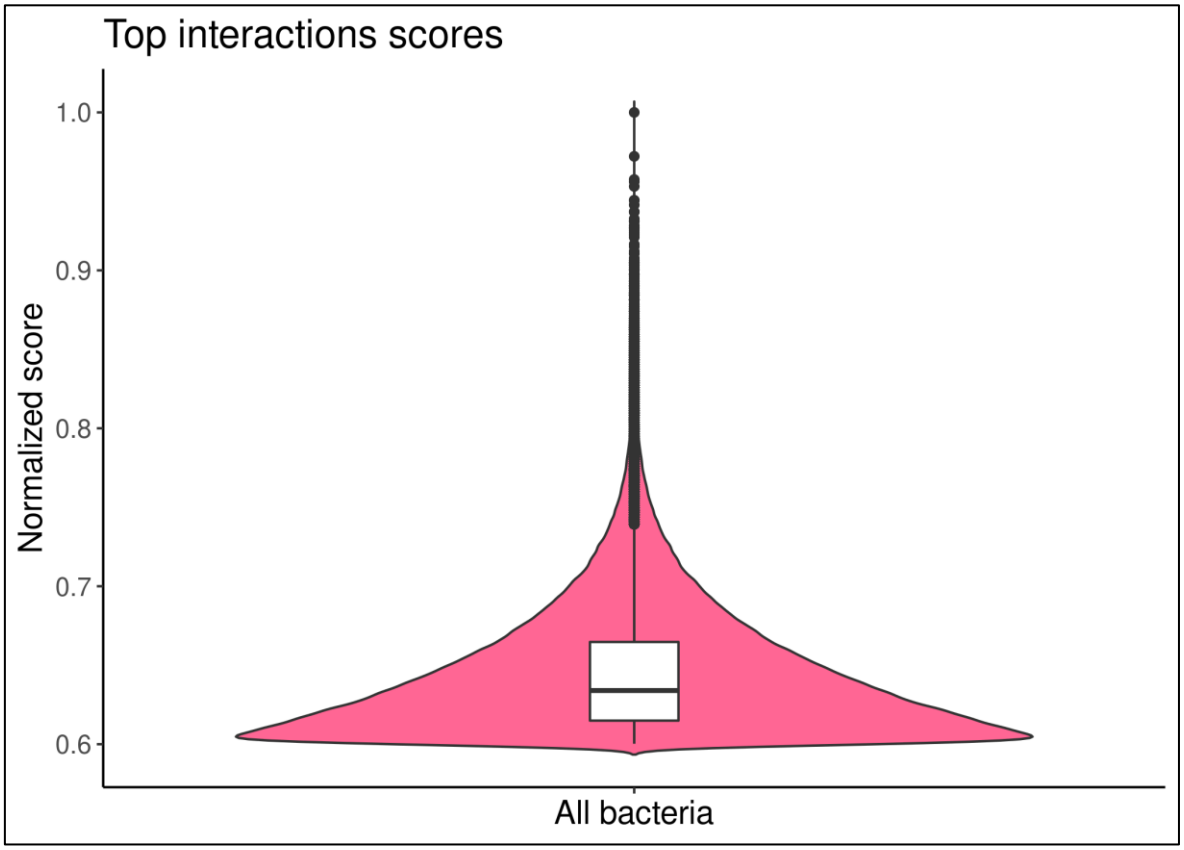

Figure 7.12: The distribution of the top human miRNA-bacterial gene interactions in terms of normalized scores (i.e., score >= 0.6) (Figure was created for the purpose of this thesis)

Next, to functionally interpret and computationally validate the filtered miRNA-gene interactions, we started by categorizing the genes participating in the top interactions from every bacterial genome based on their essentiality. Generally, essential genes are defined as the set of genes that are critical for the survival of an organism. Mostly, they are genes that are necessary for the cell to grow, proliferate and survive. Deletion of an essential gene from a cell eventually leads to its death (i.e., it is lethal/deleterious) or to a severe proliferation defect. On the contrary, non-essential genes, even though they may carry important

biological roles, are not fundamental and/or necessary for a cell to survive. To this end, we hypothesized that if the discovered miRNA-gene interacting pairs represent true positive results with functional consequences for the bacteria under regulation, there must be a significant enrichment towards the targeting of essential bacterial genes from the hosts miRNAs.

To label bacterial genes as "essential" and "non-essential" we utilized GepTop2[283]. This computational method estimates gene essentiality for prokaryotes based on orthology and phylogeny by probabilistically assigning an essentiality score in the bacterial genes of interest and then labeling the genes based on that score. We calculated the essentiality of all genes (default parameters of GepTop2 were used) from all bacterial genomes that were part of the top interactions. Expectedly, most of the times, more genes were labeled as non-essential (*Figure 7.13*, *Figure 7.14*, *Figure 7.15*, *Figure 7.16*, *Figure 7.17*).

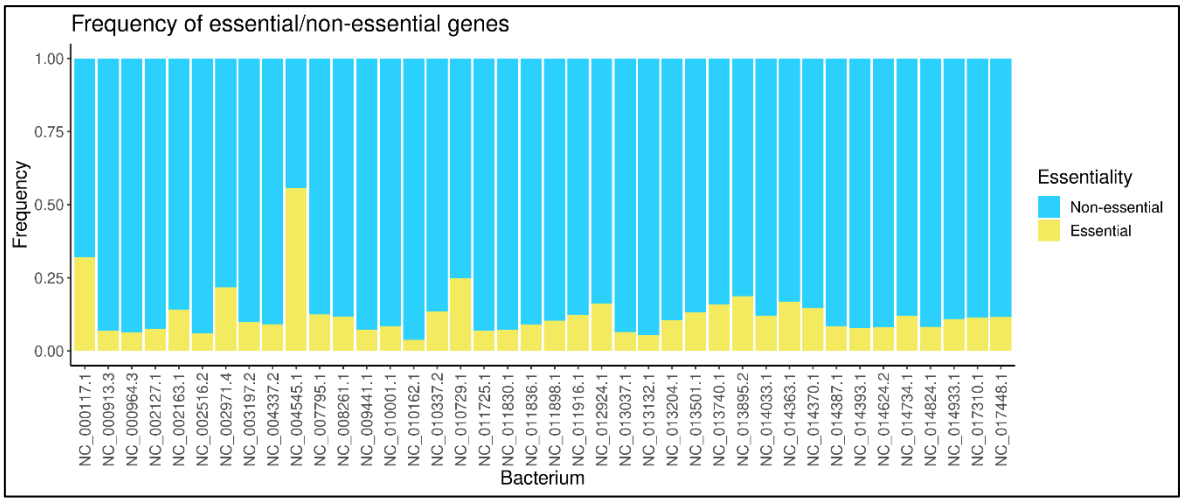

Figure 7.13: Frequency of essential and non-essential genes per bacterial genome (subset 1) (Figure was created for the purpose of this thesis)

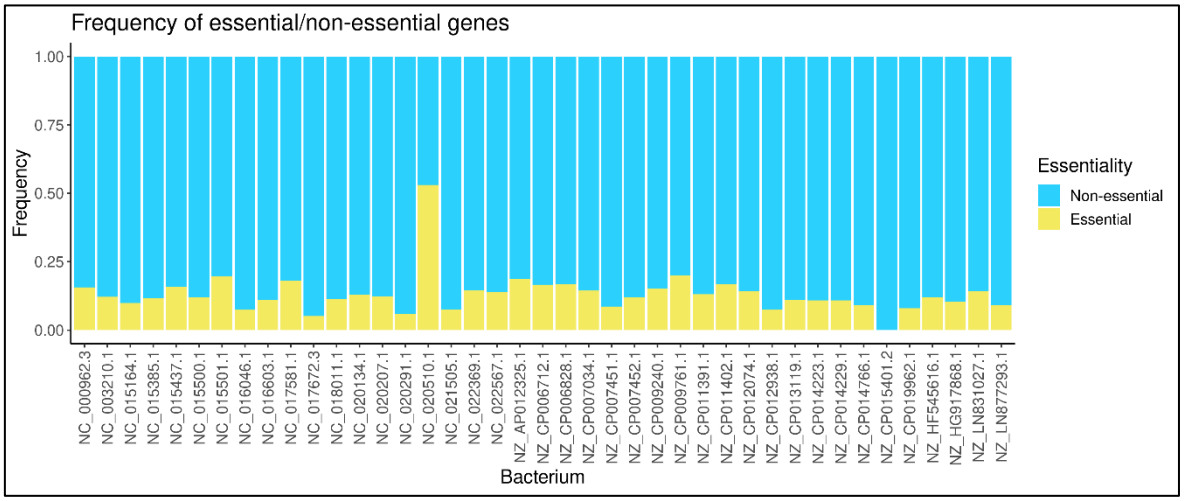

Figure 7.14: Frequency of essential and non-essential genes per bacterial genome (subset 2) (Figure was created for the purpose of this thesis)

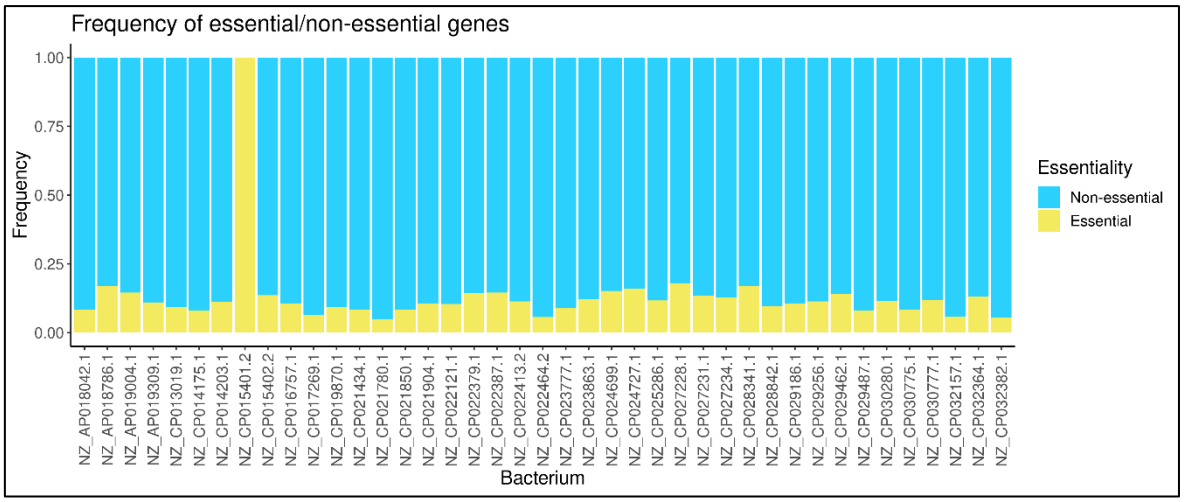

Figure 7.15: Frequency of essential and non-essential genes per bacterial genome (subset 3) (Figure was created for the purpose of this thesis)

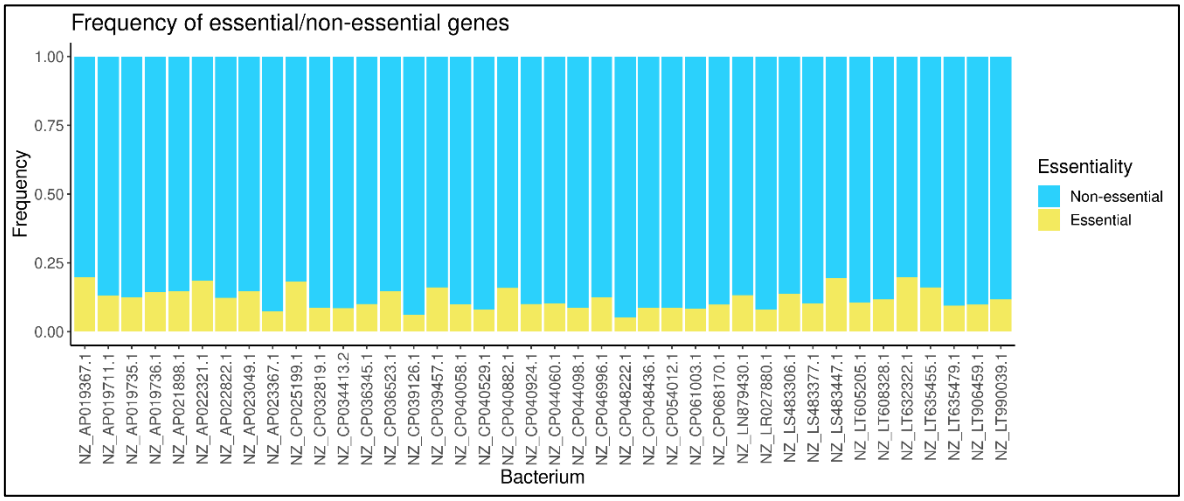

Figure 7.16: Frequency of essential and non-essential genes per bacterial genome (subset 4) (Figure was created for the purpose of this thesis)

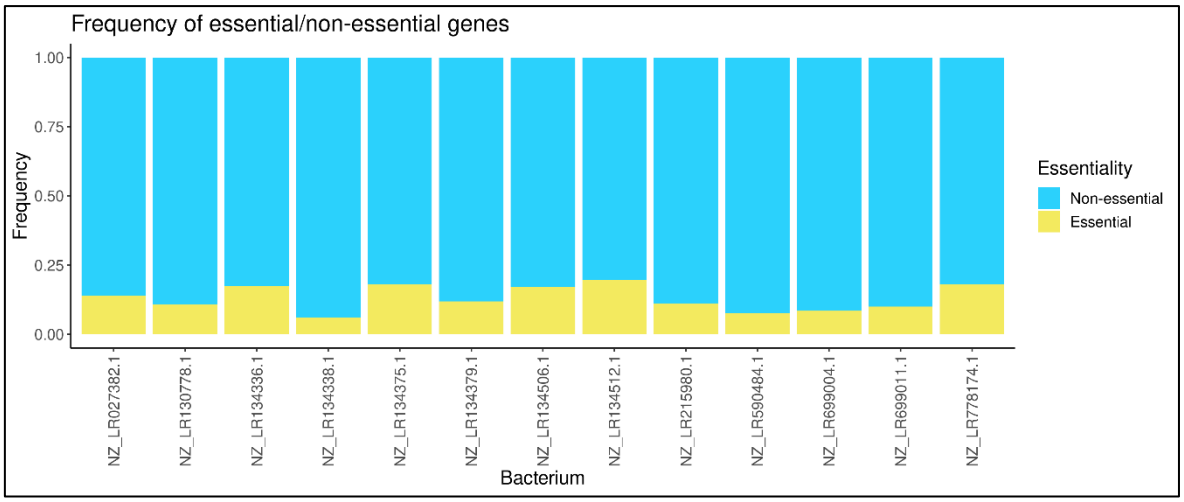

Figure 7.17: Frequency of essential and non-essential genes per bacterial genome (subset 5) (Figure was created for the purpose of this thesis)

Next, we created a table comprising the top interactions (n = 670,000) coupled with information on the host miRNA, the bacterial gene, the bacterial genome and whether the gene is labeled as essential or non-essential. Subsequently, for each bacterial genome we performed a goodness-of-fit test (chi-squared test with one dimensional contingency table) to assess the significance of imbalance between essential and non-essential genes. Prior to the statistical test, we normalized the number of interactions in essential and non-essential genes based on the total number of essential/non-essential genes per bacterial genome. Intriguingly, in 172 out of the 174 bacterial genomes tested, there is a significant enrichment in miRNA targeting events (interactions) towards essential genes (Table 2) implying a functional role of the predicted interactions (e.g., control of bacterial growth-rate by host miRNAs).

Bacterium	Interactions (essential genes)	Interactions (non-essential genes)	Test statistic	P-value
NC_000117.1	4042	2733	253.1	5.37899E-57
NC_000913.3	5084	4616	22.6	1.96534E-06
NC_000962.3	10845	5873	1478.8	0
NC_000964.3	3144	3176	0.2	0.684189061
NC_002127.1	5238	3797	229.9	6.36023E-52
NC_002163.1	1000	929	2.6	0.10541758
NC_002516.2	11667	4969	2696.1	0
NC_002971.4	4471	1047	2125.2	0
NC_003197.2	3580	2829	88.0	6.62794E-21
NC_003210.1	1842	405	919.2	6.5685E-202
NC_004337.2	5164	3598	279.7	8.84201E-63
NC_004545.1	1349	844	116.4	3.8722E-27
NC_007795.1	477	61	321.3	7.66894E-72

NC_008261.1	833	210	371.4	9.29405E-83
NC_009441.1	1132	195	661.9	5.742E-146
NC_010001.1	1205	234	654.2	2.6891E-144
NC_010162.1	10432	1256	7205.1	0
NC_010337.2	7217	2202	2670.4	0
NC_010729.1	2206	1128	348.4	9.6165E-78
NC_011725.1	1542	381	701.2	1.6155E-154
NC_011830.1	3175	740	1514.2	0
NC_011836.1	1246	260	644.6	3.2761E-142
NC_011898.1	1679	387	808.2	8.8151E-178
NC_011916.1	5269	1868	1619.9	0
NC_012924.1	1863	613	630.2	4.4441E-139
NC_013037.1	3943	536	2592.1	0
NC_013132.1	2562	310	1766.1	0
NC_013204.1	6698	1302	3640.5	0
NC_013501.1	7060	1951	2896.8	0
NC_013740.1	5271	1520	2072.3	0
NC_013895.2	2410	656	1003.7	2.8199E-220
NC_014033.1	2972	559	1648.7	0
NC_014363.1	8689	1891	4368.2	0
NC_014370.1	1715	384	843.4	2.0094E-185
NC_014387.1	1402	251	801.1	3.1886E-176
NC_014393.1	968	180	540.0	1.8925E-119
NC_014624.2	3525	654	1972.2	0
NC_014734.1	2311	433	1286.1	1.1908E-281
NC_014824.1	2761	398	1767.7	0
NC_014933.1	2679	477	1536.6	0
NC_015164.1	2477	591	1159.6	3.6704E-254
NC_015385.1	2294	320	1489.9	0
NC_015437.1	4287	1327	1559.8	0
NC_015500.1	3577	623	2078.1	0
NC_015501.1	4785	1508	1706.2	0
NC_016046.1	5582	909	3364.1	0
NC_016603.1	2491	1924	73.0	1.32022E-17
NC_017310.1	7372	1829	3338.7	0
NC_017448.1	4355	582	2883.4	0
NC_017581.1	1818	478	782.3	3.7714E-172
NC_017672.3	4385	821	2440.4	0
NC_018011.1	6085	897	3855.1	0
NC_020134.1	2526	595	1195.5	5.7998E-262
NC_020207.1	1689	414	773.0	4.0656E-170
NC_020291.1	895	164	504.0	1.2679E-111
NC_020510.1	921	442	168.1	1.9702E-38
NC_021505.1	8261	2327	3325.8	0
NC_022369.1	1423	411	559.0	1.4003E-123
NC_022567.1	123	40	42.2	8.43571E-11

NZ_AP012325.1	5113	981	2801.7	0
NZ_AP018042.1	1267	229	721.0	8.2133E-159
NZ_AP018786.1	8725	2290	3759.2	0
NZ_AP019004.1	2144	583	894.1	1.9157E-196
NZ_AP019309.1	1370	307	674.2	1.1922E-148
NZ_AP019367.1	8954	2644	3433.4	0
NZ_AP019711.1	967	170	558.8	1.5456E-123
NZ_AP019735.1	6514	1036	3974.9	0
NZ_AP019736.1	6355	1094	3715.0	0
NZ_AP021898.1	6885	1194	4008.4	0
NZ_AP022321.1	1673	580	529.5	3.5832E-117
NZ_AP022822.1	1480	337	718.4	2.9276E-158
NZ_AP023049.1	4935	1020	2573.8	0
NZ_AP023367.1	1660	276	988.8	4.9477E-217
NZ_CP006712.1	6185	1342	3116.4	0
NZ_CP006828.1	1444	394	599.9	1.7312E-132
NZ_CP007034.1	4991	1481	1903.7	0
NZ_CP007451.1	1833	369	973.9	8.3265E-214
NZ_CP007452.1	2387	684	944.4	2.2239E-207
NZ_CP009240.1	5175	1765	1675.1	0
NZ_CP009761.1	753	169	371.1	1.08523E-82
NZ_CP011391.1	3927	910	1881.8	0
NZ_CP011402.1	173	17	127.7	1.31488E-29
NZ_CP012074.1	2088	499	976.2	2.737E-214
NZ_CP012938.1	2099	292	1365.2	7.7411E-299
NZ_CP013019.1	1111	239	562.3	2.6724E-124
NZ_CP013119.1	5642	1662	2168.9	0
NZ_CP014175.1	949	163	554.7	1.205E-122
NZ_CP014203.1	861	207	400.8	3.69207E-89
NZ_CP014223.1	1494	392	643.6	5.6109E-142
NZ_CP014229.1	8233	1732	4241.0	0
NZ_CP014766.1	4714	842	2699.2	0
NZ_CP015401.2	2151	483	1055.6	1.4724E-231
NZ_CP015402.2	5328	1080	2815.9	0
NZ_CP016757.1	5588	1378	2543.7	0
NZ_CP017269.1	1724	361	890.7	1.0545E-195
NZ_CP019870.1	1156	256	574.3	6.4767E-127
NZ_CP019962.1	3105	645	1613.3	0
NZ_CP021434.1	4141	854	2163.0	0
NZ_CP021780.1	3630	524	2322.2	0
NZ_CP021850.1	2378	498	1228.3	4.3332E-269
NZ_CP021904.1	1920	395	1004.9	1.5741E-220
NZ_CP022121.1	2598	658	1155.3	3.1383E-253
NZ_CP022379.1	2894	518	1655.4	0
NZ_CP022387.1	1429	306	726.3	5.8383E-160
NZ_CP022413.2	1555	309	832.4	4.7922E-183

NZ_CP022464.2	3268	535	1963.8	0
NZ_CP023777.1	2390	360	1498.9	0
NZ_CP023863.1	1944	309	1185.6	8.4098E-260
NZ_CP024699.1	892	181	470.2	2.9459E-104
NZ_CP024727.1	2762	507	1554.8	0
NZ_CP025199.1	5247	1065	2770.7	0
NZ_CP025286.1	5362	1099	2813.2	0
NZ_CP027228.1	1953	447	945.9	1.0356E-207
NZ_CP027231.1	3474	726	1798.2	0
NZ_CP027234.1	3807	687	2166.4	0
NZ_CP028341.1	5881	1182	3126.9	0
NZ_CP028842.1	988	224	481.4	1.0578E-106
NZ_CP029186.1	2746	462	1626.0	0
NZ_CP029256.1	2778	611	1385.3	3.312E-303
NZ_CP029462.1	4398	1469	1461.8	0
NZ_CP029487.1	3165	657	1644.9	0
NZ_CP030280.1	1907	481	851.6	3.3303E-187
NZ_CP030775.1	845	205	389.7	9.78225E-87
NZ_CP030777.1	5920	1399	2792.8	0
NZ_CP032157.1	3946	442	2798.3	0
NZ_CP032364.1	1165	271	557.4	3.0442E-123
NZ_CP032382.1	4457	507	3143.2	0
NZ_CP032819.1	2017	357	1160.0	2.9981E-254
NZ_CP034413.2	6281	1067	3699.5	0
NZ_CP036345.1	2472	501	1306.6	4.2392E-286
NZ_CP036523.1	1329	285	676.2	4.5381E-149
NZ_CP039126.1	2509	420	1490.6	0
NZ_CP039457.1	1366	406	520.3	3.6784E-115
NZ_CP040058.1	2088	426	1099.8	3.7427E-241
NZ_CP040529.1	1599	403	714.8	1.804E-157
NZ_CP040882.1	6604	1801	2743.9	0
NZ_CP040924.1	1479	395	627.4	1.7985E-138
NZ_CP044060.1	8809	3272	2538.1	0
NZ_CP044098.1	5081	2865	618.5	1.5824E-136
NZ_CP046996.1	2317	726	831.6	7.3626E-183
NZ_CP048222.1	2549	314	1745.8	0
NZ_CP048436.1	8789	1515	5135.9	0
NZ_CP054012.1	2847	453	1736.6	0
NZ_CP061003.1	5056	1305	2211.5	0
NZ_CP068170.1	860	175	454.1	9.4161E-101
NZ_HF545616.1	1991	427	1010.7	8.4789E-222
NZ_HG917868.1	861	201	409.7	4.22989E-91
NZ_LN831027.1	934	133	602.4	5.1537E-133
NZ_LN877293.1	2141	403	1187.0	4.1789E-260
NZ_LN879430.1	1580	411	686.8	2.2453E-151
NZ_LR027382.1	6781	1177	3946.2	0

NZ_LR027880.1	1684	329	912.6	1.7896E-200
NZ_LR130778.1	1449	318	723.5	2.2725E-159
NZ_LR134336.1	1984	704	609.8	1.2649E-134
NZ_LR134338.1	2362	544	1137.6	2.2113E-249
NZ_LR134375.1	1490	534	451.9	2.7353E-100
NZ_LR134379.1	6438	1105	3770.6	0
NZ_LR134506.1	2554	951	732.5	2.5463E-161
NZ_LR134512.1	1076	462	244.5	4.14334E-55
NZ_LR215980.1	2765	636	1333.3	6.5315E-292
NZ_LR590484.1	2108	392	1178.4	2.9986E-258
NZ_LR699004.1	2003	456	973.6	9.6932E-214
NZ_LR699011.1	2235	482	1131.2	5.3495E-248
NZ_LR778174.1	1444	481	481.0	1.3022E-106
NZ_LS483306.1	1478	362	676.5	3.917E-149
NZ_LS483377.1	8137	2001	3713.7	0
NZ_LS483447.1	2766	777	1116.5	8.5335E-245
NZ_LT605205.1	2665	507	1467.8	0
NZ_LT608328.1	3113	684	1554.3	0
NZ_LT632322.1	2714	632	1295.3	1.1768E-283
NZ_LT635455.1	11487	2648	5526.7	0
NZ_LT635479.1	3023	732	1396.8	1.0401E-305
NZ_LT906459.1	2202	419	1212.7	1.0599E-265
NZ_LT990039.1	3065	850	1252.7	2.2019E-274

Table 7.2: Results of goodness-of-fit test (chi-squared test with one dimensional contingency table) to assess the significance of imbalance between essential and non-essential genes. 172 out of the 174 bacterial genomes showed a significant enrichment in miRNA targeting events (interactions) towards essential genes implying a functional role of the predicted interactions (e.g., control of bacterial growth-rate by host miRNAs).

Furthermore, we calculated the total number of interactions in each of the bacterial genes. A gene may participate in multiple interaction pairs because (i) multiple different host miRNAs target the same gene, (ii) the same host miRNA targets the gene in different regions and/or (iii) the gene is conserved across bacterial species and is part of different genomes. *Figure 7.18* presents the top bacterial genes in terms of the number of interactions they participate in.

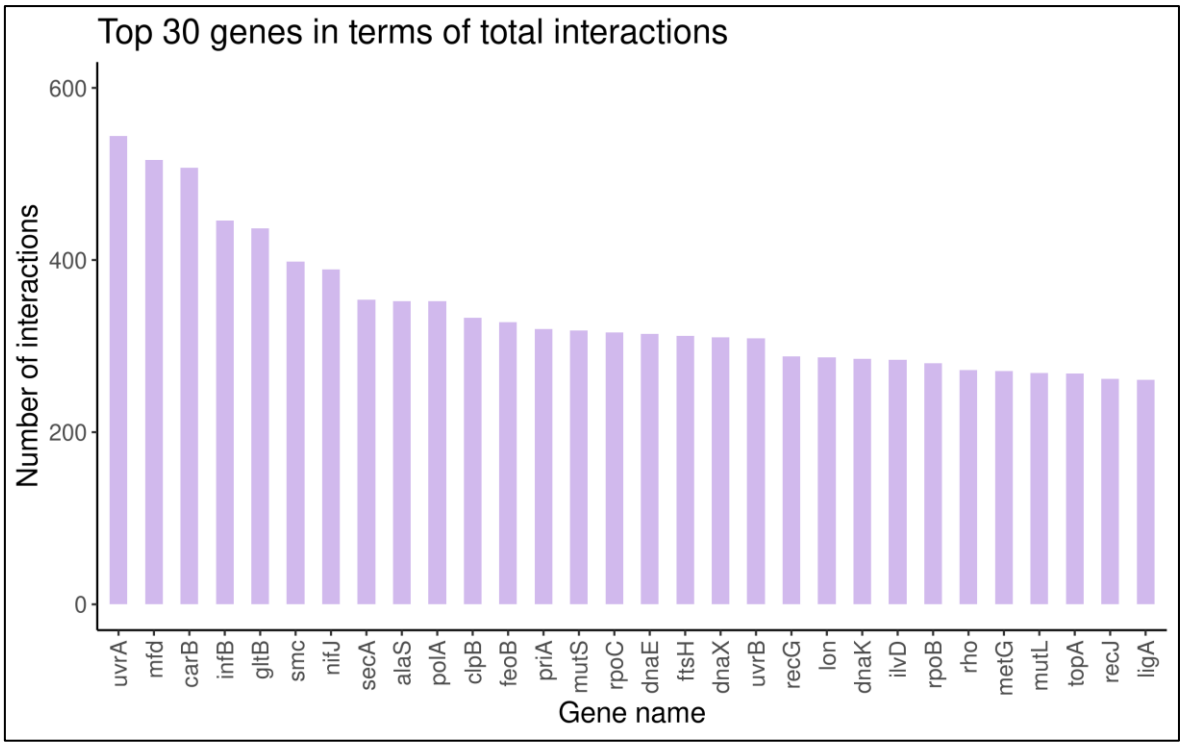

Figure 7.18: Top 30 bacterial genes in terms of total number of interactions (Figure was created for the purpose of this thesis)

Interestingly, most of the top genes participate in fundamental molecular and cellular processes of bacteria such as DNA repair, iron metabolism, DNA transcription, RNA metabolism, Glutamate synthesis and protein export.

Finally, we chose three human miRNA-bacterial gene interactions and visualized the RNA duplex they form upon interaction (*Figure 7.19*).

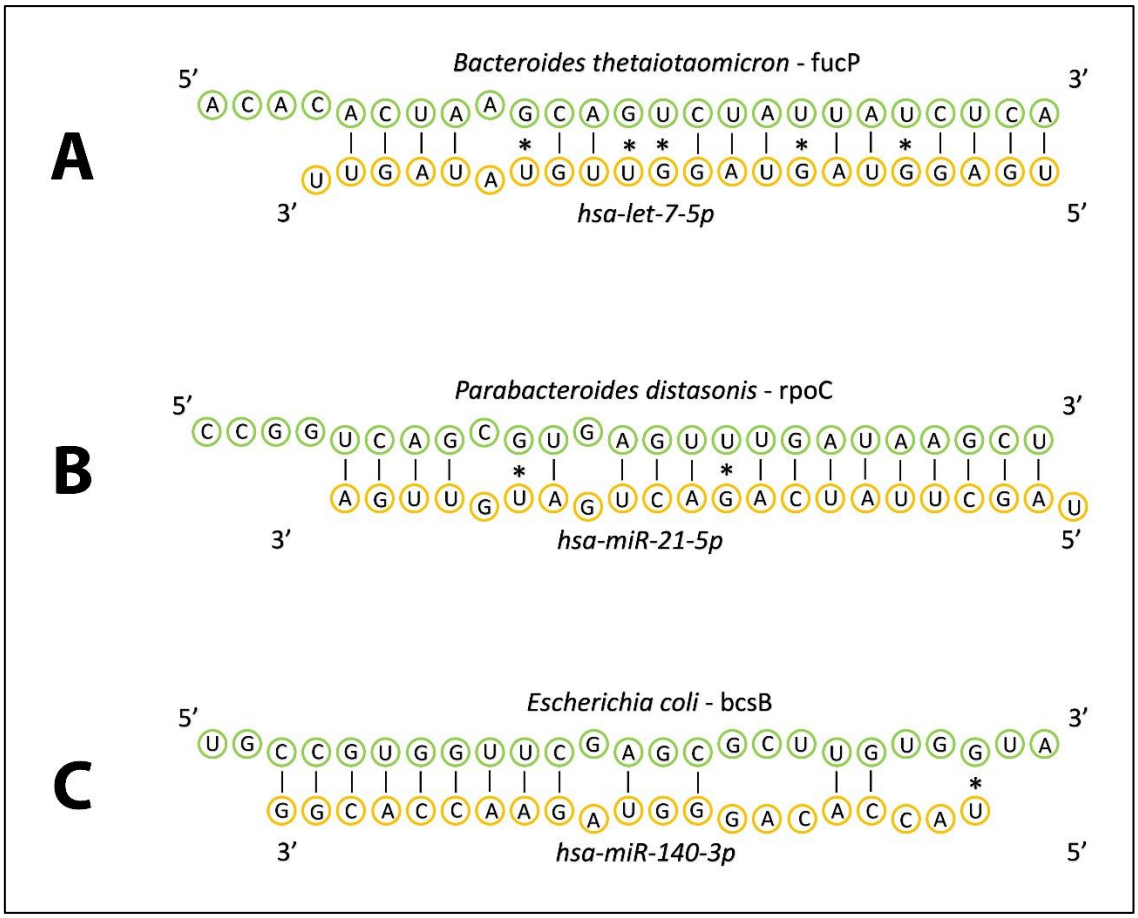

Figure 7.19: Examples of human miRNA-bacterial gene interactions (Figure was created for the purpose of this thesis)

CHAPTER 8 – Discussion and conclusions

The decreased cost and increased availability of metagenomic and metatranscriptomic sequencing experiments have revealed the importance of the human microbiome and its role in shaping health and disease. The accurate identification and quantification of microbial abundances in such experiments are the first crucial steps in the *in-silico* analysis of microbial communities. The study of host-microbiome interactions at the molecular level constitutes a major challenge. On the other hand, (non-coding) RNA research has revealed that the abundance and function of RNAs is implicated in homeostatic imbalance, disease occurrence and progression. During this PhD dissertation, effort has been placed on (i) expanding the landscape of RNA-RNA interactions in both eukaryotes and prokaryotes, (ii) performing quantification of microbial abundances in Shotgun metagenomics datasets, (iii) implementing a database linking bacteria with human pathologies and (iv) conducting single-cell and bulk-level transcriptomic studies with a focus on the tumor microenvironment and cancer therapeutic interventions.

(i) A variety of RNAs participate in direct RNA–RNA interactions with regulatory potential in both prokaryotes and eukaryotes. These interactions are sometimes interconnected and create complex cellular networks. The implementation of biomedical databases and computational methods to identify such interactions is invaluable. ***DIANA-TarBase v8*** (http://www.microrna.gr/tarbase) is a reference database comprising more than a million experimentally supported microRNA (miRNA) targets. ***DIANA-LncBase v3.0*** (www.microrna.gr/LncBase) is a reference repository with experimentally supported miRNA targets on non-coding transcripts. Its third version provides approximately 500,000 entries, corresponding to ~240 000 unique tissue and cell type specific miRNA–lncRNA pairs. ***Agnodice*** comprises a curated collection of bacterial sRNA-RNA interactions. Its first version comprises ~22,000 entries which are annotated in strain-level resolution and pertain to ~390 small RNAs (sRNAs) and ~6,630 target RNAs identified in 78 bacterial strains. The database content is exclusively experimentally supported, incorporating interactions derived via low yield as well as state-of-the-art high-throughput methods. ***Agnodice*** offers a variety of functionalities aiming to enhance user experience and enable insightful utilization. Finally, a computational method for the discovery of potential host microRNA-bacterial gene interactions have been implemented and applied to produce for the first time, thousands of potential miRNA-gene interactions and computationally explore them.

(ii) During this dissertation, we developed ***AGAMEMNON***, an *in-silico* framework for the analysis of metagenomic/metatranscriptomic samples providing highly accurate microbial abundance estimates at genus, species, and even strain resolution. Its novel indexing scheme

and analysis engine enables us to go beyond genus- and species-level analyses with the provision of microbial abundance estimates, while bypassing the vast memory requirements of similar alignment-based quantification approaches. ***AGAMEMNON*** can index the whole human microbiome or even the complete NCBI compendium using CPU/RAM specifications available to most labs. Importantly, the employed iterative, mass-preserving filtering tackles effectively the very common problem of false positive counts in metagenomic analyses. This series of innovations enable ***AGAMEMNON*** to perform hypothesis-free quantification of diverse samples without requiring the creation of custom-tailored indexes, while exhibiting higher or equally good accuracy between all the state-of-the-art methods that were tested. Importantly, on top of ***AGAMEMNON's*** quantification results, an R-Shiny application offers numerous downstream analyses modules that will push the envelope further, enabling users to explore and visualize microbial abundances but also conduct differential abundance and diversity index analyses through a user-friendly graphical interface.

(iii) Over the last decade, the role of the microbiome in human pathology is under active investigation. We developed ***Peryton*** (https://dianalab.e-ce.uth.gr/peryton/), to provide to the community a database of experimentally supported microbe-disease associations. Its first version constitutes a novel resource hosting more than 7900 entries linking 43 diseases with 1396 microorganisms. Several functionalities are provided to enhance user experience and enable ingenious use of ***Peryton***.

(iv) During the present dissertation, focus has also been placed on the study of protein-coding genes, their expression in tissues and cell-types and their role in a plethora of disease phenotypes. To this end, effort has been made for the analyses of one transcriptomic dataset for the study of the tumor microenvironment using single-cell RNA Sequencing, one transcriptomic study using bulk RNA Sequencing to investigate and compare FLASH radiation *versus* Standard radiation for the treatment of solid tumors, one transcriptomic study using dual RNA Sequencing to investigate Host-Leishmania interactions, one single-cell RNA Sequencing study to investigate the effects of FLASH radiation *versus* Standard radiation in the healthy full-thickness skin of treated mice and one single-cell RNA Sequencing study to investigate the effects of FLASH radiation *versus* Standard radiation in healthy tissue from the small intestine of irradiated mice.

CHAPTER 9 – Publications

1. AGAMEMNON: an Accurate metaGenomics And MEtatranscriptoMics quaNtificatiON analysis suite

Giorgos Skoufos, Fatemeh Almodaresi, Mohsen Zakeri, Joseph N. Paulson, Rob Patro, Artemis G. Hatzigeorgiou & Ioannis S. Vlachos

Genome biology (IF: 17.91) - https://doi.org/10.1186/s13059-022-02610-4

2. Peryton: a manual collection of experimentally supported microbe-disease associations

Giorgos Skoufos, Filippos S Kardaras, Athanasios Alexiou, Ioannis Kavakiotis, Anastasia Lambropoulou, Vasiliki Kotsira, Spyros Tastsoglou, Artemis G Hatzigeorgiou

Nucleic Acids Research (IF: 19.16) - https://doi.org/10.1093/nar/gkaa902

3. Transcriptional Profiling of Leishmania infantum Infected Dendritic Cells: Insights into the Role of Immunometabolism in Host-Parasite Interaction

Maritsa Margaroni, Maria Agallou, Athina Vasilakaki, Dimitra Karagkouni, **Giorgos Skoufos**, Artemis G. Hatzigeorgiou, Evdokia Karagouni

MDPI microorganisms (IF: 4.92) - https://doi.org/10.3390/microorganisms10071271

4. A stromal Integrated Stress Response activates perivascular cancer-associated fibroblasts to drive angiogenesis and tumour progression

Ioannis I Verginadis, Harris Avgousti, James Monslow, **Giorgos Skoufos**, Frank Chinga, Kyle Kim, Nektaria Maria Leli, Ilias V Karagounis, Brett I Bell, Anastasia Velalopoulou, Carlo Salas Salinas, Victoria S Wu, Yang Li, Jiangbin Ye, David A Scott, Andrei L Osterman, Arjun Sengupta, Aalim Weljie, Menggui Huang, Duo Zhang, Yi Fan, Enrico Radaelli, John W Tobias, Florian Rambow, Panagiotis Karras, Jean-Christophe Marine, Xiaowei Xu, Artemis G Hatzigeorgiou, Sandra Ryeom, J Alan Diehl, Serge Y Fuchs, Ellen Puré, Constantinos Koumenis

Nature Cell Biology (IF: 28.21) - https://doi.org/10.1038/s41556-022-00918-8

5. FLASH Proton Radiotherapy Spares Normal Epithelial and Mesenchymal Tissues While Preserving Sarcoma Response

Anastasia Velalopoulou, Ilias V Karagounis, Gwendolyn M Cramer, Michele M Kim, **Giorgos Skoufos**, Denisa Goia, Sarah Hagan, Ioannis I Verginadis, Khayrullo Shoniyozov, June Chiango, Michelle Cerullo, Kelley Varner, Lutian Yao, Ling Qin, Artemis G Hatzigeorgiou, Andy J Minn, Mary Putt, Matthew Lanza, Charles-Antoine Assenmacher, Enrico Radaelli, Jennifer Huck, Eric Diffenderfer, Lei Dong, James Metz, Constantinos Koumenis, Keith A Cengel, Amit Maity, Theresa M Busch

Cancer Research (IF: 12.07) - https://doi.org/10.1158/0008-5472.CAN-21-1500

6. PlasmiR: A Manual Collection of Circulating microRNAs of Prognostic and Diagnostic Value

Spyros Tastsoglou, Marios Miliotis, Ioannis Kavakiotis, Athanasios Alexiou, Eleni C Gkotsi, Anastasia Lambropoulou, Vasileios Lygnos, Vasiliki Kotsira, Vasileios Maroulis, Dimitrios Zisis, **Giorgos Skoufos**, Artemis G Hatzigeorgiou

MDPI cancers (IF: 6.57) - https://doi.org/10.3390/cancers13153680

7. DIANA-LncBase v3: indexing experimentally supported miRNA targets on non-coding transcripts

Dimitra Karagkouni, Maria D Paraskevopoulou, Spyros Tastsoglou, **Giorgos Skoufos**, Anna Karavangeli, Vasilis Pierros, Elissavet Zacharopoulou, Artemis G Hatzigeorgiou

Nucleic Acids Research (IF: 19.16) - https://doi.org/10.1093/nar/gkz1036

8. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions

Dimitra Karagkouni, Maria D Paraskevopoulou, Serafeim Chatzopoulos, Ioannis S Vlachos, Spyros Tastsoglou, Ilias Kanellos, Dimitris Papadimitriou, Ioannis Kavakiotis, Sofia Maniou, **Giorgos Skoufos**, Thanasis Vergoulis, Theodore Dalamagas, Artemis G Hatzigeorgiou

Nucleic Acids Research (IF: 19.16) - https://doi.org/10.1093/nar/gkx1141

CHAPTER 10 – Posters

1. Tastsoglou S, **Skoufos G.**, et al. “More than a million experimentally supported entries of microRNA interactions with coding and non-coding transcripts”, Greece, **3rd International Conference on the Long and the Short of Non-Coding RNAs, 2019**

2. **Skoufos G.**, et al “Time and space-efficient kmer-based microbial abundance quantification at strain resolution”, Greece, **HSCBB19 2019 Conference** (**best poster award**)

3. **Skoufos G.**, et al, “AGAMEMNON: an Accurate Metagenomic and Metatranscriptomic quantification analysis suite”, Greece, **European Conference on Computational Biology 2018, ECCB18**

4. Verginadis, I. I., Martinez, C. M., Lim, T., Chowdhury, P., **Skoufos**, **G.**, Kim, M. M., ... and Koumenis, C. (2022). “Single-cell RNA sequencing analysis and transgenic mouse models reveal differential effects of flash vs. standard proton radiotherapy on gastrointestinal tissues and tumors.”
Cancer Research, 82 (AACR 2022), 3320-3320.

5. Velalopoulou, A., Karagounis, I. V., **Skoufos, G.**, Verginadis, I. I., Kim, M., Shoniyozov, K., … and Busch, T. M. (2022). “Gene expression profiling of full-thickness skin after FLASH proton radiotherapy.”
Cancer Research, 82 (AACR 2022), 3304-3304.

6. Leli, N. M., Dey, S., Brady, L., Salinas, C. S., Lin, J., **Skoufos, G.**, ... and Koumenis, C. (2022). “Survivin, a novel mediator of the UPR identified by a functional genome wide CRISPR/Cas9 based knock out screen.”
Cancer Research, 82 (AACR 2022), 144-144.

References

[1] Y. Bai, X. Dai, A. Harrison, C. Johnston, and M. Chen, "Toward a next-generation atlas of RNA secondary structure," *Briefings in Bioinformatics,* vol. 17, no. 1, pp. 63-77, 2016, doi: 10.1093/bib/bbv026.

[2] S. E. Butcher and A. M. Pyle, "The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks," *Accounts of chemical research,* vol. 44, no. 12, pp. 1302-1311, 2011.

[3] J. Hurwitz, "The discovery of RNA polymerase," *Journal of Biological Chemistry,* vol. 280, no. 52, pp. 42477-42485, 2005.

[4] S. Ochoa, "Enzymatic synthesis of ribonucleic acid," in *Künstliche Radioaktive Isotope in Physiologie Diagnostik und Therapie/Radioactive Isotopes in Physiology Diagnostics and Therapy*: Springer, 1961, pp. 960-973.

[5] A. Rich and D. R. Davies, "A new two stranded helical structure: polyadenylic acid and polyuridylic acid," *Journal of the American Chemical Society,* vol. 78, no. 14, pp. 3548-3549, 1956.

[6] R. W. Holley *et al.*, "Structure of a ribonucleic acid," *Science,* vol. 147, no. 3664, pp. 1462-1465, 1965.

[7] W. Fiers *et al.*, "Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene," *Nature,* vol. 260, no. 5551, pp. 500-507, 1976.

[8] S. M. Berget, C. Moore, and P. A. Sharp, "Spliced segments at the 5′ terminus of adenovirus 2 late mRNA," *Proceedings of the National Academy of Sciences,* vol. 74, no. 8, pp. 3171-3175, 1977.

[9] S. J. Sharp, J. Schaack, L. Cooley, D. J. Burke, and D. Söll, "Structure and transcription of eukaryotic tRNA genes," (in eng), *CRC Crit Rev Biochem,* vol. 19, no. 2, pp. 107-44, 1985, doi: 10.3109/10409238509082541.

[10] B. Wightman, I. Ha, and G. Ruvkun, "Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans," (in eng), *Cell,* vol. 75, no. 5, pp. 855-62, Dec 3 1993, doi: 10.1016/0092-8674(93)90530-4.

[11] S. Brenner, F. Jacob, and M. Meselson, "An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis," *Nature,* vol. 190, no. 4776, pp. 576-581, 1961/05/01 1961, doi: 10.1038/190576a0.

[12] J. J. Furth, J. Hurwitz, and M. Anders, "The role of deoxyribonucleic acid in ribonucleic acid synthesis. I. The purification and properties of ribonucleic acid polymerase," (in eng), *J Biol Chem,* vol. 237, pp. 2611-9, Aug 1962.

[13] H. RC, "Enzymatic synthesis of RNA," *Biochemical and biophysical research communications,* vol. 3, pp. 689-694, 1960.

[14] A. Stevens, "Incorporation of the adenine ribonucleotide into RNA by cell fractions from E. coli B," *Biochemical and biophysical research communications,* vol. 3, no. 1, pp. 92-96, 1960.

[15] J. Archambault and J. D. Friesen, "Genetics of eukaryotic RNA polymerases I, II, and III," (in eng), *Microbiol Rev,* vol. 57, no. 3, pp. 703-24, Sep 1993, doi: 10.1128/mr.57.3.703-724.1993.

[16] D. Sweetser, M. Nonet, and R. A. Young, "Prokaryotic and eukaryotic RNA polymerases have homologous core subunits," (in eng), *Proc Natl Acad Sci U S A,* vol. 84, no. 5, pp. 1192-6, Mar 1987, doi: 10.1073/pnas.84.5.1192.

[17] A. I. Lamond, *Pre-mRNA processing*. Springer Science & Business Media, 2013.

[18] C. L. Will and R. Lührmann, "Spliceosome structure and function," (in eng), *Cold Spring Harb Perspect Biol,* vol. 3, no. 7, Jul 1 2011, doi: 10.1101/cshperspect.a003707.
[19] A. Shatkin, "Capping of eucaryotic mRNAs," *Cell,* vol. 9, no. 4, pp. 645-653, 1976.
[20] B. C. Stark, R. Kole, E. J. Bowman, and S. Altman, "Ribonuclease P: an enzyme with an essential RNA component," *Proceedings of the National Academy of Sciences,* vol. 75, no. 8, pp. 3717-3721, 1978.
[21] N. J. Proudfoot, A. Furger, and M. J. Dye, "Integrating mRNA processing with transcription," *Cell,* vol. 108, no. 4, pp. 501-512, 2002.
[22] X. Wu and G. Brewer, "The regulation of mRNA stability in mammalian cells: 2.0," (in eng), *Gene,* vol. 500, no. 1, pp. 10-21, May 25 2012, doi: 10.1016/j.gene.2012.03.021.
[23] K. Xiang and D. P. Bartel, "The molecular basis of coupling between poly (A)-tail length and translational efficiency," *Elife,* vol. 10, p. e66493, 2021.
[24] L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts, "An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA," *Cell,* vol. 12, no. 1, pp. 1-8, 1977.
[25] Y. Wang *et al.,* "Mechanism of alternative splicing and its regulation," (in eng), *Biomed Rep,* vol. 3, no. 2, pp. 152-158, Mar 2015, doi: 10.3892/br.2014.407.
[26] S. Schwartz, B. K. Felber, and G. N. Pavlakis, "Mechanism of translation of monocistronic and multicistronic human immunodeficiency virus type 1 mRNAs," (in eng), *Mol Cell Biol,* vol. 12, no. 1, pp. 207-19, Jan 1992, doi: 10.1128/mcb.12.1.207-219.1992.
[27] S. Hiraga and C. Yanofsky, "Hyper-labile messenger RNA in polar mutants of the tryptophan operon of Escherichia coli," (in eng), *J Mol Biol,* vol. 72, no. 1, pp. 103-10, Dec 14 1972, doi: 10.1016/0022-2836(72)90072-1.
[28] D. W. Selinger, R. M. Saxena, K. J. Cheung, G. M. Church, and C. Rosenow, "Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation," *Genome research,* vol. 13, no. 2, pp. 216-223, 2003.
[29] O. J. Plescia, N. C. Palczuk, E. Cora-Figueroa, A. Mukherjee, and W. Braun, "Production of antibodies to soluble RNA (sRNA)," (in eng), *Proc Natl Acad Sci U S A,* vol. 54, no. 4, pp. 1281-5, Oct 1965, doi: 10.1073/pnas.54.4.1281.
[30] A. G. Torres, "Enjoy the Silence: Nearly Half of Human tRNA Genes Are Silent," (in eng), *Bioinform Biol Insights,* vol. 13, p. 1177932219868454, 2019, doi: 10.1177/1177932219868454.
[31] J. M. Goodenbour and T. Pan, "Diversity of tRNA genes in eukaryotes," (in eng), *Nucleic Acids Res,* vol. 34, no. 21, pp. 6137-46, 2006, doi: 10.1093/nar/gkl725.
[32] H. Dong, L. Nilsson, and C. G. Kurland, "Co-variation of trna abundance and codon usage inescherichia coliat different growth rates," *Journal of molecular biology,* vol. 260, no. 5, pp. 649-663, 1996.
[33] M. Bulmer, "Coevolution of codon usage and transfer RNA abundance," *Nature,* vol. 325, no. 6106, pp. 728-730, 1987.
[34] E. Evguenieva-Hackenberg, "Bacterial ribosomal RNA in pieces," (in eng), *Mol Microbiol,* vol. 57, no. 2, pp. 318-25, Jul 2005, doi: 10.1111/j.1365-2958.2005.04662.x.
[35] A. K. Henras, C. Plisson-Chastang, M. F. O'Donohue, A. Chakraborty, and P. E. Gleizes, "An overview of pre-ribosomal RNA processing in eukaryotes," (in eng), *Wiley Interdiscip Rev RNA,* vol. 6, no. 2, pp. 225-42, Mar-Apr 2015, doi: 10.1002/wrna.1269.

[36] G. E. Fox, "Origin and evolution of the ribosome," (in eng), *Cold Spring Harb Perspect Biol,* vol. 2, no. 9, p. a003483, Sep 2010, doi: 10.1101/cshperspect.a003483.
[37] E. V. Bobkova, Y. P. Yan, D. B. Jordan, M. G. Kurilla, and D. L. Pompliano, "Catalytic properties of mutant 23 S ribosomes resistant to oxazolidinones," *Journal of Biological Chemistry,* vol. 278, no. 11, pp. 9802-9807, 2003.
[38] V. Ramakrishnan, "Ribosome structure and the mechanism of translation," *Cell,* vol. 108, no. 4, pp. 557-572, 2002.
[39] J. H. Davis, Y. Z. Tan, B. Carragher, C. S. Potter, D. Lyumkis, and J. R. Williamson, "Modular assembly of the bacterial large ribosomal subunit," *Cell,* vol. 167, no. 6, pp. 1610-1622. e15, 2016.
[40] Z. L. Watson *et al.*, "Structure of the bacterial ribosome at 2 Å resolution," (in eng), *Elife,* vol. 9, Sep 14 2020, doi: 10.7554/eLife.60482.
[41] D. J. Taylor *et al.*, "Comprehensive Molecular Structure of the Eukaryotic Ribosome," *Structure,* vol. 17, no. 12, pp. 1591-1604, 2009/12/09/ 2009, doi: https://doi.org/10.1016/j.str.2009.09.015.
[42] Y. Van de Peer and R. De Wachter, "Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA," *Journal of molecular evolution,* vol. 45, no. 6, pp. 619-630, 1997.
[43] G. Caetano-Anollés, "Tracing the evolution of RNA structure in ribosomes," (in eng), *Nucleic Acids Res,* vol. 30, no. 11, pp. 2575-87, Jun 1 2002, doi: 10.1093/nar/30.11.2575.
[44] A. Felske, A. Wolterink, R. Van Lis, and A. D. Akkermans, "Phylogeny of the main bacterial 16S rRNA sequences in Drentse A grassland soils (The Netherlands)," (in eng), *Appl Environ Microbiol,* vol. 64, no. 3, pp. 871-9, Mar 1998, doi: 10.1128/aem.64.3.871-879.1998.
[45] A. B. Idris *et al.*, "Molecular Phylogenetic Analysis of 16S rRNA Sequences Identified Two Lineages of Helicobacter pylori Strains Detected from Different Regions in Sudan Suggestive of Differential Evolution," (in eng), *Int J Microbiol,* vol. 2020, p. 8825718, 2020, doi: 10.1155/2020/8825718.
[46] S. Caburet, C. Conti, C. Schurra, R. Lebofsky, S. J. Edelstein, and A. Bensimon, "Human ribosomal RNA gene arrays display a broad range of palindromic structures," (in eng), *Genome Res,* vol. 15, no. 8, pp. 1079-85, Aug 2005, doi: 10.1101/gr.3970105.
[47] D. M. Stults, M. W. Killen, H. H. Pierce, and A. J. Pierce, "Genomic architecture and inheritance of human ribosomal RNA gene clusters," (in eng), *Genome Res,* vol. 18, no. 1, pp. 13-8, Jan 2008, doi: 10.1101/gr.6858507.
[48] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *cell,* vol. 116, no. 2, pp. 281-297, 2004.
[49] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *cell,* vol. 136, no. 2, pp. 215-233, 2009.
[50] N. Liu and E. N. Olson, "MicroRNA regulatory networks in cardiovascular development," *Developmental cell,* vol. 18, no. 4, pp. 510-525, 2010.
[51] N. K. Vo, X. A. Cambronne, and R. H. Goodman, "MicroRNA pathways in neural development and plasticity," *Current opinion in neurobiology,* vol. 20, no. 4, pp. 457-465, 2010.
[52] T. Takaya *et al.*, "MicroRNA-1 and MicroRNA-133 in spontaneous myocardial differentiation of mouse embryonic stem cells," *Circulation Journal,* vol. 73, no. 8, pp. 1492-1497, 2009.

[53] J. Ren, P. Jin, E. Wang, F. M. Marincola, and D. F. Stroncek, "MicroRNA and gene expression patterns in the differentiation of human embryonic stem cells," *Journal of translational medicine,* vol. 7, no. 1, pp. 1-17, 2009.
[54] V. Rottiers and A. M. Näär, "MicroRNAs in metabolism and metabolic disorders," (in eng), *Nat Rev Mol Cell Biol,* vol. 13, no. 4, pp. 239-50, Mar 22 2012, doi: 10.1038/nrm3313.
[55] L. Shi *et al.*, "MicroRNA-125b-2 confers human glioblastoma stem cells resistance to temozolomide through the mitochondrial pathway of apoptosis," *International journal of oncology,* vol. 40, no. 1, pp. 119-129, 2012.
[56] S. Maciotta, M. Meregalli, and Y. Torrente, "The involvement of microRNAs in neurodegenerative diseases," *Frontiers in cellular neuroscience,* vol. 7, p. 265, 2013.
[57] N. Wu *et al.*, "Role of microRNA-26b in glioma development and its mediated regulation on EphA2," *PloS one,* vol. 6, no. 1, p. e16264, 2011.
[58] Y. Peng and C. M. Croce, "The role of MicroRNAs in human cancer," *Signal transduction and targeted therapy,* vol. 1, no. 1, pp. 1-9, 2016.
[59] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14," (in eng), *Cell,* vol. 75, no. 5, pp. 843-54, Dec 3 1993, doi: 10.1016/0092-8674(93)90529-y.
[60] S. Vasudevan, Y. Tong, and J. A. Steitz, "Switching from repression to activation: microRNAs can up-regulate translation," *Science,* vol. 318, no. 5858, pp. 1931-1934, 2007.
[61] H. W. Hwang, E. A. Wentzel, and J. T. Mendell, "A hexanucleotide element directs microRNA nuclear import," (in eng), *Science,* vol. 315, no. 5808, pp. 97-100, Jan 5 2007, doi: 10.1126/science.1136235.
[62] R. Tang *et al.*, "Mouse miRNA-709 directly regulates miRNA-15a/16-1 biogenesis at the posttranscriptional level in the nucleus: evidence for a microRNA hierarchy system," *Cell research,* vol. 22, no. 3, pp. 504-515, 2012.
[63] A. Turchinovich, L. Weiz, A. Langheinz, and B. Burwinkel, "Characterization of extracellular circulating microRNA," (in eng), *Nucleic Acids Res,* vol. 39, no. 16, pp. 7223-33, Sep 1 2011, doi: 10.1093/nar/gkr254.
[64] S. Liu *et al.*, "The host shapes the gut microbiota via fecal microRNA," *Cell host & microbe,* vol. 19, no. 1, pp. 32-43, 2016.
[65] S. Das *et al.*, "Nuclear miRNA regulates the mitochondrial genome in the heart," *Circulation research,* vol. 110, no. 12, pp. 1596-1603, 2012.
[66] D. Lauressergues *et al.*, "Primary transcripts of microRNAs encode regulatory peptides," *Nature*, vol. 520, no. 7545, pp. 90-93, 2015/04/01 2015, doi: 10.1038/nature14346.
[67] R. Bayraktar, K. Van Roosbroeck, and G. A. Calin, "Cell-to-cell communication: microRNAs as hormones," (in eng), *Mol Oncol,* vol. 11, no. 12, pp. 1673-1686, Dec 2017, doi: 10.1002/1878-0261.12144.
[68] S. Shin *et al.*, "Urinary exosome microRNA signatures as a noninvasive prognostic biomarker for prostate cancer," *npj Genomic Medicine,* vol. 6, no. 1, p. 45, 2021/06/11 2021, doi: 10.1038/s41525-021-00212-w.
[69] L. Wang and L. Zhang, "Circulating Exosomal miRNA as Diagnostic Biomarkers of Neurodegenerative Diseases," (in eng), *Front Mol Neurosci,* vol. 13, p. 53, 2020, doi: 10.3389/fnmol.2020.00053.
[70] A. M. Denli, B. B. Tops, R. H. Plasterk, R. F. Ketting, and G. J. Hannon, "Processing of primary microRNAs by the Microprocessor complex," *Nature,* vol. 432, no. 7014, pp. 231-235, 2004.

[71] C. R. Alarcón, H. Lee, H. Goodarzi, N. Halberg, and S. F. Tavazoie, "N6-methyladenosine marks primary microRNAs for processing," (in eng), *Nature,* vol. 519, no. 7544, pp. 482-5, Mar 26 2015, doi: 10.1038/nature14281.
[72] C. Okada *et al.*, "A high-resolution structure of the pre-microRNA nuclear export machinery," *Science,* vol. 326, no. 5957, pp. 1275-1279, 2009.
[73] B. R. Cullen, "Transcription and processing of human microRNA precursors," *Molecular cell,* vol. 16, no. 6, pp. 861-865, 2004.
[74] C. L. Noland and J. A. Doudna, "Multiple sensors ensure guide strand selection in human RNAi pathways," *Rna,* vol. 19, no. 5, pp. 639-648, 2013.
[75] A. van den Berg, J. Mols, and J. Han, "RISC-target interaction: cleavage and translational suppression," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms,* vol. 1779, no. 11, pp. 668-677, 2008.
[76] J. Liu *et al.*, "Argonaute2 is the catalytic engine of mammalian RNAi," *Science,* vol. 305, no. 5689, pp. 1437-1441, 2004.
[77] E. Berezikov, W.-J. Chung, J. Willis, E. Cuppen, and E. C. Lai, "Mammalian mirtron genes," *Molecular cell,* vol. 28, no. 2, pp. 328-336, 2007.
[78] S. Cheloufi, C. O. Dos Santos, M. M. Chong, and G. J. Hannon, "A dicer-independent miRNA biogenesis pathway that requires Ago catalysis," *Nature,* vol. 465, no. 7298, pp. 584-589, 2010.
[79] M. Xie *et al.*, "Mammalian 5′-capped microRNA precursors that generate a single microRNA," *Cell,* vol. 155, no. 7, pp. 1568-1580, 2013.
[80] S. Diederichs and D. A. Haber, "Dual role for argonautes in microRNA processing and posttranscriptional regulation of microRNA expression," *Cell,* vol. 131, no. 6, pp. 1097-1108, 2007.
[81] A. Eulalio, E. Huntzinger, T. Nishihara, J. Rehwinkel, M. Fauser, and E. Izaurralde, "Deadenylation is a widespread effect of miRNA regulation," (in eng), *Rna,* vol. 15, no. 1, pp. 21-32, Jan 2009, doi: 10.1261/rna.1399509.
[82] W. H. Majoros and U. Ohler, "Spatial preferences of microRNA targets in 3' untranslated regions," *BMC Genomics,* vol. 8, no. 1, p. 152, 2007/06/07 2007, doi: 10.1186/1471-2164-8-152.
[83] J. Hausser, A. P. Syed, B. Bilen, and M. Zavolan, "Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation," (in eng), *Genome Res,* vol. 23, no. 4, pp. 604-15, Apr 2013, doi: 10.1101/gr.139758.112.
[84] M. Hafner *et al.*, "Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP," (in eng), *Cell,* vol. 141, no. 1, pp. 129-41, Apr 2 2010, doi: 10.1016/j.cell.2010.03.009.
[85] A. Helwak, G. Kudla, T. Dudnakova, and D. Tollervey, "Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding," (in eng), *Cell,* vol. 153, no. 3, pp. 654-65, Apr 25 2013, doi: 10.1016/j.cell.2013.03.043.
[86] M. D. Paraskevopoulou *et al.*, "DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows," (in eng), *Nucleic Acids Res,* vol. 41, no. Web Server issue, pp. W169-73, Jul 2013, doi: 10.1093/nar/gkt393.
[87] M. D. Paraskevopoulou, D. Karagkouni, I. S. Vlachos, S. Tastsoglou, and A. G. Hatzigeorgiou, "microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions," *Nature Communications,* vol. 9, no. 1, p. 3601, 2018/09/06 2018, doi: 10.1038/s41467-018-06046-y.
[88] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel, "Predicting effective microRNA target sites in mammalian mRNAs," *eLife,* vol. 4, p. e05005, 2015/08/12 2015, doi: 10.7554/eLife.05005.

[89] X. Wang, "Composition of seed sequence is a major determinant of microRNA targeting patterns," *Bioinformatics,* vol. 30, no. 10, pp. 1377-1383, 2014, doi: 10.1093/bioinformatics/btu045.
[90] R. Lorenz *et al.*, "ViennaRNA Package 2.0," *Algorithms for Molecular Biology,* vol. 6, no. 1, p. 26, 2011/11/24 2011, doi: 10.1186/1748-7188-6-26.
[91] R. M. Marín and J. Vaníček, "Efficient use of accessibility in microRNA target prediction," *Nucleic Acids Research,* vol. 39, no. 1, pp. 19-29, 2010, doi: 10.1093/nar/gkq768.
[92] M. Ha, M. Pang, V. Agarwal, and Z. J. Chen, "Interspecies regulation of microRNAs and their targets," (in eng), *Biochim Biophys Acta,* vol. 1779, no. 11, pp. 735-42, Nov 2008, doi: 10.1016/j.bbagrm.2008.03.004.
[93] H. Valadi, K. Ekström, A. Bossios, M. Sjöstrand, J. J. Lee, and J. O. Lötvall, "Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells," *Nature Cell Biology,* vol. 9, no. 6, pp. 654-659, 2007/06/01 2007, doi: 10.1038/ncb1596.
[94] J. A. Weber *et al.*, "The microRNA spectrum in 12 body fluids," *Clinical chemistry,* vol. 56, no. 11, pp. 1733-1741, 2010.
[95] J. Ratajczak, M. Wysoczynski, F. Hayek, A. Janowska-Wieczorek, and M. Ratajczak, "Membrane-derived microvesicles: important and underappreciated mediators of cell-to-cell communication," *Leukemia,* vol. 20, no. 9, pp. 1487-1495, 2006.
[96] C. Théry, L. Zitvogel, and S. Amigorena, "Exosomes: composition, biogenesis and function," *Nature reviews immunology,* vol. 2, no. 8, pp. 569-579, 2002.
[97] A. Turchinovich, L. Weiz, A. Langheinz, and B. Burwinkel, "Characterization of extracellular circulating microRNA," *Nucleic acids research,* vol. 39, no. 16, pp. 7223-7233, 2011.
[98] J. D. Arroyo *et al.*, "Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma," *Proceedings of the National Academy of Sciences,* vol. 108, no. 12, pp. 5003-5008, 2011.
[99] K. Wang, S. Zhang, J. Weber, D. Baxter, and D. J. Galas, "Export of microRNAs and microRNA-protective protein by mammalian cells," *Nucleic acids research,* vol. 38, no. 20, pp. 7248-7259, 2010.
[100] R. Garcia-Martin *et al.*, "MicroRNA sequence codes for small extracellular vesicle release and cellular retention," *Nature,* vol. 601, no. 7893, pp. 446-451, 2022/01/01 2022, doi: 10.1038/s41586-021-04234-3.
[101] Y. Zhang *et al.*, "Secreted monocytic miR-150 enhances targeted endothelial cell migration," *Molecular cell,* vol. 39, no. 1, pp. 133-144, 2010.
[102] S. Liu *et al.*, "Oral Administration of miR-30d from Feces of MS Patients Suppresses MS-like Symptoms in Mice by Expanding Akkermansia muciniphila," (in eng), *Cell Host Microbe,* vol. 26, no. 6, pp. 779-794.e8, Dec 11 2019, doi: 10.1016/j.chom.2019.10.008.
[103] S. Liu *et al.*, "The Host Shapes the Gut Microbiota via Fecal MicroRNA," (in eng), *Cell Host Microbe,* vol. 19, no. 1, pp. 32-43, Jan 13 2016, doi: 10.1016/j.chom.2015.12.005.
[104] E. Ueta *et al.*, "Extracellular vesicle-shuttled miRNAs as a diagnostic and prognostic biomarker and their potential roles in gallbladder cancer patients," *Scientific Reports,* vol. 11, no. 1, p. 12298, 2021/06/10 2021, doi: 10.1038/s41598-021-91804-0.
[105] D. Povero *et al.*, "Protein and miRNA profile of circulating extracellular vesicles in patients with primary sclerosing cholangitis," *Scientific Reports,* vol. 12, no. 1, p. 3027, 2022/02/22 2022, doi: 10.1038/s41598-022-06809-0.

[106] S. Tastsoglou *et al.*, "PlasmiR: A Manual Collection of Circulating microRNAs of Prognostic and Diagnostic Value," *Cancers*, vol. 13, no. 15, p. 3680, 2021. [Online]. Available: https://www.mdpi.com/2072-6694/13/15/3680.
[107] J. R. Chevillet *et al.*, "Quantitative and stoichiometric analysis of the microRNA content of exosomes," *Proceedings of the National Academy of Sciences*, vol. 111, no. 41, pp. 14888-14893, 2014, doi: doi:10.1073/pnas.1408301111.
[108] D. Karagkouni *et al.*, "DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions," (in eng), *Nucleic Acids Res*, vol. 46, no. D1, pp. D239-d245, Jan 4 2018, doi: 10.1093/nar/gkx1141.
[109] D. Karagkouni *et al.*, "DIANA-LncBase v3: indexing experimentally supported miRNA targets on non-coding transcripts," *Nucleic Acids Research*, vol. 48, no. D1, pp. D101-D110, 2019, doi: 10.1093/nar/gkz1036.
[110] I. S. Vlachos *et al.*, "DIANA-miRPath v3.0: deciphering microRNA function with experimental support," (in eng), *Nucleic Acids Res*, vol. 43, no. W1, pp. W460-6, Jul 1 2015, doi: 10.1093/nar/gkv403.
[111] V. Olive, I. Jiang, and L. He, "mir-17-92, a cluster of miRNAs in the midst of the cancer network," *The international journal of biochemistry & cell biology*, vol. 42, no. 8, pp. 1348-1354, 2010.
[112] C. E. Condrat *et al.*, "miRNAs as Biomarkers in Disease: Latest Findings Regarding Their Role in Diagnosis and Prognosis," (in eng), *Cells*, vol. 9, no. 2, Jan 23 2020, doi: 10.3390/cells9020276.
[113] S. Naidu, P. Magee, and M. Garofalo, "MiRNA-based therapeutic intervention of cancer," (in eng), *J Hematol Oncol*, vol. 8, p. 68, Jun 11 2015, doi: 10.1186/s13045-015-0162-0.
[114] K. C. Vickers, B. T. Palmisano, B. M. Shoucri, R. D. Shamburek, and A. T. Remaley, "MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins," *Nature cell biology*, vol. 13, no. 4, pp. 423-433, 2011.
[115] A. M. Eiring *et al.*, "miR-328 functions as an RNA decoy to modulate hnRNP E2 regulation of mRNA translation in leukemic blasts," *Cell*, vol. 140, no. 5, pp. 652-665, 2010.
[116] S. L. Ameres *et al.*, "Target RNA–directed trimming and tailing of small silencing RNAs," *Science*, vol. 328, no. 5985, pp. 1534-1539, 2010.
[117] C. Y. Shi, E. R. Kingston, B. Kleaveland, D. H. Lin, M. W. Stubna, and D. P. Bartel, "The ZSWIM8 ubiquitin ligase mediates target-directed microRNA degradation," (in eng), *Science*, vol. 370, no. 6523, Dec 18 2020, doi: 10.1126/science.abc9359.
[118] J. Han, C. A. LaVigne, B. T. Jones, H. Zhang, F. Gillett, and J. T. Mendell, "A ubiquitin ligase mediates target-directed microRNA decay independently of tailing and trimming," (in eng), *Science*, vol. 370, no. 6523, Dec 18 2020, doi: 10.1126/science.abc9546.
[119] T. R. Mercer *et al.*, "The human mitochondrial transcriptome," *Cell*, vol. 146, no. 4, pp. 645-658, 2011.
[120] S. Das *et al.*, "Nuclear miRNA regulates the mitochondrial genome in the heart," (in eng), *Circ Res*, vol. 110, no. 12, pp. 1596-603, Jun 8 2012, doi: 10.1161/circresaha.112.267732.
[121] A. Paramasivam and J. Vijayashree Priyadharsini, "MitomiRs: new emerging microRNAs in mitochondrial dysfunction and cardiovascular disease," *Hypertension Research*, vol. 43, no. 8, pp. 851-853, 2020/08/01 2020, doi: 10.1038/s41440-020-0423-3.

[122] A. Rencelj, N. Gvozdenovic, and M. Cemazar, "MitomiRs: their roles in mitochondria and importance in cancer cell metabolism," (in eng), *Radiol Oncol,* vol. 55, no. 4, pp. 379-392, Nov 19 2021, doi: 10.2478/raon-2021-0042.
[123] S. Shahid *et al.*, "MicroRNAs from the parasitic plant Cuscuta campestris target host messenger RNAs," *Nature,* vol. 553, no. 7686, pp. 82-85, 2018/01/01 2018, doi: 10.1038/nature25027.
[124] L. Yang *et al.*, "Overexpression of potato miR482e enhanced plant sensitivity to Verticillium dahliae infection," *Journal of integrative plant biology,* vol. 57, no. 12, pp. 1078-1088, 2015.
[125] S. Ouyang *et al.*, "MicroRNAs suppress NB domain genes in tomato that confer resistance to Fusarium oxysporum," *PLoS Pathogens,* vol. 10, no. 10, p. e1004464, 2014.
[126] W. Li, C. He, J. Wu, D. Yang, and W. Yi, "Epstein barr virus encodes miRNAs to assist host immune escape," (in eng), *J Cancer,* vol. 11, no. 8, pp. 2091-2100, 2020, doi: 10.7150/jca.42498.
[127] Y. Fu *et al.*, "Enterovirus 71 suppresses miR-17-92 cluster through up-regulating methylation of the miRNA promoter," *Frontiers in microbiology,* vol. 10, p. 625, 2019.
[128] L. Ma *et al.*, "LncBook: a curated knowledgebase of human long non-coding RNAs," (in eng), *Nucleic Acids Res,* vol. 47, no. D1, pp. D128-d134, Jan 8 2019, doi: 10.1093/nar/gky960.
[129] J. J. Quinn and H. Y. Chang, "Unique features of long non-coding RNA biogenesis and function," *Nature Reviews Genetics,* vol. 17, no. 1, pp. 47-62, 2016.
[130] C. Ziegler and M. Kretz, "The more the merrier—complexity in long non-coding rna loci," *Frontiers in endocrinology,* vol. 8, p. 90, 2017.
[131] H. Wu, L. Yang, and L.-L. Chen, "The diversity of long noncoding RNAs and their generation," *Trends in genetics,* vol. 33, no. 8, pp. 540-552, 2017.
[132] C.-Y. Guh, Y.-H. Hsieh, and H.-P. Chu, "Functions and properties of nuclear lncRNAs—from systematically mapping the interactomes of lncRNAs," *Journal of Biomedical Science,* vol. 27, no. 1, p. 44, 2020/03/17 2020, doi: 10.1186/s12929-020-00640-3.
[133] L. Statello, C.-J. Guo, L.-L. Chen, and M. Huarte, "Gene regulation by long non-coding RNAs and its biological functions," *Nature Reviews Molecular Cell Biology,* vol. 22, no. 2, pp. 96-118, 2021/02/01 2021, doi: 10.1038/s41580-020-00315-9.
[134] M. Sebastian-delaCruz, I. Gonzalez-Moro, A. Olazagoitia-Garmendia, A. Castellanos-Rubio, and I. Santin, "The Role of lncRNAs in Gene Expression Regulation through mRNA Stabilization," (in eng), *Noncoding RNA,* vol. 7, no. 1, Jan 5 2021, doi: 10.3390/ncrna7010003.
[135] D. Karakas and B. Ozpolat, "The Role of LncRNAs in Translation," (in eng), *Noncoding RNA,* vol. 7, no. 1, Feb 20 2021, doi: 10.3390/ncrna7010016.
[136] M. Cesana *et al.*, "A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA," *Cell,* vol. 147, no. 2, pp. 358-369, 2011.
[137] A. N. Kallen *et al.*, "The imprinted H19 lncRNA antagonizes let-7 microRNAs," *Molecular cell,* vol. 52, no. 1, pp. 101-112, 2013.
[138] J. Ruiz-Orera, X. Messeguer, J. A. Subirana, and M. M. Alba, "Long non-coding RNAs as a source of new peptides," *eLife,* vol. 3, p. e03523, 2014/09/16 2014, doi: 10.7554/eLife.03523.

[139] Z. Ji, R. Song, A. Regev, and K. Struhl, "Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins," *eLife,* vol. 4, p. e08890, 2015/12/19 2015, doi: 10.7554/eLife.08890.
[140] R. W. Henry, V. Mittal, B. Ma, R. Kobayashi, and N. Hernandez, "SNAP19 mediates the assembly of a functional core promoter complex (SNAPc) shared by RNA polymerases II and III," (in eng), *Genes Dev,* vol. 12, no. 17, pp. 2664-72, Sep 1 1998, doi: 10.1101/gad.12.17.2664.
[141] S. Valadkhan and L. S. Gunawardane, "Role of small nuclear RNAs in eukaryotic gene expression," (in eng), *Essays Biochem,* vol. 54, pp. 79-90, 2013, doi: 10.1042/bse0540079.
[142] J. Kufel and P. Grzechnik, "Small nucleolar RNAs tell a different tale," *Trends in Genetics,* vol. 35, no. 2, pp. 104-117, 2019.
[143] S. Massenet, E. Bertrand, and C. Verheggen, "Assembly and trafficking of box C/D and H/ACA snoRNPs," (in eng), *RNA Biol,* vol. 14, no. 6, pp. 680-692, Jun 3 2017, doi: 10.1080/15476286.2016.1243646.
[144] C. Ender *et al.*, "A human snoRNA with microRNA-like functions," (in eng), *Mol Cell,* vol. 32, no. 4, pp. 519-28, Nov 21 2008, doi: 10.1016/j.molcel.2008.10.017.
[145] H. Dana *et al.*, "Molecular Mechanisms and Biological Functions of siRNA," (in eng), *Int J Biomed Sci,* vol. 13, no. 2, pp. 48-57, Jun 2017.
[146] A. G. Seto, R. E. Kingston, and N. C. Lau, "The coming of age for Piwi proteins," *Molecular cell,* vol. 26, no. 5, pp. 603-609, 2007.
[147] A. Arif *et al.*, "GTSF1 accelerates target RNA cleavage by PIWI-clade Argonaute proteins," *Nature,* 2022/06/30 2022, doi: 10.1038/s41586-022-05009-0.
[148] A. A. Aravin *et al.*, "A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice," (in eng), *Mol Cell,* vol. 31, no. 6, pp. 785-99, Sep 26 2008, doi: 10.1016/j.molcel.2008.09.003.
[149] Y. W. Iwasaki, M. C. Siomi, and H. Siomi, "PIWI-Interacting RNA: Its Biogenesis and Functions," *Annual Review of Biochemistry,* vol. 84, no. 1, pp. 405-433, 2015, doi: 10.1146/annurev-biochem-060614-034258.
[150] L. S. Kristensen, M. S. Andersen, L. V. W. Stagsted, K. K. Ebbesen, T. B. Hansen, and J. Kjems, "The biogenesis, biology and characterization of circular RNAs," *Nature Reviews Genetics,* vol. 20, no. 11, pp. 675-691, 2019/11/01 2019, doi: 10.1038/s41576-019-0158-7.
[151] N. R. Pamudurti *et al.*, "Translation of CircRNAs," (in eng), *Mol Cell,* vol. 66, no. 1, pp. 9-21.e7, Apr 6 2017, doi: 10.1016/j.molcel.2017.02.021.
[152] T. Shao, Y. H. Pan, and X. D. Xiong, "Circular RNA: an important player with multiple facets to regulate its parental gene expression," (in eng), *Mol Ther Nucleic Acids,* vol. 23, pp. 369-376, Mar 5 2021, doi: 10.1016/j.omtn.2020.11.008.
[153] H. J. Evans, "Mutation as a cause of genetic disease," (in eng), *Philos Trans R Soc Lond B Biol Sci,* vol. 319, no. 1194, pp. 325-40, Jun 15 1988, doi: 10.1098/rstb.1988.0054.
[154] T. A. Cooper, L. Wan, and G. Dreyfuss, "RNA and disease," (in eng), *Cell,* vol. 136, no. 4, pp. 777-93, Feb 20 2009, doi: 10.1016/j.cell.2009.02.011.
[155] V. Bernat and M. D. Disney, "RNA Structures as Mediators of Neurological Diseases and as Drug Targets," (in eng), *Neuron,* vol. 87, no. 1, pp. 28-46, Jul 1 2015, doi: 10.1016/j.neuron.2015.06.012.
[156] H. Liu, L. Ma, L. Wang, and Y. Yang, "MicroRNA-937 is overexpressed and predicts poor prognosis in patients with colon cancer," *Diagnostic Pathology,* vol. 14, no. 1, p. 136, 2019/12/19 2019, doi: 10.1186/s13000-019-0920-3.

[157] A. M. Khalil and J. L. Rinn, "RNA-protein interactions in human health and disease," (in eng), *Semin Cell Dev Biol,* vol. 22, no. 4, pp. 359-65, Jun 2011, doi: 10.1016/j.semcdb.2011.02.016.
[158] L. Ma *et al.*, "Therapeutic silencing of miR-10b inhibits metastasis in a mouse mammary tumor model," *Nature Biotechnology,* vol. 28, no. 4, pp. 341-347, 2010/04/01 2010, doi: 10.1038/nbt.1618.
[159] T. Hideyama and S. Kwak, "When does ALS start? ADAR2–GluA2 hypothesis for the etiology of sporadic ALS," *Frontiers in molecular neuroscience,* vol. 4, p. 33, 2011.
[160] T. Hideyama *et al.*, "Profound downregulation of the RNA editing enzyme ADAR2 in ALS spinal motor neurons," *Neurobiology of disease,* vol. 45, no. 3, pp. 1121-1128, 2012.
[161] G. Berg *et al.*, "Microbiome definition re-visited: old concepts and new challenges," (in eng), *Microbiome,* vol. 8, no. 1, p. 103, Jun 30 2020, doi: 10.1186/s40168-020-00875-0.
[162] J. C. Wooley, A. Godzik, and I. Friedberg, "A primer on metagenomics," (in eng), *PLoS Comput Biol,* vol. 6, no. 2, p. e1000667, Feb 26 2010, doi: 10.1371/journal.pcbi.1000667.
[163] J. A. Eisen, "Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes," (in eng), *PLoS Biol,* vol. 5, no. 3, p. e82, Mar 2007, doi: 10.1371/journal.pbio.0050082.
[164] J. Peterson *et al.*, "The NIH human microbiome project," *Genome research,* vol. 19, no. 12, pp. 2317-2323, 2009.
[165] D. Gevers *et al.*, "The treatment-naive microbiome in new-onset Crohn's disease," *Cell host & microbe,* vol. 15, no. 3, pp. 382-392, 2014.
[166] G. D. Sepich-Poore, L. Zitvogel, R. Straussman, J. Hasty, J. A. Wargo, and R. Knight, "The microbiome and human cancer," *Science,* vol. 371, no. 6536, p. eabc4552, 2021, doi: doi:10.1126/science.abc4552.
[167] G. D. Sepich-Poore, H. Carter, and R. Knight, "Intratumoral bacteria generate a new class of therapeutically relevant tumor antigens in melanoma," *Cancer Cell,* vol. 39, no. 5, pp. 601-603, 2021.
[168] G. D. Poore *et al.*, "Microbiome analyses of blood and tissues suggest cancer diagnostic approach," (in eng), *Nature,* vol. 579, no. 7800, pp. 567-574, Mar 2020, doi: 10.1038/s41586-020-2095-1.
[169] Y. Fan and O. Pedersen, "Gut microbiota in human metabolic health and disease," *Nature Reviews Microbiology,* vol. 19, no. 1, pp. 55-71, 2021/01/01 2021, doi: 10.1038/s41579-020-0433-9.
[170] T. R. Sampson *et al.*, "Gut microbiota regulate motor deficits and neuroinflammation in a model of Parkinson's disease," *Cell,* vol. 167, no. 6, pp. 1469-1480. e12, 2016.
[171] A. L. Dunlop, J. G. Mulle, E. P. Ferranti, S. Edwards, A. B. Dunn, and E. J. Corwin, "The maternal microbiome and pregnancy outcomes that impact infant health: a review," *Advances in neonatal care: official journal of the National Association of Neonatal Nurses,* vol. 15, no. 6, p. 377, 2015.
[172] M. Wheelis, *Principles of modern microbiology*. Jones & Bartlett Publishers, 2007.
[173] J.-M. Volland *et al.*, "A centimeter-long bacterium with DNA contained in metabolically active, membrane-bound organelles," *Science,* vol. 376, no. 6600, pp. 1453-1458, 2022, doi: doi:10.1126/science.abb3634.
[174] M. T. Cabeen and C. Jacobs-Wagner, "Bacterial cell shape," (in eng), *Nat Rev Microbiol,* vol. 3, no. 8, pp. 601-10, Aug 2005, doi: 10.1038/nrmicro1205.

[175] D. Claessen, D. E. Rozen, O. P. Kuipers, L. Søgaard-Andersen, and G. P. van Wezel, "Bacterial solutions to multicellularity: a tale of biofilms, filaments and fruiting bodies," (in eng), *Nat Rev Microbiol,* vol. 12, no. 2, pp. 115-24, Feb 2014, doi: 10.1038/nrmicro3178.
[176] S. Melnikov, A. Ben-Shem, N. Garreau de Loubresse, L. Jenner, G. Yusupova, and M. Yusupov, "One core, two shells: bacterial and eukaryotic ribosomes," *Nature Structural & Molecular Biology,* vol. 19, no. 6, pp. 560-567, 2012/06/01 2012, doi: 10.1038/nsmb.2313.
[177] U. Rinas, E. Garcia-Fruitós, J. L. Corchero, E. Vázquez, J. Seras-Franzoso, and A. Villaverde, "Bacterial inclusion bodies: discovering their better half," *Trends in biochemical sciences,* vol. 42, no. 9, pp. 726-737, 2017.
[178] H. Barreteau, A. Kovac, A. Boniface, M. Sova, S. Gobec, and D. Blanot, "Cytoplasmic steps of peptidoglycan biosynthesis," (in eng), *FEMS Microbiol Rev,* vol. 32, no. 2, pp. 168-207, Mar 2008, doi: 10.1111/j.1574-6976.2008.00104.x.
[179] J. W. Bartholomew and T. Mittwer, "The Gram stain," (in eng), *Bacteriol Rev,* vol. 16, no. 1, pp. 1-29, Mar 1952, doi: 10.1128/br.16.1.1-29.1952.
[180] R. M. Maier, "Chapter 3 - Bacterial Growth," in *Environmental Microbiology (Second Edition)*, R. M. Maier, I. L. Pepper, and C. P. Gerba Eds. San Diego: Academic Press, 2009, pp. 37-54.
[181] D. Sun, "Pull in and Push Out: Mechanisms of Horizontal Gene Transfer in Bacteria," (in eng), *Front Microbiol,* vol. 9, p. 2154, 2018, doi: 10.3389/fmicb.2018.02154.
[182] B. J. Arnold, I. T. Huang, and W. P. Hanage, "Horizontal gene transfer and adaptive evolution in bacteria," *Nature Reviews Microbiology,* vol. 20, no. 4, pp. 206-218, 2022/04/01 2022, doi: 10.1038/s41579-021-00650-4.
[183] D. Sun, K. Jeannot, Y. Xiao, and C. W. Knapp, "Horizontal gene transfer mediated bacterial antibiotic resistance," vol. 10, ed: Frontiers Media SA, 2019, p. 1933.
[184] D. G. Quispe-Huamanquispe, G. Gheysen, and J. F. Kreuze, "Horizontal Gene Transfer Contributes to Plant Evolution: The Case of Agrobacterium T-DNAs," (in eng), *Front Plant Sci,* vol. 8, p. 2015, 2017, doi: 10.3389/fpls.2017.02015.
[185] E. R. Bejarano, A. Khashoggi, M. Witty, and C. Lichtenstein, "Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution," (in eng), *Proc Natl Acad Sci U S A,* vol. 93, no. 2, pp. 759-64, Jan 23 1996, doi: 10.1073/pnas.93.2.759.
[186] J. A. Schwartz, N. E. Curtis, and S. K. Pierce, "FISH labeling reveals a horizontally transferred algal (Vaucheria litorea) nuclear gene on a sea slug (Elysia chlorotica) chromosome," (in eng), *Biol Bull,* vol. 227, no. 3, pp. 300-12, Dec 2014, doi: 10.1086/BBLv227n3p300.
[187] C. Hall, S. Brachat, and F. S. Dietrich, "Contribution of horizontal gene transfer to the evolution of Saccharomyces cerevisiae," (in eng), *Eukaryot Cell,* vol. 4, no. 6, pp. 1102-15, Jun 2005, doi: 10.1128/ec.4.6.1102-1115.2005.
[188] A. Crisp, C. Boschetti, M. Perry, A. Tunnacliffe, and G. Micklem, "Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes," (in eng), *Genome Biol,* vol. 16, no. 1, p. 50, Mar 13 2015, doi: 10.1186/s13059-015-0607-3.
[189] D. Rothman, S. Marcus, and S. Kiceluk, "On the extension of the germ theory to the etiology of certain common diseases," *Medicine and western civilization,* pp. 253-257, 1995.
[190] J. L. Casanova and L. Abel, "The genetic theory of infectious diseases: a brief history and selected illustrations," (in eng), *Annu Rev Genomics Hum Genet,* vol. 14, pp. 215-43, 2013, doi: 10.1146/annurev-genom-091212-153448.

[191] T. D. Luckey, "Introduction to intestinal microecology," vol. 25, ed: Oxford University Press, 1972, pp. 1292-1294.
[192] R. Sender, S. Fuchs, and R. Milo, "Revised Estimates for the Number of Human and Bacteria Cells in the Body," (in eng), *PLoS Biol,* vol. 14, no. 8, p. e1002533, Aug 2016, doi: 10.1371/journal.pbio.1002533.
[193] K. Aagaard, J. Ma, K. M. Antony, R. Ganu, J. Petrosino, and J. Versalovic, "The Placenta Harbors a Unique Microbiome," *Science Translational Medicine,* vol. 6, no. 237, pp. 237ra65-237ra65, 2014, doi: doi:10.1126/scitranslmed.3008599.
[194] H. Wang *et al.*, "Comprehensive human amniotic fluid metagenomics supports the sterile womb hypothesis," *Scientific Reports,* vol. 12, no. 1, p. 6875, 2022/04/27 2022, doi: 10.1038/s41598-022-10869-7.
[195] M. G. Dominguez-Bello *et al.*, "Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns," (in eng), *Proc Natl Acad Sci U S A,* vol. 107, no. 26, pp. 11971-5, Jun 29 2010, doi: 10.1073/pnas.1002601107.
[196] C. J. Hill *et al.*, "Evolution of gut microbiota composition from birth to 24 weeks in the INFANTMET Cohort," *Microbiome,* vol. 5, no. 1, p. 4, 2017/01/17 2017, doi: 10.1186/s40168-016-0213-y.
[197] C. L. Maynard, C. O. Elson, R. D. Hatton, and C. T. Weaver, "Reciprocal interactions of the intestinal microbiota and immune system," *Nature,* vol. 489, no. 7415, pp. 231-241, 2012.
[198] T. Yatsunenko *et al.*, "Human gut microbiome viewed across age and geography," *nature,* vol. 486, no. 7402, pp. 222-227, 2012.
[199] E. A. Grice *et al.*, "Topographical and temporal diversity of the human skin microbiome," *science,* vol. 324, no. 5931, pp. 1190-1192, 2009.
[200] N. Hasan and H. Yang, "Factors affecting the composition of the gut microbiota, and its modulation," (in eng), *PeerJ,* vol. 7, p. e7502, 2019, doi: 10.7717/peerj.7502.
[201] N. Salazar, L. Valdés-Varela, S. González, M. Gueimonde, and C. G. de Los Reyes-Gavilán, "Nutrition and the gut microbiome in the elderly," (in eng), *Gut Microbes,* vol. 8, no. 2, pp. 82-97, Mar 4 2017, doi: 10.1080/19490976.2016.1256525.
[202] S. Huang *et al.*, "Human skin, oral, and gut microbiomes predict chronological age," *Msystems,* vol. 5, no. 1, pp. e00630-19, 2020.
[203] I. Cho and M. J. Blaser, "The human microbiome: at the interface of health and disease," (in eng), *Nat Rev Genet,* vol. 13, no. 4, pp. 260-70, Mar 13 2012, doi: 10.1038/nrg3182.
[204] J. A. Gilbert, M. J. Blaser, J. G. Caporaso, J. K. Jansson, S. V. Lynch, and R. Knight, "Current understanding of the human microbiome," (in eng), *Nat Med,* vol. 24, no. 4, pp. 392-400, Apr 10 2018, doi: 10.1038/nm.4517.
[205] C. Huttenhower *et al.*, "Structure, function and diversity of the healthy human microbiome," *Nature,* vol. 486, no. 7402, pp. 207-214, 2012/06/01 2012, doi: 10.1038/nature11234.
[206] E. M. Bik *et al.*, "Molecular analysis of the bacterial microbiota in the human stomach," *Proceedings of the National Academy of Sciences,* vol. 103, no. 3, pp. 732-737, 2006.
[207] X.-X. Li *et al.*, "Bacterial Microbiota Profiling in Gastritis without Helicobacter pylori Infection or Non-Steroidal Anti-Inflammatory Drug Use," *PLOS ONE,* vol. 4, no. 11, p. e7985, 2009, doi: 10.1371/journal.pone.0007985.
[208] T. Osaki *et al.*, "Comparative analysis of gastric bacterial microbiota in Mongolian gerbils after long-term infection with Helicobacter pylori," *Microbial Pathogenesis,*

vol. 53, no. 1, pp. 12-18, 2012/07/01/ 2012, doi: https://doi.org/10.1016/j.micpath.2012.03.008.
[209] E. A. Grice *et al.*, "Topographical and Temporal Diversity of the Human Skin Microbiome," *Science*, vol. 324, no. 5931, pp. 1190-1192, 2009, doi: doi:10.1126/science.1171700.
[210] J. Oh *et al.*, "Biogeography and individuality shape function in the human skin metagenome," *Nature*, vol. 514, no. 7520, pp. 59-64, 2014/10/01 2014, doi: 10.1038/nature13786.
[211] J. Oh, Allyson L. Byrd, M. Park, Heidi H. Kong, and Julia A. Segre, "Temporal Stability of the Human Skin Microbiome," *Cell*, vol. 165, no. 4, pp. 854-866, 2016/05/05/ 2016, doi: https://doi.org/10.1016/j.cell.2016.04.008.
[212] C. Urbaniak *et al.*, "Microbiota of human breast tissue," (in eng), *Appl Environ Microbiol*, vol. 80, no. 10, pp. 3007-14, May 2014, doi: 10.1128/aem.00242-14.
[213] E. C. Martens, H. C. Chiang, and J. I. Gordon, "Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont," *Cell host & microbe*, vol. 4, no. 5, pp. 447-457, 2008.
[214] M. Roberfroid, F. Bornet, C. e. Bouley, and J. Cummings, "Colonic microflora: nutrition and health. Summary and conclusions of an International Life Sciences Institute (ILSI)[Europe] workshop held in Barcelona, Spain," *Nutrition reviews*, vol. 53, no. 5, pp. 127-130, 1995.
[215] K. Hou *et al.*, "Microbiota in health and diseases," *Signal Transduction and Targeted Therapy*, vol. 7, no. 1, p. 135, 2022/04/23 2022, doi: 10.1038/s41392-022-00974-4.
[216] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome biology*, vol. 15, no. 3, pp. 1-12, 2014.
[217] F. Beghini *et al.*, "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3," *Elife*, vol. 10, p. e65088, 2021.
[218] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature methods*, vol. 9, no. 4, pp. 357-359, 2012.
[219] P. Menzel, K. L. Ng, and A. Krogh, "Fast and sensitive taxonomic classification for metagenomics with Kaiju," *Nature communications*, vol. 7, no. 1, pp. 1-9, 2016.
[220] L. Schaeffer, H. Pimentel, N. Bray, P. Melsted, and L. Pachter, "Pseudoalignment for metagenomic read assignment," *Bioinformatics*, vol. 33, no. 14, pp. 2082-2088, 2017.
[221] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nature biotechnology*, vol. 34, no. 5, pp. 525-527, 2016.
[222] C. Zhang *et al.*, "Identification of low abundance microbiome in clinical samples using whole genome sequencing," *Genome biology*, vol. 16, no. 1, pp. 1-16, 2015.
[223] D. Ribet and P. Cossart, "How bacterial pathogens colonize their hosts and invade deeper tissues," *Microbes and infection*, vol. 17, no. 3, pp. 173-183, 2015.
[224] J. N. Weinstein *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, pp. 1113-1120, 2013.
[225] G. Skoufos *et al.*, "AGAMEMNON: an Accurate metaGenomics And MEtatranscriptoMics quaNtificatiON analysis suite," *Genome biology*, vol. 23, no. 1, pp. 1-27, 2022.
[226] M. S. Lindner and B. Y. Renard, "Metagenomic abundance estimation and diagnostic testing on species level," *Nucleic acids research*, vol. 41, no. 1, pp. e10-e10, 2013.
[227] D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster, "Integrative analysis of environmental sequences using MEGAN4," *Genome research*, vol. 21, no. 9, pp. 1552-1560, 2011.

[228] J. Lu, F. P. Breitwieser, P. Thielen, and S. L. Salzberg, "Bracken: estimating species abundance in metagenomics data," *PeerJ Computer Science*, vol. 3, p. e104, 2017.
[229] M. Fischer, B. Strauch, and B. Y. Renard, "Abundance estimation and differential testing on strain level in metagenomics data," *Bioinformatics*, vol. 33, no. 14, pp. i124-i132, 2017.
[230] F. Almodaresi, H. Sarkar, A. Srivastava, and R. Patro, "A space and time-efficient index for the compacted colored de Bruijn graph," *Bioinformatics*, vol. 34, no. 13, pp. i169-i177, 2018.
[231] J. Cheng, "Shiny: Easy web applications in R," in *The R User Conference, useR! 2013 July 10-12 2013 University of Castilla-La Mancha, Albacete, Spain*, 2013, vol. 10, no. 30, p. 93.
[232] D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with Kraken 2," *Genome biology*, vol. 20, no. 1, pp. 1-13, 2019.
[233] D. R. Mende *et al.*, "Assessment of metagenomic assembly using simulated next generation sequencing data," *PloS one*, vol. 7, no. 2, p. e31386, 2012.
[234] V. Sevim *et al.*, "Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies," *Scientific data*, vol. 6, no. 1, pp. 1-9, 2019.
[235] D. Kim, B. Langmead, and S. L. Salzberg, "HISAT: a fast spliced aligner with low memory requirements," *Nature methods*, vol. 12, no. 4, pp. 357-360, 2015.
[236] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype," *Nature biotechnology*, vol. 37, no. 8, pp. 907-915, 2019.
[237] M. A. Walker *et al.*, "GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts," *Bioinformatics*, vol. 34, no. 24, pp. 4287-4289, 2018.
[238] W. Huang, L. Li, J. R. Myers, and G. T. Marth, "ART: a next-generation sequencing read simulator," *Bioinformatics*, vol. 28, no. 4, pp. 593-594, 2012.
[239] A. Almeida *et al.*, "A unified catalog of 204,938 reference genomes from the human gut microbiome," *Nature biotechnology*, vol. 39, no. 1, pp. 105-114, 2021.
[240] H. Integrative, "The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease," *Cell host & microbe*, vol. 16, no. 3, pp. 276-289, 2014.
[241] T. Ohkusa, N. Sato, T. Ogihara, K. Morita, M. Ogawa, and I. Okayasu, "Fusobacterium varium localized in the colonic mucosa of patients with ulcerative colitis stimulates species-specific antibody," *Journal of gastroenterology and hepatology*, vol. 17, no. 8, pp. 849-853, 2002.
[242] T. H. Luu, C. Michel, J.-M. Bard, F. Dravet, H. Nazih, and C. Bobin-Dubigeon, "Intestinal proportion of Blautia sp. is associated with clinical stage and histoprognostic grade in patients with early-stage breast cancer," *Nutrition and cancer*, vol. 69, no. 2, pp. 267-275, 2017.
[243] S. Ohigashi *et al.*, "Changes of the intestinal microbiota, short chain fatty acids, and fecal pH in patients with colorectal cancer," *Digestive diseases and sciences*, vol. 58, no. 6, pp. 1717-1726, 2013.
[244] T. Toya *et al.*, "Coronary artery disease is associated with an altered gut microbiome composition," *PloS one*, vol. 15, no. 1, p. e0227147, 2020.
[245] M. R. Rubinstein *et al.*, "Fusobacterium nucleatum promotes colorectal cancer by inducing Wnt/β-catenin modulator Annexin A1," *EMBO reports*, vol. 20, no. 4, p. e47638, 2019, doi: https://doi.org/10.15252/embr.201847638.

[246] A. C. Wong and M. Levy, "New Approaches to Microbiome-Based Therapies," (in eng), *mSystems,* vol. 4, no. 3, Jun 4 2019, doi: 10.1128/mSystems.00122-19.
[247] G. Skoufos *et al.*, "Peryton: a manual collection of experimentally supported microbe-disease associations," *Nucleic acids research,* vol. 49, no. D1, pp. D1328-D1333, 2021.
[248] S. Federhen, "The NCBI Taxonomy database," (in eng), *Nucleic Acids Res,* vol. 40, no. Database issue, pp. D136-43, Jan 2012, doi: 10.1093/nar/gkr1178.
[249] S. Gottesman and G. Storz, "Bacterial small RNA regulators: versatile roles and rapidly evolving variations," (in eng), *Cold Spring Harb Perspect Biol,* vol. 3, no. 12, Dec 1 2011, doi: 10.1101/cshperspect.a003798.
[250] S. Gottesman *et al.*, "Small RNA regulators and the bacterial response to stress," (in eng), *Cold Spring Harb Symp Quant Biol,* vol. 71, pp. 1-11, 2006, doi: 10.1101/sqb.2006.71.016.
[251] E. Holmqvist and E. G. H. Wagner, "Impact of bacterial sRNAs in stress responses," (in eng), *Biochem Soc Trans,* vol. 45, no. 6, pp. 1203-1212, Dec 15 2017, doi: 10.1042/bst20160363.
[252] J. J. González Plaza, "Small RNAs as Fundamental Players in the Transference of Information During Bacterial Infectious Diseases," (in eng), *Front Mol Biosci,* vol. 7, p. 101, 2020, doi: 10.3389/fmolb.2020.00101.
[253] H. Eichner, J. Karlsson, and E. Loh, "The emerging role of bacterial regulatory RNAs in disease," (in eng), *Trends Microbiol,* Apr 1 2022, doi: 10.1016/j.tim.2022.03.007.
[254] D. G. Mediati, S. Wu, W. Wu, and J. J. Tree, "Networks of Resistance: Small RNA Control of Antibiotic Resistance," (in eng), *Trends Genet,* vol. 37, no. 1, pp. 35-45, Jan 2021, doi: 10.1016/j.tig.2020.08.016.
[255] S. K. Eisenbart, M. Alzheimer, S. R. Pernitzsch, S. Dietrich, S. Stahl, and C. M. Sharma, "A Repeat-Associated Small RNA Controls the Major Virulence Factors of Helicobacter pylori," (in eng), *Mol Cell,* vol. 80, no. 2, pp. 210-226.e7, Oct 15 2020, doi: 10.1016/j.molcel.2020.09.009.
[256] S. Melamed *et al.*, "Global Mapping of Small RNA-Target Interactions in Bacteria," (in eng), *Mol Cell,* vol. 63, no. 5, pp. 884-97, Sep 1 2016, doi: 10.1016/j.molcel.2016.07.026.
[257] A. E. Saliba, C. S. S, and J. Vogel, "New RNA-seq approaches for the study of bacterial pathogens," (in eng), *Curr Opin Microbiol,* vol. 35, pp. 78-87, Feb 2017, doi: 10.1016/j.mib.2017.01.001.
[258] K. Han, B. Tjaden, and S. Lory, "GRIL-seq provides a method for identifying direct targets of bacterial small regulatory RNA by in vivo proximity ligation," (in eng), *Nat Microbiol,* vol. 2, p. 16239, Dec 22 2016, doi: 10.1038/nmicrobiol.2016.239.
[259] M. Hafner *et al.*, "CLIP and complementary methods," *Nature Reviews Methods Primers,* vol. 1, no. 1, p. 20, 2021/03/04 2021, doi: 10.1038/s43586-021-00018-1.
[260] A. Santos-Zavaleta *et al.*, "RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12," (in eng), *Nucleic Acids Res,* vol. 47, no. D1, pp. D212-d220, Jan 8 2019, doi: 10.1093/nar/gky1077.
[261] H. Y. Huang *et al.*, "sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes," (in eng), *Nucleic Acids Res,* vol. 37, no. Database issue, pp. D150-4, Jan 2009, doi: 10.1093/nar/gkn852.
[262] J. Pischimarov *et al.*, "sRNAdb: a small non-coding RNA database for gram-positive bacteria," (in eng), *BMC Genomics,* vol. 13, p. 384, Aug 10 2012, doi: 10.1186/1471-2164-13-384.

[263] L. Li, D. Huang, M. K. Cheung, W. Nong, Q. Huang, and H. S. Kwan, "BSRD: a repository for bacterial small regulatory RNA," (in eng), *Nucleic Acids Res,* vol. 41, no. Database issue, pp. D233-8, Jan 2013, doi: 10.1093/nar/gks1264.
[264] J. Wang *et al.*, "sRNATarBase 3.0: an updated database for sRNA-target interactions in bacteria," (in eng), *Nucleic Acids Res,* vol. 44, no. D1, pp. D248-53, Jan 4 2016, doi: 10.1093/nar/gkv1127.
[265] G. Matera, Y. Altuvia, M. Gerovac, Y. El Mouali, H. Margalit, and J. Vogel, "Global RNA interactome of Salmonella discovers a 5' UTR sponge for the MicF small RNA that connects membrane permeability to transport capacity," (in eng), *Mol Cell,* vol. 82, no. 3, pp. 629-644.e4, Feb 3 2022, doi: 10.1016/j.molcel.2021.12.030.
[266] E. P. Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature,* vol. 489, no. 7414, p. 57, 2012.
[267] C. A. Davis *et al.*, "The Encyclopedia of DNA elements (ENCODE): data portal update," *Nucleic acids research,* vol. 46, no. D1, pp. D794-D801, 2018.
[268] T. Barrett *et al.*, "NCBI GEO: archive for functional genomics data sets—update," *Nucleic acids research,* vol. 41, no. D1, pp. D991-D995, 2012.
[269] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet. journal,* vol. 17, no. 1, pp. 10-12, 2011.
[270] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature methods,* vol. 14, no. 4, pp. 417-419, 2017.
[271] D. Mas-Ponte, J. Carlevaro-Fita, E. Palumbo, T. H. Pulido, R. Guigo, and R. Johnson, "LncATLAS database for subcellular localization of long noncoding RNAs," *Rna,* vol. 23, no. 7, pp. 1080-1087, 2017.
[272] F. Cunningham *et al.*, "Ensembl 2019," *Nucleic acids research,* vol. 47, no. D1, pp. D745-D751, 2019.
[273] N. A. O'Leary *et al.*, "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation," *Nucleic acids research,* vol. 44, no. D1, pp. D733-D745, 2016.
[274] M. N. Cabili *et al.*, "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses," *Genes & development,* vol. 25, no. 18, pp. 1915-1927, 2011.
[275] N. Daniel, E. Lécuyer, and B. Chassaing, "Host/microbiota interactions in health and diseases—Time for mucosal microbiology!," *Mucosal Immunology,* vol. 14, no. 5, pp. 1006-1016, 2021/09/01 2021, doi: 10.1038/s41385-021-00383-w.
[276] E. L. Gulliver *et al.*, "Review article: the future of microbiome-based therapeutics," (in eng), *Aliment Pharmacol Ther,* vol. 56, no. 2, pp. 192-208, Jul 2022, doi: 10.1111/apt.17049.
[277] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "miRBase: from microRNA sequences to function," *Nucleic acids research,* vol. 47, no. D1, pp. D155-D162, 2019.
[278] J. E. Handzlik, S. Tastsoglou, I. S. Vlachos, and A. G. Hatzigeorgiou, "Manatee: detection and quantification of small non-coding RNAs from next-generation sequencing data," *Scientific reports,* vol. 10, no. 1, pp. 1-10, 2020.
[279] B. Bushnell, "BBMap: a fast, accurate, splice-aware aligner," Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014.
[280] G. Varani and W. H. McClain, "The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems," (in eng), *EMBO Rep,* vol. 1, no. 1, pp. 18-23, Jul 2000, doi: 10.1093/embo-reports/kvd001.

[281] R. Lorenz *et al.*, "ViennaRNA Package 2.0," *Algorithms for molecular biology,* vol. 6, no. 1, pp. 1-14, 2011.
[282] M. Zuker and P. Stiegler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," *Nucleic acids research,* vol. 9, no. 1, pp. 133-148, 1981.
[283] Q.-F. Wen, S. Liu, C. Dong, H.-X. Guo, Y.-Z. Gao, and F.-B. Guo, "Geptop 2.0: an updated, more precise, and faster Geptop server for identification of prokaryotic essential genes," *Frontiers in microbiology,* vol. 10, p. 1236, 2019.