ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ
ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

Χαρακτηρισμός Μικροβιακών Γονιδιωμάτων με Τεχνικές Αλληλούχισης Επόμενης Γενιάς

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Συγγραφέας:

Δανιηλίδης Απ. Κωνσταντίνος (Α.Μ.: 2018091)

Επιβλέπουσα:

Χατζηγεωργίου Άρτεμις, Καθηγήτρια

Λαμία, 2022

"A gene is a long sequence of coded letters,
like computer information. Modern biology is
becoming very much a branch of information technology."

- Richard Dawkins

"Knowledge of sequences
could contribute much to our
understanding of living matter..."

- Frederick Sanger

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις[1], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια.*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: **7/6/2022**

Ο Δηλών,

[1] «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Χαρακτηρισμός Μικροβιακών Γονιδιωμάτων με Τεχνικές Αλληλούχισης Επόμενης Γενιάς

Συγγραφέας:

Δανιηλίδης Απ. Κωνσταντίνος

Τριμελής Επιτροπή:

Χατζηγεωργίου Άρτεμις,
(Επιβλέπουσα) Καθηγήτρια του Τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική της Σχολής Θετικών Επιστημών του Πανεπιστημίου Θεσσαλίας

Μπάγκος Παντελεήμων,
Καθηγητής του Τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική της Σχολής Θετικών Επιστημών του Πανεπιστημίου Θεσσαλίας

Μπράλιου Γεωργία,
Επίκουρος Καθηγήτρια του Τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική της Σχολής Θετικών Επιστημών του Πανεπιστημίου Θεσσαλίας

Περιεχομενα

Σελίδα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κατάλογος Συντομογραφιών

NGS Next Genaration Sequencing

TSS Transcription Start Site

UTR Untranslated region

***lm*RNA** Leaderless *m*RNA

LCFAs Long Chain Fatty Acids

TLRs Toll like receptors

CDC Centers for Disease Control and Prevention

VCE Vaccinia Capping Enzyme

PCR Polymerase chain reaction

CMOS Complementary Metal-oxide Semiconductor

HGP Human Genome Project

SMRT-seq Single Molecule Real Time Sequencing

ZMWs Zero-mode Waveguides

dNMP Deoxynucleoside Monophosphate

DEG Differentially Expressed Genes

ORFs Open Reading Frames

***d*RNA-seq** Differential RNA Sequencing

TAP Tobacco Acid Pyrophosphatase

TEX Terminator™ 5' monophosphate-dependent exonuclease

CBC Cap Binding Complex

TRCF Transcription Repair Coupling Factor

SD Shine – Dalgarno

RBS Ribosome Binding Site

***i*GV** Integrative Genomics Viewer

BED Browser Extensible Data

BAC Bacterial Artificial Chromosome

RRS Relative Read Score

MSE Mean squared error

PDF Probability Density Function

GLM Generalized Linear Models

GAM Generalized Additive Models

sORF Small Open Reading Frames

ASVs Amplicon Sequence Variants

OTUs Operational Taxonomic Units

RDP Ribosomal Database Project

Ευχαριστιες

Η παρούσα πτυχιακή εργασία, εκπονήθηκε από το Φεβρουάριο του 2022 έως τον Ιούνιο του ίδιου έτους, στα πλαίσια της απόκτησης του βασικού πτυχίου από το τμήμα Πληροφορικής με εφαρμογές στη Βιοϊατρική, του Πανεπιστημίου Θεσσαλίας.

Με την ολοκλήρωση αυτής της πτυχιακής εργασίας, σηματοδοτείται ταυτόχρονα και η ολοκλήρωση ενός κύκλου σπουδών, ίσως του πιο σημαντικού για τον ερευνητικό προσανατολισμό, αυτού, του προπτυχιακού επιπέδου. Οφείλω ένα πολύ μεγάλο ευχαριστώ όλα αυτά τα χρόνια, σε όλους τους διδάσκοντες ανεξαιρέτως, οι οποίοι μας στήριξαν ως φοιτητές, παρέχοντάς μας καθημερινά το κίνητρο να ανακαλύπτουμε με τον καλύτερο δυνατό τρόπο, το ερευνητικό μας αντικείμενο και τις ενδιαφέρουσες προοπτικές του. Φυσικά, τα γνωστικά αντικείμενα που έχει την ευκαιρία να ακολουθήσει ένας μαθητευόμενος στο ευρύτερο πεδίο της πληροφορικής, είναι αρκετά σε αριθμό και σε αυτή τη διαδικασία του ερευνητικού «προσανατολισμού», το σημαντικότερο μερίδιο ευθύνης το κατέχουν οι ίδιοι οι εκπαιδευτικοί, πέρα από τη φύση του κάθε μαθήματος.

Η παρούσα πτυχιακή εργασία υλοποιήθηκε με την υποστήριξη ορισμένων ανθρώπων, για τους οποίους θα ήταν παράλειψη να μην εκφράσω τις θερμότερες ευχαριστίες μου. Ιδιαίτερη μνεία ευχαρίστησης, οφείλω στην επιβλέπουσα καθηγήτριά μου, την κυρία Χατζηγεωργίου Άρτεμις, για την εξαιρετική μας συνεργασία και την υποστήριξή της, καθώς και επειδή αποτέλεσε για εμένα σταθμό έμπνευσης, διαρκούς προβληματισμού αλλά και πηγή γνώσης, παρέχοντάς μου σύγχρονα εφόδια προσέγγισης βιοπληροφορικών ζητημάτων. Η καθοδήγηση που μου παρείχε, η βοήθεια και η εμπιστοσύνη που μου έδειξε, ήταν άκρως σημαντικές για την ομαλή διεξαγωγή της εργασίας αυτής και σίγουρα υπάρχουν μόνο θετικές αναμνήσεις.

Επιπλέον δε, νιώθω την ανάγκη να εκφράσω την ευγνωμοσύνη μου στον Γεώργιο Σκούφο, διδακτορικό ερευνητή του εργαστηρίου Μοριακής και Υπολογιστικής Βιολογίας και Γενετικής, Diana Lab, για την συμβολή του σε κάθε ζήτημα που προέκυπτε, καθώς και για την επίσης εξαιρετική συνεργασία που είχαμε και ήταν δίπλα μου, από την αρχή αυτής της διαδρομής.

Περιληψη

Την τελευταία δεκαετία, ο συνδυασμός πειραμάτων Αλληλούχισης Επόμενης Γενιάς (Next Generation Sequencing, NGS) και οι μέθοδοι βιοπληροφορικής ανάλυσης, έχουν καταστήσει δυνατό, το λεπτομερή χαρακτηρισμό της βιοποικιλότητας της μικροχλωρίδας, της σχέσης της με τη φυσιολογική λειτουργία του ανθρώπινου οργανισμού, αλλά και της εμπλοκής της σε παθολογικές καταστάσεις όπως ο καρκίνος. Μία πληθώρα διαφόρων ταξινομικών ομάδων βακτηριακών κυττάρων, επιβιώνει και αλληλεπιδρά συνεχώς, τόσο με τον ανθρώπινο οργανισμό, όσο και με κάθε άλλο πολυκύτταρο ευκαρυωτικό οργανισμό. Έως σήμερα, το επίπεδο οργάνωσης πολλών από αυτών των βακτηριακών κυττάρων, παραμένει σχεδόν άγνωστο και δεν είναι ακόμη πλήρως κατανοητή και τεκμηριωμένη η λειτουργία και η πολυπλοκότητα των γονιδιωμάτων τους, καθώς και πολλών βιολογικών διεργασιών. Στόχος της παρούσας πτυχιακής εργασίας, είναι η πραγματοποίηση βιοπληροφορικής ανάλυσης σε NGS δεδομένα, με απώτερο στόχο ως επί το πλείστον, το χαρακτηρισμό βακτηριακών γονιδιωμάτων, μέσα από υπολογιστική μεθοδολογία. Πιο αναλυτικά, με τη βοήθεια σύγχρονων τεχνικών αλληλούχισης, όπως η *Cappable-seq* και η *Term-seq*, δύνανται να προκύψουν δεδομένα σχολιασμού γονιδιωμάτων, αναφορικά με τις 5′ και 3′ αμετάφραστες περιοχές (Untranslated Regions, UTRs), σε επίπεδο ανάλυσης ενός νουκλεοτιδίου (single nucleotide resolution). Τέλος, με την ανάλυση αυτή, έχουν σχολιαστεί ρυθμιστικές περιοχές βακτηριακών μεταγράφων (transcripts) που ελέγχουν τη γονιδιακή έκφραση, οι οποίες θα ενισχύσουν, πέρα από την ίδια τη μεταγραφωμική (transcriptomics), την πρωτεομική (proteomics) αλλά και μεταβολομική (metabolomics) προέκταση των μελετών.

Λέξεις Κλειδιά: Μικροβίωμα, Μεταγονιδίωμα, Αλληλούχιση Επόμενης Γενιάς, Cappable-seq, Βιοπληροφορική ανάλυση, Θέση Έναρξης Μεταγραφής.

Abstract

In the last decade, the combination of Next Generation Sequencing (NGS) experiments and bioinformatics analysis methods, has made possible the detailed characterization of the microflora biodiversity, its relationship with the normal functioning of the human body, but also its involvement in pathological conditions, such as cancer. A variety of different taxonomic groups of bacterial cells, survive and interact continuously, both with the human body and with any other multicellular eukaryotic organism. To date, the level of organization of many of these bacterial cells, remains almost unknown and the function and complexity of their genomes, as well as many biological processes, are not yet fully understood and documented. The aim of this study, is to carry out bioinformatics analysis of NGS data, with the ultimate goal mostly, the characterization of bacterial genomes, through computational methodology. More analytically, with the help of modern sequencing techniques, such as *Cappable-seq* and *Term-seq*, can be generated genomic annotation data, regarding the 5' and 3' Untranslated regions (UTRs), at single nucleotide resolution. Finally, with this analysis, have annotated regulatory elements of bacterial transcripts that control gene expression, which will enhance, in addition to the transcriptomic, the proteomic and metabolomic extension of the studies.

Keywords— Microbiome, Metagenome, Next Generation Sequencing, Cappable-seq, Bioinformatic analysis, Transcription Start Site.

1 | Εισαγωγη

Η σημαντική πρόοδος της τεχνολογίας, σε συνδυασμό με την πρωτοποριακή έρευνα στο χώρο της βιολογίας, έχουν κάνει πραγματικότητα τη λύση, σε ένα μεγάλο σύνολο ζητημάτων, τα οποία απασχολούσαν ολόκληρη την ανθρωπότητα επί αιώνες. Προβλήματα, όπως πολλές πανδημίες, διάφορες κληρονομικές ασθένειες και νόσοι, έχουν εδώ και χρόνια αντιμετωπιστεί, χάρις στις ραγδαίες εξελίξεις της επιστήμης. Μία από αυτές τις εξελίξεις, είναι και η αλληλούχιση του γενετικού υλικού, σε κλίμακα υψηλής διεκπεραιωτικής ικανότητας (High Throughput Sequencing). Κατ΄ επέκταση, τα τελευταία χρόνια έχουν δημιουργηθεί νέοι ερευνητικοί ορίζοντες, αναφορικά με τον τρόπο ή την τεχνική αλληλούχισης, με βάση το βιολογικό πρόβλημα που τίθεται και τον οργανισμό που μελετάται. Μία από αυτές τις κατευθύνσεις, είναι και η αλληλούχιση του μεταγονιδιώματος, δηλαδή του γονιδιωματικού γενετικού υλικού των μικροβιακών κοινοτήτων. Με βάση αυτή, εκτιμάται η βαθύτερη κατανόηση της οργάνωσης και λειτουργίας των διάφορων μικροβιακών οργανισμών, γεγονός που μπορεί με τη σειρά του να βοηθήσει την έρευνα, τόσο σε επίπεδο μεταγραφωμικής (transcriptomics), όσο και σε επίπεδο πρωτεομικής (proteomics) και μεταβολομικής (metabolomics) (Lee Ki-Hyun *et al.*, 2013), σηματοδοτώντας ταυτόχρονα ένα πολλά υποσχόμενο μέλλον.

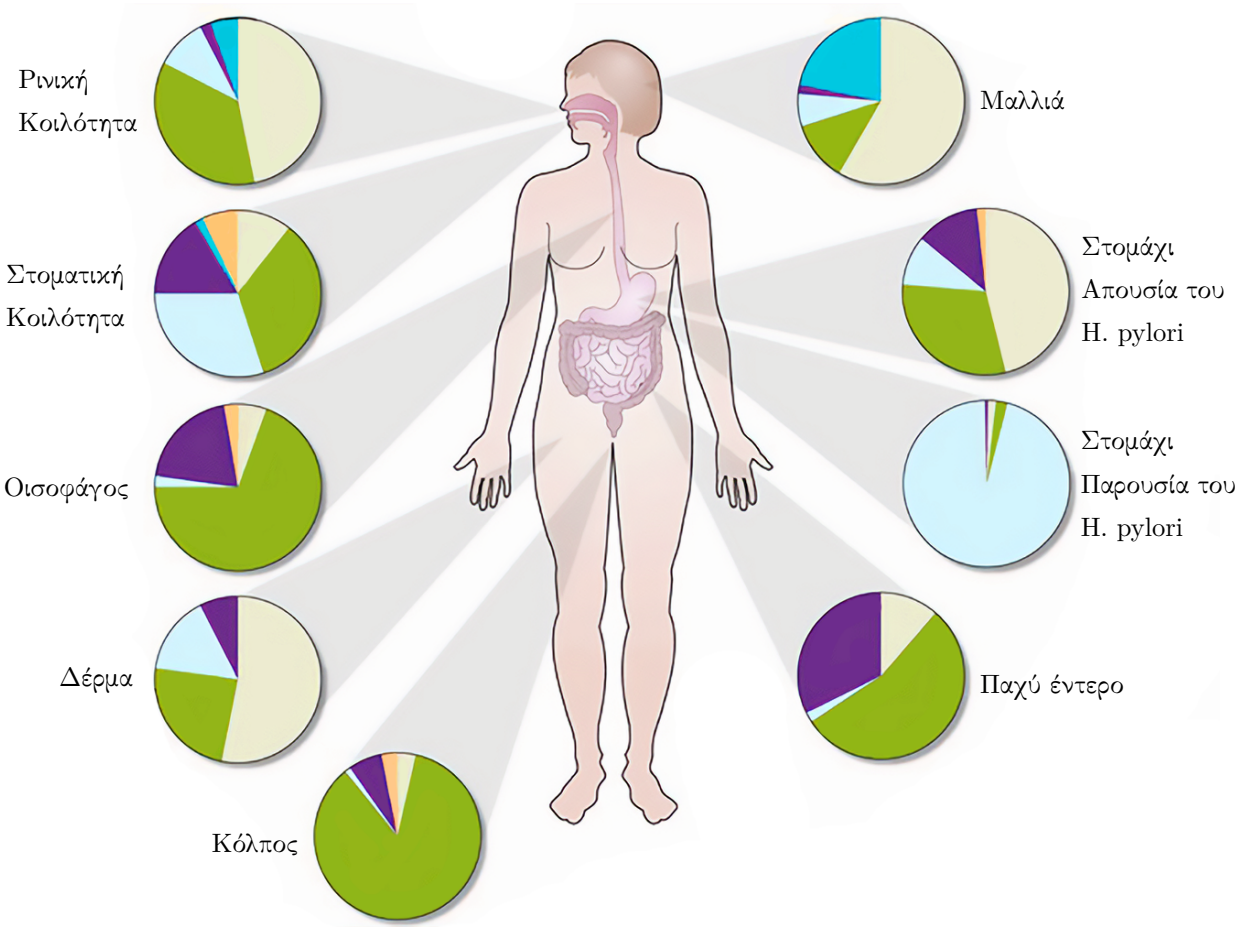
Η εργασία αυτή χωρίζεται σε δύο μέρη. Στο πρώτο μέρος γίνεται μία βιβλιογραφική ανασκόπηση, βασισμένη σε δύο άξονες: στις τρέχουσες γνώσεις της επιστήμης για τη λειτουργία, το ρόλο και την οργάνωση του μικροβιώματος και στην ιστορική αναδρομή της διαδικασίας αλληλούχισης γονιδιωμάτων, από τα πρώτα χρόνια αυτού του εγχειρήματος, έως όπως τη γνωρίζουμε σήμερα. Ακόμα, αναφέρονται σύγχρονες τεχνικές αλληλούχισης, καθεμία από τις οποίες δίνει λύση σε συγκεκριμένα βιολογικά προβλήματα και επισημαίνονται τρέχουσες προκλήσεις για περαιτέρω ανάλυση. Στο δεύτερο μέρος υπάρχει η υπολογιστική μεθοδολογία που αναπτύχθηκε, η οποία στοχεύει στην ανάλυση και τον εντοπισμό νέων σημείων έναρξης της μεταγραφής (Transcription Start Site (TSS)), καθώς και στο σχολιασμό των συνολικών αποτελεσμάτων και τη γενικευμένη εφαρμογή αυτής της μεθοδολογίας σε κάθε βακτηριακό οργανισμό. Πριν όμως, αναφερθούν αναλυτικά οι τεχνικές αλληλούχισης, γίνεται αναφορά στο μικροβίωμα, το ρόλο που αυτό κατέχει στην εκδήλωση και ανάπτυξη ασθενειών, στα οφέλη που παρέχει στον ανθρώπινο οργανισμό, καθώς και στο επίπεδο οργάνωσής του.

1.1. Μικροβιωμα – Μεταγονιδιωμα

Το ανθρώπινο σώμα αποτελείται από τρισεκατομμύρια κύτταρα, τα οποία υποστηρίζουν την ανάπτυξη, τη σωστή λειτουργία και την ανοσοπροστασία του. Ωστόσο, δεν είναι τα μόνα δομικά συστατικά που περιέχονται σε αυτό. Τα διάφορα μέρη του σώματος περιέχουν μικροβιακές κοινότητες και μικροβιώματα, τα οποία διαφέρουν ως προς την σύσταση και τη λειτουργία και ζουν σε συμβίωση με τα κύτταρα του ξενιστή (Huttenhower *et al.*, 2012). Οι πρώτοι κατάλογοι μικροβιακών γονιδίων, έγιναν διαθέσιμοι από το εθνικό ινστιτούτο υγείας των ΗΠΑ και το MetaHIT (METAgenomics of the Human Intestinal Tract), όπου προέκυψε πως τα διάφορα μικροβιακά είδη είναι προσαρμοσμένα σε 15 και 18 διαφορετικούς σωματικούς τόπους σε άντρες και γυναίκες, αντίστοιχα (Aagaard *et al.*, 2013; Peterson *et al.*, 2009). Αυτές ήταν ορισμένες από τις πρώτες προσεγγίσεις στην ανάλυση της σύστασης του μικροβιώματος και στη συμβίωσή του με το ανθρώπινο σώμα. Ο όρος συμβίωση αναφέρεται στο ότι οι μικροοργανισμοί, συνήθως δεν προκαλούν κάποια μόλυνση όσο το σώμα βρίσκεται σε ομοιόσταση, δηλαδή ζουν σε αρμονία με αυτό, επιτελώντας σημαντικές λειτουργίες, όπως θα αναλυθεί εκτενώς παρακάτω.

Τα βακτήρια ήταν η πρώτη μορφή ζωής που εμφανίστηκαν στη Γη πριν από 3.8 δισεκατομμύρια χρόνια και 1,4 δισεκατομμύρια χρόνια πριν από την εμφάνιση της ευκαρυωτικής γενεαλογίας, που περιλαμβάνει τους ανθρώπους και γενικότερα όλους τους οργανισμούς, των οποίων τα κύτταρα είναι εμπύρηνα. Ορισμένα από τα ιδιαίτερα χαρακτηριστικά τους, είναι το απλό επίπεδο οργάνωσης, καθώς και το ότι μπορούν να προσαρμόζονται σε νέα περιβάλλοντα μέσα σε μικρό χρονικό διάστημα, ρυθμίζοντας κατάλληλα την ανοσολογική τους απόκριση (Rolfe *et al.*, 2012). Η συμβίωση ξενιστή - μικροβιώματος, συνεπάγεται τη μετάδοση μικροβίων από το είδος του, σε δι-

άφορες γενιές. Μέσω της διαδικασίας της φυσικής επιλογής, οι μεταλλάξεις οδηγούν σε εξελικτικές προσαρμογές στις περιβαλλοντικές συνθήκες. Πράγματι, το ανθρώπινο περιβάλλον έχει αλλάξει δραματικά κατά την ανθρώπινη εξέλιξη και οι διατροφικές αλλαγές άσκησαν σημαντικές επιλεκτικές πιέσεις, ώστε να διαμορφωθεί η προσαρμογή των οργανισμών στο περιβάλλον τους, όπως τη γνωρίζουμε σήμερα. Ενώ υπάρχουν ενδείξεις προσαρμογής στο γονιδίωμα, ή ακόμα και χαρακτηριστικών επιβίωσης κάτω από συνθήκες ασιτίας, οι προσαρμογές του μικροβιώματος που προσφέρουν εξοικονόμηση ενέργειας στον ξενιστή, παραμένουν άγνωστες (Moeller *et al.*, 2014; Muegge *et al.*, 2011).



Σχήμα 1.1: *Διαφορές σύνθεσης του μικροβιώματος στον άνθρωπο, ανά ανατομική θέση. Με μωβ χρώμα απεικονίζονται τα βακτήρια της κλάσης Bacteroidetes, με πράσινο τα βακτήρια της συνομοταξίας Bacillota, με ανοιχτό πράσινο τα βακτήρια της συνομοταξίας Actinobacteria, με μπλε τα Cyanobacteria, με ανοιχτό μπλε τα Proteobacteria και με κίτρινο τα Fusobacteria. Η εικόνα τροποποιήθηκε από (*Aagaard et al., *2013).*

Οι άνθρωποι, είναι ουσιαστικά γενετικά πανομοιότυποι μεταξύ τους και κληρονομούν τα ίδια μορφολογικά χαρακτηριστικά. Ωστόσο, λόγω ορισμένων αλλαγών στο DNA, υπάρχει αυτή η τεράστια γενετική ποικιλομορφία, ανάμεσα στα άτομα των πληθυσμών. Αντίθετα, το γονιδίωμα του μικροβιώματος είναι περισσότερο μεταβλητό, με το 1/3 των γονιδίων τους να βρίσκεται στη πλειονότητα των υγειών ατόμων. Συνεπώς, η κατανόηση της μεταβλητότητας του υγιούς μικροβιώματος είναι μια σημαντική πρόκληση που χρονολογείται από τη δεκαετία του '60. Βασική προϋπόθεση για το θεραπευτικό ρόλο του μικροβιώματος, είναι ο μηχανισμός που καθορίζει τη διαμόρφωση αυτού. Υπάρχουν δύο κατηγορίες μηχανισμών λειτουργίας, οι οποίες αποτελούν τη συναρμολόγηση του μικροβιώματος και αυτές είναι ο τρόπος μετάδοσής του και το φιλτράρισμα. Ο πρώτος τρόπος, διακρίνεται στην οριζόντια μετάδοση και στην κάθετη μετάδοση του μικροβιώματος.

Στη κάθετη μετάδοση, οι μικροοργανισμοί μεταφέρονται απευθείας από τους γονείς στους απογόνους. Η συμβίωση με πληθυσμό ξενιστών που αποκλίνουν γενετικά, έχει ως αποτέλεσμα την υψηλή πιστότητα της σχέσης

ξενιστή - μικροβίου σε μακρές χρονικές κλίμακες. Τέτοιου είδους αλληλεπιδράσεις, αφήνουν υπογραφές στα γονιδιώματα των συμβιόντων. Στα θηλαστικά, αυτή η μετάδοση γίνεται κατά τη γέννηση και τη διάρκεια του θηλασμού. Περίπου το 75% του μικροβιώματος του ανθρώπου προέρχεται από τη μητέρα. Κατά τη γέννηση, ουσιαστικά λαμβάνουμε κολπικά μικρόβια, παρόλο που η μήτρα θεωρείται στείρο περιβάλλον. Αυτή η βάπτιση μικροβίων είναι ζωτικής σημασίας για ένα υγιές ξεκίνημα της ζωής (Asnicar *et al.*, 2017). Επιπλέον, θεωρείται πως τα παιδιά που γεννιούνται με καισαρική τομή, είναι περισσότερο πιθανό να αναπτύξουν άσθμα, αλλεργίες, κοιλιοκάκη και παχυσαρκία στη μετέπειτα ζωή τους, δηλαδή ασθένειες που συνδέονται ίσως με αδύναμο ή εξασθενημένο μικροβίωμα, λόγω αυτού του τρόπου γέννησης.

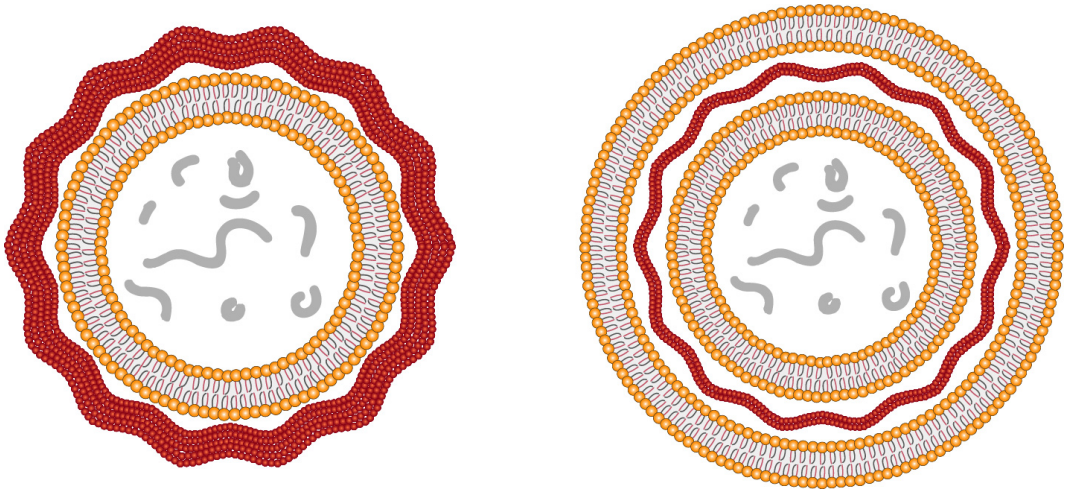
Στην οριζόντια μετάδοση, η σχέση ποιότητας ξενιστή - μικροβίου μειώνεται, διότι οι μικροοργανισμοί που μεταφέρονται αργότερα στη ζωή, αντιμετωπίζουν μεγαλύτερη δυσκολία ενσωμάτωσης στη χλωρίδα. Η οριζόντια μετάδοση, λαμβάνει χώρα μέσω διαφορετικών τρόπων, όπως οι κοινωνικές αλληλεπιδράσεις και η συμβίωση. Κατά την επαφή με διαφορετικούς ανθρώπους και μέρη, η έκθεση αυξάνεται. Σε κάθε σπίτι υπάρχει ένα μοναδικό αποτύπωμα μικροβιώματος, από τα άτομα που κατοικούν σε αυτό και αν συνυπολογιστούν και ενδεχομένως κατοικίδια στο νοικοκυριό, τότε η έκθεση σε μικροοργανισμούς είναι πολύ μεγαλύτερη, και αυτό δεν είναι απαραίτητα κάτι κακό (Tung *et al.*, 2015; Song *et al.*, 2013).

1.1.1. Ταξινομικεσ Βαθμιδεσ Βακτηριων

Τα βακτήρια αντιπροσωπεύουν την πλειοψηφία των μικροοργανισμών που ζουν στο ανθρώπινο σώμα και είναι ευρέως γνωστό, πως η ποσότητα των βακτηριακών κυττάρων υπερτερεί έναντι των ανθρώπινων. Τα βακτήρια είναι μικροσκοπικοί οργανισμοί διαφόρων σχημάτων και μεγεθών και ζουν σε διάφορα μέρη του σώματος, όπως αναφέρθηκε και νωρίτερα. Σύμφωνα με έρευνες, το συνολικό ανθρώπινο μικροβίωμα, μπορεί να περιέχει ακόμα και 10 φορές περισσότερα κύτταρα, σε σχέση με το συνολικό αριθμό των ανθρώπινων κυττάρων (Sender *et al.*, 2016), καθώς επίσης, περιέχει περίπου 400 φορές πιο μοναδικά γονίδια, σε σχέση με το ανθρώπινο γονιδίωμα (X. C. Morgan *et al.*, 2013). Με τον τρόπο αυτό, το μικροβίωμα, ή αλλιώς το *μεταγονιδίωμα*, είναι ικανό να αποτελέσει σημαντικό στοιχείο για πρόσθετη γενετική ποικιλομορφία, συμβάλλοντας μαζί με το γονιδίωμα στον καθορισμό των φαινοτυπικών αποκρίσεων (E. A. Grice and J. A. Segre, 2012). Τα νούμερα αυτά, είναι αρκετά αυξημένα και χρήζουν ιδιαίτερης αναφοράς, δεδομένου ότι το μέσο μέγεθος ενός προκαρυωτικού γονιδιώματος, είναι μικρότερο από το μέγεθος των ευκαρυωτικών γονιδιωμάτων, σε κλίμακα που μπορεί να φτάσει ακόμα και τις 3 τάξεις μεγέθους. Η διαφορά αυτή, σε συνδυασμό με τα νούμερα που αναφέρθηκαν παραπάνω για τον αριθμό των μικροβιακών κυττάρων και μοναδικών γονιδίων, μπορούν να ερμηνευθούν, λαμβάνοντας υπόψιν τις πολυάριθμες ταξινομικές ομάδες των βακτηρίων, που εμπεριέχονται σε ένα τυπικό μικροβιακό δείγμα.

Με τον όρο «ταξινομικές ομάδες», εννοείται η κατηγοριοποίηση των βακτηρίων, λαμβάνοντας υπόψιν διάφορα χαρακτηριστικά τους. Τέτοια είναι, το σχήμα, το μέγεθος, η λειτουργία που επιτελούν, το περιβάλλον καλλιέργειας που είναι ευνοϊκό για αυτά, τα προϊόντα που παράγουν με τις διάφορες μεταβολικές οδούς, την ανθεκτικότητά τους σε διάφορες καταστάσεις, αλλά και αν είναι θετικά ή αρνητικά κατά τη χρώση τους με τη μεθοδολογία Gram (Madigan *et al.*, 2006). Η τελευταία, επιτυγχάνει το διαχωρισμό των βακτηρίων σε δύο μεγάλες κατηγορίες: τα θετικά και τα αρνητικά κατά Gram βακτήρια. Ο διαχωρισμός αυτός, βασίζεται στη δομή του εκάστοτε κυτταρικού τοιχώματος, αναφορικά με το εάν υπάρχει ή όχι η πρωτεΐνη *πεπτιδογλυκάνη* σε αυτό, μία δομική πρωτεΐνη της εξωτερικής μεμβράνης πλάσματος, η οποία σχηματίζει μία άκαμπτη προστατευτική στοιβάδα, που διατηρεί τη σταθερότητα των κυττάρων. Κατ΄ αυτό τον τρόπο, διασφαλίζει αντοχή απέναντι σε διάφορες καταστάσεις και τη μη λύση των κυττάρων (Vollmer *et al.*, 2008). Η βασική διαφορά των θετικών και αρνητικών κατά Gram βακτηρίων, έγκειται στη διαφοροποίηση της κυτταρικής μεμβράνης, με τα θετικά κατά Gram βακτήρια να περιέχουν ως εξωτερικό κάλυμμα, μία παχιά στρώση της *πεπτιδογλυκάνης*, ενώ τα αρνητικά κατά Gram βακτήρια, μία αρκετά λεπτότερη στρώση αυτής της πρωτεΐνης, η οποία αυτή τη φορά δεν είναι εξωτερική, αλλά είναι εσωτερική σε αρκετές

στρώσεις κυτταρικής μεμβράνης (Σχήμα 1.2). Το γεγονός αυτό, προσθέτει περισσότερη αντοχή στα αρνητικά κατά Gram βακτήρια, λόγω του ότι ουσίες, όπως τα διάφορα απορρυπαντικά, τοξίνες ή αντιβιοτικά, δεν έχουν την ίδια δυνατότητα να εισαχθούν στο εσωτερικό των κυττάρων, αποικοδομώντας τα πολλαπλά επίπεδα μεμβρανών, σε σχέση με τα θετικά κατά Gram βακτήρια, στα οποία το κάλυμμα πεπτιδογλυκάνης διασπάται εύκολα.



Σχήμα 1.2: *Διαφορές της μεμβράνης πλάσματος, των θετικών και αρνητικών κατά Gram βακτηρίων. Αριστερά απεικονίζεται ένα θετικό κατά Gram βακτήριο, όπου ένα μεγάλο στρώμα πεπτυδογλυκάνης καλύπτει το εξωτερικό του. Δεξιά, απεικονίζεται ένα αρνητικό κατά Gram βακτήριο, όπου υπάρχει εναλλαγή ανάμεσα στις μεμβρανικές στοιβάδες και στην λεπτή στρώση πεπτυδογλυκάνης.*

Αφού αναφέρθηκαν αρκετές από τις κύριες κατηγορίες διαχωρισμού των βακτηρίων, πλέον μένει ο χαρακτηρισμός των ομάδων βακτηρίων, ανάλογα με αυτά τα χαρακτηριστικά. Αρχικά, όπως είναι προφανές όλα τα βακτήρια, όπως και όλοι οι οργανισμοί, ανήκουν στο ευρύτερο πλαίσιο των έμβιων οργανισμών. Αυτό το επίπεδο, διαχωρίζεται ανάλογα με την επικράτεια/είδος (Domain) του κάθε οργανισμού, στους Ευκαρυωτικούς, Προκαρυωτικούς και στους Αρχαίους οργανισμούς. Στη συνέχεια, η επικράτεια ανάλογα με τον οργανισμό, χωρίζεται και αυτή σε υποκατηγορίες με την ονομασία «Βασίλειο» (Kingdom) (C R Woese *et al.*, 1990). Στη συγκεκριμένη περίπτωση, μόνο οι Ευκαρυωτικοί οργανισμοί χωρίζονται σε κατηγορίες, που περιλαμβάνουν τους ευκαρυωτικούς μύκητες, τα φυτά, τα ζώα, και τα Πρώτιστα, όπου τα τελευταία αναφέρονται στα εμπύρηνα κύτταρα, όπως οι οργανισμοί φαιοφύκη, ροδοφύκη και άλλοι. Η ανάλυση συνεχίζεται με την κατηγορία της συνομοταξίας (Phylum), όπου οι οργανισμοί πλέον διακρίνονται ανάλογα με τη φυλογενετική τους ταυτότητα, όπως είναι οι κατηγορίες των *Actinobacteria*, *Acidobacteria* και άλλες. Συνήθως, προστίθεται η κατάληξη `-archaeota` για τα αρχαία και η κατάληξη `-bacteria` για τα βακτήρια. Οι κατηγορίες αυτές, χωρίζονται ανά ομοταξία (Class), δηλαδή εξειδικεύονται περαιτέρω και συνήθως προστίθεται η κατάληξη `-ia` στην λατινική ονομασία της συνομοταξίας. Ακολουθεί ο προσδιορισμός της τάξης για κάθε μία ομοταξία (Order), όπου πλέον η κατάληξη είναι η `-ales` και αφετέρου η οικογένεια (Family), με κατάληξη `-aceae` της ονομασίας ομοταξίας. Τέλος, το γένος (Genus) με κατάληξη `-ae` και το είδος (Species), προσθέτουν ακόμα περισσότερη εξειδίκευση στις ονομασίες των βακτηρίων, φτάνοντας στις τελικές και πιο στοχευμένες ονομασίες, οι οποίες είναι συνήθως δυαδικές (π.χ. *Marinithermus hydrothermalis*) (C R Woese *et al.*, 1990). Ο πίνακας 1.1 δείχνει μία τυπική ταξινόμηση του βακτηρίου *Telmatobacter*.

1.1.2. Ποικιλομορφια Μικροβιωματος

Το πιο πολυσυζητημένο περιβάλλον αποίκισης μικροοργανισμών είναι το έντερο, καθώς εμφανίζει μια αξιοσημείωτη ποικιλομορφία μεταξύ των ανθρώπινων και μικροβιακών κυττάρων, που επισκιάζει την ποικιλομορφία κάθε άλλου οργάνου. Πιο αναλυτικά, υπολογίζεται ότι 150 με 400 μικροβιακά είδη κατοικούν σε αυτό (Lloyd-Price *et al.*, 2016). Έτσι, το παχύ έντερο φιλοξενεί μακράν τη μεγαλύτερη μικροβιακή βιομάζα, με 10^{11} μικροβιακά κύτταρα

Telmatobacter	
Βασίλειο	Βακτήρια
Συνομοταξία	Acido**bacteria**
Ομοταξία	Acidobacter**iia**
Κλάση	Acidobacter**iales**
Οικογένεια	Acidobacteri**aceae**
Γένος	Telmatobact**er**
Είδος	Telmatobacter **bradus**

Πίνακας 1.1: *Ταξινόμηση του βακτηρίου Telmatobacter ανάλογα με την ιεραρχία που αναλύθηκε. Στον πίνακα φαίνονται με έντονη επισήμανση, οι καταλήξεις στην ονομασία για κάθε ομάδα, καθώς και η δυαδική ονομασία του είδους*

ανά μιλιλίτρο, σε σχέση με το λεπτό έντερο, το οποίο περιέχει το πολύ έως 10^8 κύτταρα ανά ml (Walter and R. Ley, 2011). Σε συνδυασμό με τις τεχνικές των αναερόβιων καλλιεργειών και την αλληλούχιση του γονιδίου της 16S ριβοσωμικής μονάδας (16S *r*RNA), οι ιδιότητες του εντέρου έχουν οδηγήσει σε μια έντονη εστίαση στη βιβλιογραφία, σχετικά με το βακτηριακό μικροβίωμα (Bik *et al.*, 2006). Να αναφερθεί επίσης, πως το γονίδιο της 16S ριβοσωμικής μονάδας, αποτελεί ένα αρκετά συντηρημένο γονίδιο, ανάμεσα στα διάφορα βακτηριακά είδη (Carl R. Woese and Fox, 1977). Έτσι, η αλληλούχιση αυτού του γονιδίου, εξάγει πληροφορίες σχετικά με τη συγκέντρωση και ποικιλομορφία των βακτηρίων σε ένα δείγμα, καθώς λειτουργεί ως ταυτότητα κάθε βακτηριακού γένους, παρέχοντας πληροφορία για κάθε είδος και κάνοντας διάκριση αυτών των ειδών (Kolbert and Persing, 1999; Eren *et al.*, 2013). Περισσότερα από 1000 βακτηριακά είδη, έχουν πλέον χαρακτηριστεί και μέσω της μοριακής φυλογενετικής ανάλυσης, πολλά από αυτά έχουν αναταξινομηθεί τα τελευταία 20 χρόνια. Τα είδη εντός των *Bacteroides*, που προηγουμένως θεωρούνταν ως το πιο άφθονο βακτηριακό γένος στο έντερο, έχει πλέον ταξινομηθεί σε πέντε γένη *(Prevotella, Paraprevotella, Alistipes, Parabacteroides* και *Odoribacter)* (Shapira, 2016). Σε ποσοστό άνω του 90% τα *Bacteroidetes* και *Firmicutes*, φαίνεται να κυριαρχούν σε ένα υγιές εντερικό μικροβίωμα, αν και ακόμη το επίπεδο ταξινόμησης είναι υπό εξέταση, δεδομένου ότι τα υγιή άτομα ποικίλουν κατά περισσότερο από μια τάξη μεγέθους στις *Bacteroidetes/Firmicutes* αναλογίες τους. Μέσω μοριακών τεχνικών, έχουν εντοπιστεί διαδεδομένα βακτήρια στα κόπρανα, διευρύνοντας τη λίστα.

Άλλα είδη είναι τα *Proteobacteria, Fusobacteria, Actinobacteria & Verrucomicrobia*. Οι *Lactobacillus* και το είδος *Bifidobacterium bifidum*, που ανήκουν στην οικογένεια των *Firmicutes* και *Actinobacteria*, αντίστοιχα, έχει φανεί πως είναι ωφέλιμα για την υγεία σε αντίθεση με άλλα είδη (Shapira, 2016). Πιο συγκεκριμένα, *Firmicutes* είδη όπως το *Clostridium perfringens* και ο *Staphylococcus aureus*, μπορούν να προκαλέσουν λοίμωξη εάν αυξηθεί πάνω από μία συγκέντρωση ο πληθυσμός τους. Επίσης, κάποιοι από τους ευρέως γνωστούς παθογόνους μικροοργανισμούς, όπως οι *Escherichia coli, Salmonella, Shigella, Helicobacter & Enterobacter*, ανήκουν στην ομάδα των Πρωτεοβακτηρίων *(Proteobacteria)*.

Η στοματική κοιλότητα έχει και αυτή διαφορετικό μικροβίωμα, παρόμοιο με του εντέρου και κυριαρχείται από τα *Streptococcus spp*, δηλαδή τα βακτήρια που ανήκουν στο γένος των στρεπτόκοκκων. Αντίθετα, το δέρμα διαφέρει κυρίως ως προς τις τοπικές ιδιότητες του. Μια ανασκόπηση από δημοσιεύσεις στο περιοδικό Nature, το 2018 αναφέρει πως οι πληθυσμοί των βακτηρίων ποικίλουν, ανάλογα με την περιοχή του δέρματος και είναι άκρως εξαρτώμενοι από παράγοντες, όπως η υγρασία του δέρματος και η ποσότητα του σμήγματος. Στις σμηγματογόνες περιοχές, κατοικεί το *Propionibacterium*, ενώ τα βακτήρια που ευδοκιμούν στα υγρά περιβάλλοντα ανήκουν στα είδη *Staphylococcus* και *Corynobacterium*, συμπεριλαμβανομένων περιοχών των ποδιών και των αγκώνων (Jack *et al.*, 2018).

Ένας μικρός αριθμός αρχαίων βακτηρίων έχει εντοπιστεί στο ανθρώπινο μικροβίωμα και κυρίως στο έντερο. Το πιο διαδεδομένο είδος, είναι το γένος *Methanobrevicacter*. Αυτό, έχει φανεί να είναι το πιο προσαρμοσμένο είδος στο έντερο, βελτιστοποιώντας την πέψη των διατροφικών πολυσακχαριτών από άλλα μικρόβια, και με τον τρόπο

αυτό, προσαρμόζει τη γονιδιακή του έκφραση παρουσία κοινών βακτηρίων όπως το *Bacteroides thetaiotaomicron*. Τέλος, όσον αφορά τους ιούς, είναι ως το πλείστων βακτηριοφάγοι (Gregory *et al.*, 2019; Manrique *et al.*, 2016). Από τους ευκαρυωτικούς μικροοργανισμούς, οι πιο γνωστοί στο ανθρώπινο μικροβίωμα είναι κυρίως οι μύκητες (Fiers *et al.*, 2019) και τα Πρώτιστα (Eckburg *et al.*, 2005; Scanlan and Marchesi, 2008), οι οποίοι έχουν τυπικά παθογόνο λειτουργία.

1.1.3. Μικροβιωμα και Διατηρηση της Ανθρωπινης Υγειας

Τα κυτταρικά στοιχεία του μικροβιώματος, εκτελούν σημαντικές λειτουργίες για τη διατήρηση της ομοιόστασης του οργανισμού. Μπορούν να ενισχύσουν την ανοσία ή να εμποδίσουν λοιμώξεις από παθογόνα βακτήρια, ενώ ορισμένα μικρόβια μπορεί να αποικίζουν στον ξενιστή χωρίς να προκαλούν ασθένεια, αλλά κάποια από αυτά μπορεί να γίνουν δυνητικά παθογόνα και να προκαλέσουν λοιμώξεις σε ανοσοκατεσταλμένους ξενιστές. Αυτό για να γίνει, θα πρέπει η ισορροπία του αριθμού των διαφόρων μικροοργανισμών που απαρτίζουν τη μικροχλωρίδα να διαταραχθεί και να υπάρξει υπερπληθυσμός από ένα είδος. Συνεπώς, η ισορροπία της μικροχλωρίδας είναι ζωτικής σημασίας.

Το μικροβίωμα του εντέρου εμπλέκεται σε πολλές βιολογικές διεργασίες. Όπως αναφέρθηκε νωρίτερα, ο γαστρεντερικός σωλήνας είναι η πιο διευρυμένη μικροβιακή κοινότητα. Το 97% των βακτηρίων που κατοικούν σε αυτόν είναι αυστηρά αναερόβια, ενώ το υπόλοιπο 3% αποτελούν τα αερόβια βακτήρια (Noverr and Huffnagle, 2004). Αξίζει να σημειωθεί πως υπάρχουν μεγάλες διαφορές στο μικροβιακό φορτίο στις διάφορες περιοχές του γαστρεντερικού σωλήνα. Ενδεικτικά, ο βλεννογόνος του λεπτού εντέρου κυριαρχείται από *Bacteroidetes*, ο αυλός περιέχει βακτήρια της οικογένειας *Enterobacteriaceae*, ενώ το *Helicobacter Pylori* εντοπίζεται στο στομάχι. Η μεγαλύτερη βιομάζα βρίσκεται στο παχύ έντερο, το οποίο περιέχει είδη από τα *Bacteroidetes* και *Firmicutes*, ενώ τα Proteobacteria και τα Actinobacteria αντιπροσωπεύονται πολύ λιγότερο (R. E. Ley *et al.*, 2006; Martín *et al.*, 2014).

Η μικροχλωρίδα του εντέρου και ο ξενιστής, έχουν μια σχέση άκρως αλληλοεξαρτώμενη. Η χλωρίδα, αφενός υποστηρίζει τον ξενιστή, ενισχύοντας τον μεταβολισμό και αφετέρου την ωρίμανση του ανοσοποιητικού και την προστασία του γαστρεντερικού σωλήνα από τα παθογόνα βακτήρια. Παράλληλα, ο ξενιστής παρέχει σε αυτή πλούσιο σε θρεπτικά συστατικά και φιλόξενο περιβάλλον (Gill *et al.*, 2006; Noverr and Huffnagle, 2004).

Η μικροχλωρίδα εμπλέκεται σε μια πληθώρα βασικών διεργασιών όπως η ρύθμιση του μεταβολισμού και η έμφυτη ανοσία. Ακόμα, παρέχει προστασία από εξωτερικά παθογόνα βακτήρια, μέσω παραγωγής αντιμικροβιακών παραγόντων όπως είναι οι βακτηριοσίνες. Πολλές μελέτες έχουν αποδείξει το σημαντικό ρόλο της μικροχλωρίδας του εντέρου, στην ενίσχυση της ικανότητας εξαγωγής ενέργειας από τα τρόφιμα, την αύξηση των θρεπτικών συστατικών, την αλλαγή του σήματος της όρεξης, την παραγωγή βιταμινών αλλά και το μεταβολισμό πολλών υλικών, επειδή περιέχει μια ποικιλία ενζύμων και βιοχημικών οδών. Η γενετική ποικιλότητα που βρίσκεται στη μικροχλωρίδα του εντέρου, επιτρέπει την πέψη ενώσεων μέσω μεταβολικών οδών, όπως τα σακχαρολυτικά και τα πρωτεολυτικά μεταβολικά μονοπάτια, διαδικασίες που πραγματοποιούνται αποδοτικά από τα βακτήρια (Sylvia H Duncan *et al.*, 2002). Αξίζει να σημειωθεί, ότι τα μονομερή σάκχαρα που μετατρέπονται σε λιπαρά οξέα μακράς αλύσου (Long Chain Fatty Acids (LCFAs)), μέσω μικροβιακής ζύμωσης άπεπτων τροφίμων, πέραν της βασικής τους λειτουργίας ως κύρια πηγή ενέργειας των κυττάρων του εντέρου, φαίνεται πως μειώνουν τον κίνδυνο ανάπτυξης γαστρεντερικών διαταραχών, μεταβολικών συνδρόμων, ακόμα και του καρκίνου (Wong *et al.*, 2006).

Αναφορικά με το μικροβίωμα του δέρματος, στις μέρες μας, είναι πλέον αποδεκτό πως μια ισορροπημένη μικροχλωρίδα του ανθρώπινου δέρματος είναι ευεργετική, καθώς προστατεύει από λοιμώξεις του δέρματος και άλλες δερματικές διαταραχές (Rosenthal *et al.*, 2011). Ωστόσο, λαμβάνοντας υπόψη τις παραλλαγές στη μικροβιακή ποικιλότητα, ο ορισμός της ισορροπημένης μικροχλωρίδας του δέρματος είναι δύσκολος. Παρόλα αυτά, τα τελευταία χρόνια έχει υπάρξει εξαιρετικά σημαντική πρόοδος. Για παράδειγμα, το ∼pH 5 του δέρματος, αντιπροσωπεύει τη

βασικότερη γραμμή άμυνας κατά των παθογόνων μικροοργανισμών. Το *Propionibacterium Acnes* που κατοικεί κάτω από αναερόβιες συνθήκες στους σμηγματογόνους αδένες, απελευθερώνει τα λιπαρά οξέα από τους αδένες στο δέρμα, συμβάλλοντας στον καθορισμό του όξινου pH (E. Grice and J. Segre, 2011).

Σε μία έρευνα, οι (Lai *et al.*, 2010) έδειξαν πώς το μικροβίωμα του δέρματος μπορεί να τροποποιήσει τις φλεγμονώδεις αντιδράσεις, μέσω των Toll like receptors (TLRs). Το λιποτειχοικό οξύ στο κυτταρικό τοίχωμα του *Staphylococcus Epidermidis*, έχει αποδειχθεί ότι προλαμβάνει τη δερματική φλεγμονή, μέσω αναστολής των ανοσοαποκρίσεων που βασίζονται στον TLR2 και μεσολαβεί στην αναστολή των φλεγμονώδων κυτοκινών από τα κερατινοκύτταρα. Ο *S. epidermidis* φαίνεται να παράγει αντιμικροβιακά πεπτίδια, που μπορεί να αναστείλουν την ανάπτυξη παθογόνων μικροοργανισμών του δέρματος, όπως είναι ο *Staphylococcus Aureus*, αλλά και ο *group A Streptococcus*, μέσω ενίσχυσης της γονιδιακής έκφρασης αντιμικροβιακών πεπτιδίων. Συνοπτικά, αποδείχθηκε πως ο *Staphylococcus Epidermidis* και το *Propionibacterium Acnes*, ενισχύουν σημαντικά τη δερματική ανοσία, λειτουργώντας ως τοπικό σύστημα ανοσοεπιτήρησης, πέρα της λειτουργίας τους στην εντερική οδό.

1.1.4. Διαταραχες της Μικροχλωριδας

Οποιαδήποτε μεταβολή της ισορροπίας του μικροβιώματος, οδηγεί σε δυσβίωση και αυτό με τη σειρά του επιφέρει μια σειρά από ανεπιθύμητες αντιδράσεις, όπως είναι η εμφάνιση αυτοάνοσων νοσημάτων (Chu *et al.*, 2017), χρόνιων ασθενειών, ψυχιατρικών διαταραχών (Zheng *et al.*, 2019), καρδιαγγειακών (Jie *et al.*, 2017) και μεταβολικών παθήσεων (Karlsson *et al.*, 2013), μέχρι και πιο σοβαρών ασθενειών, όπως ο καρκίνος. Υπάρχουν διάφοροι παράγοντες που μπορούν να αλλάξουν την σύνθεση και τη λειτουργία του μικροβιώματος του εντέρου. Μεταξύ αυτών των παραγόντων είναι η γενετική προδιάθεση, η επιγενετική, η ηλικία, ο τρόπος γέννησης, η διατροφή του ξενιστή και η χρήση αντιβιοτικών και ισχυρών παρεμβατικών φαρμακευτικών σκευασμάτων (Nagpal *et al.*, 2017).

1.1.4.1. Μεταβολικες Ασθενειες

Όπως αναφέρθηκε νωρίτερα, η απώλεια της ομοιόστασης προκαλεί διάφορες ασθένειες. Έτσι λοιπόν, οι αλλαγές της γαστρεντερικής μικροχλωρίδας και του ανοσοποιητικού συστήματος, μπορεί να προκαλέσει διάφορες μεταβολικές ασθένειες, όπως ο σακχαρώδης διαβήτης τύπου ΙΙ, η παχυσαρκία και οι καρδιοπάθειες (Manco *et al.*, 2010).

Η παχυσαρκία είναι μια πολυπαραγοντική διαταραχή, που εμπλέκονται τόσο η γενετική προδιάθεση, οι ορμονικές διαταραχές, ο τρόπος ζωής και οι φαρμακευτικές παρεμβάσεις, όσο και το ίδιο το μικροβίωμα του εντέρου. Μελέτες υποστηρίζουν, πως η αύξηση στο βάρος δεν είναι ποσοτικό αποτέλεσμα υπερβολικής κατανάλωσης τροφής, αλλά ποιότητας του μεταβολισμού (Raoult, 2008). Παρατηρήθηκε επίσης, ότι σε δίαιτες υψηλής περιεκτικότητας σε λιπαρά, ο λιποπολυσακχαρίτης (LPS) μέσα στον αυλό του εντέρου παχύσαρκων ατόμων, διεγείρει τη φλεγμονή. Επιπρόσθετα, δίαιτα με αυξημένη κατανάλωση λιπαρών, συμβάλει και στην ανισοκατανομή της μικροχλωρίδας του εντέρου. Σε μια έρευνα σε ποντίκια φυσιολογικού βάρους, η λιπαρή δίαιτα οδήγησε σε αύξηση των *Firmicutes* και μείωση των *Bacteroidetes*, ενώ παράλληλα η ίδια παρατήρηση έγινε και σε γενετικά παχύσαρκα ποντίκια (Sylvia H. Duncan *et al.*, 2007; Sekirov *et al.*, 2010).

1.1.4.2. Γαστρεντερικες Ασθενειες

Η εντερική μικροχλωρίδα συνήθως ζει μέσα στον ξενιστή με συναινετικό τρόπο. Ωστόσο, εξωτερικοί παράγοντες μπορούν να αλλάξουν την ισορροπία στη σύνθεση της μικροχλωρίδας και δυνητικά αυτό να οδηγήσει σε ασθένειες. Μερικοί από αυτούς τους παράγοντες, είναι οι παθογόνοι μικροοργανισμοί, συμπεριλαμβανομένων των βακτηρίων όπως η *Salmonella, Shigella, Escherichia, Campylobacter, Yersinia* και άλλα γένη, πρωτόζωα όπως η αμοιβάδα,

ιοί και πολλοί άλλοι οργανισμοί (Tsolis *et al.*, 2008). Από όλα τα παραπάνω, αξίζει να σημειωθεί πως οι λοιμώξεις από το γένος *Salmonella*, αποτελούν παγκόσμια απειλή για τη δημόσια υγεία, σύμφωνα με το κέντρο επιτήρησης της Σαλμονέλας, από το Αμερικανικό CDC (Centers for Disease Control and Prevention (CDC)). Το γένος αυτό, χωρίζεται σε δύο είδη, τα *Salmonella Enterica* και *Salmonella bongori*.

Η σαλμονέλωση μπορεί να εκδηλωθεί σε διάφορα σύνδρομα όπως φλεγμονή, εντερικό πυρετό και γαστρεντερίτιδα από σαλμονέλα (S. L. Foley and Lynne, 2008). Η τελευταία, είναι η κυρίαρχη μορφή σαλμονέλωσης και χαρακτηρίζεται από μια πληθώρα κλινικών χαρακτηριστικών όπως οι κράμπες στο στομάχι, εμετό, διάρροια και σχετίζεται συχνότερα με τη κατανάλωση μολυσμένων τροφίμων (Howard *et al.*, 2012; Shi *et al.*, 2019).

Είναι γνωστό, ότι το εντερικό επιθήλιο αποτελεί προστατευτικό φράγμα έναντι βακτηριακών λοιμώξεων, ωστόσο η γενετική της *Salmonella*, παίζει σημαντικό ρόλο στην ανάπτυξη και επιβίωση της. Η πλαστικότητα των βακτηριακών γονιδιωμάτων στα είδη *Salmonella*, είναι ιδιαίτερα γνωστή και μπορεί να επιτευχθεί με τη παρουσία των πλασμιδίων, τα οποία στο συγκεκριμένο οργανισμό, κατέχουν μείζονα ρόλο, στην ικανότητα του να επιβιώνει σε διάφορες ζωικές πηγές τροφίμων, να προκαλεί νόσο στον άνθρωπο και να αποφεύγει στρατηγικές θεραπείας (Jakočiūnė *et al.*, 2014). Η παρουσία πλασμιδίου επίσης, προκαλεί την έκφραση γονιδίων λοιμογόνου δράσης και αντίστασης προς τη μικροχλωρίδα, επιτρέποντας την αποίκιση (Steven L. Foley *et al.*, 2013; Han *et al.*, 2012). Η ικανότητα της σαλμονέλας και άλλων εντερικών παθογόνων βακτηρίων, να εισβάλει στο γαστρεντερικό σύστημα, είναι υψηλή όταν η μικροχλωρίδα του παχέος εντέρου είναι λιγότερο σταθερή, λόγω του υψηλού αριθμού των Πρωτεοβακτηρίων (Kolling *et al.*, 2012).

1.1.4.3. Αυτοανοσα Νοσηματα και Καρκινος

Το εντερικό επιθήλιο και η βλέννη που υπάρχει σε αυτό, αποτελούν τον εντερικό φραγμό που λειτουργεί ως αμυντικός μηχανισμός του εσωτερικού περιβάλλοντος του ξενιστή. Η λειτουργία του φραγμού, ρυθμίζεται από αλληλεπιδράσεις της εντερικής μικροχλωρίδας-ξενιστή μέσω ισορροπίας των γαστρεντερικών Τ-λεμφοκυττάρων (Tregs/TH17). Η ισορροπία αυτή, είναι ζωτικής σημασίας για την εντερική ομοιόσταση, συνεπώς η διαταραγμένη λειτουργία του μπορεί να αυξήσει τη διαπερατότητα του εντέρου σε μικροβιακά προϊόντα και σε άλλα συστατικά, συμβάλλοντας σε ανοσολογικές αποκρίσεις, όπως αλλεργία, φλεγμονή και άλλες ανοσολογικές διαταραχές (Barnaba and Sinigaglia, 1997). Πολλά βακτήρια όπως το *Bacteroides fragilis, Bifidobacterium infantis* και *Firmicutes*, έχει φανεί πως μπορούν να επάγουν την επέκταση των FOXP3 Τ-ρυθμιστικών κυττάρων (Tregs), που παράγουν κυτοκίνες όπως η IL-10, για την καταστολή της φλεγμονής (El Aidy *et al.*, 2012; Lawley and Walker, 2013).

Η δυσβίωση έχει κατηγορηθεί και ως παράγοντας κινδύνου για αυτοάνοσες διαταραχές όπως η κοιλιοκάκη, ωστόσο θεωρείται μια πολυπαραγοντική χρόνια διαταραχή με γενετικό υπόβαθρο, που χαρακτηρίζεται για τη δυσανεξία στη γλουτένη και σε προλαμίνες (Sanz *et al.*, 2011). Σε όλες τις φάσεις νόσου (ενεργή ή σε ύφεση), έχει παρατηρηθεί σημαντική μείωση στην αναλογία Gram+/Gram- βακτηρίων, με αξιοσημείωτη μείωση στα βακτήρια της συνομοταξίας Bifidobacteriales, που προάγουν την υγεία, καθώς και αύξηση στα μολυσματικά Gram αρνητικά βακτήρια, αντανακλώντας τον ρόλο τους στη κοιλιοκάκη (Marasco *et al.*, 2016; De Palma *et al.*, 2010).

Πολυάριθμες μεταγονιδιωματικές μελέτες αλληλούχισης, έχουν αποκαλύψει σημαντικές διαφορές στη σύνθεση μικροβιακών κοινοτήτων μεταξύ υγιών και ασθενών ατόμων. Ως συνέπεια, το μικροβίωμα έχει εμπλακεί τόσο στην πρόκληση όσο και στη πρόληψη μιας ποικιλίας ασθενειών όπως αναφέρθηκαν νωρίτερα, συμπεριλαμβανομένου και του καρκίνου (Weinstock, 2012; Goodrich *et al.*, 2014). Η δυσβίωση μπορεί να αυξήσει την παραγωγή επιβλαβών μεταβολιτών και αντιγόνων από τη μικροχλωρίδα, οδηγώντας σε ανοσολογικές αποκρίσεις. Οι διαταραχές αυτές σχετίζονται ιδιαίτερα με την ογκολογία, λαμβάνοντας υπόψη ότι βασικά χαρακτηριστικά του καρκίνου αποτελούν η φλεγμονή και ο απορρυθμισμένος μεταβολισμός (Hanahan and Weinberg, 2011).

Τα παθογόνα βακτήρια προάγουν την ανάπτυξη καρκίνου μέσω γενετικών μηχανισμών. Πιο συγκεκριμένα με

βάση τον Διεθνή Οργανισμό έρευνας για τον καρκίνο (IARC), υπάρχουν 10 βιολογικοί παράγοντες που έχουν χαρακτηριστεί ως καρκινογόνοι. Ένας από αυτούς είναι το *Helicobacter pylori*, που κατοικεί στον γαστρικό βλεννογόνο και συνδέεται στενά με γαστρικό αδενοκαρκίνωμα, ως αποτέλεσμα χρόνιας φλεγμονής με γαστρίτιδα (Marshall and Warren, 1984). Ο μηχανισμός με τον οποίο το βακτήριο αυτό προκαλεί καρκίνο στομάχου αποδίδεται στην παρουσία κυτταροτοξινών και στην έκκριση παραγόντων λοιμογόνου δράσης, που προωθούν τη χρόνια φλεγμονή και κατ' επέκταση προάγουν γενετική αστάθεια στον ξενιστή, οξειδωτικό στρες, και καρκινογένεση. Το βακτήριο αυτό, χρησιμοποιεί τύπου IV σύστημα έκκρισης για να μετατοπίσει τις σχετιζόμενες με το *CagA* γονίδιο κυτταροτοξίνες στα επιθηλιακά κύτταρα, τα οποία ρυθμίζουν την β-κατενίνη προς αύξηση της τάσης για γαστρικό καρκίνο. Ωστόσο, είναι ενδιαφέρον πως το ίδιο βακτήριο φαίνεται να σχετίζεται με μειωμένο κίνδυνο αδενοκαρκινώματος οισοφάγου και οισοφάγου Barrett (παλινδρόμηση οξέος από το στομάχι προς τον οισοφάγο), επηρεάζοντας το pH του στομάχου και τονίζοντας την πολυπλοκότητα της σχέσης μεταξύ των παθογόνων μικροβιακών επιδράσεων και ξενιστή στη καρκινογένεση (F. Wang *et al.*, 2008; Vaezi *et al.*, 2000).

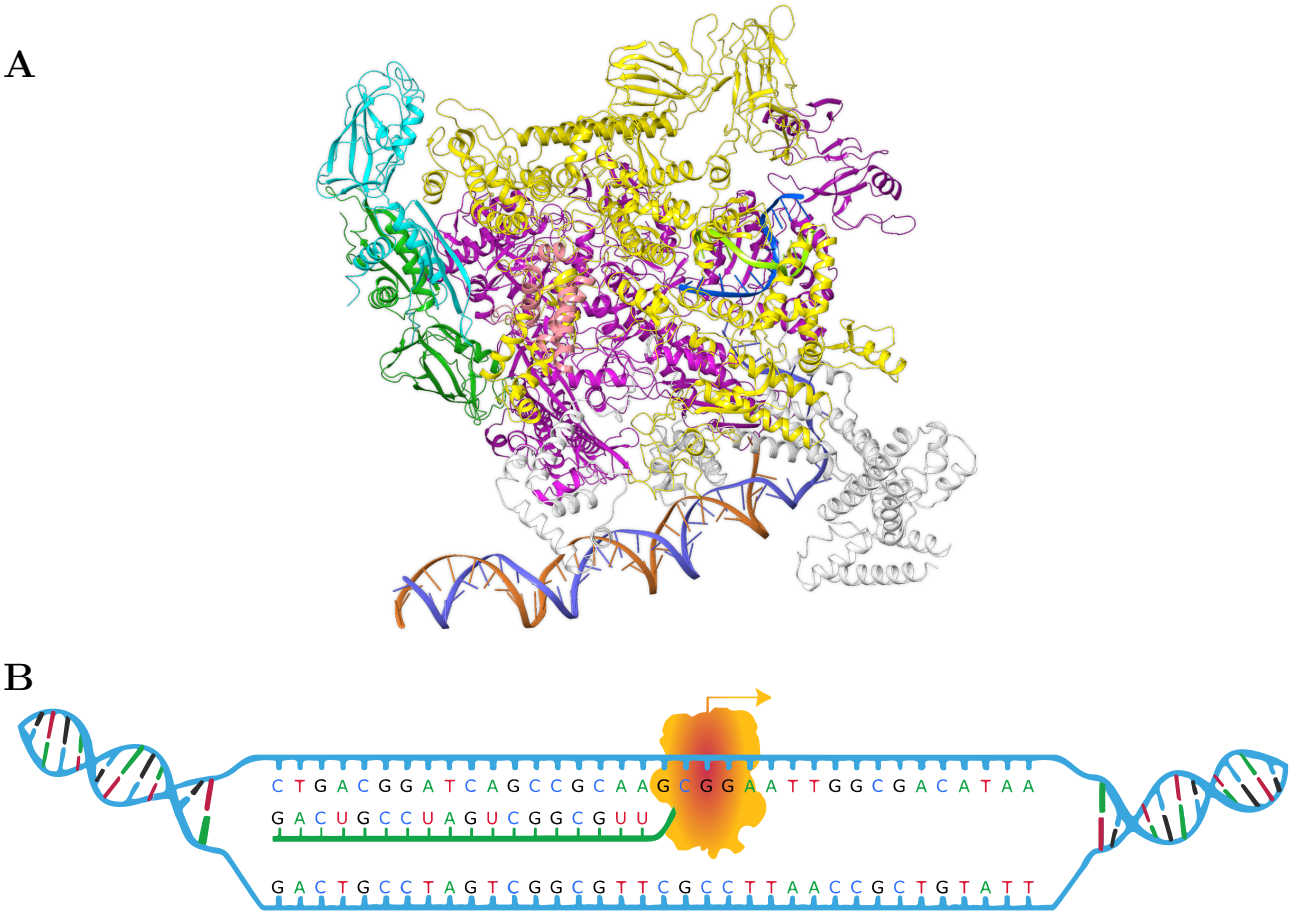
1.1.5. Γονιδιωματικα Στοιχεια και Οργανωση Βακτηριων

Σε αντίθεση με το γονιδίωμα των ευκαρυωτικών κυττάρων, το οποίο εμφανίζει εξαιρετική πολυπλοκότητα και οργάνωση, το βακτηριακό γονιδίωμα είναι δίκλωνο DNA μόριο, συνήθως κυκλικό και οργανώνεται σε ένα μοναδικό χρωμόσωμα. Το χρωμοσωμικό DNA, περιέχει όλα τα κωδικοποιητικά γονίδια, που έχουν κατά μέσο όρο 1kb μήκος (1.000 νουκλεοτιδικά ζεύγη βάσεων), χωρίς ιντρονικές περιοχές. Ένα μεγάλο ποσοστό των κωδικοποιούμενων στοιχείων στα βακτήρια, αφορά ρυθμιστικές αλληλουχίες, όπως τα μη κωδικά RNAs, δηλαδή τα *r*RNAs, *t*RNAs και *s*RNAs. Επιπρόσθετα, περιέχουν μικρά γενετικά στοιχεία, συνήθως κυκλικά που ονομάζονται πλασμίδια, τα οποία αντιγράφονται ανεξάρτητα από το χρωμοσωμικό γενετικό υλικό. Τα γονίδια των πλασμιδίων, προσδίδουν σημαντικές ιδιότητες στο κύτταρο - ξενιστή, όπως η ανθεκτικότητα στα αντιβιοτικά και γενικότερα η ενσωμάτωση εξωγενούς DNA (Madigan *et al.*, 2006).

Στα περισσότερα βακτήρια, η αντιγραφή του DNA ξεκινά από μια μόνο θέση, η οποία αναγνωρίζεται από την πρωτεΐνη εκκίνησης. Όταν οι γονικοί κλώνοι διαχωρίζονται, η DNA Ελικάση και Πριμάση συνδέονται, σχηματίζοντας το σύμπλεγμα του πριμοσώματος. Η DNA ελικάση ξετυλίγει τους κλώνους και η πριμάση συνθέτει τους RNA εκκινητές για τον κάθε κλώνο, ώστε να μπορέσει στη συνέχεια το ένζυμο της αντιγραφής, η DNA πολυμεράση III, να συνθέσει DNA προσθέτοντας νουκλεοτίδια σε κατεύθυνση από το 5' προς το 3' άκρο. Επειδή οι κλώνοι είναι αντιπαράλληλοι, ο ένας ονομάζεται κύριος κλώνος (leading strand) και συντίθεται συνεχόμενα, ενώ ο άλλος κλώνος (lagging strand) συντίθεται αντίθετα από την κατεύθυνσή της ελικάσης, σχηματίζοντας με ασυνεχή τρόπο, τα θραύσματα Okazaki. Όταν η DNA πολυμεράση III φτάσει στο άκρο του τμήματος του DNA που έχει συντεθεί, αντικαθίσταται από την DNA πολυμεράση I, η οποία υδρολύει τον προπορευόμενο RNA εκκινητή και τον αντικαθιστά με DNA. Στη συνέχεια, απομακρύνεται και το ένζυμο DNA Λιγάση, σχηματίζει φωσφοδιεστερικό δεσμό, συνδέοντας τα τμήματα DNA (Madigan *et al.*, 2006).

Η βακτηριακή μεταγραφή, έχει μελετηθεί σχεδόν αποκλειστικά στην *Escherichia coli* και χωρίζεται σε 3 διακριτές φάσεις, την έναρξη, επιμήκυνση και τον τερματισμό. Σε αντίθεση με τα ευκαρυωτικά κύτταρα, τα βακτήρια έχουν μόνο μια RNA πολυμεράση με διαφορετικούς τύπους υπομονάδων, οι οποίες αλληλεπιδρούν μεταξύ τους, σχηματίζοντας το όλοενζυμο, το οποίο απαιτείται για την έναρξη της μεταγραφής. Μία από αυτές τις υπομονάδες είναι και η σ υπομονάδα (Sigma Subunit), η οποία καθορίζει το μοτίβο πρόσδεσης του συμπλόκου, στον εκάστοτε υποκινητή. Για παράδειγμα, η σ^{70} υπομονάδα, η οποία συναντάται ευρέως στην *Escherichia coli*, καθώς και σε ορισμένα άλλα αρνητικά κατά Gram βακτήρια, στοχεύει γονίδια, που είναι απαραίτητα για τη φυσιολογική επιβίωση του κυττάρου (housekeeping factor). Η υπομονάδα σ^{70}, αναγνωρίζει και προσδένεται σε συγκεκριμένες αλληλουχίες στη περιοχή του υποκινητή, περίπου 10 ζεύγη βάσεων (πλαίσιο TATAAT box) και 35 ζεύγη βάσεων (TTGACA) ανοδικά (upstream) του σημείου όπου αρχίζει η μεταγραφή *(Σχήμα 1.3A)*, δηλαδή στις γνωστές ως -10 (Pribnow box) και -35 (Gilbert box) περιοχές, αντίστοιχα (Browning and Busby, 2004). Η Θέση

Έναρξης της Μεταγραφής (Transcription Start Site (TSS)), είναι συγκεκριμένα το πρώτο νουκλεοτίδιο, από όπου ξεκινά η κωδικοποίηση του νεοσυντιθέμενου RNA μορίου. Πολλά γονίδια, μπορεί να έχουν περισσότερες από μία θέσεις έναρξης της μεταγραφής, όπως θα αποδειχθεί στο κύριο μέρος της εργασίας αυτής. Να σημειωθεί, πως ένα μετάγραφο μπορεί να έχει έναν κλώνο με κατ΄ αντιστοιχία όμοια ακολουθία νουκλεοτιδίων με αυτό, με τη σύμβαση ότι οι βάσεις της Ουρακίλης (U), έχουν μετατραπεί σε Θυμίνη (T), επομένως αυτός ο κλώνος χαρακτηρίζεται ως sense κλώνος (sense strand) ή κωδική αλυσίδα (coding strand), ή αλυσίδα μη-πρότυπο (non-teamplate strand). Έτσι, ένα παραγόμενο RNA θα είναι συμπληρωματικό και αντιπαράλληλο με τη μη κωδική (non-coding strand) ή αρνητική αλυσίδα (antisense negative strand), η οποία χρησιμοποιήθηκε ως πρότυπο (template strand) για την παραγωγή αυτού του RNA *(Σχήμα 1.3B)*.



Σχήμα 1.3: *Μεταγραφή στους προκαρυωτικούς οργανισμούς.* ***(A)*** *Αναπαράσταση της στερεοδιάταξης του ανοιχτού συμπλόκου της RNA πολυμεράσης, με τον παράγοντα σ^{70} (RpoD) των βακτηρίων (Εγγραφή με κωδικό 6CA0 στην PDB). Με άσπρο χρώμα φαίνεται η υπομονάδα σ^{70}, με πράσινο και μπλε ανοιχτό χρώμα η υπομονάδα A, με μωβ η υπομονάδα B, με κίτρινο η υπομονάδα B΄, με ροζ η υπομονάδα Ω και με μπλε και πορτοκαλί, οι δύο κλώνοι του DNA, στους οποίους προσδένεται. Η αναπαράσταση δημιουργήθηκε με το λογισμικό Schrödinger BioLuminate (Schrödinger, LLC, 2021a).* ***(B)*** *Επίπεδη αναπαράσταση του τρόπου μεταγραφής. Με πορτοκαλί χρώμα απεικονίζεται η RNA πολυμεράση και με το βέλος υποδηλώνεται η κατεύθυνση της μεταγραφής. Ο κλώνος που μεταγράφεται είναι αυτός που χρησιμοποιείται ως πρότυπο και χαρακτηρίζεται ως μη κωδικός ή antisense, γιατί δε φέρει την ίδια πληροφορία με το παραγόμενο RNA. Με όμοιο τρόπο, η απέναντι αλυσίδα χαρακτηρίζεται ως κωδική, sense ή μη πρότυπο.*

Η επιμήκυνση της μεταγραφής, συμβαίνει μόλις εκκινηθεί η έναρξή της, υπό την παρουσία όλων των απαραίτητων μεταγραφικών παραγόντων. Σε αυτό το βήμα, το γνωστό ως «σύμπλοκο επιμήκυνσης», αποσυνδέει την σ υπομονάδα από τον υποκινητή και ξεκινά τη σύνθεση του RNA, με βάση συγκεκριμένη αλυσίδα ως πρότυπο.

Ο τερματισμός της μεταγραφής, πραγματοποιείται με έναν από τους τρεις πειραματικά αποδεδειγμένους τρόπους τερματισμού. Ο πρώτος είναι ο εσωτερικός τερματισμός (intrinsic ή Rho-indipendent termination), σύμφωνα με

τον οποίο, η αλληλεπίδραση της RNA πολυμεράσης αποκλειστικά με ειδικές αλληλουχίες του DNA, σημαίνει τον τερματισμό της μεταγραφής, με την αποκόλληση της RNA πολυμεράσης. Πιο αναλυτικά, η κατάλληλη ακολουθία του DNA, οδηγεί στο σχηματισμό φουρκέτας στο νεοσυντιθεμενο μετάγραφο, η οποία συνήθως ακολουθείται από νουκλεοτίδια ουρακίλης (ουριδίνη). Η συγκεκριμένη φουρκέτα, μπορεί να λειτουργήσει με διάφορους τρόπους κατασταλτικά στην RNA πολυμεράση, όπως με το να μετακινείται η τελευταία πάνω στο DNA, αλλά να μην επιμηκύνεται το νεοσυντιθέμενο RNA, ή με το να την καταστέλλει αλλοστερικά (Gusarov and Nudler, 1999). Και οι δύο από τους παραπάνω τρόπους τερματισμού, προκαλούν αρχικά την παύση της λειτουργίας της πολυμεράσης και έπειτα την αφαίρεσή της από τη διπλή έλικα. Αυτό μπορεί να επιτευχθεί, μέσω του πρωτεϊνικού παράγοντα *nusA*, ο οποίος προσδένεται ισχυρά στη φουρκέτα και τότε γίνεται μεταγραφή των τελευταίων νουκλεοτιδίων ουρακίλης, γεγονός που συνεπάγεται τη μείωση της ενέργειας της σύνδεσης μεταξύ DNA – RNA. Με δεδομένο ότι το RNA λαμβάνει εκείνη τη στιγμή τη δευτεροταγή του δομή, η δυναμική του ενέργεια, τείνει να το αποδεσμεύει από την πολυμεράση, πράγμα που καταφέρνει εύκολα, λόγω όπως αναφέρθηκε, της χαμηλής ενέργειας στη σύνδεσή του με το DNA.

Ο δεύτερος τρόπος τερματισμού, είναι ο λεγόμενος εξαρτώμενος από τον πρωτεϊνικό παράγοντα ρ (Rho) τερματισμός. Ο συγκεκριμένος παράγοντας, δεσμεύεται στο RNA και όταν το σύμπλοκο της RNA πολυμεράσης σταματήσει σε κάποια θέση τερματισμού, εξαρτώμενη από τον Rho παράγοντα, αυτός μπορεί να οδηγήσει στην αποκόλλησή του, τερματίζοντας τη μεταγραφή (Madigan *et al.*, 2006). Επιπλέον, σύμφωνα με έρευνες, μπορεί και ελέγχει επιλεκτικά τη μεταγραφή, περιορίζοντας μη αναγκαίες μεταγραφές για το κύτταρο, δηλαδή μεταγραφές με μικρή σημαντικότητα για αυτό, κυρίως αυτές που προέρχονται από τον αρνητικό (antisense) κλώνο (Peters *et al.*, 2012). Αυτές οι μεταγραφές μπορούν να ελεγχθούν από το εάν υπάρχει υπερσυσσώρευση των κλώνων του DNA, δημιουργώντας βρόχους R, γεγονός που συνεπάγεται αρνητικές συνέπειες για το άνοιγμα της διχάλας (Raghunathan *et al.*, 2018). Η τρόπος λειτουργίας του παράγοντα ρ, ομοιάζει με αυτόν οποιασδήποτε ελικάσης, διότι η ενέργεια για τη δράση τους πηγάζει από την υδρόλυση των ATP σε ADP, η οποία απελευθερώνει μία φωσφορική ομάδα, που προσδίδει ενέργεια σε αυτές. Ακόμη, ο συγκεκριμένος παράγοντας προσδένεται στο RNA και έχει αποδειχθεί πειραματικά, ότι αναγνωρίζει δύο θέσεις σε αυτό. Η πρώτη, είναι κάποια περιοχή με υψηλή σύσταση σε C και η δεύτερη περιλαμβάνει μία περιοχή που ενεργοποιεί την ATPase, ώστε να υδρολύσει τα ATP και να λάβει ενέργεια (Richardson, 1982). Αργότερα, όταν ο παράγοντας λάβει αρκετή ενέργεια και πλησιάσει πολύ κοντά στην πολυμεράση, με κάποιο τρόπο μπορεί να προκαλέσει την παύση και εν συνεχεία, τον τερματισμό της μεταγραφής. Κλείνοντας την αναφορά στον παράγοντα ρ, έχει βρεθεί ότι υπάρχει ένας παράγοντας τερματισμού, με όνομα *nusG*, ο οποίος φαίνεται να προσδένει τον ρ σε συγκεκριμένη περιοχή του RNA, διευκολύνοντας έτσι, τον τερματισμό της μεταγραφής (J. Li *et al.*, 1992).

Τέλος, ο τρίτος τρόπος τερματισμού της μεταγραφής, είναι ο τερματισμός εξαρτώμενος από την πρωτεΐνη *Mfd*, γνωστή και ως *Transcription Repair Coupling Factor* (TRCF). Αυτή, λειτουργεί ως ένα επιδιορθωτικό ένζυμο της μεταγραφής και έχει τη δυνατότητα να αφαιρεί την RNA πολυμεράση από το DNA, μετά την παύση της μεταγραφής. Η μετακίνηση αυτής της πρωτεΐνης, γίνεται αυτή τη φορά με την πρόσδεση πάνω στο DNA και με απευθείας αλληλεπίδραση με την πολυμεράση, ενώ επίσης χρησιμοποιείται ο ίδιος μηχανισμός της αποφωσφορυλίωσης των ATP για την απόκτηση ενέργειας, όπως και στον παράγοντα ρ (Roberts and J.-S. Park, 2004). Με τον τρόπο αυτό και με δεδομένο ότι ο παράγοντας *Mfd*, συνδέεται σε περίπου 25 νουκλεοτίδια που βρίσκονται σε κάποιο σημείο ανοδικά της πολυμεράσης, όταν η τελευταία παύσει τη μεταγραφή, τότε ο παράγοντας συνεχίζει να κινείται προς την πολυμεράση, έως ότου να την φτάσει και να την απελευθερώσει από το DNA.

1.2. Η Αλληλουχιση Πρωτης Γενιας

Ο ακριβής εντοπισμός της σειράς των νουκλεϊκών οξέων σε μία πολυνουκλεοτιδική αλυσίδα, ήταν ένα ζήτημα, το οποίο απασχολούσε τους επιστήμονες, μετά την ανακάλυψη ότι το γενετικό υλικό είναι το DNA. Μερικές

από τις πρώτες προσπάθειες αλληλούχισης, δόθηκαν από αρκετούς ερευνητές και σε αυτά τα πρώτα στάδια, περιλαμβάνονται μελέτες σε RNA, από οργανισμούς εύκολα επιλεγμένους και με χαμηλή πολυπλοκότητα, όπως τα βακτήρια και οι βακτηριοφάγοι. Επιπλέον δε, κρίθηκε προτιμότερη η μελέτη σε RNA και όχι DNA, λόγω του γεγονότος ότι δεν υπάρχει συμπληρωματική αλυσίδα, ώστε να προσθέτει πολυπλοκότητα, καθώς επίσης και λόγω του συγκριτικά μικρότερου μεγέθους αυτών των αλληλουχιών, σε σχέση με μόρια DNA ευκαρυωτικών οργανισμών. Ακόμη, πολλά ένζυμα ανακαλύφθηκε ότι δρουν αποκλειστικά σε RNA, όπως οι ριβονουκλεάσες, στοιχείο που βοήθησε τον κατάλληλο χειρισμό αυτών των μορίων. Έχοντας γνωστά τα παραπάνω, οι πρώτες προσπάθειες αλληλούχισης χρησιμοποιούσαν ραδιενεργό φώσφορο (^{32}P) για τη σήμανση των νουκλεοτιδίων, τα οποία παρέχονταν ένα προς ένα σε βακτήρια και με αυτόν τον τρόπο γινόταν μέτρηση της ενσωμάτωσής τους από την DNA πολυμεράση. Με την πάροδο των χρόνων, αυτή η τεχνική γενικεύθηκε με την εισαγωγή συγκεκριμένων ολιγονουκλεοτιδίων ως εκκινητές στην DNA πολυμεράση, ανάγοντας αυτή την τεχνική, σε μία μεγαλύτερης κλίμακας διαδικασία (Heather and Chain, 2016).

Το 1977, μία καινοτόμα προσέγγιση δόθηκε από τον Βρετανό βιοχημικό Frederick Sanger και τους συνεργάτες του, κερδίζοντας το δεύτερο εν σειρά, βραβείο Νόμπελ το 1980. Να αναφερθεί πως προηγουμένως, είχαν προταθεί και άλλες ανάλογες τεχνικές, οι οποίες εκείνες αποτέλεσαν ίσως την πρώτη γενιά αλληλουχίσεων, όπως η τεχνική «συν-πλην» του Sanger ή η τεχνική των Allan Maxam & Walter Gilber, αλλά αυτή που επικράτησε περισσότερο, ήταν η λεγόμενη «σύνθεση με τερματισμό» του Sanger (Sanger *et al.*, 1977). Η βασική ιδέα, είναι η χρήση δεοξυριβονουκλεοτιδίων (dNTPs) και δι-δεοξυριβονουκλεοτιδίων (ddNTPs) για τη διαδικασία της επιμήκυνσης, με τα τελευταία να στερούνται του 3'-OH άκρου (έχουν 3'-H), το οποίο απαιτείται για τη σύνδεση του επόμενου νουκλεοτιδίου, δηλαδή με την ενσωμάτωσή τους στην αλυσίδα, η επιμήκυνση τερματίζεται. Έτσι, σε ένα δείγμα αλληλούχισης, στο οποίο περιέχονται τόσο dNTPs, όσο και ddNTPs, η αλληλούχιση της αλυσίδας μπορεί να τερματιστεί ανά πάσα στιγμή σε οποιοδήποτε σημείο. Εκτελώντας τέσσερις παράλληλες αντιδράσεις που περιέχουν κάθε μεμονωμένη βάση ddNTP και κάνοντας ηλεκτροφόρηση σε γέλη πολυακρυλαμιδίου κάθε δείγμα ξεχωριστά, επιτυγχάνεται ταξινόμηση των νέων αναγνώσεων που προέκυψαν με βάση το μήκος τους, εκτιμώντας την αλληλουχία τους, συνήθως με την τεχνική της αυτοραδιογραφίας (Sanger *et al.*, 1977). Μερικά από τα πλεονεκτήματα αυτής της μεθόδου για πολλά χρόνια, ήταν η απλότητα στην εφαρμογή της, η ακρίβεια λόγω όμως των πολλαπλών δειγμάτων, καθώς και η ευρωστία (robustness) στη χρήση της. Τέλος, η τεχνική αυτή επιδέχθηκε πολλές βελτιώσεις από το 1977 και ύστερα και η σημαντικότερη από αυτές είναι η αντικατάσταση του ραδιενεργού ισοτόπου για τη σήμανση, με φθορίζουσες χρωστικές ουσίες, διευκολύνοντας περαιτέρω τη διαδικασία προετοιμασίας του δείγματος.

1.3. Η Αλληλουχιση Επομενης Γενιας (Next Genaration Sequencing (NGS))

Η αλληλούχιση επόμενης ή δεύτερης γενιάς, ήρθε στα μέσα της δεκαετίας του ’90, με σκοπό να αντικαταστήσει τις έως τότε αλληλουχίσεις χαμηλής διεκπεραιωτικής ικανότητας, με την πρόταση καινοτόμων λύσεων που είχαν αναπτυχθεί. Σε αντίθεση με τις τεχνικές αλληλούχισης πρώτης γενιάς, όπως η αλληλούχιση κατά Sanger, πλέον με την πρόοδο της τεχνολογίας, κατέστη εφικτή η αλληλούχιση ολόκληρων γονιδιωμάτων και όχι απλά μερικών γονιδίων. Αυτό επιτεύχθηκε, με την θρυματοποίηση του γενετικού υλικού σε αλληλουχίες μικρού μήκους και την παράλληλη – μαζική αλληλούχισή τους. Έτσι, από το 2000 και έπειτα, δηλαδή από όταν ξεκίνησαν να υλοποιούνται αυτές οι τεχνικές στους εμπορικούς αλληλουχητές, ξεκίνησε μία ραγδαία μείωση στο κόστος αλληλούχισης ανά 1kb. Οι συνεχείς βελτιώσεις των τεχνικών όλα τα χρόνια που ακολούθησαν, καταδεικνύουν τη μεγάλη ανάγκη που υπάρχει έως σήμερα, για εύκολη, γρήγορη και ακριβή αλληλούχιση, διαδικασία που δημιουργεί έναν αστρονομικό όγκο δεδομένων προς διαχείριση και ανάλυση (Stephens *et al.*, 2015).

Όπως θα αναλυθεί παρακάτω, πλέον υπάρχει ένας σημαντικός αριθμός τεχνικών αλληλούχισης επόμενης γενιάς. Αυτές χρησιμοποιούνται ανάλογα με το πρόβλημα που επιθυμούμε να προσεγγίσουμε, δηλαδή το είδος εμπλουτισμού του δείγματος που θέλουμε να πετύχουμε, ή αλλιώς, το σημείο ή χαρακτηριστικό του γονιδιώματος που θέλουμε να αλληλουχηθεί.

Ιστορικά, η διαδικασία αλληλούχισης επόμενης γενιάς, διακρίθηκε σε πολλές μεθοδολογίες, από το 2000 έως και το 2010. Αυτή η γενιά χαρακτηρίστηκε για τις συνεχείς βελτιώσεις στην αλληλούχιση μεγάλου αριθμού παράλληλων αντιδράσεων, καθώς και για τη βελτίωση της απεικόνισης υψηλής ανάλυσης, αυξάνοντας ταυτόχρονα την ποιότητα της αλληλούχισης. Οι πιο διαδεδομένες μεθοδολογίες είναι η πυροαλληλούχιση (Pyrosequencing), η αλληλούχιση με σύνθεση και η αλληλούχιση με ολιγονουκλεοτιδική πρόσδεση (ligation). Η διαφορά της πυροαλληλούχισης, σε σχέση με την προηγούμενη γενιά, έγκειται στο ότι πλέον η ταυτότητα ενός νουκλεοτιδίου δεν ανιχνεύεται με χρήση φθοριζουσών ουσιών, ενσωματωμένες σε dNTPs, αλλά με χρήση της τεχνικής χημειοφωταύγειας, στην οποία η σουλφουρυλάση, χρησιμοποιείται για τη μετατροπή του πυροφωσφορικού, που απελευθερώνεται με την ενσωμάτωση του dNTPs, σε ATP, το οποίο χρησιμοποιείται ως υπόστρωμα στο ένζυμο λουσιφεράση, επομένως εκπέμπεται φως, ανάλογο με την ποσότητα του πυροφωσφορικού (Heather and Chain, 2016). Με βάση το παραγόμενο φως, μπορεί να γίνει εκτίμηση της ποσότητας των νουκλεοτιδίων που ενσωματώθηκαν, αλλά και της ταυτότητάς τους. Έτσι, αυτή η τεχνική έχει πολλαπλά πλεονεκτήματα σε σχέση με την αλληλούχιση κατά Sanger, μερικά από τα οποία είναι ότι δεν απαιτείται πλέον μετασχηματισμός των νουκλεοτιδίων σε φθορίζοντα dNTPs, καθώς επίσης και ότι δεν απαιτείται ηλεκτροφόρηση, επειδή το σήμα λαμβάνεται απευθείας από το φως που εκπέμπεται από τη λουσιφεράση και βαθμονομείται με ειδικούς σαρωτές (Ronaghi *et al.*, 1996). Επιπλέον, η διαδικασία ενίσχυσης γίνεται με τη λεγόμενη emulsion PCR, κατά την οποία τα θραύσματα γενετικού υλικού, συνδέονται με τη βοήθεια των ανταπτόρων σε σφαιρίδια και ακολουθεί αλληλούχιση πάνω σε αυτά. Η τεχνική αυτή προτάθηκε από την εταιρεία βιοτεχνολογίας 54 Life Sciences το 2005 και έκτοτε ενσωματώθηκε στον αλληλουχητή της ίδιας εταιρείας, με όνομα 454 GS 20.

Λίγο αργότερα, το 2006, προτάθηκε μία διαφοροποιημένη μεθοδολογία αλληλούχισης, η οποία βασίστηκε στη δυνατότητα αλληλούχισης κατά ζεύγη αναγνώσεων (Paired-end) και στην τεχνική της αλληλούχισης κατά σύνθεση. Η μεθοδολογία αυτή, αρχικά προτάθηκε από την εταιρεία Solexa, η οποία ένα χρόνο αργότερα αποκτήθηκε από την ευρέως γνωστή εταιρεία στο χώρο της αλληλούχισης, την Illumina. Ανάμεσα στις βελτιώσεις αυτής της νέας τεχνικής, περιλαμβάνεται μία διαφορετική μεθοδολογία στην εκτέλεση ενίσχυσης με Polymerase chain reaction (PCR), σύμφωνα με την οποία, πλέον δεν χρησιμοποιείται η emulsion PCR, αλλά μία τεχνική, κατά την οποία οι αναγνώσεις με συνδεδεμένους αντάπτορες στα 5' και 3' άκρα, συνδέονται πάνω σε επιφάνεια συμπληρωματικών ολιγονουκλεοτιδίων, σχηματίζοντας ένα σχήμα αψίδας (Heather and Chain, 2016). Κατ' αυτό τον τρόπο γίνεται η αντιγραφή των αναγνώσεων και αυτή η τεχνική, ονομάζεται Bridge PCR. Το στάδιο της αλληλούχισης, γίνεται πλέον με αλληλούχιση κατά σύνθεση, χρησιμοποιώντας φθορίζοντα dNTPs με 3′-OH (ddNTPs), τα οποία για να μπορέσει να γίνει η επιμήκυνση της αλυσίδας όταν καταλαμβάνουν μία θέση, αποκόπτεται το υδρογόνο από τη συγκεκριμένη ομάδα και μπορεί να γίνει η συνένωση του επόμενου dNTP, με φωσφοδιεστερικό δεσμό. Η διαδικασία αυτή επαναλαμβάνεται για πολλούς κύκλους και τα φθορίζοντα σήματα ανιχνεύονται με κατάλληλα λέιζερ, πριν την ενζυμική απομάκρυνση των φθοριζόντων τμημάτων και τη συνέχιση στην επόμενη θέση (Heather and Chain, 2016). Η συγκεκριμένη μεθοδολογία υλοποιήθηκε στον πρώτο αλληλουχητή της Solexa, με την ονομασία *Genome Analyzer II*, δίνοντας στους επιστήμονες τη δυνατότητα αλληλούχισης δεδομένων, όγκου 1Gb με μία μόνο εκτέλεση. Αν και αρχικά το μέγεθος των αναγνώσεων ήταν σχετικά μικρό, δηλαδή της τάξης των 35 βάσεων περίπου, η καινοτόμα δυνατότητα των κατά ζεύγη αναγνώσεων, άνοιξε νέες προοπτικές στην ακριβέστερη στοίχιση του δείγματος με το γονιδίωμα αναφοράς, καθώς και στον εντοπισμό μεταλλάξεων, όπως απαλοιφές, εισαγωγές και αντικαταστάσεις νουκλεοτιδίων ή τμημάτων.

Η τρίτη τεχνική αλληλούχισης, είναι η αλληλούχιση με ολιγονουκλεοτιδική πρόσδεση, η οποία προτάθηκε επίσης το 2006, από την εταιρεία Applied Biosystems, που στη συνέχεια μετονομάστηκε σε Life Technologies και το 2014 είχε εξαγοραστεί από την εξίσου γνωστή Thermo Fisher Scientific. Στην αλληλούχιση αυτή,

χρησιμοποιούνται ολιγονουκλεοτιδικοί ανιχνευτές, συνήθως των 8 βάσεων, οι οποίοι φέρουν στο 3' άκρο δύο οποιεσδήποτε βάσεις και οι υπόλοιπες είναι εκφυλισμένες, ενώ μία φθορίζουσα ουσία, είναι συνδεδεμένη στο 5' άκρο. Η αλληλούχιση ξεκινά με την υβριδοποίηση του εκκινητή και δίπλα σε αυτόν, προσδένεται με τη βοήθεια της DNA λιγάσης, ο κατάλληλος ανιχνευτής, με βάση την ταυτότητα των δύο πρώτων νουκλεοτιδίων της. Έπειτα, οι μη υβριδοποιημένοι ανιχνευτές απορρίπτονται με διαδικασία ξεπλύματος και τουλάχιστον τα 3 τελευταία νουκλεοτίδια μαζί με το σήμα φθορισμού, καταγράφονται και αποικοδομούνται. Η διαδικασία αυτή συνεχίζεται για κάθε ανάγνωση, ενώ σε κάθε επόμενο κύκλο, ο εκκινητής ολισθαίνει κατά μία θέση αριστερά και η διαδικασία επαναλαμβάνεται εκ νέου (Heather and Chain, 2016). Τα αποτελέσματα της στοίχισης, μας δίνουν μία ακριβή εικόνα του προσδιορισμού της αλληλουχίας, όμως για μικρό μήκος αναγνώσεων.

Τέλος, μία ευρέως χρησιμοποιούμενη τεχνική ακόμα και σήμερα, είναι η αλληλούχιση ημιαγωγών ιόντων. Αυτή προτάθηκε από την τότε εταιρεία Life Technologies το 2010 και κατ΄ αντιστοιχία με την πυροαλληλούχιση, χρησιμοποιεί αρχικά emulsion PCR για την ενίσχυση του δείγματος, αλλά πλέον η ταυτότητα των νουκλεοτιδίων εντοπίζεται με την ανίχνευση ιόντων υδρογόνου (H^+). Πιο αναλυτικά, κατά τη διαδικασία του πολυμερισμού, δηλαδή όταν οι μεμονωμένες βάσεις, συνδέονται στην επιμηκούμενη αλυσίδα, τότε εκλύονται ιόντα υδρογόνου, με αποτέλεσμα τη μεταβολή του pH του δείγματος. Η μεταβολή αυτή, είναι μετρήσιμη από αισθητήρες, όπως ο Complementary Metal-oxide Semiconductor (CMOS), οι οποίοι μπορούν να υπολογίσουν με αρκετή ακρίβεια το pH και έτσι να γίνει η πρόβλεψη του εκάστοτε νουκλεοτιδίου που ενσωματώνεται (Heather and Chain, 2016).

Επιγραμματικά, όλες αυτές οι νέες τεχνολογίες ανά τα χρόνια, οδήγησαν στην ραγδαία μείωση του κόστους της αλληλούχισης και την έχουν καταστήσει σήμερα, μία διαδικασία σχεδόν ρουτίνας. Το πέρασμα από την αλληλούχιση πρώτης στην αλληλούχιση δεύτερης γενιάς, δηλαδή οι συνεχείς βελτιώσεις των μεθοδολογιών, ήταν και αυτές που ευθύνονται για την ολοκλήρωση του προγράμματος χαρτογράφησης του ανθρώπινου γονιδιώματος (Human Genome Project (HGP)), νωρίτερα από το αναμενόμενο. Με μία σύντομη ανασκόπηση στο πεδίο αυτό, προκύπτει ότι οι νέες προσεγγίσεις έχουν αυξήσει τη δυνατότητα της αλληλούχισης, σε βαθμό που ξεπερνά το νόμο του Moore. Ο συγκεκριμένος νόμος, περιγράφει εμπειρικά μια συνεχή τάση των νέων μικροεπεξεργαστών, να έχουν διπλασιασμένη ισχύ, άρα και διπλάσιο αριθμό τρανζίστορ σε σχέση με τους προηγούμενους, κάθε δύο χρόνια. Έτσι, τεχνολογικές εξελίξεις που υπακούουν σε αυτόν, θεωρείται ότι εμφανίζουν εξαιρετική πρόοδο, καθιστώντας το νόμο αυτό, χρήσιμο για σύγκριση. Αυτό σημαίνει, ότι παρόλο που ο αριθμός των τρανζίστορ διπλασιαζόταν κάθε δύο χρόνια, ο όγκος δεδομένων αλληλούχισης διπλασιαζόταν κάθε 5 μόλις μήνες, ανάμεσα στα έτη 2004 και 2010 (Stein, 2010). Πλέον, γίνεται αντιληπτό ότι με μία αποτίμηση της αλληλούχισης δεύτερης γενιάς, έφερε στο προσκήνιο ένα νέο μοντέλο στη μικροβιολογία, αναφορικά με το χαρακτηρισμό των παθογόνων καταστάσεων, το οποίο περιλαμβάνει το γονιδιωματικό ορισμό αυτών των καταστάσεων, σε σχέση με τα κριτήρια μορφολογίας, ιδιοτήτων χρώσης και μεταβολικών κριτηρίων των ιστών, τα οποία παλαιότερα ήταν η μόνη επιλογή των επιστημόνων. Αυτό λοιπόν το γονιδιωματικό μοντέλο που βασίζεται στην αλληλούχιση, είναι σε θέση να εξάγει πιο ακριβή αποτελέσματα των παθολογικών καταστάσεων, να αποκαλύψει και να τεκμηριώσει με ισχυρότερο τρόπο αστοχίες στις βιολογικές διεργασίες, να εκτιμήσει συγκεκριμένη γενετική προδιάθεση, καθώς και να δράσει προσωποποιημένα, για κάθε ασθενή ατομικά, όπως συμβαίνει για παράδειγμα, στον κλάδο της φαρμακογενετικής.

1.4. Η Αλληλουχιση Τριτης Γενιας

Από το 2007 και μετά, τα βιολογικά δεδομένα που προκύπτουν από την αλληλούχιση είναι πλέον καθημερινό φαινόμενο και ο όγκος τους έχει παρουσιάσει αξιοσημείωτη αύξηση, πολλών τάξεων μεγέθους, σε σχέση με το παρελθόν. Με τους τρέχοντες ρυθμούς ανάπτυξης της τεχνολογίας και τον εντοπισμό λύσεων σε υπάρχοντα προβλήματα που παραμένουν επίκαιρα ακόμη και σήμερα, η έρευνα στο πεδίο της αλληλούχισης εκτιμάται ότι θα δώσει νέα ευρήματα, αναφορικά με την ακόμη μεγαλύτερη κλιμακωσιμότητα των νέων μεθόδων. Σε συνδυασμό όμως με την κλιμακωσιμότητα, η αύξηση της ποιότητας και της ευαισθησίας της αλληλούχισης, μαζί με το χαμηλό

κόστος ανά 1Mb, παραμένουν μία σημαντική πρόκληση.

Αρχικά, όπως αναφέρθηκε παραπάνω, κατά την προετοιμασία της βιβλιοθήκης το δείγμα θρυμματοποιείται σε αναγνώσεις ορισμένου μήκους, προσδένεται σε αντάπτορες ανάλογα με τον αλληλουχητή που θα χρησιμοποιηθεί και ενισχύεται μέσα από την PCR. Όλα αυτά τα βήματα της προετοιμασίας, προσθέτουν το καθένα και κάποια πόλωση (bias) ή αλλιώς, αλλοίωση των αρχικών δεδομένων (Dohm *et al.*, 2008). Γενικά, συνήθως πάντα η χρήση της PCR οδηγεί στην απόκλιση της αλληλούχισης από τα πραγματικά δεδομένα και αυτό συμβαίνει για πολλούς λόγους. Ο πιο βασικός, είναι ότι οι αναγνώσεις δεν είναι δυνατό να ακολουθήσουν ομοιόμορφη αντιγραφή, δηλαδή να διατηρηθούν οι αναλογίες του δείγματος για κάθε γονιδιωματική περιοχή. Με αυτόν τον τρόπο, παρά το γεγονός ότι πλέον οι ακολουθίες είναι υπερπολλαπλάσιες και αυτό μπορεί να οδηγήσει σε καλύτερη στατιστική σημαντικότητα την εξαγωγή συμπερασμάτων, οι νέες ποσότητες αναγνώσεων μπορεί να μην είναι πραγματικά έγκυρες. Αναγνώσεις με υψηλό ποσοστό AT ή GC δεν μπορούν μοριακά να αντιγραφούν με την ίδια ταχύτητα, γεγονός που οδηγεί σε αλλοίωση των δεδομένων, κυρίως στο στάδιο της εκθετικής φάσης. Ορισμένες έρευνες (Van Dijk *et al.*, 2014a; Oyola *et al.*, 2012) υποστηρίζουν, ότι η αλλαγή της DNA πολυμεράσης, είναι ικανή να οδηγήσει σε πιο ακριβή αποτελέσματα, αλλά αυτό είναι κάτι που εξαρτάται από τη σύσταση του κάθε οργανισμού, ως προς τα ποσοστά των τεσσάρων βάσεων του γονιδιώματος και παραμένει απλά μία βελτίωση, χωρίς να αποτελεί τη λύση του προβλήματος.

Προς αυτή την κατεύθυνση βρίσκονται τα ευρήματα από το εργαστήριο του Stephen Quake, ήδη από το 2003 (Braslavsky *et al.*, 2003), τα οποία παρουσίασαν μία μεθοδολογία παρόμοια με εκείνη της εταιρείας Solexa (αλληλούχιση με σύνθεση), αλλά χωρίς τη χρήση της PCR. Η συγκεκριμένη μεθοδολογία, ονομάζεται αλληλούχιση μονού βιομορίου (Single-molecule Sequencing). Η εξάλειψη της PCR, εκτιμάται πλέον ότι είναι ένα από τα σημεία - κλειδιά της αλληλούχισης $3^{ης}$ γενιάς, γεγονός που συνεπάγεται τη μεγαλύτερη ευαισθησία των αλληλουχητών, στο να αλληλουχούν αξιόπιστα αναγνώσεις σε πολύ λιγότερα αντίγραφα. Σύμφωνα με αυτή τη μεθοδολογία, τα τμήματα DNA που χρησιμοποιούνται ως πρότυπο, συνδέονται σε μια επίπεδη επιφάνεια και η επιμήκυνση της νέας αλυσίδας γίνεται κατά μία βάση τη φορά, χρησιμοποιώντας *αναστρέψιμα φθορίζοντα dNTPs* τερματισμού. Τα τελευταία μοιάζουν πολύ με τα ddNTPs που αναφέρθηκαν παραπάνω, όμως παρουσιάζουν δύο ιδιαίτερα χαρακτηριστικά (Bowers *et al.*, 2009). Το πρώτο είναι η ικανότητά τους να μετασχηματίζονται από τερματικά νουκλεοτίδια, σε νουκλεοτίδια με ικανότητα σύνδεσης από το 3′ άκρο τους, το οποίο αρχικά είναι ελεύθερο υδροξυλομάδας και με διαδικασία πλύσης, μπορεί να δεχθεί τη σύνδεση και άλλων νουκλεοτιδίων. Το δεύτερο χαρακτηριστικό, αφορά μία τροποποίηση στις βάσεις τους με προπαργυλαμίνη, η οποία δίνει τη δυνατότητα στην εκάστοτε αζωτούχο βάση, να συνδέεται με έναν συνδετήρα διάσπασης, ο οποίος με τη σειρά του είναι συνδεδεμένος με την χρωστική φθορισμού. Κατά την αλληλούχιση, στον κάθε κύκλο προστίθεται ένα τροποποιημένο νουκλεοτίδιο, το οποίο απεικονίζεται με βάση την ουσία φθορισμού και έπειτα ακολουθεί μία διαδικασία πλύσης, ώστε να αποικοδομηθεί το σήμα φθορισμού και το 3′ άκρο να υδροξυλιωθεί, επιτρέποντας τη σύνδεση και άλλων βάσεων σε αυτό. Η εν λόγω μεθοδολογία, αν και ήταν η πρώτη που δεν χρησιμοποιούσε ενίσχυση με PCR, έχει κύρια μειονεκτήματα το χρόνο, το κόστος καθώς και το μικρό μήκος των αναγνώσεων της αλληλούχισης.

Μία βελτίωση της συγκεκριμένης μεθοδολογίας, αποτελεί η τεχνική που εφαρμόζεται στους αλληλουχητές της εταιρείας Pacific Biosciences, με την ονομασία Single Molecule Real Time Sequencing (SMRT-seq) (Van Dijk *et al.*, 2014b). Κατά τη συγκεκριμένη μεθοδολογία, το DNA ενσωματώνεται σε συστοιχίες νανοδομών, τις γνωστές ως Zero-mode Waveguides (ZMWs), οι οποίες εκμεταλλεύονται τις ιδιότητες του φωτός, λόγω του ότι έχουν διάμετρο μικρότερη από το μήκος κύματος του φωτός. Με αυτόν τον τρόπο, η ενσωμάτωση κάθε τροποποιημένου νουκλεοτιδίου από την DNA πολυμεράση, οδηγεί στην εκπομπή φωτός, το οποίο αποσυντίθεται εκθετικά, φωτίζοντας τον πυθμένα της νανοδομής (Eid *et al.*, 2009). Η επέκταση της αλυσίδας και συγκεκριμένα, ο φθορισμός που εκλύεται μπορεί να καταγραφεί σε πραγματικό χρόνο με μεγαλύτερη ευκολία σε σχέση με την προηγούμενη μεθοδολογία. Αυτό οδηγεί σε πολλαπλά πλεονεκτήματα, όπως η ταχύτητα αλληλούχισης που πλέον είναι ίση με την ταχύτητα της DNA πολυμεράσης, το χαμηλό κόστος και η υψηλή ποιότητα αλληλούχισης, σε μη ενισχυμένο δείγμα, καθώς και το μήκος των αναγνώσεων, το οποίο μπορεί να φτάσει και τις 10kb.

Τέλος, μία πολλά υποσχόμενη τεχνολογία, είναι και αυτή της αλληλούχισης νανοπόρων (Nanopore sequencing). Η συγκεκριμένη μεθοδολογία, κέρδισε μεγάλο ενδιαφέρον για την ανίχνευση μεμονωμένων μορίων, όπως ιόντα, νουκλεοτίδια, ουσίες φαρμάκων, καθώς και ακολουθίες όπως RNA, DNA και πολυπεπτίδια. Έχει επεκταθεί στην ιατρική διάγνωση και αποτελεί επίσης, μία σύγχρονη λύση για τον προσδιορισμό αλληλουχίας DNA τρίτης γενιάς. Οι αρχικές έρευνες που οδήγησαν στην ανακάλυψη της συγκεκριμένης μεθοδολογίας, έγιναν πριν την εδραίωση των τεχνικών αλληλούχισης δεύτερης γενιάς και περιλαμβάνουν την παρατήρηση ότι τα νουκλεϊκά οξέα, μπορούν να κατευθυνθούν στην περιοχή της λιπιδικής διπλοστοιβάδας και συγκεκριμένα σε ιοντικούς διαύλους του κυττάρου, εφαρμόζοντας ηλεκτροφόρηση. Την παρατήρηση αυτή χρησιμοποίησαν λίγο αργότερα επιστήμονες, ώστε εφαρμόζοντας ηλεκτρική τάση σε ηλεκτρικώς ανθεκτική συνθετική μεμβράνη, να καταφέρει να περάσει το DNA, βάση προς βάση, μέσα από τεχνητούς διαύλους και έτσι να επιτευχθεί αλληλούχιση. Πιο αναλυτικά, η είσοδος και οι επακόλουθες μετατοπίσεις των νουκλεοτιδίων δια μέσω ενός διαύλου, μπορούν να χαρακτηριστούν μετρώντας τη διαφοροποίηση (αναταραχή) που παρουσιάζει η τάση του ρεύματος στη μεμβράνη και έτσι, να μπορέσουν να ταυτοποιηθούν οι βάσεις που περνάνε μέσα από το δίαυλο. Να αναφερθεί επίσης, ότι οι ιοντικοί δίαυλοι συναντώνται ευρέως στη φύση και ποικίλουν ανάλογα με τα χαρακτηριστικά τους, τα οποία ευθύνονται και για τη λειτουργικότητα που παρουσιάζει ο κάθε ένας. Για παράδειγμα, στη συγκεκριμένη περίπτωση της αλληλούχισης, ο τεχνητός δίαυλος θα πρέπει να έχει διάμετρο τέτοια, ώστε να εισχωρεί σε αυτόν μία αλυσίδα DNA και όχι μεγαλύτερα βιομόρια, όπως πρωτεΐνες, ενώ θα πρέπει να έχει επίσης, κατάλληλη χημική και θερμική σταθερότητα έναντι των λιπιδικών μεμβρανών και καλή ηλεκτρική αγωγιμότητα (Haque *et al.*, 2013). Η διπλή έλικα του DNA διαχωρίζεται, λόγω του ότι δεν χωράνε να εισχωρήσουν στο δίαυλο και οι δύο και μία εξωνουκλεάση, η οποία είναι τοποθετημένη ακριβώς πάνω από το δίαυλο, κάνει αποκοπή μία βάση τη φορά και την αφήνει να διαπεράσει το δίαυλο, ώστε να γίνει η καταμέτρηση του ρεύματος. Κατ΄ επέκταση, η αναγνώριση του κάθε νουκλεοτιδίου (Deoxynucleoside Monophosphate (dNMP)), γίνεται από τον προσαρμογέα αμινοκυκλοδεξτρίνης, ο οποίος συνδέεται ομοιοπολικά εντός του διαύλου. Όταν λοιπόν, ένα dNMP περνά μέσα από το σύμπλοκο δίαυλος – αμινοκυκλοδεξτρίνη και κατευθύνεται στη λιπιδική διπλοστοιβάδα, το ιοντικό ρεύμα δια μέσω του πόρου μειώνεται σε κάποιο επίπεδο, το οποίο είναι αντιπροσωπευτικό για την εκάστοτε βάση (Branton *et al.*, 2008).

Η παρούσα μεθοδολογία είναι υλοποιημένη σε όλους τους αλληλουχητές της Oxford Nanopore Technologies (Eisenstein, 2012) από το 2008, με ονόματα MinION, GridION και PromethION. Ιδιαίτερη εντύπωση προκαλεί ο πρώτος από αυτούς, ο οποίος έχει διαστάσεις ενός κινητού τηλεφώνου (ίσως και μικρότερες πλέον) και κυκλοφόρησε το 2014. Αποτελεί μία πολύ προσιτή λύση στην αλληλούχιση κυρίως μικρών γονιδιωμάτων, όπως τα βακτηριακά γονιδιώματα. Μάλιστα, έχει χρησιμοποιηθεί στην αλληλούχιση τέτοιων γονιδιωμάτων αναφοράς, με μεγάλη επιτυχία, δηλαδή φθηνό κόστος, μεγάλο μήκος αναγνώσεων και γρήγορη ταχύτητα, καθώς η αλληλούχιση δεν γίνεται πλέον, ούτε με PCR, αλλά ούτε και με τη βοήθεια της DNA πολυμεράσης (Quick *et al.*, 2014).

1.5. Τεχνικες Αλληλουχισης Επομενης Γενιας

Με την εδραίωση της δεύτερης γενιάς αλληλούχισης και λαμβάνοντας υπόψιν, πως η αλληλούχιση του ανθρώπινου γονιδιώματος, είχε φτάσει στο πέρας της, γεννήθηκε η ανάγκη εκτέλεσης πολλών διαφορετικών πειραμάτων αλληλούχισης από την επιστημονική κοινότητα. Τα πειράματα αυτά, ήταν διαφορετικά ως προς το εκάστοτε βιολογικό πρόβλημα, δηλαδή ο προς αλληλούχιση οργανισμός και η περιοχή ενδιαφέροντος του κάθε γονιδιώματος, εμφάνιζαν μεγάλη ετερογένεια. Σταδιακά, δημιουργήθηκαν νέες τεχνικές αλληλούχισης πάνω στην τυπική μεθοδολογία, προσαρμοσμένες στα επιμέρους προβλήματα και έτσι, επιτυγχάνεται μεγαλύτερη ακρίβεια και λιγότερος θόρυβος, με την εστίαση σε περιοχές ενδιαφέροντος, καθώς και λιγότερος χρόνος αλληλούχισης, αφού εξαλείφονται μη απαραίτητες περιοχές. Με τον τρόπο αυτό, δημιουργήθηκαν ειδικά πρωτόκολλα, σύμφωνα με τα οποία, καλύπτεται ένα μεγάλο εύρος βιολογικών προβλημάτων, όπως η αναζήτηση και ο εντοπισμός μεταλλάξεων ή πολυμορφισμών,

η ανίχνευση διαφορικά εκφρασμένων γονιδίων ανάμεσα σε δείγματα (Differentially Expressed Genes (DEG)), η αλληλούχιση εξωνικών περιοχών και ο εντοπισμός γονιδιακών στοιχείων. Παρακάτω, αναφέρονται οι πιο γνωστές τεχνικές, οι οποίες χρησιμοποιούνται ευρέως, ανάλογα με τα προβλήματα που αναφέρθηκαν.

1.5.1. DNA Sequencing (Αλληλουχιση DNA)

Η αλληλούχιση DNA χρησιμοποιείται για την εύρεση της ακολουθίας γονιδίων, χρωμοσωμάτων ή γονιδιωμάτων διαφόρων ειδών, όπως η αλληλούχιση γονιδιωμάτων διαφόρων ζώων, φυτών και μικροβίων. Η συγκεκριμένη τεχνική, περιλαμβάνει δύο επιμέρους μεθοδολογίες: τη map-based και τη BAC-based αλληλούχιση DNA. Στη map-based αλληλούχηση, η ανάλυση της αλληλουχίας γίνεται με ιεραρχική τμηματική ανάλυση (hierarchical shotgun sequencing), η οποία αναφέρεται και ως ανάλυση, βασισμένη σε χάρτες ή βασισμένη σε τεχνητό βακτηριακό γενετικό υλικό (Bacterial Artificial Chromosome (BAC)). Έγινε χρήση της συγκεκριμένης προσέγγισης για την ανάλυση του ανθρώπινου γονιδιώματος, το οποίο είναι πλούσιο σε επαναληπτικές ακολουθίες, στο Πρόγραμμα του Ανθρώπινου Γονιδιώματος (HGP) (Lander *et al.*, 2001).

Η προσέγγιση αυτή, αφορά την δημιουργία και οργάνωση συνόλου κλώνων που καλύπτουν το γονιδίωμα, την αλληλούχιση τους και την συναρμολόγηση δεδομένων αλληλουχιών από τους κλώνους για τον προσδιορισμό της αλληλοεπικάλυψης και τη δημιουργία μιας συνεχούς αλληλουχίας. Οι κλώνοι προέρχονται από βιβλιοθήκες BAC, οι οποίες δημιουργούνται επειδή το γενωμικό DNA κατακερματίζεται τυχαία μέσω πέψης με ένζυμα περιορισμού (τις γνωστές DNAάσες) και ενσωματώνονται σε φορείς βακτηριακών τεχνητών χρωμοσωμάτων, οι οποίοι εισάγονται στα βακτήρια. Στη συνέχεια, χαρτογραφούνται contigs μεγάλων κλώνων, τα οποία αποτελούνται από πολλαπλούς επικαλυπτόμενους κλώνους βιβλιοθηκών BAC, που καλύπτουν καθένα από τα ανθρώπινα χρωμοσώματα. Τα contigs ταυτοποιούνται μέσω τεχνικών που αναγνωρίζουν τα τμήματα DNA των κλώνων που επικαλύπτονται (Lander *et al.*, 2001).

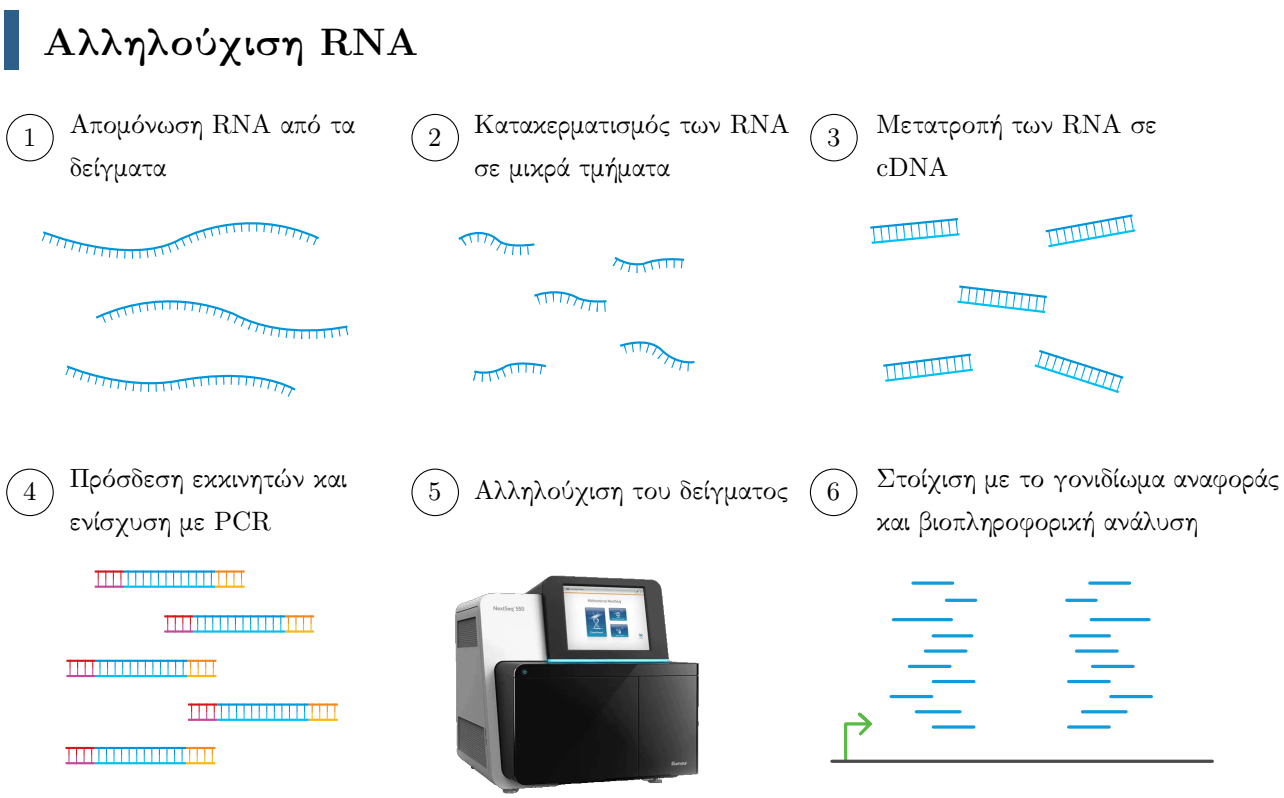
Έπειτα, δημιουργείται βιβλιοθήκη *shotgun* από κλώνους BAC που έχουν επιλεχθεί για αλληλούχιση. Συγκεκριμένα, οι κλώνοι BAC που επιλέχθηκαν για αλληλούχιση DNA κατακερματίζονται πάλι και δημιουργούνται μικρότερα θραύσματα γονιδιωματικού DNA, τα οποία κλωνοποιούνται σε φορείς για να δημιουργήσουν την συγκεκριμένη βιβλιοθήκη. Ακολουθεί βιοπληροφορική ανάλυση των δεδομένων που προκύπτουν από την αλληλούχιση DNA, ώστε να δημιουργηθεί σύνολο αλληλουχιών που καλύπτουν πολλαπλούς κλώνους BAC και συναρμολόγηση των δεδομένων αυτών, για τον προσδιορισμό των αλληλοεπικαλυπτόμενων αλληλουχιών και τη δημιουργία μιας συνεχούς αλληλουχίας (Venter *et al.*, 2001).

Όσον αναφορά τον καθορισμό της νουκλεοτιδικής αλληλουχίας, αξίζει να αναφερθεί η μέθοδος που χρησιμοποιήθηκε κατά το πρόγραμμα Αλληλούχησης του Ανθρώπινου Γονιδιώματος. Σε αυτή, χρησιμοποιείται DNA πολυμεράση, η οποία προσδένεται σε μονόκλωνο DNA και συνθέτει την συμπληρωματική αλυσίδα του. Στη συνέχεια, η DNA πολυμεράση ενσωματώνει τυχαία μια φθορίζουσα βάση στο 3' άκρο και τότε η σύνθεση του DNA τερματίζεται, δημιουργώντας μείγμα νεοσυντιθέμενων μορίων, που διαφέρουν σε μήκος κατά ένα νουκλεοτίδιο. Το μείγμα αυτό διαχωρίζεται με ηλεκτροφόρηση και με την βοήθεια υπολογιστικών προγραμμάτων προκύπτουν χρωματογράμματα, τα οποία δείχνουν κορυφές που αντιπροσωπεύουν το χρώμα και την ένταση του σήματος κάθε φθορίζουσας βάσης και έτσι προσδιορίζεται η αλληλουχία του κλώνου DNA (Hood and Galas, 2003).

1.5.2. RNA Sequencing (Αλληλουχιση RNA)

Το πρώτο πρωτόκολλο αλληλούχισης RNA, εμφανίστηκε το 2009 και έκτοτε αποτελεί μία από τις πιο ευέλικτες τεχνικές, με την οποία μπορεί να επιτευχθεί αλληλούχιση του συνολικού γονιδιώματος, κερδίζοντας μεγάλη δημοφιλία. Ανήκει στις τεχνικές της αλληλούχισης του ολικού μεταγραφώματος (Whole Transcriptome Sequ-

encing), και των ισόμορφων, τα οποία προέρχονται από γειτονικά ανοιχτά πλαίσια ανάγνωσης (Open Reading Frames (ORFs)) και εμφανίζει πολλαπλά πλεονεκτήματα, σε σχέση με την αλληλούχιση DNA. Ένα ισχυρό πλεονέκτημα της συγκεκριμένης τεχνικής, είναι το γεγονός ότι είναι ικανή να διακρίνει τα μετάγραφα που προέρχονται από τον αντιπαράλληλο κλώνο (antisense), αλλά και να οδηγήσει στην αναγνώριση νέων μεταγράφων, τα οποία δεν έχουν σχολιαστεί στο παρελθόν (Z. Wang *et al.*, 2009). Επομένως, υπάρχει η δυνατότητα για διαχωρισμό των μεταγράφων, ανάλογα με την αλυσίδα που αυτά προέκυψαν, γεγονός που αυξάνει την ευαισθησία της αλληλούχισης, καθώς έχει βρεθεί ότι μερικά μετάγραφα, προκύπτουν από το συμπληρωματικό κλώνο (Mills *et al.*, 2013). Η βασική προετοιμασία του δείγματος, περιλαμβάνει την απομόνωση όλων των μορίων RNA και εν συνεχεία τη θρυμματοποίηση αυτών, όπως αναφέρθηκε παραπάνω. Ακολούθως, ανάλογα με τον αλληλουχητή που χρησιμοποιείται, επιλέγονται αντάπτορες και συνδέονται στα άκρα των θραυσμάτων, ενώ απορρίπτονται τα θραύσματα με μικρό μήκος, ή ελεύθερα άκρα, δηλαδή άκρα χωρίς αντάπτορες. Στη συνέχεια, με τη βοήθεια της PCR, το δείγμα ενισχύεται και αλληλουχείται *(Σχήμα 1.4)*. Το περιεχόμενο των ανταπτόρων μπορεί μερικές φορές να είναι πολύπλοκο, δεδομένου ότι μπορεί να περιέχει τμήματα για τη σύνδεσή τους στην επιφάνεια που πραγματοποιείται η ενίσχυση ή τμήματα που χρησιμοποιούνται ως barcodes για τη μετέπειτα ευρετηρίαση των αναγνώσεων. Αξίζει επίσης να αναφερθεί, ότι η ποσότητα προς αλληλούχιση εξαρτάται από την αρχική ποσότητα των RNA του δείγματος και αυτό κατ΄ επέκταση στηρίζεται στο μεταγραφικό προφίλ των κυττάρων, το οποίο μπορεί να έχει επηρεαστεί από τις τρέχουσες ανάγκες κάθε κυττάρου και την κατάσταση της ανοσολογικής τους απόκρισης. Έτσι, η αλληλούχιση γίνεται συναρτήσει των τρεχουσών αναγκών του κυττάρου και αυτή είναι η βάση για την ανάλυση διαφορικής έκφρασης γονιδίων, μεταξύ δειγμάτων.



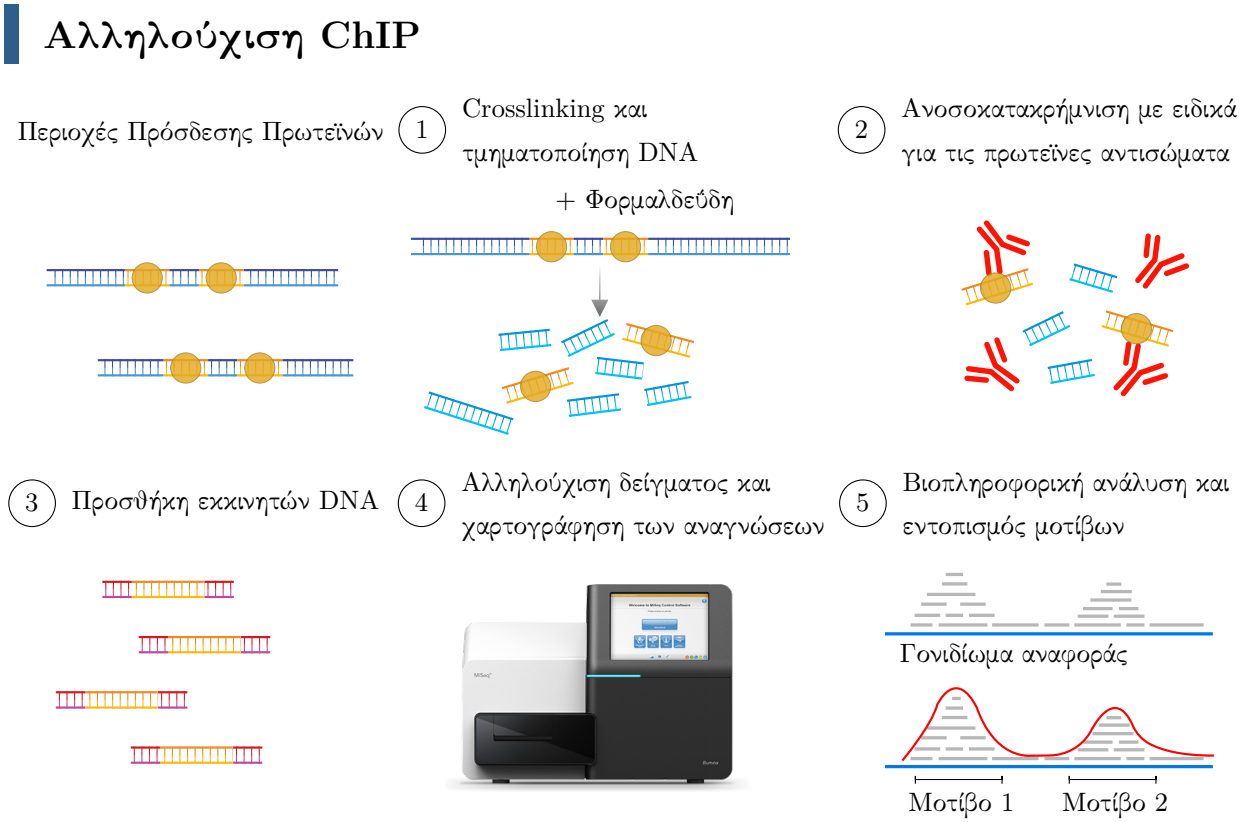
Σχήμα 1.4: *Μεθοδολογία μιας RNA αλληλούχισης, από το αρχικό δείγμα, έως την ανάλυση των δεδομένων.*

Μερικά ακόμη από τα πλεονεκτήματα της αλληλούχισης RNA που την έχουν καταστήσει μία τόσο σημαντική τεχνική στη βιολογική έρευνα, είναι ότι υποστηρίζει δυναμικό εύρος των επιπέδων έκφρασης, σε σχέση με τις μικροσυστοιχίες. Οι τάξεις μεγέθους των αναγνώσεων σε μία γονιδιακή περιοχή, μπορεί να διαφέρει ακόμα και κατά χιλιάδες αναγνώσεις, σε σχέση με κάποια άλλη γονιδιακή περιοχή, γεγονός που δεν ήταν εφικτό στις μικροσυστοιχίες. Έτσι, με προσαρμοστική κανονικοποίηση των δεδομένων, μία τεχνική που όπως θα αναλυθεί στη

μεθοδολογία είναι μείζονος σημασίας για τη διαχείριση τέτοιων δειγμάτων, μπορούν να εξαχθούν συμπεράσματα, είτε για τη στατιστική σημαντικότητα των χαμηλά εκφρασμένων περιοχών, εάν γίνεται σύγκριση με κάποιο δείγμα ελέγχου, είτε απλά για την ποσοτικοποίηση αυτών των περιοχών, χωρίς να υπάρχει σε μεγάλο βαθμό η υπερεκπροσώπηση κάποιας άλλης υψηλά εκφρασμένης περιοχής (Z. Wang *et al.*, 2009).

1.5.3. ChIP Sequencing (Αλληλοτχιση με Ανοσοκατακρημνιση Χρωματινης)

Η τεχνική της ανοσοκατακρήμνισης χρωματίνης, ανήκει σε μία διαφορετική κατηγορία τεχνικών σε σχέση με την RNA-seq, και χρησιμοποιείται για τον προσδιορισμό των αλληλεπιδράσεων, μεταξύ πρωτεϊνών και DNA (Johnson *et al.*, 2007). Αν και οι πρώτες έρευνες στον εντοπισμό αλληλεπιδράσεων πρωτεϊνών - DNA υπάρχουν στη βιβλιογραφία από τα μέσα της δεκαετίας του ’80, η αλληλούχιση ChIP προτάθηκε το 2007, ως τεχνική που βασίζεται στην αλληλούχιση δεύτερης γενιάς για την ανάλυση αυτών των αλληλεπιδράσεων σε ολόκληρο το γονιδίωμα, με ένα απλό πείραμα και με αρκετά ακριβή ανάλυση, που φτάνει τη μία βάση (Marinov, 2018). Από τότε έχουν υπάρξει συνεχείς βελτιώσεις στη συγκεκριμένη τεχνική, οι οποίες εγγυώνται τον ακριβή εντοπισμό των σημείων, στα οποία προσδένονται ρυθμιστικά στοιχεία της μεταγραφής, όπως οι μεταγραφικοί παράγοντες, η RNA πολυμεράση και τα διάφορα στοιχεία των ενισχυτών στους ευκαρυωτικούς οργανισμούς. Μερικές από αυτές, έχουν χαρακτηριστεί ως ChIP-exo, ChIP-nexus και άλλες (Van Dijk *et al.*, 2014b).



Σχήμα 1.5: *Μεθοδολογία της διαδικασίας αλληλούχισης με Ανοσοκατακρήμνιση χρωματίνης, από το αρχικό δείγμα, έως την ανάλυση των δεδομένων.*

Η βασική ιδέα είναι ότι με την απομόνωση του δείγματος, το γονιδίωμα περιέχει ή δύνανται να περιέχει σε διακριτά σημεία του, προσκολλημένες πρωτεΐνες και το βιολογικό πρόβλημα που σχετίζεται με αυτή την τεχνική, είναι ο εντοπισμός μοτίβων που φέρουν αυτές οι περιοχές πρόσδεσης πρωτεϊνών. Στην αρχική εκδοχή της ChIP

αλληλούχισης, τα κύτταρα επωάζονται σε φορμαλδεΰδη, ώστε να προσδεθούν ισχυρότερα στο DNA και το γενετικό υλικό τμηματοποιείται σε θραύσματα μήκους 200-600 ζευγών βάσεων. Στο επόμενο στάδιο, χρησιμοποιείται ένα ειδικό αντίσωμα που στοχεύει μία πρωτεΐνη ενδιαφέροντος, ώστε να επιτευχθεί η ανοσοκατακρήμνιση του συμπλόκου DNA - πρωτεΐνη και να διαχωριστεί από μη δεσμευμένο σε πρωτεΐνη DNA *(Σχήμα 1.5)*. Το τελικό στάδιο πριν την αλληλούχιση, είναι η επιλογή μόνο εκείνων των συμπλόκων με μήκος περίπου 150 ή παραπάνω ζευγών βάσεων (P. J. Park, 2009). Αντίθετα, στην τεχνική ChIP-exo, ακολουθείται ένα διαφορετικό πρωτόκολλο, στο οποίο μετά την σύνδεση των ανταπτόρων με τα άκρα των αναγνώσεων, γίνεται χρήση 5′ - 3′ εξωνουκλεασών στο δείγμα, αποικοδομούνται όλα τα ελεύθερα σημεία του DNA και μένουν μόνο αυτά που είναι προσδεδεμένα στις πρωτεΐνες. Με πλύση του δείγματος, αφαίρεση των πρωτεϊνών και αλληλούχιση DNA, μπορούν να εξαχθούν αυτές οι περιοχές και να αναλυθούν *in silico*, εξάγοντας πιθανά μοτίβα (Johnson *et al.*, 2007).

1.5.4. Small RNA Sequencing (Αλληλουχιση Small RNA)

Η αλληλούχιση μικρών μη-κωδικών RNA (Small RNA-Seq), αποτελεί κατηγορία αλληλούχισης RNA, η οποία χρησιμοποιεί τεχνολογίες αλληλούχισης επόμενης γενιάς. Η ανάλυση των δεδομένων αλληλούχισης αντλεί πληροφορίες για την ανάλυση, ποσοτικοποίηση και ανίχνευση μικρών μη-κωδικών RNA, όπως μικρά-RNA (*mi*RNAs), piwi-αλληλεπιδρώντα-RNA (*pi*RNAs), μικροπυρηνικά-RNA (*sn*RNAs) και μικρά παρεμβαλλόμενα-RNA (*si*RNAs) των ευκαρυωτικών οργανισμών (Z. Wang *et al.*, 2009). Η αρχή της μεθόδου περιλαμβάνει απομόνωση των μικρών RNA-στόχων από τα δείγματα, μέσω της χρήσης λιγάσης, κατασκευή βιβλιοθήκης μικρών RNA-στόχων και αλληλούχιση αυτών. Μετά την αλληλούχιση, εφαρμόζονται εξειδικευμένα βιοπληροφορικά και στατιστικά εργαλεία για την ταυτοποίηση και ποσοτικοποίηση των μικρών RNA-στόχων (Giurato *et al.*, 2013).

Τα μικρά RNA συμμετέχουν σε σημαντικές κυτταρικές διεργασίες όπως η παρεμβολή RNA, το μάτισμα, η μετάφραση και η μετα-μεταγραφική ρύθμιση της γονιδιακής έκφρασης (Wells *et al.*, 1998). Η εφαρμογή της τεχνολογίας αλληλούχισης μικρών μη-κωδικών RNA, καθίσταται αναγκαία για την μελέτη του ρόλου τους στις κυτταρικές διαδικασίες που ρυθμίζουν. Συγκεκριμένα, οι εφαρμογές επικεντρώνονται στην ανακάλυψη νέων μορφών μικρών RNA και στον προσδιορισμό της λειτουργίας τους, στην διαφοροποίηση των μικρών RNA από τα υπόλοιπα RNA, στην διαφορική έκφραση όλων των μικρών RNA στο εκάστοτε δείγμα και στην ανάλυση τους με ακρίβεια και μεγάλη απόδοση (Z. Wang *et al.*, 2009; Zhou *et al.*, 2011).

1.5.5. Cappable Sequencing (Αλλ/ση σεσημασμενων ακρων)

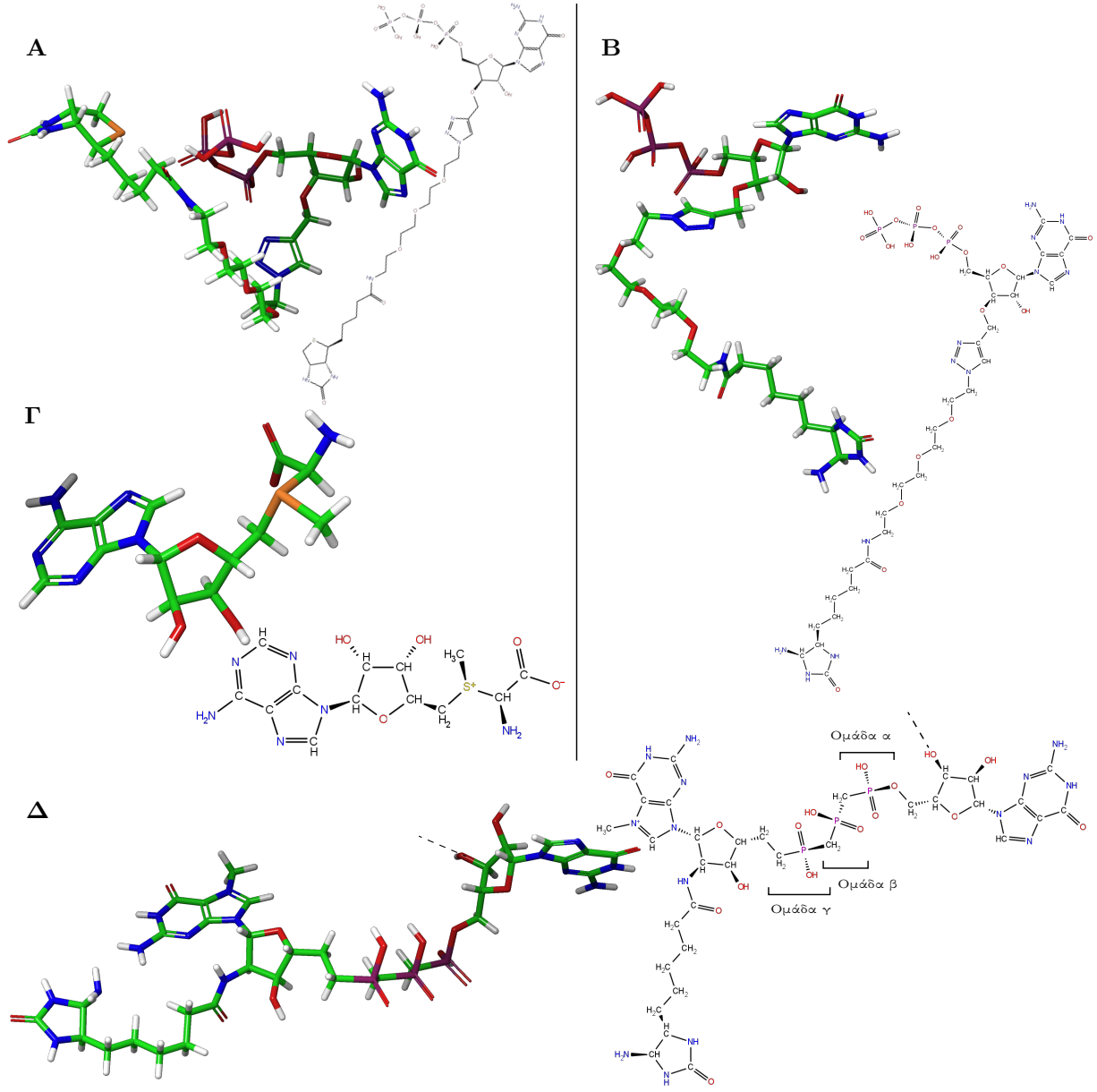
Ο εντοπισμός της σήμανσης των RNA στο 5’ άκρο τους, ήταν μία προσπάθεια που χρονολογείται από τα μέσα του 1970. Οι πρώτες εκτιμήσεις, έκαναν λόγο ότι αυτή η σήμανση δηλώνει την ταυτότητα, ή τη διαδοχή των βιολογικών διαδικασιών ενός μορίου RNA, αναφορικά με τη λειτουργία που αυτό επιτελεί. Οι πρώτες βιοχημικές αναλύσεις, έδειξαν ότι οι ευκαρυωτικοί οργανισμοί και ορισμένοι ιοί, σημαίνουν τα *m*RNA, με τέτοιο τρόπο, δημιουργώντας ένα πρόσθετο άκρο 5’ τριφωσφορικής γουανοσίνης (5’-Gppp) (Wei *et al.*, 1975). Η συγκεκριμένη σήμανση, όπως θα αναλυθεί και σε παρακάτω ενότητα, σχετίζεται με την αναγνώριση αυτών των RNA, έναντι των υπολοίπων, ώστε να επιτελούνται στοχευμένα διαδικασίες, όπως το μάτισμα ή το εναλλακτικό μάτισμα, η πολυαδενυλίωση των *m*RNA, η έξοδός τους από το κυτταρόπλασμα και η πρωτεϊνοσύνθεση (Topisirovic *et al.*, 2011). Έως σήμερα, η σήμανση των RNA στο 5’ άκρο, αλλά και οι γενικότερες τροποποιήσεις αυτών, σε όλο το μήκος τους, παραμένουν ένα πολύ ενεργό αντικείμενο μελέτης, το οποίο μπορεί να οδηγήσει στην περαιτέρω κατανόηση του τρόπου οργάνωσης και λειτουργίας τους. Σε αυτή τη σήμανση στηρίχθηκαν αρκετές μεθοδολογίες ενίσχυσης των σεσημασμένων άκρων, όπως η *d*RNA-seq (C. Sharma *et al.*, 2010), η NAD Capture-seq και η αρκετά βελτιωμένη μέθοδος *Cappable-seq*, η οποία αναλύεται εκτενώς παρακάτω.

Η αλληλούχιση *Cappable-seq* (Ettwiller *et al.*, 2016), χρησιμοποιείται για τον εντοπισμό των RNAs που δεν έχουν υποστεί επεξεργασία στα άκρα τους, δηλαδή μετα-μεταγραφικές τροποποιήσεις. Πρόκειται για μία αρκετά νέα και καινοτόμα τεχνική αλληλούχισης επόμενης γενιάς, η οποία έχει χαρακτηριστεί για την ευαισθησία της, στο να εντοπίζει με βιοχημικό τρόπο, τα σημεία έναρξης της μεταγραφής, των προκαρυωτικών (αλλά και ευκαρυωτικών (Yan *et al.*, 2022)) οργανισμών. Η βασική έννοια για τον εντοπισμό άθικτων μορίων RNA έναντι των επεξεργασμένων, έγκειται στο γεγονός ότι στους προκαρυωτικούς οργανισμούς, το πρώτο νουκλεοτίδιο (5' άκρο) των ανεπεξέργαστων RNA, είναι πάντα τριφωσφορυλιωμένο (5' triphosphorylated transcript), όπως ακριβώς προκύπτει από την RNA πολυμεράση (Σχήμα 1.6A). Έτσι, εντοπίζοντας τα βιομόρια με τριφωσφορυλιωμένο (ή διφωσφορυλιωμένο) 5' άκρο και κάνοντας στοίχιση με το γονίδιο από το οποίο προήλθε, μπορεί να εντοπιστεί η περιοχή από την οποία ξεκίνησε ακριβώς η μεταγραφή, δηλαδή το σημείο έναρξης της μεταγραφής (TSS). Η τεχνική αυτή προτάθηκε το 2016, από τα New England Biolabs (Ettwiller *et al.*, 2016).

Η μεθοδολογία αλληλούχισης *Cappable-seq* που παρουσιάζεται, διακρίνει αυτά τα RNAs από το εάν έχουν συντηχθεί με ένα παράγωγο βιοτίνης, δηλαδή αν έχουν βιοτινυλιωμένο 5' άκρο, γεγονός που σημαίνει ότι είναι άθικτα (Ettwiller *et al.*, 2016). Το παράγωγο βιοτίνης που χρησιμοποιείται, είναι η λεγόμενη 3' Δεθειοβιοτίνη-GTP (3'-DTB-GTP), όπου GTP είναι η 5'-τριφωσφορική γουανοσίνη. Για να γίνουν οι ουσίες αυτές πιο κατανοητές, η αρχική τους δομή είναι ένα απλό νουκλειτίδιο Γουανίνης (G), όμως φέρουν στο 3' άκρο τους, δεθειοβιοτίνη, ενώ ο 5' άνθρακας είναι συνδεδεμένος με μία τριφωσφορική ομάδα (Σχήμα 1.6B). Να αναφερθεί επίσης, πως η βιοτίνη *(Σχήμα 1.6A)* είναι η γνωστή βιταμίνη Η ή Β7, ενώ η δεθειοβιοτίνη, έχει παρόμοια χημική δομή με τη βιοτίνη, με τη μόνη διαφορά ότι δεν περιέχει Θείο (S). Κατά την επώαση του δείγματος με δεθειοβιοτίνη, χρησιμοποιείται το σύστημα σήμανσης Vaccinia (Vaccinia Capping Enzyme (VCE)), το οποίο κάνει ενζυμική σήμανση των τριφωσφορυλιωμένων άκρων, με τη βοήθεια των τρανσφερασών, RNA τριφωσφατάση (RNA Triphosphatase), RNA γουανυλυλοτρανσφεράση (RNA Guanylyltransferase) και Μεθυλοτρανσφεράση Γουανίνης-N7 (Guanine-N7 Methyltransferase), μέσα από 3 αντιδράσεις που καταλύονται. Πιο αναλυτικά, η RNA τριφωσφατάση, υδρολύει 1 μόριο φωσφόρου, αυτό της ομάδας Γάμα, μετατρέποντας το RNA από τριφωσφορυλιωμένο, σε διφωσφορυλιωμένο (Takagi *et al.*, 1997). Στη συνέχεια, η RNA γουανυλυλοτρανσφεράση, καταλύει την αντίδραση της 3'-Δεθειοβιοτίνη-GTP με το διφωσφορυλιωμένο RNA και έτσι παράγεται RNA με 5' άκρο το 3'-Δεθειοβιοτίνη-5'-τριφωσφορική-γουανοσίνη και απελευθερώνεται μία πυροφωσφορική ένωση (Gross and Shuman, 1998) *(Σχήμα 1.7)*. Τέλος, η Μεθυλοτρανσφεράση Γουανίνης, καταλύει την αντίδραση ανάμεσα στο RNA που παράχθηκε από την RNA γουανυλυλοτρανσφεράση και την S-αδενοσυλμεθειονίνη (S-adenosylmethionine) *(Σχήμα 1.6Γ)* και παράγει το τελικό επικαλυμμένο και μεθυλιωμένο πλέον RNA, με επιπλέον κάλυμμα στο 5' άκρο, το 5'-m^7-μεθυλογουανίνη, άρα τελικό συνολικό 5' άκρο, το 3'-Δεθειοβιοτίνη-5'-m^7-μεθυλογουανίνη ή ομοίως 3'-DTB-5'-m^7Gppp *(Σχήμα 1.6Δ)* (Gross and Shuman, 1998). Η παρουσία της S-αδενοσυλμεθειονίνης, πέρα από την αντίδραση που καταλύει, βοηθά και στην αποδοτική σήμανση του RNA με την 3'-DTB-GTP.

Αναφορά, επίσης αξίζει και στην 7-μεθυλογουανίνη (m^7G), η οποία είναι ένα τυπικό, αλλά μεθυλιωμένο νουκλεοτίδιο Γουανίνης. Σύμφωνα με έρευνα (Reynaud *et al.*, 1992), χρησιμοποιείται ως δείκτης προγνωστικής αξίας σε διάφορους τύπους καρκίνου, με βάση τα ποσοστά αυτού, αλλά και άλλων μεθυλιωμένων ή ακετυλιωμένων νουκλεοτιδίων. Ακόμη, το σύμπλοκο ενζύμων σήμανσης Vaccinia, αποτελείται από δύο υπομονάδες πεπτιδίων, τις D1 και D12, οι οποίες καταλύουν τις 3 αντιδράσεις που αναφέρθηκαν παραπάνω. Πιο συγκεκριμένα, η υπομονάδα D1, που είναι μικρότερη σε σχέση με την D12, περιέχει τις τρανσφεράσες RNA τριφωσφατάση, RNA γουανυλυλοτρανσφεράση και RNA Μεθυλοτρανσφεράση Γουανίνης-N7, ενώ η χρησιμότητα της υπομονάδας D12, είναι να δεσμεύει και να προσδίδει λειτουργικότητα στην μεθυλοτρανσφεράση, την οποία και διεγείρει αλλοστερικά (Kyrieleis *et al.*, 2014). Ο μηχανισμός σήμανσης VCE, προέρχεται από τον ιό Vaccinia και σχετίζεται με τη σήμανση προκαρυωτικών RNA με το ευκαρυωτικό πρότυπο σήμανσης, ώστε αυτά να μην αποικοδομηθούν από τις εξωνουκλεάσες όταν εισαχθούν στον ευκαρυωτικό ξενιστή και να μπορέσουν να δράσουν σε αυτόν.

Το επόμενο στάδιο μετά τη σήμανση των τριφωσφορυλιωμένων RNA, είναι ο εμπλουτισμός του δείγματος, ώστε με συγκεκριμένο βιοχημικό τρόπο, να διακριθούν τα επικαλυμμένα με βιοτίνη RNA, από αυτά χωρίς επι-

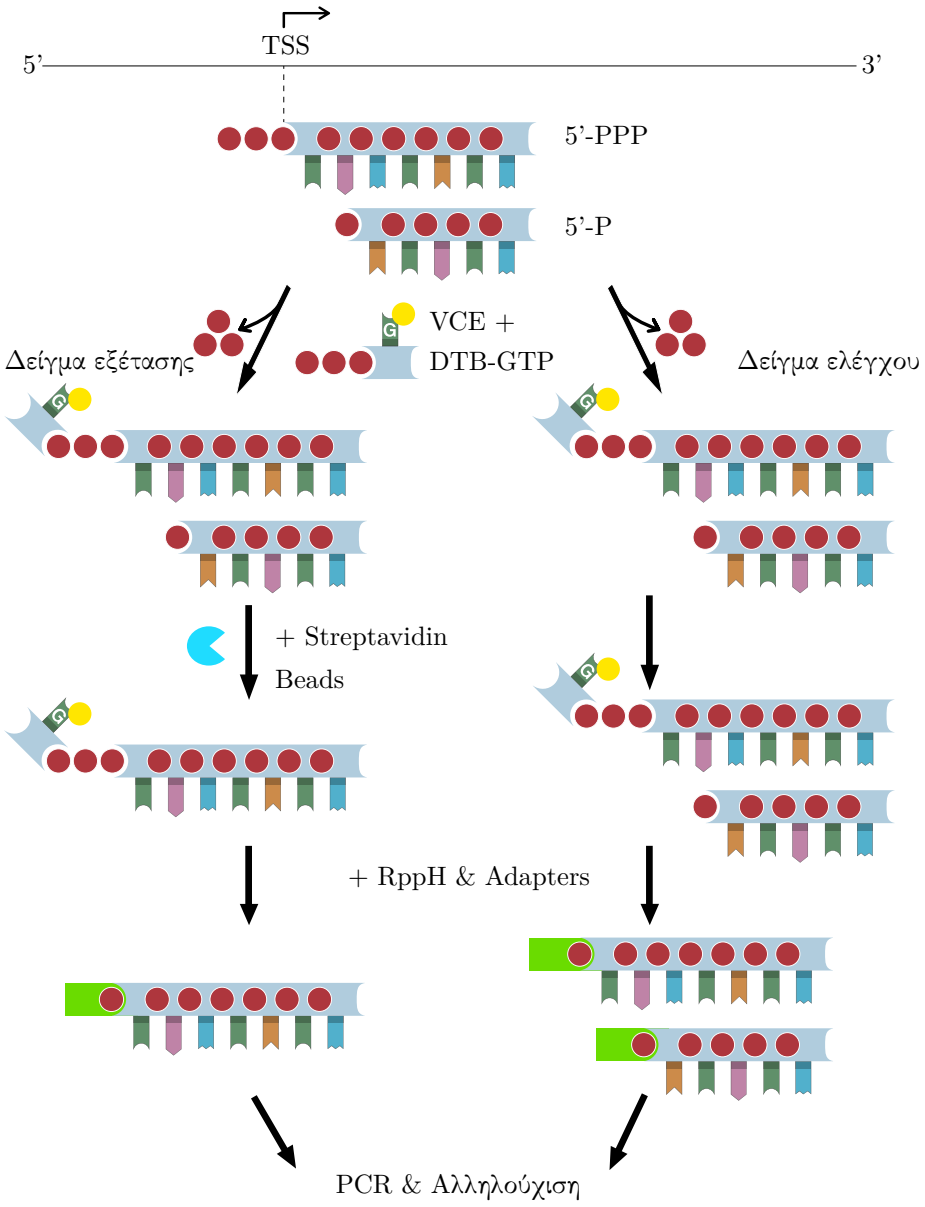


Σχήμα 1.6: *Βιοχημική σύσταση και στερεοδιάταξη των μορίων που χρησιμοποιούνται για τη σήμανση των RNA.* ***A****: 3'-Βιοτίνη-GTP.* ***B****: 3'-Δεθειοβιοτίνη-GTP.* ***Γ****: S-Adenosyl methionine.* ***Δ****: 3'-Δεθειοβιοτίνη-5'-m^7-μεθυλογουανίνη. Η υπόλοιπη RNA αλυσίδα, επεκτείνεται από τη δεξιά ριβόζη, στο 3' ελεύθερο άκρο της (διακεκομμένη γραμμή). Με πράσινο χρώμα απεικονίζονται οι άνθρακες, με κόκκινο το οξυγόνο, με μπλε το άζωτο, με πορτοκαλί το θείο, με λευκό το υδρογόνο και με μωβ ο φώσφορος. Οι παραπάνω εικόνες δημιουργήθηκαν με το λογισμικό Schrödinger Maestro (Schrödinger, LLC, 2021b).*

κάλυμμα. Για το λόγο αυτό, το δείγμα εκλούεται με υδρόφιλα μαγνητικά σφαιρίδια στρεπταβιδίνης (και βιοτίνη), τα οποία έχουν τη δυνατότητα να προσδένονται με ιδιαίτερα ισχυρό τρόπο στο 5'-βιοτινυλιωμένο άκρο, σχηματίζοντας το αντίστοιχο σύμπλοκο. Πλέον, τα RNA που μας ενδιαφέρει να διαχωρίσουμε, έχουν αποκτήσει μαγνητικές ιδιότητες και μπορούν να διαχωριστούν με φιλτράρισμα σε μαγνητικές πλάκες. Έτσι, είναι πολύ πιο εύκολος ο διαχωρισμός των σεσημασμένων RNA, γεγονός που συνεπάγεται τον ταυτόχρονο διαχωρισμό των ώριμων (επεξεργασμένων) *r*RNA & *t*RNA, τα οποία αντιστοιχούν σε ένα πολύ μεγάλο ποσοστό του δείγματος, ακόμα και της τάξης του 95%, χωρίς να περιέχουν τόσο σημαντική πληροφορία, όσο τα *m*RNA και Small RNA (*s*RNA). Το επόμενο στάδιο, αφού πραγματοποιήθηκε η διατήρηση των σεσημασμένων RNA, είναι η αφαίρεση του επικαλύμματος (de capping) και η προετοιμασία του δείγματος για αλληλούχιση, με την προσθήκη 5' και 3' ανταπτόρων. Για το σκοπό αυτό, χρησιμοποιείται η πρωτεΐνη RNA πυροφωσφοϋδρολάση (RppH) των προκαρυωτικών οργανισμών,

η οποία έχει τη δυνατότητα να αποικοδομεί το ριβονουκλεϊκό επικάλυμμα, τόσο το 3'-DTB, όσο και το 5'-m^7Gpp, δημιουργώντας 5' μονοφωσφορυλιωμένα άκρα, αφού πρώτα αφαιρεθούν τα σφαιρίδια στρεπταβιδίνης με τη χρήση διαλύματος αιθανόλης. Τέλος, το δείγμα ενισχύεται σε Polymerase chain reaction (PCR) για περίπου 15 κύκλους και είναι έτοιμο για αλληλούχιση (Ettwiller *et al.*, 2016).

Για να επιτευχθεί σύγκριση και να εξαχθούν συμπεράσματα, αναφορικά με την ευαισθησία αυτής της τεχνικής αλληλούχισης, σε σχέση με ένα μη εμπλουτισμένο δείγμα, χρησιμοποιείται ένα δείγμα ελέγχου. Η διαδικασία προετοιμασίας αυτού, είναι όμοια με αυτήν του δείγματος *Cappable-seq*, με τη μόνη διαφορά να είναι στο ότι μετά τη σήμανση με δεθειοβιοτινυλιωμένη καλύπτρα, δεν γίνεται έκλουση του δείγματος με σφαιρίδια στρεπταβιδίνης. Αυτό σημαίνει, πως δεν γίνεται μαγνητικός διαχωρισμός των RNA με βάση το εάν φέρουν ή όχι επικάλυμμα στο 5' άκρο και έτσι, λαμβάνονται και αλληλουχούνται τα συνολικά θραύσματα RNA *(Σχήμα 1.7)*.



Σχήμα 1.7: *Διαδικασία εμπλουτισμού του δείγματος με τη μεθοδολογία Cappable-seq. Χρησιμοποιείται το σύμπλοκο ενζύμων VCE, το οποίο αρχικά υδρολύει με την RNA τριφωσφατάση (TPase) την φωσφορική ομάδα γ, μετά η RNA γουανυλυλοτρανσφεράση (GTase) καταλύει την αντίδραση μεταξύ των δεθειοβιοτινυλιωμένων GTP, απελευθερώνοντας ένα πυροφωσφορικό προϊόν και τέλος, η Μεθυλοτρανσφεράση Γουανίνης-N7 (N7MTase) μεθυλιώνει το 5΄άκρο. Ακολούθως, στο δείγμα εξέτασης γίνεται επώαση με υδρόφιλα σφαιρίδια στρεπταβιδίνης, για το διαχωρισμό των σεσημασμένων από τα μη σεσημασμένα RNA και με τη βοήθεια της πρωτεΐνης RppH, γίνεται η αφαίρεση της καλύπτρας και αλληλούχιση.*

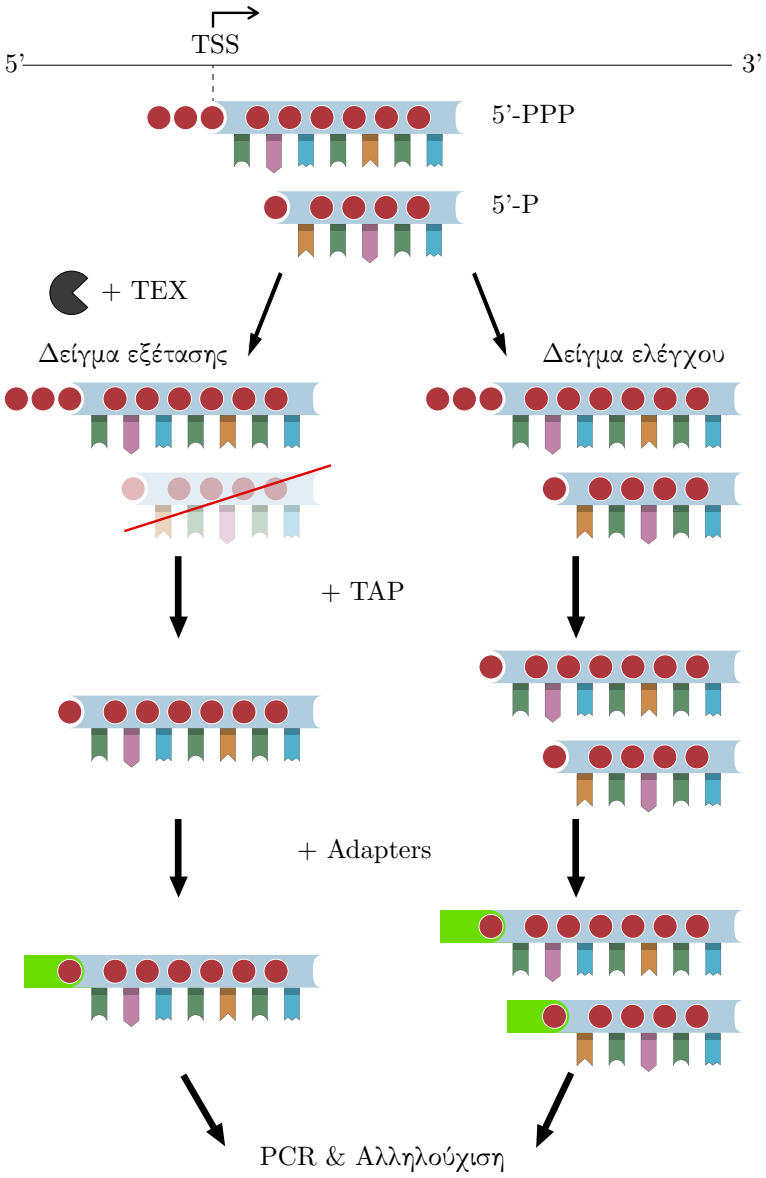
1.5.6. *d*RNA Sequencing (Διαφορικη Αλληλουχιση RNA)

Μία προγενέστερη τεχνική της *Cappable-seq* είναι η Differential RNA Sequencing (*d*RNA-seq), η οποία προτάθηκε το 2010 και επιτυγχάνει και αυτή το διαχωρισμό των 5' τριφωσφορυλιωμένων άκρων, αλλά με διαφορετικό ενζυμικό τρόπο. Η συγκεκριμένη τεχνική, χρησιμοποιήθηκε με πολλή επιτυχία για πρώτη φορά στον εντοπισμό των non-coding RNA, στο βακτήριο *Helicobacter pylori* και οδήγησε στο σχολιασμό πολλών small RNA, οπερονίων και TSS (C. Sharma *et al.*, 2010). Πιο αναλυτικά, χρησιμοποιείται συνήθως η εξωνουκλεάση με όνομα Terminator™ 5' monophosphate-dependent exonuclease (TEX), ή η Xrn1 και οι δύο εκ των οποίων, στοχεύουν και αποικοδομούν μόνο RNA με 5' μονοφωσφορυλιωμένα άκρα, δηλαδή αυτά που έχουν υποστεί μετα-μεταγραφική επεξεργασία *(Σχήμα 1.8)*, κάνοντας την ενίσχυση του δείγματος με τα άθικτα (primary) μετάγραφα (δείγμα TEX+) (C. Sharma *et al.*, 2010). Παράλληλα, ως δείγμα ελέγχου χρησιμοποιείται μη επωασμένο με την εξωνουκλεάση TEX δείγμα, το οποίο χαρακτηρίζεται ως TEX–. Στο αρχικό εμπλουτισμένο δείγμα, εισάγεται το ένζυμο Tobacco Acid Pyrophosphatase (TAP), το οποίο μετατρέπει τα τριφωσφορυλιωμένα άκρα των RNA, σε μονοφωσφορυλιωμένα, ώστε να ακολουθήσει η διαδικασία της αντίστροφης μεταγραφής (C. M. Sharma and Vogel, 2014). Με αυτό τον τρόπο, μπορούν να εξαχθούν συμπεράσματα, με βάση κάποια στατιστική κατανομή, σχετικά με τη στατιστική σημαντικότητα των αναγνώσεων στην κάθε γονιδιακή περιοχή, συγκρίνοντας τον αριθμό αυτών στα δύο δείγματα. Να αναφερθεί επίσης, ότι όπως και στην *Cappable-seq*, έτσι και στην *d*RNA-seq, γίνεται αποικοδόμηση των ριβοσωμικών RNA και των *t*RNA, λόγω του ότι συνήθως φέρουν επεξεργασμένο 5' άκρο, γεγονός που είναι σημαντικό για τον εμπλουτισμό του δείγματος, καθώς αθροιστικά αυτά τα RNA, αντιπροσωπεύουν περίπου ακόμα και το 90% του δείγματος.

Αν και η *d*RNA-seq ήταν η πρώτη προσέγγιση που επέτρεψε τη μαζική ανίχνευση και σχολιασμό, κυρίως των TSS, αλλά και των οπερονίων, εμφανίζει ορισμένα μειονεκτήματα στη γενίκευση της τεχνικής σε όλα τα βακτήρια, καθώς και στην ευαισθησία της. Πρώτων, η ενίσχυση του δείγματος με την αποικοδόμηση των ριβοσωμικών RNA, δεν είναι τόσο αποδοτική σε σχέση με την *Cappable-seq*, όπου στην τελευταία χρησιμοποιούνται ισχυρότερα ένζυμα και το δείγμα υφίσταται καλύτερη πέψη αυτών, το οποίο συνεπάγεται με ακριβέστερα και πιο αξιόπιστα αποτελέσματα. Επιπλέον, υπάρχει ένα σύνολο βακτηρίων, στα οποία η *d*RNA-seq δεν ενδείκνυται να εφαρμοστεί, λόγω της μη ικανοποιητικής ακρίβειάς της. Αυτό συμβαίνει, γιατί βακτήρια όπως τα *Pseudomonas aeruginosa, Mycobacterium tuberculosis, Bifidobacterium longum* και κυρίως το *Streptomyces coelicolor*, περιέχουν στο γονιδίωμά τους υψηλό ποσοστό GC, τα λεγόμενα *CpG islands (Σχήμα 1.9)* που κυμαίνεται από 61% έως και 72%. Αυτό συνεπάγεται μία πιο ισχυρή δευτεροταγή δομή των RNA, τα οποία πλέον σχηματίζουν περισσότερες φουρκέτες (Stem-loops) και σε σημεία πιο κοντά στα άκρα τους. Για το λόγο αυτό, οι διάφορες εξωνουκλεάσες, όπως η TEX, δεν είναι τόσο αποδοτικές στο να εντοπίζουν αυτά τα άκρα και να αποκοδομούν τα 5′ μονοφωσφορυλιωμένα μόρια, οδηγώντας σε λανθασμένες συγκρίσεις μεταξύ των δύο δειγμάτων (Romero *et al.*, 2014; Jäger *et al.*, 2014). Τέλος, η *d*RNA-seq δεν μπορεί να χρησιμοποιηθεί αποτελεσματικά για την ανάλυση των TSS στους ευκαρυωτικούς οργανισμούς, επειδή σε αυτούς, η σήμανση των RNA είναι περισσότερο πολύπλοκη και δεν μπορεί να δράσει η εξωνουκλεάση TEX, ώστε επιλεκτικά να τα αποικοδομήσει, ή μπορεί λόγω της ύπαρξης φουρκέτας στο 5′ άκρο, δηλαδή ενισχυμένης δευτεροταγούς δομής, να μη μπορέσει να προσκολληθεί και να δράσει σε αυτό (C. Sharma *et al.*, 2010).

1.5.7. Term Sequencing (Αλληλουχιση 3' Αμ/στων περιοχων)

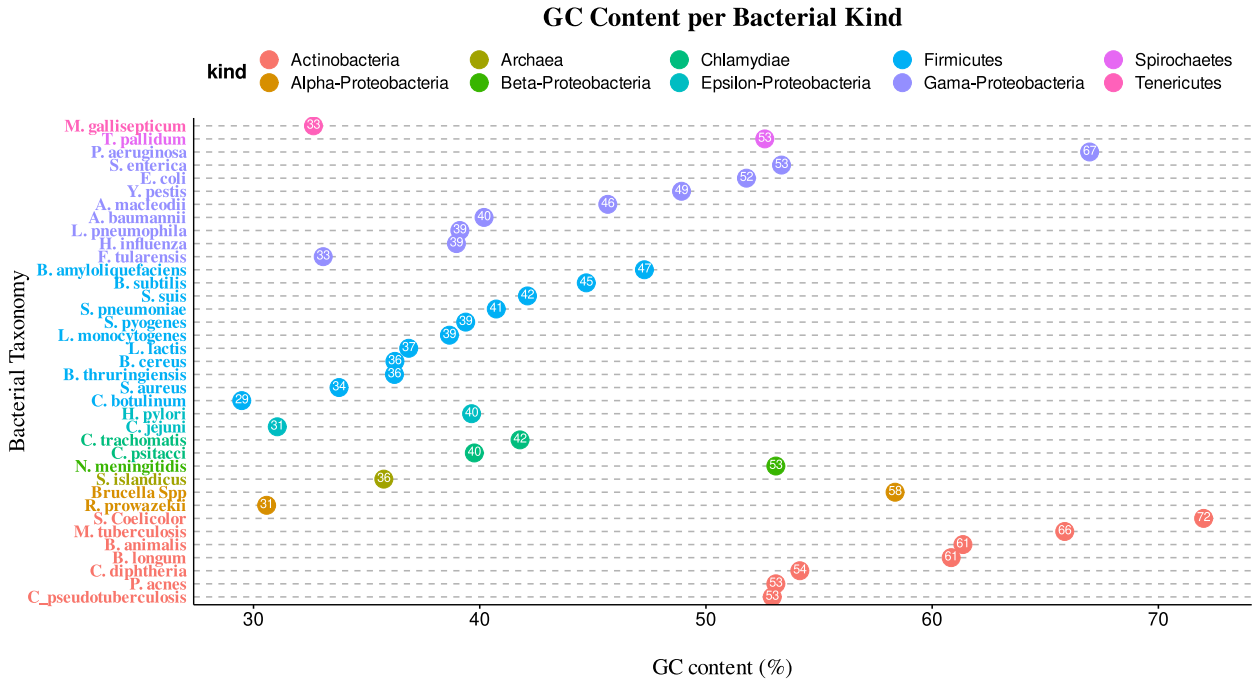
Η 3' αμετάφραστη περιοχή (3' UTR), ονομάζεται και μετατερματική αλληλουχία και αποτελεί περιοχή *m*RNA, που ακολουθεί μετά το κωδικόνιο λήξης. Η αλληλούχιση 3' αμετάφραστων περιοχών (Term-seq), έχει ως στόχο να αποκαλύψει τον ρόλο που παίζει στην γονιδιακή έκφραση, συμβάλλοντας στη σταθερότητα, στην εξαγωγή και στην αποτελεσματικότητα της μετάφρασης του *m*RNA. Η μέθοδος περιλαμβάνει ολική απομόνωση των RNA



Σχήμα 1.8: *Διαδικασία εκτέλεσης της differential RNA-seq. Το δείγμα εξέτασης επωάζεται αρχικά με τερματική εξωνουκλεάση (TEX), η οποία αποικοδομεί μη σεσημασμένα RNA και στη συνέχεια το επικάλυμμα αφαιρείται με τη χρήση του ενζύμου Tobacco Acid Pyrophosphatase (TAP). Αφετέρου, οι αναγνώσεις συνδέονται με αντάπτορες και το δείγμα ενισχύεται και αλλλουχείται.*

και πρόσδεση ανταπτόρων στα 3' άκρα αυτών, κατασκευή cDNA βιβλιοθήκης, αλληλούχιση μέσω μεθοδολογιών δεύτερης γενιάς και βιοπληροφορική ανάλυση για την ταυτοποίηση των 3' άκρων.

Η αλληλούχιση 3' αμετάφραστων περιοχών σε βακτήρια, έχει ως στόχο τον ποσοτικό προσδιορισμό των 3' άκρων των βακτηριακών RNA. Μέσω αυτής της μεθόδου, βρέθηκαν σημαντικά *cis*-ρυθμιστικά στοιχεία του RNA, τα οποία μπορούν να ελέγχουν την γονιδιακή έκφραση μέσω της αναστολής της πρόωρης λήξης της μεταγραφής. Η εφαρμογή της μεθόδου έγινε σε βακτήρια – μοντέλα, όπως τα *Bacillus subtilis, Listeria monocytogenes* και *Enterococcus faecalis* και αποκαλύφθηκαν γονίδια που ρυθμίζονται από την πρόωρη λήξη της μεταγραφής (Dar *et al.*, 2016). Πρόσφατη έρευνα στο βακτήριο *Escherichia coli*, έδειξε πως η χαρτογράφηση των 3' άκρων μέσω της αλληλούχισης 3' αμετάφραστων περιοχών, αποκαλύπτει νέες πληροφορίες που συμβάλλουν στην καλύτερη κατανόηση της γονιδιακής έκφρασης σε ένα από τα πιο μελετημένα βακτήρια. Όπως αναφέρθηκε παραπάνω, είναι γνωστό πως ο τερματισμός της μεταγραφής στα βακτήρια γίνεται είτε με Rho – εξαρτώμενο, είτε με ανεξάρτητο μηχανισμό. Στον ανεξάρτητο μηχανισμό, ο τερματισμός της μεταγραφής επιτυγχάνεται μέσω δομής στελέχους - βρόχου του RNA, ακολουθούμενος από αλληλουχίες πλούσιες σε ουριδίνη, προστατεύοντας τα μετάγραφα από τις 3' - 5' εξωνουκλεάσες, ενώ για τα Rho – εξαρτώμενα μετάγραφα δεν υπήρχαν πληροφορίες για τον τρόπο



Σχήμα 1.9: *Ποσοστό των CpG islands για τις διάφορες βακτηριακές συνομοταξίες που απεικονίζονται. Το διάγραμμα έχει τροποποιηθεί από τη δημοσίευση των (Bohlin* et al.*, 2017).*

προστασίας από αυτές. Μέσω μελέτης των 3' άκρων των Rho – εξαρτώμενων μεταγράφων, αποκαλύφθηκε πως τα συγκεκριμένα μετάγραφα προστατεύονται από την δράση 3' - 5' εξωνουκλεάσων, μέσω ενεργειακά σταθερών δομών στελέχων – βρόχων στα άκρα τους, χωρίς να ακολουθούνται από αλληλουχίες πλούσιες σε ουριδίνη (Dar and Sorek, 2018).

Στους ευκαρυωτικούς οργανισμούς, η αλληλούχιση 3' αμετάφραστων περιοχών μπορεί να χρησιμοποιηθεί για τον ποσοτικό προσδιορισμό των διάφορων ισομορφών των 3' αμετάφραστων περιοχών, που εντοπίζονται σε δείγματα ανθρώπινων ιστών. Η κυριότερη εφαρμογή της, αφορά την μελέτη διαφορικής έκφρασης πρωτεϊνών σε δείγματα ανθρώπινων ιστών (Lianoglou *et al.*, 2013).

1.6. RNA Capping (Επικαλυμμα RNA)

Όπως αναφέρθηκε παραπάνω, για τη διεξαγωγή της *Cappable-seq* χρησιμοποιείται το σύμπλοκο ενζύμων VCE, το οποίο εκ φύσεως, έχει παρατηρηθεί ότι σημαίνει και μεθυλιώνει τα *m*RNA του ιού Vaccinia. Κατά έναν ανάλογο τρόπο γίνεται και η σήμανση των *m*RNA των ευκαρυωτικών οργανισμών, όπου σε αυτά η προσθήκη καλύπτρας στο 5' άκρο, τους προσδίδει πολλαπλές ιδιότητες. Πέρα λοιπόν από την απλή 5' τριφωσφορική σήμανση που έχει το πρώτο νουκλεοτίδιο κάθε άθικτου RNA, η μελέτη της καλύπτρας στους διάφορους οργανισμούς, συμπεριλαμβανομένων και των ιών, έχει οδηγήσει σε καινοτόμες μεθόδους διαχωρισμού των επεξεργασμένων RNA από τα άθικτα, βοηθώντας εν προκειμένω, τον εντοπισμό των TSS.

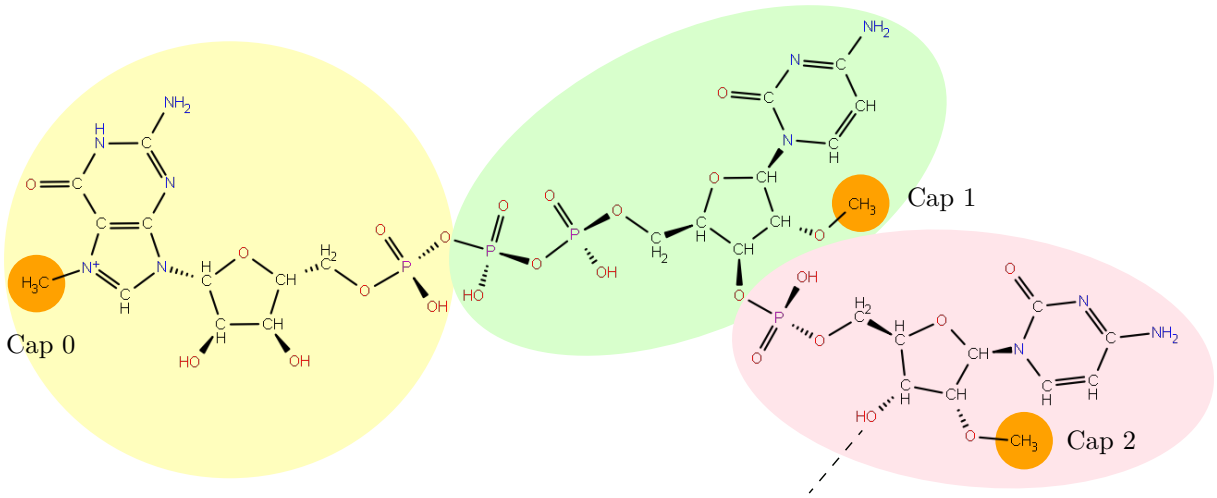
1.6.1. RNA Capping στους Ευκαρυωτικους Οργανισμους

Ένα από τα σημαντικότερα στάδια του ελέγχου της γονιδιακής έκφρασης, αποτελεί η σύνθεση ευκαρυωτικού *m*RNA μέσω της διαδικασίας της μεταγραφής. Κατά την διαδικασία αυτή, το πρώιμο *m*RNA δίνει το ώριμο *m*RNA μέσω της διαδικασία της ωρίμανσης, με τη συμβολή μικρών πυρηνικών RNA (*sn*RNA). Σύμφωνα με αυτή, απομακρύνονται τα ιντρόνια και συρράφονται τα εξώνια μεταξύ τους, για να μπορέσει στην συνέχεια το *m*RNA να

εξέλθει από τον πυρήνα, να οδηγηθεί στο κυτταρόπλασμα, να αναγνωριστεί από το ριβόσωμα και να ξεκινήσει η διαδικασία της μετάφρασης. Η προσθήκη σήμανσης ή καλύπτρας 7-μεθυλογουανοσίνης στο 5' άκρο του πρώιμου *m*RNA (RNA capping), αποτελεί ένα από τα στάδια ωρίμανσης και είναι κομβικής σημασίας, διότι συμβάλλει στην σταθερότητα του RNA. Τα υπόλοιπα στάδια της ωρίμανσης περιλαμβάνουν την προσθήκη πολύ(A) ουράς (πολυαδενυλίωση του *m*RNA) στο 3' άκρο και το εναλλακτικό μάτισμα (Ho *et al.*, 1998; Aaron Shatkin and Manley, 2000).

Ο σχηματισμός της 5' καλύπτρας, περιλαμβάνει την δημιουργία 5' - 5' δεσμού μεταξύ μιας τριφωσφορικής γουανοσίνης (GTP) και της πρώτης βάσης του μεταγράφου που είναι ένα τριφωσφορικό νουκλεοσίδιο, συνήθως πουρίνη (A.J. Shatkin, 1976). Η αντίδραση αυτή, όπως αναφέρθηκε και στην *Cappable-seq*, καταλύεται από το πυρηνικό ένζυμο RNA γουανυλυλοτρανσφεράση. Η προσθήκη 5' καλύπτρας αποτελεί υπόστρωμα για μεθυλιώσεις, δημιουργώντας τύπους καλύπτρας με βάση το πλήθος μεθυλιώσεων. Η πρώτη προσθήκη μεθυλομάδας γίνεται στη θέση N7 της ακραίας γουανίνης, καταλύεται από την 7-μέθυλοτρανσφεράση και εμφανίζεται σε όλους τους ευκαρυωτικούς οργανισμούς. Όταν μια καλύπτρα έχει μόνο αυτή την μεθυλομάδα τότε λέγεται καλύπτρα 0. Ακόμα, υπάρχει η καλύπτρα 1 που έχει μία επιπλέον μεθυλομάδα και εντοπίζεται στους πολυκύτταρους οργανισμούς και η καλύπτρα 2 που περιέχει 2 επιπλέον μεθυλομάδες και εντοπίζεται στο *m*RNA ορισμένων σπονδυλωτών *(Σχήμα 1.10)* (Langberg and Moss, 1981). Η καλύπτρα 1, η οποία εμφανίζεται στο 2′–O του +1 νουκλεοτιδίου, ευθύνεται για τη διάκριση των RNA από ξένους οργανισμούς, γεγονός που επηρεάζει την ανοσολογική απόκριση του οργανισμού, όπως για παράδειγμα όταν πρόκειται για ιικά μετάγραφα (Daffis *et al.*, 2010).

Μία από τις κύριες λειτουργίες της 5' καλύπτρας είναι να ρυθμίζει την έξοδο του *m*RNA από τον πυρήνα, μέσω της πρόσδεσης του πρωτεϊνικού συμπλόκου Cap Binding Complex (CBC) σε αυτήν. Στη συνέχεια, το σύμπλοκο CBC αναγνωρίζεται από το σύμπλοκο πυρηνικού πόρου, επάγεται η έξοδος από τον πυρήνα και η είσοδος στο κυτταρόπλασμα (Lewis and Izaurflde, 1997). Το επικάλυμμα 7-μεθυλογουανοσίνης στο 5' άκρο του *m*RNA, αποτελεί *cis*-δραστικό στοιχείο και επηρεάζει την ρύθμιση του *m*RNA, μέσω του ελέγχου της σταθερότητας και της αποικοδόμησής του, καθώς το προστατεύει από ριβονουκλεάσες (Kowtoniuk *et al.*, 2009; Aaron Shatkin and Manley, 2000). Ακόμα, η 5' καλύπτρα συμβάλει στον έλεγχο της πρωτεϊνοσύνθεσης, μέσω του ελέγχου της έναρξης της πρωτεϊνοσύνθεσης, καθώς ο παράγοντας έναρξης της μετάφρασης *eIF4e* αναγνωρίζει την 5' καλύπτρα (Marcotrigiano *et al.*, 1997). Μια ακόμα σημαντική λειτουργία της 5' καλύπτρας, αφορά την ρύθμιση της επανέναρξης της μετάφρασης μέσω της δημιουργίας κυκλοποιημένου *m*RNA, το οποίο διευκολύνει την πρόσδεση του ριβοσώματος σε αυτό (Amrani *et al.*, 2008; Wells *et al.*, 1998).



Σχήμα 1.10: *N7 μεθυλιωμένο άκρο ενός RNA ευκαρυωτικών ή ιικών οργανισμών. Με κίτρινο χρώμα φαίνεται το μεθυλιωμένο νουκλεοτίδιο γουανοσίνης, καθώς και η καλύπτρα 0 που έχει (πορτοκαλί χρώμα), ενώ με πράσινο και ροζ χρώμα, φαίνεται το πρώτο και δεύτερο νουκλεοτίδιο του RNA, με τις μεθυλιωμένες σημάνσεις ως καλύπτρα (Cap) 1 και 2, αντίστοιχα. Οι αζωτούχες βάσεις των δύο τελευταίων νουκλεοτιδίων, μπορούν να είναι οποιεσδήποτε, όπως ορίζεται για τα ριβονουκλεϊκά οξέα. Από τη διακεκομμένη γραμμή συνδέονται τα υπόλοιπα νουκλεοτίδια της αλυσίδας.*

1.6.2. RNA Capping στους Προκαρυωτικους Οργανισμους

Η μελέτη του μηχανισμού προσθήκης καλύπτρας στο 5' άκρο στους ευκαρυωτικούς οργανισμούς, έχει ξεκινήσει εδώ και πάρα πολλά χρόνια, σε αντίθεση με αυτή στους προκαρυωτικούς. Στους τελευταίους, δεδομένα έχουν προκύψει μόνο πρόσφατα, καθώς η απουσία καλύπτρας στο RNA θεωρούνταν χαρακτηριστικό της έκφρασης των προκαρυωτικών γονιδίων.

Η αρχική ένδειξη ύπαρξης τέτοιου μηχανισμού, αφορούσε τον εντοπισμό παραγώγων συνενζύμου A (CoA) ή νικοτιναμίδο-αδενίνο-δινουκλεοτίδιου (NAD) σε αρκετά 5' άκρα βακτηριακών RNA (Chen *et al.*, 2009; Kowtoniuk *et al.*, 2009). Ο πρώτος μηχανισμός ενσωμάτωσης 5' καλύπτρας σε βακτήρια εδραιώθηκε πρόσφατα, αποκαλύπτοντας πως η βακτηριακή RNA πολυμεράση προσθέτει καλύπτρα στο RNA, διότι αναγνωρίζει τα παράγωγα του συνενζύμου A και του νικοτιναμίδο-αδενίνο-δινουκλεοτίδιου, ως μη-κανονικά νουκλεοτίδια έναρξης της μεταγραφής (Bird *et al.*, 2016). Νέα δεδομένα δείχνουν πως υπάρχει ένα επιπλέον είδος καλύπτρας, στο οποίο εντοπίζεται τετραφωσφορικό νουκλεοτίδιο (Np_4) στο 5' RNA άκρο, όταν η RNA πολυμεράση χρησιμοποιεί τετραφωσφορικό δινουκλεοτίδιο (Np_4N) ως νουκλεοτίδιο έναρξης της μεταγραφής (Luciano and Belasco, 2020), δηλαδή δύο νουκλεοτίδια που συνδέονται με 5' - 5' δεσμό και το κάθε ένα φέρει μία πυροφωσφορική ομάδα. Πρόσφατες μελέτες υποδεικνύουν πως η λειτουργία της 5' καλύπτρας στους προκαρυωτικούς οργανισμούς, ενδεχομένως να σχετίζεται με τη σταθερότητα και τον ρυθμό αποικοδόμησης του *m*RNA. Επίσης, υπάρχουν ενδείξεις ότι ο ρόλος της είναι να προστατεύει το *m*RNA από την δράση των εξωνουκλεασών (Frindert *et al.*, 2018).

1.7. Leaderless *m*RNAs (*lm*RNAs)

Είναι ευρέως γνωστό, ότι τόσο στους ευκαρυωτικούς, όσο και στους προκαρυωτικούς οργανισμούς, η μετάφραση ξεκινά από περιοχές – οδηγούς, οι οποίες αποτελούν το σημείο πρόσδεσης του ριβισώματος στο 5′ άκρο των *m*RNA (Ribosome Binding Site (RBS)). Σε όλα σχεδόν τα *m*RNA των βακτηρίων και σε αρκετά των αρχαίων, περιέχεται στο 5' άκρο η γνωστή περιοχή Shine – Dalgarno (SD). Ως *SD αλληλουχία*, θεωρείται οποιοδήποτε *m*RNA που περιέχει 5' αμετάφραστη περιοχή, η οποία οριοθετείται από το σημείο έναρξης της μεταγραφής (TSS), έως και πριν το κωδικόνιο έναρξης της μετάφρασης. Επιπλέον, θα πρέπει μέσα στην περιοχή αυτή και συνήθως 8 – 10 νουκλεοτίδια ανοδικά του κωδικονίου έναρξης της μετάφρασης, να περιέχεται ένα συγκεκριμένο μοτίβο πρόσδεσης του *m*RNA στις ριβοσωμικές υπομονάδες (Shine and Dalgarno, 1974). Το μοτίβο αυτό, είναι συνήθως το `AGGAGG` στα βακτήρια, ενώ είναι πιο σπάνιο στα αρχαία. Έχει αποδειχθεί ότι αποτελεί ένα σημείο σύνδεσης, ανάμεσα στο 16S ριβοσωμικό RNA και στην 5' αμετάφραστη περιοχή ενός *m*RNA, εκκινώντας τη μετάφρασή του (Steitz and Jakes, 1975).

Μία ειδική κατηγορία *m*RNA, είναι τα λεγόμενα *Leaderless mRNA*, ή *mRNA χωρίς οδηγό*. Σε αυτά, η 5' αμετάφραστη περιοχή είτε δεν υπάρχει καθόλου και η ακολουθία ξεκινά απευθείας από το κωδικόνιο έναρξης της μετάφρασης, είτε έχει μήκος λίγων (≤ 8) νουκλεοτιδικών βάσεων. Η μετάφραση γίνεται με έναν διαφορετικό πλέον τρόπο, λόγω του ότι δεν υπάρχει περιοχή πρόσδεσης για το ριβόσωμα, ο οποίος περιλαμβάνει αρχικά, τη σύνδεση των παραγόντων εκκίνησης της μετάφρασης (συνήθως των IF2 ή IF3) και του *t*RNA – μεθειονίνη σε αυτούς και στη συνέχεια την πρόσδεση αυτού του συμπλόκου απευθείας στο κωδικόνιο έναρξης (Moll *et al.*, 2002). Διάφορες έρευνες έχουν καταδείξει, ότι η μη ύπαρξη 5' αμετάφραστης περιοχής, σχετίζεται με την κωδικοποίηση διάφορων πρωτεϊνών που έχουν τροποποιητική δράση στα RNA, όπως οι ριβονουκλεάσες και τα διάφορα τροποποιητικά ένζυμα (Romero *et al.*, 2014). Επιπλέον, έχει παρατηρηθεί ότι αυτά τα RNA κωδικοποιούν διάφορες ρυθμιστικές πρωτεΐνες των μεταθετών γονιδιακών στοιχείων στα βακτήρια, όπως οι φάγοι, τα τρανσποζόνια και τα πλασμίδια. Τέλος, το *leaderless* μεταφραστικό μοντέλο με παράγοντα τον IF2 (ή IF3), εξαρτώμενο από *t*RNA, φαίνεται να εμφανίζει συντήρηση και σε άλλα είδη πέρα των προκαρυωτικών οργανισμών, υποδηλώνοντας ενδεχομένως κάποιο κοινό πρόγονο, ο οποίος με τις εκάστοτε εξελικτικές πιέσεις, οδηγήθηκε στο έως σήμερα γνωστό, γονιδιακό επίπεδο οργάνωσης της μετάφρασης (Moll *et al.*, 2002).

2 | Μεθοδολογια

Στο δεύτερο μέρος της εργασίας, γίνεται αρχικά μία πρότυπη ανάλυση των δεδομένων αλληλούχισης στο τοπικό σύστημα του λειτουργικού Linux και στη συνέχεια παρουσιάζεται η μεθοδολογία του χαρακτηρισμού των βακτηριακών γονιδιωμάτων, με έμφαση στις περιοχές της 5′ αμετάφραστης περιοχής των γονιδίων.

2.1. Βασικες Εννοιες Αναλυσης NGS Αλληλουχισης στα Linux

Το λειτουργικό σύστημα Linux, αποτελεί εδώ και πολλά χρόνια, ένα περιβάλλον με πολλά οφέλη για την επιστημονική και όχι μόνο, κοινότητα. Σχεδόν όλα τα εργαλεία και τα λογισμικά που χρησιμοποιούνται στις βιοπληροφορικές αναλύσεις, είναι διαθέσιμα στις διάφορες διανομές των Linux, προωθώντας έτσι το λογισμικό ανοιχτού κώδικα. Ως στόχος του τελευταίου, είναι η συνεχή βελτίωση και ανάπτυξή του από το κοινό, καθώς και η κατανόηση των μηχανισμών λειτουργίας του, ώστε να αποτελέσει θεμέλιο για την ανάπτυξη νέων εργαλείων και με αυτό τον τρόπο, ο κάθε ένας ερευνητής να συμβάλλει στην βελτίωση αυτών των εργαλείων. Είναι εύλογο να ισχυριστεί κανείς, πως οι γνωστές *Αρχές των Βερμούδων*, αποτέλεσαν την πρώτη «αλλαγή υποδείγματος» για την τότε επιστημονική κοινότητα και σήμερα υπάρχουν κατά συρροή ελεύθερα διαθέσιμα, τόσο βιολογικά δεδομένα προς ανάλυση από πολυάριθμες βάσεις δεδομένων, όσο και εργαλεία και μέθοδοι για αυτές τις αναλύσεις.

Παρακάτω πραγματοποιείται μία βασική βιοπληροφορική ανάλυση σε δεδομένα αλληλούχισης. Τα εργαλεία που χρησιμοποιούνται, διατίθενται ελεύθερα, είτε ως τοπικά εγκαταστάσιμα στο σύστημα, είτε ως διαδικτυακές εφαρμογές. Τα αρχεία που χρησιμοποιήθηκαν, προέρχονται από τον άνθρωπο και συγκεκριμένα από την κυτταρική σειρά MCF-7 του μαστού, η οποία αφορά αποκλειστικά καρκινικά κύτταρα. Επιπρόσθετα, χρησιμοποιήθηκαν δύο αντίγραφα σε κάθε συνθήκη, όπου η πρώτη συνθήκη αφορά την κυτταρική σειρά ως έχει, ενώ η δεύτερη υπέστη συνθήκες υποξίας, δηλαδή έλλειψης οξυγόνου. Τα συνολικά δείγματα αλληλουχήθηκαν μέσω της Illumina HiSeq 2000, 48 ώρες μετά τη δράση της υποξίας. Τέλος, παρότι δεν προέρχονται από βακτηριακές κοινότητες, χρησιμοποιούνται οι ίδιες βασικές τεχνικές που παρουσιάζονται και στα βακτηριακά δεδομένα αλληλούχισης.

2.1.1. Ποσοτικοποιηση της Ποιοτητας Αλληλουχισης

Το πρώτο βήμα της ανάλυσης, είναι να εισάγουμε τα αρχικά δεδομένα που προέκυψαν από την αλληλούχιση, στο πρόγραμμα ανάλυσης της ποιότητας αλληλούχισης, FastQC (Andrews, 2010). Αφού εισαχθεί κάθε αρχείο, αποθηκεύεται η αναφορά που έχει προκύψει σε ένα αρχείο HTML, για μελλοντική έρευνα. Με βάση αυτές τις αναφορές, είναι εφικτή η ανάλυση των μετρικών και η εκτίμηση της ποιότητας και αξιοπιστίας της αλληλούχισης.

Η πρώτη ενότητα, με όνομα *Basic Statistics*, παρέχει πληροφορίες για τα βασικά στατιστικά στοιχεία του δείγματος. Συγκρίνοντας και τους 4 πίνακες από τα δείγματα, προκύπτει ότι το μήκος της κάθε ανάγνωσης είναι ίσο με 37 νουκλεοτίδια, ενώ οι συνολικές αναγνώσεις σε κάθε δείγμα είναι: 34.686.701 για το αντίγραφο 1 που καλλιεργήθηκε σε κανονικές οξυγόνου και 34.822.872 για το αντίγραφο 2. Αντίστοιχα, τα δείγματα τα οποία στερήθηκαν επαρκούς οξυγόνωσης (φαινόμενο της υποξίας), παρουσιάζουν παραπλήσιο πλήθος αναγνώσεων, με αριθμό 33.472.066 για το πρώτο αντίγραφο και 33.354.166 για το δεύτερο. Με αυτόν τον τρόπο, προκύπτει πως όλες οι βιβλιοθήκες που χρησιμοποιούνται στην ανάλυση, έχουν συγκρίσιμο μέγεθος, επειδή η τυπική απόκλιση του μεγέθους των reads στα φυσιολογικά και μη φυσιολογικά δείγματα, παραμένει παραπλήσια και ελάχιστη.

Αναφορικά με την ποιότητα της αλληλούχισης, μία ευρέως χρησιμοποιούμενη μετρική, είναι το λεγόμενο σκορ *Phred*, το οποίο εκφράζει σε λογαριθμική κλίμακα την ακρίβεια της αλληλούχισης και δίνεται από τον τύπο 2.1. Για παράδειγμα, για σκορ *Phred* ίσο με 38, η πιθανότητα επιλογής λάθος βάσης κατά την αλληλούχιση, είναι ίση

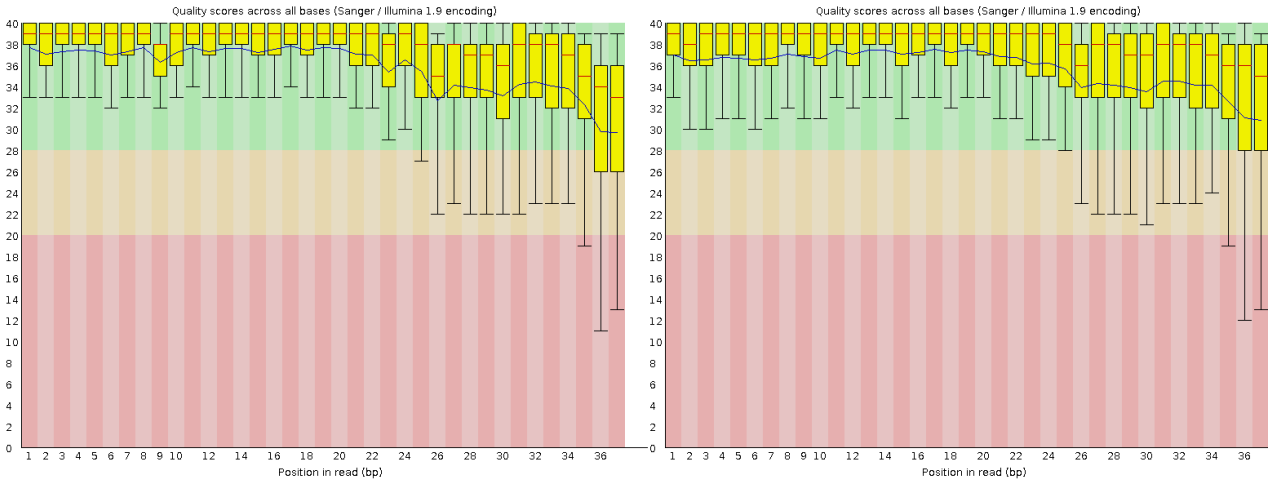
με $P = 10^{-\frac{38}{10}} = 10^{-3.8} = 0.000158$ ή 0.0158%, επομένως η ακρίβεια θα είναι της τάξης του 99.984%.

$$Q = -10 \log_{10} P \tag{2.1}$$

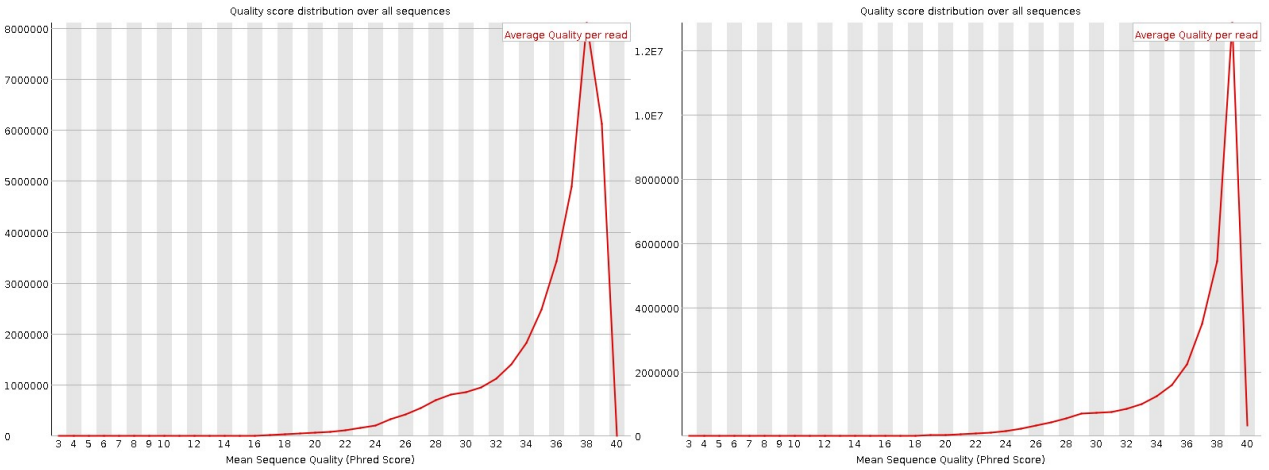
$$P = 10^{\frac{-Q}{10}} \tag{2.2}$$

Παρατηρώντας τα διαγράμματα που αναπαριστούν την ποιότητα της αλληλούχισης ανά βάση *(Σχήμα 2.1)*, τότε γίνεται αντιληπτό, πως στο τέλος των αναγνώσεων, η ποιότητα της αλληλούχισης μειώνεται. Το γεγονός αυτό, συμβαίνει επειδή η ποιότητα εξαρτάται από την ένταση και την καθαρότητα του σήματος φθορισμού. Αυτό σημαίνει, πως χαμηλή ένταση στα σήματα φθορισμού ή θορυβώδη σήματα που προέρχονται πολλές φορές από υπέρθεση, μπορούν να οδηγήσουν σε ανακρίβειες τον νουκλεοτιδικό προσδιορισμό και άρα χαμηλή ακρίβεια αλληλούχισης. Κατά τους διάφορους κύκλους λοιπόν, της αλληλούχισης, η ένταση του σήματος φθορισμού σταδιακά μειώνεται, είτε λόγω μείωσης της δραστικότητας του φθορισμοφόρου, είτε λόγω του ότι κάποιοι κλώνοι παύουν να επιμηκύνονται. Άλλες αιτίες της χαμηλής ποιότητας αλληλούχισης, μπορεί να είναι η ύπαρξη προβλημάτων με τα όργανα μέτρησης της illumina, αλλά και το πρόβλημα της υπέρ και υπό-ομαδοποίησης (underclustering/overclustering) δηλαδή της κακής εκτίμησης του νουκλεοτιδίου, λόγω μικρής απόστασης μεταξύ των cluster του εκτιμητή, που έχει ως αποτέλεσμα το μη διαχωρισμό των σημάτων.

Σύμφωνα με τα διαγράμματα ποιότητας, δε φαίνεται να υπάρχει κάποιο πρόβλημα στην ποιότητα αλληλούχισης των αναγνώσεων. Τα περισσότερα θηκογράμματα που έχουν προκύψει, βρίσκονται στην πράσινη περιοχή, δηλαδή μεταξύ της τιμής 28 και 40 της κλίμακας σκορ *Phred*. Μία παρατήρηση που μπορεί να εξαχθεί, είναι πως τα υποξιασμένα δείγματα έχουν και αυξημένη ποιότητα αλληλούχισης σε όλες τις νουκλεοτιδικές θέσεις των αναγνώσεων, σε σύγκριση με τα αντίστοιχα φυσιολογικά τους. Το γεγονός αυτό μπορεί να το ερμηνεύσει καλύτερα το διάγραμμα ποιότητας σκορ ανά αλληλουχία *(Σχήμα 2.2)*. Το συγκεκριμένο διάγραμμα, δείχνει τη συσχέτιση που υπάρχει σε κάθε τιμή του σκορ, με το αντίστοιχο πλήθος αναγνώσεων που αντιστοιχούν σε αυτό ακριβώς το σκορ. Γίνεται εύκολα αντιληπτό, πως το φυσιολογικό δείγμα 1 έχει περισσότερες αλληλουχίες με σκορ από 22 έως 36 και η μέγιστη κορυφή της καμπύλης με σκορ 39, έχει περίπου 10^7 αναγνώσεις, σε αντίθεση με το υποξιασμένο δείγμα 1, το οποίο έχει λιγότερες αναγνώσεις με σκορ από 22 έως 36 και η μέγιστη κορυφή της καμπύλης έχει επίσης σκορ 39, αλλά αυτή τη φορά περιέχει $1.1 \cdot 10^7$ αναγνώσεις. Η αντίθεση είναι πιο εμφανής στο δείγμα 2, όπου η μέγιστη κορυφή της καμπύλης στο φυσιολογικό δείγμα, έχει σκορ 38, το οποίο αντιστοιχεί σε $8 \cdot 10^6$ αναγνώσεις, ενώ στο αντίστοιχο υποξιασμένο δείγμα, η κορυφή αυτή έχει σκορ 39 και αντιστοιχεί σε $13 \cdot 10^6$ αναγνώσεις.

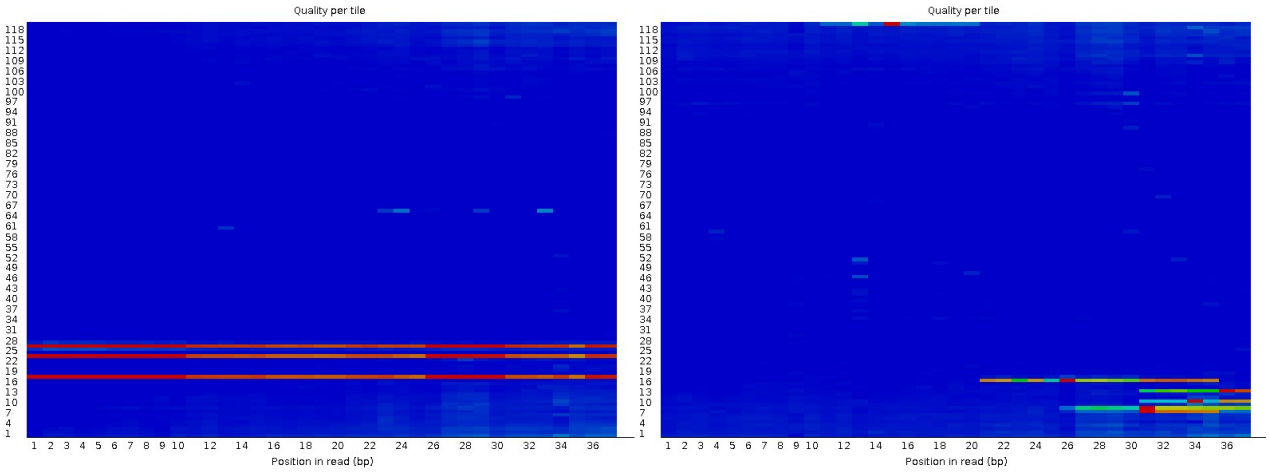


Σχήμα 2.1: *Διαγράμματα ποιότητας αλληλούχισης του αντιγράφου 2, μεταξύ του φυσιολογικού (αριστερά) και υποξιασμένου δείγματος (δεξιά).*



Σχήμα 2.2: *Διαγράμματα σκορ ανά πλήθος αλληλουχιών του αντιγράφου 2, μεταξύ του φυσιολογικού (αριστερά) και υποξιασμένου δείγματος (δεξιά).*

Ιδιαίτερη αναφορά αξίζει το διάγραμμα της ποιότητας ανά πλάκα (tile). Το διάγραμμα αυτό είναι ένας χάρτης θερμότητας (heatmap), ο οποίος περιέχει τις νουκλεοτιδικές θέσεις των αναγνώσεων στον άξονα x και τον αριθμό των πλακών στον άξονα y και εμφανίζεται μόνο όταν χρησιμοποιούνται οι βιβλιοθήκες της illumina. Αυτό που αναπαρίσταται, είναι η μέση ποιότητα αλληλούχισης σε κάθε θέση των αναγνώσεων και για κάθε tile, και έτσι μπορεί να συμπερανθεί, εάν υπήρξε μείωση της ποιότητας σε κάποιο συγκεκριμένο τμήμα του flowcell της αλληλούχισης. Ψυχρά χρώματα διαβάθμισης, αντιστοιχούν σε τιμές κοντά στη μέση ποιότητα ή μεγαλύτερες από αυτή και θερμές αποχρώσεις, σε τιμές χαμηλότερες από ότι άλλα tiles. Παρατηρώντας τα δεδομένα, βλέπουμε πως τα δύο φυσιολογικά δείγματα (αντίγραφο 1 & 2) εμφανίζουν σε ορισμένες περιοχές ανομοιογένεια στην ποιότητα ορισμένων tiles. Στο φυσιολογικό αντίγραφο 1, η ανομοιογένεια αυτή καταλαμβάνει όλη την έκταση τριών tiles, ενώ στο αντίγραφο 2 εμφανίζεται σε μερικές περιοχές του χάρτη *(Σχήμα 2.3)*. Και στις δύο περιπτώσεις, ο πιο συχνός λόγος εμφάνισης αυτών των αστοχιών, είναι η υπερφόρτωση του flowcell, συνήθως λόγω της ανακριβούς ποσοτικοποίησης της βιβλιοθήκης που χρησιμοποιείται. Η πιθανότητα όμως να είναι υπεύθυνη αυτή η αιτία, είναι μεγαλύτερη στην πρώτη περίπτωση, λόγω του ότι η ανομοιομορφία καλύπτει όλη την έκταση των αναγνώσεων, από ότι στη δεύτερη περίπτωση. Άρα, ως συμπέρασμα προκύπτει ότι θα πρέπει να γίνει ακριβέστερη ποσοτικοποίηση της βιβλιοθήκης, ώστε να μην υπάρχουν προβλήματα υπερφορτώσεων στα clusters.

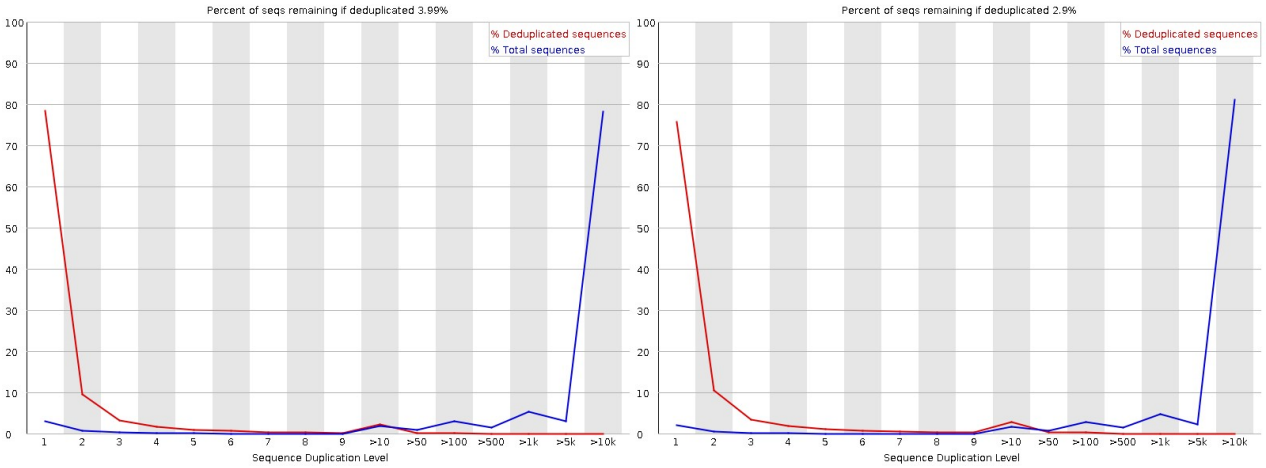


Σχήμα 2.3: *Χάρτες θερμότητας της ποιότητας ανά tile των φυσιολογικών αντιγράφων, μεταξύ του πρώτου (αριστερά) και δεύτερου αντιγράφου (δεξιά).*

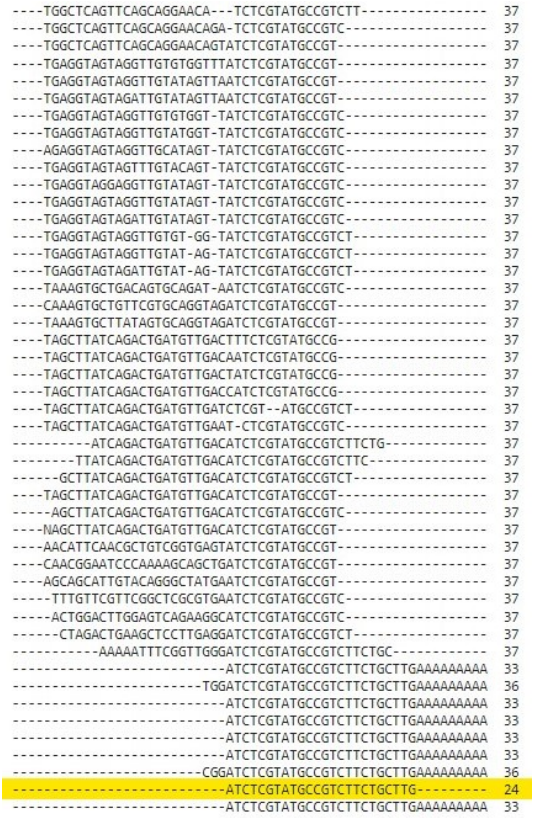
2.1.2. Διαχειριση των Διπλοτυπων Αναγνωσεων

Αναφορικά με το ζήτημα των πολλαπλών αντιγράφων όμοιων αναγνώσεων, υπάρχει μία πολύ εκτενής θεωρία πίσω από αυτές και μπορεί να έχουν προκύψει από διαφορετικές αιτίες. Μία πολύ συνηθισμένη αιτία, είναι η αντίδραση PCR, η οποία μπορεί να παράξει τα γνωστά τμήματα *PCR duplicates*, λόγω της μείωσης της πολυπλοκότητας της βιβλιοθήκης του δείγματος, που μπορεί να έχει προκύψει από πολλούς κύκλους ενίσχυσης, ή από μικρή ποσότητα αρχικού υλικού καλλιέργειας. Επίσης, η αλληλούχιση μπορεί να παράξει και αυτή διπλότυπες αναγνώσεις, όταν μία βιβλιοθήκη που χρησιμοποιείται ως πρότυπο για την παραγωγή συστάδων, δημιουργήσει πανομοιότυπες συστάδες που έχουν ως αποτέλεσμα την παραγωγή αναγνώσεων με ίδια 3′ και 5′ άκρα και ίδιες νουκλεοτιδικές συντεταγμένες, ένα φαινόμενο αρκετά συχνό τους αλληλουχητές. Μία λύση που μπορεί να προταθεί για τη βελτίωση των διπλότυπων αναγνώσεων, είναι η χρήση της τεχνικής paired-end αλληλούχισης, η οποία συνδέει τις αναγνώσεις ανά ζεύγη, μειώνοντας κατά πολύ την πιθανότητα των απρόσκοπτα τυχαίων διπλότυπων ταιριασμάτων, αλλά έτσι, αυξάνεται η πολυπλοκότητα της αλληλούχισης.

Μετά την κάλυψη των αιτιών που οδηγούν στην παραγωγή των διπλότυπων αναγνώσεων, μεταβαίνουμε στην επεξήγηση των διαγραμμάτων των επιπέδων διπλότυπων αλληλουχιών. Όπως εμφανώς γίνεται αντιληπτό, όλα αυτά τα γραφήματα εμφανίζουν μία τάση αντίθετη από την ιδανική, η οποία ορίζει ότι ο αριθμός των αναγνώσεων που είναι μοναδικές, θα πρέπει να τείνει κοντά στο 100% του δείγματος *(Σχήμα 2.4)*. Στην προκειμένη περίπτωση, η μπλε γραμμή που αντιστοιχεί στις συνολικές ανεπεξέργαστες αναγνώσεις, σημαίνει ότι περίπου το 80% του δείγματος, εμφανίζει επίπεδο διπλασιασμού μεγαλύτερο της τάξης των 10.000 αναγνώσεων. Αντιθέτως, εάν φιλτράρουμε τις διπλότυπες αναγνώσεις, βλέπουμε πλέον ότι οι μοναδικές, αντιστοιχούν περίπου στο 80% του νέου δείγματος (κόκκινη γραμμή). Όπως προκύπτει και από τους πίνακες των υπερεκπροσωπούμενων αλληλουχιών, αλλά και από τη δεδομένη ακολουθία αντάπτορα των δειγμάτων, γίνεται ακόμη περισσότερο κατανοητό το πως προέκυψαν οι τιμές των διαγραμμάτων του σχήματος 2.4. Μία χρήσιμη πρακτική, είναι ο εντοπισμός της αλληλουχίας, η οποία είναι παρούσα σε ένα μεγάλο πλήθος αναγνώσεων και να γίνει στοίχιση με όλες τις διπλότυπες αλληλουχίες του πίνακα *(Σχήμα 2.5)*. Πειραματικά, θα επιλέξουμε τις αλληλουχίες μόνο του πρώτου φυσιολογικού δείγματος, όμως παρόμοια είναι τα ευρήματα και από τις υπόλοιπες πολλαπλές στοιχίσεις. Η στοίχιση έγινε χρησιμοποιώντας το εργαλείο Clustal Omega και η επισημασμένη με κίτρινο αλληλουχία, είναι η δεδομένη αλληλουχία του αντάπτορα.



Σχήμα 2.4: *Διαγράμματα του επιπέδου διπλασιασμού των συνολικών (μπλε γραμμή) και των μοναδικών αναγνώσεων (κόκκινη γραμμή), για το φυσιολογικό αντίγραφο 1 (αριστερά) και το υποξιδωμένο αντίγραφο 1 (δεξιά).*



Σχήμα 2.5: *Πολλαπλή στοίχιση των διπλότυπων αλληλουχιών με την ακολουθία στόχο του adaptor (κίτρινη επισήμανση)*

2.1.3. Αποκοπη Ανταπτορων/Εκκινητων

Ένα από τα επόμενα βήματα της ανάλυσης, είναι ο εντοπισμός και η αποκοπή των ανταπτόρων, που η Illumina προσθέτει στα άκρα των αναγνώσεων, ώστε αυτοί να λειτουργήσουν ως εκκινητές (*primers*) και να ολοκληρωθεί επιτυχώς η αντιγραφή των τμημάτων στο στάδιο της ενίσχυσης με PCR. Η αντιγραφή αυτή, γίνεται με τη βοήθεια του ενζύμου DNA πολυμεράση, η οποία απομονώνεται από το βακτήριο *Thermus aquaticus*, που ζει σε ιαματικές πηγές και για το λόγο αυτό, η συγκεκριμένη DNA πολυμεράση, εμφανίζει αντοχή στη θερμότητα και συγκεκριμένα ονομάζεται *Taq DNA πολυμεράση*.

Εκτός από την αφαίρεση αυτής της ακολουθίας από τις αναγνώσεις, θεωρείται χρήσιμη η αφαίρεση αλληλουχιών μικρότερων από μήκος 18 νουκλεοτιδίων, αλλά και η αποκοπή βάσεων με ποιότητα αλληλούχισης μικρότερη από σκορ 10 στην κλίμακα *Phred*. Αυτές οι αναγνώσεις πιθανώς να μην στοιχισθούν σωστά στο γονιδίωμα αναφοράς, δηλαδή να περιέχουν λάθος βάσεις, ή να στοιχισθούν σε λάθος σημείο στο γονιδίωμα, αλλοιώνοντας την αξιοπιστία της στοίχισης. Για το σκοπό αυτό, χρησιμοποιείται το εργαλείο cutadapt (Martin, 2011), το οποίο παρέχει μία διεπαφή στη γραμμή εντολών, από την οποία θα γίνει η εν λόγω επεξεργασία. Για την αφαίρεση της αλληλουχίας του αντάπτορα που βρίσκεται στο 3′ άκρο, εκτελείται η παρακάτω εντολή, αλλάζοντας κάθε φορά το αρχείο εισόδου τύπου FastQ του εκάστοτε δείγματος. Το όρισμα `-a`, ψάχνει για αυτή την αλληλουχία στο 3′ άκρο, η επιλογή `min_overlap`, ορίζει τον ελάχιστο αριθμό ταιριασμάτων μεταξύ των στοιχίσεων με αυτή την αλληλουχία του αντάπτορα, ενώ το όρισμα `--cores`, χρησιμοποιείται για να ορίσει το μέγιστο αριθμό των πυρήνων του επεξεργαστή, που θα χρησιμοποιηθούν στην επεξεργασία. Η τιμή ίση με μηδέν στο όρισμα αυτό, συνεπάγεται την χρήση όλων των δυνατών πυρήνων του συστήματος.

```
cutadapt -a "ATCTCGTATGCCGTCTTCTGCTTG;min_overlap=3" --cores=8 -o output.fastq input.fastq
```

Με αυτόν τον τρόπο, η παραπάνω εντολή, θεωρεί ως παράμετρο (flag) την αλληλουχία του αντάπτορα και

ορίζει ως κατώτατο όριο ταιριάσματος τα 2 νουκλεοτιδικά ταιριάσματα, δηλαδή εάν βρεθούν κάτω από 2 ταυτίσεις μεταξύ της αλληλουχίας του αντάπτορα και μίας ανάγνωσης, τότε η ανάγνωση δεν αποκόπτεται στο 3' άκρο της. Αντιθέτως, για 3 και πάνω ταυτίσεις, η ανάγνωση μπορεί να αποκοπεί. Ακόμη, τα δύο επόμενα ορίσματα ρυθμίζουν το αρχείο εξόδου, ώστε να μην γίνει επικάλυψη με το αρχείο εισόδου.

Μετά το στάδιο αυτό όμως, μπορεί να προκύψουν δεδομένα χαμηλότερης ποιότητας και μήκους, λόγω του ότι αφαιρέθηκε ένα μεγάλο ποσοστό των βάσεων σχεδόν από όλες τις αναγνώσεις και άρα η ποιότητα της αλληλούχισης για τις θέσεις που παραμένουν, θα είναι μειωμένη[1]. Θα πρέπει λοιπόν, να γίνει αποκοπή των νουκλεοτιδικών θέσεων με πολύ χαμηλό σκορ ποιότητας και να εξαλειφθούν οι πολύ μικρές αλληλουχίες. Ο κώδικας για το πρόγραμμα cutadapt για αυτές τις δύο ενέργειες, ακολουθεί παρακάτω:

```
cutadapt --minimum-length=18 --quality-cutoff=10 -o output.fastq input.fastq
```

Συνδυάζοντας τις δύο παραπάνω εντολές σε μία για πιο γρήγορη εκτέλεση σε αρχεία αλληλούχισης, προκύπτει μία συνολική τελική εντολή, η οποία εκτελεί και τις τρεις ενέργειες, όπως φαίνεται παρακάτω:

```
cutadapt -a "ATCTCGTATGCCGTCTTCTGCTTG;min_overlap=3" --minimum-length=18
    --quality-cutoff=10 --cores=0 -o output.fastq input.fastq
```

Πλέον, μένει ο έλεγχος της αφαίρεσης των διπλότυπων αλληλουχιών και αν η νέα ποιότητα των δειγμάτων είναι ικανοποιητική. Για αυτό το λόγο, χρησιμοποιείται εκ νέου το λογισμικό FastQC, ώστε να ληφθούν οι καινούριες αναφορές και να αναλυθούν.

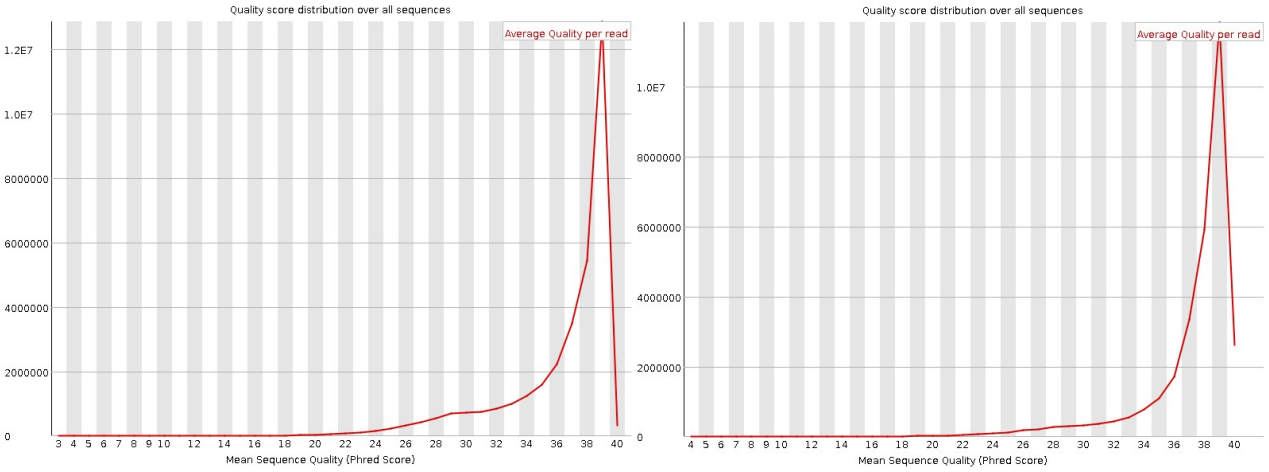
Συγκρίνοντας τις στατιστικές πληροφορίες από τα αρχικά και τελικά δείγματα, προκύπτει μία μείωση στον αριθμό των αναγνώσεων, της τάξης περίπου των 3 έως 4 εκατομμυρίων, ενώ από το μήκος των 37 νουκλεοτιδίων για κάθε ανάγνωση, έχουμε πλέον από 18 έως 37, λόγω της αποκοπής που έγινε στους αντάπτορες στο 3′ άκρο, αλλά και του φιλτραρίσματος αναγνώσεων, μικρότερων από 18 νουκλεοτιδίων. Επιπρόσθετα, υπάρχει και μία μείωση στις νησίδες CpG της τάξης του 3 - 4%, λόγω ενδεχομένως, της απουσίας των ανταπτόρων, οι οποίοι είναι πλούσιοι σε νουκλεοτίδια κυτοσίνης και γουανίνης.

Παρατηρώντας όλες τις αναφορές που προέκυψαν, γίνεται κατανοητό πως η ποιότητα των αναγνώσεων, έχει μειωθεί κυρίως προς το 3′ άκρο, λόγω του ότι πλέον, οι αναγνώσεις μήκους πάνω από 24 νουκλεοτίδια, είναι κατά πολύ λιγότερες και μικρότερης ποιότητας. Επίσης, η συνολική ποιότητα μεταξύ των αρχικών και τελικών δειγμάτων αυξήθηκε, με βάση το το διάγραμμα σκορ ανά πλήθος αναγνώσεων *(Σχήμα 2.6)*, το οποίο δείχνει, πως η καμπύλη σε πιο χαμηλά σκορ έχει μειωθεί, καθώς επίσης και ότι η μέγιστη κορυφή της για σκορ 38 ή 39 έχει αυξηθεί ακόμη και κατά 4 εκατομμύριες αναγνώσεις, παρά τη μείωση στον συνολικό αριθμό αυτών. Η γενική φιλοσοφία για την ποσοτικοποίηση της ποιότητας, είναι πως όσο η αναγνώσεις μειώνονται, λόγω αποκοπής ή και φιλτραρίσματος βάσει μήκους, η καμπύλη του σκορ ποιότητας ανά πλήθος αναγνώσεων, τείνει προς τα δεξιά, άρα πλήθος αναγνώσεων με μικρά σκορ μειώνεται και για μεγάλα σκορ, είτε παραμένει ίδιο, είτε αυξάνεται επιπλέον. Το γεγονός αυτό μπορεί να επαληθευθεί για όλα τα τελικά δείγματα, σε σύγκριση με τα αρχικά, ενώ αξίζει να αναφερθεί πως σε όλα τα τελικά δείγματα, το πλήθος αναγνώσεων με σκορ 40, εμφανίζεται υπερτετραπλασιασμένο *(Σχήμα 2.6)*, και έτσι δύνανται να εξαχθεί ως τελικό συμπέρασμα, ότι η ποιότητα της αλληλούχισης ενισχύθηκε.

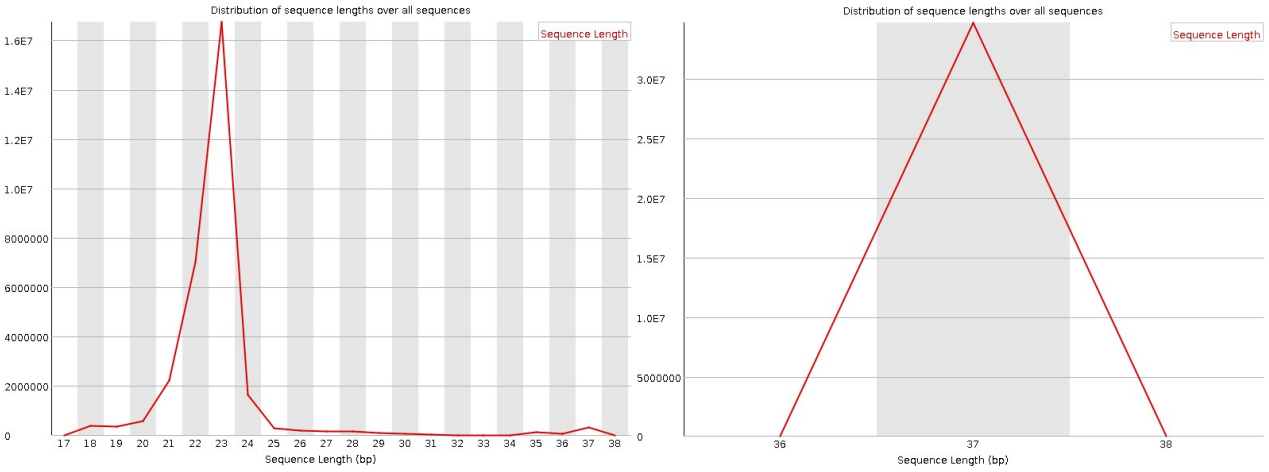
Όπως συνοπτικά αναφέρθηκε, στα τελικά δείγματα το πλήθος των αναγνώσεων δεν είναι πλέον τα 37 νουκλεοτίδια, αλλά ένα εύρος από διάφορα μήκη, στο σύνολο από 18 έως 37 βάσεις. Το γεγονός αυτό, αλλάζει πλέον την καλοσχηματισμένη τριγωνική μορφή του διαγράμματος της κατανομής μήκους των αναγνώσεων, σε μία πιο νατουραλιστική μορφή, που μας μεταφέρει το μήνυμα ότι πλέον η πλειοψηφία των αναγνώσεων είναι μήκους 23 νουκλεοτιδίων και πλήθους μεταξύ 1.6 με $1.7 \cdot 10^7$ αναγνώσεων. Το γεγονός ότι θα μειωνόταν το μήκος των αναγνώσεων, είναι σχετικά προφανές, εάν ληφθεί υπόψιν η στοίχιση του σχήματος 2.5, η οποία προερμήνευσε ότι μερικές ακολουθίες είχαν ως αντάπτορα όλο το μήκος της δοθείσας ακολουθίας, ενώ άλλες είχαν μόνο ένα μέρος

[1] Όπως αναφέρθηκε και παραπάνω, όσο κατευθυνόμαστε προς το τέλος των αλληλουχιών, η ποιότητα φθίνει.

της, με έκταση κυρίως από 14 έως 16 βάσεις (και με υψηλό αριθμό ταιριασμάτων, ώστε να γίνεται η αποκοπή) *(Σχήμα 2.7)*. Συνεπώς, τα δεδομένα αυτά, επιβεβαιώνουν ότι η διαδικασία επεξεργασίας των δεδομένων που πραγματοποιήθηκε, μπορεί να τεκμηριωθεί και ίσως ήταν προς ορθόλογη κατεύθυνση.



Σχήμα 2.6: *Διαγράμματα ποιότητας σκορ ανά πλήθος αλληλουχιών για το αρχικό υποξιασμένο δείγμα του αντιγράφου 2 (αριστερά) και το ίδιο τελικό δείγμα μετά την επεξεργασία (δεξιά)*

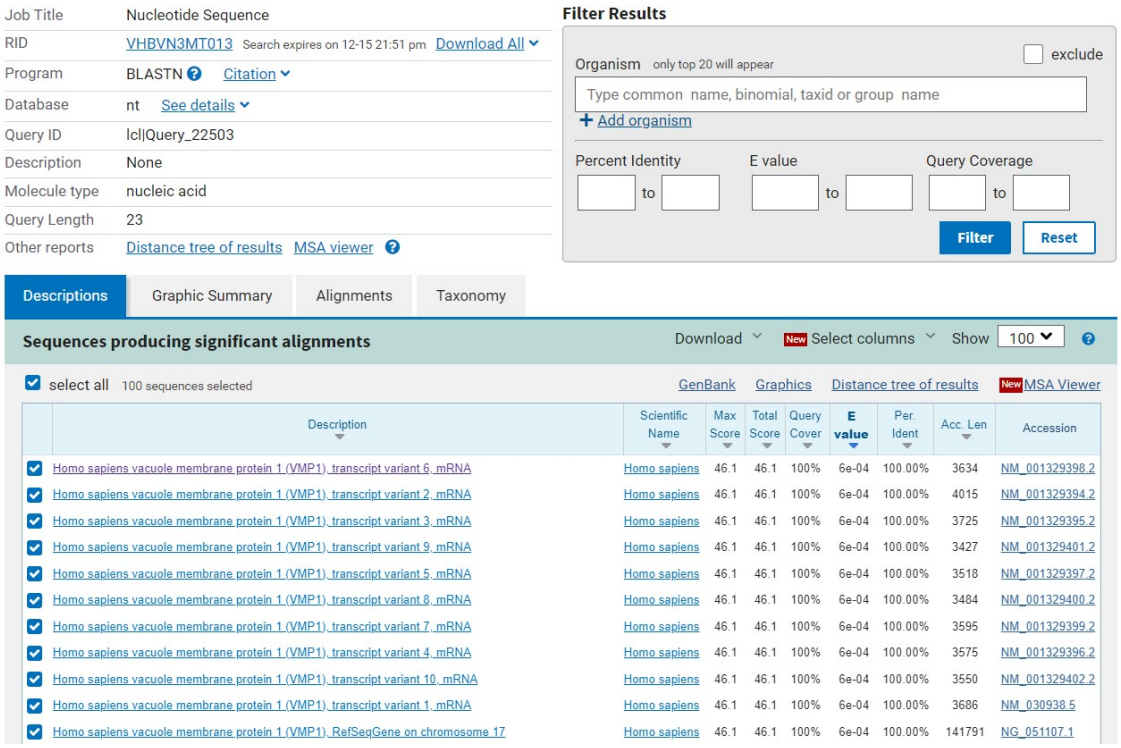


Σχήμα 2.7: *Διαγράμματα κατανομής μήκους των αναγνώσεων, στο φυσιολογικό αντίγραφο 1 του τελικού δείγματος (αριστερά) και του αρχικού δείγματος (δεξιά)*

Κλείνοντας την ανάλυση, θα πρέπει να επισημανθούν και τα δεδομένα που υπάρχουν από τις νέες διπλασιασμένες αλληλουχίες. Αρχικά, συγκρίνοντας τους νέους πίνακες με τις υπερεπροσωπούμενες αλληλουχίες, γίνονται αντιληπτά δύο πράγματα: το ένα είναι ότι ακολουθείται το ίδιο μοτίβο, δηλαδή ότι η πρώτη αλληλουχία εμφανίζει μεγάλο ποσοστό εμφάνισης και οι επόμενες πολύ μικρότερο, όπως αντίστοιχα στην αρχική ανάλυση, ενώ πλέον σε κανέναν τελικό πίνακα δεν αναφέρεται πλέον στο πεδίο της πιθανής πηγής, ο όρος *adaptor*, δείχνοντας έτσι, ότι από την αποκοπή δεν παρέμεινε κανένας άλλος αντάπτορας. Όμως, επειδή τώρα στο δείγμα υπάρχουν πολύ μικρότερες αλληλουχίες, από στατιστικής άποψης αυτές είναι περισσότερο πιθανές να εντοπιστούν πολλές φορές (ακόμα και τυχαία), και άρα τα ποσοστά εμφάνισης είναι λίγο αυξημένα σε σχέση με τα αρχικά δεδομένα. Εάν, βέβαια συγκριθούν τα διαγράμματα των επιπέδων διπλότυπων αλληλουχιών, παρατηρείται πως δεν έχουν κάποια ουσιαστική διαφορά, με τη μόνη επισήμανση ότι τα επίπεδα διπλότυπων μεταξύ 10 και 1000, είναι λίγο μειωμένα σε σχέση με τα αρχικά διαγράμματα, γεγονός που οφείλεται στην εν μέρει διαφοροποίησή τους μετά την αποκοπή του αντάπτορα. Επιστρέφοντας στους πίνακες των διπλότυπων αναγνώσεων, η πρώτη αλληλουχία που έχει και

το μεγαλύτερο ποσοστό εμφάνισης (41 έως 43%), εμφανίζεται κοινή και στους υπόλοιπους πίνακες των επεξεργασμένων δειγμάτων. Το γεγονός αυτό δεν μπορεί να οφείλεται σε τυχαία φαινόμενα, επομένως για αυτή την αλληλουχία, θα αναζητήσουμε τις ομόλογές της με τη βοήθεια του BLASTn (Zhang *et al.*, 2000), ώστε να εξακριβωθεί η λειτουργία που μπορεί να έχει το γονίδιο που τη φέρει.

Στο BLASTn λοιπόν, εκτελείται αναζήτηση στη βάση *Nucleotide collection (nt)* και ως οργανισμός, επιλέγεται ο κοινός Homo sapiens (taxid: 9606), επειδή τα δεδομένα προέρχονται από την ανθρώπινη κυτταρική σειρά, με κωδικό MCF7. Παρατηρείται από τα αποτελέσματα *(Σχήμα 2.8)*, ότι οι εγγραφές με το μικρότερο *e-value*, αντιστοιχούν στο γονίδιο, το οποίο κωδικοποιεί μία διαμεμβρανική πρωτεΐνη κενοτόπιου και τους πολυμορφισμούς αυτού. Σύμφωνα με έρευνες, αυτή η διαμεμβρανική πρωτεΐνη διαδραματίζει βασικό ρυθμιστικό ρόλο στη διαδικασία της αυτοφαγίας, μία ιδιότητα των κυττάρων να ανακυκλώνουν κατεστραμμένα κύτταρα, ώστε να ανανεώνονται και να διατηρούνται λειτουργικά. Με δεδομένο ότι η κυτταρική σειρά MCF-7 είναι αποκλειστικά καρκινική, εύλογα μπορεί να γίνει ο ισχυρισμός, ότι το εύρημα της υπερέκφρασης αυτού του γονιδίου, είναι και αυτό που ίσως ευθύνεται περισσότερο στην ανάπτυξη του καρκίνου. Έτσι, όταν η αυτοφαγία δεν πραγματοποιείται όπως φυσιολογικά θα έπρεπε, υπάρχει κίνδυνος εμφάνισης ασθενειών όπως ο διαβήτης, ο καρκίνος και άλλες.



Σχήμα 2.8: *Αποτελέσματα του BLASTn για τις πιο πιθανές αλληλουχίες μεταγράφων, σύμφωνα με την ομολογία της υπερεκπροσωπούμενης αλληλουχίας με αλληλουχίες της βάσης δεδομένων Nucleotide του NCBI.*

2.1.4. Στοιχιση Δεδομενων στο Γονιδιωμα Αναφορας

Σε αυτή την ενότητα, γίνεται στοίχιση των αρχείων της αλληλούχισης με το ανθρώπινο γονιδίωμα, μέσα από μία σειρά υπολογιστικών εργαλείων, προκειμένου να επιτευχθεί η στοίχιση, αλλά και η οπτικοποίηση γονιδιακών περιοχών και των αντίστοιχων αναγνώσεων που υπάρχουν. Για το σκοπό αυτό, αρχικά θα χρησιμοποιηθεί το εργαλείο στοίχισης *Bowtie2* (Langmead and Salzberg, 2012), το οποίο με δεδομένο ένα γονιδίωμα αναφοράς, στοιχίζει σε αυτό όλες τις αναγνώσεις. Το αποτέλεσμα σχεδόν όλων των εργαλείων στοίχισης, είναι ένα αρχείο τύπου SAM (Sequence Alignment/Map) ή BAM (Binary Alignment/Map), όπου το τελευταίο αποτελεί τη δυαδική κωδικοποίηση του πρώτου. Αφού ολοκληρωθεί η στοίχιση, η περαιτέρω ανάλυση επαφίεται πλέον σε

πιο απλές διεργασίες text manipulation, ώστε να γίνει η μετατροπή του αρχείου SAM σε BAM για τη μείωση του όγκου των δεδομένων, η ταξινόμηση των αναγνώσεων ανά χρωμόσωμα για την καλύτερη διαχείριση αυτών των αρχείων, καθώς και η ευρετηρίαση του ταξινομημένου αρχείου BAM, σε βοηθητικό αρχείο τύπου BAI (Binary Alignment Indexes) για την οπτικοποίηση των αποτελεσμάτων. Όλες αυτές οι ενέργειες, μπορούν να πραγματοποιηθούν με το εργαλείο *Samtools* (H. Li *et al.*, 2009), το οποίο παρέχει μία διεπαφή για την εκτύπωση και το χειρισμό του περιεχομένου αυτών των αρχείων. Τέλος, θα χρησιμοποιηθεί το εργαλείο *Integrative Genomics Viewer (iGV)* (Thorvaldsdóttir *et al.*, 2012), που αποτελεί μία τοπικά εγκαταστάσιμη εφαρμογή περιηγητή γονιδιωμάτων, σε αντίθεση με άλλες διαδικτυακές εφαρμογές, όπως η *UCSC Genome Browser* και η *ensembl Genome Browser*.

Για τη στοίχιση των αναγνώσεων, θα πρέπει να ληφθεί το ανθρώπινο γονιδίωμα ως ακολουθία, το οποίο βρίσκεται συνήθως σε μορφή fasta και με τη βοήθεια της ρουτίνας *build* του εργαλείου *bowtie2*, να γίνει ευρετηρίαση αυτού. Για το σκοπό αυτό, γίνεται λήψη του γονιδιώματος αναφοράς από το σύνδεσμο https://www.ncbi.nlm.nih.gov/genome/guide/human/ και εκτελώντας την παρακάτω εντολή, παράγονται τα αρχεία της ευρετηρίασης.

```
bowtie2-build -f GRCh38_latest_genomic.fna GRCh38_latest_genomic
```

Η παράμετρος `-f`, δείχνει ότι το αρχείο εισόδου είναι σε μορφή fasta. Τα αρχεία ευρετηρίασης του γονιδιώματος που παράγονται, αποτελούνται από πολλά αρχεία τύπου `.bt2`, και χρησιμοποιούνται ως γονιδίωμα αναφοράς. Η εντολή για τη δημιουργία του αρχείου στοίχισης SAM είναι η εξής:

```
bowtie2 -x RefGenome/GRCh38_noalt_as -U input.fastq.gz -S output.sam -p 8,
```

όπου η παράμετρος `-x` ορίζει το γονιδίωμα αναφοράς, εισάγοντας το κοινό όνομα (basename) όλων των αρχείων `.bt2`, ενώ το όρισμα `-U` ορίζει μία λίστα αρχείων Fastq, τα οποία περιέχουν αναγνώσεις χωρίς ζεύγη (Single-end reads). Τέλος, το όρισμα `-S` εξάγει τη στοίχιση σε ένα αρχείο SAM, παρακάμπτοντας την επιλογή για εκτύπωση της στοίχισης στην κονσόλα, ενώ με την παράμετρο `-p` ορίζεται ο υπολογισμός να καταλαμβάνει ορισμένο αριθμό από threads, για παραλληλοποίηση και πιο γρήγορο υπολογισμό.

Στη συνέχεια, γίνεται μετατροπή του αρχείου SAM σε BAM, ώστε να μειωθεί ο όγκος των δεδομένων, αλλά και η ταξινόμηση των εγγραφών ανά χρωμόσωμα. Θα χρησιμοποιηθεί και εδώ η σουίτα εργαλείων *Samtools*, ώστε να γίνει η διαχείριση αυτών των αρχείων υψηλής διεκπεραιωτικής αλληλούχισης. Ο κώδικας για τη μετατροπή, φαίνεται παρακάτω:

```
samtools view -b -o output.bam input.sam -@ 8,
```

όπου οι παράμετροι `-b -o` εξάγουν το αρχείο εισόδου σε μορφή BAM και ο χαρακτήρας `@`, ορίζει το μέγιστο αριθμό πυρήνων.

Με την παρακάτω εντολή, μπορεί να εκτυπωθεί το περιεχόμενο του δυαδικού αρχείου στην κονσόλα, ενώ με την παράμετρο `-H` πριν το όνομα του αρχείου, εκτυπώνεται μόνο η κεφαλίδα του.

```
samtools view output.bam -@ 8
```

Συνεχίζοντας την ανάλυση, για την καλύτερη διαχείριση αυτού του αρχείου, θα πρέπει να ταξινομήσουμε τα περιεχόμενά του, με αύξουσα σειρά των χρωμοσωμάτων και των χρωμοσωμικών του συντεταγμένων, στις οποίες στοιχίζονται οι αναγνώσεις στο γονιδίωμα αναφοράς. Η ταξινόμηση, γίνεται με την παρακάτω εντολή, που δημιουργεί ένα νέο ταξινομημένο αρχείο ως έξοδο:

```
samtools sort input.bam -o output.sort.bam -@ 8
```

Ακολούθως, η δημιουργία του αρχείου BAI γίνεται με την εντολή `index`, η οποία δέχεται το ταξινομημένο αρχείο και παράγει τις τελικές συντεταγμένες αυτού, ώστε να επιτρέπει στα προγράμματα οπτικοποίησης, να μεταβαίνουν σε ένα συγκεκριμένο σημείο της στοίχισης, χωρίς να διαβάζουν το συνολικό αρχείο. Τα δύο τελευταία αρχεία (`xx.sort.bam & xx.sort.bam.bai`), συνήθως χρησιμοποιούνται ως αλληλένδετα αρχεία στα προγράμματα οπτικοποίησης της στοίχισης, δηλαδή, όταν φορτώνουμε το αρχείο `.sort.bam`, τότε είναι προϋπόθεση να έχουμε παράξει και το αρχείο `.sort.bam.bai`, το οποίο εντοπίζεται αυτόματα, με την περαιτέρω προϋπόθεση να έχουν το ίδιο βασικό όνομα (*basename*).

```
samtools index input.sort.bam
```

Το σχήμα 2.9, αναπαριστά τις πρώτες γραμμές από το αρχικό SAM αρχείο, το συμπιεσμένο BAM αρχείο, καθώς και το αντίστοιχο ταξινομημένο.

```
kdan@apollo:/mnt/raid0/kdan$ samtools view Processed_Hypoxia_Rep1.sam | head -n 5
SRR873388.1     0       chr17   59841273        42      23M     *       0       0       NAGCTTATCAGACTGATGTTGAC #/2..55855000000000000 AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0T22       YT:Z:UU
SRR873388.3     0       chr17   59841273        42      23M     *       0       0       NAGCTTATCAGACTGATGTTGAC #333387778@@@C@@@C@@@C AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0T22       YT:Z:UU
SRR873388.4     0       chr17   59841273        42      23M     *       0       0       NAGCTTATCAGACTGATGTTGAC #333386666C@CC@CC@@C AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0T22       YT:Z:UU
SRR873388.5     0       chr17   59841273        42      23M     *       0       0       NAGCTTATCAGACTGATGTTGAC #--/-535530000000000000 AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0T22       YT:Z:UU
SRR873388.6     0       chr17   59841273        42      23M     *       0       0       NAGCTTATCAGACTGATGTTGAC #1121543335@@@CC@@CC@@@C AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0T22       YT:Z:UU
kdan@apollo:/mnt/raid0/kdan$ samtools view Processed_Hypoxia_Rep1.bam | head -n 5
SRR873388.1     0       chr17   59841273        42      23M     *       0       0       NAGCTTATCAGACTGATGTTGAC #/2..55855000000000000 AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0T22       YT:Z:UU
SRR873388.3     0       chr17   59841273        42      23M     *       0       0       NAGCTTATCAGACTGATGTTGAC #333387778@@@C@@@C@@@C AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0T22       YT:Z:UU
SRR873388.4     0       chr17   59841273        42      23M     *       0       0       NAGCTTATCAGACTGATGTTGAC #333386666C@CC@CC@@C AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0T22       YT:Z:UU
SRR873388.5     0       chr17   59841273        42      23M     *       0       0       NAGCTTATCAGACTGATGTTGAC #--/-535530000000000000 AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0T22       YT:Z:UU
SRR873388.6     0       chr17   59841273        42      23M     *       0       0       NAGCTTATCAGACTGATGTTGAC #1121543335@@@CC@@CC@@@C AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0T22       YT:Z:UU
kdan@apollo:/mnt/raid0/kdan$ samtools view Processed_Hypoxia_Rep1.sorted.bam | head -n 5
SRR873388.26803920      16      chr1    17408   0       24M     *       0       0       ATGTCCTGAGCCCATGTTCCTCTC GGGFGGBGGHBDHHGGGGGGEGGG       AS:i:-5 XS:i:-5 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0C23       YT:Z:UU
SRR873388.21418947      16      chr1    17409   1       23M     *       0       0       TGTCCTGAGCCCATGTTCCTCTC BIIHHIHIIIIGIIDBGAGGGGEG AS:i:0  XS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:23 YT:Z:UU
SRR873388.855607        16      chr1    17410   1       22M     *       0       0       GTCCTGAGCCCATGTTCCTCTC HHGHHHGHHHHHHHHHGHHHHH   AS:i:0  XS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:22 YT:Z:UU
SRR873388.12978055      16      chr1    17410   1       21M     *       0       0       GTCCTGAGCCCATGTTCCTCT    G<GG@GHGHDHEGGEBGGDE@   AS:i:0  XS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:21 YT:Z:UU
SRR873388.12709944      16      chr1    18801   1       19M     *       0       0       CCTGAGTGTCGTCACCTAA  GGGGGGGGG=C=B:@;;-A        AS:i:0  XS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:19 YT:Z:UU
kdan@apollo:/mnt/raid0/kdan$ samtools view -H Processed_Hypoxia_Rep1.bam
@HD     VN:1.0  SO:unsorted
@SQ     SN:chr1 LN:248956422
@SQ     SN:chr2 LN:242193529
@SQ     SN:chr3 LN:198295559
@SQ     SN:chr4 LN:190214555
@SQ     SN:chr5 LN:181538259
@SQ     SN:chr6 LN:170805979
@SQ     SN:chr7 LN:159345973
@SQ     SN:chr8 LN:145138636
@SQ     SN:chr9 LN:138394717
@SQ     SN:chr10        LN:133797422
@SQ     SN:chr11        LN:135086622
@SQ     SN:chr12        LN:133275309
@SQ     SN:chr13        LN:114364328
@SQ     SN:chr14        LN:107043718
@SQ     SN:chr15        LN:101991189
@SQ     SN:chr16        LN:90338345
@SQ     SN:chr17        LN:83257441
@SQ     SN:chr18        LN:80373285
@SQ     SN:chr19        LN:58617616
@SQ     SN:chr20        LN:64444167
@SQ     SN:chr21        LN:46709983
@SQ     SN:chr22        LN:50818468
@SQ     SN:chrX LN:156040895
@SQ     SN:chrY LN:57227415
@SQ     SN:chrM LN:16569
```

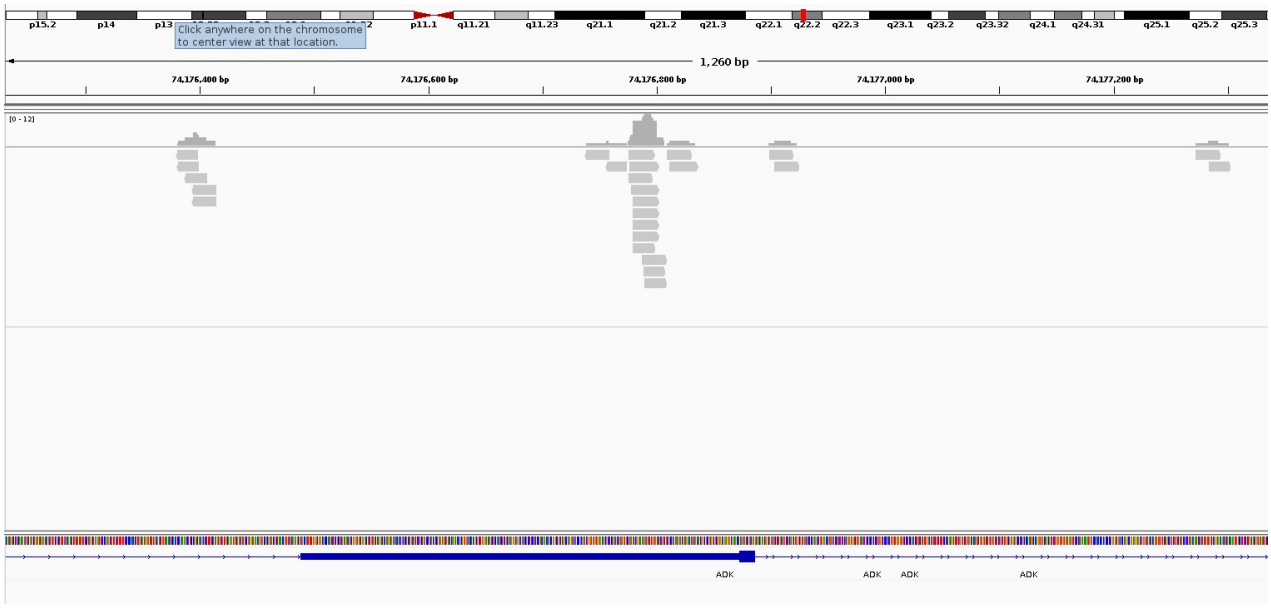
Σχήμα 2.9: *Πρώτες 5 εγγραφές των αρχείων που παρήχθησαν. Πρώτα, εμφανίζονται τα περιεχόμενα του ασυμπίεστου αρχείου SAM, δεύτερα τα περιεχόμενα του συμπιεσμένου αρχείου BAM και τρίτα τα περιεχόμενα του ταξινομημένου αρχείου BAM, τα οποία ξεκινούν, όπως είναι αναμενόμενο, από το χρωμόσωμα 1. Στο τέλος, εμφανίζονται τα δεδομένα της κεφαλίδας (header), του αρχείου BAM.*

Το τελευταίο βήμα πριν την οπτικοποίηση, είναι η εξαγωγή ορισμένων στατιστικών αναφορών από τις στοιχισμένες αναγνώσεις, ώστε να ποσοτικοποιηθεί η ποιότητα της στοίχισης. Για το σκοπό αυτό, εκτελείται η εντολή:

```
samtools flagstat input.sort.bam,
```

η οποία εμφανίζει ότι από τις συνολικά 29.440.658 αναγνώσεις, έχουν στοιχηθεί επιτυχώς, οι 28.112.553 με μηδενικό ποσοστό σφαλμάτων, με βάση την ποιότητα αλληλούχισης (QC failure rate).

Το τελικό στάδιο της ανάλυσης, είναι η οπτικοποίηση της στοίχισης στο πρόγραμμα *iGV*, η οποία προϋποθέτει την ύπαρξη του ταξινομημένου αρχείου BAM και του αρχείου BAI. Αρχικά, με την εκκίνηση του προγράμματος, φορτώνεται το γονιδίωμα αναφοράς του ανθρώπου (Human [GRCh38/hg38]) και φορτώνονται οι αναγνώσεις της αλληλούχισης από το ταξινομημένο αρχείο BAM. Επειδή το μήκος των αναγνώσεων (περίπου 23 βάσεις) είναι απειροελάχιστο σε σχέση με το συνολικό μήκος ενός χρωμοσώματος, το οποίο έχει μέσο μήκος περίπου 130 Μεγαβάσεις (Mb), θα πρέπει να γίνει μεγέθυνση έως το επίπεδο που εμφανίζονται στην οθόνη το πολύ 2.500 βάσεις, ώστε να είναι ορατές οι αναγνώσεις *(Σχήμα 2.10)*.



Σχήμα 2.10: *Στοιχισμένες αναγνώσεις που εμφανίζονται στο χρωμόσωμα 10, στην περιοχή μεταξύ των θέσεων 74.176.700 έως 74.177.000.*

2.1.5. Peak Calling και Εξαγωγη Συμπερασματων

Πριν γίνει αναφορά στη μέθοδο Peak Calling, είναι αναγκαίο να αναφερθεί ένας ακόμη τύπος αρχείου, ο οποίος χρησιμοποιείται σε αυτήν. Τα αρχεία Browser Extensible Data (BED) είναι μια μορφή αρχείου κειμένου, που χρησιμοποιείται για την αποθήκευση γονιδιωματικών περιοχών ως συντεταγμένων. Τα δεδομένα παρουσιάζονται με τη μορφή στηλών, διαχωρισμένα με το χαρακτήρα διαστήματος ή στηλοθέτη και αντιπροσωπεύουν συντεταγμένες, αντί για ακολουθίες νουκλεοτιδίων. Το αρχείο αυτό δεν είναι δυαδικό, αλλά ένα αρχείο κειμένου, το οποίο διαφέρει σε σχέση με τα αρχεία BAI, ως προς την κωδικοποίηση και ως προς το περιεχόμενο, καθώς τα αρχεία BAI αποτελούν έναν πίνακα περιεχομένων των αρχείων BAM. Επιπλέον, το αρχείο αυτό περιέχει τόσες εγγραφές, όσες και το αντίστοιχο πλήθος των επιτυχώς στοιχισμένων αναγνώσεων.

Για να επιτευχθεί αυτή τη μετατροπή, μέσα από το εργαλείο *Bedtools* (Quinlan and Hall, 2010) και συγκεκριμένα με τη ρουτίνα *bamtobed* εκτελείται η παρακάτω εντολή και προκύπτει το αρχείο τύπου BED *(Σχήμα 2.11)*.

```
bedtools bamtobed -i input.sort.bam > output.bed
```

```
kdan@apollo:/mnt/raid0/kdan$ cat Processed_Hypoxia_Rep1.sorted.bam.bed | head -n 10
chr1    17407   17431   SRR873388.26803920      0       -
chr1    17408   17431   SRR873388.21418947      1       -
chr1    17409   17431   SRR873388.855607        1       -
chr1    17409   17430   SRR873388.12978055      1       -
chr1    18800   18819   SRR873388.12709944      1       -
chr1    20369   20387   SRR873388.20653811      1       -
chr1    29277   29295   SRR873388.19599080      1       -
chr1    29281   29303   SRR873388.21790238      1       -
chr1    29283   29311   SRR873388.20582746      1       -
chr1    29307   29331   SRR873388.17494577      1       -
```

Σχήμα 2.11: *Πρώτες 5 εγγραφές του αρχείου BED, χρησιμοποιώντας το εργαλείο Bedtools*

Η μέθοδος Peak Calling είναι μια υπολογιστική μέθοδος, η οποία χρησιμοποιείται για τον προσδιορισμό περιοχών σε ένα γονιδίωμα, όπου υπάρχει υψηλός εμπλουτισμός με στοιχισμένες αλληλουχούμενες αναγνώσεις. Για το σκοπό αυτό, χρησιμοποιείται η ρουτίνα *genomecov* του εργαλείου *Bedtools*, η οποία με δεδομένο ένα αρχείο των γονιδιωματικών χρωμοσωμικών μηκών και ενός αρχείου αναγνώσεων μορφής BAM ή BED, υπολογίζει την

1	Χρωμόσωμα	Αριθμός Χρωμοσώματος
2	Θέση Έναρξης	Έναρξη συντεταγμένων στο χρωμόσωμα για την εξεταζόμενη ανάγνωση (η πρώτη βάση στο χρωμόσωμα αριθμείται ως 0)
3	Θέση Λήξης	Τελική συντεταγμένη στο χρωμόσωμα για την εξεταζόμενη ακολουθία
4	Όνομα	Ονομασία/κωδικός ανάγνωσης
5	Σκορ	Βαθμολογία μεταξύ 0 και 1000
6	Προσανατολισμός	Προσανατολισμός ανάγνωσης DNA (θετικός [+] ή αρνητικός [-] ή [.] εάν δεν υπάρχει προσανατολισμός)

Πίνακας 2.1: *Πίνακας επεξήγησης βασικών στηλών των εγγραφών στο αρχείο BED*

επικάλυψη των αναγνώσεων σε κάθε θέση. Ο κώδικας 1, περιγράφει ένα παράδειγμα χρήσης αυτής της εντολής, ενώ παρακάτω φαίνεται η εντολή που έγινε εκτέλεση, για την παραγωγή του αρχείου επικάλυψης των αναγνώσεων.

```
$ cat A.bed
chr1  10  20
chr1  20  30
chr2  0   500

$ cat my.genome
chr1  1000
chr2  500

$ bedtools genomecov -i A.bed -g my.genome
chr1   0  980  1000  0.98
chr1   1  20   1000  0.02
chr2   1  500  500   1
genome 0  980  1500  0.653333
genome 1  520  1500  0.346667
```

Κώδικας 1: *Παράδειγμα χρήσης της ρουτίνας genomecov για τον υπολογισμό της επικάλυψης αλληλουχιών*

```
samtools genomecov -i input.bed -g h19.genome > output.new_file.bed
```

Το αρχείο των χρωμοσωμικών μηκών, λήφθηκε από τον σύνδεσμο https://github.com/arq5x/bedtools/blob/master/genomes/human.hg19.genome και τα αποτελέσματα αυτής της εκτέλεσης, φαίνονται στο σχήμα 2.12. Η πρώτη στήλη αυτού του σχήματος αντιπροσωπεύει τον αριθμό του χρωμοσώματος, στο οποίο αναφέρεται η εγγραφή, η δεύτερη στήλη τον αριθμό του βάθους κάλυψης και η τρίτη στήλη τον αριθμό των βάσεων που έχουν βάθος κάλυψης ίσο με αυτό της δεύτερης στήλης για το αντίστοιχο χρωμόσωμα της πρώτης στήλης. Επιπλέον, η τέταρτη στήλη αντιστοιχεί στο συνολικό μήκος του χρωμοσώματος, ενώ η πέμπτη στήλη είναι ένα σκορ, το οποίο προκύπτει από τον αριθμό των βάσεων με σκορ ίσο με αυτό της δεύτερης στήλης, προς το συνολικό μήκος του χρωμοσώματος. Τέλος, στην πρώτη εγγραφή, αντιστοιχούν οι περιοχές του χρωμοσώματος 1, στις οποίες δεν υπάρχει καμία ανάγνωση και φαίνεται ότι αυτές κατέχουν το 98% της συνολικής έκτασης αυτού του χρωμοσώματος.

2.2. Αναλυση NGS Αλληλουχισης με τη Γλωσσα ℝ

Τα εργαλεία που αναφέρθηκαν στην παραπάνω ενότητα, καθώς και αρκετά άλλα εναλλακτικά εργαλεία, τα οποία είναι ελεύθερα διαθέσιμα, μπορούν να πραγματοποιήσουν ένα μεγάλο εύρος αναλύσεων, είτε μέσω της γραμμής εντολών, είτε μέσω διαδικτυακών και γραφικών διεπαφών. Ωστόσο, δεν ενδείκνυται η χρήση τους όταν η ανάλυση

```
kdan@apollo:/mnt/raid0/kdan$ cat file.bed | head -n 20
chr1    0       248804169       249250621       0.998209
chr1    1       339189  249250621       0.00136084
chr1    2       50578   249250621       0.00020292
chr1    3       16135   249250621       6.4734e-05
chr1    4       7154    249250621       2.8702e-05
chr1    5       4963    249250621       1.99117e-05
chr1    6       3397    249250621       1.36289e-05
chr1    7       2274    249250621       9.12335e-06
chr1    8       1752    249250621       7.02907e-06
chr1    9       1325    249250621       5.31593e-06
chr1    10      1236    249250621       4.95886e-06
chr1    11      946     249250621       3.79538e-06
chr1    12      859     249250621       3.44633e-06
chr1    13      700     249250621       2.80842e-06
chr1    14      641     249250621       2.57171e-06
chr1    15      588     249250621       2.35907e-06
chr1    16      560     249250621       2.24673e-06
chr1    17      380     249250621       1.52457e-06
chr1    18      349     249250621       1.4002e-06
chr1    19      345     249250621       1.38415e-06
kdan@apollo:/mnt/raid0/kdan$
```

Σχήμα 2.12: *Πρώτες 20 εγγραφές του νέου αρχείου κάλυψης BED, χρησιμοποιώντας τη ρουτίνα genomecov του εργαλείου Bedtools*

που πρέπει να διεξαχθεί, αποτελείται από πολλά σύνθετα επιμέρους βήματα, τα οποία συνδυάζουν πολλά και διαφορετικά εργαλεία. Για αυτό το λόγο, με τη σημαντική πρόοδο των γλωσσών προγραμματισμού έως σήμερα, δίνεται η δυνατότητα εκτέλεσης αυτών των αναλύσεων, μέσα από ένα πιο ευέλικτο και ενιαίο περιβάλλον. Μερικές από τις πλέον γνωστές γλώσσες προγραμματισμού, είναι η Python, η Perl, η Julia, καθώς και η ℝ, η οποία χρησιμοποιείται κατά κόρον στην παρούσα εργασία. Πιο αναλυτικά, η ℝ (R Core Team, 2021) σχεδιάστηκε το 1993 ως μία διερμηνευμένη γλώσσα για στατιστικές αναλύσεις και πλέον αποτελεί μία από τις πιο χρησιμοποιούμενες γλώσσες, όχι μόνο για στατιστικούς υπολογισμούς, αλλά και για ανάλυση δεδομένων και εξόρυξη γνώσης, βιοπληροφορικές αναλύσεις, μηχανική μάθηση και άλλα. Η μεγάλη κοινότητα χρηστών που έχει δημιουργήσει η ℝ, έχει οδηγήσει στην ανάπτυξη ενός μεγάλου πλήθους πακέτων, τα οποία έχουν προταθεί για να προσφέρουν λύσεις στις ολοένα και αυξανόμενες ανάγκες των προβλημάτων.

2.2.1. Το Project ℝ/Bioconductor

Το Project Bioconductor (Gentleman *et al.*, 2004; M. Morgan, 2021) είναι μία πρωτοβουλία, η οποία ξεκίνησε το 2001, με στόχο τη δημιουργία βιβλιοθηκών ανοιχτού λογισμικού στην ℝ, οι οποίες σχετίζονται με την ανάλυση βιολογικών δεδομένων. Σήμερα, αποτελεί μία από τις πιο καλά τεκμηριωμένες πλατφόρμες πακέτων και περιέχει πάνω από 2.000 πακέτα για σχολιασμό γονιδιωμάτων, ανάλυση ποιότητας και διαχείριση NGS δεδομένων, δομικό και λειτουργικό σχολιασμό βιομορίων, στοιχίσεις ακολουθιών, γενετικές και επιγενετικές αναλύσεις, εντοπισμό μεταλλάξεων, στατιστικές τεχνικές και πολλά άλλα. Αυτή η μεγάλη ανάπτυξη που έχει γνωρίσει το συγκεκριμένο project, πιθανόν να πηγάζει από τα πολλαπλά πλεονεκτήματα χρήσης της ℝ για ανάλυση δεδομένων, καθώς και επειδή τα τελευταία χρόνια, προτάσσονται συνεχώς ως τρέχουσες εξελίξεις, ένα μεγάλο πλήθος από μεθοδολογίες αλληλούχισης (όπως RNA-seq, *d*RNA-seq, ChIP-seq, Term-seq, κ.α.), των οποίων τα δεδομένα απαιτούν μία πιο εξατομικευμένη διαχείριση.

2.2.2. Αναλυση Αλληλουχισης RNA με την ℝ

Όπως αναφέρθηκε, στην ℝ είναι διαθέσιμα πολλά πακέτα για βιοπληροφορικές αναλύσεις, οι οποίες περιλαμβάνουν και την εκτίμηση ποιότητας και προ-επεξεργασία των δεδομένων που αναλύθηκαν παραπάνω. Μέσα από αυτό το περιβάλλον, η διαχείριση των δεδομένων δεν γίνεται πλέον με τη μορφή ογκωδών αρχείων τοπικά στο σύστημα,

αλλά με την αποθήκευση των δεδομένων σε Data Frames. Αυτά, αν και απαιτούν αρκετή μνήμη σε σχέση με τα αρχεία (SAM, BAM, BED κ.α.), δεν περιέχουν επαναλαμβανόμενη και μη απαραίτητη για την ανάλυση, πληροφορία και μπορούν να επεξεργαστούν με απευθείας προσπέλαση στη μνήμη, ώστε να προκύψει ένα τελικό Data Frame με τα επιθυμητά δεδομένα.

Για τους σκοπούς του χαρακτηρισμού γονιδιωμάτων, χρησιμοποιήθηκαν δεδομένα αλληλούχισης που προέρχονται από μονοκαλλιέργεια της *Escherichia Coli*, σύμφωνα με το πρωτόκολλο αλληλούχισης *Cappable-seq*. Για την προ-επεξεργασία αυτών των δεδομένων, χρησιμοποιήθηκε το Bioconductor πακέτο *QuasR* (Gaidatzis *et al.*, 2014), το οποίο παρέχει μία ροή εργασίας για τη διαχείριση των αναγνώσεων υψηλής διεκπεραιωτικής αλληλούχισης, δηλαδή τη στοίχιση με το γονιδίωμα αναφοράς, την ανάλυση της ποιότητας στοίχισης και αλληλούχισης και την ποσοτικοποίηση συγκεκριμένων περιοχών. Το συγκεκριμένο πακέτο ενσωματώνει τη λειτουργικότητα άλλων δημοφιλών πακέτων, όπως το *Rbowtie*, το *Rhisat2*, το *Rsamtools* και το *IRanges*, τα οποία αποτελούν επίσης διεπαφές στην ℝ των αντίστοιχων τοπικά εγκαταστάσιμων προγραμμάτων. Με τις συναρτήσεις *preprocessReads, qAlign, qQCReport* και *alignmentStats*, εκτελείται η ροή εργασίας για την προ-επεξεργασία και στοίχιση των δεδομένων και παράγονται τα νέα αρχεία FastQ, BAM, καθώς και οι αναφορές τους. Για τα δεδομένα της *Cappable-seq*, χρησιμοποιήθηκαν ως παράμετροι η ακολουθία αντάπτορα `AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC`, το ελάχιστο μήκος των νέων αναγνώσεων ίσο με 26 νουκλεοτίδια, το γονιδίωμα αναφοράς της *Escherichia Coli K-12 MG1655* (με ονομασία *NCBI.20080805.NC_000913* στο πακέτο γονιδιωμάτων *BSgenome* (Pagès, 2021)) και η ελάχιστη τιμή πολυπλοκότητας των αναγνώσεων, ίση με 0.4, η οποία υπολογίζεται σε επίπεδο δινουκλεοτιδίων, η τιμή της έχει σχέση με την πολυπλοκότητα του ανθρώπινου γονιδιώματος και αποτελεί ένα μέτρο απόρριψης αναγνώσεων με μικρή εντροπία κατά Shannon. Με εύκολο τρόπο λοιπόν, μέσα από το περιβάλλον της ℝ μπορούν να γίνουν αναλύσεις, οι οποίες σε αντίθετη περίπτωση θα απαιτούσαν τη λήψη και εγκατάσταση πολλών διαφορετικών εργαλείων, παράγοντας περισσότερα ενδιάμεσα αρχεία δεδομένων. Ως αποτέλεσμα, προκύπτουν δύο αρχεία BAM, ένα για το δείγμα εξέτασης και ένα για το δείγμα ελέγχου, τα οποία περιέχουν συνολικά 8.378.849 και 10.579.644 επιτυχώς χαρτογραφημένες αναγνώσεις, αντίστοιχα.

2.2.3. Εντοπισμος των Θεσεων Εναρξης της Μεταγραφης

Με την κατάλληλη ανάλυση των δεδομένων που προέκυψαν με την τεχνική *Cappable-seq*, μπορούν να εξαχθούν οι γονιδιωματικές συντεταγμένες των εν δυνάμει θέσεων έναρξης της μεταγραφής. Για την επίτευξη αυτού, θα πρέπει να μετασχηματιστούν οι αναγνώσεις, ώστε να διατηρηθεί μόνο η πρώτη βάση σε καθεμία, ανάλογα με τον κλώνο που αυτές ανήκουν. Έτσι, γίνεται αρχικά η μετατροπή του αρχείου BAM σε αρχείο BED, το οποίο όπως αναφέρθηκε, περιέχει τις πληροφορίες θέσεις και μήκους των αναγνώσεων και όχι πληροφορίες αλληλουχίας. Η μετατροπή του αρχείου μπορεί να γίνει με το πακέτο *bedr* (Haider *et al.*, 2019), που ενσωματώνει λειτουργίες από τα αντίστοιχα εργαλεία *BEDTools, BEDOPS & Tabix* και πλέον το περιεχόμενο αυτού του αρχείου αποθηκεύεται ως Data Frame, όπου κάθε στήλη του αντιπροσωπεύει μία μεταβλητή. Ανάλογα με το αν η κάθε ανάγνωση βρίσκεται στον θετικό ή αρνητικό κλώνο, διατηρείται η πρώτη βάση από αριστερά ή δεξιά, αντίστοιχα, λόγω της αντιπαραλληλίας στη διαδοχή των βάσεων. Το πλαίσιο δεδομένων ομαδοποιείται, ανάλογα με το πλήθος των αναγνώσεων σε κάθε θέση έναρξης και τον κλώνο που αυτή ανήκει, δημιουργώντας μία επιπρόσθετη στήλη, η οποία αντιπροσωπεύει το πλήθος των αναγνώσεων, με βάση αυτά τα κριτήρια. Για την ποσοτικοποίηση του πλήθους των μοναδιαίων βάσεων, χρησιμοποιείται η μετρική *Relative Read Score (RRS)*, η οποία εκφράζει τον αριθμό των επικαλύψεων των βάσεων σε μία συντεταγμένη, ανά 1.000.000 βάσεις συνολικά. Κατ΄ αυτό τον τρόπο, προκύπτει ένας αριθμός για κάθε υπαρκτή θέση, ο οποίος ποσοτικοποιεί την ύπαρξη των εν δυνάμει TSS. Η ίδια μεθοδολογία ακολουθείται και για το δείγμα ελέγχου του πειράματος, και αφετέρου τα δύο Data Frames συνδυάζονται με την τεχνική της *πλήρους συνένωσης* και προκύπτει ένα τελικό πλαίσιο δεδομένων.

Μέσα σε αυτό το πλαίσιο, έχουν συνδυαστεί τα TSS και των δύο δειγμάτων, τα οποία είτε είναι παρόντα

και στα δύο δείγματα, είτε βρίσκονται μόνο στο ένα από αυτά, άρα οι τιμές σκορ και επαναλήψεων στο άλλο δείγμα δεν περιέχουν τιμές. Στη συνέχεια, εφαρμόζεται ξανά ομαδοποίηση των δεδομένων, αλλά αυτή τη φορά είναι προσαρμοστική, ανάλογα με το εάν τα γειτονικά TSS έχουν απόσταση μικρότερη ή ίση με 5 νουκλεοτίδια και η νέα ομάδα που προκύπτει, αντιπροσωπεύεται από το TSS με το μεγαλύτερο πλήθος αναγνώσεων. Η συγκεκριμένη ομαδοποίηση γίνεται για την αφαίρεση του θορύβου στα δεδομένα, συνήθως λόγω των ανακριβειών στη θέση πρόσδεσης της RNA πολυμεράσης. Πιο αναλυτικά, καθώς η υπομονάδα της πολυμεράσης προσδένεται στα συντηρημένα μοτίβα του υποκινητή, η θέση πρόσδεσης πολλές φορές δεν είναι απόλυτα καθορισμένη και μπορεί να αποκλίνει κατά ένα μικρό αριθμό νουκλεοτιδίων, συνήθως όχι περισσότερα από 5 (Salgado *et al.*, 2006; Robb *et al.*, 2013). Με αυτό τον τρόπο διασφαλίζεται ότι κάθε TSS είναι μοναδικό, καθώς δεδομένα με μη σημαντική βιολογική υπόσταση, απαλείφονται.

2.2.4. Στατιστικη Αναλυση Δεδομενων Αλληλουχισης

Για να μπορέσουν αυτά τα δεδομένα να αντιπροσωπεύσουν με στατιστικά σημαντικό τρόπο TSS, θα πρέπει επιπρόσθετα να ακολουθηθεί ορισμένη στατιστική ανάλυση. Σε αυτή, λαμβάνονται για κάθε TSS οι τιμές του πλήθους των αναγνώσεων στα δύο δείγματα και με βάση μία δεδομένη κατανομή, υπολογίζεται η τιμή *p-value*, η οποία εκφράζει την πιθανότητα της τυχαίας παρατήρησης των τιμών πλήθους (count data) στα δείγματα. Όσον αφορά την κατανομή που χρησιμοποιείται, υπάρχει μία εκτενής αναφορά στη βιβλιογραφία, σχετικά με την κατανομή που φαίνεται να ακολουθούν οι αναγνώσεις μίας RNA αλληλουχισης δεύτερης γενιάς.

Ιστορικά, με τη διαθεσιμότητα των πρώτων δεδομένων αλληλούχισης νέας γενιάς, υπήρξε ένας μαζικός αριθμός επιστημόνων, οι οποίοι χρησιμοποιούσαν κατανομή Poisson για τη στατιστική ανάλυση αυτών των δεδομένων. Σύμφωνα με αυτή, μια τυχαία μεταβλητή απαριθμεί τον αριθμό των γεγονότων που συμβαίνουν σε μια συγκεκριμένη περιοχή ή χρονική στιγμή. Ένα ιδιαίτερο χαρακτηριστικό αυτής της κατανομής, είναι ότι εμφανίζει *ισοδιασπορά*, δηλαδή η μέση τιμή και η διακύμανσή της, ταυτίζονται. Το γεγονός αυτό οδήγησε λίγα χρόνια αργότερα στην απόρριψή της για την ανάλυση δεδομένων αλληλούχισης, καθώς τα συγκεκριμένα δεδομένα παρουσιάζουν υπερδιακύμανση (overdispersion), δηλαδή η διακύμανση είναι κατά πολύ, ή απλώς μεγαλύτερη από τη μέση τιμή τους *(Σχήμα 2.13A)*, γεγονός που δεν προβλέπεται να υπάρχει από την κατανομή Poisson, καθώς η μοναδική ελεύθερη παράμετρος λ, δεν επιτρέπει η διακύμανση να είναι ανεξάρτητη από τη μέση τιμή.

Ως πρόταση βελτίωσης της στατιστικής διαχείρισης αυτών των δεδομένων, υπήρξαν οι μεμειγμένες κατανομές. Αυτές, μπορούν γενικότερα να προσεγγίσουν με αξιόλογη ακρίβεια πιο ρεαλιστικά δεδομένα και χρησιμοποιούνται ευρέως στη λύση πολλών προβλημάτων, όπου τα δεδομένα διαχωρίζονται επιλεκτικά σε μικρότερους υποπληθυσμούς, οι οποίοι έχουν μεταβαλλόμενο κάποιο χαρακτηριστικό, σε σχέση με το συνολικό πληθυσμό. Ως απόρροια αυτών των κατανομών, είναι η Αρνητική Διωνυμική κατανομή, η οποία προτάθηκε για την εξάλειψη του προβλήματος παραδοχής της ισοδιασποράς και επί της ουσίας πρόκειται για την ίδια κατανομή Poisson, αλλά με τη μόνη τροποποίηση ότι ο συντελεστής λ ακολουθεί την κατανομή Γάμμα, άρα πρόκειται για κατανομή δύο παραμέτρων. Η συγκεκριμένη κατανομή χρησιμοποιείται κατά κόρον από τα σύγχρονα εργαλεία ανάλυσης διαφορικής έκφρασης γονιδίων, όπως το *DESeq2* (Love *et al.*, 2014), το *edgeR* (Robinson *et al.*, 2009) και το *Voom* (Law *et al.*, 2014), τα οποία αν και προσφέρουν ισχυρούς αλγορίθμους για την προσαρμοστική κανονικοποίηση των δεδομένων, ισοβαθμίζοντας περιοχές με υπερεκπροσώπηση αναγνώσεων έναντι των υπόλοιπων περιοχών, κάνουν την παραδοχή ότι όλες οι αναγνώσεις κάθε δείγματος, ακολουθούν ανεξαιρέτως αρνητική διωνυμική κατανομή. Επιπλέον, για την ποσοτικοποίηση των διαφορικά εκφρασμένων γονιδίων, εφαρμόζουν υπολογισμό της διασποράς, είτε με τον υπολογισμό της Μέγιστης Πιθανοφάνειας, είτε με τη χρήση του εμπειρικού εκτιμητή Bayes και εκτελούν τους *ακριβείς ελέγχους* (exact tests) μέσω της αρνητικής διωνυμικής κατανομής με βάση τις εκτιμώμενες παραμέτρους. Αυτό το γεγονός, μπορεί να οδηγήσει ορισμένες φορές σε αρνητικές παραδοχές, οι οποίες εξαρτώνται από τα χαρακτηριστικά του κάθε δείγματος, όπως το μέγεθος του δείγματος, τη διακύμανση των count data και άλλα.

Ειδικά για τη διακύμανση των δεδομένων, η αρνητική διωνυμική κατανομή κάνει επίσης την παραδοχή ότι δύο γονιδιακές περιοχές με παρόμοιο αριθμό αναγνώσεων, θα έχουν και την ίδια διακύμανση, λόγω της παρόμοιας φύσης αυτών των δύο δειγμάτων σε μία RNA-seq μεθοδολογία (δηλαδή των σχετικά παρόμοιων βιολογικών αντιγράφων), σε σχέση για παράδειγμα με τη μεθοδολογία *Cappable-seq*, στην οποία τα δύο δείγματα, εξέτασης και ελέγχου, έχουν πολύ διαφορετικά χαρακτηριστικά, άρα και διακύμανση.

2.2.4.1. Το Βητα – Διωνυμικο Μοντελο

Λαμβάνοντας υπόψιν τα ιδιαίτερα χαρακτηριστικά του δείγματος που προκύπτει από την τεχνική *Cappable-seq*, γίνεται αντιληπτό ότι το μέγεθος του δείγματος είναι αρκετά μικρότερο σε αντίθεση με τα δείγματα ολικής RNA αλληλούχισης, καθώς διατηρούνται συγκεκριμένες μόνο αναγνώσεις των άθικτων 5′ άκρων των γονιδίων. Επίσης, τα συνολικά TSS δεν μπορεί να είναι άπειρα, όπως ενδεχομένως οι αναγνώσεις, αλλά καθορισμένα σε αριθμό, σύμφωνα με τον αριθμό των συνολικών γονιδίων συμπεριλαμβανομένου και της διάχυτης μεταγραφής (pervasive transcription) (Lybecker *et al.*, 2014), η οποία ορίζει ότι συνήθως η μεταγραφή ξεκινά από πολλές και διαφορετικές θέσεις για ένα γονίδιο. Επιπλέον, ενώ στην ανάλυση διαφορικής έκφρασης, η ποσοτικοποίηση γινόταν ανάλογα με τον αριθμό των αναγνώσεων σε μία ολόκληρη γονιδιακή περιοχή, δηλαδή σε κάθε γονιδιακό *locus*, στην ανάλυση των TSS, κάθε ένα ποσοτικοποιείται ανάλογα με τη μοναδιαία γονιδιακή συντεταγμένη στην οποία ανήκει. Άρα, θα πρέπει να χρησιμοποιηθεί μία περισσότερο ευαίσθητη κατανομή, η οποία να προσαρμόζεται καλύτερα πάνω στα δεδομένα αυτά και να έχει το μικρότερο δυνατό μέσο τετραγωνικό σφάλμα (Mean squared error (MSE)). Μία τέτοια κατανομή, φαίνεται να είναι η Βήτα – Διωνυμική Κατανομή, η οποία ανήκει επίσης στην κατηγορία των μεμειγμένων κατανομών και είναι η συζευγμένη προηγούμενη (conjugate prior) της αρνητικής διωνυμικής κατανομής. Έχει μία επιπλέον ελεύθερη παράμετρο, άρα συνολικά 4 παραμέτρους και καλή προσαρμοστικότητα σε αριθμητικά δεδομένα, καθώς δεν θεωρεί δεδομένη την πιθανότητα επιτυχίας των διωνυμικών δοκιμών σε μία γονιδιακή συντεταγμένη. Το βήτα – διωνυμικό μοντέλο περιγράφει άμεσα τη διακύμανση στις πιθανότητες επιτυχίας των αναγνώσεων και έτσι απλοποιεί την προσαρμογή του μοντέλου στα δεδομένα και με αυτή την έννοια μπορεί να παρέχει μια πιο άμεση ερμηνεία της υπερδιασποράς στα δεδομένα. Η συνάρτηση πυκνότητας πιθανότητας (Probability Density Function (PDF)) αυτής, δίνεται παρακάτω *(Εξίσωση 2.3)*, η οποία μπορεί εναλλακτικά να γραφτεί συναρτήσει της Γάμμα συνάρτησης της ομώνυμης κατανομής *(Εξίσωση 2.4)*.

$$f(k|N^+, n, N^-) = \binom{n}{k} \frac{B(k + N^+, n - k + N^-)}{B(N^+ + N^-)} \tag{2.3}$$

$$f(k|N^+, n, N^-) = \frac{\Gamma(n+1)\Gamma(k+N^+)}{\Gamma(k+1)\Gamma(n-k+1)} \frac{\Gamma(n-k+N^-)\Gamma(N^+ + N^-)}{\Gamma(n+N^+ + N^-)\Gamma(N^+)\Gamma(N^-)} \tag{2.4}$$

Κατ′ αντιστοιχία με την αρνητική διωνυμική κατανομή, ο αριθμός των επιτυχιών έχει αντικατασταθεί με την πιθανότητα επιτυχιών και ως επιτυχία λαμβάνεται κάθε επιπλέον ανάγνωση μίας βάσης στο εμπλουτισμένο *Cappable-seq* δείγμα, ενώ ως αποτυχία, κάθε επιπλέον ανάγνωση μονής βάσης στο δείγμα ελέγχου, επειδή πρόκειται για μία κατανομή διπλής δειγματοληψίας. Θα πρέπει όμως να ληφθεί υπόψιν στον τρόπο διαχείρισης των δεδομένων, ότι το δείγμα ελέγχου περιέχει όλες τις αναγνώσεις του εμπλουτισμένου δείγματος, συν τις αναγνώσεις που φέρεται να έχουν επεξεργασμένα άκρα, άρα ισχύει ότι για οποιαδήποτε θέση, το δείγμα ελέγχου θα περιέχει το ίδιο ή μεγαλύτερο πλήθος αναγνώσεων, σε σχέση με το αντίστοιχο εμπλουτισμένο, γεγονός που δεν ισχύει πάντα για τα TSS. Έτσι, επιτυγχάνεται υψηλή απόδοση αυτού του μοντέλου, ακόμα και σε πολύ μικρά δείγματα ή γονίδια με χαμηλή έκφραση.

Για τον υπολογισμό των *p-values* μέσα από αυτό το μοντέλο, χρησιμοποιήθηκε το ℝ πακέτο VGAM (Yee, 2010), το οποίο περιέχει υλοποιήσεις για γενικευμένα γραμμικά μοντέλα (Generalized Linear Models (GLM)),

όπως η αρνητική και η βήτα διωνυμική κατανομή, καθώς και υλοποιήσεις για γενικευμένα προσθετικά μοντέλα (Generalized Additive Models (GAM)). Ως είσοδοι, δίνονται για κάθε TSS οι τιμές των RRS σκορ των δύο δειγμάτων, οι οποίες αντιπροσωπεύουν την ποσοτικοποίηση αυτών, έπειτα από κανονικοποίηση προς συνολικά 1 εκατομμύριο TSS. Τέλος, παρατηρώντας τις συνολικές στοιχισμένες αναγνώσεις των δύο δειγμάτων, καθώς και τα σκορ RRS, προκύπτει ότι τη μεγαλύτερη συνεισφορά στο παραγόμενο *p-value* την έχει όπως είναι αναμενόμενο, το δείγμα εξέτασης, λόγω των λιγότερων αναγνώσεων του, άρα η ίδια μεταβολή στο πλήθος των αντιγράφων ενός TSS στα δύο δείγματα, θα έχει ως αποτέλεσμα μεγαλύτερη μεταβολή στο RRS σκορ του δείγματος εξέτασης.

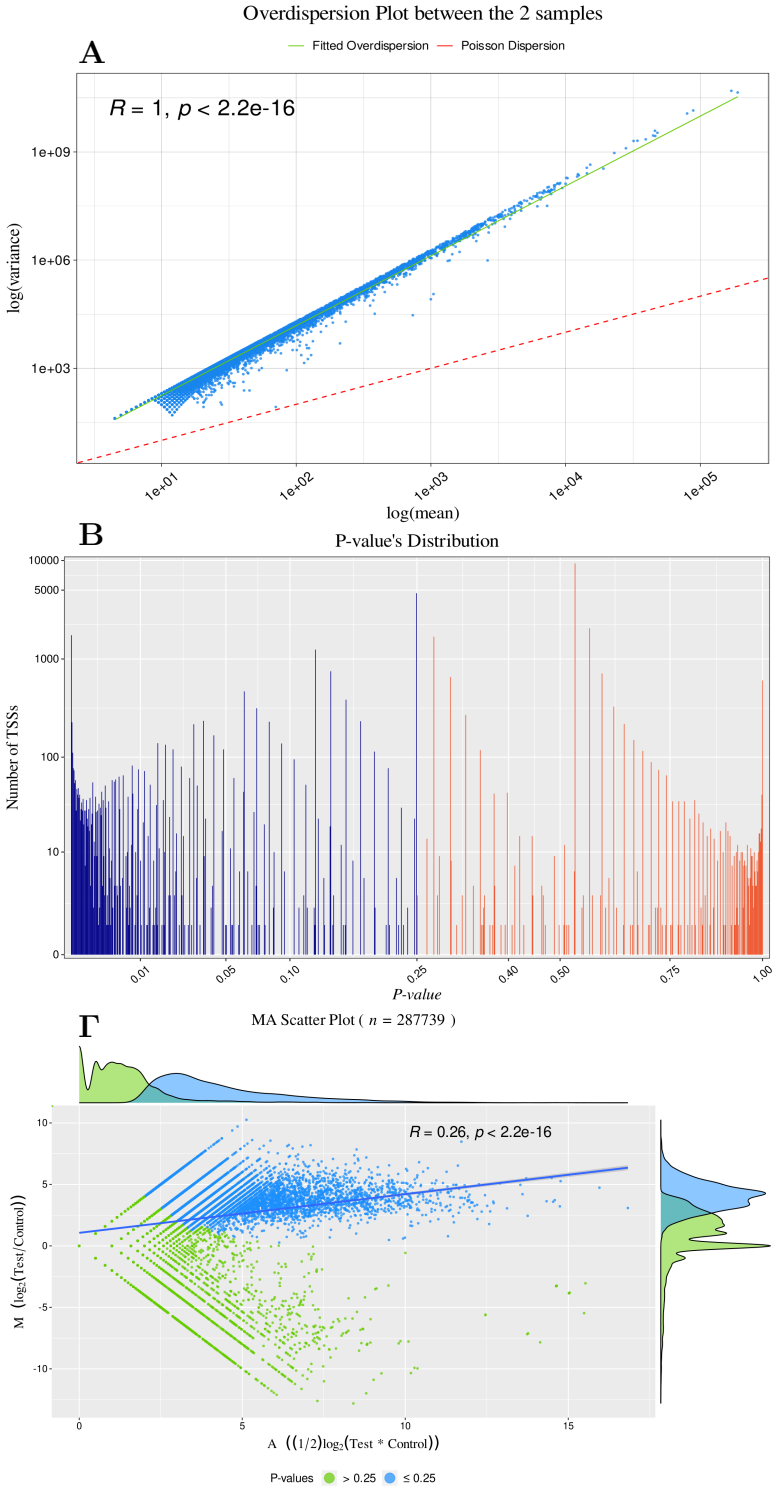
2.2.4.2. Αξιολογηση του Μοντελου

Με την παραπάνω στατιστική ανάλυση και εφαρμόζοντας ως μέγιστο κατώφλι *p-value* την τιμή 0.25, προκύπτουν συνολικά 15.085 TSS *(Σχήμα 2.13B)*. Η συγκεκριμένη τιμή *p-value* αν και φαίνεται φαινομενικά μεγάλη σε σχέση με τα συνήθη επίπεδα σημαντικότητας της τάξης του 1 ή 5%, αποτελεί μία ικανοποιητική επιλογή για το δείγμα. Η γραφική αναπαράσταση του διαγράμματος MA *(Σχήμα 2.13)*, επισημαίνει πως η συγκεκριμένη τιμή διατηρεί τα TSS με τιμές M και A μεγαλύτερες από το 0, με προσαρμοστικό τρόπο στα δεδομένα. Έτσι, εξασφαλίζεται πιο βέλτιστος διαχωρισμός των δεδομένων, σε σχέση με τεχνικές, οι οποίες αναφέρονται συχνά στη βιβλιογραφία, όπως τα αυθαίρετα κατώφλια, ή η διατήρηση των θετικών τιμών, μέσα από το διάγραμμα MA. Τέλος, τα TSS με την επιθυμητή στατιστική σημαντικότητα, εμφανίζουν κατά μέσο όρο μικρότερη τιμή M από την τιμή A, γεγονός που συνεπάγεται μεγαλύτερη ποσότητα των RNA με επεξεργασμένα 5′ άκρα *in vivo*, σε σχέση με τα αντίστοιχα άθικτα του δείγματος εξέτασης.

2.2.5. Χαρακτηρισμος TSS και 5′ Αμεταφραστων Περιοχων

Τα στατιστικώς σημαντικά TSS του προηγούμενη βήματος, κατηγοριοποιούνται σε 5 κατηγορίες σύμφωνα με τη θέση και τον προσανατολισμό τους, ανάμεσα ή εσωτερικά στα γονίδια *(Σχήμα 2.14A)*. Οι κατηγορίες αυτές περιγράφονται παρακάτω:

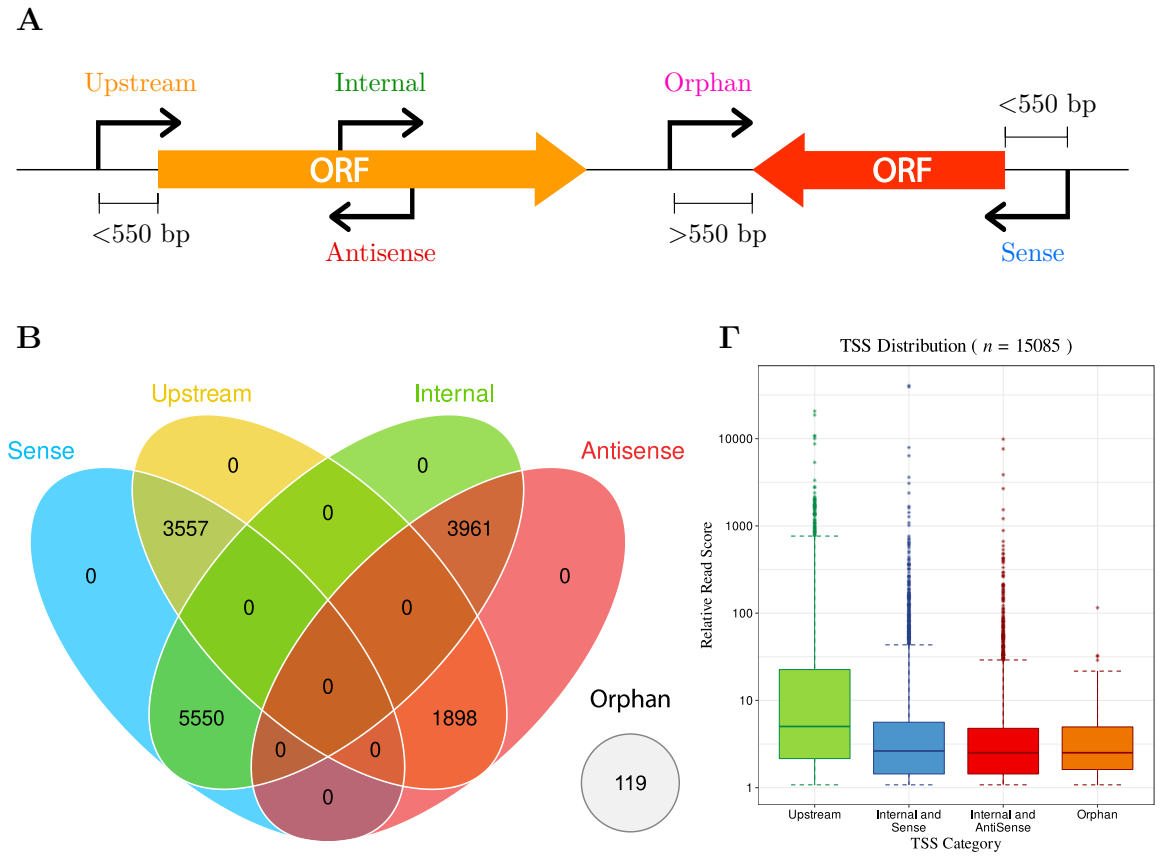
- **Ανοδικά (Upstream):** όσα βρίσκονται πριν το κωδικόνιο έναρξης των *m*RNA ή πριν από το 5′ άκρο των *s*RNA, σε απόσταση μικρότερη από 550 νουκλεοτίδια από αυτές τις περιοχές.
- **Εσωτερικά (Internal):** όσα βρίσκονται μέσα σε κάποιο σχολιασμένο γονίδιο, προς οποιαδήποτε κατεύθυνση. Αυτή η κατηγορία απαιτεί ιδιαίτερη προσοχή, διότι εάν κάποιο TSS είναι εσωτερικό ενός γονιδίου αντίθετης κατεύθυνσης, αλλά έχει απόσταση μικρότερη από 550 βάσεις από ένα ομόρροπο καθοδικό και γειτονικό γονίδιο, τότε το συγκεκριμένο TSS θα θεωρείται ανοδικό αυτού και όχι εσωτερικό, εμφανίζοντας επικάλυψη των υποκινητών με άλλο γονίδιο.
- **Ομόρροπα (Sense):** εάν ο προσανατολισμός των TSS είναι σύμφωνος με το υπάρχον γονίδιο στην αντίστοιχη θέση. Εάν το TSS είναι ανοδικό ενός γονιδίου, τότε η κατεύθυνση λαμβάνεται σύμφωνα με το πιο γειτονικό γονίδιο που βρίσκεται καθοδικά αυτού. Εάν υπάρχουν γονίδια με διαφορετικές κατευθύνσεις, δηλαδή γονίδια που βρίσκονται σε διαφορετικούς κλώνους, τότε ως μέτρο θεωρείται το ομόρροπο προς το TSS γονίδιο, όπως επίσης συμβαίνει και σε επικαλυπτόμενα γονίδια ενός κλώνου, δηλαδή γονίδια που παρουσιάζουν υπέρθεση, κυρίως λόγω εισαγωγής ακολουθιών, δηλαδή μεταλλαξογένεσης.
- **Αντίρροπα (Antisense):** εάν ο προσανατολισμός των TSS είναι αντίθετος με αυτόν του γονιδίου. Κατ΄ αντιστοιχία με την προηγούμενη περίπτωση, εάν υπάρχουν γονίδια με διαφορετικές κατευθύνσεις, τότε ως μέτρο θεωρείται το ομόρροπο προς το TSS γονίδιο, όπως συμβαίνει και στα επικαλυπτόμενα γονίδια. Ένα TSS θα είναι αντίρροπο, μόνο εάν δεν υπάρχει σε εγγύτητα κάποιο ομόρροπο γονίδιο, το οποίο μπορεί να συσχετιστεί με αυτό.



Σχήμα 2.13: *Στατιστική ανάλυση του τελικού πλαισίου δεδομένων.* ***(Α)*** *Αναπαράσταση της υπερδιασποράς του δείγματος. Η κόκκινη γραμμή απεικονίζει την ισοδιασπορά των δεδομένων, η οποία είναι η παραδοχή στην κατανομή Poisson και τα δεδομένα με μπλε, δείχνουν την πραγματική υπερδιασπορά των δεδομένων. Η πράσινη γραμμή δείχνει την ευθεία της παλινδρόμησης, η οποία έχει προέλθει από το γενικευμένο γραμμικό μοντέλο (GLM), σύμφωνα με το οποίο, ο εκτιμητής ελαχίστων τετραγώνων που χρησιμοποιείται στην κλασσική γραμμική παλινδρόμηση, έχει αντικατασταθεί με εκτιμητές μέγιστης πιθανοφάνειας. Επιπλέον, ο συντελεστής Pearson συγκλίνει στην τιμή 1.* ***(Β)*** *Κατανομή των τιμών p-value. Οι άξονες είναι μετασχηματισμένοι κατά λογαριθμική κλίμακα, συν 1 (log1p). Με μπλε χρώμα φαίνονται οι διατηρηθείσες παρατηρήσεις ($n = 15.085$ έναντι των 287.739, $p - value \leq 0.25$).* ***(Γ)*** *Διάγραμμα MA των δύο δειγμάτων. Κάθε σημείο αντιπροσωπεύει ένα TSS και με μπλε χρώμα φαίνονται αυτά που έχουν στατιστική σημαντικότητα, με p-value μικρότερο του 0.25. Όπως προκύπτει, η διατήρηση των δεδομένων είναι προσαρμοστική και έχουν διατηρηθεί TSS με τιμές M και A άνω του 0.*

- **Ορφανά (Orphans):** όσα βρίσκονται ανοδικά και σε απόσταση μεγαλύτερη των 550 νουκλεοτιδίων από τη θέση έναρξης ενός οποιουδήποτε γειτονικού γονιδίου. Σε αυτή την περίπτωση, γίνεται η θεώρηση ότι δεν υπάρχει κάποιο σχολιασμένο γονίδιο σε εγγύτητα με το TSS, άρα το τελευταίο κατηγοριοποιείται ως *ορφανό*.

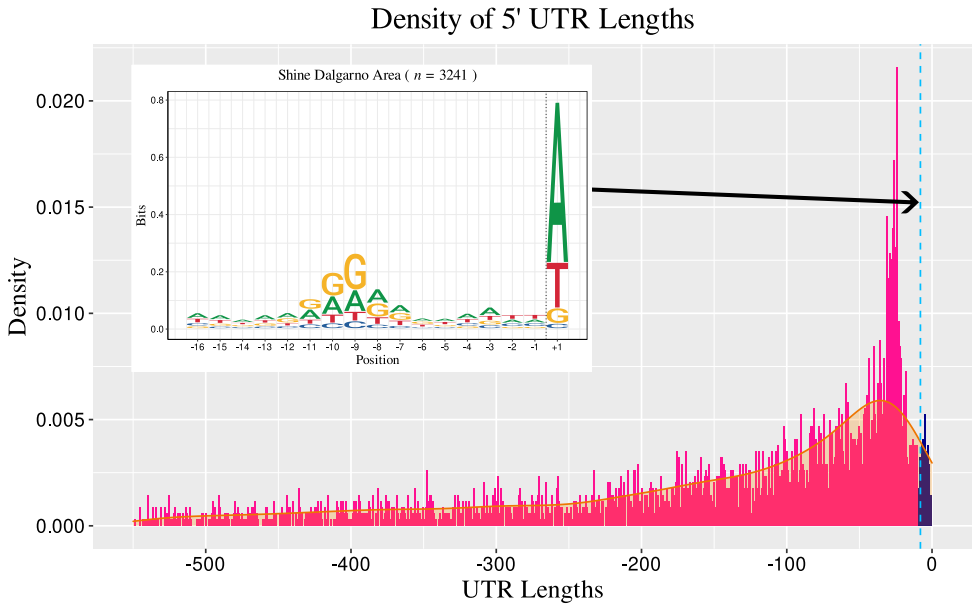
Όπως γίνεται αντιληπτό, ένα TSS μπορεί να ανήκει σε περισσότερες από μία κατηγορίες, δηλαδή ένα ομόρροπο TSS μπορεί να είναι είτε ανοδικό, είτε εσωτερικό. Ο παραπάνω σχολιασμός υλοποιήθηκε με τη βοήθεια της R, αφού πρώτα ελήφθησαν από τις βάσεις δεδομένων *RegulonDB* (Santos-Zavaleta *et al.*, 2018) και *EcoCyc* (Keseler *et al.*, 2021), τα συνολικά γονίδια της *Escherichia Coli* με μαζικό τρόπο συνδυαστικά, συμπεριλαμβανομένων των γονιδιακών τους συντεταγμένων και του κλώνου στον οποίο ανήκουν, καθώς και πληροφορίες για μη κωδικά RNA σε ξεχωριστό αρχείο. Τα σχολιασμένα δεδομένα αναπαραστάθηκαν σε διάγραμμα Venn *(Σχήμα 2.14B)*, το οποίο επισημαίνει τις επικαλύψεις στις κατηγορίες των TSS και τα ποσοτικοποιεί σε κάθε μία κατηγορία. Τέλος, τα TSS που ανήκουν στις κατηγορίες Sense και Upstream, μπορούν συνολικά να χαρακτηριστούν ως πρωταρχικά (ή αντιπροσωπευτικά), λόγω του γεγονότος ότι περιέχονται ανάμεσα στους υποκινητές και στην 5′ αμετάφραστη περιοχή των κωδικών γονιδίων, καθώς και επειδή συνήθως το πρωταρχικό μοντέλο της γονιδιακής μεταγραφής, περιέχει ένα ή περισσότερα TSS, που ακολουθούνται από καθορισμένη αμετάφραστη περιοχή και εν συνεχεία το γονίδιο ξεκινά με το κωδικόνιο έναρξης.



Σχήμα 2.14: *Χαρακτηρισμός των TSS σύμφωνα με τον προσανατολισμό και τη θέση τους στο γονιδίωμα.* ***(Α)*** *Αναπαράσταση κατηγοριών των TSS για ορισμένες περιπτώσεις θέσης και προσανατολισμού. Οποιεσδήποτε άλλες περιπτώσεις μπορούν να σχολιαστούν κατ΄ αυτό τον τρόπο.* ***(Β)*** *Διάγραμμα Venn των χαρακτηρισμένων δεδομένων.* ***(Γ)*** *Κατηγοριοποίηση των TSS με βάση την τιμή RRS. Φαίνεται ότι τα ανοδικά TSS κατέχουν υψηλότερες τιμές RRS, σε σχέση με τις υπόλοιπες κατηγορίες.*

Αναφορικά με την 5′ αμετάφραστη περιοχή των RNA, έχει αποδειχτεί ότι καθορίζει την αποδοτικότητα της μετάφρασης στα *m*RNA, μέσω του μεταβλητού μήκους της. Για να μπορέσουν να εξαχθούν αυτές οι περιοχές από τα *m*RNA, γράφτηκε επιπλέον πρόγραμμα στην R, το οποίο αναθέτει στα σχολιασμένα TSS, το γονίδιο στο οποίο φέρεται να αντιστοιχούν, άρα κατόπιν σύγκρισής τους με το αρχείο των μη κωδικών RNA, μπορεί

να εξακριβωθεί εάν πρόκειται για γονίδιο που κωδικοποιεί πρωτεΐνη ή όχι. Λαμβάνοντας μόνο τα ανοδικά και ομόρροπα TSS (Upstream & Sense) που αντιστοιχούν σε γονίδιο που παράγει *m*RNA, δύνανται να εξαχθεί αυτή η περιοχή από το γονιδίωμα αναφοράς και να εντοπιστούν κοινά μοτίβα, η περιοχή Shine – Dalgarno καθώς και Leaderless *m*RNA και έτσι, μπορούν να προκύψουν αξιόπιστα συμπεράσματα για την κατανομή του μήκους αυτής της περιοχής *(Σχήμα 2.15)*.

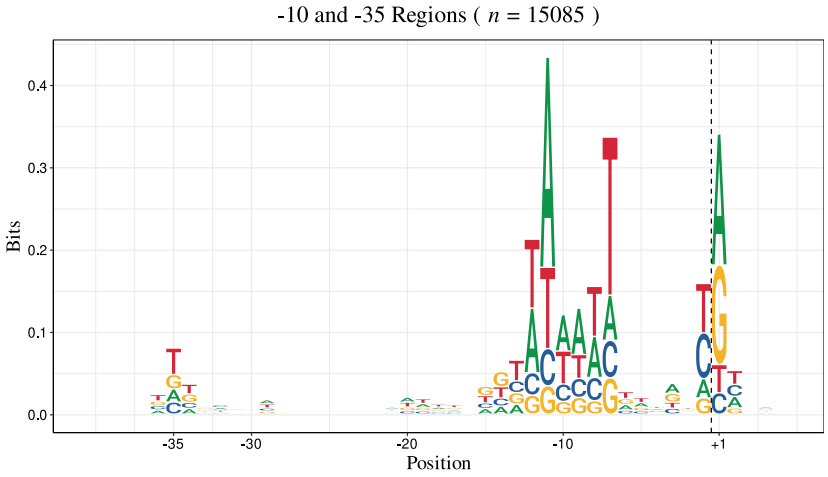


Σχήμα 2.15: *Κατανομή πυκνότητας των 5′ αμετάφραστων περιοχών των mRNA. Η συγκεκριμένη περιοχή οριοθετείται από το TSS έως το κωδικόνιο έναρξης του εκάστοτε γονιδίου. Με μπλε χρώμα απεικονίζονται τα υποψήφια ως Leaderless mRNA, ενώ με τη γαλάζια κάθετη γραμμή, φαίνεται η περιοχή Shine – Dalgarno, η οποία βρίσκεται περίπου 8 βάσεις ανοδικά από το κωδικόνιο έναρξης. Το εσωτερικό παράθυρο, δείχνει το βαθμό συντήρησης της περιοχής SD. Σε αυτή φαίνονται συντηρημένα με συγκριτικά μεγαλύτερη πιθανότητα, κάποια δινουκλεοτίδια γουανίνης, διαχωρισμένα με βάσεις Αδενίνης, παρουσιάζοντας συσχέτιση με το μοτίβο Shine – Dalgarno (SD).*

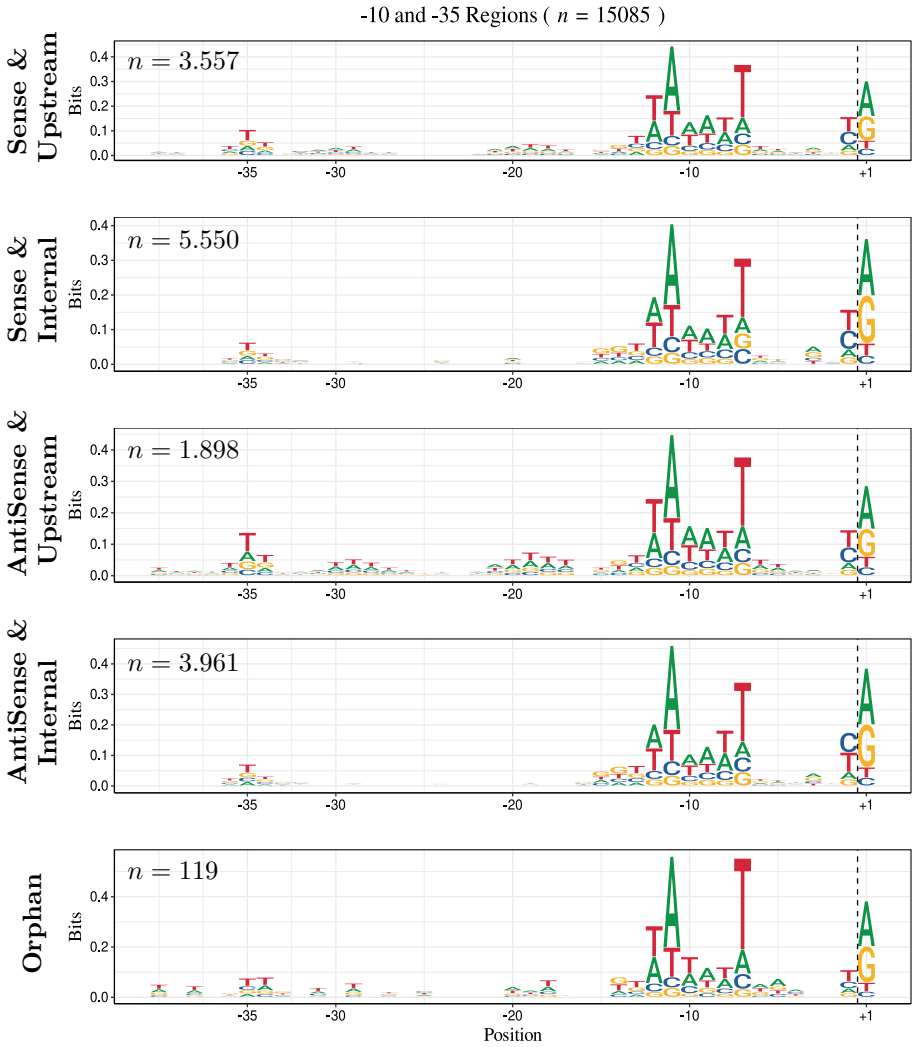
Στο ίδιο μήκος κύματος κυμαίνονται και οι περιοχές -10 και -35 της υπομονάδας σ^{70} της RNA πολυμεράσης, οι οποίες βρίσκονται ανοδικά του εκάστοτε TSS. Σύμφωνα με τη βάση δεδομένων *EcoCyc*, περισσότερα από το 90% των υποκινητών των γονιδίων περιέχουν περιοχές πρόσδεσης για τον παράγοντα σ^{70} (πρωτεΐνη RpoD), ενώ τα υπόλοιπα γονίδια ελέγχονται από εναλλακτικούς παράγοντες Σίγμα, οι οποίοι εκφράζονται κάτω από συγκεκριμένες συνθήκες στρες, θερμότητας ή ορισμένης συγκέντρωσης ουσιών. Τα συντηρημένα μοτίβα -10 και -35 *(Σχήμα 2.16)* που μπορούν να εξαχθούν, είναι τα `ntTAaatT` και `nTTn`, αντίστοιχα. Στην περιοχή -10 υπάρχει μεγαλύτερη συντήρηση στις νουκλεοτιδικές θέσεις -11 και -7, ενώ η θέση -35 είναι αυτή με τη μεγαλύτερη συντήρηση στην ομώνυμη περιοχή, υποδεικνύοντας τα κοινά μεταγραφικά μοτίβα των αρνητικών κατά Gram βακτηρίων (Schlüter *et al.*, 2013). Το *Σχήμα 2.17*, δείχνει τις ανωτέρω συντηρημένες περιοχές για κάθε κατηγορία TSS ξεχωριστά, με τα ορφανά να εμφανίζουν παρόμοια συντήρηση με τις υπόλοιπες κατηγορίες TSS.

Όπως προκύπτει από το παραπάνω γράφημα *(Σχήμα 2.16)*, η περιοχή των TSS εμφανίζει και αυτή συντήρηση, ακολουθώντας το μοτίβο `TA` στη θέση -1 έως +1. Ωστόσο, μπορούν να προκύψουν επιπλέον δεδομένα από αυτή την περιοχή με περαιτέρω ανάλυση. Διαχωρίζοντας τη συγκεκριμένη θέση, ανάλογα με τον τύπο των δινουκλεοτιδίων που περιέχει, δηλαδή αν αντιστοιχούν σε πουρίνες (A ή G) ή πυριμιδίνες (C ή T) *(Σχήμα 2.18)*, προκύπτει μία εμφανής προτίμηση στο μοτίβο πυριμιδίνη – πουρίνη (YR), στο οποίο αντιστοιχούν 9.033 (59.88%) TSS (μπλε χρώμα θηκογραμμάτων). Αντιθέτως, ο συνδυασμός πουρίνης – πυριμιδίνης (RY), εμφανίζεται να κατέχει τα λιγότερα TSS, με αριθμό 705 (4.67%) (ροζ χρώμα θηκογραμμάτων).

Τα TSS που ταυτοποιήθηκαν με την *Cappable-seq* και εντοπίζονται μέσα σε γονίδια, αντιπροσωπεύουν 9.511 (63.04%), από τα συνολικά 15.085. Από τα συνολικά, μόλις 346 βρίσκονται σε γονίδια που δεν κωδικοποιούν πρωτεΐνες, αλλά *t*RNA, *r*RNA ή *s*RNA, όπου τα τελευταία κωδικοποιούνται κυρίως στα Small Open Reading

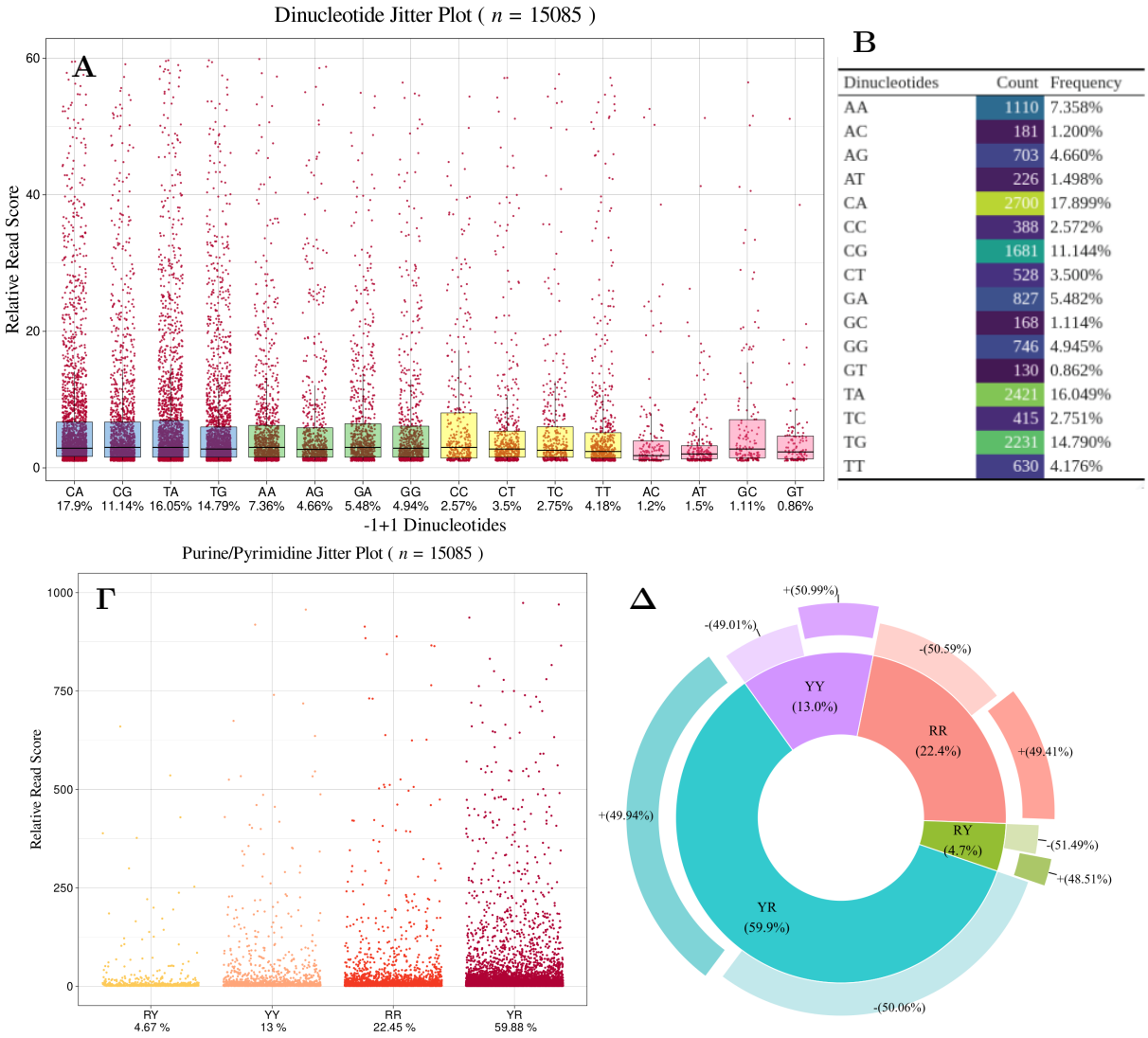


Σχήμα 2.16: *Συντηρημένα μοτίβα ανοδικά των συνολικών TSS της Escherichia Coli. Σε αυτά φαίνονται οι περιοχές -10 (γνωστή και ως Pribnow box) και η περιοχή -35. Με την κάθετη διακεκομμένη γραμμή, φαίνεται η συντήρηση των TSS (περιοχή +1).*

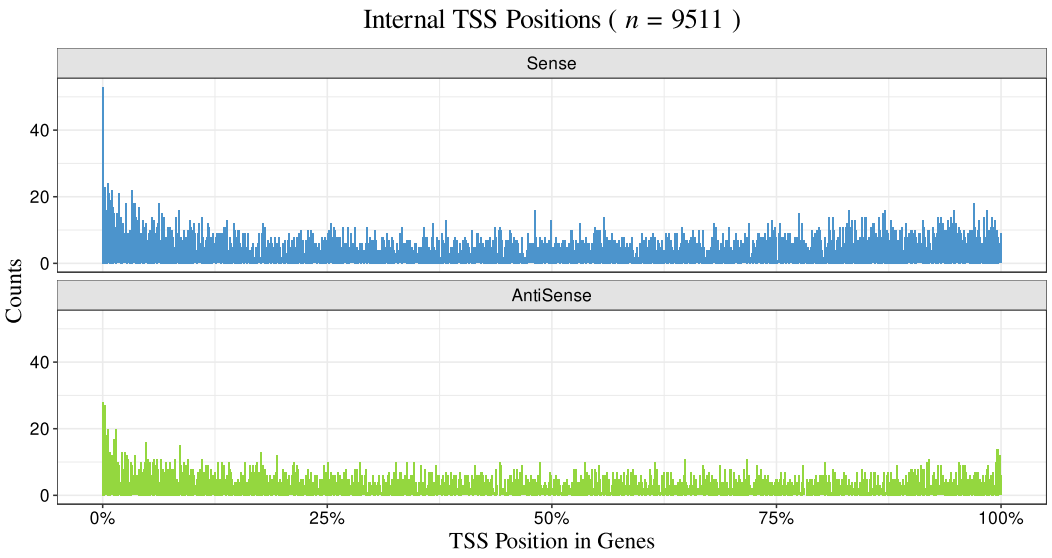


Σχήμα 2.17: *Συντηρημένες περιοχές ανοδικά των TSS για κάθε μία κατηγορίας. Στα διαγράμματα logo διακρίνονται οι περιοχές -10 και -35. Μεγαλύτερη συντήρηση σε αυτό το μήκος, φαίνεται να υπάρχει στα Sense & Upstream, Antisense & Upstream, καθώς και στα Orphan TSS. Η συντήρηση των τελευταίων προκαλεί ιδιαίτερη εντύπωση, επειδή όπως αναφέρθηκε, τα Orphan TSS βρίσκονται σε μεγάλη απόσταση από κάποιο σχολιασμένο γονίδιο. Το γεγονός αυτό προμηνύει την ύπαρξη νέων γονιδίων, τα οποία δεν έχουν ακόμη σχολιαστεί στο γονιδίωμα αναφοράς.*

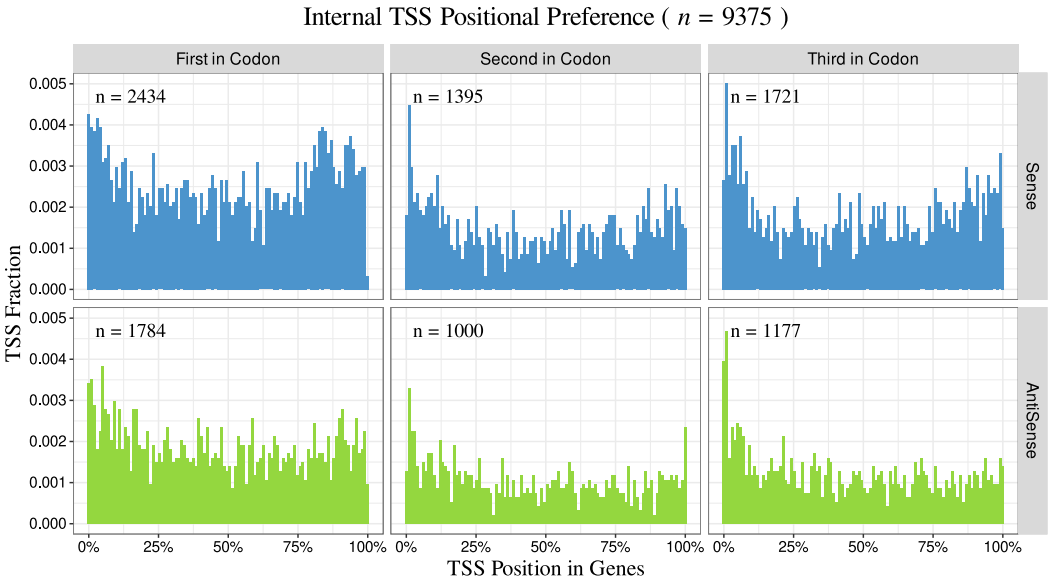
Frames (sORF) που αναλύονται παρακάτω. Επίσης, από τα 9.511 εσωτερικά TSS, τα 5.550 (58.35%) έχουν τον ίδιο προσανατολισμό με το γονίδιο στο οποίο ανήκουν και τα υπόλοιπα 3.961 (41.64%) αντίθετο προσανατολισμό. Τα TSS με ίδιο προσανατολισμό με τα γονίδια, τείνουν να εμφανίζουν επικρατέστερη θέση στην έναρξη των γονιδίων *(Σχήμα 2.19)*, καθώς και με λίγο μικρότερη τάση στην 3′ περιοχή αυτών. Επίσης, οι θέσεις των εσωτερικών TSS με αντίθετο προσανατολισμό, τείνουν να είναι πιο ομοιόμορφα κατανεμημένες σε όλο το μήκος των γονιδίων. Ανάλυση της θέσης των TSS στα κωδικά γονίδια, έχει δείξει ότι υπάρχει προτίμηση των ομόρροπων και αντίρροπων TSS, στο να βρίσκονται κυρίως στην πρώτη θέση και με μικρότερη πιθανότητα στη δεύτερη και τρίτη θέση, αναφορικά με το βήμα τριπλέτας του ανοιχτού πλαισίου ανάγνωσης, δηλαδή σχετικά με τη θέση στο κωδικόνιο που ανήκουν μέσα σε ένα ανοιχτό πλαίσιο ανάγνωσης *(Σχήμα 2.20)*. Τέλος, διατηρώντας μόνο το αντιπροσωπευτικό (representative) TSS κάθε γονιδίου, δηλαδή αυτό που είναι ανοδικό, ομόρροπο και αντιστοιχεί στο μικρότερο *p-value* αυτών, προέκυψαν 1.900 TSS.



Σχήμα 2.18: *Ανάλυση της σύνθεσης της περιοχής των TSS, μαζί με την πρώτη βάση ανοδικά αυτών.* ***(Α)*** *Αναπαράσταση των πιθανών συνδυασμών δινουκλεοτιδίων. Με μπλε χρώμα φαίνεται ο συνδυασμός πυριμιδίνης – πουρίνης (YR), με πράσινο ο συνδυασμός πυριμιδίνης – πυριμιδίνης (YY), με κίτρινο ο συνδυασμός πουρίνης – πουρίνης (RR) και με ροζ ο συνδυασμός πουρίνης – πυριμιδίνης (RY). Ο κάθετος άξονας αντιστοιχεί στο σκορ που κατέχει κάθε δινουκλεοτίδιο, το οποίο έχει προέλθει από κανονικοποίηση του πλήθους των TSS μιας συγκεκριμένης περιοχής, προς 1.000.000 συνολικά TSS.* ***(Β)*** *Πίνακας συχνότητας των πιθανών δινουκλεοτιδίων.* ***(Γ)*** *Αναπαράσταση της ομαδοποίησης των συνδυασμών, με βάση τις κατηγορίες που αναφέρθηκαν.* ***(Δ)*** *Κυκλικό διάγραμμα, το οποίο στον εσωτερικό κύκλο αναπαριστά την κατανομή σε βάσεις της ίδιας περιοχής που αναφέρθηκε και στον εξωτερικό κύκλο εμφανίζεται το ποσοστό που κατέχει ο κάθε κλώνος, για κάθε κατηγοριοποίηση.*



Σχήμα 2.19: *Εσωτερικά TSS. Κατανομή του αριθμού των TSS, τα οποία είναι ομόρροπα (πάνω) και αντίρροπα (κάτω), σε σχέση με τον προσανατολισμό και τη θέση στο γονίδιο που αντιστοιχούν, εκφρασμένη ως ποσοστό.*



Σχήμα 2.20: *Προτίμηση των εσωτερικών TSS στα γονίδια που κωδικοποιούν πρωτεΐνες. Οι τρεις στήλες αντιπροσωπεύουν την απόσταση των εσωτερικών TSS από την αρχή των γονιδίων, εκφρασμένη ως θέση στο ανοιχτό πλαίσιο ανάγνωσης, στο πρώτο, δεύτερο και τρίτο νουκλεοτίδιο ενός κωδικονίου, αντίστοιχα. Οι δύο γραμμές αντιστοιχούν στα ομόρροπα και αντίρροπα TSS. Παρατηρείται ότι τόσο τα ομόρροπα, όσο και τα αντίρροπα TSS τείνουν να εμφανίζονται κυρίως στο πρώτο νουκλεοτίδιο ενός κωδικονίου και με μικρότερη πιθανότητα στο δεύτερο και τρίτο. Ο κάθετος άξονας εκφράζει τη συχνότητα αυτών των TSS προς τα συνολικά εσωτερικά, για κάθε θέση μέσα στα γονίδια και κάθε κατηγοριοποίηση που αναφέρθηκε.*

2.2.6. Αναλυση των Leaderless mRNA

Λαμβάνοντας τις συνολικές 5′ αμετάφραστες περιοχές των mRNA, προκύπτει με μία αρχική ανάλυση, ότι τα συνολικά μετάγραφα με περιοχή μικρότερη των 10 νουκλεοτιδίων, είναι ίσα με 115. Κρίνεται όμως απαραίτητος ο περαιτέρω σχολιασμός τους, για το εάν ο εντοπισμός της πρώτης τριάδας `ATG` οφείλεται στην τύχη ή πράγματι επιτελεί κωδικόνιο έναρξης σε κάποιο γονίδιο. Για να επιτευχθεί ο συγκεκριμένος σχολιασμός, απομονώνονται τα mRNA με αμετάφραστη περιοχή μικρότερη από 10 βάσεις και σε αυτά αναζητούνται σε όλο το μήκος τους, συγκεκριμένες τριπλέτες έναρξης της μεταγραφής, οι οποίες να ταυτίζονται με το σχολιασμό των γονιδίων που

λήφθηκε από τις βάσεις δεδομένων. Σύμφωνα με έρευνες (Villegas and Kropinski, 2008; Kearse and Wilusz, 2017), αυτές οι τριπλέτες έναρξης της μετάφρασης, έχει αποδειχθεί ότι μπορεί να είναι οι `ATG, CTG, GTG & TTG`, οι οποίες είναι ικανές να ξεκινήσουν τη μετάφραση, αντιπροσωπεύοντας το αμινοξύ της Μεθειονίνης. Πιο αναλυτικά, ανάλυση διαφορετικών βακτηριακών γονιδιωμάτων, έφερε στο προσκήνιο ένα ποσοστό γονιδίων της τάξης του 20%, τα οποία φέρουν ως κωδικόνιο έναρξης κάποια διαφορετική τριπλέτα από την κοινή `ATG`. Επίσης, είναι γνωστό ότι ορισμένα *Small Open Reading Frames* (sORF), που κωδικοποιούν ηγετικές πρωτεΐνες (leader peptides) με ρυθμιστικό και λειτουργικό ρόλο σε βιολογικές διεργασίες, όπως η μεταγραφή και η μετάφραση, μπορούν και αυτά να ξεκινήσουν τη μετάφραση από κάποιο κωδικόνιο που διαφέρει κατά μία βάση (συνήθως την πρώτη) από το `ATG` (Kearse and Wilusz, 2017).

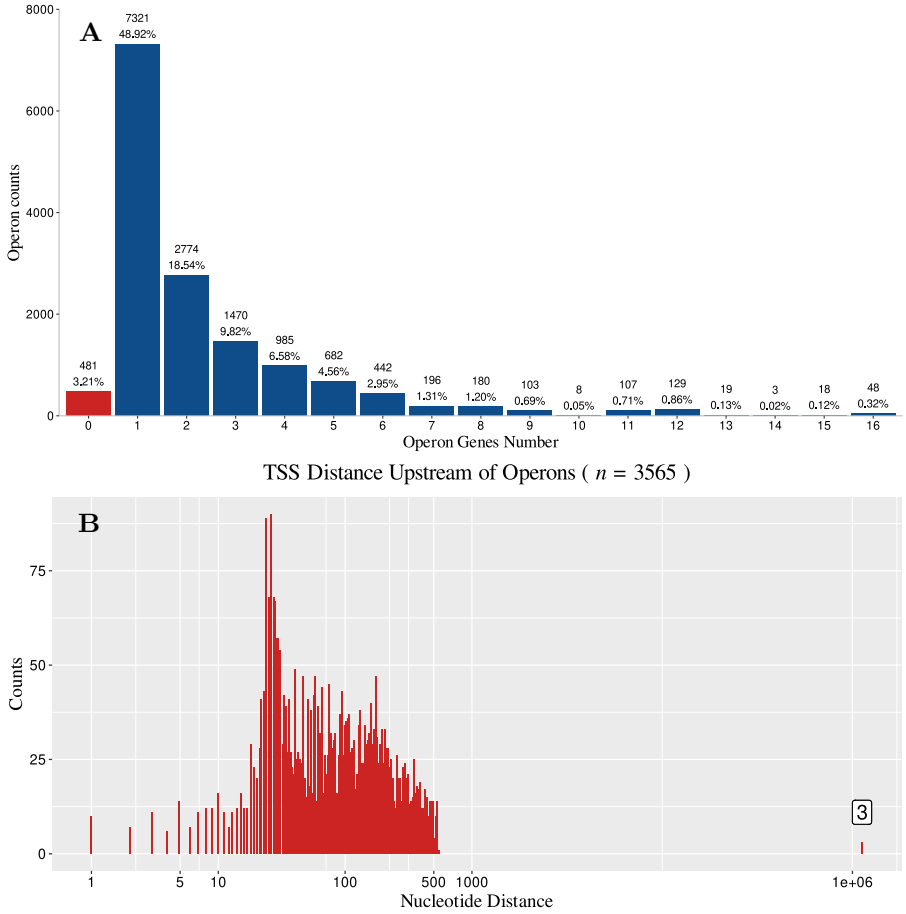
Λαμβάνοντας υπόψιν τα παραπάνω, γίνεται εντοπισμός των θέσεων των κωδικονίων `ATG, CTG, GTG & TTG` και διατηρούνται μόνο οι θέσεις που ταυτίζονται με το πρώτο κωδικόνιο του γονιδίου που έχει αντιστοιχηθεί ένα TSS. Σε δεύτερο επίπεδο ανάλυσης, τα υποψήφια ως Leaderless mRNA ελέγχονται, ώστε να διασφαλιστεί ότι δεν υπάρχουν άλλα γειτονικά TSS, τα οποία να είναι επίσης εξωτερικά και ομόρροπα, να αντιστοιχούν στο ίδιο γονίδιο και να απέχουν από τη θέση έναρξης του γονιδίου, απόσταση μεγαλύτερη από 10 βάσεις. Με αυτόν τον τρόπο, γίνεται έλεγχος εάν ένα γονίδιο δεν έχει TSS με αμετάφραστη περιοχή μεγαλύτερη των 10 βάσεων και μόνο τότε μπορεί να χαρακτηριστεί ως leaderless. Το αποτέλεσμα της συγκεκριμένης ανάλυσης είναι ο χαρακτηρισμός 36 (0.24% των συνολικών TSS) από τα αρχικά 115 TSS, ως leaderless.

2.2.7. Ανάλυση και Χαρακτηρισμος Οπερονιων

Τα γονίδια που εμπλέκονται σε ένα κοινό μονοπάτι ή έχουν ταυτόσημη βιολογική λειτουργία, είναι συχνά διατεταγμένα σε οπερόνια με έναν κοινό υποκινητή. Τα δεδομένα των TSS που ελήφθησαν με την ανάλυση των *Cappable-seq* δεδομένων, προσδιορίζουν αυτούς τους υποκινητές και έτσι μπορούν να ορίσουν την αρχή των οπερονίων.

Για να εξακριβωθεί εάν τα γονίδια και κατ΄ επέκταση τα TSS που αντιστοιχήθηκαν σε αυτά, ανήκουν σε κάποιο οπερόνιο, χρησιμοποιήθηκε συναινετικός σχολιασμός από τις βάσεις δεδομένων *RegulonDB*, και *ODB 4* (Okuda and Yoshizawa, 2010), ο οποίος συγκρίθηκε με τη βοήθεια της ℝ, με τα δεδομένα της παραπάνω ανάλυσης. Το αποτέλεσμα ήταν ο εντοπισμός 14.484 (96.01%) TSS *(Σχήμα 2.21A)*, τα οποία ανήκουν σε ένα οπερόνιο, ενώ 482 (3.2%) εμφανίζονται μεμονωμένα στο γονιδίωμα, χωρίς να περιέχονται σε κάποιο οπερόνιο (κόκκινη μπάρα). Από την ανάλυση εξαιρέθηκαν τα ορφανά TSS, τα οποία δεν φάνηκε να ανήκουν σε κάποιο σχολιασμένο, σύμφωνα με τα τρέχοντα δεδομένα, γονίδιο και κατ΄ επέκταση ούτε σε κάποιο οπερόνιο, αντιπροσωπεύοντας τα υπόλοιπα 119 TSS (0.79%) της *Cappable-seq*. Πιο αναλυτικά, η πλειοψηφία των TSS ανήκουν σε οπερόνια που περιέχουν ακριβώς 1 γονίδιο, με ποσοστό 48.9% και όσο αυξάνεται ο αριθμός των γονιδίων που περιέχονται σε αυτά, τόσο μειώνεται και το πλήθος αυτών των οπερονίων και των TSS που αντιστοιχούν σε αυτά. Επιπλέον, η απόσταση των ανοδικών και ομόρροπων TSS από τη θέση έναρξης των οπερονίων, κυμαίνεται για το 98% αυτών των TSS, από 20 έως 500 βάσεις *(Σχήμα 2.21B)*, ενώ η ανάλυση έδειξε ότι υπάρχουν 3 TSS με απόσταση μεγαλύτερη από 1.000.000 βάσεις από το αντίστοιχο οπερόνιό τους, γεγονός που είναι μη πραγματικό και βιολογικά μη επιτεύξιμο. Αυτό συνέβη λόγω της ύπαρξης μεταθετών αλληλουχιών γονιδίων, δηλαδή τρανσποζονίων (transposons), τα οποία αποτελούνται από ένα γονίδιο τρανσποζάσης, που κωδικοποιεί το ομώνυμο ένζυμο (Tase), πλαισιωμένο από δύο τελικές ανεστραμμένες επαναλήψεις. Το ένζυμο που καταλύει αυτή τη μεταφορά των αλληλουχιών, κόβει και συρράφει συγκεκριμένες αλληλουχίες από μία θέση του γονιδιώματος σε μία άλλη, οδηγώντας τα διάφορα κύτταρα στη γενετική διαφοροποίησή τους, χωρίς αυτό να είναι πάντα ένα αρνητικό γεγονός (Bucher *et al.*, 2012). Το γονίδιο που αντιστοιχούν αυτά τα 3 TSS, ονομάζεται *insI2* και ενώ στο σχολιασμό γονιδίων που λήφθηκε από τις βάσεις δεδομένων, εμφανίζεται σε δύο θέσεις στο γονιδίωμα με συντεταγμένες 279.931 – 280.104 και 1.469.358 – 1.470.509, στο σχολιασμό των οπερονίων, περιέχεται μόνο το δεύτερο αντίγραφο, οδηγώντας τα TSS

του πρώτου αντιγράφου στο λάθος χαρακτηρισμό και επισημαίνοντας την ανάγκη για τον επανασχολιασμό αυτών.

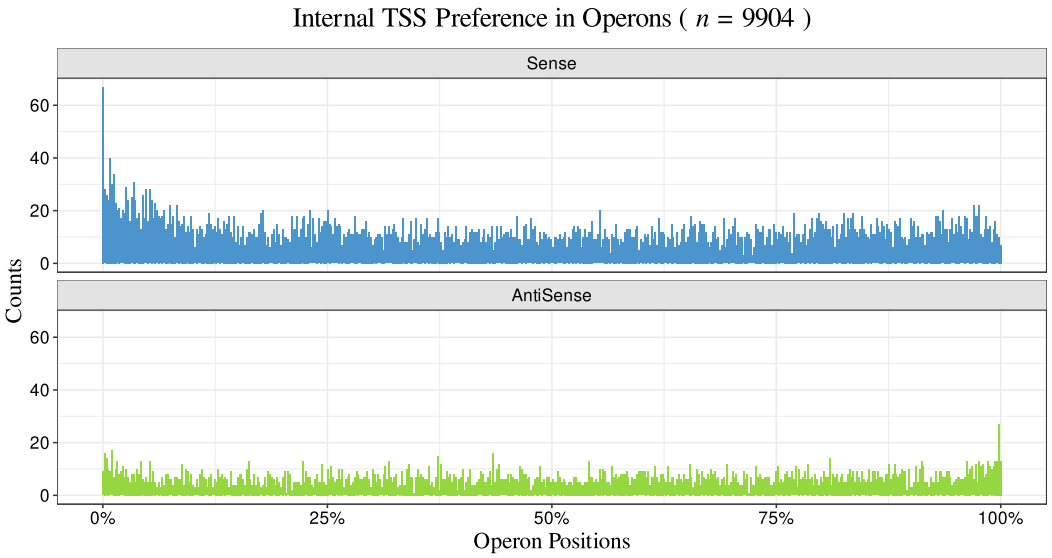


Σχήμα 2.21: *Συσχέτιση των TSS με τα σχολιασμένα οπερόνια της Escherichia Coli.* ***(A)*** *Αναπαράσταση του αριθμού των γονιδίων που περιέχει κάθε αντιστοιχούμενο οπερόνιο. Η κόκκινη μπάρα επισημαίνει όλα τα TSS που δεν ανήκουν σε κάποιο οπερόνιο.* ***(B)*** *Κατανομή των αποστάσεων των ανοδικών TSS από τη θέση έναρξης των αντιστοιχισμένων οπερονίων. Παρατηρείται ότι τα περισσότερα TSS έχουν απόσταση από 20 έως 500 βάσεις, ενώ υπάρχουν 3 TSS, των οποίων η απόσταση υπολογίστηκε λάθος, λόγω της ύπαρξης μεταθετών γονιδίων και εμφανίζουν απόσταση μεγαλύτερη από 1.000.000 βάσεις.*

Αναφορικά με τη θέση των TSS μέσα στα οπερόνια, εκτιμήθηκε ότι το 14.9% των ομόρροπων TSS που έχουν αντιστοιχηθεί και βρίσκονται εσωτερικά ενός οπερονίου, εμφανίζουν ως προτιμώμενη θέση να βρίσκονται το πολύ στο πρώτο 10% του συνολικού τους μήκους, από τη θέση έναρξης των οπερονίων. Το υπόλοιπο μήκος των οπερονίων, παραμένει περισσότερο ομοιόμορφα κατανεμημένο σχετικά με την πιθανότητα ύπαρξης TSS σε κάθε θέση τους, τόσο στα ομόρροπα, όσο και στα αντίρροπα TSS *(Σχήμα 2.22)*.

2.3. Αναλυση των TSS στις Μικροβιακες Κοινοτητες

Μία σημαντική πρόκληση που είχε να αντιμετωπίσει η βιοπληροφορική ανάλυση των δεδομένων αλληλούχισης από μικροβιακές κοινότητες, ήταν ο εντοπισμός της σύνθεσης του δείγματος. Πιο αναλυτικά, η ταυτοποίηση των βακτηρίων που ανήκουν σε ένα δεδομένο δείγμα, είναι μείζονος σημασίας για την μετέπειτα ανάλυση και το χειρισμό αυτών των δεδομένων. Αφού εντοπιστούν οι πιθανοί οργανισμοί με ακρίβεια γένους, τότε τα δεδομένα μπορούν να στοιχισθούν στο κάθε ένα γονιδίωμα αναφοράς και να εφαρμοστεί η παραπάνω ανάλυση του πρώτου μέρους της μεθοδολογίας, όπως και στην *Escherichia Coli*.



Σχήμα 2.22: *TSS εσωτερικά των οπερονίων. Κατανομή του αριθμού των TSS, τα οποία είναι ομόρροπα (πάνω) και αντίρροπα (κάτω), σε σχέση με τον προσανατολισμό και τη θέση τους στο οπερόνιο που ανήκουν, εκφρασμένη ως ποσοστό.*

2.3.1. Amplicon Sequence Variants και Operational Taxonomic Units

Για την επίτευξη του παραπάνω σκοπού, θα πρέπει αρχικά να ακολουθηθεί το πρωτόκολλο που ορίζει η αλληλούχιση του γονιδίου της 16S ριβοσωμικής υπομονάδας, ώστε με βάση τη συντήρηση και την αλληλουχία του, να ταυτοποιηθούν τα υποψήφια βακτηριακά γένη. Επίσης, τρέχουσες εξελίξεις αυτών των αναλύσεων, προτείνουν ότι η μεθοδολογία που θα ακολουθηθεί, ξεκινά έχοντας γνωστό το συγκεκριμένο γονίδιο για κάθε γένος, και συνεχίζει με τον υπολογισμό του αριθμού των μεταλλάξεων και τη σύγκριση αυτού, με τα αντίστοιχα γονίδια του δείγματος, ώστε να καθοριστούν ποια γένη περιέχονται. Η συγκεκριμένη μεθοδολογία είναι επιρρεπής σε λανθασμένες μεταλλάξεις, οι οποίες μπορούν να προκύψουν είτε από την αλληλούχιση, είτε από την εξειδίκευση του γονιδίου σε επίπεδο είδους (species). Άρα, θα πρέπει να χρησιμοποιηθεί μία μεθοδολογία, η οποία να κάνει λιγότερες αρνητικές παραδοχές στην ποσοτικοποίηση αυτών των μεταλλάξεων και να προσαρμόζεται καλύτερα στα δεδομένα. Τη λύση στο πρόβλημα αυτό, έχουν δώσει δύο νέες μεθοδολογίες, με ονόματα *Operational Taxonomic Units* (OTUs) και *Amplicon Sequence Variants* (ASVs), οι οποίες βασίζονται στα μαθηματικά και τη στατιστική. Πιο αναλυτικά, βασίζονται στη συσταδοποίηση (clustering) των δεδομένων, λαμβάνοντας υπόψιν πως οργανισμοί που βρίσκονται φυλογενετικά σε μικρή απόσταση, θα έχουν μεγαλύτερη συντήρηση στο γονίδιο 16S *r*RNA. Κατ΄ αυτό τον τρόπο, σφάλματα αλληλουχιών, θεωρούνται ως outliers και έχουν σχεδόν ασήμαντη συμβολή στη συναινετική αλληλουχία κάθε συστάδας.

Στη μεθοδολογία των OTUs, ως κατώφλι ομοιότητας θεωρείται συνήθως το 97% του αλληλουχικού συγκρινόμενου μήκους (Blaxter *et al.*, 2005). Ωστόσο, σύμφωνα με την βαθμονόμηση αυτού του μοντέλου, αυτό το κατώφλι δεν επαρκεί για τη σύγκριση πολύ κοντινών φυλογενετικά ειδών, άρα οδηγεί σε ορισμένες περιπτώσεις τη συσταδοποίηση διαφορετικών οργανισμών στην ίδια συστάδα. Αντιθέτως, ένα μεγαλύτερο κατώφλι μπορεί να μην αγνοήσει τα σφάλματα αλληλουχιών και η ακρίβεια στην αξιοπιστία της συσταδοποίησης να μειωθεί, με την εισαγωγή ψευδών αλληλουχιών ως νέα είδη (Kunin *et al.*, 2010). Το κυριότερο πρόβλημα στη διαχείριση αυτών των δεδομένων, παραμένει η μη εκ των προτέρων γνώση, έστω και περιορισμένη, της μικροχλωρίδας που μπορεί να περιέχεται στο δείγμα. Για αυτό το λόγο, η μέθοδος των OTUs μπορεί άλλοτε να δίνει μία ακριβή προσέγγιση των οργανισμών, εάν αυτοί έχουν ξεκάθαρη φυλογενετική απόσταση, ενώ μπορεί να δίνει μεροληπτικά και ανακριβή αποτελέσματα, όταν η απόσταση αυτή είναι μικρή (Edgar, 2017).

Σε αντίθεση με τα OTUs, η μεθοδολογία των ASVs προσδιορίζει ποιες ακολουθίες και πόσες φορές αυτές διαβάστηκαν, ώστε με βάση μία συνάρτηση σφάλματος, να καθοριστεί η πιθανότητα μια δεδομένη ανάγνωση, η οποία έχει μία δεδομένη συχνότητα, να μην οφείλεται σε σφάλμα ακολουθίας, δηλαδή σε τυχαία ομοιότητα. Έτσι, δημιουργείται ένα *p-value* για κάθε μία αλληλουχία και εφαρμόζοντας ένα κατώφλι ως διάστημα εμπιστοσύνης, μπορούν να διατηρηθούν μόνο οι στατιστικά σημαντικές αλληλουχίες. Αυτές μπορούν αργότερα να συγκριθούν με τα γονίδια στόχους, δηλαδή τα 16S *r*RNA και να εξακριβωθεί για το ποιοι οργανισμοί περιέχονται στο δείγμα, *in silico* (B. J. Callahan *et al.*, 2019).

Σύμφωνα με συγκρίσεις αυτών των δύο μεθοδολογιών, φαίνεται πως τα ASVs δίνουν τη δυνατότητα για έναν πιο ακριβή εντοπισμό των μικροβιακών ειδών, δηλαδή τον υπολογισμό του δείκτη της ποικιλομορφίας Άλφα (α-diversity). Επίσης, μειώνουν τα σφάλματα των μη έγκυρων αλληλουχιών, δίνοντας μία πιο αντιπροσωπευτική εικόνα, σχετικά με τη σύνθεση του δείγματος (Chiarello *et al.*, 2022). Η δυνατότητα για *de novo* χαρακτηρισμό των αλληλουχιών, είναι ακριβέστερη στα ASVs, αλλά και στις δύο μεθοδολογίες προσθέτει περισσότερη πολυπλοκότητα, λόγω του ότι οι μη στοιχισμένες αλληλουχίες, δηλαδή αυτές που δεν έχουν στοιχισθεί σε κάποια αλληλουχία αναφοράς στόχου, υφίστανται διαφορετικό σχολιασμό και έτσι αυξάνεται εκθετικά η αλγοριθμική πολυπλοκότητα του προβλήματος (Chiarello *et al.*, 2022).

Σύμφωνα με τα παραπάνω, για την ανάλυση του μεταγονιδιώματος σε αρχείο αλληλούχισης, χρησιμοποιήθηκε το πρωτόκολλο των ASVs και η ανάλυση έγινε στην $\mathbb{R}$, με το πακέτο *Dada2* (B. Callahan *et al.*, 2016; B. J. Callahan *et al.*, 2019), το οποίο παρέχει μία από τις πιο πολυχρησιμοποιημένες ροές εργασίας για αναλύσεις υψηλής ακρίβειας σε δεδομένα αλληλούχισης παραλλαγών αμπλικονίων (ASVs). Το συγκεκριμένο πακέτο περιλαμβάνει μία σειρά αναλύσεων του επίσης γνωστού εργαλείου *Qiime2* και η ροή εργασίας ξεκινά, όπως και στη βασική ανάλυση αλληλούχισης, με το φιλτράρισμα και την αποκοπή των εκκινητών των αναγνώσεων. Αφετέρου, τα αρχεία FastQ μετατρέπονται σε αρχεία Fasta, τα οποία περιέχουν μόνο τις αλληλουχίες των αναγνώσεων, απαλείφοντας κάθε άλλη πληροφορία, και εν συνεχεία γίνεται ένα dereplication των νέων δεδομένων, σύμφωνα με το οποίο διατηρούνται μόνο οι μοναδικές αναγνώσεις, οδηγώντας στην απόρριψη των πολλαπλών αντιγράφων. Η ροή συνεχίζεται με την αφαίρεση αλληλουχιών, γνωστές ως *chimeras*, οι οποίες είναι λανθασμένες αλληλουχίες που προέρχονται από την PCR, μέσω της σύνδεσης δύο ή περισσότερων αλληλουχιών που δεν έχουν ολοκληρώσει επιτυχώς την επέκτασή τους από τη DNA πολυμεράση. Ως αποτέλεσμα αυτού, ένας μερικώς επεκτεινόμενος κλώνος μπορεί να συνδεθεί στο εκμαγείο κάποιας άλλης αλληλουχίας, ο μερικώς επεκτεινόμενος κλώνος να δράσει ως εκκινητής και με επέκταση να σχηματιστεί μια χιμαιρική αλληλουχία. Το γεγονός αυτό οδηγεί στη δημιουργία θορύβου, εάν ληφθεί υπόψιν ότι αυτή η λανθασμένη αλληλουχία, μπορεί να αντιγραφεί στους επόμενους κύκλους, αυξάνοντας το σφάλμα. Οι χιμαιρικές αλληλουχίες, παρατηρούνται αρκετά σε μεικτά δείγματα αλληλούχισης και είναι σημαντική η αφαίρεσή τους, ώστε να μην επηρεάσουν τις αναγνώσεις που αντιστοιχούν σε γονίδια 16S *r*RNA (Haas *et al.*, 2011). Οι αρχικές αναγνώσεις ήταν ίσες με 30.172, ενώ μετά την αφαίρεση των χιμαιρικών, ανήλθαν στις 26.682, δηλαδή οι χιμαιρικές αντιστοιχούν σε ποσοστό περίπου ίσο με 11.56%.

2.3.2. Αναθεση Ταξινομησεων στο Δειγμα

Μετά την προετοιμασία και προ-επεξεργασία των δεδομένων της αλληλούχισης των μικροβιακών κοινοτήτων, θα πρέπει να γίνει ο εντοπισμός αυτών, δηλαδή ο υπολογισμός της α-ποικιλομορφίας, καθώς και ο υπολογισμός της β-ποικιλομορφίας, δηλαδή ο ακριβής ποσοτικός εντοπισμός αυτών των ειδών.

Η συγκεκριμένη εκτίμηση, γίνεται με την κατηγοριοποίηση των αναγνώσεων, με βάση τα μη ταιριάσματα αυτών, με ένα σετ που περιέχει τις αλληλουχίες των 16S *r*RNA γονιδίων. Για αυτό το σκοπό, στη βιβλιογραφία είναι αρκετά δημοφιλής μία συγκεκριμένη εκδοχή του απλού Μπεϋζιανού κατηγοριοποιητή *(Naive Bayes Classifier)* (Q. Wang *et al.*, 2007), η οποία συγκρίνει αυτές τις μεταλλάξεις με ένα σετ εκπαίδευσης που ελήφθη από τη βάση δεδομένων *Ribosomal Database Project* (RDP) (RDP v18 training set (Cole *et al.*, 2013)). Πιο συγκε-

κριμένα, ο χώρος χαρακτηριστικών αυτού του κατηγοριοποιητή, είναι συνήθως κάθε 8-μερής υποακολουθία των αναγνώσεων, υπό τον αλγόριθμο k-mers και το κάθε συνολικό πλήθος αυτών των 8-μερών λέξεων, χρησιμοποιείται για τον υπολογισμό της αναμενόμενης πιθανότητας να ανήκει σε γονίδιο της 16S ριβοσωμικής υπομονάδας και κατ΄ επέκταση της πιθανότητας να ανήκει σε κάποιο συγκεκριμένο βακτηριακό γένος. Με τον τρόπο αυτό, η συγκεκριμένη μεθοδολογία είναι αρκετά ακριβής στις προβλέψεις της, επειδή δε βασίζεται σε αυθαίρετες τιμές σκορ, όπως γίνεται στη στοίχιση ακολουθιών και έχει στέρεο μαθηματικό υπόβαθρο.

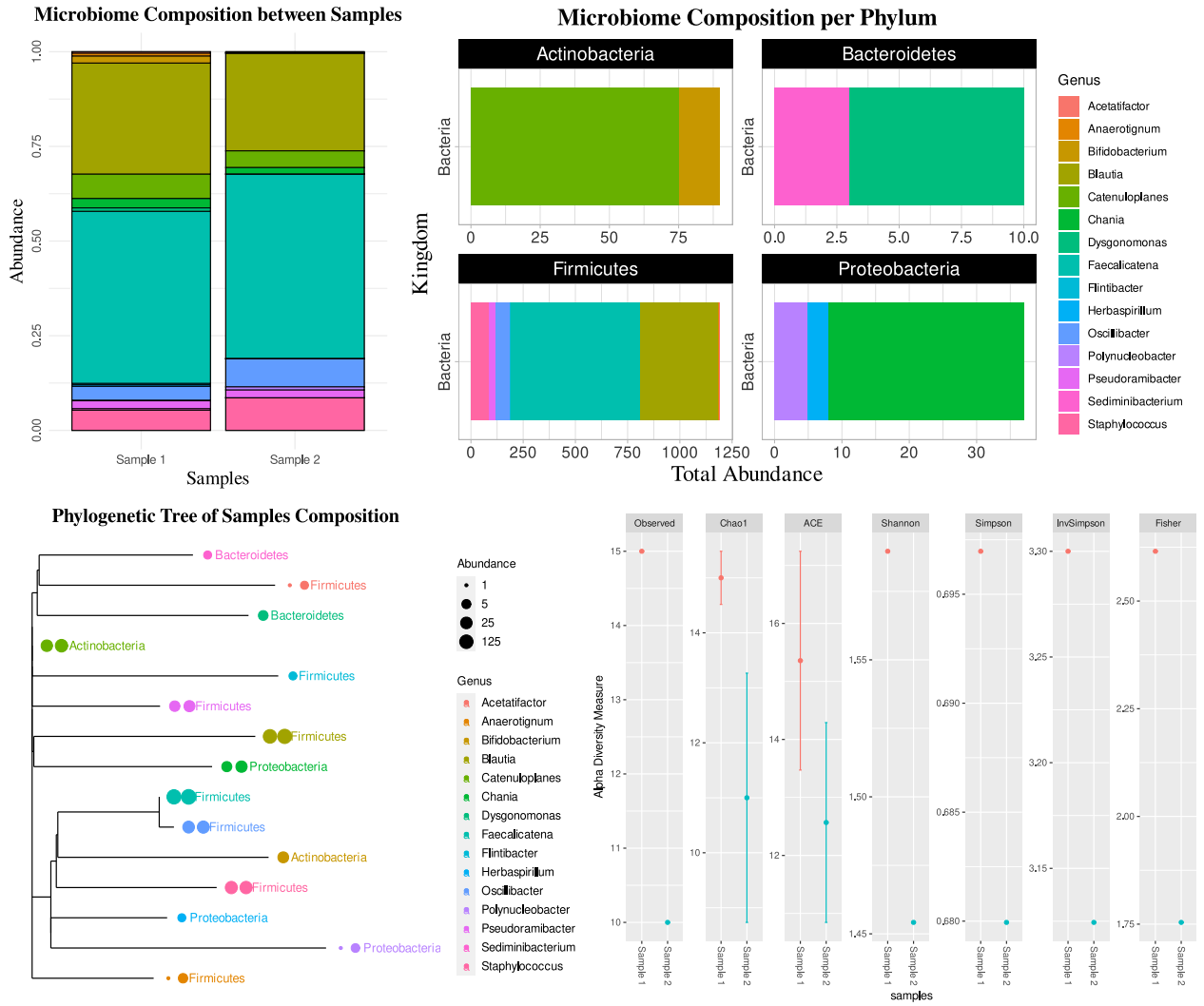
2.3.3. Εντοπισμος των Θεσεων Εναρξης της Μεταγραφης

Η παραπάνω ανάλυση εφαρμόστηκε σε δύο μικροβιακά αντίγραφα δειγμάτων και είχε ως αποτέλεσμα τον εντοπισμό 6 κύριων μικροβιακών ειδών, των *Escherichia Coli MG1655, Oscillibacter valericigenes, Staphylococcus aureus, Blautia marasmi, Faecalicatena contorta & [Clostridium] indolis DSM 755 (Σχήμα 2.23)*, των οποίων το πιο αντιπροσωπευτικό *(representative)* γονιδίωμα, χρησιμοποιήθηκε για τη στοίχιση των δεδομένων. Πιο αναλυτικά, βακτηριακά είδη που δεν είναι διαθέσιμο το γονιδίωμα αναφοράς τους, παρά μόνο πολλά και διαφορετικά *assemblies*, τα οποία εμφανίζουν μικρότερη ακρίβεια από τα γονιδιώματα αναφοράς, έχουν το λεγόμενο αντιπροσωπευτικό γονιδίωμα, που είναι εν ολίγοις το πιο αξιόπιστο γονιδίωμα από όλα τα *assemblies* με βάση ορισμένα κριτήρια καταλληλότητας, όπως ο αριθμός των ψευδών κωδικών περιοχών, η ύπαρξη πλασμιδίου και άλλα. Επιπλέον, στα βακτηριακά είδη που δεν έχουν ούτε αντιπροσωπευτικό γονιδίωμα, χρησιμοποιήθηκε ένα τυχαίο γονιδίωμα *(contig)* για τη στοίχιση, συνήθως κάποιο από αυτά με το μεγαλύτερο ακολουθιακό μήκος. Τα αρχεία των γονιδιωμάτων για κάθε είδος ξεχωριστά, δίνονται μαζί με την εργασία αυτή. Πριν τη στοίχιση, εφαρμόστηκε προ-επεξεργασία και φιλτράρισμα των δεδομένων, όπως και προηγουμένως και οι χαρτογραφημένες αναγνώσεις μετασχηματίστηκαν, διατηρώντας μόνο την πρώτη βάση από το 5′ άκρο τους ως υποψήφιο TSS. Τα δεδομένα αυτά σχολιάστηκαν για κάθε βακτήριο ξεχωριστά, με τη χρήση τις ίδιας βιοπληροφορικής ροής εργασίας του πρώτου μέρους της μεθοδολογίας.

2.4. Το Προβλημα Διαχειρισης των Dependencies

Η ταχύρρυθμη ανάπτυξη βιβλιοθηκών στις σύγχρονες γλώσσες προγραμματισμού, έχει οδηγήσει στην επίσης γρήγορη ανάπτυξη εφαρμογών, πάνω σε ήδη κατασκευασμένες προγραμματιστικές διεπαφές και πρότυπα. Ωστόσο, μέσα σε αυτές τις βιβλιοθήκες εμπεριέχονται άλλες, οι οποίες καλούνται εσωτερικά και θα πρέπει και αυτές να εγκατασταθούν και να εκτελεστούν, ώστε να μπορέσει να χρησιμοποιηθεί μία βιβλιοθήκη. Κατ΄ αυτό τον τρόπο, δημιουργείται ένα δίκτυο αλληλεπιδράσεων, όπου πολλά πακέτα εξαρτώνται από κάποια άλλα, τα οποία πολλές φορές μπορεί να απαιτείται να βρίσκονται σε μία συγκεκριμένη έκδοση, ώστε να καταστεί εφικτή η χρησιμοποίησή τους από τις εξαρτώμενες βιβλιοθήκες. Το γεγονός αυτό δημιουργεί το *πρόβλημα διαχείρισης των dependencies* ή *dependency hell*, που εντείνεται όταν η εκτέλεση αφορά πολλαπλά λειτουργικά συστήματα.

Τη λύση σε αυτά τα προβλήματα έρχονται να δώσουν ορισμένες τεχνολογίες πακετοποίησης *(containerization)*, κάθε μία εκ των οποίων προσπαθεί με την εκάστοτε τεχνική που εφαρμόζει, να συλλέξει τα απαιτούμενα dependencies και να προσφέρει ένα προϋλοποιημένο και αυτόνομο περιβάλλον για την εκτέλεση της τελικής εφαρμογής. Έτσι, μειώνεται η πιθανότητα σφάλματος, επειδή η τελική εικόνα εκτελείται σε δικό της λειτουργικό σύστημα και δεν εξαρτάται από το λειτουργικό του κάθε χρήστη. Μία πολύ γνωστή τεχνολογία, είναι αυτή του Docker, η οποία αναλύεται παρακάτω.



Σχήμα 2.23: *Μικροβιακές κοινότητες σε δείγμα ποντικιού, στο οποίο ακολουθήθηκε η τεχνική εμπλουτισμού αλληλούχισης Cappable-seq. Στα διαγράμματα αναπαριστάται η μικροβιακή σύνθεση για κάθε δείγμα, καθώς και για κάθε συνομοταξία ξεχωριστά. Στο φυλογενετικό δέντρο φαίνονται οι αποστάσεις των διάφορων ειδών που έχουν σχολιαστεί σε επίπεδο είδους, η συγκέντρωση που αυτά αντιπροσωπεύουν σε κάθε δείγμα (ως κύκλοι) και η συνομοταξία που ανήκουν. Το δέντρο κατασκευάστηκε με τη μέθοδο Neighbor joining και με τρόπο de novo, δηλαδή χωρίς κάποιο πρότυπο ή προηγούμενη γνώση, αποκλειστικά και μόνο από την φυλογενετική απόσταση των ταξινομικών ομάδων των συγκεκριμένων δειγμάτων. Στο γράφημα κάτω και δεξιά, φαίνονται ορισμένες στατιστικές πληροφορίες α-ποικιλομορφίας για τα δύο δείγματα ξεχωριστά.*

2.4.1. Εισαγωγη στο Docker

Το Docker (Merkel, 2014) είναι μία πλατφόρμα ανοιχτού κώδικα, η οποία υποστηρίζει την εικονικοποίηση (Virtualization) πολλών εικόνων (Images) στο επίπεδο ενός λειτουργικού συστήματος. Πιο αναλυτικά, το Docker χρησιμοποιεί κυρίως το βασικό πυρήνα του Linux, ώστε να εκτελεί συγκεκριμένες διεργασίες, που χαρακτηρίζουν την εκάστοτε εικόνα. Επίσης, ένα Docker Container μπορεί να εκτελεί ταυτόχρονα πολλές εικόνες και μπορούν να εκτελούνται πολλά, ανεξάρτητα μεταξύ τους Containers, από το ίδιο λειτουργικό σύστημα. Η χρησιμότητα του Docker έγκειται στο γεγονός, ότι μπορεί να λύσει το πρόβλημα διαχείρισης των Dependencies, και αυτό επιτυγχάνεται με πολλούς και διαφορετικούς τρόπους. Αρχικά, το Docker είναι διαθέσιμο για κάθε λειτουργικό σύστημα και μπορεί να εκτελέσει σε κάθε σύστημα Linux Containers. Ακόμα, δίνει τη δυνατότητα πρόσβασης σε έτοιμες προϋλοποιημένες εικόνες, οι οποίες μπορούν να ληφθούν και να χρησιμοποιηθούν δωρεάν, μέσω του Docker Hub. Οι συγκεκριμένες εικόνες έχουν καθορισμένη χρησιμότητα, που μπορεί να αφορά μία γλώσσα προγραμματισμού, μία βάση δεδομένων, ένα προγραμματιστικό πλαίσιο και πολλά άλλα, τα οποία βοηθούν τον

χρήστη στην κατασκευή του δικού του Container με μεγάλη ευκολία. Τέλος, ένας επίσης σημαντικός λόγος που βοηθά την εξάλειψη του προβλήματος των dependencies, είναι το αυτοματοποιημένο περιβάλλον διαχείρισης που παρέχει, γεγονός που βοηθά την ενσωμάτωση συγκεκριμένων εκδόσεων των χρησιμοποιούμενων εικόνων, με ιδιαίτερα απλό τρόπο, απλά παραμετροποιώντας κατάλληλα ένα αρχείο που ονομάζεται *Dockerfile*.

Η αρχιτεκτονική που έκανε το Docker ενα τόσο σημαντικό εγχείρημα, είναι παρόμοια με αυτή που χρησιμοποιείται από τις εικονικές μηχανές (Virtual Machines), αλλά έχει και πολλές διαφορές σε σύγκριση με αυτές. Μία εμφανής διαφορά οφείλεται στη διαχείριση των πόρων, με τις εικονικές μηχανές να υποστηρίζουν αυτόνομα λειτουργικά συστήματα, τα οποία με τη σειρά τους εκτελούνται πάνω σε συγκεκριμένο λειτουργικό σύστημα του κεντρικού διακομιστή (hypervisor). Αυτό το μοντέλο, απαιτεί τη χρήση ξεχωριστής μνήμης και επεξεργαστικής ισχύος, για κάθε μία εικονική μηχανή, επειδή κάθε μία από αυτές είναι μία ανεξάρτητη οντότητα. Αντιθέτως, τα Docker Container μοιράζονται τον ίδιο πυρήνα και δεν απαιτούνται επιπρόσθετα λειτουργικά συστήματα για κάθε Container. Έτσι, η απαίτηση για υπολογιστικούς πόρους είναι μικρότερη, εάν επίσης ληφθεί υπόψιν ότι απλά απαιτείται ένας πυρήνας λειτουργικού συστήματος και όχι πολλά λειτουργικά συστήματα στην ολότητά τους. Τέλος, η αρχιτεκτονική του Docker βασίζεται στην επικοινωνία του χρήστη με τα containers, αλλά και μεταξύ των διαφόρων containers, μέσω ενός REST – API του μοντέλου client – server.

2.4.2. Οργανωση των Docker Containers

Τα Docker images, είναι η στοιχειώδης μονάδα οργάνωσης ενός Docker Container. Συνήθως περιέχουν ορισμένες βιβλιοθήκες, βασικά προγραμματιστικά λογισμικά και τον πυρήνα του Linux. Για να μπορέσουν, όμως να αλληλεπιδράσουν οι διάφορες εικόνες μεταξύ τους και να οργανωθεί καλύτερα η συνολική εφαρμογή, πρέπει να χρησιμοποιηθεί ως αρχή λειτουργίας αυτών ένα ενιαίο μοντέλο αποθήκευσης, το οποίο να αποτελείται από όλα τα επιμέρους συστήματα αποθήκευσης των εικόνων. Λόγω αυτής της απαίτησης, το σύστημα αρχείων που χρησιμοποιείται δεν δίνει τη δυνατότητα αποθήκευσης δεδομένων μέσα στις εικόνες, αλλά μόνο την ανάγνωση δεδομένων από αυτές. Το συγκεκριμένο σύστημα αρχείων που χρησιμοποιείται σε κάθε εικόνα, βασίζεται στη λειτουργία του UnionFS με περισσότερο χρησιμοποιούμενα συστήματα, τα AUFS και OverlayFS. Τα κριτήρια επιλογής κάποιου από αυτά τα συστήματα αρχείων, ποικίλουν ανάλογα με το λειτουργικό σύστημα, το είδος της εφαρμογής, την οργάνωση των αρχείων που αυτή πρέπει να έχει και άλλα.

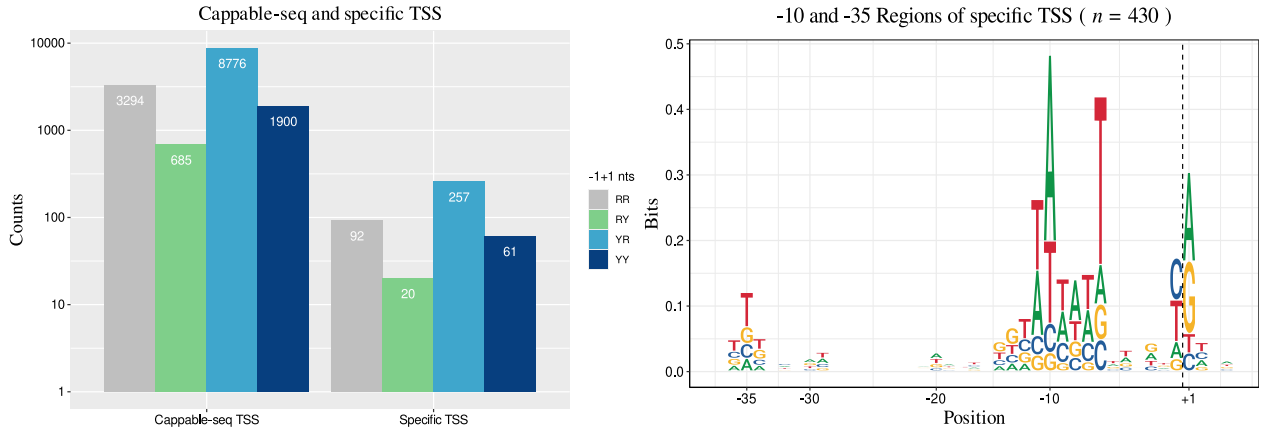
Με αυτόν τον τρόπο, στο επίπεδο ενός container γίνεται συνένωση όλων των επιμέρους συστημάτων αρχείων, μέσω ενός container layer, ο οποίος επιτυγχάνει τη διασύνδεση των images με το container. Παρόλο που δεν είναι εφικτή η άμεση αποθήκευση δεδομένων σε μία εικόνα, μέσα από ένα container είναι εφικτή η αποθήκευση, με τη δημιουργία νέων εικόνων που φέρουν τα νέα δεδομένα και έτσι αυτές οι εικόνες μπορούν ξανά να προσπελαστούν από ένα container.

Η ενορχήστρωση των διαφορετικών containers δηλαδή η διαχείριση μίας εφαρμογής που αποτελείται από πολλά containers, γίνεται συνήθως με τη βοήθεια του εργαλείου Docker Compose. Με αυτό το εργαλείο, ο χρήστης μπορεί εύκολα και με τη μικρότερη δυνατή παραμετροποίηση, να κατασκευάσει την εφαρμογή και να εκτελέσει ή να τροποποιήσει τα containers. Αυτή η παρέμβαση γίνεται με τη χρήση ενός συγκεντρωτικού αρχείου τύπου *yaml*, το οποίο καθορίζει ποια containers θα εκτελεστούν και με ποιες παραμέτρους, πως θα γίνει η επικοινωνία των containers, καθώς και η αποθήκευση των νέων δεδομένων σε αυτά. Το αρχείο *yaml* χωρίζεται σε δύο μέρη, όπου στο πρώτο μέρος υπάρχουν κάποιες βασικές πληροφορίες για τη διεύθυνση σύνδεσης και την έκδοση του Docker Compose που θα χρησιμοποιηθεί και το δεύτερο μέρος περιέχει την παραμετροποίηση για τη λειτουργικότητα της εφαρμογής.

Με βάση όλα τα παραπάνω, δημιουργήθηκε μία εφαρμογή Docker, η οποία περιέχει το πακέτο που υλοποιήθηκε για την ανάλυση των θέσεων έναρξης και λήξης της μεταγραφής. Η εφαρμογή στηρίχθηκε στη βασική εικόνα της γλώσσας R, η οποία είναι διαθέσιμη στο Docker hub και με τη διαδικασία του containerization, η εφαρμογή μαζί με όλα τα depedencies που απαιτεί, έχει πακετοποιηθεί και μπορεί να διανεμηθεί μέσω του Docker hub.

3 Αποτελεσματα - Συζητηση

Συνοψίζοντας τα συνολικά αποτελέσματα αυτής της εργασίας, έχει αποδειχθεί ότι η *Cappable-seq* είναι μία πρωτοποριακή τεχνική, η οποία μπορεί να απομονώσει το πρωταρχικό μεταγράφωμα των κυττάρων, με ισχυρό ενζυμικό τρόπο. Το γεγονός αυτό, οδήγησε στον εντοπισμό 15.085 TSS συνολικά στην *Escherichia Coli*, νούμερο το οποίο χρήζει ιδιαίτερης ερμηνείας, εάν ληφθεί υπόψιν ότι ακόμα και η πιο σχολιασμένη βάση δεδομένων της *Escherichia Coli*, δεν περιέχει περισσότερα από 5.500 πειραματικά αποδεδειγμένα TSS. Σύμφωνα με πειραματικά δεδομένα από το εργαστήριο του Dr. Morett, τα οποία έχουν κατατεθεί στη *RegulonDB*, μόνο 430 TSS είναι κοινά με τα δεδομένα της παρούσας ανάλυσης, σε εύρος ταύτισης ίσο με ±10 νουκλεοτιδικές βάσεις (specific TSS). Επίσης, ένα μεγάλο ποσοστό των συνολικών TSS της τάξης του 60%, αποδείχθηκε ότι είναι εσωτερικά, εγείροντας διάφορα ερωτήματα για την έναρξη της μεταγραφής σε συγκεκριμένα σημεία εντός των γονιδίων.



Σχήμα 3.1: *Ποσοτικοποίηση των κοινών TSS που εντοπίστηκαν με τη μέθοδο Cappable-seq, συναρτήσει των συνδυασμών πουρίνης ή πυριμιδίνης στις θέσεις -1 έως +1. Στο δεξιό διάγραμμα, φαίνονται τα συντηρημένα μοτίβα των υποκινητών, ανοδικά των specific TSS.*

Εσωτερικά TSS παρατηρούνται και στα σχολιασμένα οπερόνια, όπου η εσωτερική έναρξη της μεταγραφής παραμένει μία ελάχιστα μελετημένη διαδικασία, υποδηλώνοντας, όπως αποδείχθηκε *(Σχήμα 2.17)*, την ύπαρξη υποκινητών στα σημεία αυτά. Πιο αναλυτικά, τα εσωτερικά και αντίρροπα TSS φάνηκε να έχουν σημαντικά λιγότερη συντήρηση σε σχέση με τα ομόρροπα και εσωτερικά *(Σχήμα 2.19)* και η συντήρηση στα μοτίβα των υποκινητών τους, τα καθιστά να έχουν λειτουργικούς ρόλους, όπως το να κωδικοποιούν συγκεκριμένα ORFs. Έτσι, τα εσωτερικά TSS στις κωδικές περιοχές των γονιδίων, καταδεικνύουν μια πολύπλοκη μεταγραφική αρχιτεκτονική του γονιδιώματος της πολύ μελετημένης, αλλά ακόμα άγνωστης σε πολλές βιολογικές διαδικασίας, *Escherichia Coli*. Η παρουσία των εσωτερικών υποκινητών σε σχολιασμένα οπερόνια, είναι κάτι που χρήζει ιδιαίτερης αναφοράς, καθώς με αυτό τον τρόπο, μπορεί να μειωθεί η ποσότητα παραγωγής ενιαίων πρωτεϊνών μέσω των συμμεταγραφόμενων γονιδίων. Αυτό, διότι με τη διαφορική έκφραση γονιδίων εντός οπερονίων ή πιο περίπλοκων ρυθμιστικών διεργασιών, που σχετίζονται με την εναλλακτική δομή οπερονίου και συμμεταγραφική σχέση μεταξύ γονιδίων, μπορούν να εκφραστούν μεμονωμένα γονίδια (Shao *et al.*, 2014). Επιπρόσθετα, έρευνες σχετικά με τα ορφανά TSS έχουν αποδείξει ότι συσχετίζονται με γονίδια που κωδικοποιούν *nc*RNA (Cho *et al.*, 2009), γεγονός που ενισχύεται, εάν ληφθούν υπόψιν οι συντηρημένες περιοχές των φερόμενων υποκινητών τους *(Σχήμα 2.17)*.

Ανάλυση της συντήρησης των TSS και των υποκινητών για κάθε κατηγορία, έφερε στο προσκήνιο μία εμφανή συντήρηση σε αυτές τις περιοχές, γεγονός που υποδηλώνει ότι τόσο η επιλογή των στατιστικά σημαντικών TSS, όσο και ο σχολιασμός αυτών, έγινε με ορθολογικό τρόπο. Επιπλέον, παρόμοια επίπεδα συντήρησης προέκυψαν και στα ορφανά TSS, καταδεικνύοντας το γεγονός, ότι ίσως υπάρχουν γονίδια σε αυτές τις περιοχές, που δεν έχουν ακόμη σχολιαστεί, δείχνοντας έτσι τη δύναμη της μεταγραφωμικής ανάλυσης. Ακόμη, λόγω του διαφορετικού ενζυμικού τρόπου στον εμπλουτισμό του δείγματος, σε σχέση με τη γενική RNA-seq μεθοδολογία, έχει εξαχθεί ότι το 91,7% (4.767) των σχολιασμένων TSS από το σύνολο δεδομένων, δεν έχει προσδιοριστεί από την *Cappable-seq*,

αλλά αντιθέτως, μόνο το 2,85% των *Cappable-seq* δεδομένων, έχουν βρεθεί κοινά στα έως τώρα σχολιασμένα TSS. Ο λόγος για αυτή την απόκλιση των δεδομένων, μπορεί να εξηγηθεί από τις διαφορές στις συνθήκες ανάπτυξης των δειγμάτων, τον τρόπο απομόνωσης του πρωταρχικού RNA, το βάθος της αλληλούχισης, τη διακύμανση στις *in vitro* βιοχημικές διαδικασίες και στα αντιδραστήρια.

Όπως αναφέρθηκε, οι 5′ αμετάφραστες περιοχές των *m*RNA μπορούν να οδηγήσουν σε σημαντικά συμπεράσματα, σε σχέση με την αποκάλυψη της μεταφραστικής αποδοτικότητας σε αυτά. Από αυτή την ανάλυση, προκύπτει ότι σχεδόν τα μισά από τα συνολικά *m*RNA (51,82%), έχουν 5′ αμετάφραστη περιοχή μικρότερη από 100 βάσεις, ενώ το 27,11% μεγαλύτερη από 200 βάσεις. Επιπλέον, υπάρχει συντήρηση στην περιοχή που βρίσκεται περίπου 9 βάσεις ανοδικά του κωδικωνίου έναρξης, η οποία αντιπροσωπεύει την περιοχή Shine – Dalgarno και η απόσταση αυτή είναι η διάμεσος των αποστάσεων του μοτίβου `aaGGGAA`. Ακόμη, βρέθηκε συσχέτιση των αμετάφραστων περιοχών, με συγκεκριμένη λειτουργικότητα που επιτελούν τα αντίστοιχα γονίδια. Για παράδειγμα, γονίδια που κωδικοποιούν ρυθμιστικές πρωτεΐνες της μεταγραφής, βρέθηκε ότι σε ποσοστό 51,04%, έχουν περιοχή μικρότερη από 80 βάσεις, δηλαδή σχετικά μικρή 5′ αμετάφραστη περιοχή, ενώ γονίδια που κωδικοποιούν πρωτεΐνες και παράγοντες που συμβάλλουν στην κυτταρική διαίρεση, έχουν περιοχή μεγαλύτερη από 100 βάσεις (μέσος όρος οι 213 βάσεις), σε ποσοστό 62,07%. Στο ίδιο μήκος κύματος κυμαίνονται και τα γονίδια που κωδικοποιούν νουκλεοσιδάσες ή νουκλεοτιδάσες, με το 61,11% αυτών να έχουν αμετάφραστη περιοχή μεγαλύτερη των 110 βάσεων (μέσος όρος οι 350 βάσεις). Έτσι, με την ανάλυση αυτών των περιοχών σε επίπεδο ριβοδιακοπτών, δύνανται να αποκαλυφθούν μεταγραφικοί μηχανισμοί για κάθε γονίδιο ή οπερόνιο, καθώς και να ταυτοποιηθεί εάν επιδρούν *cis*-μεταγραφικοί ριβοδιακόπτες ή *trans*-acting *s*RNA.

Αναφορικά με τα *lm*RNA, χαρτογράφηση των TSS στις θέσεις έναρξης των γονιδίων, έχει οδηγήσει στον εντοπισμό 115 *m*RNA, τα οποία μπορούν να χαρακτηριστούν ως υποψήφια *lm*RNA, δηλαδή περιέχουν 5′ αμετάφραστη περιοχή, με μήκος μικρότερο από 10 βάσεις. Ωστόσο, για την περαιτέρω διασφάλιση ότι οι θέσεις των TSS δεν είναι τυχαίες και δεν εμφανίζουν γειτνίαση με άλλα ανοδικά TSS που αντιστοιχούν στο ίδιο γονίδιο, σε απόσταση μεγαλύτερη από 10 βάσεις από τη θέση έναρξης της μετάφρασης του γονιδίου, κρίνεται απαραίτητη η ταυτοποίησή τους σε δεύτερο χρόνο. Με την τελευταία ανάλυση, προέκυψαν μόλις 36 *lm*RNA, στα οποία εξετάζοντας τα γονίδια που αντιστοιχούν, καθώς και το βασικό λειτουργικό σχολιασμό αυτών, έχουν προκύψει ορισμένα σημαντικά συμπεράσματα.

Ένα από αυτά, έρχεται να ενισχύσει τα ήδη υπάρχοντα ευρήματα της βιβλιογραφίας, ότι συνήθως τα *lm*RNA αντιστοιχούν σε γονίδια που κωδικοποιούν ρυθμιστές μεταγραφής κατά τη διάρκεια της πρόφασης (Qin & e14 prophage) του κυτταρικού κύκλου, ή ορισμένους άλλους ρυθμιστές γονιδίων (Romero *et al.*, 2014). Βρέθηκε λοιπόν, ότι τα γονίδια με κωδική ονομασία *ybcM, rcdA, dicA* και *yfeR*, σχετίζονται με αυτές τις λειτουργίες και παράγουν RNA χωρίς οδηγό. Επίσης, η συγκεκριμένη ανάλυση έφερε στο προσκήνιο και 2 γονίδια που κωδικοποιούν πρωτεΐνες για την 50S ριβοσωμική υπομονάδα, τα *rplE* και *rplV*, καθώς και κάποια γονίδια που κωδικοποιούν φωσφατάσες, όπως τα *nudJ* και *ybjG*, οπως ακριβώς έχει αναφερθεί στη βιβλιογραφία. Τα τελευταία, δηλαδή τα προϊόντα των γονιδίων που κωδικοποιούν φωσφατάσες, ανήκουν στην κατηγορία των ενζύμων που αποικοδομούν RNA και έχει επίσης βρεθεί ότι μερικά από αυτά τα ένζυμα, κωδικοποιούνται από Leaderless *m*RNA, πιθανώς για την εξασφάλιση αποτελεσματικής μετάφρασης σε διάφορες συνθήκες, όπως κατά τη διάρκεια στρες των κυττάρων (Romero *et al.*, 2014). Με αυτό τον τρόπο, ενισχύονται οι γνώσεις της μεταγραφικής ρύθμισης στην *Escherichia Coli*, η οποία παρόλο που είναι το πιο καλά σχολιασμένο βακτήριο έως τώρα, εξακολουθεί να προσφέρει νέους ορίζοντες βιολογικής γνώσης, οι οποίοι μπορούν να γενικευτούν και σε άλλα βακτηριακά είδη, ακόμα και σε ευκαρυωτικούς οργανισμούς.

Είναι επίσης κατανοητό, ότι ορισμένα *Small Open Reading Frames* είναι συσχετισμένα με την κωδικοποίηση *s*RNA, τα οποία όπως αναφέρθηκε δύνανται να έχουν ρυθμιστικό ρόλο και να κωδικοποιούνται από περιοχές που προηγούνται (ή έπονται) των σχολιασμένων γονιδίων. Ανάλυση των TSS που αντιστοιχούν σε *s*RNA, οδήγησε στον εντοπισμό 51 γνωστών *s*RNA, με *trans* ρυθμιστική δράση, καθώς και 21 *s*RNA, με *cis* ρυθμιστική δράση (συμπληρωματικά αρχεία), δηλαδή τα τελευταία κωδικοποιούνται ανοδικά (ή καθοδικά) των γειτονικών

γονιδίων, στον ίδιο κλώνο και ρυθμίζουν την έκφραση αποκλειστικά αυτών των γονιδίων. Για την ταυτοποίηση και κατηγοριοποίηση αυτών των περιοχών, λήφθηκαν οι συγκεκριμένες αλληλουχίες καθοδικά των TSS και αναζητήθηκε η πιθανή λειτουργία τους με τη χρήση του BLAST. Από τα δεδομένα που προέκυψαν, διατηρήθηκαν οι πιθανές ακολουθίες, οι οποίες ρυθμίζουν την έκφραση κάποιων γονιδίων και ο συγκεκριμένος σχολιασμός συνδυάστηκε με δεδομένα από τη βάση *EcoCyc*, παρουσιάζοντας μεγάλη επικάλυψη. Από αυτά τα *s*RNA, 4 βρέθηκαν να σχετίζονται με έκφραση υπό συνθήκες stress με *cis* ρυθμιστικό ρόλο και 5 με *trans* ρυθμιστικό ρόλο. Ωστόσο, δεν μπόρεσε να προγνωσθεί η ύπαρξη μη σχολιασμένων έως τώρα *s*RNA, πιθανόν λόγω της φύσης της τεχνικής εμπλουτισμού *Cappable-seq* του δείγματος, σε σχέση με άλλες τεχνικές ολικής RNA αλληλούχισης. Η βασική ανάλυση σχολιασμού έγινε με τη χρήση της ℝ και η επιπρόσθετη ταυτοποίηση των *s*RNA, έγινε χειροκίνητα και με βάση τη βιβλιογραφία.

Σχετικά με τα μικροβιακά δείγματα του δεύτερου μέρους της μεθοδολογίας, ανάλυση των 6 αρχείων αλληλούχισης (ένα για κάθε βακτήριο) που προέκυψαν, έφερε στο προσκήνιο συνολικά 12.530 TSS, εκ των οποίων τα 2.015 αντιστοιχούν στο βακτήριο *Blautia marasmi*, τα 1.580 στο *Oscillibacter valericigenes*, τα 3.870 στην *Escherichia Coli MG1655*, τα 1.250 στο *Faecalicatena contorta*, τα 2.807 στο *indolis DSM 755* και τα υπόλοιπα 1.008 στο *Staphylococcus aureus*. Η ομαδοποίηση των δεδομένων έγινε χρησιμοποιώντας ως τιμή ομαδοποίησης το 10, αντί του 5 που χρησιμοποιήθηκε παραπάνω, ώστε να εξαλειφθούν τα πολύ κοντινά TSS με πιο αυστηρό τρόπο, καθώς πλέον δεν υπάρχει δείγμα ελέγχου για τη σύγκριση και τη στατιστική συμπερασματολογία των δεδομένων. Τα συγκεκριμένα TSS σχολιάστηκαν για κάθε είδος ξεχωριστά, σχετικά με τον προσανατολισμό και τη θέση και τον προσανατολισμό τους και αντιστοιχήθηκαν σε γονίδια, όπως ακριβώς έγινε και στο πρώτο μέρος της μεθοδολογίας.

Τέλος, όλες οι βιοπληροφορικές ροές εργασίας που χρησιμοποιήθηκαν για το χαρακτηρισμό των βακτηριακών γονιδιωμάτων στη μεθοδολογία της εργασίας, έχουν συμπεριληφθεί σε πακέτο της ℝ και μέσα από αυτό μπορεί να γίνει παρόμοια ανάλυση σε οποιοδήποτε βακτηριακό είδος που έχει αλληλουχηθεί με την τεχνική *Cappable-seq*, είτε σε μονοκαλλιέργεια, είτε σε ποικιλόμορφο βακτηριακό δείγμα. Επιπλέον, έχει αναπτυχθεί ℝ/Shiny εφαρμογή (Chang *et al.*, 2021), η οποία δύνανται να βοηθήσει σε αυτές τις αναλύσεις, μέσα από παραθυρικό και πιο διαδραστικό περιβάλλον.

4 | Επιλογος

Κλείνοντας, αναφορικά με το πρώτο μέρος αυτής της εργασίας, είναι γενικά εύκολο να εκτιμήσει κανείς τη συμβολή της αλληλούχισης του DNA στη βιολογική έρευνα. Σε αυτό το εγχείρημα, ήδη περίπου από τη δεκαετία του ΄80 έως σήμερα, προτάθηκε ένα μεγάλο πλήθος νέων πειραμάτων και μεθοδολογιών, για την εύρεση τεχνολογιών που προσδιορίζουν το DNA, απονεμήθηκαν σε πολλούς ερευνητές σπουδαίοι τίτλοι και διδακτορικά για τα ευρήματα των μελετών τους, ιδρύθηκαν πολλές εταιρείες (αλλά και έκλεισαν άλλες τόσες), δημοσιεύθηκαν χιλιάδες άρθρα και μέσα σε λίγα χρόνια, το συγκεκριμένο πρόβλημα, που έως πρότινος φάνταζε ασύλληπτων διαστάσεων και πολλών εμποδίων στη διαδικασία επίλυσής του, κατάφερε να γίνει μία διαδικασία ρουτίνας. Και όλα αυτά, σε μισό μόλις αιώνα αφότου υπήρξε το εύρημα ότι το DNA αποτελεί το γενετικό υλικό, από τους Alfred Hershey και Martha Chase, δηλαδή από όταν τέθηκε στο προσκήνιο το πρόβλημα του προσδιορισμού της αλληλουχίας αυτού.

Λαμβάνοντας υπόψιν τα παραπάνω, οι ερευνητές κατάφεραν, σε συνδυασμό με την πρόοδο της τεχνολογίας και της επιστήμης, να περάσουν από την αλληλούχιση λίγων μόνο βάσεων, στην αλληλούχιση ολόκληρων γονιδιωμάτων, να ελαχιστοποιήσουν το κόστος και να αναπτύξουν αυτόματα μηχανήματα αλληλούχισης, περνώντας από την εργαστηριακή έρευνα των wet labs, σε υπολογιστικά συστήματα και αλγορίθμους. Σήμερα, σε έναν κόσμο όπου τα μέσα υλοποίησης είναι διαθέσιμα και στο χαμηλότερο κόστος από ποτέ, έχουν προταθεί νέες, πολλά υποσχόμενες και βελτιωμένες μεθοδολογίες αλληλούχισης, όπως η αλληλούχιση τρίτης γενιάς, σε συνδυασμό με ένα τεράστιο πλήθος τεχνικών αλληλούχισης, οι οποίες εστιάζουν στον ειδικό εμπλουτισμό συγκεκριμένων γονιδιακών περιοχών, με γνώμονα συγκεκριμένα βιολογικά προβλήματα.

Όσον αφορά το δεύτερο μέρος, τα ευρήματα που έχουν προκύψει από την *Cappable-seq*, ερχονται να ενισχύσουν τα έως τώρα δεδομένα που υπάρχουν στις βάσεις δεδομένων των διάφορων βακτηρίων. Έτσι, έχοντας γνωστό και σχολιασμένο το συνολικό μεταγραφικό προφίλ των βακτηρίων, δύνανται να ενισχυθούν οι γνώσεις των επόμενων βιολογικών επιπέδων οργάνωσης, όπως είναι η πρωτεομική και η μεταβολομική. Επιπλέον, ο εντοπισμός των θέσεων έναρξης της μεταγραφής και στη συνέχεια, η συσχέτιση με τη γονιδιωματική αλληλουχία που έχει προέλθει από πειράματα RNA-seq, παρέχει μια πληθώρα πληροφοριών σχετικά με τη ρύθμιση της γονιδιακής έκφρασης, την εξαγωγή των μοτίβων Shine – Dalgarno και πρόσδεσης του παράγοντα Σίγμα (σ), καθώς και την οριοθέτηση των 5′ αμετάφραστων περιοχών.

Επομένως, ορισμένες από τις προεκτάσεις της συγκεκριμένης εργασίας, είναι η χρήση τεχνικών εντοπισμού ριβοσωμικού προφίλ (Ribosome profiling), ώστε να προκύψουν περαιτέρω πληροφορίες για τη θέση έναρξης της μετάφρασης, την κατανομή των ριβοσωμάτων στα mRNA και γενικότερα την ανακάλυψη της ρύθμισης στο μεταφραστικό προφίλ των βακτηρίων, καθώς και ο συνδυασμός των 5′ άκρων, με δεδομένα των 3′ άκρων των γονιδίων. Ο συγκεκριμένος συνδυασμός, ουσιαστικά συνδυάζει τα δεδομένα που προκύπτουν από την *Cappable-seq* και την *Term-seq*, με στόχο τη μαζική οριοθέτηση γονιδίων ή μεταγραφικών μονάδων, άρα και τον εντοπισμό νέων περιοχών και γονιδίων που δεν έχουν ακόμη σχολιαστεί. Έτσι, αυτά τα ευρήματα μπορούν αργότερα να ταυτοποιηθούν με εργαλεία όπως το BLAST, ωστε να επιτευχθεί ο λειτουργικός σχολιασμός με βάση την ομολογία αυτών των αλληλουχιών, σε σύγκριση με αυτές από άλλους οργανισμούς.

Το βιοχημικό πρωτόκολλο που χρησιμοποιείται στην *Cappable-seq*, όπως έχει αποδειχθεί, είναι ισχυρότερο στη σήμανση των 5′ άκρων σε σχέση με το προγενέστερο πρωτόκολλο της dRNA-seq και η μέθοδος εμφανίζει υψηλή συσχέτιση μεταξύ των διάφορων βιολογικών αντιγράφων *(Σχήμα Α΄.2)*, συνεπώς και σταθερή επαναληψιμότητα, υποδεικνύοντας ότι η συγκεκριμένη τεχνική μπορεί να γενικευτεί σε κάθε βακτηριακό είδος. Αυτό είναι και το ισχυρότερο πλεονέκτημα της μεθόδου, εάν ληφθεί υπόψιν πως ένα μεγάλο ποσοστό γονιδιωμάτων πολλών βακτηριακών ειδών, τόσο παθογόνων, όσο και ωφέλιμων για τους οργανισμούς, δεν έχει ακόμη χαρακτηρισθεί πλήρως, ώστε να ερευνηθεί το πρωτεομικό και μεταβολομικό προφίλ τους, άρα και η συσχέτισή τους με την ανάπτυξη ασθενειών, να αναπτυχθεί η έρευνα αντιβιοτικών και πολλές ακόμη προεκτάσεις αυτού του γνωστικού πεδίου.

5 | Διαθεσιμοτητα Δεδομενων

Τα αρχεία αλληλούχισης *Cappable-seq* που χρησιμοποιήθηκαν σε αυτή την εργασία, είναι ελεύθερα διαθέσιμα στη βάση δεδομένων *European Nucleotide Archive*, με αριθμό προσπέλασης PRJEB9717. Τα συμπληρωματικά αρχεία μπορούν να ληφθούν από αυτό το σύνδεσμο και το πακέτο που αναπτύχθηκε είναι διαθέσιμο στο σύνδεσμο `https://github.com/daniilidisK/TSSextractR`. Τέλος, το στιγμιότυπο Docker που κατασκευάστηκε βάσει του παραπάνω πακέτου στην $\mathbb{R}$, μπορεί να βρεθεί μέσα από την επίσημη σελίδα στιγμιοτύπων του Docker, στον ακόλουθο σύνδεσμο: `https://hub.docker.com/r/58924855/tssextractr`.

Βιβλιογραφια

1. Lee Ki-Hyun, Kim Mincheol adnd, Yoon Seok-Whan, Kim Bong-Soo, Chun Jongsik, and Yi Hana (2013). «Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era». In: *Genomics Inform* 11.3, pp. 102–113. doi: 10.5808/GI.2013.11.3.102. url: http://genominfo.org/journal/view.php?number=44.
2. Huttenhower, Curtis, Dirk Gevers, Rob Knight, *et al.* (June 2012). «The Human Microbiome Project (HMP) Consortium. Structure, function and diversity of the healthy human microbiome. Nature 486: 207-214». In: *Nature* 486, pp. 207–214. doi: 10.1038/nature11234.
3. Aagaard, Kjersti, Joseph Petrosino, Wendy Keitel, *et al.* (2013). «The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters». In: *The FASEB Journal* 27.3, pp. 1012–1022. doi: https://doi.org/10.1096/fj.12-220806. url: https://faseb.onlinelibrary.wiley.com/doi/abs/10.1096/fj.12-220806.
4. Peterson, Jane, Susan Garges, Maria Giovanni, *et al.* (Oct. 2009). «The NIH human microbiome project». In: *Genome research* 19, pp. 2317–23. doi: 10.1101/gr.096651.109.
5. Rolfe, Matthew D., Christopher J. Rice, Sacha Lucchini, *et al.* (2012). «Lag Phase Is a Distinct Growth Phase That Prepares Bacteria for Exponential Growth and Involves Transient Metal Accumulation». In: *Journal of Bacteriology* 194.3, pp. 686–701. doi: 10.1128/JB.06112-11. url: https://journals.asm.org/doi/abs/10.1128/JB.06112-11.
6. Moeller, Andrew H., Yingying Li, Eitel Mpoudi Ngole, *et al.* (2014). «Rapid changes in the gut microbiome during human evolution». In: *Proceedings of the National Academy of Sciences* 111.46, pp. 16431–16435. doi: 10.1073/pnas.1419136111. url: https://www.pnas.org/doi/abs/10.1073/pnas.1419136111.
7. Muegge, Brian D., Justin Kuczynski, Dan Knights, *et al.* (2011). «Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans». In: *Science* 332.6032, pp. 970–974. doi: 10.1126/science.1198719. url: https://www.science.org/doi/abs/10.1126/science.1198719.
8. Asnicar, Francesco, Serena Manara, Moreno Zolfo, *et al.* (2017). «Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling». In: *mSystems* 2.1, e00164–16. doi: 10.1128/mSystems.00164-16. url: https://journals.asm.org/doi/abs/10.1128/mSystems.00164-16.
9. Tung, Jenny, Luis B Barreiro, Michael B Burns, *et al.* (2015). «Social networks predict gut microbiome composition in wild baboons». In: *eLife* 4. Ed. by Eric Alm, e05224. issn: 2050-084X. doi: 10.7554/eLife.05224. url: https://doi.org/10.7554/eLife.05224.
10. Song, Se Jin, Christian Lauber, Elizabeth K Costello, *et al.* (2013). «Cohabiting family members share microbiota with one another and with their dogs». In: *eLife* 2. Ed. by Detlef Weigel, e00458. issn: 2050-084X. doi: 10.7554/eLife.00458. url: https://doi.org/10.7554/eLife.00458.
11. Sender, Ron, Shai Fuchs, and Ron Milo (Aug. 2016). «Revised Estimates for the Number of Human and Bacteria Cells in the Body». In: *PLOS Biology* 14.8, pp. 1–14. doi: 10.1371/journal.pbio.1002533. url: https://doi.org/10.1371/journal.pbio.1002533.
12. Morgan, Xochitl C., Nicola Segata, and Curtis Huttenhower (2013). «Biodiversity and functional genomics in the human microbiome». In: *Trends in Genetics* 29.1, pp. 51–58. issn: 0168-9525. doi: https://doi.org/10.1016/j.tig.2012.09.005. url: https://www.sciencedirect.com/science/article/pii/S016895251200145X.
13. Grice, Elizabeth A. and Julia A. Segre (2012). «The Human Microbiome: Our Second Genome». In: *Annual Review of Genomics and Human Genetics* 13.1. PMID: 22703178, pp. 151–170. doi: 10.1146/annurev-genom-090711-163814. url: https://doi.org/10.1146/annurev-genom-090711-163814.
14. Madigan, Michael T., John M. Martinko, and Thomas D. Brock (2006). *Brock Biology of Microorganisms*. Pearson/Prentice Hall.
15. Vollmer, Waldemar, Didier Blanot, and Miguel A. De Pedro (Feb. 2008). «Peptidoglycan structure and architecture». In: *FEMS Microbiology Reviews* 32.2, pp. 149–167. issn: 0168-6445. doi: 10.1111/j.1574-6976.2007.00094.x. eprint: https://academic.oup.com/femsre/article-pdf/32/2/149/8431303/32-2-149.pdf. url: https://doi.org/10.1111/j.1574-6976.2007.00094.x.

16. Woese, C R, O Kandler, and M L Wheelis (1990). «Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.» In: *Proceedings of the National Academy of Sciences* 87.12, pp. 4576–4579. doi: 10.1073/pnas.87.12.4576. url: https://www.pnas.org/doi/abs/10.1073/pnas.87.12.4576.
17. Lloyd-Price, Jason, Galeb abu-ali, and Curtis Huttenhower (Apr. 2016). «The healthy human microbiome». In: *Genome Medicine* 8. doi: 10.1186/s13073-016-0307-y.
18. Walter, Jens and Ruth Ley (2011). «The Human Gut Microbiome: Ecology and Recent Evolutionary Changes». In: *Annual Review of Microbiology* 65.1. PMID: 21682646, pp. 411–429. doi: 10.1146/annurev-micro-090110-102830. url: https://doi.org/10.1146/annurev-micro-090110-102830.
19. Bik, Elisabeth M., Paul B. Eckburg, Steven R. Gill, *et al.* (2006). «Molecular analysis of the bacterial microbiota in the human stomach». In: *Proceedings of the National Academy of Sciences* 103.3, pp. 732–737. doi: 10.1073/pnas.0506655103. url: https://www.pnas.org/doi/abs/10.1073/pnas.0506655103.
20. Woese, Carl R. and George E. Fox (1977). «Phylogenetic structure of the prokaryotic domain: The primary kingdoms». In: *Proceedings of the National Academy of Sciences* 74.11, pp. 5088–5090. doi: 10.1073/pnas.74.11.5088. url: https://www.pnas.org/doi/abs/10.1073/pnas.74.11.5088.
21. Kolbert, Christopher P and David H Persing (1999). «Ribosomal DNA sequencing as a tool for identification of bacterial pathogens». In: *Current Opinion in Microbiology* 2.3, pp. 299–305. issn: 1369-5274. doi: https://doi.org/10.1016/S1369-5274(99)80052-6. url: https://www.sciencedirect.com/science/article/pii/S1369527499800526.
22. Eren, A. Murat, Loïs Maignien, Woo Jun Sul, Leslie G. Murphy, Sharon L. Grim, Hilary G. Morrison, and Mitchell L. Sogin (2013). «Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data». In: *Methods in Ecology and Evolution* 4.12, pp. 1111–1119. doi: https://doi.org/10.1111/2041-210X.12114. url: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12114.
23. Shapira, Michael (2016). «Gut Microbiotas and Host Evolution: Scaling Up Symbiosis». In: *Trends in Ecology & Evolution* 31.7, pp. 539–549. issn: 0169-5347. doi: https://doi.org/10.1016/j.tree.2016.03.006. url: https://www.sciencedirect.com/science/article/pii/S0169534716000859.
24. Jack, Gilbert, Martin Blaser, J Caporaso, Janet Jansson, Susan Lynch, and Rob Knight (Apr. 2018). «Current understanding of the human microbiome». In: *Nature Medicine* 24, pp. 392–400. doi: 10.1038/nm.4517.
25. Gregory, Ann C., Olivier Zablocki, Allison Howell, Benjamin Bolduc, and Matthew B. Sullivan (2019). «The human gut virome database». In: *bioRxiv*. doi: 10.1101/655910. eprint: https://www.biorxiv.org/content/early/2019/07/02/655910.full.pdf. url: https://www.biorxiv.org/content/early/2019/07/02/655910.
26. Manrique, Pilar, Benjamin Bolduc, Seth T. Walk, John van der Oost, Willem M. de Vos, and Mark J. Young (2016). «Healthy human gut phageome». In: *Proceedings of the National Academy of Sciences* 113.37, pp. 10400–10405. doi: 10.1073/pnas.1601060113. url: https://www.pnas.org/doi/abs/10.1073/pnas.1601060113.
27. Fiers, William D, Iris H Gao, and Iliyan D Iliev (2019). «Gut mycobiota under scrutiny: fungal symbionts or environmental transients?» In: *Current Opinion in Microbiology* 50. Microbiota, pp. 79–86. issn: 1369-5274. doi: https://doi.org/10.1016/j.mib.2019.09.010. url: https://www.sciencedirect.com/science/article/pii/S1369527419300542.
28. Eckburg, Paul B., Elisabeth M. Bik, Charles N. Bernstein, *et al.* (2005). «Diversity of the Human Intestinal Microbial Flora». In: *Science* 308.5728, pp. 1635–1638. doi: 10.1126/science.1110591. url: https://www.science.org/doi/abs/10.1126/science.1110591.
29. Scanlan, Pauline D and Julian R Marchesi (2008). «Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces». In: *The ISME Journal* 2.12, pp. 1183–1193. issn: 1751-7362. doi: 10.1038/ismej.2008.76.
30. Noverr, Mairi C. and Gary B. Huffnagle (2004). «Does the microbiota regulate immune responses outside the gut?» In: *Trends in Microbiology* 12.12, pp. 562–568. issn: 0966-842X. doi: https://doi.org/10.1016/j.tim.2004.10.008. url: https://www.sciencedirect.com/science/article/pii/S0966842X04002409.

31. Ley, Ruth E., Daniel A. Peterson, and Jeffrey I. Gordon (2006). «Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine». In: *Cell* 124.4, pp. 837–848. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2006.02.017. URL: https://www.sciencedirect.com/science/article/pii/S0092867406001929.
32. Martín, R., S. Miquel, J. Ulmer, P. Langella, and L.G. Bermúdez-Humarán (2014). «Gut ecosystem: how microbes help us». In: *Beneficial Microbes* 5.3, pp. 219–233. DOI: 10.3920/BM2013.0057. URL: https://doi.org/10.3920/BM2013.0057.
33. Gill, Steven R., Mihai Pop, Robert T. DeBoy, *et al.* (2006). «Metagenomic Analysis of the Human Distal Gut Microbiome». In: *Science* 312.5778, pp. 1355–1359. DOI: 10.1126/science.1124234. URL: https://www.science.org/doi/abs/10.1126/science.1124234.
34. Duncan, Sylvia H, Georgina L Hold, Adela Barcenilla, Colin S Stewart, and Harry J Flint (2002). «Roseburia intestinalis sp. nov., a novel saccharolytic, butyrate-producing bacterium from human faeces.» In: *International Journal of Systematic and Evolutionary Microbiology* 52.5, pp. 1615–1620. ISSN: 1466-5034. DOI: https://doi.org/10.1099/00207713-52-5-1615. URL: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-52-5-1615.
35. Wong, Julia Wai Ming, Russell J. de Souza, Cyril W. C. Kendall, Azadeh Emam, and David J A Jenkins (2006). «Colonic Health: Fermentation and Short Chain Fatty Acids». In: *Journal of Clinical Gastroenterology* 40, pp. 235–243.
36. Rosenthal, Mariana, Deborah Goldberg, Allison Aiello, Elaine Larson, and Betsy Foxman (Apr. 2011). «Skin microbiota: Microbial community structure and its potential association with health and disease». In: *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases* 11, pp. 839–48. DOI: 10.1016/j.meegid.2011.03.022.
37. Grice, Elizabeth and Julia Segre (Apr. 2011). «The skin microbiome». In: *Nature reviews. Microbiology* 9, pp. 244–53. DOI: 10.1038/nrmicro2537.
38. Lai, Yuping, Anna Cogen, Katherine Radek, *et al.* (May 2010). «Activation of TLR2 by a Small Molecule Produced by Staphylococcus epidermidis Increases Antimicrobial Defense against Bacterial Skin Infections». In: *The Journal of investigative dermatology* 130, pp. 2211–21. DOI: 10.1038/jid.2010.123.
39. Chu, Derrick, Jun Ma, Amanda Prince, Kathleen Antony, Maxim Seferovic, and Kjersti Aagaard (Jan. 2017). «Maturation of the Infant Microbiome Community Structure and Function Across Multiple Body Sites and in Relation to Mode of Delivery». In: *Nature Medicine* 23. DOI: 10.1038/nm.4272.
40. Zheng, Peng, Benhua Zeng, Meiling Liu, *et al.* (2019). «Correction for the Research Article: The gut microbiome from patients with schizophrenia modulates the glutamate-glutamine-GABA cycle and schizophrenia-relevant behaviors in mice». In: *Science Advances* 5.6, eaay2759. DOI: 10.1126/sciadv.aay2759. URL: https://www.science.org/doi/abs/10.1126/sciadv.aay2759.
41. Jie, Zhuye, Xia Huihua, Shi-Long Zhong, *et al.* (Oct. 2017). «The gut microbiome in atherosclerotic cardiovascular disease». In: *Nature Communications* 8. DOI: 10.1038/s41467-017-00900-1.
42. Karlsson, Fredrik, Valentina Tremaroli, Jens Nielsen, and Fredrik Bäckhed (Sept. 2013). «Assessing the Human Gut Microbiota in Metabolic Diseases». In: *Diabetes* 62.10, pp. 3341–3349. ISSN: 0012-1797. DOI: 10.2337/db13-0844. eprint: https://diabetesjournals.org/diabetes/article-pdf/62/10/3341/569702/3341.pdf. URL: https://doi.org/10.2337/db13-0844.
43. Nagpal, Ravinder, Hirokazu Tsuji, Takuya Takahashi, Koji Nomoto, Kazunari Kawashima, Satoru Nagata, and Yuichiro Yamashiro (2017). «Ontogenesis of the Gut Microbiota Composition in Healthy, Full-Term, Vaginally Born and Breast-Fed Infants over the First 3 Years of Life: A Quantitative Bird's-Eye View». In: *Frontiers in Microbiology* 8. ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.01388. URL: https://www.frontiersin.org/article/10.3389/fmicb.2017.01388.
44. Manco, Melania, Lorenza Putignani, and Gianfranco Bottazzo (Dec. 2010). «Gut Microbiota, Lipopolysaccharides, and Innate Immunity in the Pathogenesis of Obesity and Cardiovascular Risk». In: *Endocrine reviews* 31, pp. 817–44. DOI: 10.1210/er.2009-0030.

45. Raoult, D (Apr. 2008). «Obesity pandemics and the modification of digestive bacterial flora». In: *European journal of clinical microbiology & infectious diseases: official publication of the European Society of Clinical Microbiology* 27, pp. 631–4. doi: 10.1007/s10096-008-0490-x.
46. Duncan, Sylvia H., Alvaro Belenguer, Grietje Holtrop, Alexandra M. Johnstone, Harry J. Flint, and Gerald E. Lobley (2007). «Reduced Dietary Intake of Carbohydrates by Obese Subjects Results in Decreased Concentrations of Butyrate and Butyrate-Producing Bacteria in Feces». In: *Applied and Environmental Microbiology* 73.4, pp. 1073–1078. doi: 10.1128/AEM.02340-06. url: https://journals.asm.org/doi/abs/10.1128/AEM.02340-06.
47. Sekirov, Inna, Shannon L. Russell, L. Caetano M. Antunes, and B. Brett Finlay (2010). «Gut Microbiota in Health and Disease». In: *Physiological Reviews* 90.3. PMID: 20664075, pp. 859–904. doi: 10.1152/physrev.00045.2009. url: https://doi.org/10.1152/physrev.00045.2009.
48. Tsolis, Renée M, Glenn M Young, Jay V Solnick, and Andreas J Bäumler (2008). «From bench to bedside: stealth of enteroinvasive pathogens». In: *Nature Reviews Microbiology* 6.12, pp. 883–892.
49. Foley, S. L. and A. M. Lynne (Apr. 2008). «Food animal-associated Salmonella challenges: Pathogenicity and antimicrobial resistance». In: *Journal of Animal Science* 86.suppl_14, E173–E187. issn: 0021-8812. doi: 10.2527/jas.2007-0447. eprint: https://academic.oup.com/jas/article-pdf/86/suppl_14/E173/23673709/e173.pdf. url: https://doi.org/10.2527/jas.2007-0447.
50. Howard, Zoe R., Corliss A. O'Bryan, Philip G. Crandall, and Steven C. Ricke (2012). «Salmonella Enteritidis in shell eggs: Current issues and prospects for control». In: *Food Research International* 45.2, pp. 755–764. issn: 0963-9969. doi: https://doi.org/10.1016/j.foodres.2011.04.030. url: https://www.sciencedirect.com/science/article/pii/S0963996911002602.
51. Shi, Zhaohao, Michael J. Rothrock Jr., and Steven C. Ricke (2019). «Applications of Microbiome Analyses in Alternative Poultry Broiler Production Systems». In: *Frontiers in Veterinary Science* 6. issn: 2297-1769. doi: 10.3389/fvets.2019.00157. url: https://www.frontiersin.org/article/10.3389/fvets.2019.00157.
52. Jakočiūnė, D., M. Bisgaard, K. Pedersen, and J.E. Olsen (2014). «Demonstration of persistent contamination of a cooked egg product production facility with Salmonella enterica serovar Tennessee and characterization of the persistent strain». In: *Journal of Applied Microbiology* 117.2, pp. 547–553. doi: https://doi.org/10.1111/jam.12536. url: https://sfamjournals.onlinelibrary.wiley.com/doi/abs/10.1111/jam.12536.
53. Foley, Steven L., Timothy J. Johnson, Steven C. Ricke, Rajesh Nayak, and Jessica Danzeisen (2013). «Salmonella Pathogenicity and Host Adaptation in Chicken-Associated Serovars». In: *Microbiology and Molecular Biology Reviews* 77.4, pp. 582–607. doi: 10.1128/MMBR.00015-13. url: https://journals.asm.org/doi/abs/10.1128/MMBR.00015-13.
54. Han, Jing, Aaron M. Lynne, Donna E. David, *et al.* (Dec. 2012). «DNA Sequence Analysis of Plasmids from Multidrug Resistant Salmonella enterica Serotype Heidelberg Isolates». In: *PLOS ONE* 7.12, pp. 1–8. doi: 10.1371/journal.pone.0051160. url: https://doi.org/10.1371/journal.pone.0051160.
55. Kolling, Glynis, Martin Wu, and Richard Guerrant (2012). «Enteric pathogens through life stages». In: *Frontiers in Cellular and Infection Microbiology* 2. issn: 2235-2988. doi: 10.3389/fcimb.2012.00114. url: https://www.frontiersin.org/article/10.3389/fcimb.2012.00114.
56. Barnaba, Vincenzo and Francesco Sinigaglia (May 1997). «Molecular Mimicry and T Cell–mediated Autoimmune Disease». In: *Journal of Experimental Medicine* 185.9, pp. 1529–1532. issn: 0022-1007. doi: 10.1084/jem.185.9.1529. eprint: https://rupress.org/jem/article-pdf/185/9/1529/1111872/5520.pdf. url: https://doi.org/10.1084/jem.185.9.1529.
57. El Aidy, Sahar, Peter Baarlen, Muriel Derrien, *et al.* (May 2012). «Temporal and spatial interplay of microbiota and intestinal mucosa drive establishment of immune homeostasis in conventionalized mice». In: *Mucosal immunology* 5, pp. 567–79. doi: 10.1038/mi.2012.32.
58. Lawley, Trevor D. and Alan W. Walker (2013). «Intestinal colonization resistance». In: *Immunology* 138.1, pp. 1–11. doi: https://doi.org/10.1111/j.1365-2567.2012.03616.x. url: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2567.2012.03616.x.

59. Sanz, Yolanda, Giada De Palma, and Moisés Laparra (2011). «Unraveling the Ties between Celiac Disease and Intestinal Microbiota». In: *International Reviews of Immunology* 30.4, pp. 207–218. doi: 10.3109/08830185.2011.599084. url: https://doi.org/10.3109/08830185.2011.599084.
60. Marasco, Giovanni, Anna Biase, Ramona Schiumerini, *et al.* (June 2016). «Gut Microbiota and Celiac Disease». In: *Digestive Diseases and Sciences* 61. doi: 10.1007/s10620-015-4020-2.
61. De Palma, Giada, Inmaculada Nadal, Marcela Medina, Ester Donat, Carmen Ribes-Koninckx, Miguel Calabuig, and Yolanda Sanz (Feb. 2010). «Intestinal dysbiosis and reduced immunoglobulin-coated bacteria associated with coeliac disease in children». In: *BMC microbiology* 10, p. 63. doi: 10.1186/1471-2180-10-63.
62. Weinstock, George (Sept. 2012). «Genomic approaches to studying the human microbiota». In: *Nature* 489, pp. 250–6. doi: 10.1038/nature11553.
63. Goodrich, Julia K, Sara C. Di Rienzi, Angela C. Poole, Omry Koren, William A Walters, Gregory J Caporaso, Rob Knight, and Ruth E. Ley (July 2014). «Conducting a microbiome study». English (US). In: *Cell* 158.2, pp. 250–262. issn: 0092-8674. doi: 10.1016/j.cell.2014.06.037.
64. Hanahan, Douglas and Robert A. Weinberg (2011). «Hallmarks of Cancer: The Next Generation». In: *Cell* 144.5, pp. 646–674. issn: 0092-8674. doi: https://doi.org/10.1016/j.cell.2011.02.013. url: https://www.sciencedirect.com/science/article/pii/S0092867411001279.
65. Marshall, Barry J and J.Robin Warren (1984). «Unidentified Curved Bacilli In The Stomach Of Patients With Gastritis And Peptic Ulceration». In: *The Lancet* 323.8390. Originally published as Volume 1, Issue 8390, pp. 1311–1315. issn: 0140-6736. doi: https://doi.org/10.1016/S0140-6736(84)91816-6. url: https://www.sciencedirect.com/science/article/pii/S0140673684918166.
66. Wang, Fengsong, Peng Xia, Fang Wu, *et al.* (2008). «Helicobacter pylori VacA Disrupts Apical Membrane-Cytoskeletal Interactions in Gastric Parietal Cells». In: *Journal of Biological Chemistry* 283.39, pp. 26714–26725. issn: 0021-9258. doi: https://doi.org/10.1074/jbc.M800527200. url: https://www.sciencedirect.com/science/article/pii/S0021925820524149.
67. Vaezi, Michael, Gary Falk, Richard Peek, *et al.* (Oct. 2000). «CagA-positive strains of Helicobacter pylori may protect against Barrett's esophagus». In: *The American journal of gastroenterology* 95, pp. 2206–11. doi: 10.1111/j.1572-0241.2000.02305.x.
68. Browning, Douglas and Stephen Busby (Feb. 2004). «The regulation of bacterial transcription initiation». In: *Nature reviews. Microbiology* 2, pp. 57–65. doi: 10.1038/nrmicro787.
69. Schrödinger, LLC (2021a). «Schrödinger Release 2022-2: BioLuminate, New York».
70. Gusarov, Ivan and Evgeny Nudler (1999). «The Mechanism of Intrinsic Transcription Termination». In: *Molecular Cell* 3.4, pp. 495–504. issn: 1097-2765. doi: https://doi.org/10.1016/S1097-2765(00)80477-3. url: https://www.sciencedirect.com/science/article/pii/S1097276500804773.
71. Peters, Jason, Rachel Mooney, Jeffrey Grass, Erik Jessen, Frances Tran, and Robert Landick (Dec. 2012). «Rho and NusG suppress pervasive antisense transcription in Escherichia coli». In: *Genes & development* 26, pp. 2621–33. doi: 10.1101/gad.196741.112.
72. Raghunathan, Nalini, Rajvardhan M Kapshikar, Jakku K Leela, Jillella Mallikarjun, Philippe Bouloc, and Jayaraman Gowrishankar (Feb. 2018). «Genome-wide relationship between R-loop formation and antisense transcription in Escherichia coli». In: *Nucleic Acids Research* 46.7, pp. 3400–3411. issn: 0305-1048. doi: 10.1093/nar/gky118. eprint: https://academic.oup.com/nar/article-pdf/46/7/3400/24677497/gky118.pdf. url: https://doi.org/10.1093/nar/gky118.
73. Richardson, J P (1982). «Activation of rho protein ATPase requires simultaneous interaction at two kinds of nucleic acid-binding sites». In: *Journal of Biological Chemistry* 257.10, pp. 5760–5766. issn: 0021-9258. doi: https://doi.org/10.1016/S0021-9258(19)83844-9. url: https://www.sciencedirect.com/science/article/pii/S0021925819838449.
74. Li, J, R Horwitz, S McCracken, and J Greenblatt (1992). «NusG, a new Escherichia coli elongation factor involved in transcriptional antitermination by the N protein of phage lambda.» In: *Journal of Biological Chemistry* 267.9, pp. 6012–6019. issn: 0021-9258. doi: https://doi.org/10.1016/S0021-9258(18)42655-5. url: https://www.sciencedirect.com/science/article/pii/S0021925818426555.

75. Roberts, Jeffrey and Joo-Seop Park (2004). «Mfd, the bacterial transcription repair coupling factor: translocation, repair and termination». In: *Current Opinion in Microbiology* 7.2, pp. 120–125. ISSN: 1369-5274. DOI: https://doi.org/10.1016/j.mib.2004.02.014. URL: https://www.sciencedirect.com/science/article/pii/S136952740400027X.
76. Heather, James M. and Benjamin Chain (2016). «The sequence of sequencers: The history of sequencing DNA». In: *Genomics* 107.1, pp. 1–8. ISSN: 0888-7543. DOI: https://doi.org/10.1016/j.ygeno.2015.11.003. URL: https://www.sciencedirect.com/science/article/pii/S0888754315300410.
77. Sanger, F., S. Nicklen, and A. R. Coulson (1977). «DNA sequencing with chain-terminating inhibitors». In: *Proceedings of the National Academy of Sciences* 74.12, pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463. URL: https://www.pnas.org/doi/abs/10.1073/pnas.74.12.5463.
78. Stephens, Zachary D., Skylar Y. Lee, Faraz Faghri, *et al.* (July 2015). «Big Data: Astronomical or Genomical?» In: *PLOS Biology* 13.7, pp. 1–11. DOI: 10.1371/journal.pbio.1002195. URL: https://doi.org/10.1371/journal.pbio.1002195.
79. Ronaghi, Mostafa, Samer Karamohamed, Bertil Pettersson, Mathias Uhlén, and Pål Nyrén (1996). «Real-Time DNA Sequencing Using Detection of Pyrophosphate Release». In: *Analytical Biochemistry* 242.1, pp. 84–89. ISSN: 0003-2697. DOI: https://doi.org/10.1006/abio.1996.0432. URL: https://www.sciencedirect.com/science/article/pii/S0003269796904327.
80. Stein, Lincoln (May 2010). «The case for cloud computing in genome informatics». In: *Genome biology* 11, p. 207. DOI: 10.1186/gb-2010-11-5-207.
81. Dohm, Juliane C., Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer (July 2008). «Substantial biases in ultra-short read data sets from high-throughput DNA sequencing». In: *Nucleic Acids Research* 36.16. e105. ISSN: 0305-1048. DOI: 10.1093/nar/gkn425. eprint: https://academic.oup.com/nar/article-pdf/36/16/e105/39554537/nar_36_16_e105.pdf. URL: https://doi.org/10.1093/nar/gkn425.
82. Van Dijk, Erwin L., Yan Jaszczyszyn, and Claude Thermes (2014a). «Library preparation methods for next-generation sequencing: Tone down the bias». In: *Experimental Cell Research* 322.1, pp. 12–20. ISSN: 0014-4827. DOI: https://doi.org/10.1016/j.yexcr.2014.01.008. URL: https://www.sciencedirect.com/science/article/pii/S0014482714000160.
83. Oyola, Samuel, Thomas Otto, Yong Gu, *et al.* (Jan. 2012). «Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes». In: *BMC genomics* 13, p. 1. DOI: 10.1186/1471-2164-13-1.
84. Braslavsky, Ido, Benedict Hebert, Emil Kartalov, and Stephen R. Quake (2003). «Sequence information can be obtained from single DNA molecules». In: *Proceedings of the National Academy of Sciences* 100.7, pp. 3960–3964. DOI: 10.1073/pnas.0230489100. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0230489100.
85. Bowers, Jayson L, Judith Mitchell, Eric Beer, *et al.* (2009). «Virtual Terminator nucleotides for next generation DNA sequencing». In: *Nature methods* 6, pp. 593–595.
86. Van Dijk, Erwin L., Hélène Auger, Yan Jaszczyszyn, and Claude Thermes (2014b). «Ten years of next-generation sequencing technology». In: *Trends in Genetics* 30.9, pp. 418–426. ISSN: 0168-9525. DOI: https://doi.org/10.1016/j.tig.2014.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0168952514001127.
87. Eid, John, Adrian Fehr, Jeremy Gray, *et al.* (2009). «Real-Time DNA Sequencing from Single Polymerase Molecules». In: *Science* 323.5910, pp. 133–138. DOI: 10.1126/science.1162986. URL: https://www.science.org/doi/abs/10.1126/science.1162986.
88. Haque, Farzin, Jinghong Li, Hai-Chen Wu, Xing-Jie Liang, and Peixuan Guo (2013). «Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA». In: *Nano Today* 8.1, pp. 56–74. ISSN: 1748-0132. DOI: https://doi.org/10.1016/j.nantod.2012.12.008. URL: https://www.sciencedirect.com/science/article/pii/S1748013212001454.
89. Branton, Daniel, David W Deamer, Andre Marziali, *et al.* (2008). «The potential and challenges of nanopore sequencing». In: *Nature biotechnology* 26.10, pp. 1146–1153.
90. Eisenstein, Michael (Apr. 2012). «Oxford Nanopore announcement sets sequencing sector abuzz». In: *Nature biotechnology* 30, pp. 295–6. DOI: 10.1038/nbt0412-295.

91. Quick, Joshua, Aaron R Quinlan, and Nicholas J Loman (Oct. 2014). «A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer». In: *GigaScience* 3.1. 2047-217X-3-22. issn: 2047-217X. doi: 10.1186/2047-217X-3-22. url: https://doi.org/10.1186/2047-217X-3-22.
92. Lander, Eric, Lauren Linton, Bruce Birren, *et al.* (Mar. 2001). «Initial sequencing and analysis of the human genome». In: *Nature* 409, pp. 860–921. doi: 10.1038/35057062.
93. Venter, J. Craig, Mark D. Adams, Eugene W. Myers, *et al.* (2001). «The Sequence of the Human Genome». In: *Science* 291.5507, pp. 1304–1351. doi: 10.1126/science.1058040. url: https://www.science.org/doi/abs/10.1126/science.1058040.
94. Hood, Leroy and David Galas (Feb. 2003). «The digital code of DNA». In: *Nature* 421, pp. 444–8. doi: 10.1038/nature01410.
95. Wang, Zhong, Mark B. Gerstein, and Michael Snyder (Dec. 2009). «RNA-Seq: A Revolutionary Tool for Transcriptomics». In: *Nature reviews. Genetics* 10, pp. 57–63. doi: 10.1038/nrg2484.
96. Mills, James, Yoshihiro Kawahara, and Michael Janitz (May 2013). «Strand-Specific RNA-Seq Provides Greater Resolution of Transcriptome Profiling». In: *Current genomics* 14, pp. 173–81. doi: 10.2174/1389202911314030003.
97. Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold (2007). «Genome-Wide Mapping of in Vivo Protein-DNA Interactions». In: *Science* 316.5830, pp. 1497–1502. doi: 10.1126/science.1141319. eprint: https://www.science.org/doi/pdf/10.1126/science.1141319. url: https://www.science.org/doi/abs/10.1126/science.1141319.
98. Marinov, Georgi K (Mar. 2018). «A decade of ChIP-seq». In: *Briefings in Functional Genomics* 17.2, pp. 77–79. issn: 2041-2657. doi: 10.1093/bfgp/ely012. eprint: https://academic.oup.com/bfg/article-pdf/17/2/77/24502568/ely012.pdf. url: https://doi.org/10.1093/bfgp/ely012.
99. Park, Peter J. (2009). «ChIP–seq: advantages and challenges of a maturing technology». In: *Nature Reviews Genetics* 10.10, pp. 669–680. issn: 1471-0056. doi: 10.1038/nrg2641. url: https://dx.doi.org/10.1038/nrg2641.
100. Giurato, Giorgio, maria rosaria de filippo maria rosaria, Antonio Rinaldi, *et al.* (Dec. 2013). «IMir: An integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq». In: *BMC bioinformatics* 14, p. 362. doi: 10.1186/1471-2105-14-362.
101. Wells, Sandra E., Paul E. Hillner, Ronald D. Vale, and Alan B. Sachs (1998). «Circularization of mRNA by Eukaryotic Translation Initiation Factors». In: *Molecular Cell* 2.1, pp. 135–140. issn: 1097-2765. doi: 10.1016/s1097-2765(00)80122-7.
102. Zhou, Linglin, Xueying Li, Qi Liu, Fangqing Zhao, and Jinyu Wu (2011). «Small RNA transcriptome investigation based on next-generation sequencing technology». In: *Journal of Genetics and Genomics* 38.11, pp. 505–513. issn: 1673-8527. doi: https://doi.org/10.1016/j.jgg.2011.08.006. url: https://www.sciencedirect.com/science/article/pii/S1673852711001470.
103. Wei, Cha-Mer, Alan Gershowitz, and Bernard Moss (1975). «Methylated nucleotides block 5' terminus of HeLa cell messenger RNA». In: *Cell* 4.4, pp. 379–386. issn: 0092-8674. doi: https://doi.org/10.1016/0092-8674(75)90158-0. url: https://www.sciencedirect.com/science/article/pii/0092867475901580.
104. Topisirovic, Ivan, Yuri V. Svitkin, Nahum Sonenberg, and Aaron J. Shatkin (2011). «Cap and cap-binding proteins in the control of gene expression». In: *WIREs RNA* 2.2, pp. 277–298. doi: https://doi.org/10.1002/wrna.52. url: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wrna.52.
105. Sharma, Cynthia, Steve Hoffmann, Fabien Darfeuille, *et al.* (Feb. 2010). «The Primary Transcriptome of the Major Human Pathogen Helicobacter Pylori». In: *Nature* 464, pp. 250–5. doi: 10.1038/nature08756.
106. Ettwiller, Laurence, John Buswell, Erbay Yigit, and Ira Schildkraut (Mar. 2016). «A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome». In: *BMC Genomics* 17. doi: 10.1186/s12864-016-2539-z.
107. Yan, Bo, George Tzertzinis, Ira Schildkraut, and Laurence Ettwiller (2022). «Comprehensive determination of transcription start sites derived from all RNA polymerases using ReCappable-seq». In: *Genome Research* 32.1, pp. 162–174. doi: 10.1101/gr.275784.121. url: http://genome.cshlp.org/content/32/1/162.abstract.

108. Takagi, Toshimitsu, Christine R Moore, Felix Diehn, and Stephen Buratowski (1997). «An RNA 5'-Triphosphatase Related to the Protein Tyrosine Phosphatases». In: *Cell* 89.6, pp. 867–873. ISSN: 0092-8674. DOI: https://doi.org/10.1016/S0092-8674(00)80272-X. URL: https://www.sciencedirect.com/science/article/pii/S009286740080272X.

109. Gross, Christian H. and Stewart Shuman (1998). «RNA 5'-Triphosphatase, Nucleoside Triphosphatase, and Guanylyltransferase Activities of Baculovirus LEF-4 Protein». In: *Journal of Virology* 72.12, pp. 10020–10028. DOI: 10.1128/JVI.72.12.10020-10028.1998. URL: https://journals.asm.org/doi/abs/10.1128/JVI.72.12.10020-10028.1998.

110. Schrödinger, LLC (2021b). «Schrödinger Release 2022-2: Maestro, New York».

111. Reynaud, C., C. Bruno, P. Boullanger, J. Grange, S. Barbesti, and A. Niveleau (1992). «Monitoring of urinary excretion of modified nucleosides in cancer patients using a set of six monoclonal antibodies». In: *Cancer Letters* 61.3, pp. 255–262. ISSN: 0304-3835. DOI: https://doi.org/10.1016/0304-3835(92)90296-8. URL: https://www.sciencedirect.com/science/article/pii/0304383592902968.

112. Kyrieleis, Otto J.P., Jonathan Chang, Marcos de la Peña, Stewart Shuman, and Stephen Cusack (2014). «Crystal Structure of Vaccinia Virus mRNA Capping Enzyme Provides Insights into the Mechanism and Evolution of the Capping Apparatus». In: *Structure* 22.3, pp. 452–465. ISSN: 0969-2126. DOI: https://doi.org/10.1016/j.str.2013.12.014. URL: https://www.sciencedirect.com/science/article/pii/S0969212614000069.

113. Sharma, Cynthia M and Jörg Vogel (2014). «Differential RNA-seq: the approach behind and the biological insight gained». In: *Current Opinion in Microbiology* 19. Ecology and industrial microbiology • Special Section: Novel technologies in microbiology, pp. 97–105. ISSN: 1369-5274. DOI: https://doi.org/10.1016/j.mib.2014.06.010. URL: https://www.sciencedirect.com/science/article/pii/S1369527414000800.

114. Romero, David A., Ayad H. Hasan, Yu-fei Lin, *et al.* (2014). «A comparison of key aspects of gene regulation in Streptomyces coelicolor and Escherichia coli using nucleotide-resolution transcription maps produced in parallel by global and differential RNA sequencing». In: *Molecular Microbiology* 94.5, pp. 963–987. DOI: https://doi.org/10.1111/mmi.12810. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mmi.12810.

115. Jäger, Dominik, Konrad Förstner, Cynthia Sharma, Thomas Santangelo, and John Reeve (Aug. 2014). «Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis». In: *BMC genomics* 15, p. 684. DOI: 10.1186/1471-2164-15-684.

116. Bohlin, Jon, Vegard Eldholm, John Pettersson, Ola Brynildsrud, and Lars Snipen (Feb. 2017). «The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes». In: *BMC Genomics* 18. DOI: 10.1186/s12864-017-3543-7.

117. Dar, Daniel, Maya Shamir, J. R. Mellin, Mikael Koutero, Noam Stern-Ginossar, Pascale Cossart, and Rotem Sorek (2016). «Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria». In: *Science* 352.6282, aad9822. DOI: 10.1126/science.aad9822. URL: https://www.science.org/doi/abs/10.1126/science.aad9822.

118. Dar, Daniel and Rotem Sorek (Apr. 2018). «High-resolution RNA 3'-ends mapping of bacterial Rho-dependent transcripts». In: *Nucleic Acids Research* 46.13, pp. 6797–6805. ISSN: 0305-1048. DOI: 10.1093/nar/gky274. eprint: https://academic.oup.com/nar/article-pdf/46/13/6797/25228000/gky274.pdf. URL: https://doi.org/10.1093/nar/gky274.

119. Lianoglou, S., V. Garg, J. L. Yang, C. S. Leslie, and C. Mayr (2013). «Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression». In: *Genes & Development* 27.21, pp. 2380–2396. ISSN: 0890-9369. DOI: 10.1101/gad.229328.113. URL: https://dx.doi.org/10.1101/gad.229328.113.

120. Ho, C. Kiong, Verl Sriskanda, Susan Mccracken, David Bentley, Beate Schwer, and Stewart Shuman (1998). «The Guanylyltransferase Domain of Mammalian mRNA Capping Enzyme Binds to the Phosphorylated Carboxyl-terminal Domain of RNA Polymerase II». In: *Journal of Biological Chemistry* 273.16, pp. 9577–9585. ISSN: 0021-9258. DOI: 10.1074/jbc.273.16.9577.

121. Shatkin, Aaron and James Manley (Nov. 2000). «The ends of the affair: Capping and polyadenylation». In: *Nature structural biology* 7, pp. 838–42. DOI: 10.1038/79583.

122. Shatkin, A.J. (1976). «Capping of eucaryotic mRNAs». In: *Cell* 9.4, Part 2, pp. 645–653. ISSN: 0092-8674. DOI: https://doi.org/10.1016/0092-8674(76)90128-8. URL: https://www.sciencedirect.com/science/article/pii/0092867476901288.

123. Langberg, S.R. and B. Moss (1981). «Post-transcriptional modifications of mRNA. Purification and characterization of cap I and cap II RNA (nucleoside-2'-)-methyltransferases from HeLa cells.» In: *Journal of Biological Chemistry* 256.19, pp. 10054–10060. ISSN: 0021-9258. DOI: https://doi.org/10.1016/S0021-9258(19)68740-5. URL: https://www.sciencedirect.com/science/article/pii/S0021925819687405.

124. Daffis, Stephane, Kristy Szretter, Jill Schriewer, *et al.* (Nov. 2010). «2'-O-Methylation of the viral mRNA cap evades host restriction by IFIT family members». In: *Nature* 468, pp. 452–6. DOI: 10.1038/nature09489.

125. Lewis, Joe D. and Elisa Izaurfldе (1997). «The Role of the Cap Structure in RNA Processing and Nuclear Export». In: *European Journal of Biochemistry* 247.2, pp. 461–469. DOI: https://doi.org/10.1111/j.1432-1033.1997.00461.x. URL: https://febs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1432-1033.1997.00461.x.

126. Kowtoniuk, Walter E., Yinghua Shen, Jennifer M. Heemstra, Isha Agarwal, and David R. Liu (2009). «A chemical screen for biological small molecule–RNA conjugates reveals CoA-linked RNA». In: *Proceedings of the National Academy of Sciences* 106.19, pp. 7768–7773. DOI: 10.1073/pnas.0900528106. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0900528106.

127. Marcotrigiano, Joseph, Anne-Claude Gingras, Nahum Sonenberg, and Stephen K. Burley (1997). «Cocrystal Structure of the Messenger RNA 5' Cap-Binding Protein (eIF4E) Bound to 7-methyl-GDP». In: *Cell* 89.6, pp. 951–961. ISSN: 0092-8674. DOI: 10.1016/s0092-8674(00)80280-9.

128. Amrani, Nadia, Shubhendu Ghosh, David Mangus, and Allan Jacobson (July 2008). «Translation factors promote the formation of two states of the closed-loop mRNP». In: *Nature* 453, pp. 1276–80. DOI: 10.1038/nature06974.

129. Chen, Y Grace, Walter E Kowtoniuk, Isha Agarwal, Yinghua Shen, and David R Liu (2009). «LC/MS analysis of cellular RNA reveals NAD-linked RNA». In: *Nature Chemical Biology* 5.12, pp. 879–881. ISSN: 1552-4450. DOI: 10.1038/nchembio.235. URL: https://dx.doi.org/10.1038/nchembio.235.

130. Bird, Jeremy G., Yu Zhang, Yuan Tian, *et al.* (2016). «The mechanism of RNA 5' capping with NAD+, NADH and desphospho-CoA». In: *Nature* 535.7612, pp. 444–447. ISSN: 0028-0836. DOI: 10.1038/nature18622. URL: https://dx.doi.org/10.1038/nature18622.

131. Luciano, Daniel J. and Joel G. Belasco (2020). «Np_4A alarmones function in bacteria as precursors to RNA caps». In: *Proceedings of the National Academy of Sciences* 117.7, pp. 3560–3567. DOI: 10.1073/pnas.1914229117. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1914229117.

132. Frindert, Jens, Yaqing Zhang, Gabriele Nübel, *et al.* (2018). «Identification, Biosynthesis, and Decapping of NAD-Capped RNAs in B. subtilis». In: *Cell Reports* 24.7, 1890–1901.e8. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2018.07.047. URL: https://www.sciencedirect.com/science/article/pii/S2211124718311501.

133. Shine, J. and L. Dalgarno (1974). «The 3'-Terminal Sequence of Escherichia coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites». In: *Proceedings of the National Academy of Sciences* 71.4, pp. 1342–1346. DOI: 10.1073/pnas.71.4.1342. URL: https://www.pnas.org/doi/abs/10.1073/pnas.71.4.1342.

134. Steitz, J A and K Jakes (1975). «How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli.» In: *Proceedings of the National Academy of Sciences* 72.12, pp. 4734–4738. DOI: 10.1073/pnas.72.12.4734. URL: https://www.pnas.org/doi/abs/10.1073/pnas.72.12.4734.

135. Moll, Isabella, Sonja Grill, Claudio O. Gualerzi, and Udo Bläsi (2002). «Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control». In: *Molecular Microbiology* 43.1, pp. 239–246. DOI: https://doi.org/10.1046/j.1365-2958.2002.02739.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2958.2002.02739.x.

136. Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

137. Martin, Marcel (2011). «Cutadapt removes adapter sequences from high-throughput sequencing reads». In: *EMBnet.journal* 17.1, pp. 10–12. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200. URL: https://journal.embnet.org/index.php/embnetjournal/article/view/200.

138. Zhang, Zhijiao, Scott Schwartz, Lukas Wagner, and Webb Miller (Feb. 2000). «Greedy Algorithm for Aligning DNA Sequences». In: *Journal of computational biology : a journal of computational molecular cell biology* 7, pp. 203–14. DOI: 10.1089/10665270050081478.

139. Langmead, Ben and Steven Salzberg (Mar. 2012). «Langmead B, Salzberg SL.. Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359». In: *Nature methods* 9, pp. 357–9. DOI: 10.1038/nmeth.1923.

140. Li, Heng, Bob Handsaker, Alec Wysoker, *et al.* (June 2009). «The Sequence Alignment/Map format and SAMtools». In: *Bioinformatics* 25.16, pp. 2078–2079. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp352. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/16/2078/531810/btp352.pdf. URL: https://doi.org/10.1093/bioinformatics/btp352.

141. Thorvaldsdóttir, Helga, James T. Robinson, and Jill P. Mesirov (Apr. 2012). «Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration». In: *Briefings in Bioinformatics* 14.2, pp. 178–192. ISSN: 1467-5463. DOI: 10.1093/bib/bbs017. eprint: https://academic.oup.com/bib/article-pdf/14/2/178/546734/bbs017.pdf. URL: https://doi.org/10.1093/bib/bbs017.

142. Quinlan, Aaron R. and Ira M. Hall (Jan. 2010). «BEDTools: a flexible suite of utilities for comparing genomic features». In: *Bioinformatics* 26.6, pp. 841–842. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq033. eprint: https://academic.oup.com/bioinformatics/article-pdf/26/6/841/16897802/btq033.pdf. URL: https://doi.org/10.1093/bioinformatics/btq033.

143. R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

144. Gentleman, Robert C, Vincent J Carey, Douglas M Bates, *et al.* (2004). «Bioconductor: open software development for computational biology and bioinformatics». In: *Genome Biology* 5.10, R80. ISSN: 1465-6906. DOI: 10.1186/gb-2004-5-10-r80.

145. Morgan, Martin (2021). *BiocManager: Access the Bioconductor Project Package Repository*. R package version 1.30.16. URL: https://CRAN.R-project.org/package=BiocManager.

146. Gaidatzis, Dimos, Anita Lerch, Florian Hahne, and Michael B. Stadler (Dec. 2014). «QuasR: quantification and annotation of short reads in R». In: *Bioinformatics* 31.7, pp. 1130–1132. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu781. eprint: https://academic.oup.com/bioinformatics/article-pdf/31/7/1130/17124757/btu781.pdf. URL: https://doi.org/10.1093/bioinformatics/btu781.

147. Pagès, Herve (2021). *Software infrastructure for efficient representation of full genomes and their SNPs*. R package version 1.62.0. URL: https://bioconductor.org/packages/BSgenome.

148. Haider, Syed, Daryl Waggott, Emilie Lalonde, Clement Fung, and Paul C. Boutros (2019). *bedr: Genomic Region Processing using Tools Such as 'BEDTools', 'BEDOPS' and 'Tabix'*. R package version 1.0.7. URL: https://CRAN.R-project.org/package=bedr.

149. Salgado, Heladia, Socorro Gama-Castro, Martín Peralta-Gil, *et al.* (Jan. 2006). «RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions». In: *Nucleic Acids Research* 34.suppl_1, pp. D394–D397. ISSN: 0305-1048. DOI: 10.1093/nar/gkj156. URL: https://doi.org/10.1093/nar/gkj156.

150. Robb, Nicole C., Thorben Cordes, Ling Chin Hwang, *et al.* (2013). «The Transcription Bubble of the RNA Polymerase–Promoter Open Complex Exhibits Conformational Heterogeneity and Millisecond-Scale Dynamics: Implications for Transcription Start-Site Selection». In: *Journal of Molecular Biology* 425.5, pp. 875–885. ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2012.12.015. URL: https://www.sciencedirect.com/science/article/pii/S0022283612009485.

151. Love, Michael, Wolfgang Huber, and Simon Anders (Dec. 2014). «Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2». In: *Genome Biol* 15, p. 550. DOI: 10.1186/s13059-014-0550-8.

152. Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth (Nov. 2009). «edgeR: a Bioconductor package for differential expression analysis of digital gene expression data». In: *Bioinformatics* 26.1, pp. 139–140.

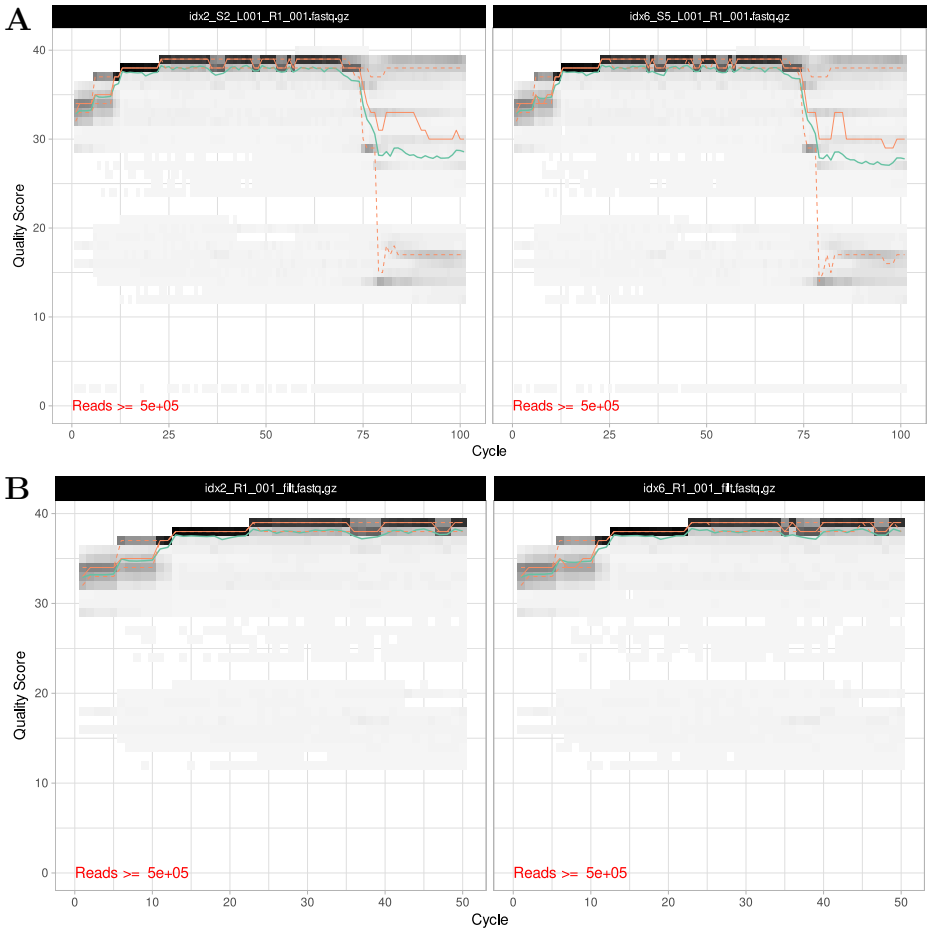
ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp616. eprint: https://academic.oup.com/bioinformatics/article-pdf/26/1/139/443156/btp616.pdf. URL: https://doi.org/10.1093/bioinformatics/btp616.

153. Law, Charity, Yunshun Chen, Wei Shi, and Gordon Smyth (Feb. 2014). «Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts». In: *Genome biology* 15, R29. DOI: 10.1186/gb-2014-15-2-r29.
154. Lybecker, Meghan, Ivana Bilusic, and Rahul Raghavan (Sept. 2014). «Pervasive transcription: Detecting functional RNAs in bacteria». In: *Transcription* 5, pp. 1–5. DOI: 10.4161/21541272.2014.944039.
155. Yee, Thomas W. (2010). «The VGAM Package for Categorical Data Analysis». In: *Journal of Statistical Software* 32.10, pp. 1–34. DOI: 10.18637/jss.v032.i10. URL: https://www.jstatsoft.org/index.php/jss/article/view/v032i10.
156. Santos-Zavaleta, Alberto, Heladia Salgado, Socorro Gama-Castro, *et al.* (Nov. 2018). «RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12». In: *Nucleic Acids Research* 47.D1, pp. D212–D220. ISSN: 0305-1048. DOI: 10.1093/nar/gky1077. eprint: https://academic.oup.com/nar/article-pdf/47/D1/D212/27437535/gky1077.pdf. URL: https://doi.org/10.1093/nar/gky1077.
157. Keseler, Ingrid M., Socorro Gama-Castro, Amanda Mackie, *et al.* (2021). «The EcoCyc Database in 2021». In: *Frontiers in Microbiology* 12. ISSN: 1664-302X. DOI: 10.3389/fmicb.2021.711077. URL: https://www.frontiersin.org/article/10.3389/fmicb.2021.711077.
158. Schlüter, Jan-Philip, Jan Reinkensmeier, Melanie Barnett, Claus Lang, Elizaveta Krol, Robert Giegerich, Sharon Long, and Anke Becker (Mar. 2013). «Global mapping of transcription start sites and promoter motifs in the symbiotic α-proteobacterium Sinorhizobium meliloti 1021». In: *BMC genomics* 14, p. 156. DOI: 10.1186/1471-2164-14-156.
159. Villegas, Andre and Andrew M. Kropinski (2008). «An analysis of initiation codon utilization in the Domain Bacteria – concerns about the quality of bacterial genome annotation». In: *Microbiology* 154.9, pp. 2559–2661. ISSN: 1465-2080. DOI: https://doi.org/10.1099/mic.0.2008/021360-0. URL: https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.2008/021360-0.
160. Kearse, Michael G. and Jeremy E. Wilusz (2017). «Non-AUG translation: a new start for protein synthesis in eukaryotes». In: *Genes & Development* 31.17, pp. 1717–1731. DOI: 10.1101/gad.305250.117. URL: http://genesdev.cshlp.org/content/31/17/1717.abstract.
161. Okuda, Shujiro and Akiyasu C. Yoshizawa (Nov. 2010). «ODB: a database for operon organizations, 2011 update». In: *Nucleic Acids Research* 39.suppl_1, pp. D552–D555. ISSN: 0305-1048. DOI: 10.1093/nar/gkq1090. eprint: https://academic.oup.com/nar/article-pdf/39/suppl_1/D552/7628337/gkq1090.pdf. URL: https://doi.org/10.1093/nar/gkq1090.
162. Bucher, Etienne, Jon Reinders, and Marie Mirouze (2012). «Epigenetic control of transposon transcription and mobility in Arabidopsis». In: *Current Opinion in Plant Biology* 15.5, pp. 503–510. ISSN: 1369-5266. DOI: https://doi.org/10.1016/j.pbi.2012.08.006. URL: https://www.sciencedirect.com/science/article/pii/S1369526612001094.
163. Blaxter, Mark, Jenna Mann, Tom Chapman, Fran Thomas, Claire Whitton, Robin Floyd, and Eyualem Abebe (2005). «Defining operational taxonomic units using DNA barcode data». In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1462, pp. 1935–1943. DOI: 10.1098/rstb.2005.1725. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2005.1725.
164. Kunin, Victor, Anna Engelbrektson, Howard Ochman, and Philip Hugenholtz (2010). «Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates». In: *Environmental Microbiology* 12.1, pp. 118–123. DOI: https://doi.org/10.1111/j.1462-2920.2009.02051.x. URL: https://sfamjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1462-2920.2009.02051.x.
165. Edgar, Robert (Oct. 2017). «Accuracy of microbial community diversity estimated by closed- and open-reference OTUs». In: *PeerJ* 5, e3889. DOI: 10.7717/peerj.3889.
166. Callahan, Benjamin J, Joan Wong, Cheryl Heiner, Steve Oh, Casey M Theriot, Ajay S Gulati, Sarah K McGill, and Michael K Dougherty (July 2019). «High-throughput amplicon sequencing of the

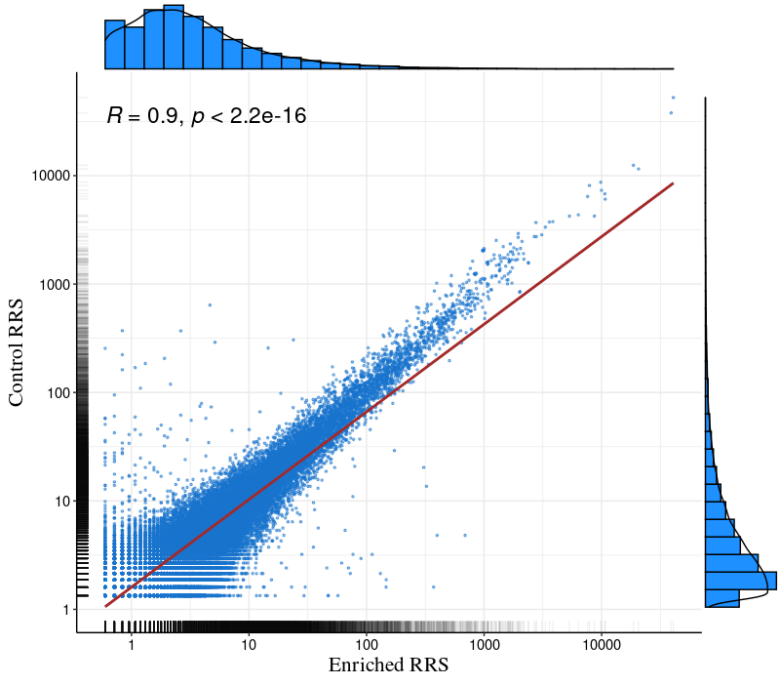
full-length 16S rRNA gene with single-nucleotide resolution». In: *Nucleic Acids Research* 47.18, e103–e103. ISSN: 0305-1048. DOI: 10.1093/nar/gkz569. URL: https://doi.org/10.1093/nar/gkz569.

167. Chiarello, Marlène, Mark McCauley, Sébastien Villéger, and Colin R. Jackson (Feb. 2022). «Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold». In: *PLOS ONE* 17.2, pp. 1–19. DOI: 10.1371/journal.pone.0264443. URL: https://doi.org/10.1371/journal.pone.0264443.
168. Callahan, Benjamin, Paul Mcmurdie, Michael Rosen, Andrew Han, Amy Jo Johnson, and Susan Holmes (Apr. 2016). «DADA2: High resolution sample inference from amplicon data». In: DOI: https://doi.org/10.1038/nmeth.3869.
169. Haas, Brian J., Dirk Gevers, Ashlee M. Earl, *et al.* (2011). «Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons». In: *Genome Research* 21.3, pp. 494–504. ISSN: 1088-9051. DOI: 10.1101/gr.112730.110. URL: https://dx.doi.org/10.1101/gr.112730.110.
170. Wang, Qiong, George M. Garrity, James M. Tiedje, and James R. Cole (2007). «Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy». In: *Applied and Environmental Microbiology* 73.16, pp. 5261–5267. DOI: 10.1128/AEM.00062-07. URL: https://journals.asm.org/doi/abs/10.1128/AEM.00062-07.
171. Cole, James R., Qiong Wang, Jordan A. Fish, *et al.* (Nov. 2013). «Ribosomal Database Project: data and tools for high throughput rRNA analysis». In: *Nucleic Acids Research* 42.D1, pp. D633–D642. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1244. URL: https://doi.org/10.1093/nar/gkt1244.
172. Merkel, Dirk (2014). «Docker: lightweight linux containers for consistent development and deployment». In: *Linux journal* 2014.239, p. 2.
173. Shao, Wenjun, Morgan N. Price, Adam M. Deutschbauer, Margaret F. Romine, Adam P. Arkin, Joerg Vogel, and Caroline S. Harwood (2014). «Conservation of Transcription Start Sites within Genes across a Bacterial Genus». In: *mBio* 5.4, e01398–14. DOI: 10.1128/mBio.01398-14. URL: https://journals.asm.org/doi/abs/10.1128/mBio.01398-14.
174. Cho, Byung Kwan, Karsten Zengler, Yu Qiu, Young Park, Eric Knight, Christian Barrett, Yuan Gao, and Bernhard Palsson (Nov. 2009). «The Transcription Unit Architecture of the Escherichia Coli Genome». In: *Nature biotechnology* 27, pp. 1043–9. DOI: 10.1038/nbt.1582.
175. Chang, Winston, Joe Cheng, JJ Allaire, *et al.* (2021). *shiny: Web Application Framework for R*. R package version 1.7.1. URL: https://CRAN.R-project.org/package=shiny.
176. Hilker, Rolf, Kai Bernd Stadermann, Oliver Schwengers, Evgeny Anisiforov, Sebastian Jaenicke, Bernd Weisshaar, Tobias Zimmermann, and Alexander Goesmann (Aug. 2016). «ReadXplorer 2—detailed read mapping analysis and visualization from one single source». In: *Bioinformatics* 32.24, pp. 3702–3708. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw541. URL: https://doi.org/10.1093/bioinformatics/btw541.

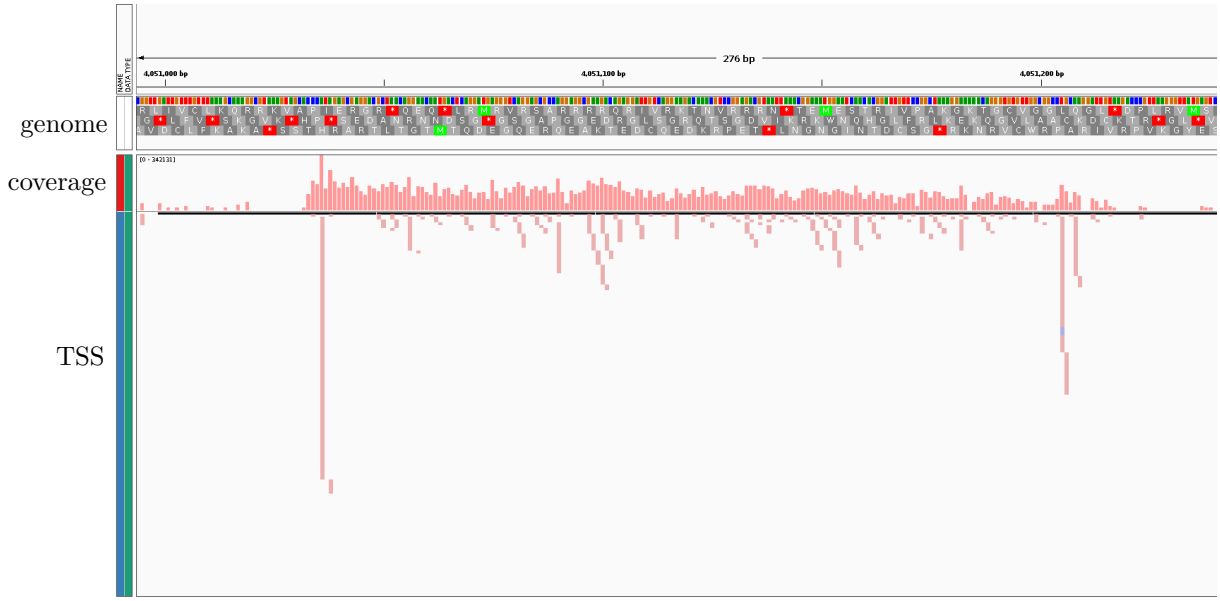
Παραρτημα



Σχήμα Α′.1: *Διαγράμματα ποιότητας αλληλούχισης των μικροβιακών δειγμάτων, για κάθε ένα από τα δύο δείγματα.* ***(Α)*** *Πριν από την προ–επεξεργασία και φιλτράρισμα των δεδομένων.* ***(Β)*** *Μετά από την προ–επεξεργασία των δεδομένων. Όπως είναι αναμενόμενο, η ποιότητα μετά την προ-επεξεργασία αυξάνεται και τα μήκη των αναγνώσεων μειώνονται.*



Σχήμα Α′.2: *Ανάλυση βιολογικών αντιγράφων. Ο συντελεστής συσχέτισης των τιμών RRS, μεταξύ του αντιγράφου 1 και του αντιγράφου 2 είναι 0,9.*



Σχήμα Α′.3: *Αναπαράσταση της περιοχής με το μεγαλύτερο βάθος αλληλούχισης, μέσα από τον Integrative Genomics Viewer (iGV) και με τη βοήθεια του προγράμματος ReadXplorer 2 (*Hilker et al., *2016). Η συγκεκριμένη περιοχή, αντιστοιχεί στο γονίδιο που κωδικοποιεί το sRNA με κωδική ονομασία csrC και σε αυτό το μήκος φαίνονται όλα τα εσωτερικά TSS που υπάρχουν, ως σημεία μονής βάσης. Το διάγραμμα επικάλυψης εμφανίζει σε λογαριθμική κλίμακα την ποσοτικοποίηση των TSS για κάθε θέση.*