ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ- ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Βιοπληροφορική – Υπολογιστική Βιολογία»

«Δημιουργία βάσης δεδομένων πανγονιδιώματος (pangenome) *Proteus mirabilis* και ορισμός του core genome ως εργαλείου φυλογενετικής ανάλυσης»

Διπλωματική Εργασία
Σκουλάκης Ανάργυρος
Α.Μ. 71812

Επιβλέπων Καθηγητής
Παντελής Μπάγκος

Αθήνα, 2021

Ευχαριστίες

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στο εργαστήριο Βακτηριολογίας του Ελληνικού Ινστιτούτου Παστέρ, υπό την επίβλεψη του καθηγητή κ. Παντελή Μπάγκου, σε συνεργασία με τους ερευνητές του Ελληνικού Ινστιτούτου Παστέρ κ. Βιβή Μυριαγκού και κ. Στάθη Κωτσάκη, Όλους τους παραπάνω ευχαριστώ βαθύτατα για τη βοήθεια και την καθοδήγηση. Ιδιαίτερα θα ήθελα να ευχαριστήσω τον κ. Κωτσάκη, ο οποίος καθόλη τη διάρκεια της εκπόνησης της διπλωματικής εργασίας ήταν δίπλα μου σε κάθε απορία και δυσκολία, έτοιμος να προσφέρει κάθε δυνατή βοήθεια.
Επίσης, ευχαριστίες απευθύνονται στους καθηγητές κ. Ιωάννη Τρουγκάκο και κ. Βασιλική Οικονομίδου, για τη συμμετοχή τους στην τριμελή συμβουλευτική επιτροπή της παρούσας διπλωματικής.

Ανάργυρος Σκουλάκης

Περιεχόμενα

Περίληψη

Ο *Proteus mirabilis* είναι ένα Gram αρνητικό βακτήριο, το οποίο ανήκει στην οικογένεια *Morganellaceae* και το οποίο στον άνθρωπο, ενοχοποιείται κυρίως για λοιμώξεις του ουροποιητικού συστήματος. Το πανγονιδίωμα του *P. mirabilis* είναι το σύνολο γονιδίων που συναντώνται στα στελέχη του είδους αυτού και περιλαμβάνει το core genome, δηλαδή τα κοινά γονίδια τα οποία συναντώνται σε όλα τα στελέχη, το dispensable genome, δηλαδή τα γονίδια που συναντώνται σε παραπάνω από ένα στελέχη αλλά όχι σε όλα, και το unique genome, δηλαδή τα γονίδια που βρίσκονται μόνο σε ένα στέλεχος. Η αύξηση και η διασπορά των ανθεκτικών σε αντιβιοτικά στελεχών του *P. mirabilis* και η ανάγκη για την επιδημιολογική επιτήρηση τους απαιτεί την ανάπτυξη νέων πιο ευαίσθητων φυλογενετικών τεχνικών και εργαλείων. Έτσι σκοπός της συγκεκριμένης διπλωματικής εργασίας είναι ο χαρακτηρισμός του core genome και του πανγονιδιώματος του βακτηρίου *P. mirabilis* με στόχο την πιθανή χρήση του core genome ως εργαλείο για τη φυλογενετική μελέτη των στελεχών του βακτηριακού αυτού είδους.

Για την εύρεση του core genome και του πανγονιδιώματος, έγινε επιλογή, από το σύνολο των γενωμάτων που είναι κατατεθειμένα στη βάση δεδομένων GenBank, των πιο αξιόπιστων γενωμάτων για ανάλυση, ενώ στη συνέχεια με τη χρήση του εργαλείου Prokka βρέθηκαν οι πρωτεΐνες που κωδικοποιούνται σε κάθε γένωμα, και με τη χρήση του εργαλείου CD-HIT ομαδοποιήθηκαν σε ομάδες ορθόλογων πρωτεϊνών (clusters) με βάση την ομοιότητα και το ποσοστό κάλυψης. Τα clusters στα οποία ταξινομούνται πρωτεΐνες από όλα τα γενώματα, που χρησιμοποιήσαμε στην ανάλυση, θεωρούνται τα clusters του core genome και το core genome αποτελείται από τα αντιπροσωπευτικά γονίδια των clusters αυτών. Σε κάθε cluster του core genome έγινε στοίχιση των διαφορετικών γονιδίων που συναντώνται στο cluster αυτό, και στη συνέχεια δημιουργήθηκε για κάθε στέλεχος το ψευδογονιδίωμά του, δηλαδή ενωμένες στη σειρά οι αλληλουχίες των γονιδίων του στελέχους, που εμπεριέχονται στα clusters του core genome. Συγκρίνοντας τα ψευδογενώματα των διαφόρων στελεχών, δημιουργήθηκε το φυλογενετικό δέντρο του *P. mirabilis* με τη χρήση του εργαλείου RAxML. Επίσης, βρέθηκαν οι core genome Multilocus Locus Sequence Types (cgMLST τύποι) των διαφόρων στελεχών του *P. mirabilis*, και συγκρίθηκαν μεταξύ τους.

Αποτέλεσμα της παρούσας εργασίας αποτελεί ο χαρακτηρισμός του core genome και του πανγονιδιώματος του βακτηρίου *P. mirabilis*, η ταξινόμηση των διαφόρων στελεχών σε cgMLST τύπους και η δημιουργία του φυλογενετικού δέντρου όλων των στελεχών του *P. mirabilis,* που είναι κατατεθειμένα στη βάση δεδομένων Genbank. Επίσης, από την ανάλυση μας φαίνεται ότι οι φυλογενετικές αναλύσεις με τη χρήση του core genome είναι αξιόπιστες και μεγάλης ακρίβειας, και μπορούν να χρησιμοποιηθούν για επιδημιολογική επιτήρηση των διαφόρων επιδημικών στελεχών. Τέλος, στα πλαίσια αυτής της διπλωματικής εργασίας δημιουργήθηκαν αυτόματα εργαλεία για τη πραγματοποίησή πανγονιδιωματικών αναλύσεων.

Abstract

Proteus mirabilis is a Gram negative bacterium that belongs in the family of Morganellaceae, and in humans is responsible mainly for urinary tract infections. The pangenome of *P. mirabilis* is the sum of all genes, that are observed in *P. mirabilis* strains and includes the core genome, corresponding to the common genes that are found in all the *P. mirabilis* strains, the dispensable genome, i.e. the set of genes that are found in more than one strain but not in all strains, and lastly the genes that are strain specific and are found in only one strain forming the unique genome. The growth and spread of antibiotic resistant strains of *P. mirabilis* and the need for their epidemiological surveillance requires the development of new, more sensitive phylogenetic techniques and tools. Thus, the purpose of this thesis is the characterization of the core genome and the pangenome of the bacterium *P. mirabilis*, in order to investigate the possible use of the core genome as a tool for phylogenetic studies.

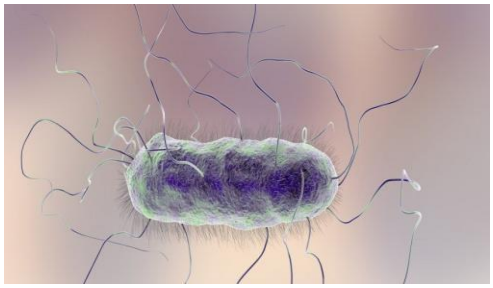
To find the core genome and the pangenome, among the assembled genomes stored in Genbank database, only the most reliable assemblies were used for analysis. Then using the tool Prokka the assemblies were annotated and using the CD-HIT tool their proteins were grouped into groups of orthologous proteins (clusters) based on similarity and coverage rates. Clusters, in which all the genomes used in the analysis are found, are considered the clusters of core genome, and the core genome consists of their representative genes. In each cluster of the core genome, the different genes found in this cluster were aligned. Using the aligned data, the pseudogenome of every assembly was created; the pseudogenome of an assembly consists of the aligned sequences of all the genes of this assembly contained in the clusters of the core genome, joined in series. Comparing the different pseudogenomes, the phylogenetic tree of *P. mirabilis* was created using the RAxML tool. Also, the core genome Multilocus Locus Sequence Types (cgMLST types) of the different strains of *P. mirabilis* were found, and compared with each other.

The results of the present work are the characterization of the core genome and the pangenome of the bacterium *P. mirabilis*, the classification of the various strains into cgMLST types and the creation of the phylogenetic tree of all *P. mirabilis* genomes stored in Genbank database. Also, from our analysis, it appears that phylogenetic analyses using the core genome are reliable and highly accurate, and can be used for epidemiological surveillance of various epidemic strains. Finally, in the context of this thesis, automatic tools for pangenomic analyses were created.

1. Εισαγωγή

1.1. Γενικά χαρακτηριστικά του *Proteus mirabilis*

Το είδος *Proteus mirabilis* είναι ένα αερόβιο και προαιρετικά αναερόβιο Gram αρνητικό βακτήριο, το οποίο ανήκει γένος *Proteus*, στην οικογένεια *Morganellaceae*, στην ομοταξία *Enterobacterales*, στη τάξη *Gammaproteobacteria*, στη συνομοταξία *Proteobacteria*, και στο βασίλειο Bacteria (Adeolu et al., 2016).



Εικόνα 1. Φωτογραφία του βακτηρίου Proteus mirabilis. Έχει παρθεί από την ιστοσελίδα https://www.hygiene-in-practice.com/pathogen/proteus-mirabilis-en/

Η οικογένεια *Morganellaceae* περιλαμβάνει gram αρνητικά βακτήρια και περιλαμβάνει 8 γένη: *Morganella*, *Arsenophonus, Cosenzaea, Moellerella, Photorhabdus, Proteus, Providencia* και *Xenorhabdus* (Adeolu et al., 2016). Το γένος *Arsenophonus* περιλαμβάνει βακτήρια ενδοσυμβιοτικά με διάφορα έντομα (Nováková et al., 2009). Τα γένη *Photorhabdus* και *Xenorhabdus* προκαλούν λοιμώξεις σε έντομα σε συνεργασία με εντομοπαθογόνα νηματοειδή (Jaffuel et al., 2019). Το γένος *Providencia* είναι συχνά συναντούμενο βακτήριο σε μύγες, και ιδιαίτερα στην *Drosophila melanogaster* (Martinson et al., 2017). Το γένος *Morganella* περιλαμβάνει μόνο ένα είδος, την *M. morganii,* και συναντάται στην εντερική χλωρίδα του ανθρώπου και άλλων θηλαστικών (Golemi-Kotra 2008). Το γένος *Cosenzaea* περιλαμβάνει και αυτό μόνο ένα είδος, το *C. myxofaciens*, που παλαιότερα ήταν χαρακτηρισμένο ως *Proteus* (Giammanco et al., 2011). Τέλος, το γένος *Moellerella* περιλαμβάνει μόνο ένα είδος και αυτό, την *M. wisconsensis* και συναντάται στο νερό, σε τρόφιμα, σε κόπρανα πουλιών και σπάνια σε κόπρανα ανθρώπου *(*Aller et al., 2009).
Το γένος *Proteus* περιλαμβάνει 13 είδη: *P. cibi*, *P. faecis*, *P. mirabilis, P. vulgaris, P. penneri, P. hauseri, P. terrae, P. cibarius, P. columbae, P. alimentorum,* και 3 genomospecies, *P. genomospecies 4, 5 and 6* (Dai et al., 2019). Το γένος *Proteus* περιγράφηκε για πρώτη φορά από τον Hauser το 1885 και ονομάστηκε *Proteus* από τον θεό Πρωτέα, λόγω της ικανότητας του βακτηρίου για μορφολογικές αλλαγές των αποικιών του. Το ενδιαίτημα των *Proteus* είναι το έδαφος, το νερό, ο γαστρεντερικός σωλήνας και τα κόπρανα θηλαστικών. Όλα τα είδη του γένους *Proteus* μοιράζονται τα παρακάτω χαρακτηριστικά:

Πίνακας 1: Κοινά χαρακτηριστικά που απαντώνται σε όλα τα βακτήρια που ανήκουν στο γένος *Proteus.*

Κοινά χαρακτηριστικά βακτηρίων γένους *Proteus*:		
Gram αρνητικά	Μη σπορογόνα βακτήρια	Προαιρετικά αναερόβια
Απαμινώνουν την φαινυλαλανίνη σε φαινυλοπυροσταφιλικό οξύ	Οξειδάση αρνητικά	Μεταβολίζουν την γλυκόζη και όχι τη λακτόζη
Ουρεάση θετικά		

Η ταξινόμηση των στελεχών του *Proteus* είναι μία διαδικασία που έχει διαρκέσει πολλά χρόνια. Μέχρι τη δεκαετία του '60 η ταξινόμηση των βακτηρίων στηρίζονταν σε φαινοτυπικά χαρακτηριστικά και σε παρατηρήσεις των αποικιών τους. Παρακάτω παρουσιάζονται τα φαινοτυπικά-βιοχημικά χαρακτηριστικά των διαφορετικών ειδών *Proteus,* βάσει των οποίων γίνεται ο διαχωρισμός τους. (Dai et al., 2019).

Table 2. The growth conditions and biochemical characteristics of *P. cibi*, *P. faecis* and other *Proteus* species

Strains: 1, *P. cibi* FJ2001126-3^T; 2, *P. faecis* TJ1636^T and *P. faecis* 08MAS3143; 3, *P. faecis* 08MAS1600 and *P. faecis* 08MAS1603; 4, *P. faecis* CA120921; 5, *P. faecis* 08MAS2231; 6, *P. faecis* 08MAS2631; 7, *P. mirabilis* KCTC 2566^T; 8, *P. vulgaris* KCTC 2579^T; 9, *P. penneri* ATCC 33519^T; 10, *P. hauseri* JCM 1668^T; 11, *P. terrae* LMG 28659^T; 12, *P. cibarius* JCM 30699^T; 13, *Proteus* genomospecies 4 ATCC 51469^T; 14, *Proteus* genomospecies 5 ATCC 51470^T; 15, *Proteus* genomospecies 6 ATCC 51471^T; 16, *P. columbae* 08MAS2615^T; 17, *P. alimentorum* 08MAS0041^T. +, Positive; –, negative. All test data were from this study and the growth condition in this study. NT, Not tested.

Growth condition/biochemical characteristics	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Swarming (under 1.5% agar)	+	+	+	+	+	+	+	+	–	+	+	+	+	+	+	+	+
Optimum temperature for growth (°C)	37	37	37	37	37	37	37	37	37	37	37	37	37	37	37	37	37
Temperature range for growth (°C)	10–45	10–45	10–45	10–45	10–45	10–45	10–45	10–45	10–45	10–45	10–45	10–45	10–45	10–45	10–45	10–45	10–45
Optimum NaCl for growth (%, w/v)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Nacl range for growth (%, w/v)	0–15	0–12	0–15	0–15	0–12	0–12	0–15	0–10	0–12	0–10	0–15	0–12	0–15	0–15	0–15	0–8	0–6
Optimum pH for growth	7	7	7	7	7	7	7	7	8	7	7	7	7	7	7	NT	NT
pH range for growth	4–10 (48 h)	4–9	4–9	4–9	4–9	4–9	4–9*	4–9*	4–9*	4–9*	4–9*	4–9*	4–9	4–9	4–9	NT	NT
API 20E results:																	
Ornithine decarboxglase	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	–	–
Citrate uilization	–	–	–	–	–	–	+	+	–	–	–	–	–	–	–	–	–
H_2S production	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–	+	+
Indole production	+	+	+	+	–	–	–	+	–	+	+	+	+	+	+	+	+
Amydalin	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	+
API 20NE test results:																	
Aesculin ferric citrate	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	+
Maltose	+	+	+	+	+	+	–	+	+	+	+	+	+	+	+	+	+
API 50CH test results:																	
D-Arabinose	+	+	+	+	+	+	+	+	+	+	+	+	–	–	–	+	+
Xylose	+	+	+	+	+	+	+	+	–	+	+	+	+	+	+	+	+
D-Fructose	+	+	+	+	+	+	+	+	+	+	–	+	–	–	–	+	+
L-Rhamnose	–	–	–	–	–	–	–	–	–	–	+	–	+	–	–	–	–
Methyl α-D-glucopyranoside	+	+	+	+	+	+	–	+	–	+	+	+	–	–	–	+	+
Salicin	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	+
D-Saccharine	+	+	+	+	+	+	–	+	+	+	+	+	+	–	+	+	+
Trehalose	+	+	+	+	+	+	+	+	–	–	–	+	–	–	–	+	+
Melezitose	+	+	+	+	–	+	–	–	+	–	–	+	–	–	–	+	+
Turanose	+	+	+	+	+	+	–	+	+	+	–	+	–	–	–	+	+
Potassium 2-ketogluconate	–	–	–	+	–	+	+	+	–	+	–	+	–	–	–	+	+
API ZYM test results:																	
Esterase	+	–	–	–	–	–	+	+	+	+	+	+	+	+	+	+	+
Esterase lipase	–	+	–	–	–	–	–	–	–	–	–	–	+	+	+	–	+
Lipase	–	–	–	–	–	–	–	–	–	–	–	–	–	+	–	–	–
Valine arylamidase	+	–	–	–	–	–	+	–	–	–	–	–	–	–	–	+	–

Εικόνα 2. Φαινοτυπικά χαρακτηριστικά των διαφόρων ειδών *Proteus.* Πάρθηκε από το άρθρο "Proteus faecis sp. nov., and Proteus cibi sp. nov., two new species isolated from food and clinical samples in China" των Dai et al., το 2019.

Στις δεκαετίες που ακολούθησαν, τεχνικές όπως DNA-DNA υβριδισμός και σύγκριση ποσοστού GC διευκόλυναν τη σωστή ταξινόμηση των βακτηρίων στα διάφορα είδη. Για την τοποθέτηση στελεχών στο ίδιο είδος θεωρήθηκε ότι πρέπει τα στελέχη να έχουν τουλάχιστον 70% DNA-DNA ομοιότητα και μικρότερη από 5% διαφορά στο άθροισμα (G+C) (Wayne et al., 1987). Πλέον στην εποχή της βιοπληροφορικής, ο *in silico* DNA-DNA υβριδισμός έχει αντικαταστήσει τον παραδοσιακό DNA-DNA υβριδισμό (Auch et al., 2010).
Παρακάτω βλέπουμε την σχέση μεταξύ των διαφόρων ειδών του *Proteus* σε ένα φυλογενετικό δέντρο, που έχει προκύψει από την ανάλυση των νουκλεοτιδικών αλληλουχιών του 16S rRNA γονιδίου διαφόρων ειδών του *Proteus*.

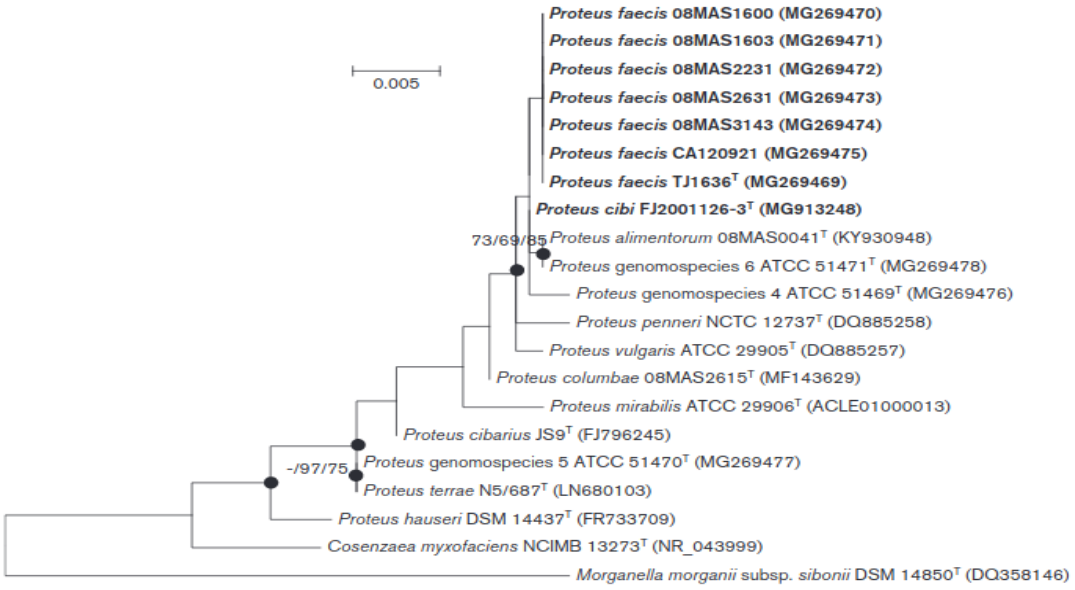


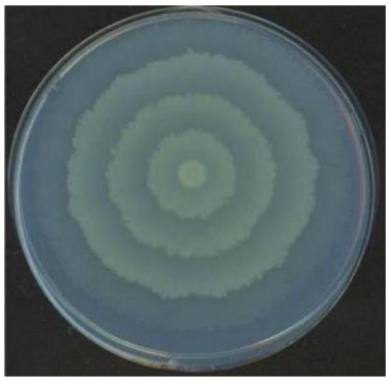
Fig. 1. Phylogenetic tree reconstructed by the maximum-likelihood, neighbour-joining and maximum-parsimony methods based on the nucleotide sequences of the 16S rRNA gene. The phylogenetic tree was reconstructed from an alignment of 1380 nt and 11 known type strains. Bootstrap values were calculated from 1000 replications and values>70 % are shown at branch points. All nodes marked with a dot (•) were also recovered in the three trees. Bar, 0.005 substitutions per nucleotide position. Bold font indicates *P. cibi* and *P. faecis*. GenBank accession numbers of 16S rRNA gene sequences are in brackets. *Morganella morganii* subsp. *sibonii* DSM 14850^T (GenBank accession No. DQ358146) was used as an outgroup.

Εικόνα 3. Φυλογενετικό δέντρο των διαφόρων ειδών *Proteus* όπως προέκυψε από την ανάλυση του 16S rRNA γονιδίου. Πάρθηκε από το άρθρο "Proteus faecis sp. nov., and Proteus cibi sp. nov., two new species isolated from food and clinical samples in China" των Dai et al., το 2019**.**

Παρόλο που η ανάλυση του 16S rRNA γονιδίου είναι η πιο συχνή μέθοδος για διαχωρισμό των βακτηριακών ειδών και παραδοσιακά στις φυλογενετικές μελέτες χρησιμοποιούνται τα γονίδια του rRNA για τη ταξινόμηση των βακτηριακών ειδών, στην ομοταξία των *Enterobacterales* η χρήση του 16S rRNA παρουσιάζει ορισμένα προβλήματα. Στην ομοταξία *Enterobacterales* το γονίδιο αυτό παρουσιάζεται πολύ συντηρημένο και έτσι ορισμένα κοντινά εξελικτικά βακτηριακά είδη δεν μπορούν να διαχωριστούν (Giammanco et al., 2011). Για παράδειγμα με τη χρήση του γονιδίου 16S rRNA δεν είναι εφικτός ο διαχωρισμός των ειδών *Proteus vulgaris* και *Proteus penneri* (Cao et al., 2009). Γι' αυτό αναζητήθηκαν άλλα γονίδια για τη χρήση τους για φυλογενετικές αναλύσεις, όπως είναι το γονίδιο rpoB από τους Giammanco et al. το 2011.
Το βακτήριο *P. mirabilis* είναι το περισσότερο μελετημένο βακτήριο του γένους *Proteus.* Ο *P. mirabilis* συναντάται σε διάφορα μέρη στο περιβάλλον, μπορεί να βρεθεί στο νερό και στο έδαφος και ανήκει στη φυσιολογική χλωρίδα του εντέρου του ανθρώπου και άλλων θηλαστικών. Στον άνθρωπο είναι ευκαιριακά παθογόνο, και οι πιο συχνές λοιμώξεις που προκαλεί είναι οι ουρολοιμώξεις (πιο συγκεκριμένα κυστίτιδες, πυελονεφρίτιδες, προστατίτιδες), και σπάνια προκαλεί λοιμώξεις του αναπνευστικού, του δέρματος, και

βακτηριαιμίες. Ο *P. mirabilis* ευθύνεται για το 3-5% των ουρολοιμώξεων στον άνθρωπο και για το 90% των λοιμώξεων που προκαλούνται από βακτήρια του γένους *Proteus* (Manos et al., 2006). Στην οικογένεια *Morganellaceae,* λίγα μόνο είδη προκαλούν ευκαιριακές λοιμώξεις στον άνθρωπο, πιο συγκεκριμένα τα: *Proteus hauseri*, *Proteus mirabilis*, *Proteus myxofaciens*, *Proteus penneri*, *Proteus vulgaris*, *Morganella morganii, Providencia alcalifaciens*, *Providencia heimbachae*, *Providencia rettgerii*, *Providencia rustigianii*, *Providencia stuartii* (Gajdács et al., 2019), ενώ δεν έχει εξακριβωθεί ακόμα εάν η *Moellerella wisconsensis* ευθύνεται για λοιμώξεις στον άνθρωπο *(*Aller et al., 2009).

Δύο πολύ σημαντικά χαρακτηριστικά του *P. mirabilis* είναι ο ερπυσμός και η ικανότητα να διασπά την ουρία. Ο ερπυσμός αφορά στην ικανότητα του βακτηρίου να μετακινείται σε στερεές επιφάνειες σαν ένα οργανωμένο σύνολο. Αυτή η δυνατότητα του βακτηρίου, δημιουργεί στα τρυβλία, όπου καλλιεργείται το βακτήριο, ένα χαρακτηριστικό μοτίβο, που ονομάζεται bull's-eye pattern (Schaffer et al., 2015). Επίσης ο *P. mirabilis* διαθέτει το κυτταροπλασματικό μεταλλοένζυμο ουρέαση, το οποίο διασπά την ουρία σε αμμωνία και διοξείδιο του άνθρακα, με αποτέλεσμα όταν υπάρχει στα ούρα του ανθρώπου να αυξάνει το pH των ούρων και να προδιαθέτει στο σχηματισμό ουρόλιθων (Jones et al., 1988).



Εικόνα 4. Αποικία του βακτήριο *P. mirabilis*, με χαρακτηριστικό το μοτίβο bull's eye pattern. Πάρθηκε από το άρθρο "*Proteus mirabilis* and Urinary Tract Infections" των Jones et al., 1988**.**

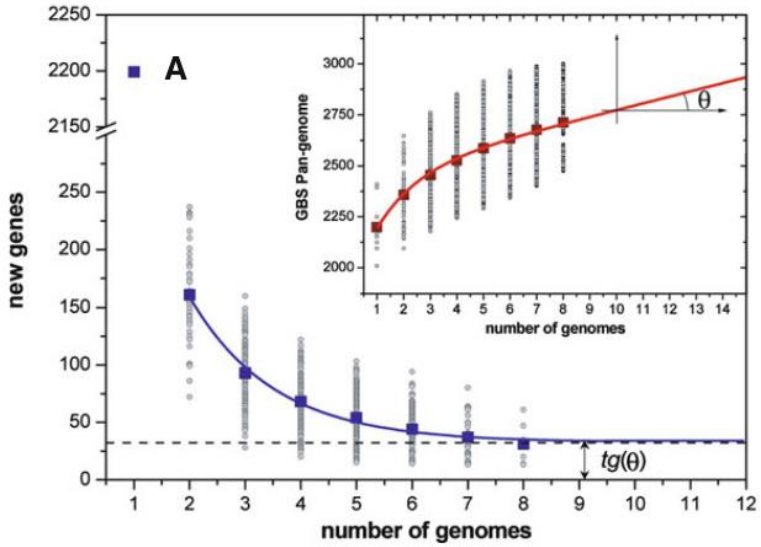
Το 2008 ολοκληρώθηκε πλήρως το γονιδίωμα του *P. mirabilis* από την Melanie M. Pearson. Ο *P. mirabilis* διαθέτει πάνω από 3.658 κωδικές περιοχές με 7 rRNA οπερόνια. Το συνολικό μέγεθος του γονιδιώματος του είναι 4.063 Mb με 28,8 % αναλογία GC (Pearson et al., 2008).

Η παγκόσμια αύξηση των ανθεκτικών στελεχών σε συνδυασμό με την τάση του βακτηρίου αυτού να δημιουργεί λίθους του ουροποιητικού και να φωλιάζει στο εσωτερικό τους, καθιστά την αντιμετώπιση των ουρολοιμώξεων από *P. mirabilis* ιδιαίτερα δύσκολη (Schaffer et al., 2015). Το βακτήριο *P. mirabilis* έχει φυσική αντοχή σε τετρακυκλίνη, τιγεκυκλίνη και κολιστίνη (EUCAST, 2020). Όσον αφορά στην επίκτητη αντοχή του βακτηρίου στα αντιβιοτικά, τα τελευταία χρόνια παγκοσμίως φαίνεται να υπάρχει μία αύξηση των ανθεκτικών στελεχών *P. mirabilis* σε διάφορες τάξεις αντιβιοτικών, όπως β-λακταμικά, αμινογλυκοσίδες και φθοριοκινολόνες (Girlich et al., 2020). Αν και δεν είναι ιδιαίτερα συχνές οι επιδημίες από πολυανθεκτικά στελέχη *P. mirabilis,* το 2015-2016 απομονώθηκαν από 14 νοσηλευόμενους ασθενείς στην Ελλάδα 27 πολυανθεκτικά στελέχη *P. mirabilis,* που παρήγαγαν την VIM-4 μεταλλο-β-λακταμάση (MBL) και την TEM-2 β-λακταμάση (Protonotariou et al., 2019). Το 2004-2005 στην Ιαπωνία παρατηρήθηκε μία έξαρση από ανθεκτικά στελέχη *P. mirabilis* στην κεφοταξίμη (Nakano et al., 2012), και στην Ινδία παρατηρήθηκε μία επιδημία σε νεογνά της μονάδας εντατικής θεραπείας από στελέχη *P. mirabilis* που παρήγαγαν ευρέου φάσματος β-λακταμάση (ESBL) (Jain et al., 2016).

1.2. Πανγονιδίωμα

Η ανάπτυξη και η χρήση της τεχνολογίας αλληλούχισης νέας γενιάς (next-generation sequencing) έκανε εφικτή τη δυνατότητα σύγκρισης πολλών διαφορετικών γονιδιωμάτων του ίδιου βακτηριακού είδους. Έτσι παρατηρήθηκε ότι υπήρχαν διαφορές στο γονιδίωμα των διαφόρων στελεχών του ίδιου βακτηρίου. Ο όρος “pangenome”, πανγονιδίωμα, αναφέρθηκε πρώτη φορά από τους Tettelin et al. το 2005, καθώς μελετούσαν 8 στελέχη του βακτηρίου *Streptococcus agalactiae*. Ο ορισμός που δώσανε για το πανγονιδίωμα είναι ότι το πανγονιδίωμα ενός βακτηριακού είδους περιλαμβάνει το **core genome**, δηλαδή τα γονίδια τα οποία συναντώνται σε όλα τα στελέχη του είδους αυτού, το **dispensable genome**, δηλαδή τα γονίδια που συναντώνται σε παραπάνω από ένα στελέχη του είδους αυτού αλλά όχι σε όλα τα στελέχη, και το **unique genome** (strain-specific genome), δηλαδή τα γονίδια που βρίσκονται μόνο σε ένα στέλεχος (Tettelin et al., 2005).

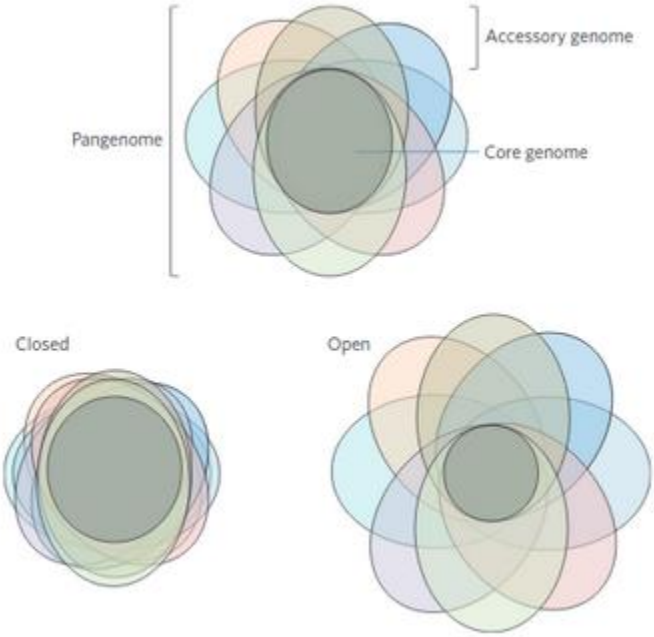
Η πολύ σημαντική τομή που έφερε η εργασία των Tettelin et al. ήταν ότι ανέτρεψε την μέχρι τότε θεώρηση ότι το γονιδίωμα ενός στελέχους ενός βακτηριακού είδους είναι ενδεικτικό για το γονιδίωμα όλων των στελεχών του είδους αυτού. Για την πλήρη αποκάλυψη του γονιδιώματος ενός είδους απαιτείται η αλληλούχιση των γονιδιωμάτων πολλών στελεχών του και η μεταξύ τους σύγκριση. Η αμέσως επόμενη ερώτηση που προσπάθησαν να απαντήσουν οι Tettelin et al. ήταν ποιος είναι ο απαιτούμενος αριθμός των στελεχών για την πλήρη εύρεση του γονιδιώματος ενός βακτηριακού είδους. Εδώ έρχεται και η δεύτερη μεγάλη τομή της μελέτης τους, που είναι ότι για το βακτήριο *Streptococcus agalactiae*, που μελετούσαν, όσο περισσότερα στελέχη αλληλουχούνταν, τόσο αύξανε το πανγονιδίωμα του και συνεχώς νέα γονίδια προστίθονταν.



Εικόνα 5. Μια αναπαράσταση των νέων γονιδίων (μπλε γραμμή) και του πανγονιδιώματος (κόκκινη γραμμή), σε σχέση με τον αριθμό των γονιδιωμάτων που έχουν αναλυθεί. ‘Οσο αυξάνεται ο αριθμός των γονιδιωμάτων, ο αριθμός των νέων γονιδίων μειώνεται αλλά σταθεροποιείται πολύ πάνω από την τιμή μηδέν. Στην εικόνα πάνω δεξιά βλέπουμε και την ανοδική τάση του πανγονιδιώματος του *Streptococcus agalactiae* με την προσθήκη στην ανάλυση επιπλέον γονιδιωμάτων. Πάρθηκε από το βιβλίο “The Pangenome DIversity, Dynamics and Evolution of Genomes” των Tettelin και Medini.

Με βάση την παραπάνω εικόνα καταλαβαίνουμε ότι το πανγονιδίωμα του *S. agalactiae* δεν μπορεί να προβλεφθεί ολοκληρωτικά καθώς όσο περισσότερα γονιδιώματα στελεχών του αλληλουχούνται τόσο θα αυξάνεται και το πανγονιδίωμα του. Αυτό το πανγονιδίωμα ονομάστηκε ανοικτό (open pangenome), σε αντίθεση με το κλειστό πανγονιδίωμα, στο οποίο μετά από ένα συγκεκριμένο αριθμό στελεχών που έχουν αλληλουχηθεί, όσα επιπλέον στελέχη και να αλληλουχηθούν, δεν προστίθενται νέα γονίδια (Tettelin και Medini 2020).

Παρακάτω παρουσιάζεται μία σχηματική αναπαράσταση του πανγονιδιώματος, του ανοικτού πανγονιδιώματος και του κλειστού πανγονιδιώματος. Στις περιπτώσεις του κλειστού πανγονιδιώματος, το unique genome είναι πολύ μικρότερο και δεν υπάρχει τόσο μεγάλη διαφορά μεγέθους μεταξύ του πανγονιδιώματος και του core-genome.



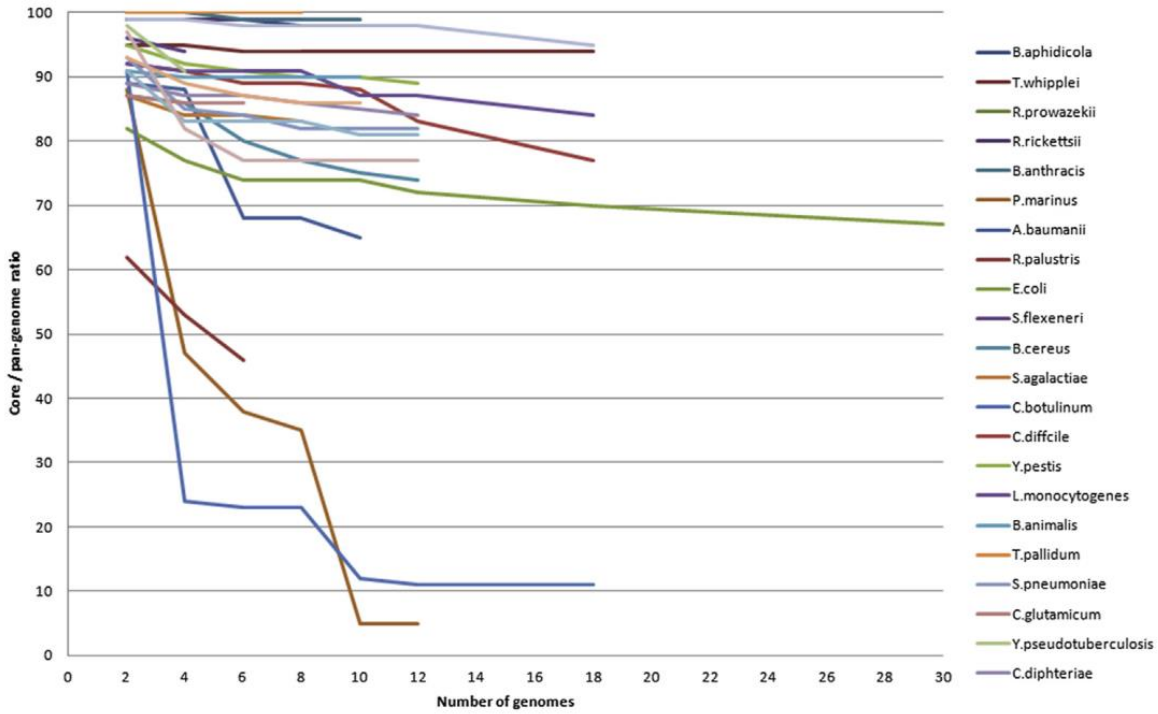
Εικόνα 6. Σχηματική αναπαράσταση του ανοικτού/κλειστού πανγονιδιώματος. Πάρθηκε από το άρθρο "Why prokaryotes have pangenomes" των Mclnerney et al., 2017.

Από το 2005 μέχρι το 2013, είχε δημοσιευθεί το πανγονιδίωμα από 41 βακτηριακά είδη και 12 βακτηριακά γένη. Παρακάτω παρουσιάζονται τα βακτηριακά είδη που έχει δημοσιευθεί το πανγονιδίωμά τους μέχρι το 2013 (Rouli et al., 2015).

Πίνακας 2. Περιλαμβάνει τα βακτηριακά είδη των οποίων το πανγονιδίωμά τους έχει δημοσιευθεί μέχρι το 2013.

Βακτηριακά είδη με δημοσιευμένο πανγονιδίωμα					
Escherichia coli	*Streptococcus pneumoniae*	*Salmonella enterica*	*Staphylococcus aureus*	*Helicobacter pylori*	*Mycobacterium tuberculosis*
Vibrio cholerae	*Yersinia pestis*	*Bacillus cereus*	*Chlamydia trachomatis*	*Streptococcus pyogenes*	*Acinetobacter baumannii*
Enterococcus faecium	*Clostridium difficile*	*Francisella tularensis*	*Pseudomonas aeruginosa*	*Haemophilus influenzae*	*Listeria monocytogenes*
Bacillus anthracis	*Clostridium botulinum*	*Buchnera aphidicola*	*Streptococcus agalactiae*	*Legionella pneumophila*	*Campylobacter jejuni*
Bifidobacterium animalis	*Streptococcus suis*	*Ralstonia solanacearum*	*Sinorhizobium meliloti*	*Prochlorococcus marinus*	*Actinobacillus pleuropneumoniae*
Lactobacillus casei	*Coxiella burnetii*	*Erwinia amylovora*	*Corynebacterium pseudotuberculosis*	*Rhodopseudomonas palustris*	*Aggregatibacter actinomycetemcomitans*
Salmonella paratyphi	*Staphylococcus epidermidis*	*Oenococcus oeni*	*Corynebacterium diphtheriae*	*Tropheryma whipplei*	

Παρακάτω βλέπουμε ένα διάγραμμα για το λόγο core/pangenome ratio, δηλαδή το λόγο αριθμό γονιδίων του core-genome προς τον αριθμό γονιδίων του pangenome, για 27 βακτηριακά είδη. Εάν ο λόγος core/pangenome φτάσει σε σταθερό επίπεδο (plateau), τότε το πανγονιδίωμα του βακτηριακού είδους είναι κλειστό. Από την παρακάτω εικόνα συμπεραίνουμε εύκολα ότι για παράδειγμα το πανγονιδίωμα του *B. anthracis* είναι κλειστό (Rouli et al., 2015).



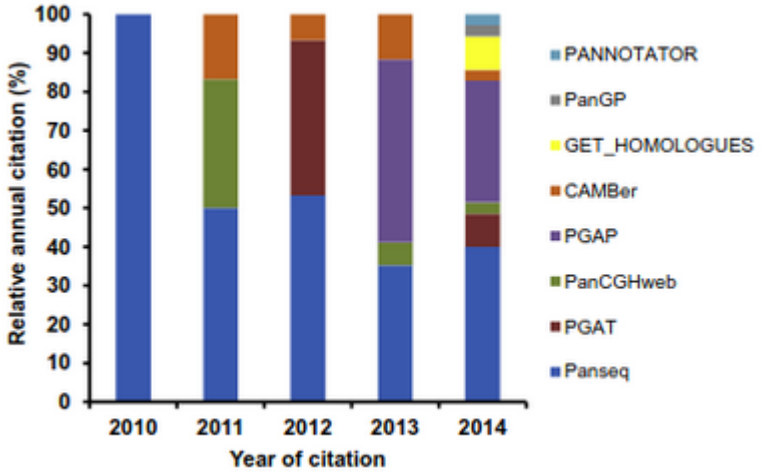
Εικόνα 7. Παρουσιάζεται για διάφορα βακτήρια ο λόγος core/pangenome σε σχέση με τον αριθμό των αναλυμένων γονιδιωμάτων. Πάρθηκε από το άρθρο “The bacterial pangenome as a new tool for analysing pathogenic bacteria” των Rouli et al., 2015

1.3. Εφαρμογές και εργαλεία για την ανάλυση του πανγονιδιώματος

Οι εφαρμογές του πανγονιδιώματος είναι πολλές και αφορούν διάφορους επιστημονικούς κλάδους. Αρχικά με την εύρεση του unique genome, γίνεται εφικτός ο εντοπισμός και ο φαινοτυπικός χαρακτηρισμός ενός συγκεκριμένου στελέχους του βακτηριακού είδους, για παράδειγμα η ανεύρεση παθογονικών παραγόντων ή γονιδίων αντοχής που υπάρχουν σε ένα συγκεκριμένο βακτηριακό στέλεχος. Η χρήση του unique genome ή του dispensable genome είναι ιδιαίτερα σημαντική τόσο για την ανάπτυξη εμβολίων όσο και φαρμάκων στοχευμένων σε συγκεκριμένα στελέχη του βακτηριακού είδους. Πολλά βακτηριακά είδη συμβιώνουν με τον άνθρωπο και μόνο ορισμένα στελέχη τους είναι παθογόνα για τον άνθρωπο. Τα συμβιωτικά στελέχη μοιράζονται το core genome με τα παθογόνα στελέχη, έτσι στοχεύοντας στο unique genome ή στο dispensable genome των παθογόνων στελεχών είτε με φάρμακα είτε με εμβόλια, δεν επηρεάζουμε την φυσιολογική βακτηριακή χλωρίδα, αλλά συνάμα αντιμετωπίζουμε και τις πιθανές λοιμώξεις από τα παθογόνα στελέχη (Innamorati et al., 2020).

Επίσης, το πανγονιδίωμα των βακτηρίων μπορεί να χρησιμοποιηθεί σε μελέτες για τη διευρεύνηση πιθανής οριζόντιας μεταφοράς γονιδίων μεταξύ διαφόρων στελεχών και βακτηριακών ειδών. Μπορεί να βρει επίσης εφαρμογή σε μελέτες γονιδιώματος και παθογονικότητας, αποκαλύπτοντας σχέσεις μεταξύ γονιδίων και παθογονικών χαρακτηριστικών. Ενώ μπορεί να χρησιμοποιηθεί και σε μελέτες σχετικά με τα απαιτούμενα

για την επιβίωση του κάθε βακτηρίου μεταβολικά γονίδια, καθώς το core genome αφορά σε απαραίτητα γονίδια για το κάθε βακτηριακό είδος.
Τέλος, μία πολύ σημαντική εφαρμογή του πανγονιδιώματος / core genome είναι ο έλεγχος επιδημιών και η επιδημιολογική επιτήρηση των διαφόρων στελεχών των βακτηρίων. Με τη χρήση του core genome αποκαλύφθηκε ότι η επιδημία χολέρας, που ξέσπασε στη Αϊτή το 2010, οφείλοταν σε ένα μόνο στέλεχος από την Νοτιοανατολική Ασία (Reimer et al., 2011). Με τη χρήση του core genome μπορεί να αποκαλυφθούν με μεγαλύτερη ακρίβεια οι εξελικτικές σχέσεις μεταξύ διαφόρων στελεχών και έτσι να γίνεται παρακολούθηση των στελεχών. Οι Chen et al., το 2016 χρησιμοποίησαν το core genome για να παρακολουθήσουν και να ομαδοποιήσουν σε ομάδες, τα στελέχη *Listeria monocytogenes* που προκάλεσαν επιδημίες σε διάφορα τρόφιμα. Επίσης, οι Venditti et al., χρησιμοποίησαν το core genome για να μελετήσουν την έξαρση ανθεκτικών στην καρβαπενεμάση, *Acinetobacter baumannii* στη Μονάδα Εντατικής Θεραπείας, και παρατήρησαν ότι τα στελέχη αυτά ομαδοποιούνταν σε δύο ομάδες, το οποίο υποδεικνύει δύο διαφορετικές πηγές διασποράς των ανθεκτικών στελεχών. Τέλος, οι Pearce et al., στο άρθρο τους το 2018 κατέληξαν στο συμπέρασμα ότι το core genome μπορεί να χρησιμοποιηθεί για την ανάλυση επιδημιών από *Salmonella*.
Από το 2010 και μετά, έχουν δημιουργηθεί διάφορα βιοπληροφορικά εργαλεία που αναλύουν δεδομένα πανγονιδιώματος. Ανάλογα με την επιθυμητή ανάλυση, τα δεδομένα εισόδου και εξόδου, αλλά και τη μεθοδολογία, υπάρχουν πλέον πολλές επιλογές εργαλείων. Παρακάτω βλέπουμε το ποσοστό σε αναφορές για τα διάφορα βιοπληροφορικά εργαλεία μέχρι το 2014. Μέχρι το 2014 το Panseq και το PGAP ήταν τα δύο προγράμματα με τις περισσότερες αναφορές, πλέον το πρόγραμμα με τις περισσότερες αναφορές είναι το Roary.



Εικόνα 8. Εκατοστιαίο ποσοστό βιβλιογραφικών αναφορών για τα διάφορα βιοπληροφορικά εργαλεία για την ανάλυση του πανγονιδιώματος από το 2010-2014. Πάρθηκε από το άρθρο “A brief Review of Software Tools for Pangenomics” των Xiao et al., 2015.

Το πρόγραμμα Panseq δημοσιεύτηκε το 2010 και βρίσκει το πανγονιδίωμα, το core genome, το dispensable genome και το unique genome, επίσης εντοπίζει και σημειακές μεταλλάξεις (SNPs) στο core genome. Δέχεται ως δεδομένα εισόδου ολόκληρα γονιδιώματα, και το core genome που δημιουργεί είναι ουσιαστικά οι περιοχές του γονιδιώματος που είναι παρόμοιες σε όλα τα γονιδιώματα που έχουν εισαχθεί στο πρόγραμμα. Είναι ελεύθερα διαθέσιμο (Laing et al., 2010).
Το εργαλείο Pan-Genomes Analysis Pipeline (PGAP) δημοσιεύτηκε το 2012 και η πιο εξελιγμένη μορφή του, το PGAP-X το 2018. Το PGAP υποστηρίζει την εύρεση του πανγονιδίωματος, αλλά και άλλες επιπλέον αναλύσεις, όπως ανάλυση της εξέλιξης των ειδών και λειτουργική ανάλυση των γονιδίων του πανγονιδιώματος. Το PGAP-X είναι πιο εξελιγμένο ως προς την ανάλυση και την παρουσίαση των αποτελεσμάτων. Τόσο το PGAP όσο και το PGAP-X είναι ελεύθερα προσβάσιμα. Τα δύο αυτά εργαλεία

βρίσκουν το core genome ως το σύνολο των γονιδίων που μοιράζονται από όλα τα στελέχη, και σαν δεδομένα είσοδο δέχονται ολόκληρα γονιδιώματα (Zhao et al., 2012).
Το εργαλείο PanOCT δημοσιεύτηκε το 2012 και βρίσκει το πανγονιδίωμα για κοντινούς εξελικτικά προκαρυωτικούς οργανισμούς, καθώς χρησιμοποιεί ως πληροφορία για την ομαδοποίηση των γονιδίων και την συντήρηση των περιοχών που γειτνιάζουν με τα αντίστοιχα γονίδια. Λόγω του ότι κάνει και all-versus-all συγκρίσεις των γονιδιωμάτων είναι αρκετά πιο αργό και απαιτεί περισσότερη υπολογιστική ισχύ από τα άλλα προγράμματα. Είναι και αυτό ελεύθερα διαθέσιμο (Fouts et al., 2012).
Το εργαλείο LS-BSR δημοσιεύτηκε το 2014, και επιτρέπει την γρήγορη σύγκριση του γενετικού περιεχομένου πολλών βακτηριακών γονιδιωμάτων μεταξύ τους. Το πρόγραμμα αυτό δεν βρίσκει το πανγονιδίωμα, αλλά επιτρέπει την γρήγορη σύγκριση των κωδικών περιοχών κάθε γονιδιώματος μεταξύ τους. Με την παραπάνω πληροφορία βεβαία είναι εφικτό να δημιουργηθεί στη συνέχεια το πανγονιδίωμα ή να δημιουργηθεί ένα φυλογενετικό δέντρο. Είναι και αυτό ελεύθερα προσβάσιμο (Sahl et al., 2014).
Το λογισμικό πακέτο Harvest δημοσιεύθηκε το 2014 και επιτρέπει την ανάλυση πάρα πολλών βακτηριακών γονιδιωμάτων και την εύρεση του core genome σε σύντομο χρονικό διάστημα. Το Harvest βρίσκει το core genome ως τις περιοχές του γονιδιώματος που είναι παρόμοιες σε όλα τα γονιδιώματα που έχουν εισαχθεί σε αυτό, στη συνέχεια μπορεί να στοιχίζει τις περιοχές του core genome των διαφόρων στελεχών και οπτικοποιεί τα αποτελέσματα. Επίσης επιτρέπει τη δημιουργία φυλογενετικών δέντρων με τη χρήση single nucleotide polymorphisms (SNPs) που έχουν προκύψει από το core genome. Είναι ελεύθερα διαθέσιμο (Treangen et al., 2014)
Το εργαλείο Roary δημοσιευθηκε το 2015 και επιτρέπει την κατασκευή μεγάλων πανγονιδιωμάτων σε σύντομο χρονικό διάστημα, εντοπίζοντας το core genome, δηλαδή τα γονίδια που υπάρχουν σε όλα τα στελέχη, και το dispensable genome, δηλαδή τα γονίδια που βρίσκονται σε ορισμένα μόνο στελέχη από το σύνολο των βακτηριακών στελεχών. Σαν είσοδο δέχεται το σχολιασμένο γονιδίωμα (annotated assembly) για κάθε στέλεχος. Το Roary είναι αρκετά γρήγορο και ελεύθερα διαθέσιμο (Page et al., 2015).
Τέλος, το πρόγραμμα Bacterial Pangenome Analysis (BGPA) δημοσιεύτηκε το 2016, και δίνει τη δυνατότητα για ανάλυση του πανγονιδιώματος αλλά και επιπλέον δυνατότητες, όπως φυλογένεση με τη χρήση του core genome, εύρεση γονιδίων που λείπουν σε συγκεκριμένα στελέχη και άλλες δυνατότητες. Και σε αυτό το πρόγραμμα ως core genome θεωρείται το σύνολο από τα γονίδια που συναντώνται σε όλα τα στελέχη που έχουν εισαχθεί. Είναι ιδιαίτερα γρήγορο και είναι ελεύθερα προσβάσιμο (Chaudhari et al., 2016).

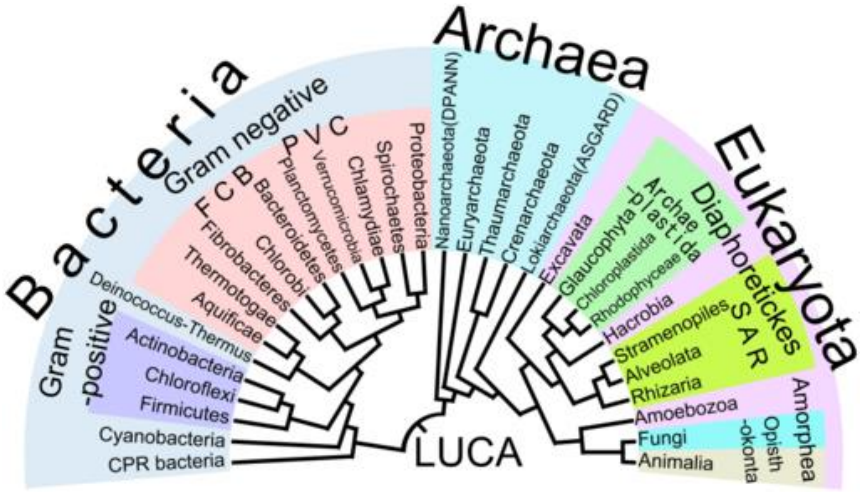
1.4. MLST και core genome MLST

Το Multilocus Sequence Typing (MLST) είναι μία τεχνική, που προτάθηκε πρώτη φορά το 1998, για τον χαρακτηρισμό πολλαπλών γενετικών τόπων (loci), ώστε να είναι εφικτή η γενετική σύγκριση / τυποποίηση των διαφόρων βακτηριακών στελεχών ενός είδους. Η τεχνική αυτή βασίζεται στο χαρακτηρισμό των διαφόρων στελεχών χρησιμοποιώντας μικρά τμήματα (loci) από ορισμένα συντηρημένα γονίδια. Από κάθε γονίδιο χρησιμοποιούνται περιοχές (loci) μήκους 450-500bp, και χρησιμοποιούνται γονίδια, τα οποία είναι πολύ συντηρημένα και ως εκ τούτου υπάρχουν υποχρεωτικά στα στελέχη του είδους (housekeeping genes) που μελετάμε. Έτσι για κάθε στέλεχος που θέλουμε να χαρακτηρίσουμε, αλληλουχείται η περιοχή από κάθε γονίδιο που έχει επιλεγεί, και στη συνέχεια μετά από σύγκριση με την MLST database, η οποία περιέχει όλες τις αλληλουχίες που έχουν κατατεθεί για το συγκεκριμένο γονίδιο, η αλληλουχία παίρνει έναν αριθμό που αντιστοιχεί σε συγκεκριμένο αλλήλιο του εξεταζόμενου γονιδίου. Σε περίπτωση που η αλληλουχία διαφέρει από τις ήδη κατατεθειμένες θεωρείται ότι πρόκειται για νέο αλλήλιο του γονιδίου και της δίνεται ένας καινούργιος αριθμός. Για κάθε στέλεχος το σύνολο των αλληλίων του για όλα τα γονίδια που εξετάζονται στο MLST σχήμα είναι το allelic profile του ή το multilocus sequence type (ST) του (Maiden et al., 1998). Ο πρώτος οργανισμός για τον οποίο δημιουργήθηκε MLST είναι η *Neisseria meningitis*, αλλά

στη συνέχεια το MLST χρησιμοποιήθηκε για διάφορα βακτήρια και ευκαρυωτικούς οργανισμούς, με στόχο την επιδημιολογική επιτήρηση των διαφόρων παθογόνων (Belen et al., 2014).
Συνήθως για το MLST ενός βακτηριακού είδους χρησιμοποιούνται περιοχές 6 ή 7 καλά συντηρημένων γονιδίων (Dingle και MAcCannel, 2015). Η χρήση της γενετικής πληροφορίας από μόνο 6-7 γονίδια περιορίζει την διακριτική ικανότητα της μεθοδου, έτσι σήμερα στην εποχή του Next-Generation Sequencing όπου είναι δυνατή η αλληλούχιση ολόκληρων των γονιδιωμάτων των βακτηρίων, εισήχθη η έννοια του core genome MLST από τους Maiden et al., το 2013. Το core genome MLST (cgMLST) επεκτείνει την τεχνική του MLST, καθώς κάνει χρήση πολλών περισσότερων γενετικών τόπων για την τυποποίηση των διαφόρων στελεχών (Pearce et al., 2018). Για το cgMLST μπορούν να χρησιμοποιηθούν τα γονίδια που υπάρχουν στο 95% (Bratcher et al., 2014), στο 99% (Moura et al., 2016) ή και στο 100% των στελεχών που αναλύονται. Έτσι για ένα βακτηριακό στέλεχος, ο cgMLST τύπος του θα αποτελείται από το κάθε αλλήλιό του για κάθε γονίδιο του core genome. Το cgMLST φιλοδοξεί να αυξήσει τη διακριτική ικανότητα του MSLT για βακτηριακά στελέχη εξελικτικά κοντινά μεταξύ τους, ώστε να είναι δυνατή η ακριβέστερη τυποποίησή τους.

1.5. Φυλογένεση

Η φυλογένεση είναι ο κλάδος της βιολογίας που ασχολείται με την αποκάλυψη των εξελικτικών σχέσεων μεταξύ διαφόρων οργανισμών. Τα ευρήματα της φυλογενετικής ανάλυσης παρουσιάζονται σε ένα φυλογενετικό δέντρο που ουσιαστικά είναι ένα διάγραμμα που δείχνει τις εξελικτικές σχέσεις μεταξύ των οργανισμών ή γενικότερα των ταξινομικών βαθμίδων (taxa) που συγκρίνονται (Roy et al., 2014). Παρακάτω παρουσιάζεται ένα φυλογενετικό δέντρο, με τις κυριότερες ομάδες οργανισμών.



Εικόνα 9. Το δέντρο της ζωής, που δείχνει τις εξελικτικές σχέσεις μεταξύ των 3 βασιλείων της ζωής. Πάρθηκε απο : http://commons.wikimedia.org/wiki/File:Phylogenetic_Tree_of_Life.png.

Πριν την ανάπτυξη της γονιδιακής αλληλούχισης και της επιστήμης των υπολογιστών, τα φυλογενετικά δέντρα κατασκευάζονταν κυρίως με βάση μορφολογικά χαρακτηριστικά. Πλέον, γονιδιακά δεδομένα αναλύονται με υπολογιστικά βιοπληροφορικά εργαλεία, και χρησιμοποιούνται για τη δημιουργία φυλογενετικών δέντρων.
Τα φυλογενετικά δέντρα μπορεί να είναι δέντρα με ρίζα, όπως το παραπάνω, δηλαδή να μπορούμε να προσδιορίσουμε τον πιο παλιό κοινό πρόγονο, και δέντρα χωρίς ρίζα, δηλαδή να μην μπορουμε να

προσδιορίσουμε την εξελικτική κατεύθυνση στο χρόνο και να μπορούμε μόνο να δούμε τις σχέσεις μεταξύ των οργανισμών αλλά όχι ποιος είναι ο αρχαίος κοινός πρόγονος.
Για τη δημιουργία μιας φυλογενετικής ανάλυσης από γονιδιακά δεδομένα είναι απαραίτητα τέσσερα βασικά βήματα. Αρχικά, μία πολλαπλή στοίχιση των αλληλουχιών, που θα χρησιμοποιήσουμε, είναι απαραίτητη. Στη συνέχεια, πρέπει να γίνει επιλογή του μαθηματικού μοντέλου αντικαταστάσεων. Υπάρχουν διάφορα μοντέλα και η επιλογή εξαρτάται και από τι είδους δεδομένα έχουμε, δηλαδή υπάρχουν διαφορετικά μοντέλα για αλληλουχίες DNA και διαφορετικά για πρωτεϊνικά δεδομένα. Για τις αλληλουχίες DNA, το πιο γενικό μοντέλο είναι το Generalised time reversible (GTR) μοντέλο. Στην συνέχεια, μετά την επιλογή του μοντέλου, πρέπει να επιλεγεί ο αλγόριθμος για την κατασκευή του δέντρου. Υπάρχουν διάφοροι αλγόριθμοι με αρκετά βιοπληροφορικά εργαλεία να εφαρμόζουν κάθε ένα από αυτούς. Οι κυριότεροι αλγόριθμοι είναι : ο αλγόριθμος UPGMA (Unweighted Pair Group using Arithmetic Mean), η μέθοδος της ένωσης γειτόνων (Neighbour Joining, NJ), η μέθοδος μέγιστη φειδωλότητα (maximum parsimony), η μέθοδος της Μέγιστης Πιθανοφάνειας (Maximum Likelihood), και οι Bayesians μέθοδοι. Το τελευταίο βήμα είναι η αξιολόγηση του δέντρου, ώστε να υπολογιστεί η αξιοπιστία του (Bagos 2015).
Τέλος, τα φυλογενετικά δέντρα παρόλο που μας παρέχουν μια εικόνα για τη εξελικτική διαδικασία, μπορεί να αποκλίνουν από την πραγματικότητα, καθώς εξαρτώνται από τις διάφορες παραδοχές που κάνουμε και είναι πολύ ευαίσθητα σε πιθανά σφάλματα της ανάλυσης.

1.6. Σκοπός

Σκοπός της συγκεκριμένης διπλωματικής εργασίας είναι η εύρεση του core genome και του pangenome του βακτηρίου *P. mirabilis* και η διερεύνηση της πιθανής χρήσης του core genome ως εργαλείο για φυλογενετική μελέτη των στελεχών του, με στόχο τη χρήση του για την επιδημιολογική επιτήρηση ανθεκτικών στελεχών που δημιουργούν επιδημίες τόσο τοπικά (π.χ. σε νοσοκομεία ή κλινικές νοσοκομείων) όσο και πιο εκτεταμένα, σε ολόκληρες χώρες ή και ηπείρους. Η ανάπτυξη πιο εξελιγμένων μεθόδων, όπως το cgMLST, για την ακριβέστερη τυποποίηση και φυλογενετική ανάλυση των διαφόρων στελεχών βακτηρίων και ιδιαίτερα των πολυανθεκτικών στελεχών τους είναι απαραίτητη για την καλύτερη κατανόηση της εξέλιξης και της διασποράς τους.
Ένας επιπλέον στόχος ήταν να δημιουργηθεί ένα απλό, αυτοματοποιημένο βιοπληροφορικό εργαλείο που να επιτρέπει στο χρήστη να δημιουργεί το core genome όποιου βακτηριακού είδους επιθυμεί και να μπορεί μετέπειτα να εισάγει γονιδιώματα στελεχών του είδους αυτού ώστε να ελέγχει τη γενετική συγγένειά τους με τα υπόλοιπα στελέχη που είναι κατατεθειμένα στη βάση δεδομένων GenBank.

2. Μέθοδοι

Για τη συλλογή των γονιδιωμάτων και τον χειρισμό των δεδομένων ώστε να βρεθεί το πανγονιδίωμα και το core genome από το βακτήριο *P. mirabilis*, χρησιμοποιήθηκε η γλώσσα προγραμματισμού Perl, η γλώσσα Bash script (sh) και η γλώσσα C shell (csh). Έγινε χρήση των βιοπληροφορικών εργαλείων Prokka (Seemann, 2014) , CD-HIT (Fu et al., 2012**)**, CLUSTAL OMEGA (Sievers και Higgins, 2018), EggNOG-mapper (Huerta-Cepas et al., 2017, Huerta-Cepas et al., 2019), AMRFinder (Feldgardern et al., 2019), RAxML (Stamatakis, 2014), FastTree (Price et al., 2010), Dendroscope (Huson et al., 2012). Οι αναλύσεις πραγματοποιήθηκαν στο υπολογιστικό κέντρο του Ελληνικού Ινστιτούτου Παστέρ χρησιμοποιώντας τους servers του εργαστηρίου Βακτηριολογίας [3x Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30GHz 48 cores]

2.1. Εύρεση αξιόπιστων γονιδιωμάτων

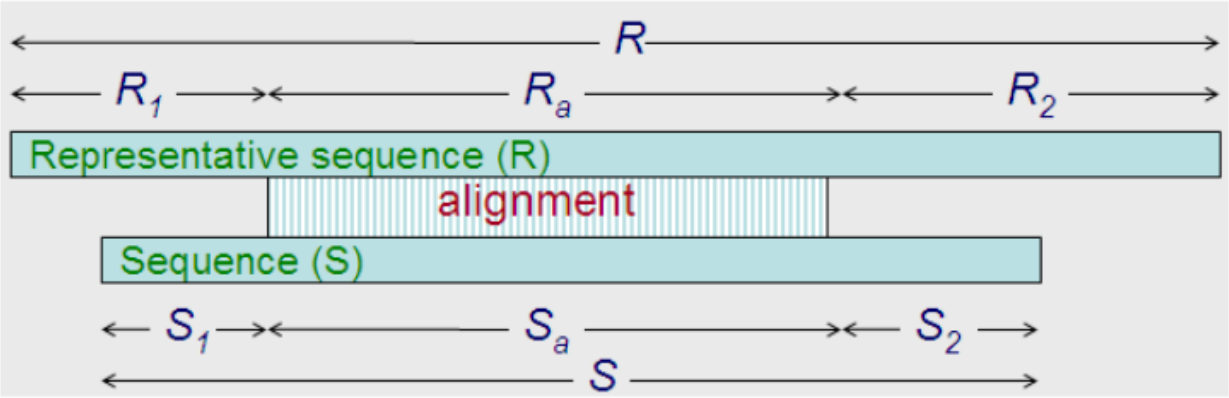
Κατεβάσαμε από τη βάση δεδομένων Genbank του NCBI το αρχείο assembly_summary.txt, το οποίο περιέχει όλα τα βακτήρια που είναι κατατεθειμένα στη βάση αυτή. Απομονώσαμε από το αρχείο αυτό όλα τα ftp αρχεία για βακτήρια τα οποία έχουν ταξινομηθεί στη βάση Genbank ως *P. mirabilis.* Για κάθε γονιδίωμα κατεβάσαμε το fasta αρχείο που έχει αποθηκευμένη ολόκληρη την αλληλουχία του και το αρχείο *assembly_stats.txt*, που περιέχει στοιχεία για την αλληλούχιση του και ορισμένα στατιστικά στοιχεία. Από τα γονιδιώματα της Genbank χρησιμοποιήθηκαν για περαιτέρω ανάλυση τα γονιδιώματα τα οποία πληρούσαν τουλάχιστον ένα από τα εξής κριτήρια:
α) η αλληλούχιση για το γονιδίωμα τους ήταν πλήρως ολοκληρωμένη, δηλαδή το assembly level του γονιδιώματος με βάση το αρχειο *assembly_stats.txt* είναι “Complete Genome”.
β) το γονιδίωμα τους δεν ήταν πλήρως ολοκληρωμένο αλλά με βάση το αρχείο *assembly_stats.txt* είχαν N50>100.000.

2.2. Χαρακτηρισμός των γονιδιωμάτων

Για τα στελέχη, που πληρούσαν τα κριτήρια του προηγουμενου βήματος, αναλύσαμε το γονιδίωμα τους με το βιοπληροφορικό εργαλείο Prokka, το οποίο κάνει σχολιασμό (annotation) γονιδιωμάτων βακτηρίων, αρχαίων και ιών. Για το σχολιασμό των γονιδιωμάτων χρησιμοποιήθηκαν οι προτεινόμενες ρυθμίσεις του Prokka.
Αρχικά, για το χαρακτηρισμό των γονιδιωμάτων, δοκιμάσαμε να χρησιμοποιήσουμε τον χαρακτηρισμό του NCBI για όσα στελέχη υπήρχε κατατεθειμένος ο χαρακτηρισμός τους στο NCBI ενώ για όσα γονιδιώματα δεν είχαν σχολιασμό να χρησιμοποιήσουμε το εργαλείο Prokka. Ο χαρακτηρισμός των γονιδιωμάτων στο NCBI έχει γίνει με τη χρήση του βιοπληροφορικού εργαλείου Prokaryotic Genome Annotation Pipeline (PGAP). Όταν, όμως συγκρίναμε σε γονιδιώματα με σχολιασμό στο NCBI, τον αριθμό των πρωτεϊνών που προέκυψε βάσει του PGAP με τον αριθμό των πρωτεϊνών βάσει του Prokka, υπήρχε αρκετά μεγάλη διαφορά. Οπότε εάν για κάποια γονιδιώματα χρησιμοποιούσαμε το σχολιασμό του NCBI και για κάποια το σχολιασμό του Prokka, πολλές διαφορές μεταξύ των γονιδιωμάτων δεν θα ήταν πραγματικές αλλά θα οφείλονταν στο διαφορετικό σχολιασμό από τα δύο εργαλεία. Γι’ αυτό και δεν κάναμε χρήση του σχολιασμού από το NCBI σε κανένα γονιδίωμα και χρησιμοποιήσαμε σε όλα τα στελέχη που δεχτήκαμε για την ανάλυση μας το εργαλείο Prokka.
Στο τέλος αυτού του βήματος, για κάθε γονιδίωμα έχουμε το σχολιασμό του, δηλαδή τις κωδικές περιοχές (coding region CDS), και τις αντίστοιχες παραγόμενες πρωτεΐνες του κάθε γονιδιώματος.

2.3. Ομαδοποίηση των πρωτεϊνών των στελεχών

Αφού δημιουργήσουμε μία πρώτη βάση δεδομένων που να περιέχει όλες τις πρωτεΐνες των γονιδιωμάτων που έχουμε αποδεχτεί, χρησιμοποιήσαμε το βιοπληροφορικό εργαλείο CD-HIT για να ομαδοποιήσουμε τις πρωτεΐνες όλων των γονιδιωμάτων με στόχο να αποκαλυφθούν τα ορθόλογα γονίδια. Εδώ το πρόβλημα είναι η επιλογή των κατάλληλων παραμέτρων για το CD-HIT. Στην βιβλιογραφία, οι περισσότεροι θέτουν ως κριτήριο για να εισαχθούν 2 πρωτεΐνες στο ίδιο cluster, να τηρείται το κριτήριο της 50% ομοιότητας για τουλάχιστον 50% της αλληλουχίας. Εμείς προσπαθήσαμε να βρούμε ένα πιο ασφαλές τρόπο για την επιλογή των κριτηρίων με βάση των οποίων θα δημιουργούνται τα clusters.
Δοκιμάσαμε να τρέξουμε το CD-HIT με ορισμένα κριτήρια, ώστε να παρατηρήσουμε ποια κριτήρια επηρεάζουν περισσότερο το μέγεθος του core genome και να αποσαφηνίσουμε και τη βιολογική σημασία του κάθε κριτηρίου. Τα κριτήρια που αρχικά θέσαμε ήταν το ελάχιστο μήκος πρωτεϊνών που θα δέχεται για την ομαδοποίηση το CD-HIT (κριτήριο -l στο CD-HIT), το ποσοστό ομοιότητας (-c στο CD-HIT), το ποσοστό στοίχισης (alignment) των αλληλουχιών που θα ταξινομηθούν στο ίδιο cluster (-aL στο CD-HIT), και ως εναλλακτική του aL το κριτήριο -s του CD-HIT, που αφορά στην διαφορά στο μήκος της αλληλουχίας των πρωτεϊνών που ταξινομούνται στο ίδιο cluster.



Εικόνα 10. Έχει παρθεί από το manual του CD-HIT, http://www.bioinformatics.org/cd-hit/cd-hit-user-guide.pdf.

Για το ελάχιστο μήκος πρωτεϊνών δοκιμάσαμε την επιλογή των 120, των 80 και των 60 αμινοξέων. Για το ποσοστό ομοιότητας την επιλογή του 60%, 70%, 80% και 90% ομοιότητα. Για το aL δοκιμάσαμε την επιλογή του 50%, 70% και 80% και για το s το 50% και 70%. Σε κάθε δοκιμή μετρούσαμε το μέγεθος του πανγονιδιώματος, το μέγεθος του core genome, τον αριθμό των clusters του core genome όπου ο λειτουργικός σχολιασμός (functional annotation) που έδινε το prokka για κάθε πρωτεΐνη του cluster διέφερε μεταξύ των μελών του cluster, δηλαδή τα clusters στα οποία έχουν ομαδοποιηθεί πρωτεΐνες που με βάση το Prokka επιτελούν διαφορετικές λειτουργίες. Η διαδικασία δημιουργίας του πανγονιδιώματος και του core genome περιγράφεται λεπτομερώς παρακάτω.
Αφού αποσαφηνίσαμε τις αλλαγές που προκαλεί κάθε αλλαγή, και με γνώμονα τη βιολογική σημασία κάθε κριτηρίου, καθορίσαμε τα κριτήρια που θα χρησιμοποιήσουμε στο CD-HIT.

2.4. Επιπλέον επιλογή γονιδιωμάτων

Αφού βρήκαμε τα κριτήρια του CD-HIT που θα χρησιμοποιήσουμε για να ομαδοποιήσουμε τις πρωτεΐνες των γονιδιωμάτων που δεχθήκαμε στο πρώτο βήμα, δημιουργήσαμε μία βάση δεδομένων με όλες τις πρωτεΐνες από τα γονιδιώματα που έχουμε δεχθεί. Αφού ομαδοποιήσαμε τις πρωτεΐνες των γονιδιωμάτων με τα κριτήρια του CD-HIT, στα οποία καταλήξαμε στο προηγούμενο βήμα, μετρήσαμε τον αριθμό των φορών που κάθε γονιδίωμα λείπει από clusters τα οποία συναντάται στην συντριπτική πλειοψηφία των γονιδιωμάτων, δηλαδή από clusters που συναντώνται στο 98% των γονιδιωμάτων. Βάση για αυτή τη σκέψη είναι ότι εάν ένα στέλεχος λείπει πολλές φορές από clusters τα οποία συναντώνται στην συντριπτική πλειοψηφία των γονιδιωμάτων, μάλλον θα είναι κακής ποιότητας η αλληλούχιση του και γι' αυτό λείπει από τόσα πολλά clusters. Έτσι αφαιρέθηκαν από την ανάλυση τα γονιδιώματα τα οποία βρέθηκε να λείπουν σε πολλά, πολύ συντηρημένα clusters.

2.5. Core genome, dispensable genome, unique genome και pangenome

2.5.1.Δημιουργία του core, dispensable, unique και pangenome

Μετά την επιλογή των γονιδιωμάτων, που θα χρησιμοποιηθούν για την περαιτέρω ανάλυση, δημιουργήσαμε τη βάση δεδομένων που περιέχει όλες τις πρωτεΐνες από όλα τα γονιδιώματα. Στη συνέχεια, χρησιμοποιήσαμε το CD-HIT για να ομαδοποιήσουμε τις πρωτεΐνες σε clusters με βάση τα κριτήρια στα οποία καταλήξαμε στο βήμα 2.3. Κάθε cluster αποτελεί μία οικογένεια ορθόλογων πρωτεϊνών. Τα clusters τα οποία περιέχουν πρωτεΐνες από όλα τα γενώματα, δηλαδή σε αυτές τις οικογένειες πρωτεϊνών ταξινομούνται πρωτεΐνες από όλα τα γενώματα, αποτελούν τα clusters του core genome / core proteome. Τα clusters αυτά εκπροσωπούνται από το αντιπροσωπευτικό τους CDS ή πρωτεΐνη στο core genome και core proteome, αντίστοιχα.
Στη συνέχεια, τα clusters τα οποία περιέχουν πρωτεΐνες από μόνο ένα γένωμα, χρησιμοποιήθηκαν για να δημιουργήσουμε το unique genome / unique proteome. Το unique genome περιλαμβάνει το CDS των γονιδίων από όλα τα clusters, στα οποία συναντάται μόνο ένα γονιδίωμα, και αντίστοιχα το unique proteome περιλαμβάνει τις πρωτεΐνες από όλα τα clusters, στα οποία συναντάται μόνο ένα γένωμα.
Τα clusters στα οποία περιλαμβάνεται παραπάνω από ένα γονιδίωμα αλλά όχι όλα τα γονιδιώματα που έχουμε αποδεχτεί, χρησιμοποιήθηκαν για να δημιουργήσουμε το dispensable genome / dispensable proteome. Δηλαδή το dispensable genome περιλαμβάνει το CDS των αντιπροσωπευτικών γονιδίων από όλα τα clusters, στα οποία συναντάται παραπάνω από ένα γένωμα αλλά όχι όλα τα γενώματα που έχουμε αποδεχτεί και το dispensable proteome περιλαμβάνει τις αντιπροσωπευτικές πρωτεΐνες από όλα τα clusters, στα οποία συναντάται παραπάνω από ένα γένωμα αλλά όχι όλα τα γενώματα που έχουμε αποδεχτεί.
Τέλος το πανγονιδίωμα (pangenome) περιλαμβάνει τα αντιπροσωπευτικά CDS του κάθε cluster, ενώ το pan proteome περιλαμβάνει τις αντιπροσωπευτικές πρωτεΐνες από κάθε cluster.

2.5.2. Κριτήριο για τα clusters του core genome

Αφού βρήκαμε το core genome του *P.mirabilis*, συλλέξαμε σε ένα συγκεντρωτικό πίνακα ορισμένα χαρακτηριστικά των clusters του core genome. Παρατηρήσαμε ότι σε ορισμένα clusters ομαδοποιούνταν μαζί αρκετές πρωτεΐνες του ίδιου γενώματος. Έτσι θεωρήσαμε ότι clusters τα οποία περιείχαν παραπάνω από μια πρωτεΐνη του ίδου γενώματος για παραπάνω από 10 γενώματα, δεν ήταν αξιόπιστα για την ένταξη

τους στο core genome, καθώς πιθανώς αφορούν σε παράλογα γονίδια, τα οποία όμως δεν θα μας βοηθήσουν στην αποσαφήνιση των εξελικτικών σχέσεων μεταξύ των στελεχών. Γι' αυτό, όσα clusters περιείχαν πάνω από μία πρωτεΐνη ίδιου στελέχους για παραπάνω από 10 στελέχη, αφαιρέθηκαν από το core genome, και core proteome και δεν χρησιμοποιήθηκαν στην περαιτέρω ανάλυση. Για τα cluster στα οποία υπήρχαν πάνω από μία πρωτεΐνη ίδιου στελέχους αλλά για λιγότερα από 10 στελέχη, δεν έγινε κάτι επιπλέον, απλά για τα στελέχη αυτά χρησιμοποιήθηκε για την ανάλυση η αλληλουχία που είχε μεγαλύτερο ποσοστό ομοιότητας με την αντιπροσωπευτική αλληλουχία του cluster ή εάν ήταν η αντιπροσωπευτική αλληλουχία του cluster, χρησιμοποιείται αυτή ως η αλληλουχία για το στέλεχος αυτό.

2.5.3.Λειτουργικός σχολιασμός του core / dispensable / unique proteome

Αφού βρήκαμε το core proteome, το unique proteome και το dispensable proteome χρησιμοποιήσαμε το βιοπληροφορικό εργαλείο EggNOG-mapper για το λειτουργικό σχολιασμό τους.
Από τα αποτελέσματα του EggNOG-mapper απομονώσαμε την στήλη με την κατηγοριοποίηση των πρωτεϊνών σε κατηγορίες από Clusters of Orthologous Groups (COG categories). Στη συνέχεια κάναμε ένα συγκεντρωτικό πίνακα με τις κατηγορίες COG των πρωτεϊνών του core proteome, του unique proteome και του dispensable proteome, ώστε να δούμε ποιες κατηγορίες πρωτεϊνών βρίσκονται συχνότερα σε κάθε proteome. Παρακάτω βλέπουμε τις κατηγορίες των COG (Tatusov et al., 2000)

Πίνακας 3. Κατηγορίες Clusters of Orthologous Groups (COG categories).

COG κατηγορίες	
Μεταφορά και μεταβολισμός Αμινοξέων [E]	Μεταβολισμός και μεταφορά ανόργανων ιόντων [P]
Μεταφορά και μεταβολισμό υδατανθράκων [G]	Ενδοκυτταρική μεταφορά και απέκκριση [U]
Κυτταρική διαίρεση, έλεγχος κυτταρικού κύκλου [D]	Μεταβολισμός και μεταφορά λιπιδίων [I]
Κυτταρική κινητικότητα [N]	Mobilome: προφάγοι και, transposons [X]
Βιοσύνθεση κυτταρικού τοιχώματος, μεμβράνης και κυτταρικού φακέλου [M]	Πυρηνική δομή [Y]
Δομή χρωματίνης [B]	Μεταβολισμός και μεταφορά νουκλεοτιδίων [F]
Μεταβολισμός και μεταφορά συνενζύμων [H]	Μεταμεταφραστικές τροποποιήσεις, αντικατάσταση πρωτεϊνών, chaperons [O]
Κυταροσκελετός [Z]	Επεξεργασία και αλλαγή του RNA [A]
Μηχανισμοί άμυνας [V]	Αντιγραφή και επιδιόρθωση [L]
Παραγωγή ενέργειας [C]	Μεταφορά, καταβολισμός και βιοσύνθεση δευτερευόντων μεταβολιτών [Q]
Εξωκυτταρικές δομές [W]	Μηχανισμοί μετάδοσης σήματος [T]
Άγνωστη λειτουργία [S]	Μεταγραφή [K]
Γενική λειτουργία [R]	Μετάφραση, δομή και βιοσύνθεση ριβοσώματος [J]

Στη συνέχεια, κάθε ένα από τα proteome, ελέγχθηκε και για γονίδια που προσδίδουν ανθεκτικότητα σε αντιβιοτικά από το βιοπληροφορικό εργαλείο AMRFinder (Feldgardern et al., 2019).

2.6. Προσδιορισμός του πανγονιδιώματος: ανοικτό ή κλειστό

Προκειμένου να διερευνήσουμε αν το pangenome που προσδιορίστηκε είναι κλειστό ή ανοικτό, ακολουθούμε την εξής διαδικασία: ξανακάνουμε ένα clustering με τα κριτήρια του CD-HIT στα οποία καταλήξαμε, μόνο που αυτή τη φορά δεν κάνουμε clustering απευθείας στο σύνολο των πρωτεϊνών όλων των γονιδιωμάτων, αλλά αρχίζουμε από δύο γονιδιώματα και βρίσκουμε το πανγονιδίωμα και το core genome για αυτά τα γονιδιώματα, και στη συνέχεια προσθέτουμε κάθε φορά ένα ακόμη γονιδίωμα, μέχρι να έχουν αναλυθεί όλα τα γονιδιώματα.
Αυτή η διαδικασία θα μας δώσει ένα πίνακα που θα έχουμε στη μία στήλη τον αριθμό των γονιδιωμάτων, στην δεύτερη στήλη το μέγεθος του πανγονιδιώματος, στη τρίτη στήλη το μέγεθος του core genome και στη τέταρτη στήλη το λόγο του core genome / pangenome. Ο λόγος core / pangenome υπολογίζεται για κάθε αριθμό γονιδιωμάτων ως το πηλίκο του μεγέθους του core genome προς το μέγεθος του pangenome για το συγκεκριμένο αριθμό γονιδιωμάτων.
Εάν ο λόγος αυτός μετά από κάποιο αριθμό γονιδιωμάτων φτάνει σε σταθερό επίπεδο (πλατό) τότε το πανγονιδίωμα του βακτηρίου αυτού θεωρείται κλειστό. Εάν δεν φτάνει σε πλατό αλλά ο λόγος συνεχώς μειώνεται τότε θεωρείται ανοικτό.

2.7. Φυλογενετική ανάλυση

Αφού καταλήξαμε στο τελικό core proteome και core genome, απομονώσαμε για κάθε cluster του core genome, τα CDS του κάθε γονιδιώματος για το cluster αυτό. Έτσι για κάθε cluster έχουμε ένα αρχείο το οποίο περιέχει τα CDS που περιλαμβάνονται στο cluster αυτό. Για τα γονιδιώματα που περιείχαν παραπάνω από ένα αντίγραφα σε κάποιο cluster, για την φυλογενετική ανάλυση χρησιμοποιείται το αντίγραφο που είχε την μεγαλύτερη ομοιότητα με το αντιπροσωπευτικό για το cluster CDS με βάση τα αποτελέσματα του CD-HIT, είτε εάν κάποιο από τα αντίγραφα είναι το αντιπροσωπευτικό του cluster, χρησιμοποιείται αυτό για την φυλογενετική ανάλυση.
Στη συνέχεια, χρησιμοποιούμε το πρόγραμμα CLUSTAL OMEGA για να στοιχίσουμε τις αλληλουχίες σε κάθε cluster του core genome. Αφού στοιχηθούν, τότε δημιουργείται για κάθε στέλεχος το ψευδογονιδίωμα του, δηλαδή ένα FASTA αρχείο το οποίο περιέχει συνεχόμενα τα στοιχισμένα CDS του του στελέχους για όλα τα clusters του core genome. Αυτό απαιτείται για την φυλογενετική ανάλυση και την κατασκευή του φυλογενετικού δέντρου.
Τα ψευδογονιδιώματα όλων των στελεχών έχουν το ίδιο μήκος και μεταξύ τους είναι στοιχισμένα. Στη συνέχεια, χρησιμοποιώντας τα ψευδογονιδιώματα, δημιουργήθηκαν φυλογενετικά δέντρα χρησιμοποιώντας τα βιοπληροφορικά εργαλεία RaxML και FastTree. Το RAxML χρησιμοποιεί την μέθοδο maximum-likelihood για τη φυλογενετική ανάλυση, ως μοντέλο αντικατάστασης χρησιμοποιήθηκε το μοντέλο GTRGAMMA και χρησιμοποιήθηκε και η partitioned analysis, στην οποία για κάθε γονίδιο δημιουργούνται ξεχωριστοί πίνακες αντικατάστασης, έτσι για την κατασκευή του δέντρου τα γονιδιώματα συγκρίνονται μεταξύ τους για κάθε γονίδιο ξεχωριστά και όχι ως συνολικό γονιδίωμα. Για την κατασκευή του φυλογενετικού δέντρου το FastTree χρησιμοποιεί την μέθοδο approximately-maximum-likelihood για την κατασκευή των δέντρων, και χρησιμοποιήσαμε τις προτεινόμενες ρυθμίσεις με μόνη διαφορά ότι επιλέξαμε ως μοντέλο αντικατάστασης το GTR-CAT. Για την οπτικοποίηση των αποτελεσμάτων των δύο εργαλείων χρησιμοποιήθηκε η online πλατφόρμα ITOL (https://itol.embl.de/).

2.8. Δημιουργία core genome multilocus sequence typing (cgMLST)

Για κάθε cluster του core genome δημιουργήθηκε ένα αρχείο που περιέχει όλα τα CDS των γονιδίων που ταξινομούνται σε αυτό το cluster. Κάθε cluster αποτελεί μία οικογένεια ορθόλογων γονιδίων και θέλουμε να βρούμε τα διαφορετικά γονίδια (αλλήλια) που υπάρχουν σε κάθε cluster. Έτσι συγκρίνουμε τα CDS του κάθε cluster του core genome με το CD-HIT, με προϋπόθεση για να ομαδοποιηθούν δύο αλληλουχίες στην ίδια ομάδα, να έχουν ομοιότητα 100% και να έχουν ακριβώς το ίδιο μήκος. Γονίδια που διαφέρουν ακόμη και σε ένα νουκλεοτίδιο ή είναι μικρότερα ακόμη και κατά ένα νουκλεοτίδιο, ομαδοποιούνται ξεχωριστά. Κάθε cluster του core genome αντιπροσωπεύει ένα γονίδιο του core genome, έτσι για κάθε cluster ή καλύτερα για κάθε γενετικό τόπο ή γονίδιο του core genome, βρήκαμε τα διαφορετικά “αλλήλια” του, δηλαδή τις διαφορετικές μορφές του που συναντώνται στα διάφορα στελέχη του *P. mirabilis*. Σε κάθε γενετικό τόπο δινεται το όνομα cgPM (core genome Proteus Mirabilis) και ένας αύξοντας αριθμός, οπότε συνολικά έχουμε τόσα cgPM όσα και είναι και τα γονίδια του core genome. Και σε κάθε cgPM, δίνεται ένας δεύτερος διαφορετικός αύξοντας αριθμός που χαρακτηρίζει κάθε διαφορετικό αλλήλιο του.

Στη συνέχεια, για τα γονιδιώματα που χρησιμοποιήσαμε στην ανάλυση, βρήκαμε το cgMLST τύπο τους. Αυτό έγινε με το να αντιστοιχίσουμε τα CDS του κάθε γονιδιώματος σε ποια αλλήλια των αντίστοιχων cgPM ανήκουν. Κάθε cgMLST τύπος συγκρίθηκε με όλους τους υπόλοιπους cgMLST τύπους για να βρούμε σε πόσα διαφορετικά cgPM διέφεραν. Παρατηρήσαμε από τη σύγκριση των διαφόρων cgMLST τύπων ότι οι διαφορές μεταξύ τους ομαδοποιούνταν σε δύο διακριτές ομάδες, δηλαδή στην μία ομάδα οι διαφορές μεταξύ των cgMLST τύπων ήταν λιγότερες από 430, ενώ στη δεύτερη ομάδα οι cgMLST τύποι διέφεραν σε παραπάνω από 800 αλλήλια. Ελάχιστοι cgMLST τύποι διέφεραν μεταξύ τους σε 430 έως 800 αλλήλια, έτσι θεωρήσαμε ότι η διαφορά μέχρι και σε 430 αλλήλια παρατηρούνταν σε εξελικτικά κοντινούς cgMLST τυπους ενώ η διαφορά σε παραπάνω από 430 αλλήλια δείχνει μεγαλύτερη εξελικτική διαφορά. Γι’ αυτό για τον ορισμό των κλωνικών συμπλεγμάτων (clonal complex), δηλαδή cgMLST τύπων κοντινών εξελικτικά μεταξύ τους, θεωρήσαμε ως όριο τη διαφορά σε λιγότερα από 430 αλλήλια. Για την ένταξη ενός cgMLST τύπου σε ένα κλωνικό σύμπλεγμα αρκεί να απέχει λιγότερο από 430 αλλήλια από ένα άλλο cgMLST τύπο που ανήκει στο κλωνικό σύμπλεγμα.

Η διαδικασία για να βρούμε το cgMLST τύπο για ένα νέο γονιδίωμα είναι η εξής: αρχικά, γίνεται ένας σχολιασμός από το εργαλείο Prokka για να βρούμε τα CDS του γονιδιώματος αυτού. Στη συνέχεια, όλα τα CDS του γονιδιώματος συγκρίνονται με την blast database που περιλαμβάνει όλα τα αλλήλια από όλα τα cgPM με το λογισμικό blastn. Εφόσον κάποιο από τα CDS του γενώματος που διερευνάται είναι εντελώς όμοιο (δηλαδή ταύτιση=100%, coverage=100% και mismatches=0), με κάποιο από τα αλλήλια του αντίστοιχου cgPM που έχουν εντοπιστεί, τότε για το συγκεκριμένο cgPM ο cgMLST τύπος του νέου γενώματος έχει το αλλήλιο αυτό. Εάν για κάποιο cgPM δεν βρεθεί CDS που να είναι ακριβώς όμοιο με κάποιο από τα αλλήλια του, αλλά βρεθεί CDS που να πληρεί τα επιλεγμένα κριτήρια (70% ταύτιση και 70% coverage) για κάποιο αλλήλιο του συγκεκριμένου cgPM, και αυτό το CDS δεν δίνει μεγαλύτερη ομοιότητα για κάποιο άλλο cgPM, αυτό θεωρείται καινούργιο αλλήλιο (novel allele). Εάν βρεθούν παραπάνω από ένα CDS για ένα συγκεκριμένο cgPM ως πιθανά νέα αλλήλια, μόνο το αλλήλιο με τη μεγαλύτερη ομοιότητα με κάποιο ήδη υπάρχων αλλήλιο χαρακτηρίζεται ως καινούργιο αλλήλιο. Το καινούργιο αλληλίο παίρνει αύξοντα αριθμό κατα ένα μεγαλύτερο από το τελευταίο αλλήλιο.

Εάν για ένα συγκεκριμένο cgPM δεν βρεθεί κάποιο CDS που να πληρεί τα κριτήρια του blastn, τότε γίνεται ένα νέο blastn με είσοδο όλο το γονιδίωμα έναντι της βάσης δεδομένων των αλληλίων του συγκεκριμένου cgPM, με κριτήριο το e-value να είναι μικρότερο του 10^{-6}. Αυτό το βήμα γίνεται ώστε να αποφύγουμε τις περιπτώσεις όπου σε ένα γονιδίωμα δεν βρίσκουμε αλλήλιο, επειδή το Prokka δεν μπόρεσε να εντοπίσει ανοιχτό πλαίσιο ανάγνωσης (Open Reading Frame, ORF) λόγω ύπαρξης ψευδογονιδίου ή χαμηλής ποιότητας δεδομένων αλληλούχισης. Η περιοχή αυτή του γονιδιώματος ελέγχεται εάν έχει μεγαλύτερη ομοιότητα με κάποιο άλλο cgPM, και εάν δεν παρουσιάζει με κάποιο άλλο cgPM μεγαλύτερη ομοιότητα τότε θεωρείται ως νέο αλλήλιο, βέβαια σημειώνεται στο όνομα του το “pseudo”. Εάν και πάλι δεν βρεθεί

ομοιότητα που να πληρεί τα κριτήρια του blastn, τότε θεωρούμε ότι το συγκεκριμένο cgPM απουσιάζει από το γονιδίωμα αυτό και στο cgMLST τύπο του το συγκεκριμένο cgPM σημειώνεται με 0. Εάν το γονιδίωμα δεν περιέχει CDS για ορισμένους cgPM, ελέγχεται εάν ο αριθμός των cgPM που λείπουν προς τον αριθμό όλων των cgPM υπερβαίνει το cut-off (που τέθηκε στο 8%). Εάν υπερβαίνει το cut-off δεν συνεχίζεται η ανάλυση και δεν θα του δοθεί κάποιος cgMLST τύπος, καθώς θεωρείται ότι λείπουν πολλά cgPM.
Με αυτή τη διαδικασία, για κάθε νέο γονιδίωμα βρίσκουμε τα αλλήλια του για κάθε γενετικό τόπο. Εάν το γονιδίωμα αυτό έχει ακριβώς τα ίδια αλλήλια για κάθε cgPM με κάποιο από τους γνωστούς cgMLST τύπους, τότε το γονιδίωμα αυτό χαρακτηρίζεται ότι έχει αυτόν το cgMLST τύπο. Εάν δεν έχει ακριβώς τα ίδια αλλήλια με κάποιο από τους γνωστούς cgMLST τύπους, τότε ανήκει σε ένα καινούργιο cgMLST τύπο, που προσδιορίζεται με αύξοντα αριθμό κατά ένα μεγαλύτερο από τον τελευταίο γνωστό cgMLST τύπο.

2.9. Δημιουργία εργαλείων για πανγονιδιωματικές αναλύσεις

Δημιουργήσαμε τέσσερα (4) εργαλεία για την πραγματοποίηση των παραπάνω βημάτων και την ευκολότερη χρήση τους για πανγονιδιωματικές αναλύσεις. Έτσι, αρχικά δημιουργήθηκε το script **Genomes_Finder.sh**, το οποίο ουσιαστικά συνοψίζει τα βήματα 2.1 και 2.2. Δηλαδή ο χρήστης επιλέγει σε ποιο βακτηριακό είδος θέλει να κάνει την ανάλυση για το core genome, και το script κατεβάζει όλα τα γονιδιώματα που πληρούν τα κριτήρια για μια ποιοτική ανάλυση, όπως την περιγράψαμε παραπάνω. Το αποτέλεσμα που βγάζει είναι ένας πίνακας με όλα τα στελέχη που πληρούν τις προϋποθέσεις για ανάλυση, το γονιδίωμα αυτών των στελεχών και τα CDS και πρωτεΐνες αυτών των στελεχών.
Στη συνέχεια με το script **Pangenome_Finder.sh**, δημιουργείται στο φάκελο που επέλεξε ο χρήστης, το core genome, το core proteome, το dispensable genome, το dispensable proteome, το unique genome, το unique proteome και το pangenome και το panproteome. Επίσης, δημιουργείται εάν το επιλέξει ο χρήστης η ανάλυση για το κλειστό/ανοικτό πανγονιδίωμα και η ανάλυση για τη δημιουργία του αρχείου με τα ψευδο-γονιδιώματα των στελεχών ώστε να είναι εφικτή η δημιουργία ενός φυλογενετικού δέντρου είτε από το RAxML είτε από το FastTree.
Στη συνέχεια, το script **cgMLST_prep.sh** δημιουργεί όλα τα απαραίτητα αρχεία και τις βάσεις δεδομένων για να λειτουργήσει το εργαλείο cg_finder. Βασικά δημιουργούνται η βάση δεδομένων που περιέχει όλους τους cgMLST τυπους των στελεχών που χρησιμοποιήσαμε για την ανάλυση μας και η βάση δεδομένων για τα αλλήλια για κάθε cgPM. Όσον αφορά στη βάση δεδομένων για τους cgMLST τύπους έχουμε δύο βάσεις δεδομένων: μία main database (cg_types) που περιέχει τους cgMLST τύπους από τα στελέχη που χρησιμοποιήσαμε για να βρούμε το core genome και τους cgMLST τύπους που περιέχουν όλα τα cgPM, και μία δεύτερη βάση δεδομένων, missing database (cg_types_missing), η οποία περιέχει ό,τι και η πρώτη βάση δεδομένων αλλά και cgMLST τύπους που εισάγουμε και δεν περιέχουν όλα τα cgPM. Πριν τη χρήση του cg_finder, οι δύο βάσεις δεδομένων είναι ίδιες, αλλά καθώς αναλύουμε και γονιδιώματα τα οποία δεν περιέχουν όλα τα cgPM, οι δύο βάσεις δεδομένων αρχίζουν να διαφέρουν.
Τέλος, το εργαλείο **cg_finder** δέχεται ως δεδομένα εισόδου, ένα γένωμα ή μία συλλογή από γενώματα, κάνει σχολιασμό του ή των γονιδιωμάτων με το εργαλείο Prokka, και στη συνέχεια βρίσκει το cgMLST τύπο τους. Εάν το γονιδίωμα που εισάγουμε περιέχει όλα τα cgPM τότε το cg_finder του δίνει ένα cgMLST τύπο με βάση την main database (cgMLST σχήμα). Για τα στελέχη που δεν περιέχουν όλα τα cgPM ο cgMLST τυπος καθορίζεται με βάση τη missing database (cgMLST missing σχήμα). Τέλος, αφού δώσει το cgMLST τύπο στα νέα γονιδιώματα, και εάν το επιλέξει ο χρήστης τότε το cg_finder δημιουργεί και φυλογενετικά δέντρα είτε με το εργαλείο RAxML είτε με το εργαλείο FastTree με βάση τη διαδικασία που ακολουθειται στο βήμα 2.7. Εάν εισάγουμε συλλογή γονιδιωμάτων που ορισμένα από αυτά δεν περιέχουν όλους τους cgPM, τότε δημιουργούνται 2 δέντρα, ένα δέντρο με τα στελέχη που περιέχουν όλους τους cgPM και ένα δεύτερο με όλα τα στελέχη που έχουμε εισάγει. Για το σχηματισμό του δεύτερου φυλογενετικού δέντρου χρησιμοποιούνται μόνο οι αλληλουχίες των cgPM που βρίσκονται σε όλα τα στελέχη.

2.10. Αναλύσεις με τη χρήση cg_finder

Αρχικά, για το σύνολο των γενωμάτων του *P. mirabilis* που είναι κατατεθειμένα στη Genbank τρέξαμε το εργαλείο cg_finder. Θεωρήσαμε ως κριτήριο για τη μη αποδοχή στελεχών την έλλειψη πάνω από 5% τόπων, και ως κριτήριο για την ένταξη ενός νέου αλληλίου την ομοιότητα τουλάχιστον 70% για τουλάχιστον 70% του μήκους ως προς ένα άλλο αλλήλιο του γενετικού τόπου. Για όλα τα στελέχη του *P. mirabilis* τρέξαμε το εργαλείο ResBar, κατασκευασμένο στο εργαστήριο Βακτηριολογίας του Ελληνικού Ινστιτούτου Παστέρ από τον Δρ. Στάθη Κωτσάκη, και το οποίο δέχεται ως είσοδο ένα γονιδίωμα και βρίσκει τα γονίδια αντοχής σε διάφορα είδη αντιβιοτικών. Επιπλέον, για όλα τα στελέχη εντοπίσαμε τη χώρα προέλευσης τους, και μαζί με τα αποτελέσματα του ResBar, τα ενσωματώσαμε στο φυλογενετικό δέντρο με όλα τα στελέχη του *P. mirabilis.*

Στη συνέχεια για να εντοπίσουμε τις εξελικτικές σχέσεις μεταξύ των διαφορετικών ειδών του *Proteus* κατεβάσαμε από τη GenBank τα γονιδιώματα όλων των στελεχών που έχουν χαρακτηριστεί ως *Proteus* και τρέξαμε το εργαλείο cg_finder με είσοδο τη συλλογή από όλα τα γονιδιώματα των *Proteus,* με κριτήριο να μην τους λείπουν παραπάνω από 8% των cgPM, και με κριτήριο για την ένταξη ενός νέου αλληλίου την ομοιότητα τουλάχιστον 60% για τουλάχιστον 60% του μήκους ως προς ένα άλλο αλλήλιο του γενετικού τόπου*.*

Τέλος, κατεβάσαμε από τη βάση δεδομένων RefSeq όλα τα στελέχη που ανήκουν στα γένη *Proteus, Providencia, Morganella, Xenorhabdus, Photorhabdus* της οικογένειας *Morganellaceae* και τρέξαμε το εργαλείο cg_finder με είσοδο τη συλλογή αυτή, με κριτήριο να μην τους λείπουν παραπάνω από 34% των cgPM, και με κριτήριο για την ένταξη ενός νέου αλληλίου την ομοιότητα τουλάχιστον 60% για τουλάχιστον 60% του μήκους ως προς ένα άλλο αλλήλιο του γενετικού τόπου*.*

2.11. Σύγκριση φυλογενετικών αναλύσεων

Χρησιμοποιήσαμε το βιοπληροφορικό εργαλείο Harvest με είσοδο τα 154 στελέχη που έγιναν αποδεκτά για τη πανγονιδιωματική ανάλυσή μας, ώστε να δημιουργήσουμε το φυλογενετικό δέντρο τους. Το πρόγραμμα δημιουργεί το core genome, εντοπίζοντας τις κοινές περιοχές με υψηλή ομολογία στη νουκλεοτιδική αλληλουχία ως προς ένα γένωμα αναφοράς, και στη συνέχεια εντοπίζει SNPs από το core genome και δημιουργεί το φυλογενετικό δέντρο, συγκρίνοντας τα διάφορα SNPs των στελεχών.

Το δέντρο που προέκυψε από το Harvest συγκρίθηκε τόσο με το δέντρο του RAxML όσο και με το δέντρο του FastTree, χρησιμοποιώντας το εργαλείο Fast Tree-Comparison Tools, που δημιουργήθηκε από τους Price et al., στο Lawrence Berkeley National Lab (http://www.microbesonline.org/fasttree/treecmp.html). Το αποτέλεσμα της σύγκρισης δύο δέντρων με το εργαλείο Fast Tree-Comparison Tools είναι ένα πηλίκο (fraction), ανάλογο με το Robinson-Foulds distance. Όσο πιο κοντά στο 1 είναι η παράμετρος fraction τόσο πιο όμοια είναι τα συγκρινόμενα δέντρα. Επίσης, χρησιμοποιήθηκε και ο αλγόριθμος tanglegram του εργαλείου Dendroscope για την οπτικοποίηση των διαφορών των δέντρων.

Τέλος, συγκρίθηκαν μεταξύ τους και τα δέντρα που προέκυψαν από το RAxML και το FastTree με τη χρήση του Fast Tree-Comparison Tools και του Dendroscope.

3. Αποτελέσματα

3.1. Επιλογή χρωμοσωμικών αλληλουχιών *P. mirabilis*

Από τα συνολικά 266 γενώματα *P. mirabilis* κατατεθειμένα στη βάση δεδομένων GenBank του NCBI (κατά την 16/05/2020), εφαρμόζοντας τα κριτήρια αποδοχής γενωμάτων για την ανάλυση του core genome, ήτοι γενώματα με **N50** > 100.000 ή πλήρως συναρμολογημένα γονιδιώματα, επιλέχθηκαν 158 γενώματα τα στοιχεία των οποίων δίνονται στον **Πίνακα 4.** Συγκεκριμένα στον πίνακα 4 αναφέρονται:

1. Το όνομα του στελέχους.
2. Το taxonomy id (Tax_Id) του βακτηρίου όπως είναι σημειωμένο στο αρχείο assembly_stats. Στη στήλη αυτή παρατηρούμε διάφορα id, αλλά όλα αφορούν στο βακτήριο *P. mirabilis*.
3. Η ημερομηνία που κατατέθηκε το κάθε γονιδίωμα.
4. Το επίπεδο κατασκευής του γονιδιώματος (Assembly Level). Σε αυτή τη στήλη έχουμε 4 πιθανές επιλογές:
   α) “complete”: σε γονιδιώματα που είναι το πλήρως συναρμολογημένο το γονιδίωμα του στελέχους, δηλαδή ένα ολοκληρωμένο χρωμόσωμα με ή χωρίς το ή τα πλασμίδια του αλλά χωρίς να υπάρχουν κενά ή μη τοποθετημένα scaffolds, β) “chromosome” που είναι συναρμολογημένο το χρωμόσωμα του βακτηρίου αλλά περιέχει κενά ή μη τοποθετημένα scaffolds, γ) “scaffold”, όπου έχουν δημιουργηθεί ορισμένα scaffolds αλλά αυτά δεν είναι τοποθετημένα σε σειρά, και δ) “contig”, όπου έχουν δημιουργηθεί μονάχα τα contigs.
5. Το accession number του στελέχους στη βάση δεδομένων GenBank.
6. Το μέγεθος του γενώματος.
7. Ο αριθμός των contigs, (στα γονιδιώματα που είναι “complete” δεν αναφέρεται ο αριθμός των contigs).
8. Το genome coverage, το οποίο αφορά στο πόσες φορές κατά μέσο όρο έχει αλληλουχηθεί κάθε βάση του γονιδιώματος. Όσο πιο μεγάλο είναι το genome coverage τόσο πιο ακριβής είναι η αλληλούχιση.
9. Το N50, που είναι το μήκος του μικρότερου contig, το οποίο μαζί με όλα τα μεγαλύτερα, από αυτό σε μήκος, contigs περιέχουν το 50% του γονιδιώματος.
10. Το L50, που είναι ο μικρότερος αριθμός contig που περιέχουν το 50% του γονιδιώματος.

Πίνακας 4. Όλα τα στελέχη *P.mirabilis* της βάσης δεδομένων GenBank, τα οποία έγιναν αρχικά αποδεκτά στην ανάλυση για το πανγονιδίωμα και το core genome.

Είδος	Στέλεχος	Tax_id	Ημερομηνία	Assembly Level	GenBank number	Μέγεθος	Contigs	Genome Coverage	N50	L50
Proteus mirabilis	HI4320	529507	2008-04-03	Complete Genome	GCA_000069965.1	4099895	-	-	-	-
Proteus mirabilis	BB2000	1266738	2013-08-06	Complete Genome	GCA_000444425.1	3846754	-	10X (overall); 100X (sections)	-	-
Proteus mirabilis	FDAARGOS_81	584	2018-01-29	Complete Genome	GCA_000783575.2	4077315	-	389.811x	-	-
Proteus mirabilis	FDAARGOS_80	584	2018-01-30	Complete Genome	GCA_000783595.2	4044233	-	458.931x	-	-
Proteus mirabilis	FDAARGOS_67	584	2018-02-13	Complete Genome	GCA_000783875.2	3955473	-	672.753x	-	-
Proteus mirabilis	FDAARGOS_60	584	2018-01-29	Complete Genome	GCA_000784015.2	4105573	-	491.256x	-	-
Proteus mirabilis	CYPM1	584	2015-09-15	Complete Genome	GCA_001281545.1	3793000	-	454 28x; Illumina 200X	-	-

Proteus mirabilis	CYPV1	584	2015-09-15	Complete Genome	GCA_001281565.1	3794983	-	454 30x; Illumina 200X	-	-
Proteus mirabilis	AOUC-001	584	2016-05-09	Complete Genome	GCA_001640985.1	4272433	-	260x	-	-
Proteus mirabilis	AR_0059	584	2017-03-21	Complete Genome	GCA_002055685.1	4191021	-	131x	-	-
Proteus mirabilis	AR_0159	584	2017-06-08	Complete Genome	GCA_002180115.1	4216982	-	263x	-	-
Proteus mirabilis	AR_0155	584	2017-06-08	Complete Genome	GCA_002180235.1	4587183	-	133x	-	-
Proteus mirabilis	AR_0156	584	2017-06-16	Complete Genome	GCA_002197405.1	4440469	-	65x	-	-
Proteus mirabilis	T21	584	2017-09-18	Complete Genome	GCA_002310875.1	4286593	-	100x	-	-
Proteus mirabilis	T18	584	2017-09-18	Complete Genome	GCA_002310895.1	4190461	-	100x	-	-
Proteus mirabilis	BC11-24	584	2018-02-11	Complete Genome	GCA_002944495.1	4021165	-	100.0x	-	-
Proteus mirabilis	GN2	584	2018-02-12	Complete Genome	GCA_002945235.1	4012640	-	48.0x	-	-
Proteus mirabilis	AR379	584	2018-04-30	Complete Genome	GCA_003073935.1	4238045	-	16.09x	-	-
Proteus mirabilis	AR_0029	584	2018-06-10	Complete Genome	GCA_003204115.1	3990122	-	210x	-	-
Proteus mirabilis	PmBC1123	584	2018-12-02	Complete Genome	GCA_003855615.1	4074828	-	100.0x	-	-
Proteus mirabilis	PmSC1111	584	2018-12-02	Complete Genome	GCA_003855635.1	4189199	-	100.0x	-	-
Proteus mirabilis	VAC	584	2019-08-21	Complete Genome	GCA_008041895.1	4090308	-	30.0x	-	-
Proteus mirabilis	CRPM10	584	2019-09-02	Complete Genome	GCA_008195605.1	4204856	-	100.0x	-	-
Proteus mirabilis	S1959	584	2019-09-21	Complete Genome	GCA_008630655.1	3957207	-	265.0x	-	-
Proteus mirabilis	K817	584	2019-09-24	Complete Genome	GCA_008705195.1	4017079	-	317.0x	-	-
Proteus mirabilis	CRE14IB	584	2019-12-11	Complete Genome	GCA_009429045.2	4120262	-	152.0x	-	-
Proteus mirabilis	ENT1157	584	2019-11-20	Complete Genome	GCA_009684665.1	4105149	-	1x	-	-
Proteus mirabilis	SCBX1.1	584	2019-12-28	Complete Genome	GCA_009806715.1	4352264	-	200.0x	-	-
Proteus mirabilis	N18-00201	584	2020-02-12	Complete Genome	GCA_010442675.1	3819372	-	107X	-	-
Proteus mirabilis	CC15031	584	2020-02-18	Complete Genome	GCA_010692865.1	4031742	-	50.0x	-	-
Proteus mirabilis	1701092	584	2020-03-02	Complete Genome	GCA_011045575.1	4131513	-	100.0x	-	-
Proteus mirabilis	ZA25	584	2020-03-02	Complete Genome	GCA_011045855.1	4108293	-	50.0x	-	-
Proteus mirabilis	PmBR607	584	2020-03-11	Complete Genome	GCA_011149675.1	4087423	-	200x	-	-
Proteus mirabilis	XH1568	584	2020-03-17	Complete Genome	GCA_011383025.1	4012915	-	200.0x	-	-
Proteus	XH1569	584	2020-03-17	Complete	GCA_01138	3998116	-	200.0x	-	-

mirabilis				Genome	3045.1					
Proteus mirabilis	STP3	584	-	Complete Genome	GCA_012516515.1	4115975	-	200.0x	-	-
Proteus mirabilis	NCTC4199	584	2018-12-20	Complete Genome	GCA_900635965.1	3875274	-	100x	-	-
Proteus mirabilis	FDAARGOS_284	584	2018-01-19	Contig	GCA_002206145.2	4155562	1	598x	4155562	1
Proteus mirabilis	NCTC13376	584	2018-08-01	Contig	GCA_900455195.1	3943574	3	100x	3922486	1
Proteus mirabilis	CCUG	584	2018-06-06	Chromosome	GCA_003194305.1	4101638	2	107.0x	3909769	1
Proteus mirabilis	MGYG-HGUT-02514	584	2019-08-16	Scaffold	GCA_902387925.1	4101638	2	10x	3909769	1
Proteus mirabilis	CLPM181223	584	2019-07-16	Chromosome	GCA_007012285.1	4049470	2	751.0x	3746249	1
Proteus mirabilis	NCTC12441	584	2018-08-01	Contig	GCA_900455055.1	4168708	6	100x	3548147	1
Proteus mirabilis	K670	584	2018-04-03	Chromosome	GCA_003030945.1	3935626	2	90.0x	3144533	1
Proteus mirabilis	AR_0377	584	2018-04-30	Contig	GCA_003075585.1	4422028	4	22.44x	3116701	1
Proteus mirabilis	NCTC6197	584	2018-08-01	Contig	GCA_900454995.1	3974058	5	100x	2544959	1
Proteus mirabilis	FDAARGOS_85	584	2018-01-19	Contig	GCA_000783465.2	3945021	9	574x	2497270	1
Proteus mirabilis	K1609	584	2018-04-15	Chromosome	GCA_003051845.1	3817795	4	90.0x	2353381	1
Proteus mirabilis	UMB0315	584	2017-12-28	Contig	GCA_002861105.1	4174256	20	224.8x	2128130	1
Proteus mirabilis	NCTC11938	584	2018-08-01	Contig	GCA_900455245.1	4074307	6	100x	1368431	2
Proteus mirabilis	NCTC12446	584	2018-06-29	Contig	GCA_900455045.1	4128674	12	100x	676006	2
Proteus mirabilis	CRK0056	584	2018-11-14	Contig	GCA_002184635.2	3923349	30	131.29x	513140	3
Proteus mirabilis	ENT1224	584	2019-11-20	Chromosome	GCA_009684595.1	4105071	16	1x	506894	3
Proteus mirabilis	ATCC	584	2014-09-17	Scaffold	GCA_000755485.1	3994012	15	331x	453955	3
Proteus mirabilis	NCTC60	584	2018-06-29	Contig	GCA_900455025.1	3888092	16	100x	352375	4
Proteus mirabilis	irhom_Sw	584	2019-04-02	Contig	GCA_004570595.1	3945692	61	30x	258819	5
Proteus mirabilis	PMS-MEIH	584	2019-03-29	Contig	GCA_004522505.1	4328332	292	30.0x	257093	5
Proteus mirabilis	L76	584	2019-12-13	Contig	GCA_009749255.1	4232672	139	200.0x	255343	6
Proteus mirabilis	11985-2-3	584	2018-06-04	Contig	GCA_003185705.1	3962003	63	117.413x	252724	6
Proteus mirabilis	CRK0329	584	2018-11-14	Contig	GCA_002936215.2	4003868	49	282.72x	245053	7
Proteus mirabilis	GER_MD10_1505_Pmi_049	584	2018-07-15	Contig	GCA_003322945.1	3930028	55	121x	244024	6
Proteus	168F7	584	2019-04-02	Contig	GCA_00457	3949182	45	100x	236082	7

mirabilis					0225.1					
Proteus mirabilis	WGLW4	1125693	2012-09-21	Scaffold	GCA_000297835.1	3960485	36	279x	232868	6
Proteus mirabilis	PmBC55	584	2018-12-02	Contig	GCA_003856175.1	3887792	48	100.0x	229517	6
Proteus mirabilis	XH1566	584	2020-03-21	Contig	GCA_011604125.1	4117579	132	200x	228709	6
Proteus mirabilis	SCDR1	584	2017-03-03	Contig	GCA_002013325.1	3815621	63	300.0x	227512	5
Proteus mirabilis	PmSDC32	584	2020-03-16	Contig	GCA_011365085.1	3902337	49	100.0x	223813	7
Proteus mirabilis	LBUEL-H11	584	2018-05-25	Scaffold	GCA_003171695.1	4525840	177	285.0x	223429	8
Proteus mirabilis	50664164	584	2015-12-09	Contig	GCA_001463025.1	4238841	126	261.0x	220007	6
Proteus mirabilis	CRPM1	584	2019-07-20	Scaffold	GCA_007197655.1	4028540	58	250.0x	219788	7
Proteus mirabilis	M-12	584	2020-03-09	Scaffold	GCA_011078105.1	3929550	55	100x	209147	5
Proteus mirabilis	GB08	584	2016-04-08	Contig	GCA_001617295.1	4047963	59	192.0x	208630	6
Proteus mirabilis	XH1550	584	2020-03-21	Contig	GCA_011602905.1	4141600	135	200x	207975	4
Proteus mirabilis	XH1548	584	2020-03-21	Contig	GCA_011602895.1	4119527	138	200x	207581	8
Proteus mirabilis	L90	584	2019-12-13	Contig	GCA_009749245.1	4245387	85	200.0x	206980	8
Proteus mirabilis	L49	584	2019-12-13	Contig	GCA_009749175.1	4209671	83	200.0x	206952	7
Proteus mirabilis	L52	584	2019-12-13	Contig	GCA_009749085.1	4123606	144	200.0x	206951	7
Proteus mirabilis	L44	584	2019-12-13	Contig	GCA_009749185.1	4101796	90	200.0x	206951	7
Proteus mirabilis	L71	584	2019-12-13	Contig	GCA_009749225.1	4126881	122	200.0x	206951	7
Proteus mirabilis	NIVEDI3-PG74	584	2016-05-07	Scaffold	GCA_001640165.1	4058222	53	97.0x	206798	6
Proteus mirabilis	SC93	584	2019-12-31	Scaffold	GCA_009821615.1	4017218	77	100x	205712	8
Proteus mirabilis	PrK	584	2020-03-07	Contig	GCA_011067125.1	3970593	60	281.0x	203813	8
Proteus mirabilis	XH1564	584	2020-03-21	Contig	GCA_011602785.1	4148589	138	200x	203254	7
Proteus mirabilis	BIOML-A3	584	2019-11-27	Scaffold	GCA_009718745.1	3970443	143	122.163x	202672	5
Proteus mirabilis	Wood	584	2016-05-13	Contig	GCA_001643755.1	3953708	113	104.0x	202584	7
Proteus mirabilis	WGLW6	1125694	2012-09-21	Scaffold	GCA_000297815.1	4101891	54	170x	199921	8
Proteus mirabilis	BIOML-A1	584	2019-11-27	Scaffold	GCA_009718785.1	4541435	91	134.332x	199551	6
Proteus mirabilis	BIOML-A4	584	2019-11-27	Scaffold	GCA_009718905.1	4445693	435	79.558x	198311	6
Proteus mirabilis	BIOML-A2	584	2019-11-27	Scaffold	GCA_009718885.1	4390195	398	202.233x	198252	6

Proteus mirabilis	CRPM11	584	2019-07-20	Scaffold	GCA_007197625.1	4122471	52	250.0x	196840	8
Proteus mirabilis	PM185	584	2017-08-22	Contig	GCA_002265385.1	3960840	62	262.4427038x	195392	9
Proteus mirabilis	GCID_CRE_0016	584	2018-12-27	Contig	GCA_003977485.1	3895174	64	230.38x	195307	7
Proteus mirabilis	XH1565	584	2020-03-21	Contig	GCA_011604135.1	4078455	132	200x	194745	8
Proteus mirabilis	18QD2AZ3W	584	2020-01-03	Contig	GCA_009829055.1	4081112	109	200x	194114	8
Proteus mirabilis	PM005	584	2017-10-03	Contig	GCA_002417445.1	4087896	96	102x	193947	8
Proteus mirabilis	PmSN55	584	2018-12-02	Contig	GCA_003856195.1	4108608	74	100.0x	189865	7
Proteus mirabilis	GB11	584	2016-04-08	Contig	GCA_001617305.1	4151667	92	327.0x	186668	8
Proteus mirabilis	NLAE-zl-G534	584	2016-11-02	Scaffold	GCA_900113495.1	3950550	51	415x	185733	9
Proteus mirabilis	T1C	584	2016-11-01	Contig	GCA_001858185.1	4060112	60	50.8x	184006	7
Proteus mirabilis	XH1556	584	2020-03-21	Contig	GCA_011604145.1	4117831	142	200x	183323	9
Proteus mirabilis	XH1562	584	2020-03-21	Contig	GCA_011602825.1	4116605	140	200x	180919	9
Proteus mirabilis	XH1549	584	2020-03-21	Contig	GCA_011602945.1	4129285	142	200x	178270	9
Proteus mirabilis	172J1	584	2019-04-02	Contig	GCA_004570745.1	3878428	67	100x	172586	6
Proteus mirabilis	PmOXA23-13	584	2019-03-12	Contig	GCA_004347585.1	3945870	89	80.0x	172545	7
Proteus mirabilis	PMWJ	584	2018-09-20	Scaffold	GCA_003583505.1	4109683	62	250.0x	171942	7
Proteus mirabilis	pmi_p1	584	2019-05-29	Scaffold	GCA_901485075.1	4136674	152	50x	171669	8
Proteus mirabilis	Pr2921	584	2016-05-02	Contig	GCA_001647515.1	3924499	56	30.0x	170494	10
Proteus mirabilis	SC89	584	2019-12-31	Scaffold	GCA_009821695.1	4019616	77	100x	169032	9
Proteus mirabilis	C05028	1245027	2013-01-02	Scaffold	GCA_000313255.1	3817619	85	330.0x	168941	9
Proteus mirabilis	XH1559	584	2020-03-21	Contig	GCA_011604105.1	4119088	152	200x	168841	9
Proteus mirabilis	XH1567	584	2020-03-21	Contig	GCA_011604155.1	4120871	150	200x	168695	9
Proteus mirabilis	XH1547	584	2020-03-21	Contig	GCA_011602865.1	4147886	148	200x	168286	10
Proteus mirabilis	M16	584	2016-04-13	Contig	GCA_001619795.1	3903767	90	173.25x	165802	9
Proteus mirabilis	UMB0038	584	2020-04-04	Contig	GCA_012030315.1	3946388	57	103x	164741	9
Proteus mirabilis	TUM11571	584	2018-05-25	Contig	GCA_003175495.1	4297382	186	41.5x	164562	10
Proteus mirabilis	NLAE-zl-C285	584	2016-10-21	Scaffold	GCA_900101725.1	3773105	38	343x	162534	8
Proteus mirabilis	NO-051/03	584	2015-11-27	Contig	GCA_001448445.1	4197318	100	130.0x	160469	11

Proteus mirabilis	NCTC10975	584	2018-06-29	Contig	GCA_900455065.1	4097539	48	100x	158741	9
Proteus mirabilis	175H8	584	2019-04-02	Contig	GCA_004570715.1	3889760	68	100x	157682	9
Proteus mirabilis	UBA3152	584	2017-09-26	Scaffold	GCA_002364645.1	3899048	46	76.28x	157636	8
Proteus mirabilis	4748	584	2019-06-30	Contig	GCA_006517685.1	3927035	96	60.0x	154551	9
Proteus mirabilis	172C2	584	2019-04-02	Contig	GCA_004570215.1	3820128	73	100x	153409	10
Proteus mirabilis	C02011	584	2016-03-30	Scaffold	GCA_001604705.1	3806618	83	150x	152517	9
Proteus mirabilis	PmDJ107	584	2018-12-02	Contig	GCA_003856215.1	3985114	102	100.0x	152164	8
Proteus mirabilis	XH1552	584	2020-03-21	Contig	GCA_011603015.1	4154500	162	200x	147903	10
Proteus mirabilis	188J6	584	2019-04-02	Contig	GCA_004570705.1	3951833	93	100x	147806	9
Proteus mirabilis	XH1557	584	2020-03-21	Contig	GCA_011604025.1	4148761	149	200x	147147	10
Proteus mirabilis	XH1561	584	2020-03-21	Contig	GCA_011604085.1	4118618	142	200x	145494	10
Proteus mirabilis	127_PMIR	584	2015-07-10	Scaffold	GCA_001060765.1	3867931	74	34x	140934	11
Proteus mirabilis	XH1553	584	2020-03-21	Contig	GCA_011604205.1	4106104	141	200x	140871	10
Proteus mirabilis	429_PMIR	584	2015-07-10	Scaffold	GCA_001076715.1	3831225	80	16x	140501	10
Proteus mirabilis	PM187	584	2017-08-22	Scaffold	GCA_002265355.1	4003276	87	269.5462x	138929	7
Proteus mirabilis	XH1555	584	2020-03-21	Contig	GCA_011604055.1	4112527	120	200x	138123	9
Proteus mirabilis	XH1554	584	2020-03-21	Contig	GCA_011602955.1	4074821	127	200x	137147	9
Proteus mirabilis	XH1560	584	2020-03-21	Contig	GCA_011604045.1	4084961	132	200x	137137	10
Proteus mirabilis	47_PMIR	584	2015-07-10	Scaffold	GCA_001065085.1	3923142	89	10x	135456	12
Proteus mirabilis	68_PMIR	584	2015-07-10	Scaffold	GCA_001065915.1	3908659	102	14x	135013	11
Proteus mirabilis	SC90	584	2019-12-31	Scaffold	GCA_009821675.1	4358502	93	100x	134783	11
Proteus mirabilis	ATCC	525369	2009-05-15	Scaffold	GCA_000160755.1	4027100	115	-	130334	11
Proteus mirabilis	XH1563	584	2020-03-21	Contig	GCA_011602885.1	4152335	154	200x	129967	11
Proteus mirabilis	51_PMIR	584	2015-07-10	Scaffold	GCA_001063575.1	3916631	129	12x	127069	12
Proteus mirabilis	430_PMIR	584	2015-07-10	Scaffold	GCA_001076755.1	3820373	87	11x	121050	11
Proteus mirabilis	PmirR	584	2020-04-15	Contig	GCA_012360255.1	3839383	171	96.0x	119534	12
Proteus mirabilis	PmBR19	584	2019-10-19	Contig	GCA_009184605.1	4055226	138	140x	117403	11
Proteus mirabilis	GED7834	584	2016-02-09	Scaffold	GCA_001553525.1	3899263	95	52x	114898	11

Proteus mirabilis	1166_PMIR	584	2015-07-10	Scaffold	GCA_001062135.1	3855586	115	24x	114393	12
Proteus mirabilis	Pm-Oxa48	584	2014-11-03	Contig	GCA_000770765.1	4137208	91	286.0x	114052	12
Proteus mirabilis	PM_178	584	2016-06-01	Contig	GCA_001653845.1	3969065	190	153.7x	112521	13
Proteus mirabilis	PmBR614	584	2019-10-19	Contig	GCA_009183735.1	4020722	97	200x	111831	11
Proteus mirabilis	PR03	1279010	2013-04-19	Contig	GCA_000372565.1	3847612	99	605.0x	111069	13
Proteus mirabilis	AHEPA923	584	2019-07-15	Scaffold	GCA_007004575.1	4296713	110	35.0x	108719	14
Proteus mirabilis	AS012459	584	2020-02-14	Scaffold	GCA_010590425.1	4098176	627	179.34x	107027	13
Proteus mirabilis	AS012369	584	2020-02-14	Scaffold	GCA_010594645.1	4657414	2890	217.27x	105329	15
Proteus mirabilis	1313_PMIR	584	2015-07-10	Scaffold	GCA_001062655.1	3853063	108	20x	105216	13
Proteus mirabilis	XH1551	584	2020-03-21	Contig	GCA_011602985.1	4124296	172	200x	104270	13
Proteus mirabilis	Pm1LENAR	584	2019-07-17	Contig	GCA_007097095.1	3864437	82	41.8x	102814	11
Proteus mirabilis	1134_PMIR	584	2015-07-10	Scaffold	GCA_001060325.1	4007366	183	17x	102633	14
Proteus mirabilis	AS012423	584	2020-02-14	Scaffold	GCA_010591085.1	4251871	1700	175.61x	100452	15

3.2. Εύρεση ορθόλογων πρωτεϊνών – ορισμός πανγονιδιώματος και core genome

Με σκοπό να οριστεί το core genome έλαβε χώρα η εύρεση των ορθόλογων πρωτεϊνών στο σύνολο του πρωτεώματος των επιλεγμένων γενωμάτων *P. mirabilis*. Η εύρεση των ορθόλογων πραγματοποιήθηκε χρησιμοποιώντας ομαδοποίηση βάσει ομολογίας με το πρόγραμμα CD-HIT. Αρχικά το πρωτέωμα αναλύθηκε χρησιμοποιώντας ως όριο ταύτισης αλληλουχίας 60% (-c 0.6), ποσοστό πρωτεΐνης που καλύπτεται από την αντιστοίχιση αλληλουχίας 60% (-aL 0.6) και μήκος πρωτεΐνης 120 αμινοξέα (-l 120). Στην συνέχεια εντοπίστηκαν οι ομάδες πρωτεϊνών (“clusters”) που έφεραν πρωτεΐνες και από τα 158 γενώματα. Τα γονίδια που κωδικοποιούν για αυτές τις πρωτεΐνες συστήνουν το core genome του είδους ενώ τα γονίδια για το σύνολο των clusters το πανγονιδιώμα. Κατά την διάρκεια αυτής της ανάλυσης παρατηρήθηκε ότι σε συγκεκριμένα γενώματα κατ’ επανάληψη δεν εντοπίζονταν clusters που ανιχνεύονταν στα υπόλοιπα γενώματα. Η παραπάνω παρατήρηση ποσοτικοποιήθηκε εντοπίζοντας τα clusters που περιέχουν γονίδια από 155-157 στελέχη και ταυτοποιώντας τα γενώματα στα οποία αυτά τα γονίδια λείπουν. Δύο στελέχη βρέθηκε ότι εμφανίζουν μεγάλο αριθμό απόντων cluster που εντοπίζονται στα υπόλοιπα γενώματα (**Πίνακας 5**).

Το παραπάνω ενδεχομένως να οφείλεται σε δεδομένα NGS κακής ποιότητας. Πράγματι το στέλεχος με accession number GCA_900455245.1 έχει αποκλειστεί από την βάση δεδομένων Refseq, καθώς περιέχει πολλές πρωτεΐνες με μετατόπιση του πλαισίου ανάγνωσης (frameshifted proteins). Με βάση τον πίνακα 5, αφαιρέσαμε από περαιτέρω ανάλυση τα στελέχη με accession number GCA_900455065.1 και GCA_900455245.1, καθώς θεωρήθηκε ότι η τόσο μεγάλη έλλειψη γονιδίων σε clusters τα οποία συναντώνται στα 155 με 157 υπόλοιπα στελέχη είναι αποτέλεσμα της όχι τόσο ακριβούς αλληλούχισής τους. Έτσι, καταλήξαμε σε 156 στελέχη τα οποία χρησιμοποιήσαμε για περαιτέρω ανάλυση.

Πίνακας 5. Στελέχη που εμφανίζουν υψηλό αριθμό απόντων cluster που εντοπίζονται στα υπόλοιπα 155-157 γενώματα. Ο αριθμός των απόντων clusters έχει υπολογιστεί χρησιμοποιώντας τα τελικά κριτήρια CD-HIT που χρησιμοποιήσαμε για την ανάλυση του core genome.

Γενώματα	Απώντα clusters
GCA_900455065.1	656
GCA_900455245.1	587
GCA_001647515.1	242
GCA_002013325.1	213
GCA_002206145.2	159
GCA_002184635.2	156
GCA_901485075.1	99
GCA_001643755.1	84
GCA_900455055.1	59

Από την πρώτη ομαδοποίηση παρατηρήθηκε ότι σε αρκετά cluster εμπεριέχονταν πρωτεΐνες με διαφορετική λειτουργία, όπως αυτή αποδόθηκε από τους αλγόριθμους του PROKKA. Δηλαδή τα κριτήρια ομαδοποίησης που χρησιμοποιήθηκαν δημιουργούσαν cluster όπου εκτός από ορθόλογες υπήρχαν και παράλογες πρωτεΐνες. Συνεπώς απαιτούνταν να βρεθούν τα βέλτιστα κριτήρια ομαδοποίησης ώστε να μειωθεί η ενσωμάτωση παράλογων πρωτεϊνών στα κοινά cluster χωρίς όμως τα ορθόλογα γονίδια να ταξινομούνται σε διαφορετικές ομάδες. Έτσι λοιπόν επαναλήφθηκαν οι αναζητήσεις μέσω CD-HIT εφαρμόζοντας διαφορετική παραμετροποίηση που αφορούσε τα κριτήρια ταύτισης αλληλουχίας (-c), το ποσοστό στοίχισης των αλληλουχιών που θα ταξινομηθούν στο ίδιο cluster (-aL), τη διαφορά στο μήκος της αλληλουχίας των πρωτεϊνών που ταξινομούνται στο ίδιο cluster (-s) και το ελάχιστο μήκος που έχουν οι πρωτεΐνες που ομαδοποιούνται από το CD-HIT (-l). Για κάθε συνδυασμό παραμέτρων προσδιορίστηκε το μέγεθος του πανγενώματος, του core genome καθώς και τα cluster του core genome που περιείχαν παράλογες πρωτεΐνες (**Πίνακας 6**).

Πίνακας 6. Εύρεση των βέλτιστων παραμέτρων του CD-HIT.

Κριτήρια				Pangenome	Core genome	Clusters με παράλογα
Ελάχιστο μήκος (-l)	Ποσοστό ομοιότητας (-c)	Ποσοστό στοίχισης (-aL)	Διαφορά μήκους πρωτεϊνών (-s)			
120	0.6	0.5	-	11033	1627	113
120	0.6	0.5	0.5	11033	1627	113
120	0.6	0.7	-	12459	1157	62
120	0.6	0.7	0.5	12459	1157	62
120	0.6	0.7	0.7	12459	1157	62
120	0.6	-	0.5	10615	1751	122
120	0.6	-	0.7	12075	1219	65
80	0.6	0.5	-	13509	1835	128
60	0.6	0.5	-	14791	1893	130

120	0.7	0.5	-	11563	1597	102
120	0.8	0.5	-	12294	1471	80
120	0.9	0.5	-	14192	991	40
120	0.7	0.7	-	12975	1137	56
120	0.8	0.7	-	13696	1046	42
120	0.9	0.7	-	15530	730	20
120	0.6	0.8	-	13354	998	49
120	0.7	0.8	-	13846	985	45
120	0.8	0.8	-	14543	908	34
120	0.9	0.8	-	16357	637	15
60	0.75	0.7	-	17772	1274	57

Από τον παραπάνω πίνακα παρατηρούμε ότι το κύριο κριτήριο το οποίο ουσιαστικά αλλάζει τον αριθμό των clusters του core genome και του pangenome είναι το aL. Το -s αν και αντίστοιχο με το -aL φαίνεται να μην είναι τόσο αυστηρό με το -aL, καθώς όταν έχουμε και τα δύο κριτήρια υπερισχύει το κριτήριο του -aL και είναι σαν να μην υπάρχει το κριτήριο -s. Έτσι, ανάμεσα σε αυτά τα δύο αποφασίστηκε να χρησιμοποιηθεί το κριτήριο -aL. Επιπλέον, βλέπουμε μία πολύ μεγάλη μείωση του core genome και των clusters που περιέχουν παράλογα για την αλλαγή του -aL από το 50% στο 70%. Αποφασίσαμε να χρησιμοποιήσουμε το πιο αυστηρό κριτήριο, δηλαδή -aL 0.7, καθώς με το 50% μπορεί το CD-HIT να ομαδοποιούσε στο ίδιο cluster πρωτεΐνες οι οποίες να είχαν απλώς μια κοινή επικράτεια (domain) αλλά να μην ήταν λειτουργικά παρόμοιες.

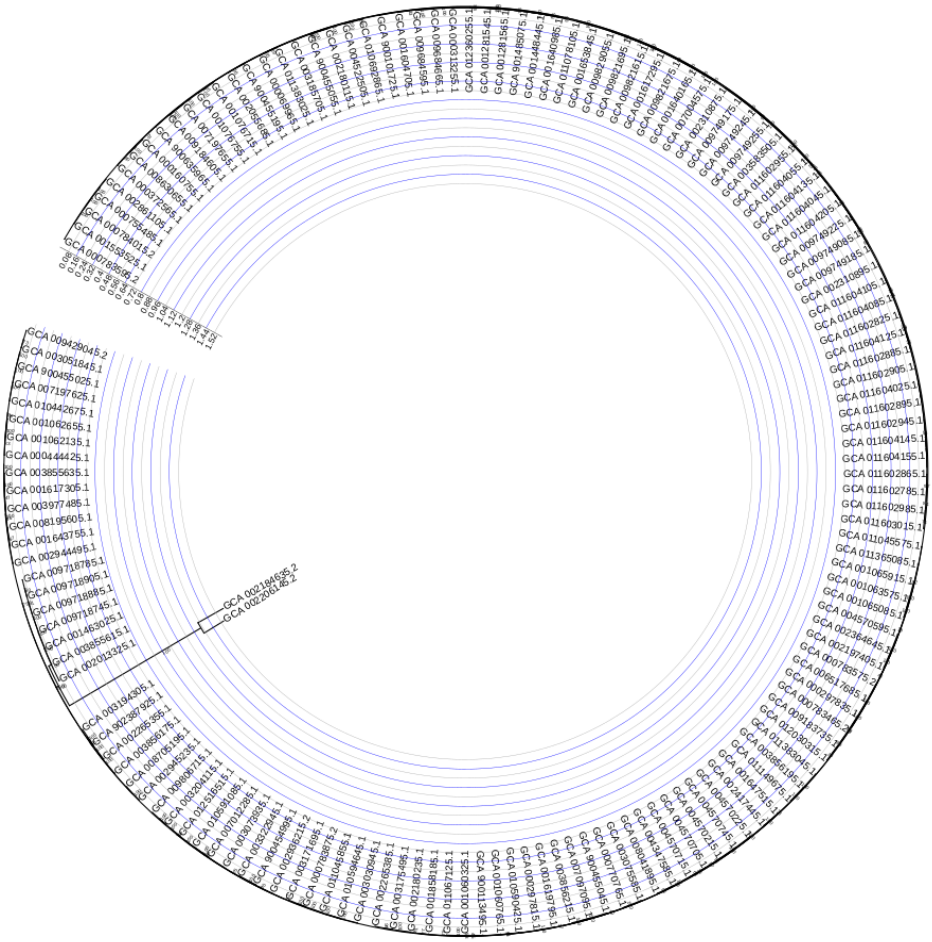
Όσον αφορά στο κριτήριο του ποσοστού ομοιότητας μεταξύ των αλληλουχιών παρατηρούμε ότι όταν αυξάνεται το ποσοστό μειώνεται ο αριθμός του core genome και των cluster με παράλογα, αλλά όχι κατά πολύ μεγάλο αριθμό. Για την τελική ανάλυση χρησιμοποιήθηκε ποσοστό ταύτισης αλληλουχίας 75% (δηλαδή κριτήριο στο cdhit -c 0.75). Στο ποσοστό ομοιότητας 75%, ελέγξαμε τα clusters με παράλογες πρωτεΐνες, και παρότι υπήρχαν διαφορές στο annotation του Prokka, οι διαφορές ήταν ανούσιες και όλες οι πρωτεΐνες σε όλα τα cluster επιτελούσαν την ίδια κατηγορία λειτουργιών.

Τέλος, για το κριτήριο του μήκους των αλληλουχιών βλέπουμε ότι όσο αυξάνουμε το κριτήριο αυτό μειώνεται το core genome. Με βάση το άρθρο των Tiessen et al., το 2012, οι οποίοι περιέγραψαν για διάφορα βακτήρια το μέγεθος των πρωτεϊνών τους, η *Escherichia coli* είχε μέσο μήκος πρωτεϊνών το 287 αμινοξέα και το 90% των πρωτεϊνών της είχαν μήκος μεγαλύτερο από 58 αμινοξέα. Έτσι καθώς η *E. coli* είναι το πιο κοντινό από τα βακτήρια που ανέλυσε το παραπάνω άρθρο στο *P. mirabilis*, θεωρήσαμε ότι παρόμοια κατανομή μήκους πρωτεϊνών θα παρουσιάζει και ο *P. mirabilis* και γι' αυτό αποφασίσαμε να έχουμε όριο το 60, δηλαδή -l 60, ώστε και να αφαιρέσουμε τα ψευδογονίδια από την ανάλυση μας, αλλά και να μη χάσουμε πολλές μικρές πρωτεΐνες.

Συνεπώς, οι τελικές παράμετροι για το CD-HIT είναι **-c 0.75 -aL 0.7 -l 60**.

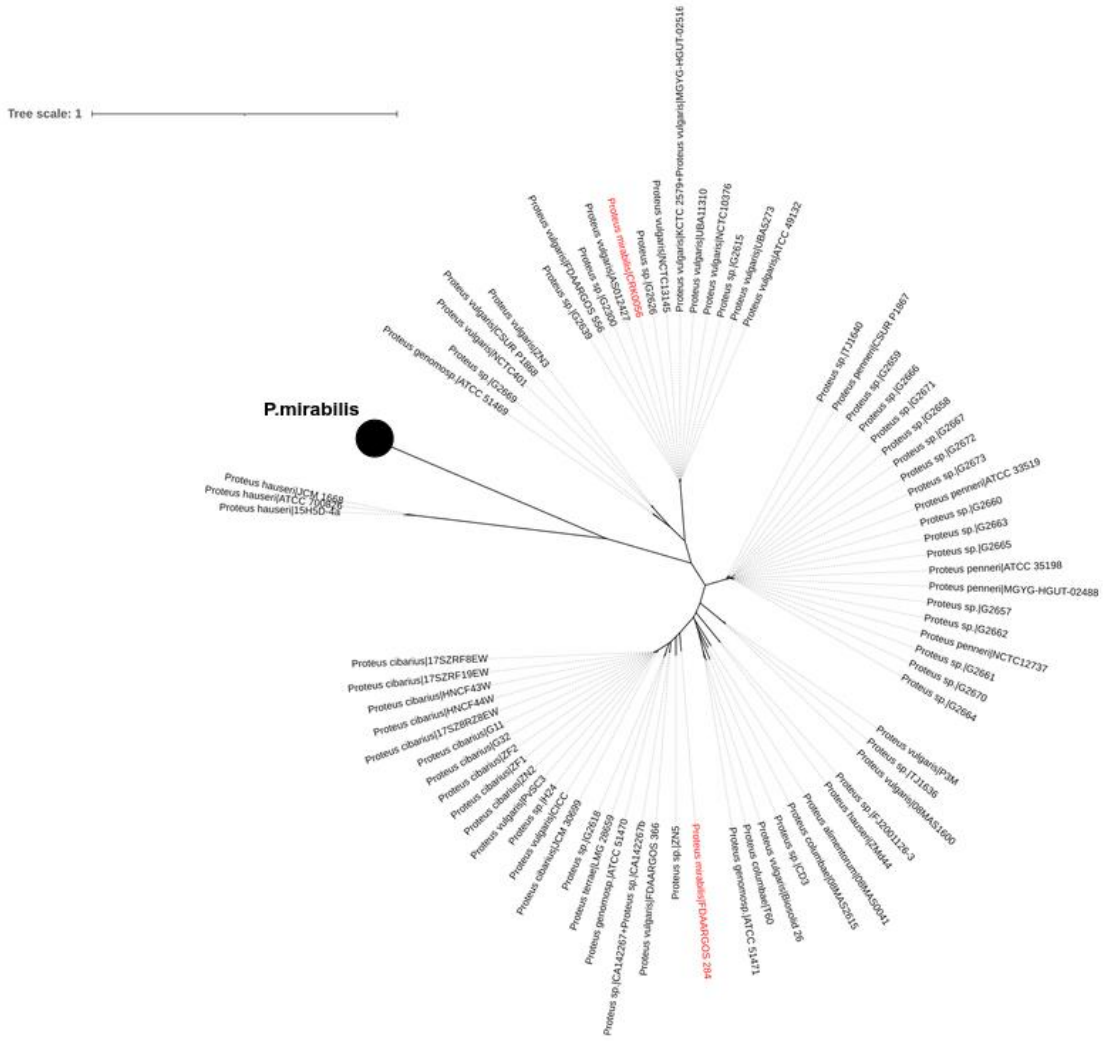
Βάσει της παραπάνω ομαδοποίησης και αφαιρώντας τα cluster που περιείχαν πολλαπλά αντίγραφα σε αρκετά γενώματα, το core genome για τα 156 γενώματα που αναλύθηκαν περιείχε 1268 γονίδια. Στην συνέχεια έγινε στοίχιση των κωδικών περιοχών του κάθε γενετικού τόπου με το πρόγραμμα Clustal Omega, για κάθε γένωμα οι στοιχισμένες αλληλουχίες για όλους τους γενετικούς τόπους ενώθηκαν, δημιουργώντας έτσι «ψευδογενώματα» core genome για κάθε στέλεχος, και οι φυλογενετικές σχέσεις εκτιμήθηκαν με φυλογένεση Maximum Likelihood (ML) χρησιμοποιώντας το πρόγραμμα RaxML. Συνολικά έλαβαν χώρα 100 προσομοιώσεις φυλογένεσης ενώ η στατιστική σημαντικότητα των διακλαδώσεων εκτιμήθηκε με 1000 βήματα bootstrap. Το δέντρο με τη μεγαλύτερη πιθανότητα δίνεται στην **εικόνα 11**.

Παρατηρήθηκε ότι 2 γενώματα (*P. mirabilis* CRK0056 με GenBank accession number GCA_002184635 και *P. mirabilis* FDAARGOS_284 με GenBank accession number GCA_002206145) εμφάνιζαν υψηλή εξελικτική απόσταση από τα υπόλοιπα στελέχη *P. mirabilis,* γεγονός που υποδείκνυε ενδεχόμενη εσφαλμένη ταξινόμησή τους από το NCBI. Σε αυτό συνηγορούσε και το γεγονός ότι και στα δύο γενώματα εντοπίζονταν γονίδιο τάξης A β-λακταμάσης που ανήκε στην φυλογενετική γραμμή των χρωμοσωμικών β-λακταμασών των άλλων ειδών του γένους, ενώ είναι γνωστό ότι τα *P. mirabilis* δε φέρουν φυσικά γονίδιο αυτού του τύπου.



Εικόνα 11. Φυλογένεση Maximum Likelihood των 156 στελεχών *P. mirabilis* χρησιμοποιώντας 1268 γενετικούς τόπους.

Με σκοπό να εξακριβωθεί η σωστή ταξινόμηση αυτών των δύο στελεχών, χρησιμοποιήσαμε το εργαλείο cg-finder με είσοδο όλα τα γονιδιώματα του γένους *Proteus* της βάσης δεδομένων RefSeq του NCBI (370 γενώματα), ώστε να δημιουργήσουμε το φυλογενετικό δέντρο του γένους Proteus. Βρέθηκε ότι αρκετά στελέχη *P. vulgaris* και *P. cibarius* έφεραν και τους 1268 τόπους, γεγονός που υποδείκνυε ότι τα γονίδια αυτά δεν αντιστοιχούν στο core genome του *P. mirabilis* αλλά σε αυτό μιας ευρύτερης ομάδας βακτηρίων του γένους Proteus. Οι φυλογενετικές αναλύσεις με το RaxML χρησιμοποιώντας τους γενετικούς τόπους που εντοπίζονταν σε όλα τα γενώματα που αναλύθηκαν (665 κοινοί γενετικοί τόποι) έδειξε ότι πράγματι τα δύο προβληματικά στελέχη δεν ταξινομούνται με τα *P. mirabilis* αλλά ανήκουν σε άλλες φυλογενετικές γραμμές (**Εικόνα 12)**. Συγκεκριμένα το *P. mirabilis* CRK0056 ομαδοποιείται με στελέχη *P. vulgaris* ενώ το *P. mirabilis* FDAARGOS_284 σχηματίζει διακριτή εξελικτική γραμμή συγγενική με τα *P. columbae* και *P. terrae* (**Εικόνα 12).**



Εικόνα 12. Φυλογένεση Maximum Likelihood των στελεχών του γένους *Proteus* με γένωμα στη βάση δεδομένων RefSeq. Το δέντρο δημιουργήθηκε χρησιμοποιώντας 665 γενετικούς τόπους που εντοπίζονταν σε όλα τα στελέχη. Τα γενώματα στα οποία έλειπε περισσότερο από 8% των γενετικών τόπων εξαιρέθηκαν από την ανάλυση. Τα προβληματικά στελέχη *P. mirabilis* CRK0056 (GCA_002184635) και *P. mirabilis* FDAARGOS_284 (GCA_002206145) σημειώνονται με κόκκινο.

Συνεπώς, για να οριστεί το core genome του *P. mirabilis* τα δύο αυτά στελέχη απαιτούνταν να εξαιρεθούν από τις αναλύσεις. Επαναλάβαμε τη διαδικασία και εφαρμόζοντας τα βέλτιστα κριτήρια CD-HIT στο πρωτέωμα των 154 τελικών επιλεχθέντων γενωμάτων καταλήξαμε στο core genome του είδους που περιγράφεται παρακάτω.

3.3. Πανγονιδίωμα και core genome του *P. mirabilis*

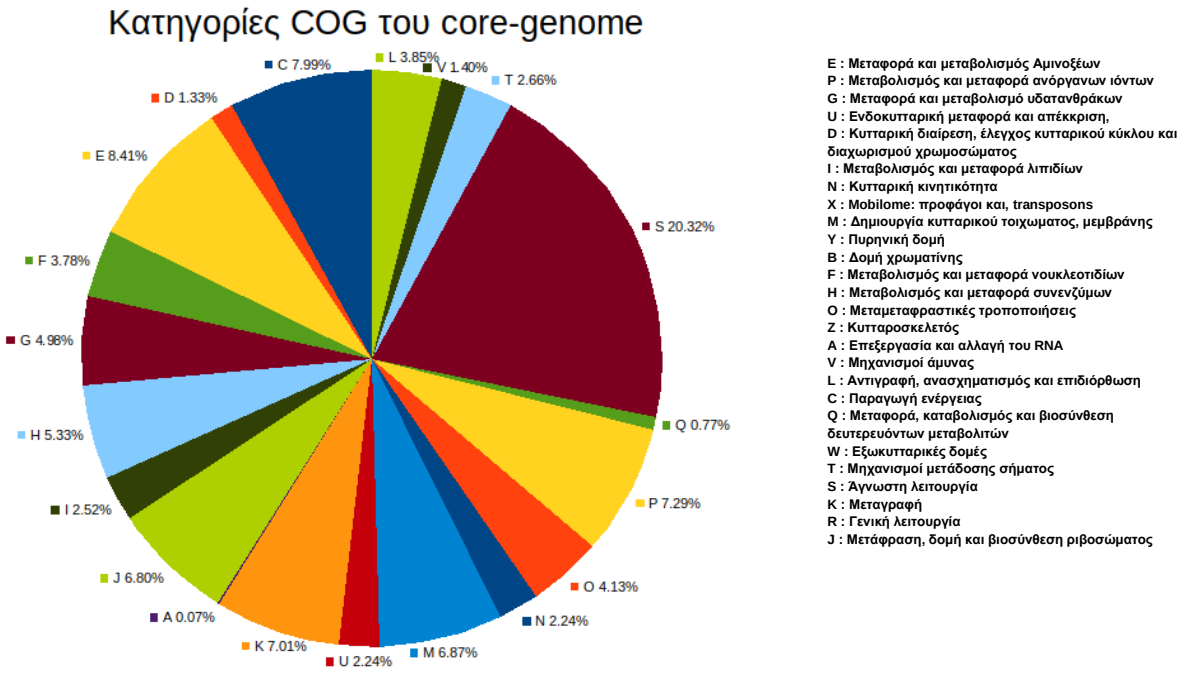
Από την ομαδοποίηση των πρωτεϊνών, προέκυψαν 16598 οικογένειες πρωτεϊνών (clusters). Συνεπώς το πανγονιδίωμα του *P. mirabilis* αποτελείται από 16598 γονίδια, τα αντιπροσωπευτικά γονίδια του κάθε cluster. Στο **παράρτημα Α** παρουσιάζεται η βάση δεδομένων των γονιδίων του πανγονιδιώματος του *P. mirabilis*. Στο **παράρτημα Β** δίνονται ορισμένα χαρακτηριστικά των clusters του πανγονιδιώματος, όπως η αντιπροσωπευτική πρωτεΐνη του κάθε cluster, το μέγιστο/ελάχιστο μήκος της, το όνομα του γονιδίου της, η λειτουργία της, τα στελέχη τα οποία περιλαμβάνονται στο cluster, και η COG κατηγορία της αντιπροσωπευτικής πρωτεΐνης.
Η ομαδοποίηση των πρωτεϊνών επίσης αποκάλυψε 1407 οικογένειες πρωτεϊνών που εντοπίζονταν και στα 154 γενώματα που αναλύσαμε. Από αυτές αφαιρέθηκαν 10 clusters, καθώς περιλάμβαναν παραπάνω από ένα αντίγραφο για παραπάνω από 10 στελέχη (**Πίνακας 7**).

Πίνακας 7. Οικογένειες πρωτεϊνών που αφαιρέθηκαν από το core genome.

Cluster	Μέγιστο μήκος	Ελάχιστο μήκος	Όνομα Γονιδίου	Λειτουργικός σχολιασμός (αριθμός πρωτεϊνών)	Αριθμός πρωτεϊνων για το cluster	Μέγιστος αριθμός πρωτεϊνών που ανήκουν στο ίδιο γένωμα	Γενώματα με παραπάνω από μία πρωτεΐνη στο cluster
14071	94	92	pduA_2,pduA_3,pduA_4,pduA_1	Propanediol utilization protein PduA(456)	456	3	153
15881	70	70	cspB,cspA_1,cspA_2,cspC,cspA_3,cspA,cspA_4,cspB_2	Cold shock protein CspA(105),Cold shock-like protein CspC(153),Cold shock-like protein CspB(152)	410	5	153
2064	450	334	cydA_2,cydA_3,cydA_1,appC_2	Cytochrome bd-II ubiquinol oxidase subunit 1(1),Cytochrome bd ubiquinol oxidase subunit 1(152),Cytochrome bd-I ubiquinol oxidase subunit 1(151)	304	2	150
3239	370	356	fliC2_1,fliC2_2,fliC2,fliC1,fliC1_2	Flagellin 2(184),Flagellin 1(118)	302	2	148
6569	239	180	-	hypothetical protein(448)	448	6	144
15898	70	69	cspE_1,cspE_2,cspC,cspA_2,cspA_1,cspA,cspA_3,cspE_3,cspB_1,cspE	Cold shock-like protein CspC(1),Cold shock-like protein CspE(155),Cold shock protein	309	3	140

				CspA(152),Cold shock-like protein CspB(1)			
10391	153	150	-	hypothetical protein(216)	216	3	56
1632	495	362	-	hypothetical protein(215)	215	3	55
10436	152	152	-	hypothetical protein(213)	213	3	54
8517	189	144	ssb_4,ssb_1,ssb_2,ssb_3,ssb,ssb_5,ssb_6	Single-stranded DNA-binding protein(217)	217	4	54
14071	94	92	pduA_2,pduA_3,pduA_4,pduA_1	Propanediol utilization protein PduA(456)	456	3	153
15881	70	70	cspB,cspA_1,cspA_2,cspC,cspA_3,cspA,cspA_4,cspB_2	Cold shock protein CspA(105),Cold shock-like protein CspC(153),Cold shock-like protein CspB(152)	410	5	153
2064	450	334	cydA_2,cydA_3,cydA_1,appC_2	Cytochrome bd-II ubiquinol oxidase subunit 1(1),Cytochrome bd ubiquinol oxidase subunit 1(152),Cytochrome bd-I ubiquinol oxidase subunit 1(151)	304	2	150

Συνεπώς συνολικά στο core genome περιλαμβάνονται οι αντιπροσωπευτικές αλληλουχίες από 1397 clusters (**Παράρτημα Γ**). Για τα στελέχη που χρησιμοποιήσαμε στην ανάλυση, το core genome αποτελούσε κατά μέσο όρο το 35% του γενώματος κάθε στελέχους, και το 8,5% του πανγενώματος του *P. mirabilis*.
Οι αντιπροσωπευτικές αλληλουχίες κάθε cluster χαρακτηρίστηκαν λειτουργικά μέσω του προγράμματος EggNOG-mapper και ταξινομήθηκαν σε λειτουργικές κατηγορίες COG (**Εικόνα 13).**

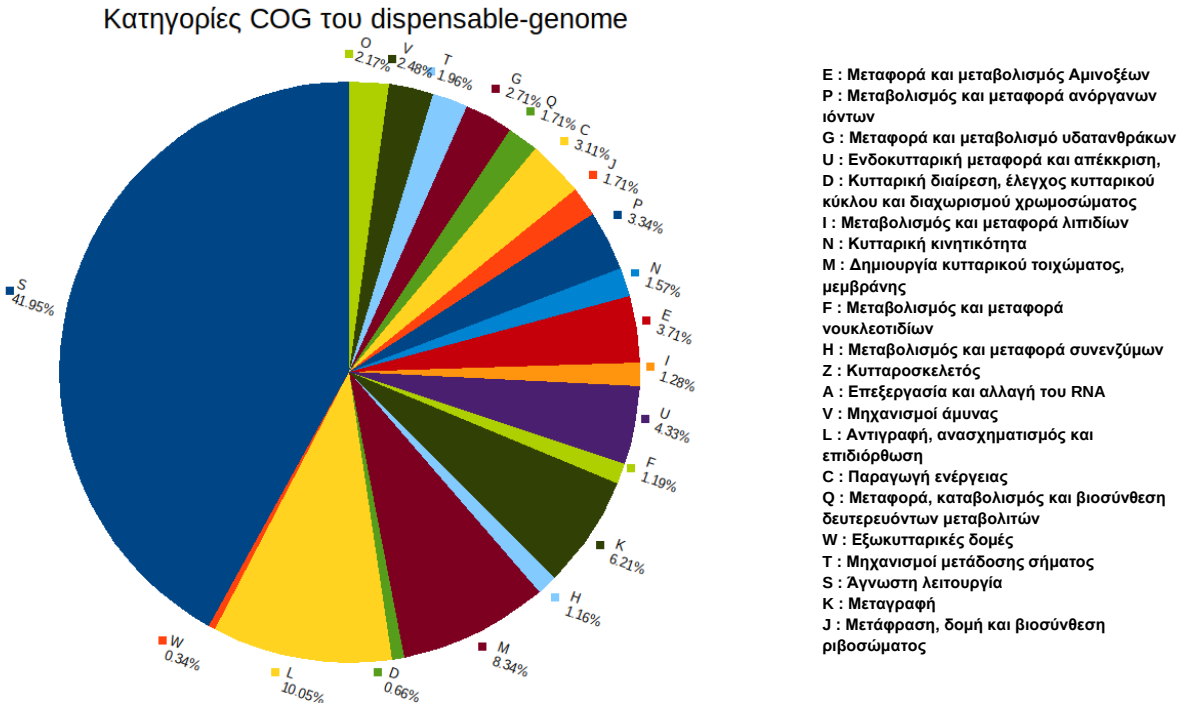


Εικόνα 13. Κατανομή των κατηγοριών COG του core genome του *P. mirabilis*

Παρατηρούμε ότι στο core genome, ένα μεγάλο ποσοστό (~20%) αφορά σε πρωτεΐνες με άγνωστη λειτουργία. Ποσοστό περίπου ~40% των γονιδίων του core genome αφορά σε λειτουργίες που σχετίζονται με το μεταβολισμό, και ταξινομείται σε κατηγορίες COG όπως μεταφορά και μεταβολισμό αμινοξέων, ανόργανων ιόντων, υδατανθράκων, νουκλεοτιδίων, λιπιδίων, συνενζύμων, δευτερευόντων μεταβολιτών και παραγωγή ενέργειας. ~14% των γονιδίων του core genome σχετίζονται με την μεταγραφή και τη μετάφραση, 7% με τη σύνθεση του κυτταρικού τοιχώματος, μόλις ~4% σχετίζεται με την αντιγραφή. Τέλος, ~14% του core genome σχετίζεται με λοιπές κυτταρικές διεργασίες, όπως κυτταρική κινητικότητα, μεταμεταφραστικές τροποποιήσεις/αντικατάσταση πρωτεϊνών, μηχανισμοί μετάδοσης σήματος, ενδοκυτταρική μεταφορά/απέκκριση και μηχανισμοί άμυνας.
Όσον αφορά στα γονίδια για ανθεκτικότητα σε αντιβιοτικά, το AMRFinder δεν εντόπισε γονίδια ανθεκτικότητας στο core genome.

3.4. Dispensable genome του *P. mirabilis*

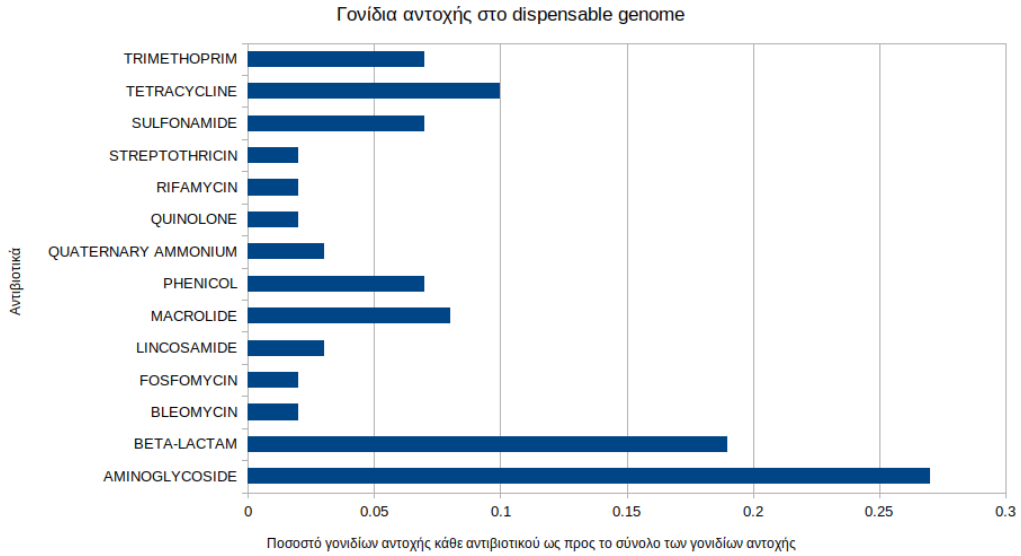
Το dispensable genome περιλαμβάνει τις αντιπροσωπευτικές αλληλουχίες από 7684 clusters, και αποτελεί το 46,3% του pangenome. Παρακάτω παρουσιάζεται το pie chart με τα ποσοστά των διαφόρων κατηγοριών COG για τα γονίδια του dispensable genome, όπως προέκυψε από το εργαλείο EggNOG-mapper.



Εικόνα 14. Ποσοστό γονιδίων του dispensable genome του *P. mirabilis* για κάθε κατηγορία COG.

Παρατηρούμε ότι στο dispensable genome, ένα μεγάλο ποσοστό (~42%) αφορά σε πρωτεΐνες με άγνωστη λειτουργία. Ποσοστό περίπου 18% των γονιδίων του dispensable genome αφορά σε λειτουργίες που σχετίζονται με το μεταβολισμό, και ταξινομείται σε κατηγορίες COG όπως μεταφορά και μεταβολισμό αμινοξέων, ανόργανων ιόντων, υδατανθράκων, νουκλεοτιδίων, λιπιδίων, συνενζύμων, δευτερευόντων μεταβολιτών και παραγωγή ενέργειας. Περίπου 8% του dispensable genome αφορά σε γονίδια για τη μεταγραφή και τη μετάφραση, ενώ 10% αφορά σε διεργασίες της αντιγραφής και της επιδιόρθωσης του γενετικού υλικού, ~8 στη σύνθεση του κυτταρικού τοιχώματος. Τέλος, ~14% του dispensable genome σχετίζεται με λοιπές κυτταρικές διεργασίες, όπως κυτταρική κινητικότητα, μεταμεταφραστικές τροποποιήσεις/αντικατάσταση πρωτεϊνών, μηχανισμοί μετάδοσης σήματος, ενδοκυτταρική μεταφορά/απέκκριση, μηχανισμοί άμυνας και εξωκυτταρικές δομές.

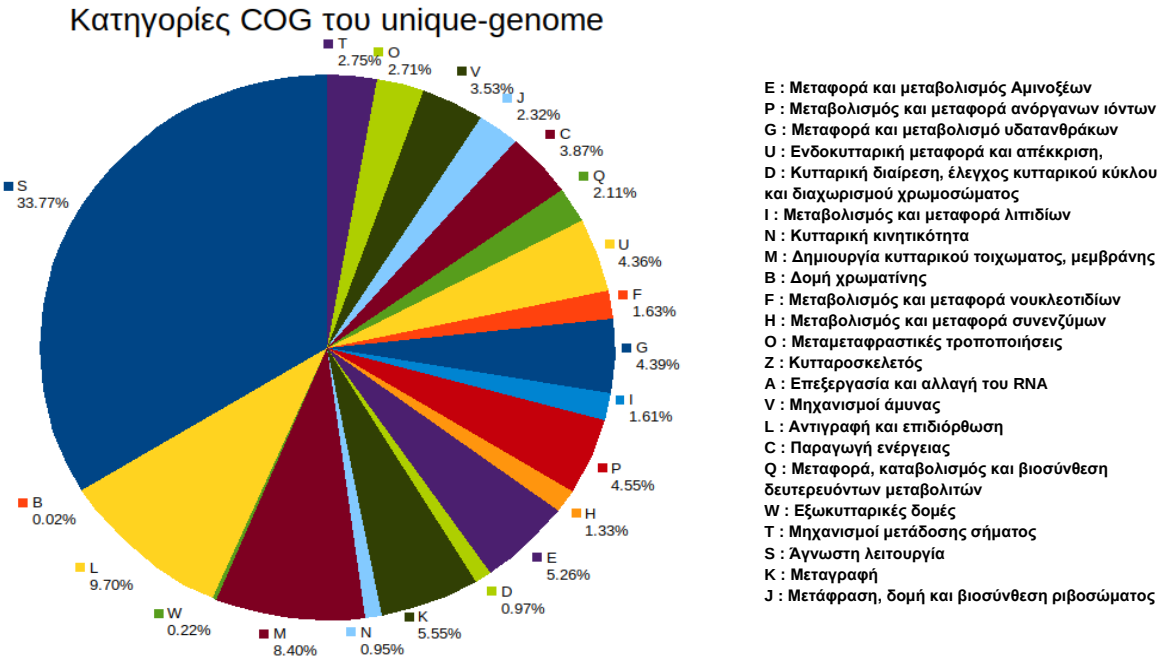
Όσον αφορά στα γονίδια για ανθεκτικότητα σε αντιβιοτικά, το AMRFinder εντόπισε 59 γονίδια ανθεκτικότητας στο dispensable genome. Από τα 59 αυτά γονίδια, 16 ήταν γονίδια αντοχής σε αμινογλυκοσίδες, 11 σε β-λακταμικά αντιβιοτικά, 1 σε μπλεομυκίνη, 1 σε φωσφομυκίνη, 2 σε λινκοσαμίδη, 5 σε μακρολίδες, 4 σε φαινικόλες, 2 σε τριτορικό αμμώνιο, 1 σε κινολόνες, 1 σε ριφαμπικίνη, 1 σε στρεπτομυκίνη, 4 σε σουλφοναμίδη, 6 σε τετρακυκλίνη και 4 σε τριμεθοπρίμη.



Εικόνα 15. Ιστόγραμμα που δείχνει τα συγκεντρωτικά αποτελέσματα για τα γονίδια αντοχής στο dispensable genome όπως πάρθηκαν από το AMRFinder.

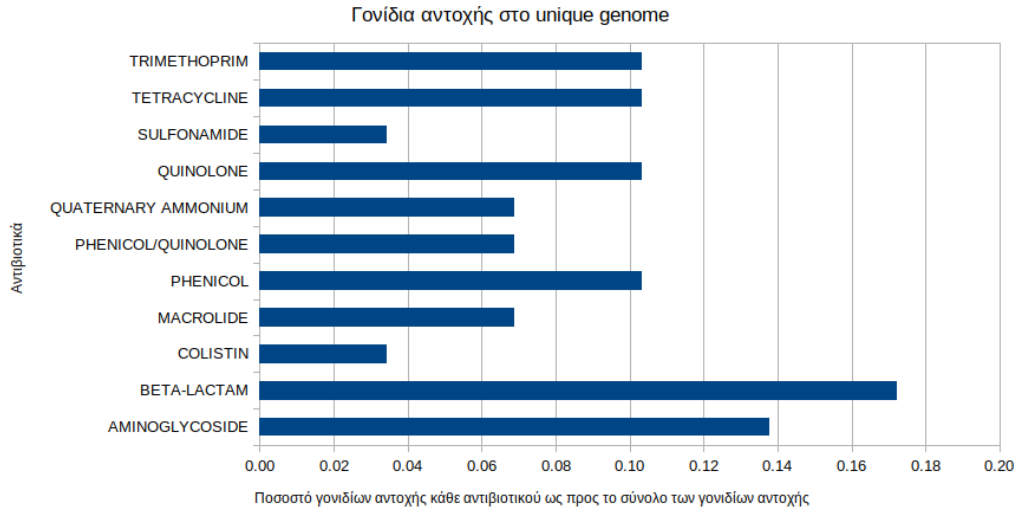
3.5. Unique genome του *P. mirabilis*

Το unique genome περιλαμβάνει τις αντιπροσωπευτικές αλληλουχίες από 7507 clusters, και αποτελεί το 45,2% του pangenome. Παρακάτω παρουσιάζεται το pie chart με το ποσοστό των διαφόρων κατηγοριών COG στο unique genome.



Εικόνα 16. Μας δείχνει το ποσοστό γονιδίων του unique genome για κάθε κατηγορία COG.

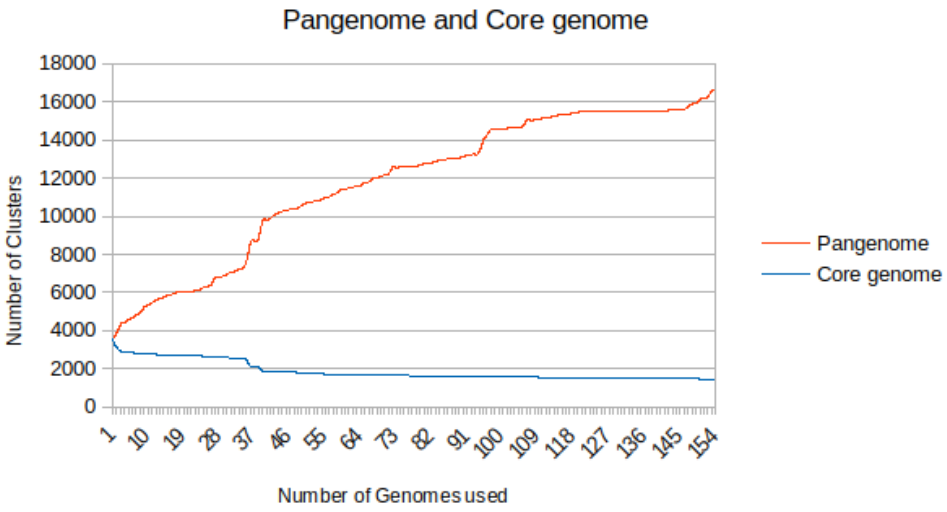
Ποσοστό περίπου 24,5% των γονιδίων του unique genome αφορά σε λειτουργίες που σχετίζονται με το μεταβολισμό, και ταξινομείται σε κατηγορίες COG όπως μεταφορά και μεταβολισμό αμινοξέων, ανόργανων ιόντων, υδατανθράκων, νουκλεοτιδίων, λιπιδίων, συνενζύμων, δευτερευόντων μεταβολιτών και παραγωγή ενέργειας. Περίπου 8% του unique genome αφορα σε γονίδια για τη μεταγραφή και τη μετάφραση, ενώ 10% αφορά σε διεργασίες της αντιγραφής και της επιδιόρθωσης του γενετικού υλικού, ~8 στη σύνθεση του κυτταρικού τοιχώματος. Ενώ ~12% σχετίζεται με λοιπές κυτταρικές διεργασίες, όπως κυτταρική κινητικότητα, μεταμεταφραστικές τροποποιήσεις/αντικατάσταση πρωτεϊνών, μηχανισμοί μετάδοσης σήματος, ενδοκυτταρική μεταφορά/απέκκριση, μηχανισμοί άμυνας και εξωκυτταρικές δομές. Τέλος, ~34% του unique genome είναι πρωτεΐνες άγνωστης λειτουργίας.
Όσον αφορά στα γονίδια για ανθεκτικότητα σε αντιβιοτικά, το AMRFinder εντόπισε 29 γονίδια αντοχής στο unique genome. Από τα γονίδια αυτά, 5 αφορούν σε β-λακταμικά αντιβιοτικά, 4 σε αμινογλυκοσίδες, 1 σε κολιστίνη, 2 σε μακρολίδες, 3 σε φαινικόλες, 2 σε φαινικόλη/κινολόνη, 3 σε κινολόνες, 1 σε σουλφοναμίδη, 3 σε τετρακυκλίνη, 3 σε τριμεθοπρίμη και 2 στο τριτορικό αμμώνιο.



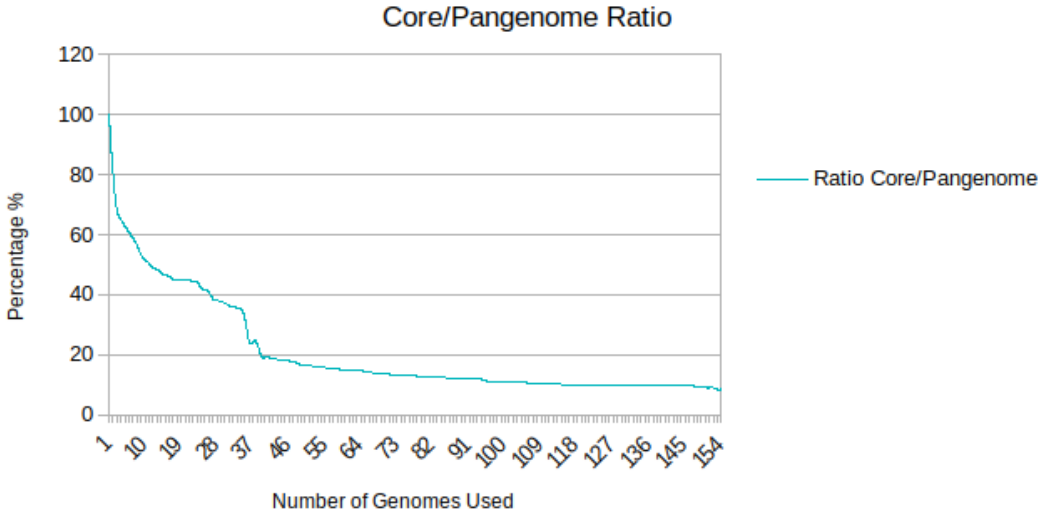
Εικόνα 17. Ιστόγραμμα που δείχνει τα συγκεντρωτικά αποτελέσματα για τα γονίδια αντοχής στο unique genome όπως πάρθηκαν από το AMRFinder.

3.6. Ανοικτό/κλειστό πανγονιδίωμα του P. mirabilis

Με την ανάλυση για το ανοικτό/κλειστό πανγονιδίωμα καταλήξαμε σε ένα συγκεντρωτικό πίνακα που περιέχει τον αριθμό των γενωμάτων, το μέγεθος του πανγονιδιώματος, το μέγεθος του core genome και το core/pangenome ratio για κάθε αριθμό χρησιμοποιούμενων γενωμάτων. Με τα στοιχεία αυτά δημιουργήσαμε τα 2 παρακάτω διαγράμματα.



Εικόνα 18. Το διάγραμμα μας δείχνει την τάση του pangenome και του core genome σε σχέση με τον αριθμό των γονιδιωμάτων που χρησιμοποιούνται στην ανάλυση.



Εικόνα 19. Το διάγραμμα μας δείχνει την τάση του core/pangenome ratio σε σχέση με τον αριθμό των γενωμάτων που χρησιμοποιούνται για την δημιουργία του core genome και του pangenome.

Και από τα δύο διαγράμματα φαίνεται ότι το πανγονιδίωμα του *P. mirabilis* είναι πιθανότατα ανοικτό, καθώς και στο πρώτο διάγραμμα τόσο το pangenome έχει ανοδική τάση συνεχόμενα, ενώ το core genome δεν φαίνεται να σταθεροποιείται σε μία τιμή. Αλλά και το κάτω διάγραμμα, που αφορά στο core/pangenome πηλίκο (core/pangenome ratio), δεν φαίνεται να σταθεροποιείται το πηλίκο γύρω από μία τιμή, αλλά έχει καθοδική τάση όσο αυξάνεται ο αριθμός των γονιδιωμάτων που χρησιμοποιούμε.

3.7. Ανάλυση πληθυσμών *P. mirabilis* βάσει core genome

3.7.1.core genome Multi-Locus Sequence Typing (cgMLST)

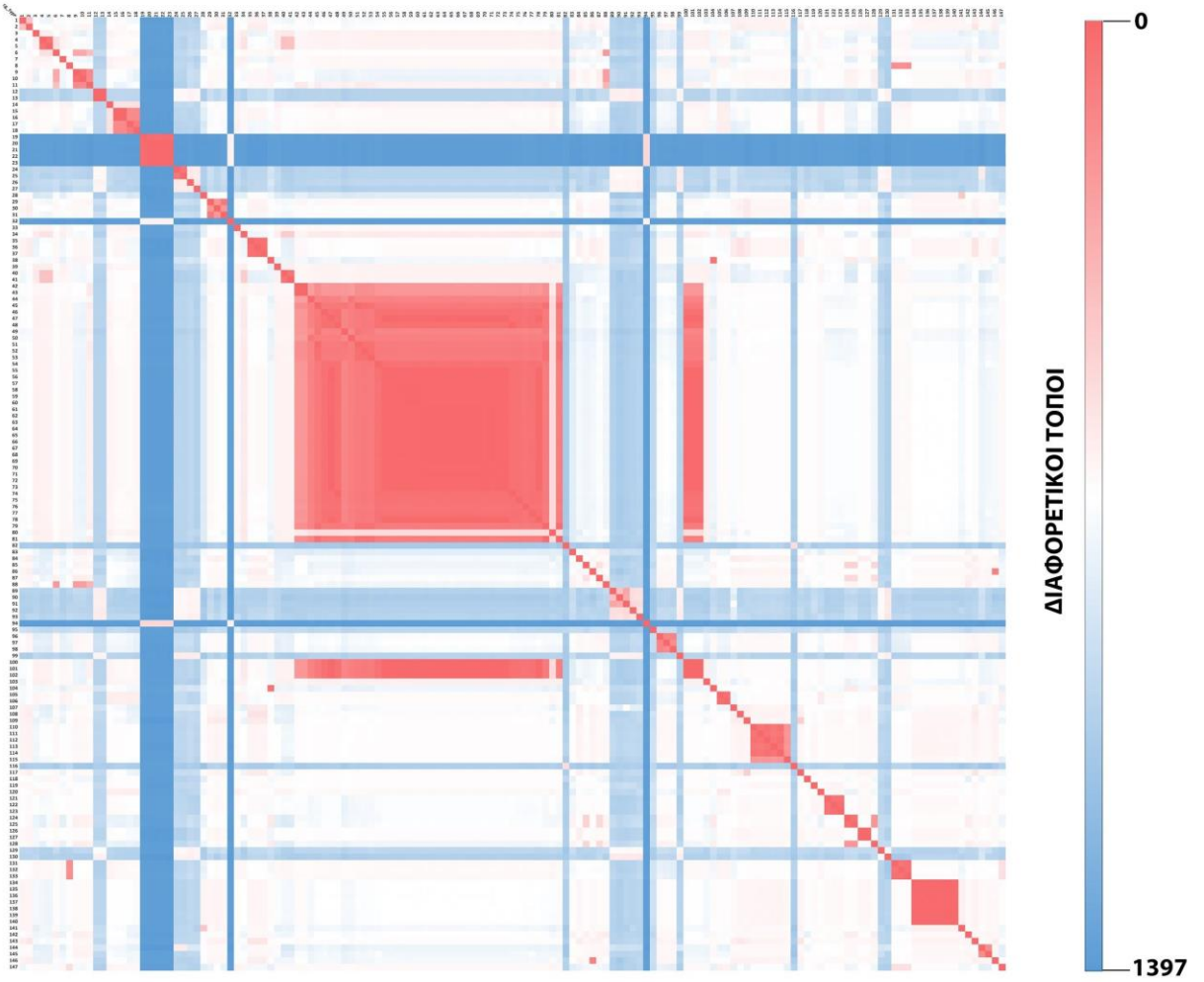
Τα διαφορετικά αλλήλια για τον κάθε γενετικό τόπο του core genome (cgPM1-cgPM1397) των 154 στελεχών *P. mirabilis* εντοπίστηκαν μέσω ομαδοποίησης των κωδικών αλληλουχιών (CDS) κάθε οικογένειας πρωτεϊνών (cluster) χρησιμοποιώντας το CD-HIT θέτοντας τις παραμέτρους –c και –s ίσες με 1. Στο **παράρτημα Δ** δίνονται ορισμένα χαρακτηριστικά για τα cgPM του core genome. Ο ελάχιστος αριθμός αλληλίων παρατηρήθηκε για τα γονίδια rplY (50S ribosomal protein), atpE (ATP synthase subunit c), acpP (Acyl carrier protein), rpsU (30S ribosomal protein), τα οποία παρουσίαζαν ένα μόνο αλλήλιο. Ενώ ο μέγιστος αριθμός αλληλίων παρατηρήθηκε για το γονίδιο gyrA (DNA gyrase subunit A), στο οποίο παρατηρήθηκαν 77 αλλήλια. Σε κάθε αλλήλιο δόθηκε ένας αριθμός και προέκυψε έτσι το core genome αλληλικό προφίλ (cgMLST τύπος) για το κάθε στέλεχος.
Η σύγκριση των 154 αλληλικών προφίλ αποκάλυψε ότι τα αναλυθέντα στελέχη ταξινομούνται σε 147 διαφορετικούς cgMLST τύπους. Στο **παράρτημα Ε** δίνεται ο πίνακας που παρουσιάζει τους cgMLST τύπους για τα 154 στελέχη που έγιναν αποδεκτά στην ανάλυση μας. Σε 7 περιπτώσεις, δύο γενώματα είχαν τον ίδιο cgMLST τύπο (cg_type).

Πίνακας 8. Στελέχη με ίδιο cgMLST τύπο

Στελέχη		**cgMLST τύπο**
P. mirabilis CCUG GCA_003194305.1	*P. mirabilis* MGYG-HGUT-02514 GCA_902387925.1	cg_type 116
P. mirabilis 51_PMIR GCA_001063575.1	*P. mirabilis* 47_PMIR GCA_001065085.1	cg_type 126
P. mirabilis XH1564 GCA_011602785.1	*P. mirabilis* XH1547 GCA_011602865.1	cg_type 61
P. mirabilis XH1561 GCA_011604085.1	*P. mirabilis* XH1559 GCA_011604105.1	cg_type 66
P. mirabilis XH1557 GCA_011604025.1	*P. mirabilis* XH1567 GCA_011604155.1	cg_type 67
P. mirabilis XH1562 GCA_011602825.1	*P. mirabilis* XH1566 GCA_011604125.1	cg_type 68
P. mirabilis XH1551 GCA_011602985.1	*P. mirabilis* XH1552 GCA_011603015.1	cg_type 69

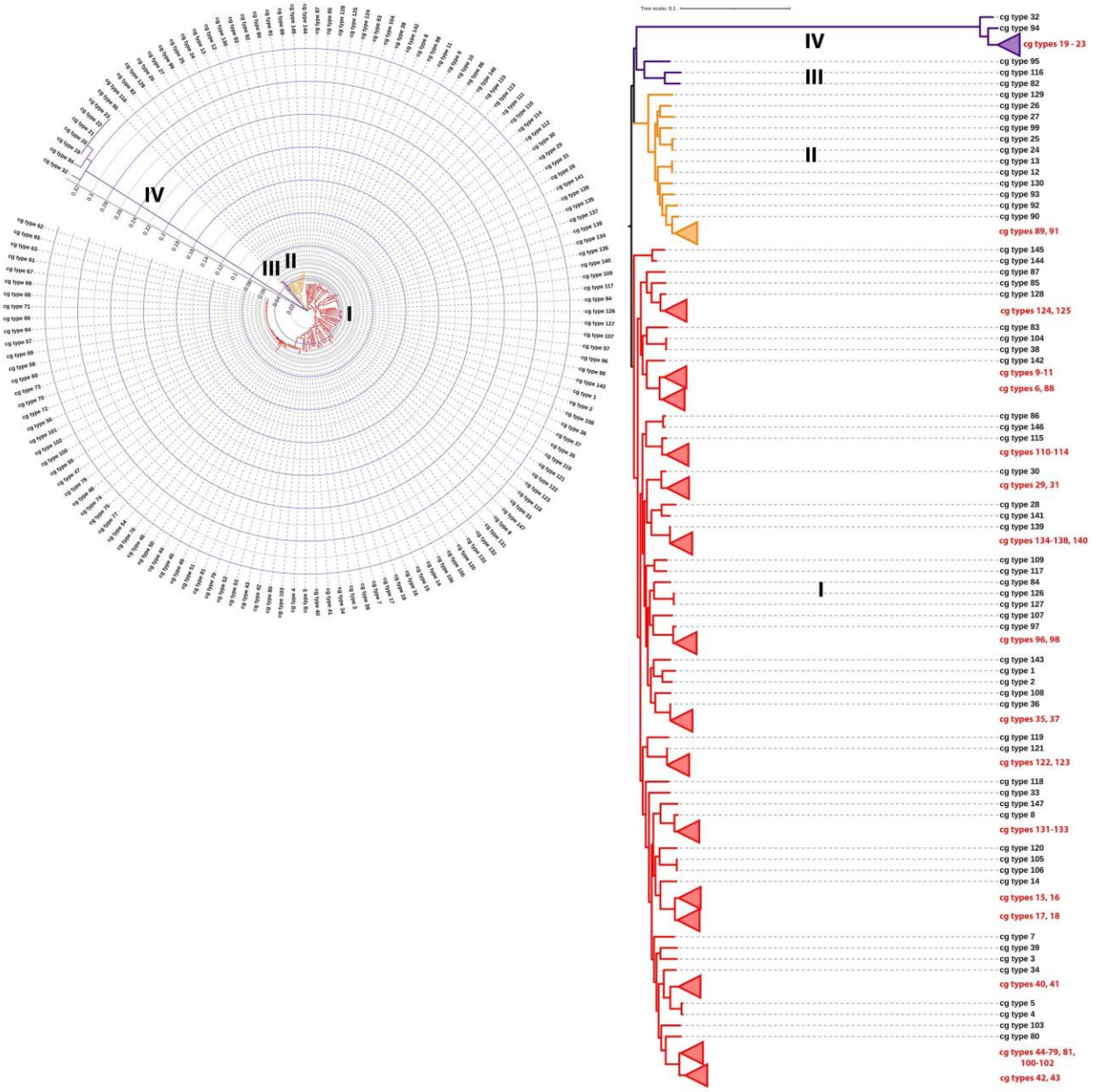
Επιπλέον η σύγκριση των αλληλικών προφίλ των διαφόρων τύπων cgMLST αποκάλυψε τα συμπλέγματα των συγγενικών κλώνων (clonal complexes), τα οποία αποτελούνται από διαφορετικούς cgMLST τύπους που διαφέρουν σε λιγότερο από 430 γενετικούς τόπους. Συνολικά παρατηρήθηκαν 21 clonal complexes (**Εικόνα 20).**
Από την παραπάνω ανάλυση προέκυψε ότι στα επιλεγμένα γενώματα, ένα σύμπλεγμα κλώνων υπέρ-αντιπροσωπεύεται. Το σύμπλεγμα αυτό αφορά στους cgMLST τύπους cg_type 42 – 79, 81 και 100-102 και περιλαμβάνει 42 cgMLST τύπους. Εντός του συμπλέγματος αυτού ένα υποσύνολο τύπων, ήτοι οι cg_type 47, 56-73, 78 και 100-102, εμφάνιζε διαφορετικά αλλήλια σε λιγότερους από 20 γενετικούς τόπους, είναι δηλαδή πολύ συγγενικοί. Παρατηρήθηκαν επίσης συμπλέγματα κλώνων μικρότερου μεγέθους (πχ cg_type 19-23, cg_type 134-140, cg_type 38/104, cg_type 105-106, cg_type 86/146). Στο σύμπλεγμα cg_type 19-23, ενώ οι διαφορές μεταξύ των τύπων που το απαρτίζουν είναι λιγότερες από 40, σε σχέση με

την πλειονότητα των υπόλοιπων τύπων διαφέρουν και στους 1397 γενετικούς τόπους, ενδεχομένως δηλαδή να πρόκειται για απομακρυσμένους εξελικτικά κλώνους σε σχέση με τα υπόλοιπα στελέχη.



Εικόνα 20: Πίνακας 147×147 που απεικονίζει τον αριθμό των διαφορετικών γενετικών τόπων μεταξύ των cgMLST τύπων. Οι περιοχές με αποχρώσεις του κόκκινου αντιστοιχούν σε διαφορές < 430 τόπων.

Οι παραπάνω παρατηρήσεις επιβεβαιώθηκαν από τη φυλογένεση ML χρησιμοποιώντας ψευδογενώματα core genome κατασκευασμένα με τις αντιστοιχίσεις των 1397 γενετικών τόπων (**Εικόνα 21**) όπου αποκαλύφθηκε ότι τα στελέχη των τύπων 19 – 23 ανήκουν σε μία απομακρυσμένη φυλογενετική γραμμή που εμπεριέχει επίσης τους τύπους 32 και 94. Τα υπόλοιπα στελέχη είναι αρκετά συγγενικά και κατανέμονται σε τρεις κύριες φυλογενετικές γραμμές. Εφαρμόζοντας ένα κριτήριο φυλογενετικής απόστασης < 0.01 σχηματίστηκαν 17 συμπλέγματα κλώνων με το περισσότερο πολυπληθές να είναι αυτό των τύπων 44-79, 81 και 100-102 (**Εικόνα 21, δεξιά**).



Εικόνα 21. Φυλογενετικές σχέσεις μεταξύ των 147 cgMLST τύπων όπως εκτιμήθηκαν με το πρόγραμμα RaxML. Ο πληθυσμός των *P. mirabilis* που αναλύθηκε χαρακτηρίζεται από τέσσερις κύριες φυλογενετικές γραμμές (αριστερά, I, II, III και IV). Ομαδοποιώντας τους cgMLST τύπους εφαρμόζοντας ένα κριτήριο φυλογενετικής απόστασης BRL < 0.01 στο iTol προέκυψαν 17 συμπλέγματα κλώνων (δεξιά).

3.7.2.Δομή του πληθυσμού P. mirabilis - επιδημικοί κλώνοι

Η καταλληλόλητα της χρήσης του core genome, ως προς την απόδοση των ορθών φυλογενετικών σχέσεων μεταξύ των διαφόρων επιδημικών κλώνων, προσδιορίστηκε με ανάλυση του συνόλου των κατατεθειμένων στην GenBank γενωμάτων *P. mirabilis* και συσχέτιση τους με ανθεκτικότητα σε αντιβιοτικά καθώς και με την γεωγραφική προέλευση. Από τα 266 γενώματα αφαιρέθηκαν αυτά των στελεχών CRK0056 και FDAARGOS_284, που όπως αναφέρθηκε έχουν εσφαλμένα ταξινομηθεί ως *P. mirabilis*, ενώ στη ανάλυση προστέθηκαν μη δημοσιευμένα δεδομένα WGS τριών ανθεκτικών στελεχών *P. mirabilis* του Εργαστηρίου Βακτηριολογίας του ΕΙΠ. Δύο από τα στελέχη αυτά είχαν απομονωθεί από κλινικά δείγματα σε ελληνικά νοσοκομεία (*P. mirabilis* ESDY17 και *P. mirabilis* EUG91) ενώ το τρίτο από λοίμωξη σε ζώο συντροφιάς (*P. mirabilis* PmE21).

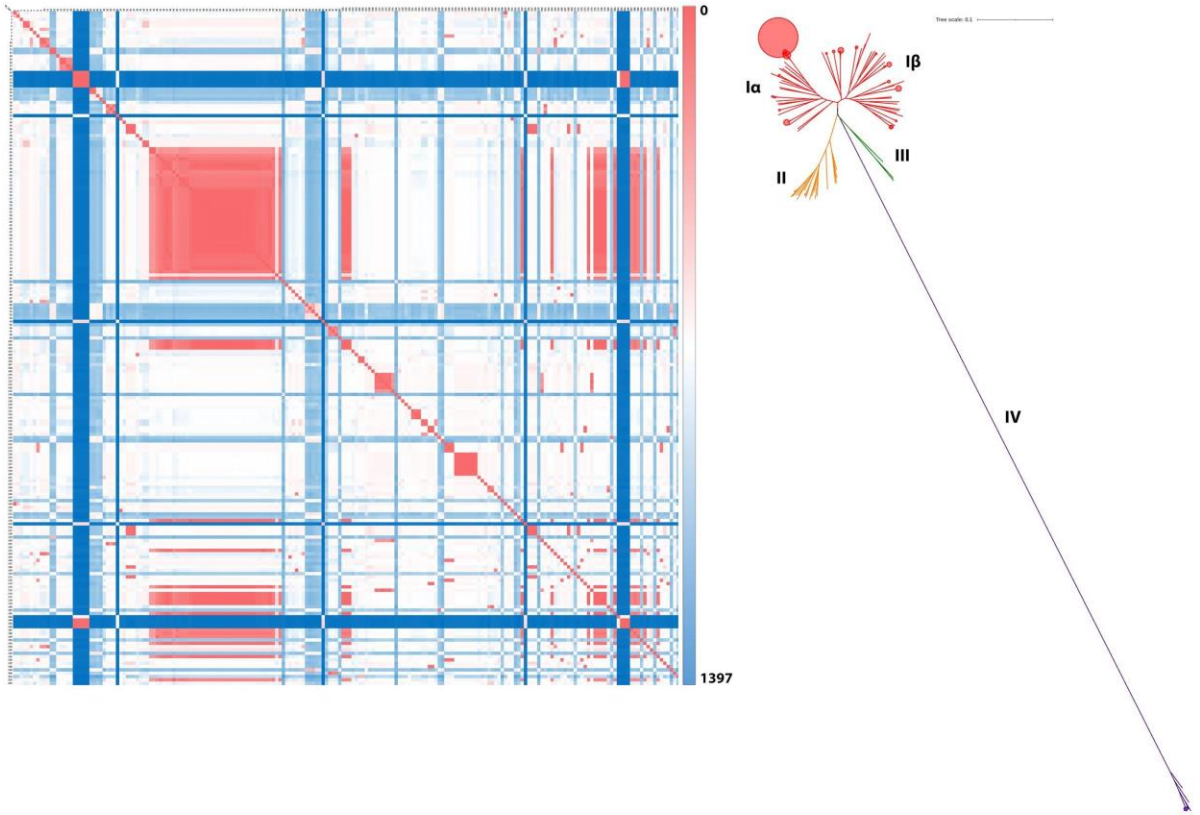
Η ανίχνευση των 1397 γονιδίων του σχήματος σε κάθε γένωμα, η ταξινόμηση σε cgMLST τύπους και η φυλογένεση των ψευδογενωμάτων core genome έγινε με το εργαλείο cg-finder που αναπτύχθηκε στην παρούσα εργασία. Γενώματα τα οποία εμφάνιζαν ποσοστό απόντων γενετικών τόπων μεγαλύτερο από 5% εξαιρούνταν από περαιτέρω ανάλυση. Από τα συνολικά 267 γενώματα που αναλύθηκαν τα επτά εξαιρέθηκαν λόγω υψηλού αριθμού απόντων τόπων ενώ στα 210 ανιχνεύθηκαν και τα 1397 γονίδια. Στα υπόλοιπα 50 γενώματα παρατηρήθηκε έλλειψη κάποιων γενετικών τόπων (**Πίνακας 9**) σύμφωνα με τα κριτήρια ανίχνευσης (blastn e-value < 10^{-5}, CDS id=70% και cov=70%). Συνήθως η έλλειψη κάποιου γενετικού τόπου αφορούσε θραύσμα με χαμηλό ποσοστό κάλυψης (%cov) που εντοπίζονταν σε κάποιο από τα δύο άκρα του αντίστοιχου contig (**Πίνακας 9**). Το παραπάνω υποδήλωνε ότι η πλειονότητα των ελλείψεων δεν ήταν φυσικές αλλά οφείλονταν σε σφάλματα των αλγορίθμων συναρμολόγησης. Συνολικά 958 γονίδια από τα 1397 ανιχνεύθηκαν σε όλα τα στελέχη *P. mirabilis* που ενσωματώθηκαν στις αναλύσεις (260 γενώματα). Η ταξινόμηση σε cgMLST τύπους και η φυλογένεση αφορούσε δύο ομάδες γενωμάτων ήτοι αυτά στα οποία ανιχνεύθηκαν όλοι οι τόποι (210, σχήμα cgMLST) και στο σύνολο των αναλυθέντων στελεχών όπου χρησιμοποιήθηκαν μόνο τα 958 γονίδια για τις αναλύσεις (σχήμα cgMLST missing).

Τα 210 στελέχη όπου ανιχνεύθηκαν και οι 1397 τόποι ταξινομήθηκαν σε 203 διαφορετικούς τύπους cgMLST. Ο 203×203 πίνακας με τους γενετικούς τόπους που διαφέρουν οι cgMLST τύποι αποκάλυψε επιπλέον στελέχη που ανήκουν στα επικρατέστερα συμπλέγματα κλώνων (**Εικόνα 22**, δεξιά). Η φυλογένεση ML έδειξε ότι τα στελέχη ταξινομούνται σε 4 κύριες φυλογενετικές γραμμές με τα αποτελέσματα να είναι ανάλογα με αυτά των αναλύσεων των 154 στελεχών (**Εικόνα 22,** αριστερά). Σε αυτή την περίπτωση η φυλογενετική γραμμή Ι, όπου εντοπίζονται οι περισσότεροι τύποι, φαίνεται ότι διαφοροποιείται σε δύο υποομάδες (Ιa και Ιβ, **Εικόνα 22**, αριστερά). Η ενσωμάτωση των στελεχών όπου δεν ανιχνεύθηκαν κάποια γονίδια δεν τροποποίησε τα συμπεράσματα για τη δομή του πληθυσμού *P. mirabilis*. Τα 260 στελέχη ταξινομήθηκαν σε 252 διαφορετικούς cgMLST τύπους με τα κύρια συμπλέγματα κλώνων να εμπλουτίζονται με περισσότερα στελέχη ενώ οι φυλογενετικές σχέσεις των στελεχών χρησιμοποιώντας τους 958 γενετικούς τόπους (**Εικόνα 23**) φαίνεται να μην διαφοροποιούνται από όταν χρησιμοποιούνται οι 1397 τόποι.

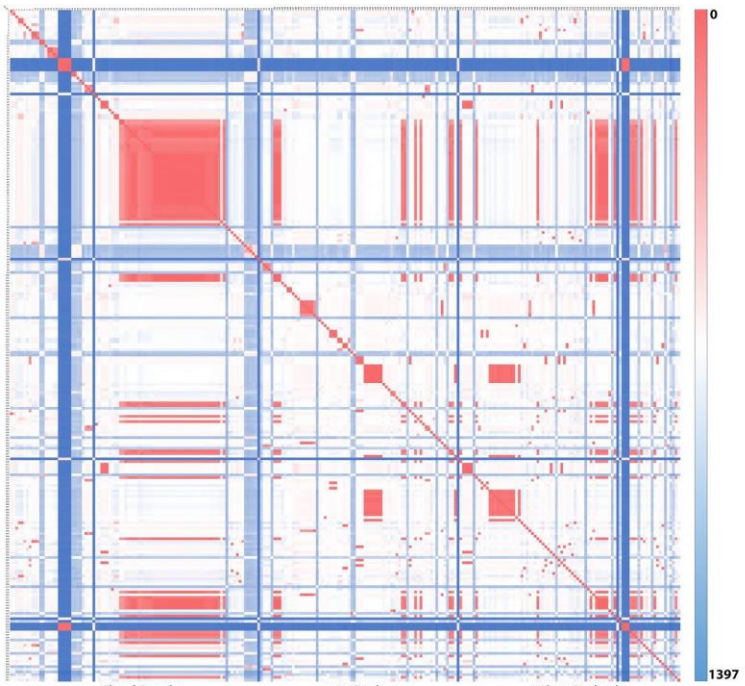
Για περαιτέρω ανάλυση των κλωνικών συμπλεγμάτων επιλέχθηκαν τα αποτελέσματα του σχήματος cgMLST missing καθώς αφορά το σύνολο των κατατεθειμένων στελεχών της GenBank. Βάσει του αριθμού των διαφορετικών γενετικών τόπων στο σχήμα cgMLST missing ορίστηκαν τα κύρια συμπλέγματα κλώνων που δίνονται στην **Εικόνα 23**.

Πίνακας 9: Στελέχη όπου παρατηρήθηκε έλλειψη κάποιων από τους γενετικούς τόπους του cgMLST σχήματος και αριθμητικά δεδομένα των απόντων γονιδίων.

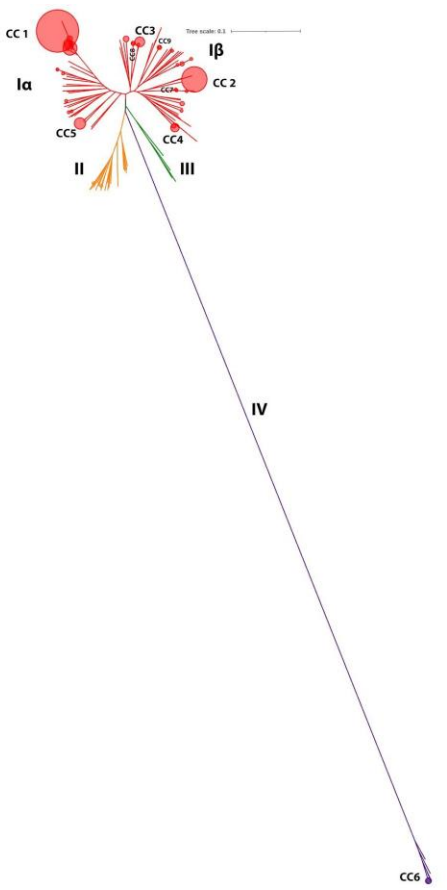
Στέλεχος P. mirabilis	**Σύνολο απόντων γονιδίων**	**Απόντα γονίδια όπου ανιχνεύθηκε θραύσμα**	**Θραύσματα σε άκρα contig (d < 100 bp)**
071H2	1	1	0
071H3	3	3	1
1023322	16	11	8
1091	8	4	1
1114_PMIR	1	1	0
130608239-L92	23	22	7
130688346-L100	41	30	17
130B9	6	5	4
1326_PMIR	12	12	5
1330_PMIR	1	1	0
160A10	15	8	6
23809	1	1	0
25_PMIR	5	5	2
292_PMIR	1	0	0
360_PMIR	1	1	1
418_PMIR	1	0	0
646_PMIR	4	4	3
672_PMIR	5	5	1
998368	1	1	1
AS012308	5	5	3
AS012310	5	4	2
AS012318	8	7	5
AS012328	9	7	5
AS012355	2	1	0
AS012360	2	2	0
AS012362	8	8	4
AS012363	3	2	2
AS012407	1	1	0
BCT11	29	23	12
BCT17	61	52	31
BOC1	1	1	1
CKTH01	12	12	6
CNR20130297	30	20	13
CNR20160617	21	19	7
CNR20160679	35	32	14
CNR20160877	43	36	17
Dog-06-37660	69	47	24
Dog-35-37761	46	22	11
ESDY17	1	1	1
EUG91	7	3	3
GEN000048	25	25	7
MH13-009N	3	3	2
MPE0734	4	0	0
MPE0767	4	0	0
NCTC10975	44	10	5
PM_125	5	5	2
PM593	41	6	1
PmirS	1	1	0
S4	30	26	13
TUM11568	1	1	0



Εικόνα 22. Αριστερά: Πίνακας 203×203 με τον αριθμό γενετικών τόπων (cgPM) με διαφορετικά αλλήλια για το κάθε ζεύγος cgMLST τύπων. Οι κόκκινες περιοχές αντιστοιχούν στα συμπλέγματα κλώνων (ίδια αλλήλια σε ≥ 967 τόπους). Δεξιά: Φυλογένεση ML με το πρόγραμμα RaxML των 203 cgMLST τύπων χρησιμοποιώντας τα 1397 γονίδια. Έχουν επισημανθεί οι κύριες φυλογενετικές γραμμές. Οι συγγενικοί κλάδοι (κύκλοι μεγέθους αναλογικού με τον αριθμό των κλάδων που τα απαρτίζουν) σύμφωνα με το κριτήριο BRL < 0.01. Η εικόνα ετοιμάστηκε στο iTol.



Clonal Complex	cgMLST τύποι	#Τύπων/Στελεχών
1	42-79, 81,100-102, 148, 149, 153, 155, 166,167, 170, 176, 199, 219, 221-225, 234-236, 238, 243, 251	63/69
2	134-140, 168, 181-189, 191	18/18
3	35-37, 171-174, 204, 208	9/9
4	110-115, 163, 195, 220	9/9
5	8, 131-133, 151, 202, 207, 209, 244	9/9
6	19-23, 231-233	8/8
7	29-31, 157, 158, 177, 197	7/7
8	121-123, 178, 180	5/5
9	96-98, 164, 248	5/5



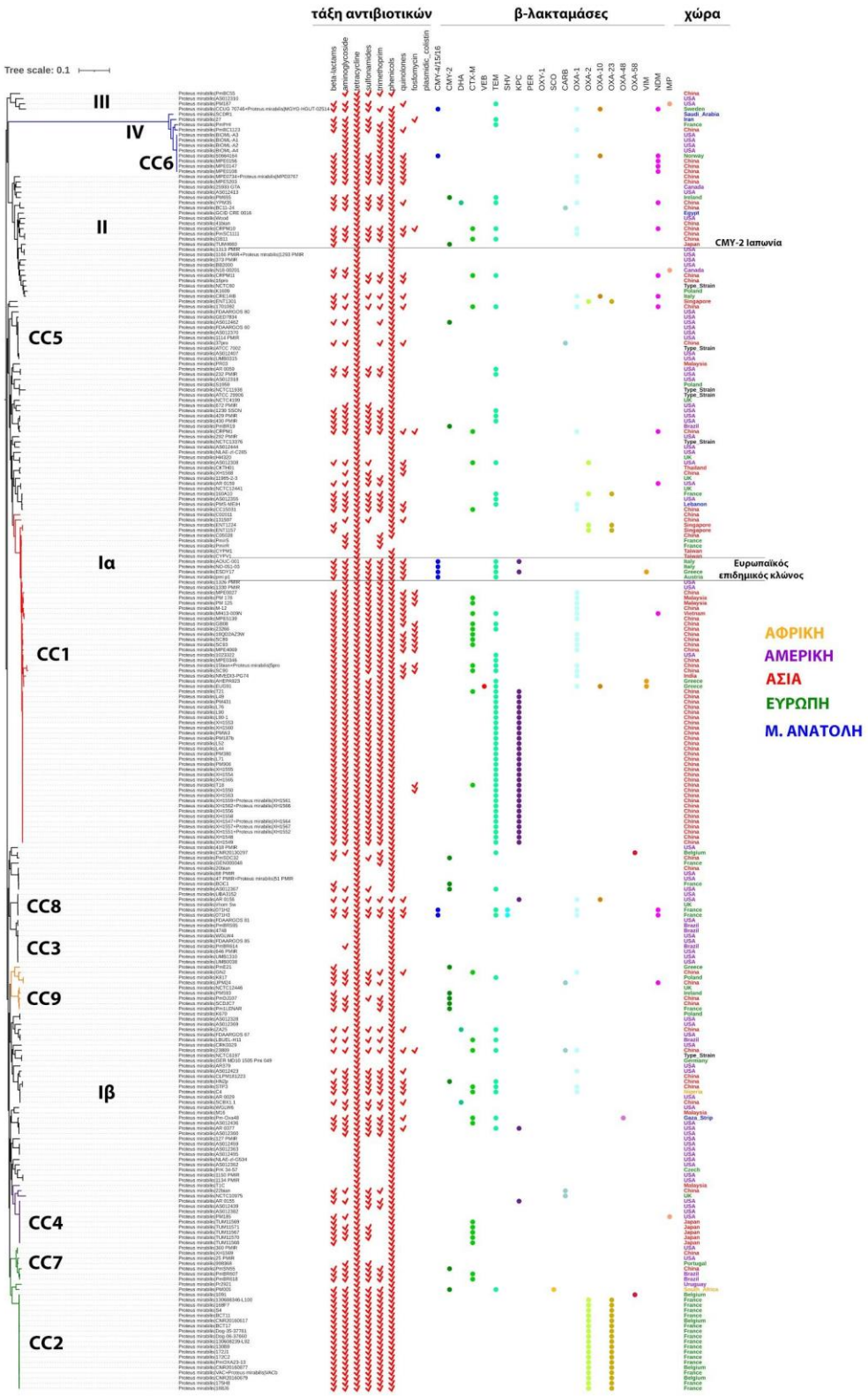
Εικόνα 23. Ομοίως με Εικόνα 22, χρησιμοποιώντας τα δεδομένα του σχήματος cgMLST missing. Ο 252×252 πίνακας κατασκευάστηκε με τα αλληλλικά προφίλ των cgMLST τύπων, όπου για τους γενετικούς τόπους που δεν ανιχνεύθηκαν αλλήλια σε ορίστηκε αλλήλιο με αριθμό «0». Η φυλογένεση εκτιμήθηκε χρησιμοποιώντας τα 958 κοινά γονίδια όλων των γενωμάτων. Ο πίνακας δίνει τα κύρια συμπλέγματα κλώνων σύμφωνα με το κριτήριο ταύτισης αλληλίων (<430 αλλήλια) τα οποία δηλώνονται και στο δέντρο.

Τα κύρια συμπλέγματα κλώνων (≥ 5 cgMLST τύποι) εντοπίζονται κυρίως στην φυλογενετική γραμμή I (8 από τα 9). Τα 69 από τα 260 στελέχη (26,5%) ανήκουν στο σύμπλεγμα κλώνων 1 (CC1) ενώ το αμέσως επόμενο πιο πολυπληθές περιέχει 18 στελέχη-cgMLST τύπους (CC2). Οι δύο αυτοί κύριοι κλώνοι ανήκουν στους συγγενικούς κλάδους Ια και Ιβ αντίστοιχα (**Εικόνα 23**, δεξιά). Το σύμπλεγμα κλώνων 6 (CC6) εντοπίζεται στην απομακρυσμένη φυλογενετική γραμμή IV. Παρατηρήθηκε ότι το κριτήριο ορισμού των συμπλεγμάτων κλώνων, δηλαδή ταύτιση αλληλίων σε 967 τόπους, συμβαδίζει με τη ομαδοποίηση βάσει φυλογενετικής απόστασης < 0.01 καθώς αν και σε μερικούς κλώνους η ομαδοποίηση αυτή δίνει πάνω από έναν κλάδους αυτοί είναι πολύ συγγενικοί (π.χ. CC1 **Εικόνα 23**, δεξιά).

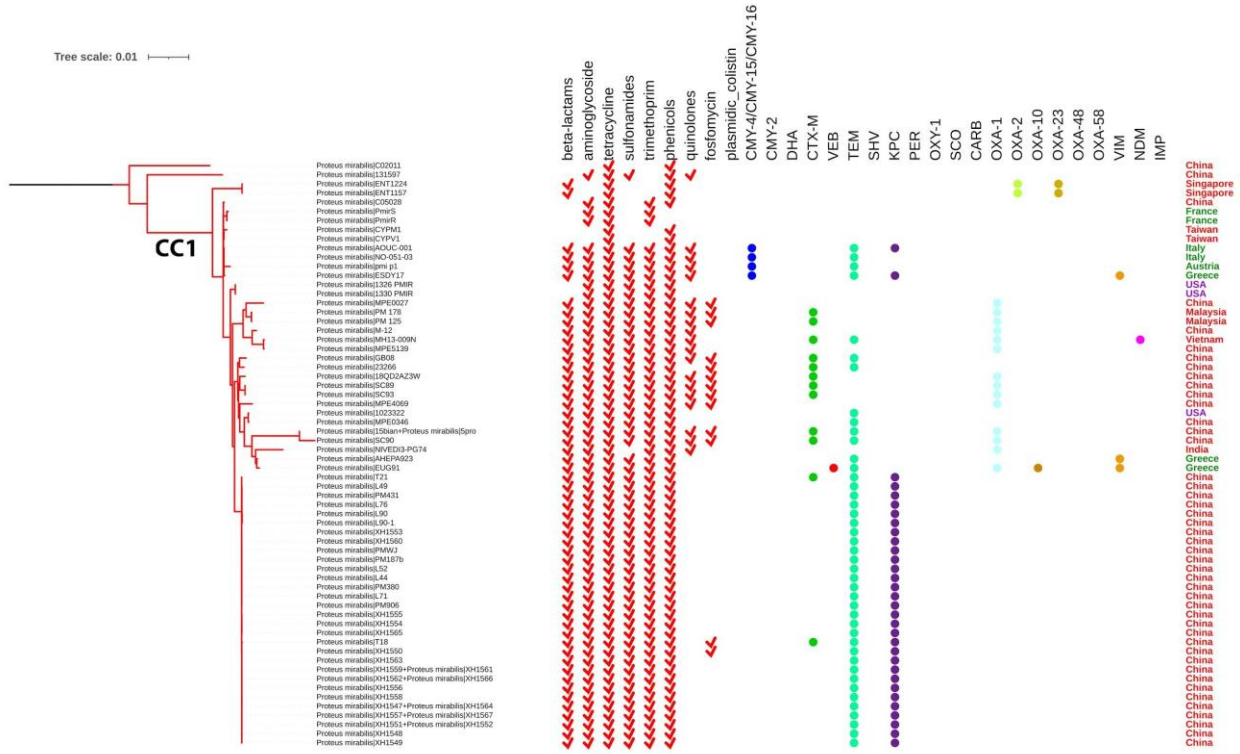
Η ανάλυση του γονοτύπου αντοχής σε αντιβιοτικά έδειξε ότι όλα τα στελέχη φέρουν γονίδιο αντοχής σε τετρακυκλίνες και η πλειονότητα τους φέρει και γονίδιο αντοχής στις φαινικόλες (**Εικόνα 24**). Ένας αυξημένος αριθμός στελεχών ήταν πολυανθεκτικά (Multi Drug Resistant, MDR) καθώς έφερε επίκτητα γονίδια αντοχής για περισσότερες από τρεις τάξεις αντιβιοτικών (**Εικόνα 24**). Το παραπάνω αντικατοπτρίζει την ετεροβαρή αλληλούχιση και ενσωμάτωση στη βάση δεδομένων GenBank γενωμάτων νοσοκομειακών στελεχών. Επίσης, τα περισσότερα στελέχη *P. mirabilis* που έχουν αλληλουχηθεί έχουν απομονωθεί στις ΗΠΑ και στην Κίνα, ενώ ακολουθούν στελέχη από την Ευρώπη (**Εικόνα 24**). Συνεπώς ο πληθυσμός *P. mirabilis* που αναλύθηκε αφορά συγκεκριμένα ενδιαιτήματα και γεωγραφικές περιοχές και ενδεχομένως να μην μπορούν να εξαχθούν ασφαλή συμπεράσματα σχετικά με τη δομή του. Μολαταύτα η

φυλογένεση βάσει των γονιδίων του core genome των φαίνεται να αποδίδει τις παρατηρήσεις σχετικά με κάποιες νοσοκομειακές επιδημίες ανθεκτικών στελεχών.

Αναφορικά με την διασπορά γονιδίων β-λακταμασών φαίνεται ότι αυτή αφορά και τις τέσσερις κύριες φυλογενετικές γραμμές (I, II, III και IV). Εκτός από τα ένζυμα περιορισμένου φάσματος τύπου TEM και OXA-1 έχουν διασπαρεί και ένζυμα που υδρολύουν αντιβιοτικά τελευταίας γραμμής (νεότερες κεφαλοσπορίνες και καρβαπενέμες) όπως είναι οι CMY κεφαλοσπορινάσες, οι εκτετεταμένου φάσματος β-λακταμάσες (Extended Spectrum beta-Lactamases, ESBLs) CTX-M, οι καρβαπενεμάσες KPC-2 (τάξη A) και OXA-23 (τάξη D) και οι μέταλλο-β-λακταμάσες VIM, NDM και IMP. Αν και η διασπορά αφορά όλες τις φυλογενετικές γραμμές φαίνεται ότι κάποια συμπλέγματα κλώνων συσσωρεύουν β-λακταμάσες όλων των τύπων. Το κλωνικό σύμπλεγμα CC1 για παράδειγμα (**Εικόνες 24** και **25**) χαρακτηρίζεται από την διασπορά σχεδόν όλων των τύπων β-λακταμασών που αναφέρθηκαν παραπάνω. Στον συγκεκριμένο κλώνο, ο οποίος υπέρ-αντιπροσωπεύεται στα γενώματα που αναλύθηκαν, εντοπίζονται κυρίως στελέχη από Ανατολική Ασία και Ευρώπη. Στα ευρωπαϊκά στελέχη συμπεριλαμβάνονται τέσσερα που φέρουν γονίδια κεφαλοσπορινάσης τύπου CMY-4 (CMY-4/16/15) τα οποία αντιστοιχούν στον κύριο κλώνο που απομονώνεται σε Ιταλία και Ελλάδα τα τελευταία 10 χρόνια (D'Andrea et al., 2011). Έτσι λοιπόν αν και τα περισσότερα δημοσιευμένα γενώματα του κλώνικού συμπλέγματος προέρχονται από την Κίνα, είναι επίσης συχνός σε νοσοκομειακά ενδιαιτήματα στην Ευρώπη. Πέρα από την κεφαλοσπορινάση τα ευρωπαϊκά στελέχη CC1 έχουν αποκτήσει γονίδια των καρβαπενεμασών KPC και VIM.

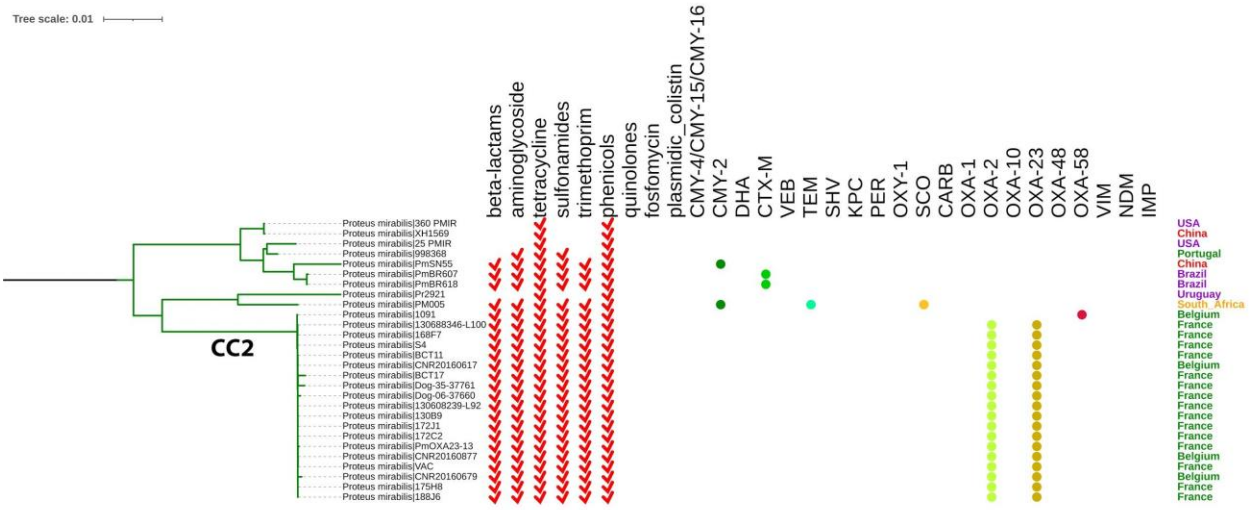


Εικόνα 24. ML φυλογενετικό δέντρο των 252 τύπων cgMLST χρησιμοποιώντας τα 958 κοινά γονίδια. Τα ονόματα των cgMLST τύπων έχουν αντικατασταθεί με ονόματα στελεχών. Για κάθε κλαδί δηλώνονται οι γονότυποι αντοχής με εστίαση στις β-λακταμάσες καθώς και η χώρα απομόνωσης. Έχουν σημανθεί οι κύριες φυλογενετικές γραμμές και τα συμπλέγματα κλώνων. Η εικόνα ετοιμάστηκε στο iTol.



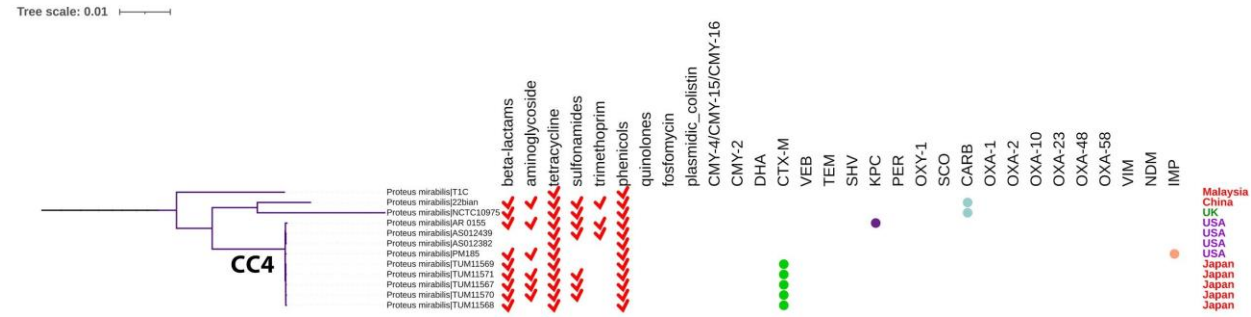
Εικόνα 25. Εστίαση στη φυλογενετική γραμμή του κλώνικού συμπλέγματος CC1. Η εικόνα προέκυψε με την εντολή prune tree για τον αντίστοιχο κλάδο τους δέντρου της Εικόνας 24 στο iTol.

Εντός του CC1 παρατηρήθηκε μία διακριτή φυλογενετική γραμμή που περιλαμβάνει δύο στελέχη από την Ελλάδα που παράγουν τη VIM ΜβΛ, τα *P. mirabilis* AHEPA923 και *P. mirabilis* EUG91, τα οποία απομονώθηκαν σε Θεσσαλονίκη και Αθήνα αντίστοιχα καθώς και ένα στέλεχος από την Ινδία, το *P. mirabilis* NIVED3-PG74, που παρήγαγε μόνο την TEM πενικιλλινάση και είχε απομονωθεί από κοτόπουλο **(Εικόνα 25)**. Το μοτίβο αυτό παρατηρείται και για τα υπόλοιπα ευρωπαϊκά στελέχη του κλώνου τα οποία εμφανίζουν στενές σχέσεις με ασιατικά τα οποία δεν εκφράζουν επίκτητη αντοχή, π.χ. *P. mirabilis* CYPM1 και *P. mirabilis* CYPV1 (Ταϊβάν) με τα 4 στελέχη που διαθέτουν την CMY-4 **(Εικόνα 25)**. Έτσι λοιπόν φαίνεται ότι το κλωνικό σύμπλεγμα πιθανώς να προέκυψε στην Ασία και με μεταφορά γονιδίων αντοχής σε αντιβιοτικά τελευταίας γραμμής να προσαρμόστηκε στο νοσοκομειακό περιβάλλον και να εξαπλώθηκε στην Ευρώπη.



Εικόνα 26. Εστίαση στη φυλογενετική γραμμή του κλώνικού συμπλέγματος CC2

Επιβεβαιώθηκε επίσης η στενή φυλογενετική σχέση στελεχών *P. mirabilis* που παρήγαγαν τις OXA-23 και OXA-58 καρβαπενεμάσες και απομονώθηκαν στην Γαλλία και το Βέλγιο σε ζώα και σε νοσοκομειακά περιβάλλοντα (Bonnin et al., 2020) προκαλώντας μία εντοπισμένη επιδημία (κλωνικό σύμπλεγμα CC2, **Εικόνα 26**) στην Ευρώπη.

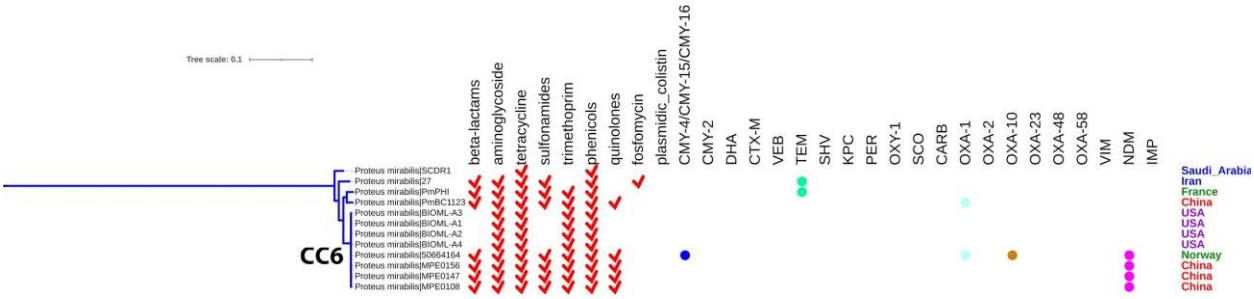


Εικόνα 27. Εστίαση στη φυλογενετική γραμμή του κλώνικού συμπλέγματος CC4.

Ομοίως στελέχη *P. mirabilis* που παράγουν την CTX-M ESBL και εμπλέκονται σε νοσοκομειακές επιδημίες στην Ιαπωνία (Nagano et al., 2003) ομαδοποιούνται στον κλωνικό σύμπλεγμα CC4 **(Εικόνα 27)**. Στον ίδιο κλωνικό σύμπλεγμα εντοπίζεται και ένα στέλεχος που δεν φέρει επίκτητα γονίδια αντοχής και είχε απομονωθεί στις ΗΠΑ (*P. mirabilis* AS012382).

Εικόνα 28. Εστίαση στη φυλογενετική γραμμή του κλώνικού συμπλέγματος CC9.

Στο κλωνικό σύμπλεγμα CC9 ομαδοποιούνται στελέχη από την Ευρώπη και την Κίνα και φέρουν τη CMY-2 κεφαλοσπορινάση, το γονίδιο της οποίας σε αυτά τα στελέχη μεταφέρεται από συζευκτικό μεταθετό γενετικό στοιχείο (ICE) τύπου SXT/R391 (π..χ. το SXT στοιχείο στο PmDJ107 εντοπίζεται στο contig42 GenBank ID: RQSH01000042.1). Στην ίδια γραμμή εντοπίζεται και ένα στέλεχος με αυτό το χαρακτηριστικό που απομονώθηκε στην Ελλάδα από λοίμωξη σε σκύλο (*P. mirabilis* PmE21, **Εικόνα 28**). Το *P. mirabilis* TUM4660 που απομονώθηκε στην Ιαπωνία (Harada et al., 2003) και αποτελεί το πρώτο στέλεχος όπου εντοπίστηκε ICE του παραπάνω τύπου με *bla*$_{CMY-2}$ (ICEPmiJpn1, GenBank ID: BGM01000062, 99% ID και 100% cov με το αντίστοιχο του PmDJ107) φαίνεται να είναι απομακρυσμένο από τα παραπάνω βακτήρια και να ανήκει στην φυλογενετική γραμμή II.



Εικόνα 29. Εστίαση στη φυλογενετική γραμμή του κλώνικού συμπλέγματος CC6.

Τέλος το κλωνικό σύμπλεγμα CC6 που ανήκει στην απομακρυσμένη φυλογενετική γραμμή IV φαίνεται ότι αρχίζει να συσσωρεύει την NDM καρβαπενεμάση και να έχει ευρεία γεωγραφική εξάπλωση (**Εικόνα 29**). Αν και αρκετά απομακρυσμένος από τα υπόλοιπα *P. mirabilis*, στη φυλογένεση του γένους *Proteus* ομαδοποιούνται μαζί με τα υπόλοιπα στελέχη του είδους και πιθανόν να αποτελεί κάποιο υποείδος.

3.8. Χρήση των γονιδίων του core genome για ταξινομικές και φυλογενετικές αναλύσεις στο γένος *Proteus* και την οικογένεια *Morganellaceae*.

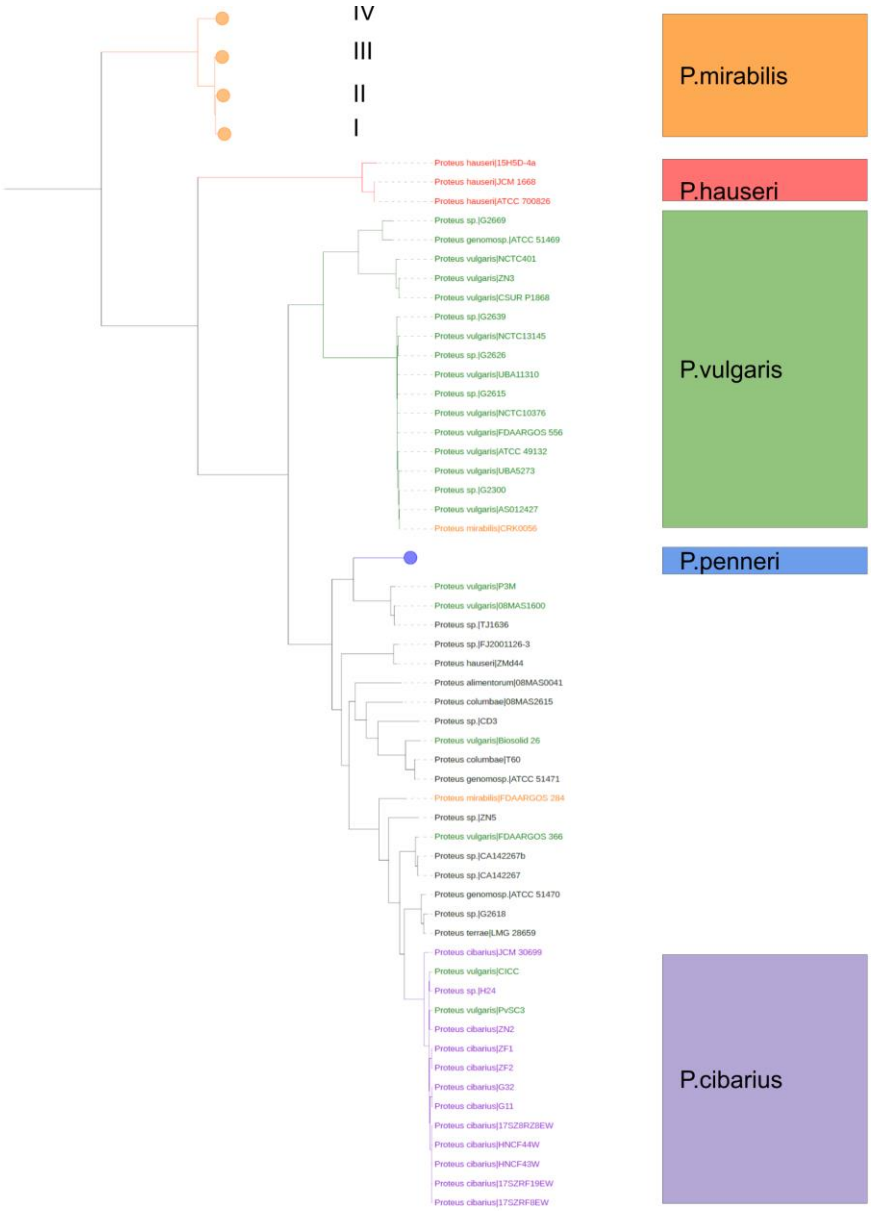
Στην συνέχεια επεκτείναμε τις φυλογενετικές αναλύσεις σε γενώματα του γένους *Proteus* της βάσης δεδομένων GenBank και της οικογένειας *Morganellaceae* της βάσης δεδομένων RefSeq του NCBI. Η ανίχνευση των 1397 τόπων έγινε με το εργαλείο cg-finder, εφαρμόζοντας λιγότερο αυστηρά κριτήρια απ' ότι στην περίπτωση των *P. mirabilis*. Για το γένος *Proteus* το κριτήριο απόντων τόπων ώστε να αναλυθούν περαιτέρω τα γενώματα (missing_loci_co) τέθηκε στο 8% ενώ τα ποσοστά ταύτισης νουκλεοτιδικής

αλληλουχίας και κάλυψης ορίστηκαν αμφότερα στο 60%. Οι αντίστοιχες παράμετροι στις αναζητήσεις στα *Morganellaceae* ήταν: missing_loci_co=34%, id=60% και cov=60%.
Στο γένος *Proteus* εντοπίστηκαν 748 από τους 1397 τόπους σε όλα τα γενώματα που αναλύθηκαν. Τα είδη *P. alimentorum*, *P. penneri* και *P. columbae* εμφάνισαν τους περισσότερους απόντες τόπους (**Πίνακας 10**). Χρησιμοποιώντας τα ψευδογενώματα core genome που αποτελούνταν από τις αντιστοιχίσεις των 748 κοινών γονιδίων προσδιορίστηκαν οι φυλογενετικές σχέσεις των στελεχών.

Πίνακας 10. Γενώματα του γένους *Proteus* και μέσος όρος cgPM που λείπουν από τα στελέχη του κάθε είδους.

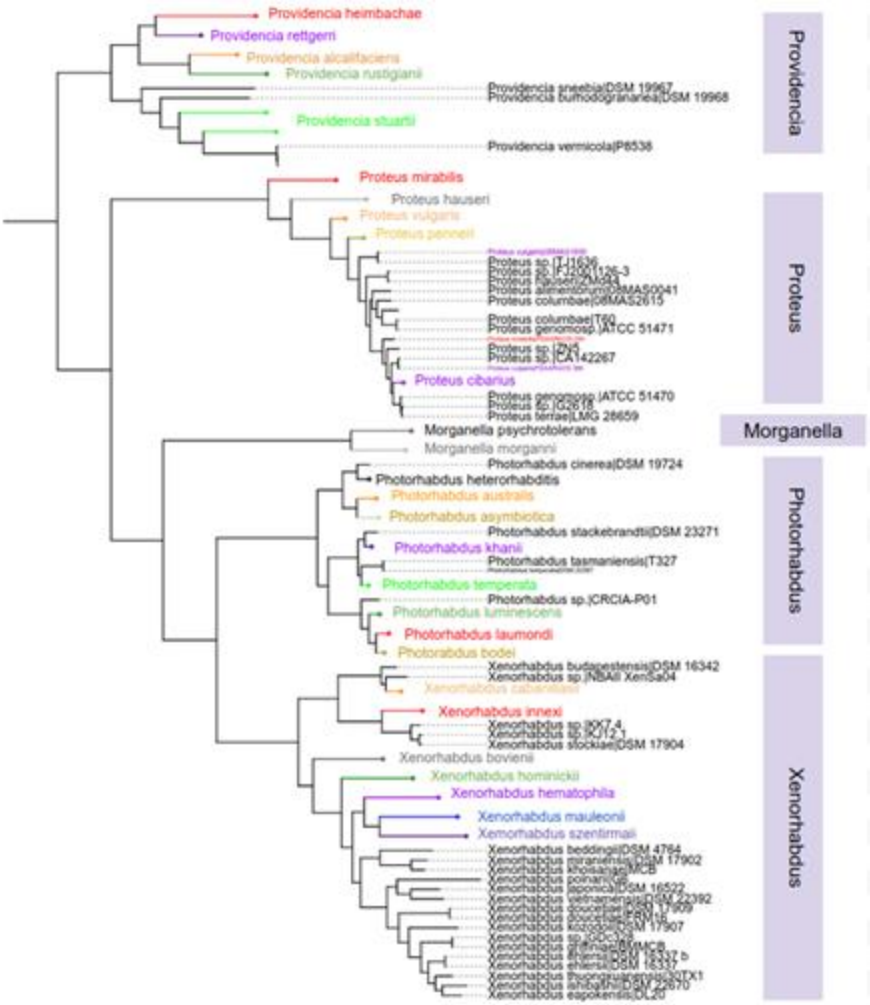
Βακτηριακό είδος	**Αριθμός γενωμάτων**	**Μέσος όρος cgPM που λείπουν**
Proteus alimentorum	1	58
Proteus cibarius	11	35
Proteus columbae	2	51
Proteus genomosp.	3	43
Proteus hauseri	4	48
Proteus penneri	5	53
Proteus sp.	59	31
Proteus terrae	1	29
Proteus vulgaris	18	47

Τα διαφορετικά είδη ταξινομήθηκαν σε διακριτούς κλάδους οπότε τα γονίδια του σχήματος μπορούν να χρησιμοποιηθούν για φυλογενετικές αναλύσεις του γένους (**Εικόνα 30**). Από τα 59 στελέχη για τα οποία δεν υπήρχε πληροφορία ως προς το είδος (*Proteus* sp.) τα 28 φαίνεται να ταξινομούνται στον κλάδο του *P. mirabilis*, τα 16 στον κλάδο *P. penneri,* τα 5 στον κλάδο *P. vulgaris* ενώ 1 ομαδοποιήθηκε με τα *P. cibarius*. Εκτός από τα 2 στελέχη *P. mirabilis* που χαρακτηρίζονταν από λανθασμένη ταξινόμηση εντοπίστηκαν και 6 *P. vulgaris* τα οποία απείχαν από τον κλάδο του είδους τους, και πιθανότατα έχουν ταξινομηθεί λανθασμένα. Τα στελέχη αυτά είναι : *P. vulgaris* P3M, *P. vulgaris* 08MAS1600, *P. vulgaris* Biosolid26*, P. vulgaris* FDAARGOS 366, *P. vulgaris* PvSC3, και *P. vulgaris* CICC. (**Εικόνα 30**).



Εικόνα 30. Approximately-maximum-likelihood φυλογένεση χρησιμοποιώντας τα 748 κοινά γονίδια του σχήματος cgMLST στο γένος *Proteus*. Οι φυλογενετικές σχέσεις εκτιμήθηκαν με το πρόγραμμα FastTree (μοντέλο GTR). Στο *P. mirabilis* σημειώνονται οι 4 κύριες φυλογενετικές γραμμές. Με πορτοκαλί σημειώνονται τα στελέχη *P. mirabilis* και με πράσινο τα στελέχη *P. vulgaris* που έχουν ταξινομηθεί πιθανότατα λάθως.

Στις αναζητήσεις στην οικογένεια *Morganellaceae* ανιχνεύθηκαν 357 κοινοί γενετικοί τόπου του σχήματος cgMLST *P. mirabilis* σε όλα τα γενώματα που αναλύθηκαν σύμφωνα με τα επιλεγμένα κριτήρια. Οι περισσότεροι απόντες τόποι εντοπίστηκαν στα γένη των *Xenorhabdus* και *Photorhabdus* τα οποία είναι συμβιωτικά βακτήρια νηματωδών σκωλήκων. Ακολουθούσαν τα γένη *Morganella* και *Providencia* **(Πίνακες 11** και **12)**. Η φυλογένεση χρησιμοποιώντας τα 357 γονίδια ομαδοποίησε τα γενώματα κατά γένη ενώ έδειξε ότι τα *Xenorhabdus* και *Photorhabdus* απέχουν από τα υπόλοιπα γένη και είναι πιο συγγενικά με τη *Morganella* **(Εικόνα 31)**.



Εικόνα 31. Approximately-maximum-likelihood φυλογένεση χρησιμοποιώντας τα 357 κοινά γονίδια του σχήματος cgMLST στην οικογένεια *Morganellaceae*. Οι φυλογενετικές σχέσεις εκτιμήθηκαν με το πρόγραμμα FastTree (μοντέλο GTR).

Πίνακας 11. Βακτηριακά είδη και ο μέσος όρος των γενετικών τόπων που λείπουν από κάθε βακτηριακό είδος.

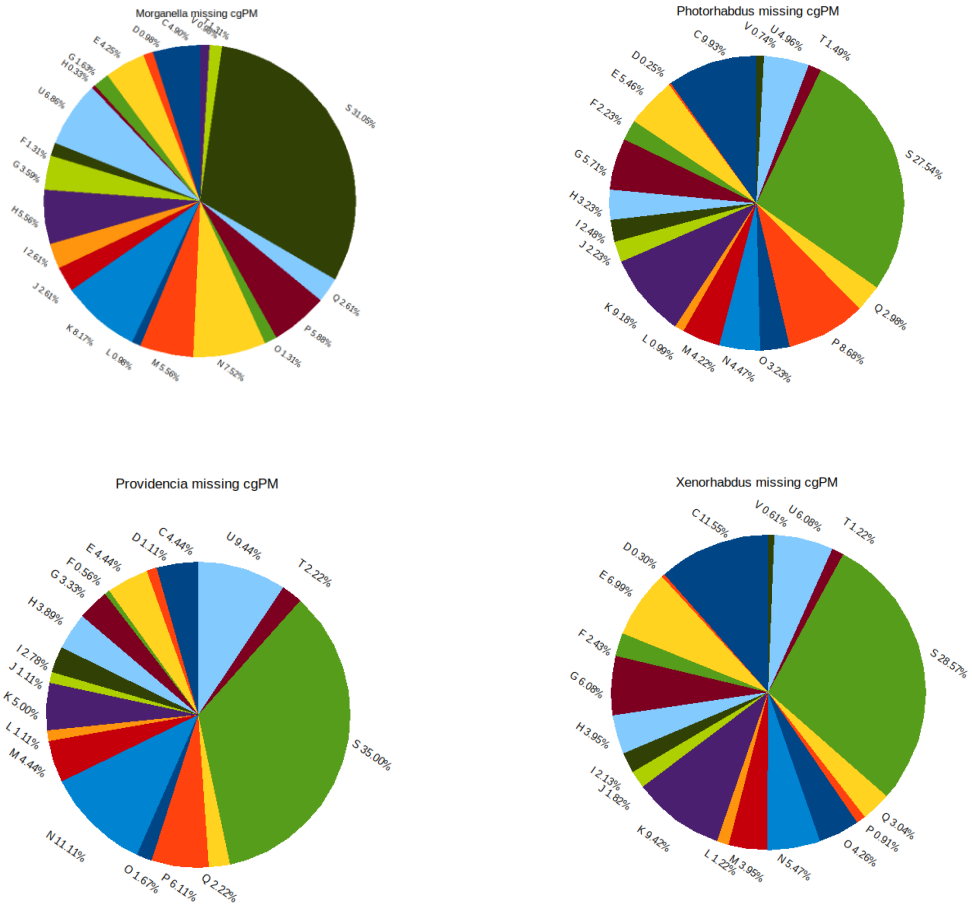
Βακτηριακό Είδος	Αριθμός Γενωμάτων	Απόντες γενετικοί τόποι	Βακτηριακό Είδος	Αριθμός Γενωμάτων	Απόντες γενετικοί τόποι
Morganella morganii	72	305.26	Providencia heimbachae	5	225.80
Morganella psychrotolerans	4	313.00	Providencia rettgeri	46	227.11
Morganella sp.	2	328.50	Providencia rustigianii	9	258.00
Photorhabdus asymbiotica	2	406.00	Providencia sneebia	1	367.00
Photorhabdus australis	2	405.50	Providencia stuartii	18	269.08
Photorhabdus bodei	4	389.25	Providencia vermicola	1	254.00

Photorhabdus cinerea	1	376.00	Xenorhabdus beddingii	1	391.00
Photorhabdus heterorhabditis	3	388.67	Xenorhabdus bovienii	11	414.70
Photorhabdus khanii	3	399.67	Xenorhabdus budapestensis	3	411.00
Photorhabdus laumondii	15	394.87	Xenorhabdus cabanillasii	2	417.67
Photorhabdus luminescens	14	394.08	Xenorhabdus doucetiae	2	400.00
Photorhabdus namnaonensis	1	399.00	Xenorhabdus eapokensis	1	418.00
Photorhabdus sp.	11	391.09	Xenorhabdus ehlersii	2	397.50
Photorhabdus stackebrandtii	1	385.00	Xenorhabdus griffiniae	1	421.00
Photorhabdus tasmaniensis	1	387.00	Xenorhabdus hominickii	2	394.00
Photorhabdus temperata	5	414.80	Xenorhabdus innexi	2	392.00
Photorhabdus thracensis	1	416.00	Xenorhabdus ishibashii	1	403.00
Proteus alimentorum	1	57.00	Xenorhabdus japonica	1	401.00
Proteus cibarius	11	33.73	Xenorhabdus khoisanae	1	389.00
Proteus columbae	2	50.00	Xenorhabdus kozodoii	1	386.00
Proteus genomosp.	3	42.00	Xenorhabdus mauleonii	2	356.50
Proteus hauseri	4	48.25	Xenorhabdus miraniensis	1	378.00
Proteus mirabilis	236	21.18	Xenorhabdus nematophila	7	431.71
Proteus penneri	2	47.00	Xenorhabdus poinarii	1	440.00
Proteus sp.	56	30.35	Xenorhabdus sp.	4	403.50
Proteus terrae	1	28.00	Xenorhabdus stockiae	1	377.00
Proteus vulgaris	14	54.50	Xenorhabdus szentirmaii	3	386.00
Providencia alcalifaciens	29	263.00	Xenorhabdus thuongxuanensis	1	405.00
Providencia burhodogranariea	1	292.00	Xenorhabdus vietnamensis	1	389.00

Πίνακας 12. Βακτηριακά γένη και ο μέσος όρος των cgPM που λείπει από κάθε βακτηριακό γένος.

Βακτηριακό Γένος	Αριθμός γενωμάτων	Μέσος όρος απόντων cgPM
Morganella	78	306.3
Photorhabdus	64	389.5
Proteus	330	12.1
Providencia	147	247.8
Xenorhabdus	52	390.1

Στη συνέχεια, απομονώσαμε τους γενετικούς τόπους που λείπουν από το 95% των γενωμάτων κάθε γένους και αναζητήσαμε την κατηγορία COG τους.



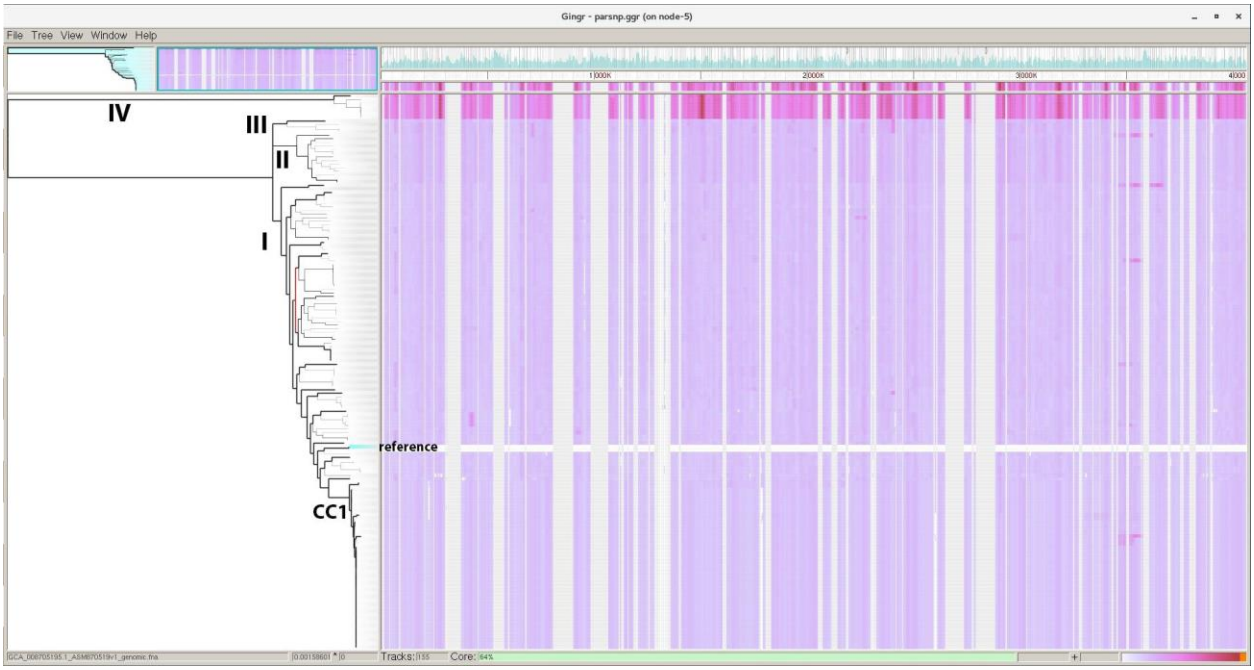
Εικόνα 32. Οι κατηγορίες COG για τα γονίδια που λείπουν στο 95% των γενωμάτων των γενών Morganella, Providencia, Photorhabdus και Xenorhabdus.

Στην **Εικόνα 32** βλέπουμε τις κατηγορίες COG των γενετικών τόπων, στους οποίους παρατηρείται έλλειψη στο 95% των στελεχών του κάθε γένους. Παρατηρούμε ότι η πλειοψηφία των τόπων και στα τέσσερα γένη αφορά σε πρωτεΐνες με άγνωστες λειτουργίες. Όσον αφορά στη *Providencia*, οι κατηγορίες COG που λείπουν σε μεγαλύτερο ποσοστό είναι οι κατηγορίες, N: κυτταρική κινητικότητα (11,11%), U: ενδοκυτταρική μεταφορά και απέκκριση (9,44%), P: Μεταβολισμός και μεταφορά ανόργανων ιόντων (6,11%) και K: μεταβολισμός και μεταφορά ανόργανων ιόντων (5%). Όσον αφορά στο *Xenorhabdus*, σε μεγαλύτερο ποσοστό λείπουν οι κατηγορίες C: Παραγωγή ενέργειας (11,59%), K: μεταβολισμός και μεταφορά ανόργανων ιόντων (9,42%), E: Μεταφορά και μεταβολισμός Αμινοξέων (6,99%) και G: Μεταφορά και μεταβολισμό υδατανθράκων (6.08%). Όσον αφορα στον *Photorhabdus*. οι κατηγορίες που λείπουν συχνότερα είναι οι C: Παραγωγή ενέργειας (9,93%), K: μεταβολισμός και μεταφορά ανόργανων ιόντων (9,18%), P: Μεταβολισμός και μεταφορά ανόργανων ιόντων (8,68%) και G: Μεταφορά και μεταβολισμό υδατανθράκων (5,71%). Τέλος, όσον αφορά στη *Morganella* οι κατηγορίες COG που λείπουν σε μεγαλύτερο ποσοστό είναι οι K: μεταβολισμός και μεταφορά ανόργανων ιόντων (8,17%), N: κυτταρική κινητικότητα (7,52%), U: ενδοκυτταρική μεταφορά και απέκκριση (6,86%), P: Μεταβολισμός και μεταφορά ανόργανων ιόντων (5,88%). Παρατηρούμε ότι οι κατηγορίες που λείπουν σε μεγαλύτερο ποσοστό στο 95% των στελεχών των γενών *Xenorhabdus* και *Photorhabdus* είναι παρόμοιες, καθώς τα γένη αυτά είναι εξελικτικά πιο κοντινά **(Εικόνα 31)**, και αφορούν κυρίως λειτουργίες του μεταβολισμού. Ενώ στη Providencia και στη Morganella οι γενετικοί τόποι που εκλείπουν αφορούν τόσο σε μονοπάτια του μεταβολισμού όσο και σε λοιπές κυτταρικές διαδικασίες, όπως η κινητικότητα του βακτηρίου και η ενδοκυττάρια απέκκριση. Οι ελλείψεις σε cgPM αντικατοπτρίζουν τις διαφορετικές φυσιολογίες των διαφορετικών γενών αλλά έμμεσα και την εξελικτική τους απόσταση. Ειδικότερα, τα γένη *Xenorhabdus* και *Photorhabdus* που απέχουν περισσότερο από τον *P. mirabilis,* φαίνεται να μην έχουν βασικά γονίδια μεταβολισμού που συναντώνται στο core genome του *P. mirabilis,* ενώ στα γένη *Providencia* και *Morganella* που είναι εξελικτικά πιο κοντά στο *P. mirabilis* παρατηρούμε μικρότερο αριθμό γονιδίων του core genome που λείπουν, και αυτά τα γονίδια δεν αφορούν μόνο σε βασικές μεταβολικές λειτουργίες του κυττάρου αλλά και σε φαινοτυπικά χαρακτηριστικά όπως η κινητικότητα. Το γεγονός ότι από τα γένη αυτά λείπουν γονίδια κινητικότητας του *P. mirabilis* αντανακλά τη δυνατότητα του τελευταίου για ερπυσμό, και σε αυτά τα cgPM θα πρέπει να αναζητηθούν τα γονίδια που καθορίζουν τον συγκεκριμένο φαινότυπο.

3.9. Σύγκριση με τον αλγόριθμο Harvest

Το core genome των 154 γενωμάτων που χρησιμοποιήθηκαν στην παρούσα εργασία αναζητήθηκε επίσης χρησιμοποιώντας το λογισμικό πακέτο Harvest. Ως γένωμα αναφοράς χρησιμοποιήθηκε το χρωμόσωμα του στελέχους *P. mirabilis* HI4320 (GenBank acc: NC_010554). Το πρόγραμμα δημιουργεί το core genome για ένα σύνολο γενωμάτων εντοπίζοντας τις κοινές περιοχές με υψηλή ομολογία στη νουκλεοτιδική αλληλουχία ως προς ένα γένωμα αναφοράς το οποίο πρέπει να είναι ολοκληρωμένο. Συνεπώς ο αλγόριθμος του Harvest ενσωματώνει στο core genome και περιοχές που δεν κωδικοποιούν πρωτεΐνες. Τα αποτελέσματα **(Εικόνα 33)** έδειξαν ότι το “core genome” του *P. mirabilis* καλύπτει το 64% του χρωμοσώματος του *P. mirabilis* HI4320 (μέγεθος χρωμοσώματος 4.06 Mb). Στην περίπτωση του σχήματος core genome που αναπτύξαμε (συνολικό μέγεθος ψευδογενώματος των 1397 τόπων: 1.348.765 bp) το αντίστοιχο ποσοστό ήταν 33,2%. Η παραπάνω διαφορά οφείλεται στο ότι το παρόν core genome ορίστηκε χρησιμοποιώντας μόνο τις κωδικές περιοχές των βακτηριών και άρα μη κωδικές περιοχές, οπερόνια rRNA, γονίδια tRNA ή άλλων μη μεταφραζόμενων RNA έχουν εξαιρεθεί σε αντίθεση με το Harvest. Επίσης, στο παρόν core genome έχουν εξαιρεθεί περιοχές στις οποίες δεν ανιχνεύθηκαν γονίδια, λόγω κακής συναρμολόγησης και χαμηλής ποιότητας αλληλουχιών.
Παρ’ όλα αυτά οι φυλογενετικές σχέσεις που προέκυψαν από το Harvest φαίνεται να συμβαδίζουν αδρά με τις παρατηρήσεις μας αναφορικά με τις κύριες φυλογενετικές γραμμές του πληθυσμού και τα συμπλέγματα

κλώνων. Για παράδειγμα επιβεβαιώθηκε η εξελικτική απόσταση της γραμμής IV από τα υπόλοιπα στελέχη και η χαμηλή ομολογία αυτών των γενωμάτων επί του συνόλου σχεδόν του χρωμοσώματος. Τα υπόλοιπα στελέχη ομαδοποιούνται και σε αυτή την περίπτωση σε τρείς κύριες φυλογενετικές γραμμές με την I να έχει το μεγαλύτερο μέγεθος ενώ επιβεβαιώθηκαν και τα συμπλέγματα κλώνων (πχ CC1, **Εικόνα 33**).



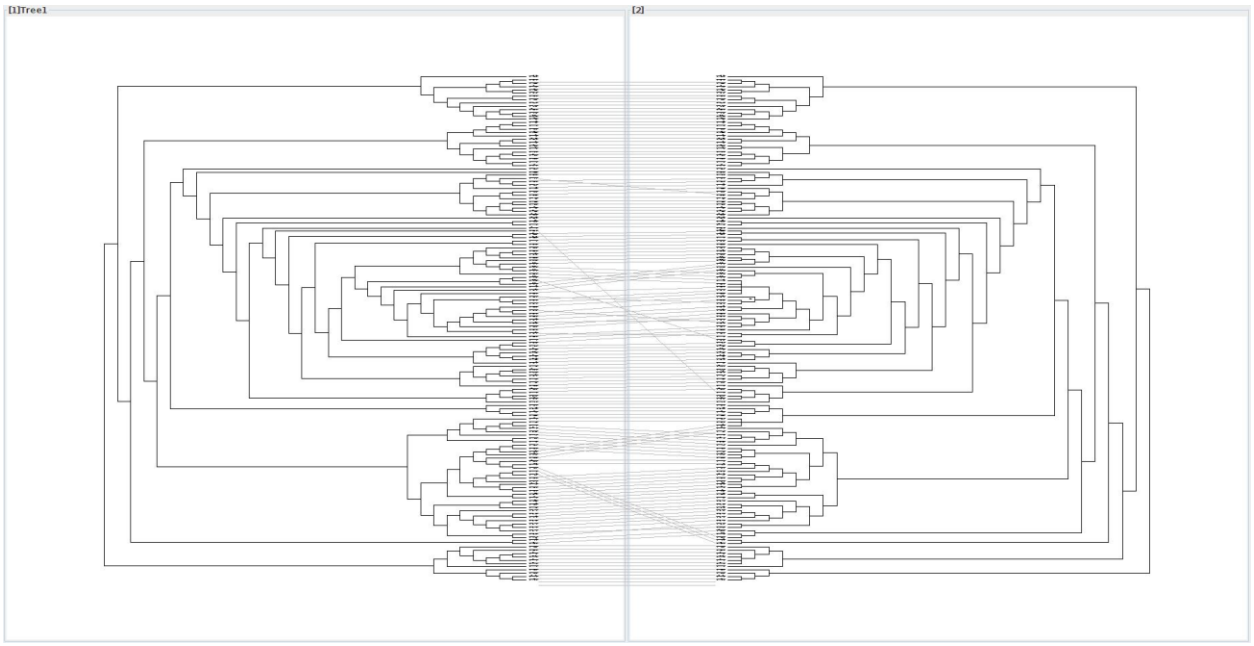
Εικόνα 33: Αποτελέσματα του αλγορίθμου Harvest. Η απεικόνιση αφορά την οθόνη που προκύπτει από την ανάλυση με το πρόγραμμα GINGR του αρχείου εξόδου .ggr του προγράμματος parsnp. Στο φυλογενετικό δέντρο τα χρώματα αντιστοιχούν στις τιμές bootstrap με το έντονο μαύρο να αντιστοιχεί σε bootstrap 1. Οι σκιάσεις αντιστοιχούν στον βαθμό διαφορών με το γένωμα αναφοράς, σύμφωνα με την χρωματική κλίμακα που δίνεται κάτω δεξιά.

Με σκοπό να εξαχθούν περισσότερο ασφαλή συμπεράσματα ακολούθησε λεπτομερής σύγκριση των δέντρων που προέκυψαν από τα 1397 κοινά γονίδια των 154 γενωμάτων με τους αλγορίθμους ML των λογισμικών RaxML (μοντέλο υποκαταστάσεων DNA GTR-GAMMA, 1397 partitions) και FastTree (μοντέλο GTR-CAT) και του δέντρου που παρήγαγε το Harvest.

Το δέντρο του FastTree συγκρινόμενο με το δέντρο του Harvest έδωσε fraction=0.68 ενώ η σύγκριση RaxML-Harvest fraction=0.69, ενώ το δέντρο του RaxML συγκρινόμενο με το δέντρο του FastTree έδωσε fraction=0.78. Όσο πιο κοντά στο 1 είναι η παράμετρος fraction τόσο πιο όμοια είναι τα συγκρινόμενα δέντρα, έτσι το δέντρο του FastTree είναι πιο κοντά με το δέντρο του RaxML, σε σχέση με το δέντρο του Harvest.

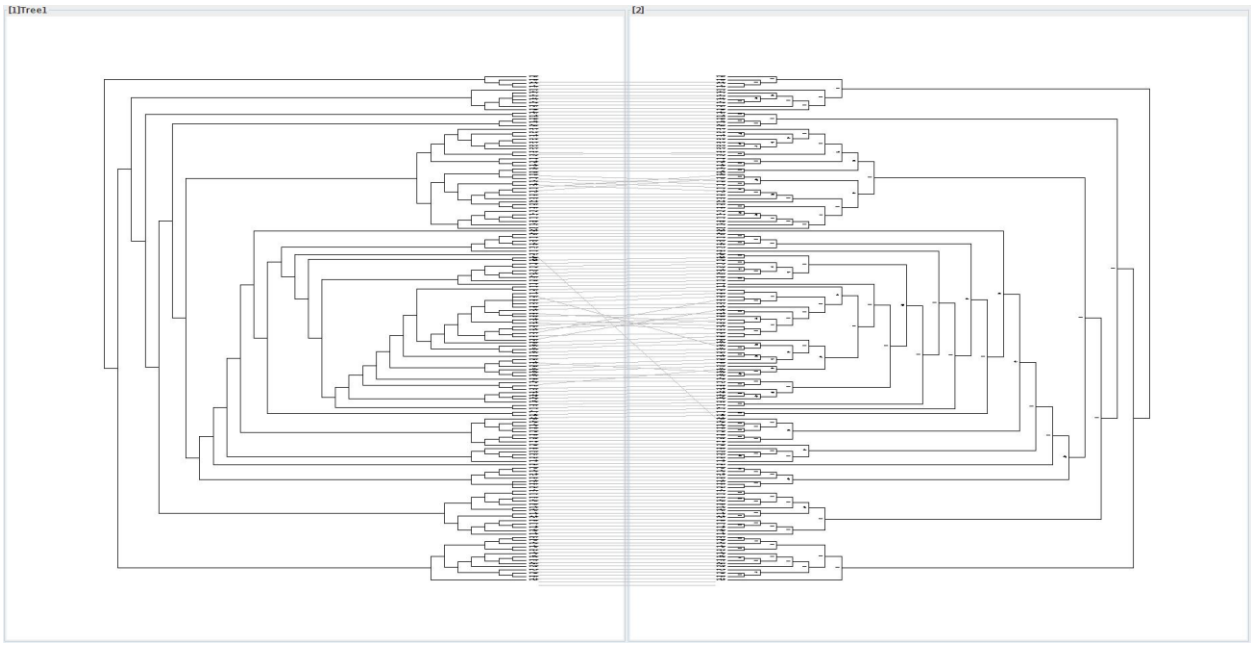
Ο αλγόριθμος tanglegram του Dendroscope επέτρεψε την οπτικοποίηση των διαφορών μεταξύ των δέντρων. Παρατηρήθηκε ότι και στις 3 συγκρίσεις οι διαφορές εντοπίζονταν κυρίως εντός του κλάδου I. Ενώ πράγματι τα δέντρα RaxML και FastTree ήταν περισσότερο όμοια (**Εικόνες 34-36**), χωρίς όμως να παρατηρούνται σημαντικές διαφορές σε καμία από τις 3 συγκρίσεις.

RaxML **Harvest**



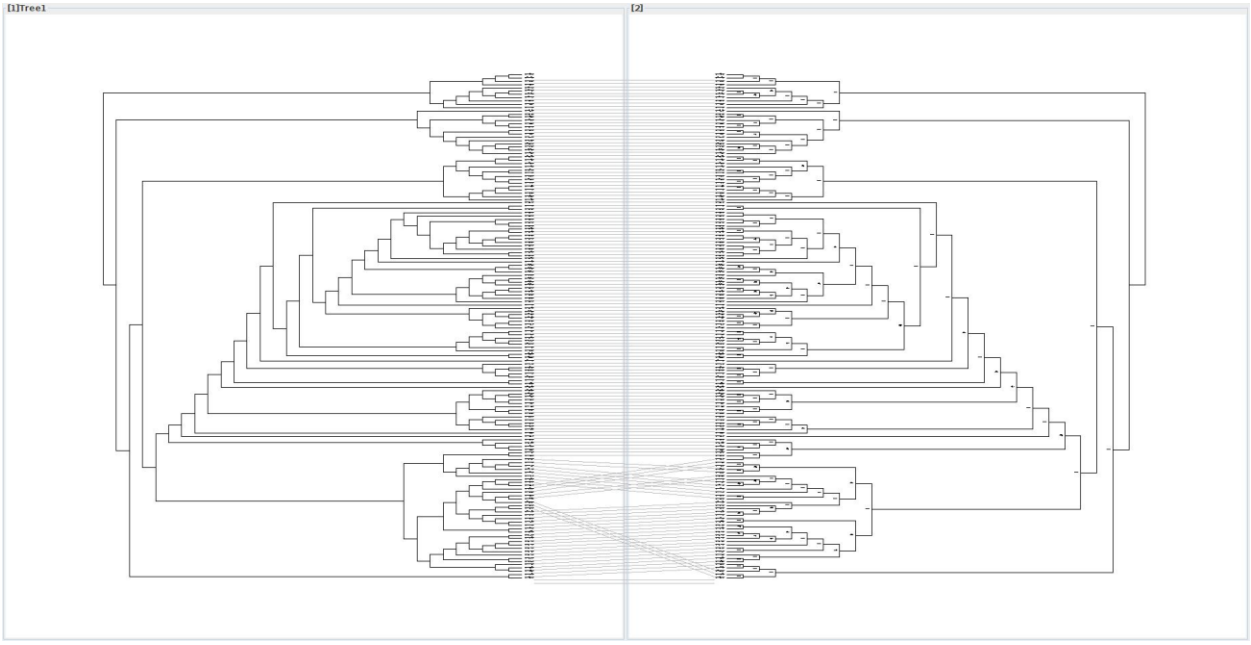
Εικόνα 34. Σύγκριση φυλογενετικών δέντρων από το RAxML και το Harvest, με τη χρήση του αλγορίθμου Tanglegram του Dendroscope.

Harvest **Fasttree**



Εικόνα 35. Σύγκριση φυλογενετικών δέντρων από το Harvest και το FastTree, με τη χρήση του αλγορίθμου Tanglegram του Dendroscope.

RaxML **FastTree**



Εικόνα 36. Σύγκριση φυλογενετικών δέντρων από το RAxML και το FastTree, με τη χρήση του αλγορίθμου Tanglegram του Dendroscope.

4. Συζήτηση

4.1. Κριτήρια για την εύρεση του core genome

Το 2005 οι Tettelin et al. παρουσίασαν την έννοια του πανγονιδιώματος, προσπαθώντας να περιγράψουν την ολότητα της γενετικής πληροφορίας ενός βακτηριακού είδους. Μεμονωμένα βακτηριακά στελέχη περιέχουν στο γονιδίωμα τους ένα συγκεκριμένο - περιορισμένο αριθμό γονιδίων, και συγκρίνοντας τα γονίδια αυτά με τα γονίδια άλλων στελεχών του ίδιου είδους παρατηρούμε μεγάλες διαφορές μεταξύ των στελεχών. Έτσι τα γονίδια ενός μόνο στελέχους είναι πολύ λιγότερα από το σύνολο των διαφόρων γονιδίων που συναντώνται στα στελέχη του είδους αυτού. Το πλήρες σύνολο των γονιδίων, ή ακριβέστερα των αντιπροσωπευτικών γονιδίων των οικογενειών ορθόλογων γονιδίων, που συναντώνται στα στελέχη ενός είδους αποτελεί το πανγονιδίωμα (pangenome), ενώ το core genome περιέχει μόνο τα γονίδια ή τα αντιπροσωπευτικά γονίδια των οικογενειών ορθόλογων γονιδίων, στις οποίες συναντώνται όλα τα στελέχη του είδους, το dispensable genome τα αντιπροσωπευτικά γονίδια οικογενειών στις οποίες συναντώνται παραπάνω από ένα αλλά όχι όλα τα στελέχη, και το unique genome περιλαμβάνει τα γονίδια που συναντώνται μόνο σε ένα στέλεχος του είδους αυτού.

Στην παραπάνω περιγραφή δεν εξηγείται ότι καθώς είναι αδύνατο να απομονώσουμε και να αλληλουχίσουμε όλα τα στελέχη ενός βακτηριακού είδους, οι έννοιες του core genome και του pangenome είναι σχετικές και αφορούν σε ένα συγκεκριμένο δείγμα στελεχών του είδους που χρησιμοποιούμε. Για να είναι όσο το δυνατόν πιο αντιπροσωπευτικό το δείγμα των στελεχών ενός βακτηριακού είδους, θα πρέπει να παρθούν στελέχη από διάφορα γεωγραφικά μέρη και ενδιαιτήματα. Επίσης, η ποιότητα των γενωμάτων των στελεχών που χρησιμοποιούμε καθορίζει σημαντικά την περαιτέρω ανάλυση, καθώς σε χαμηλής ποιότητας γενώματα είτε θα λείπουν γονίδια και λανθασμένα τα γονίδια αυτά θα ταξινομηθούν στο dispensable genome, είτε κάποια γονίδια θα παρουσιάζουν, λόγω λαθών στην αλληλούχιση, σημαντικές διαφορές με τα ορθόλογα γονίδια των υπολοίπων στελεχών και έτσι δεν θα ομαδοποιούνται μαζί και λανθασμένα πάλι τα γονίδια αυτά δεν θα τοποθετούνται στο core genome. Γι' αυτό η χρήση καλής ποιότητας γενωμάτων από διάφορες περιοχές και διαφορετικά ενδιαιτήματα είναι πολύ σημαντική για τη δημιουργία ολοκληρωμένου pangenome και core genome ενός βακτηριακού είδους.

Στην ανάλυση μας για το core genome του βακτηρίου *P. mirabilis* χρησιμοποιήσαμε τα γενώματα που είναι κατατεθειμένα στη βάση δεδομένων GenBank. Για την επιλογή των στελεχών, που θεωρήσαμε αποδεκτά, χρησιμοποιήσαμε δύο βασικά κριτήρια. Αρχικά δεχτήκαμε τα στελέχη που το γένωμα τους ήταν πλήρως ολοκληρωμένο, δηλαδή το χρωμόσωμα τους δεν περιέχει κανένα κενό και δεν υπάρχουν μη τοποθετημένα scaffolds. Και το δεύτερο κριτήριο είναι ότι για όσα γενώματα δεν είναι πλήρη, πρέπει το N50 να είναι πάνω από 100.000. Το N50 είναι ένα μετρικό μέγεθος, και ορίζεται ως το μήκος του μικρότερου contig, το οποίο μαζί με όλα τα μεγαλύτερα από αυτό contigs, καλύπτουν τουλάχιστον το 50% του γενώματος. Με το N50 μπορούμε να αξιολογήσουμε την ποιότητα ενός γενώματος, καθώς μικρά N50 δείχνουν ότι η αλληλούχιση δεν κατάφερε να δημιουργήσει contigs με μεγάλο μέγεθος. Δεν υπάρχει κάποιο συγκεκριμένο όριο του N50 που να ξεχωρίζει τα καλής ποιότητας από τα κακής ποιότητας γενώματα, γι' αυτό και στην ανάλυση μας χρησιμοποιήσαμε ένα αρκετά υψηλό όριο για το N50, ώστε όσα γενώματα χρησιμοποιήσουμε να είναι υψηλής ποιότητας. Στη συνέχεια, εκτός αυτών των δύο κριτηρίων, χρησιμοποιήσαμε και ένα τρίτο κριτήριο, το οποίο αφορά στο ότι αφαιρέσαμε γενώματα τα οποία δεν περιείχαν πολλά από τα γονίδια που συναντωνται στην συντριπτική πλειοψηφία των στελεχών (να λείπουν 30% των γονιδίων που συναντώνται στο 98% των στελεχών). Δηλαδή εάν ένα στέλεχος έχει χάσει πολλά γονίδια τα οποία συναντώνται σχεδόν σε όλα τα υπόλοιπα στελέχη, πιθανότατα αυτό οφείλεται σε λάθη της αλληλούχισης του και γι' αυτό δεν το χρησιμοποιούμε στην ανάλυση μας. Η επιλογή στελεχών με καλής ποιότητας γενώματα είναι πολύ βασική για την πληρότητα της ανάλυσης, καθώς η χρήση κακής ποιότητας γενωμάτων θα μείωνε αριθμητικά κατά πολύ το core genome. Στη βιβλιογραφία έχουν γίνει προσπάθειες για να αξιολογήθουν τα γενώματα που

χρησιμοποιούνται στις αναλύσεις, για παράδειγμα η van Vliet το 2017 χρησιμοποίησε ως κριτήριο τον αριθμό των contigs και γενώματα με παραπάνω από 100 contigs αφαιρούνταν από την ανάλυση τους, ενώ οι Lukjancenko et al. το 2010 αξιολόγησαν τα γενώματα που θα χρησιμοποιήσουν, χειροκίνητα, δίνοντας τους σκόρ από 1 έως 4, με 1 να είναι το καλύτερο δυνατό σκορ.

Εκτός από τα κριτήρια για την επιλογή των στελεχών που θα χρησιμοποιηθούν στην ανάλυση, ένα μεγάλο θέμα είναι και τα κριτήρια με βάση των οποίων θα ομαδοποιηθούν στα διάφορα clusters τα γονίδια. Η ομαδοποίηση σε clusters των γονιδίων, συνήθως γίνεται συγκρίνοντας τις πρωτεΐνες των γονιδίων αυτών. Οι πρωτοπόροι Tettelin et al. το 2005 στην εργασία τους για το πανγονιδίωμα του *Streptococcus agalactiae* χρησιμοποίησαν ως κριτήρια για την ομαδοποίηση των γονιδίων στο ίδιο cluster, οι πρωτεΐνες που κωδικοποιούνται από τα γονίδια του ίδιου cluster να εμφανίζουν τουλάχιστον 50% ομοιότητα για τουλάχιστον 50% του μήκους τους. Για τα κριτήρια για την ομαδοποίηση στο ίδιο cluster στη διεθνή βιβλιογραφία υπάρχει πληθώρα απόψεων, το 2007 ο Hiller et al. χρησιμοποίησαν ως κριτήριο 70% ομοιότητα για τουλάχιστον 70% του μήκους των γονιδίων, το 2010 οι Lukjancenko et al. χρησιμοποίησαν το ίδιο κριτήριο με την αρχική εργασία των Tettelin et al., ενώ το 2014 οι Meric et al. χρησιμοποίησαν ως κριτήριο: 70% ομοιότητα για τουλάχιστον 50% του μήκους των γονιδίων. Για την εύρεση των κριτηρίων που θέσαμε, δοκιμάσαμε διαφορα ποσοστά ομοιότητας και διάφορα ποσοστά στοίχισης των γονιδίων. Με δεδομένο ότι θέλαμε να έχουμε τον μικρότερο δυνατό αριθμό clusters, που να περιείχαν και παράλογες πρωτεΐνες, χωρίς όμως να μειώσουμε σημαντικά τον αριθμό του core genome καταλήξαμε στο να χρησιμοποιήσουμε ως κριτήριο το 75% ποσοστό ομοιότητας για 70% στοίχισης των γονιδίων.

Στη συνέχεια, εισάγαμε ένα ακόμη κριτήριο στην ομαδοποίηση των γονιδίων, το οποίο είναι το ελάχιστο μήκος αμινοξέων που πρέπει να έχει το προϊόν των γονιδίων που θα χρησιμοποιηθούν στην ανάλυση. Το Prokka εντοπίζει πολλά γονίδια, τα οποία δημιουργούν πολλές μικρές πρωτεΐνες, οι περισσότερες των οποίων είναι υποθετικές και συνήθως δεν αφορούν σε λειτουργικά γονίδια. Γι' αυτό απαιτείται ένα κριτήριο μήκους, ώστε να μην εισαχθεί στο πανγονιδίωμα μια πληθώρα από υποθετικές, μη λειτουργικές πρωτεΐνες. Για τον ορισμό του ορίου, χρησιμοποιήσαμε δεδομένα από τη βιβλιογραφία για την διακύμανση του μήκους των πρωτεϊνών της *Escherichia coli*, καθώς δεν υπήρχαν αντίστοιχα δεδομένα για τον *P. mirabilis*. Το όριο τέθηκε με γνώμονα το 90% των πρωτεϊνών να είχαν μεγαλύτερο μήκος και να γίνονταν δεκτές στην ανάλυση. Τέλος, το τελευταίο κριτήριο που θέσαμε ώστε να δημιουργήσουμε το core genome είναι τα clusters που θα περιέχονται σε αυτό να μην περιέχουν πολλαπλά αντίγραφα στα γενώματα. Από το core genome θέλαμε να αποφύγουμε τόσο τα ψευδογονίδια όσο και γονίδια με πολλά αντίγραφα, καθώς και τα δύο στοιχεία δεν βρίσκονται κάτω από εξελικτική πίεση και δεν θα μας ενίσχυαν τη διακριτική ικανότητα της φυλογενετικής ανάλυσης μας (Pearce et al., 2018).

Τα κριτήρια που θα χρησιμοποιηθούν για την ομαδοποίηση των γονιδίων είναι ιδιαίτερα σημαντική απόφαση και εξαρτάται από το είδος της ανάλυσης που πρόκειται να πραγματοποιηθεί. Στην παρούσα εργασία στόχος ήταν να δημιουργήσουμε ένα core genome ώστε να προχωρήσουμε σε φυλογενετική ανάλυση των στελεχών, γι' αυτό ήμασταν αυστηροί στα κριτήρια με τα οποία ομαδοποιούμε τις πρωτεΐνες, ώστε όσα γονίδια θα χρησιμοποιούσαμε, να ήταν πράγματι γονίδια του core genome, τα οποία είναι συντηρημένα και περιέχουν σημαντική εξελικτική πληροφορία για τα στελέχη. Πιθανώς σε μελέτες που στόχο έχουν την λειτουργική αξιολόγηση του core genome να θέτουν πιο χαλαρά κριτήρια, ώστε να μην παραληφθεί κάποιο γονίδιο. Ανάλογα με το σκοπό της μελέτης αλλάζουν και τα κριτήρια με τα οποία ομαδοποιούνται τα γονίδια, γι' αυτό και δεν υπάρχει ένας απόλυτος κανόνας με τον οποίο ομαδοποιούνται τα γονίδια για το core genome.

4.2. Λειτουργικός σχολιασμός του core / dispensable / unique / pangenome

Τα γονίδια του core genome αποτελούν τα γονίδια που είναι υπεύθυνα για τις βασικές λειτουργίες και τα φαινοτυπικά χαρακτηριστικά του βακτηριακού είδους. Στα στελέχη του είδους *P.mirabilis,* το core genome αποτελούσε ~35% του γενώματος κάθε στελέχους και το ~8% του πανγενώματος του είδους. Το υπόλοιπο 65% αφορούσε στο dispensable και unique genome, σε γονίδια δηλαδή τα οποία πρόσθεταν στο βακτήριο επιπλέον βιοχημικά μονοπάτια και λειτουργίες, επιτρέποντας του την προσαρμογή σε διαφορετικά ενδιαιτήματα (Medini et al., 2005).

Στην λειτουργική ανάλυση του core genome του *P. mirabilis*, βρήκαμε ότι ~40% του core genome αφορούσε σε λειτουργίες του μεταβολισμού, ~14% σε μετάφραση και μεταγραφή, 7% στη σύνθεση του κυτταρικού τοιχώματος, ~4% σε αντιγραφή και ~14% σε λοιπές κυτταρικές διεργασίες. Το εύρημα αυτό συνάδει και με άλλες προσπάθειες λειτουργικού σχολιασμού του core genome άλλων βακτηριακών ειδών και στελεχών. Οι Bosi et al. το 2016 βρήκαν ότι το 36% του core genome του *Streptococcus aureus* αφορά στο μεταβολισμό, ~30% σε μετάφραση και μεταγραφή, ~4% σε αντιγραφή και 5% στη βιοσύνθεση του τοιχώματος και ~25% σε λοιπές κυτταρικές διεργασίες. Οι Kim et al. το 2017 ανέλυσαν το πανγονιδίωμα του γένους *Bacillus* και βρήκαν ότι το ~43% του core genome του γένους αυτού αφορά σε λειτουργίες του μεταβολισμού, ~14% σε μετάφραση και μεταγραφή, ~6% σε βιοσύνθεση κυτταρικού τοιχώματος, ~3,5% σε αντιγραφή και ~24% σε λοιπές κυτταρικές διεργασίες, και τέλος οι Yang et al. το 2016 βρήκαν ότι το 44% του core genome της *Brucella* αφορά στο μεταβολισμό, ~17,5% σε μετάφραση και μεταγραφή, ~8% σε βιοσύνθεση κυτταρικού τοιχώματος, ~4% σε αντιγραφή και ~23% σε λοιπές κυτταρικές διεργασίες. Και στις τέσσερις αναλύσεις το μεγάλο ποσοστό του core genome αφορά σε λειτουργίες του μεταβολισμού, καθώς αυτές είναι απαραίτητες για τη συντήρηση και διατήρηση του βακτηρίου, αλλά και απαιτούν πολλές πρωτεΐνες για να πραγματοποιηθούν. Σε μικρότερο ποσοστό του core genome παρατηρούμε λειτουργίες της αντιγραφής, μεταγραφής, μετάφρασης και βιοσύνθεσης του κυτταρικού τοιχώματος, καθώς είναι απαραίτητες για την λειτουργία του κυττάρου. Σίγουρα το core genome του *P. mirabilis* δεν περιέχει μονάχα τα εντελώς απαραίτητα για την βακτηριακή επιβίωση γονίδια (essential genes), αλλά περιλαμβάνει και κάποια γονίδια που χαρακτηρίζουν το είδος *P.mirabilis.*

Το dispensable και το unique genome παρουσιάζουν παρόμοια ποσοστά γονιδίων για τις βασικές λειτουργικές κατηγορίες. Συγκεκριμένα στο dispensable και στο unique genome, ~18% και ~24,5% των γονιδίων αφορά στο μεταβολισμό, ~8% σε γονίδια για τη μεταγραφή και τη μετάφραση, ~10% σε διεργασίες της αντιγραφής και της επιδιόρθωσης του γενετικού υλικού, ~8 στη σύνθεση του κυτταρικού τοιχώματος και ~14% και 12% στις λοιπές κυτταρικές διεργασίες, αντίστοιχα. Οι Bosi et al. βρήκαν ότι το 29% του dispensable genome του *Streptococcus aureus* αφορά στο μεταβολισμό, ~24% σε μετάφραση και μεταγραφή, ~24% σε αντιγραφή, 4% στη βιοσύνθεση του τοιχώματος και ~19% σε λοιπές κυτταρικές διεργασίες. Οι διαφορές που παρατηρούνται οφείλονται στο γεγονός ότι στην εργασία των Bosi et al. δεν υπήρχαν πρωτεΐνες με άγνωστη λειτουργία. Ενώ οι Kim et al. για το γένος *Bacillus* βρήκαν ότι το ~27,5% του unique genome αφορά σε λειτουργίες του μεταβολισμού, ~8% σε μετάφραση και μεταγραφή, ~10,5% σε βιοσύνθεση κυτταρικού τοιχώματος, ~12% σε αντιγραφή και ~30% σε λοιπές κυτταρικές διεργασίες. Τα αποτελέσματα των τριών αναλύσεων συνάδουν μεταξύ τους, αν και εντύπωση προκαλεί η αρκετά μεγαλύτερη εκπροσώπηση της κατηγορίας για την αντιγραφή και την επιδιόρθωση στο unique και dispensable genome σε σχέση με το core genome. Αυτό οφείλεται κυρίως στο γεγονός ότι στην κατηγορία του COG αντιγραφή και επιδιόρθωση ανήκουν και πολλά γονίδια που σχετίζονται με μεταθετά στοιχεία. Στην δική μας ανάλυση για την κατηγορία των γονιδίων που σχετίζονται με την αντιγραφή και επιδιόρθωση (L), το μεγαλύτερο ποσοστό του dispensable genome και του unique genome αφορούσε σε γονίδια για τρανσποζάσες και ιντεγκράσες.

Όσον αφορά στα γονίδια ανθεκτικότητας σε αντιμικροβιακούς παράγοντες, παρατηρούμε ότι κυρίως γονίδια ανθεκτικότητας βρίσκονται στο dispensable και στο unique genome. Στο core genome του *P. mirabilis* δεν συναντώνται γονίδια ανθεκτικότητας, γεγονός που φαντάζει περίεργο καθώς τα στελέχη του *P. mirabilis*

έχουν φυσική αντοχή σε τετρακυκλίνη, τιγεκυκλίνη και κολιστίνη. Συγκεκριμένα η αντοχή σε τετρακυκλίνες στον *P. mirabilis* οφείλεται στο χρωμοσωμικό γονίδιο ανθεκτικότητας tetJ (Stock, 2003). Στην ανάλυση μας το γονίδιο αυτό ταξινομήθηκε στο dispensable genome και όχι στο core genome, παρόλο που με την ανάλυση από το ResBar συναντάται ανθεκτικότητα στην τετρακυκλίνη σε όλα τα γενώματα, καθώς το CD-HIT δεν κατέταξε τα γονίδια tetJ όλων των στελεχών στο ίδιο cluster αλλά σε δύο διαφορετικά clusters, σε ένα που περιείχε τα γονίδια από τα 152 γενώματα και σε ένα άλλο που περιείχε τα άλλα 2. Αυτό συνέβη καθώς στα 2 γενώματα (*P. mirabilis* SCDR1 : GCA_002013325.1 και *P .mirabilis* pmi_p1 : GCA_901485075.1) που ομαδοποιήθηκαν μαζί, το prokka προέβλεψε το γονίδιο αυτό αρκετά μικρότερο από ό,τι στα υπόλοιπα γενώματα, καθώς μία έλλειψη ενός νουκλεοτιδίου αλλάζει το πλαίσιο ανάγνωσης και δημιουργούσε ένα πρόωρο κωδικόνιο λήξης. Συγκεκριμένα στα 2 γενώματα το γονίδιο είχε 281 aa (DIADFPKC_00114) και στα υπόλοιπα 397aa (KPDLDLLF_02405). Όσον αφορά στην ανθεκτικότητα στην κολιστίνη, ο *P. mirabilis* εμφανίζει ενδογενή αντοχή στην κολιστίνη, καθώς το συγκεκριμένο αντιβιοτικό δεν μπορεί να συνδεθεί με τον λιποπολυσακχαρίτη (LPS) της εξωτερικής μεμβράνης του *P. mirabilis* και συνεπώς δεν μπορεί να ασκήσει την βακτηριοκτόνο δράση της στο *P. mirabilis* (Aghapour et al., 2019). Τέλος, όσον αφορά στην τιγεκυκλίνη, αν και δεν εντόπισε το AMRFinder στο core genome γονίδιο ανθεκτικότητας, υπάρχει στο core genome το γονίδιο acrB, το οποίο απομακρύνει το αντιβιοτικό από το εσωτερικό των βακτηρίων προσδίδοντας στο *P. mirabilis* ανθεκτικότητα στην τιγεκυκλίνη (Visalli et al., 2003).

Στη συνέχεια, από την ανάλυση μας συμπεράναμε ότι το πανγονιδίωμα του βακτηρίου *P. mirabilis* είναι ανοικτό. Ανοικτό πανγονιδίωμα εννοούμε το πανγονιδίωμα ενός βακτηρίου για το οποίο όσο περισσότερα στελέχη αλληλουχούμε τόσο περισσότερα νέα γονίδια θα προστίθενται στο πανγονιδίωμα του. Η προσθήκη νέων γονιδίων αυξάνει τη γενετική ποικιλομορφία ενός βακτηριακού είδους, και προκαλείται από την οριζόντια μεταφορά γενετικού υλικού μεταξύ των βακτηριακών στελεχών. Η οριζόντια μεταφορά γονιδίων (Horizontal Gene Transfer, HGT) πραγματοποιείται με τρεις τρόπους: το μετασχηματισμό, δηλαδή την ενσωμάτωση απευθείας εξωγενούς γενετικού υλικού από το περιβάλλον, τη μεταγωγή, δηλαδή την είσοδο της γενετικής πληροφορίας μέσα από βακτηριοφάγους, και τη σύζευξη, δηλαδή την απευθείας μεταφορά γενετικού υλικού μεταξύ βακτηριακών κυττάρων. Η γενετική ποικιλομορφία που παρουσιάζει ένα βακτηριακό είδος σχετίζεται άμεσα με τους τρόπους και τη συχνότητα που συμβαίνει οριζόντια μεταφορά γονιδίων στα στελέχη του και κατ' επέκταση σχετίζεται με την οικολογία και το ενδιαίτημα του βακτηριακού είδους (Azarian et al., 2020). Η εξερεύνηση νέων ενδιαιτημάτων επιτάσσει τη διαφοροποίηση και την προσαρμογή στο νέο περιβάλλον μέσα από την απόκτηση νέων γονιδίων και τη γενετική ποικιλομορφία (Sheppard et al., 2011). Έτσι παρατηρούμε ότι βακτήρια με πολύ περιορισμένα ενδιαιτήματα παρουσιάζουν κλειστό πανγονιδίωμα, όπως ο *Bacillus anthracis,* ενώ βακτήρια με πολλά διαφορετικά ενδιαιτήματα και μεγάλη δυνατότητα για οριζόντια μεταφορά γονιδίων παρουσιάζουν ανοικτό πανγονιδίωμα, όπως η *Escherichia coli* και ο *P. mirabilis* (Rasco et al., 2008).

4.3. core genome MLST και φυλογένεση του P. mirabilis

Η εύρεση διαφορετικών χαρακτηριστικών, είτε γενετικών είτε φαινοτυπικών, μεταξύ των βακτηριακών γενών και ειδών επιτρέπει τόσο την ταυτοποίηση των νέων βακτηρίων αλλά και τη δημιουργία φυλογενετικών δέντρων, ώστε να γίνουν εμφανείς οι εξελικτικές αποστάσεις μεταξύ των διαφορετικών ειδών και στελεχών. Στην σημερινή εποχή η ταυτοποίηση των βακτηρίων και η φυλογένεση τους πραγματοποιείται με τη χρήση κυρίως γενετικών χαρακτηριστικών, με πιο συχνές μέθοδους, τη PCR-ριβοτυπία (ribotyping), το MLST και με ηλεκτροφόρηση παλλόμενου πεδίου (PFGE). Η χρήση του 16s Rna γονιδίου με τη μέθοδο της ριβοτυπίας αν και είναι ιδιαίτερα συχνή σε φυλογενετικές μελέτες, στην ομοταξία των *Enterobacteriaceae* φαίνεται να μην μπορεί να διαχωρίσει ικανοποιητικά τα στελέχη στα διάφορα γένη και είδη (Naum et al., 2008). Η μέθοδος PFGE θεωρείται ως η πρότυπη μέθοδος για την τυποποίηση και

τη φυλογένεση βακτηριακών στελεχών, όμως το βασικό μειονέκτημα της είναι ότι απαιτεί τη ταυτόχρονη σύγκριση όλων των στελεχών μεταξύ τους, δηλαδή όσα στελέχη συγκριθούν πρέπει να ηλεκτροφορηθούν την ίδια χρονική στιγμή στο ίδιο εργαστήριο ώστε να είναι εφικτή η σύγκριση τους. Έτσι η τεχνική αυτή καθιστά αδύνατη την απομακρυσμένη σύγκριση στελεχών (Neoh et al., 2019). Στη συνέχεια, η τεχνική του MLST στηρίζεται στο γεγονός ότι τα στελέχη διαχωρίζονται με βάση τη γενετική τους ομοιότητα όχι μόνο ως προς ένα γονίδιο, όπως το 16S Rrna γονίδιο, αλλά ως προς ένα σύνολο συντηρημένων γενετικών τόπων (συνήθως 6 έως 9) κατάλληλα επιλεγμένων ώστε να μπορέσουν να αποδώσουν τις βαθιές φυλογενετικές σχέσεις των κλώνων του είδους. Στην σημερινή εποχή, με την έκρηξη που έχει προκαλέσει το NGS είναι δυνατή η σύγκριση των στελεχών όχι ως προς ένα περιορισμένο σύνολο γονιδίων αλλά ως προς όλα τα γονίδια που έχουμε χαρακτηρίσει ως core genome (μέθοδος cgMLST). Με τη χρήση του cgMLST κάθε στέλεχος αποκτά ένα συγκεκριμένο cgMLST τύπο, που περιέχει για κάθε γενετικό τόπο τον αριθμό του αλληλίου που φέρει το συγκεκριμένο στέλεχος. Εκτός από τη τυποποίηση των στελεχών με το cgMLST, συγκρίνοντας τους διάφορους cgMLST τύπους μπορούμε να βρούμε και τα διάφορα κλωνικά συμπλέγματα, δηλαδή ομάδες cgMLST τύπων που είναι εξελικτικά κοντά και διαφέρουν σε μικρό αριθμό γενετικών τόπων. Το όριο για το σε πόσους γενετικούς τόπους πρέπει να διαφέρουν δύο cgMLST τύποι για να ενταχθούν στο ίδιο κλωνικό σύμπλεγμα δεν είναι συγκεκριμένο, καθώς το core genome για κάθε βακτηριακό είδος έχει διαφορετικό μέγεθος. Οι Bialek-Davenetet al. το 2014 χρησιμοποιήσαν ως όριο για να κατατάξουν στο ίδιο κλωνικό σύμπλεγμα δύο στελέχη *Klepsiella pneumoniae*, να διαφέρουν σε λιγότερο από 100 γενετικούς τόπους (το core genome της *K.pneumoniae* στην εργασία αυτή είχε βρεθεί 694 γονίδια). Οι Been et al. το 2015 για να ομαδοποιήσουν σε κλωνικά συμπλέγματα τα στελέχη του *Enterococcus faecium* χρησιμοποίησαν ως όριο τη διαφορά σε λιγότερο από 20 γενετικούς τόπους (με core genome 1423 γονίδια). Εμείς θέσαμε ως όριο τη διαφορά σε λιγότερους από 430 γενετικούς τόπους, και παρατηρήσαμε ότι τα κλωνικά συμπλέγματα που προκύπτουν από αυτό το όριο συμφωνούν με τα δεδομένα από το φυλογενετικό δέντρο, στο οποίο οι cgMLST τύποι που είχαν διαφορά σε λιγότερους από 430 τόπους τοποθετούνταν πολύ κοντά, ενώ δεν υπήρχε σε καμία περίπτωση κάποιος cgMLST τύπος που να είχε τοποθετηθεί σε κάποιο κλωνικό σύμπλεγμα και στο δέντρο να είχε τοποθετηθεί μακριά από τα υπόλοιπα μέλη του συμπλέγματος. Το γεγονός ότι χρησιμοποιήσαμε πολύ αυστηρά κριτήρια για την ένταξη των διαφόρων γενετικών τόπων στο core genome, πιθανώς μας επιτρέπει τη χρήση πιο διευρυμένου ορίου για την κατάταξη των cgMLST τύπων σε κλωνικά συμπλέγματα. Η κατάταξη των διαφόρων στελεχών σε κλωνικά συμπλέγματα με τη χρήση του cgMLST μας δίνει τη δυνατότητα να ταξινομήσουμε τους εξελικτικά κοντινούς κλώνους, και έτσι να παρακολουθήσουμε και να ιχνηλατήσουμε τις σχέσεις των διαφόρων στελεχών σε τυχόν επιδημίες.

Επιπλέον με τη χρήση core genome εκτός από την τυποποίηση των διαφόρων στελεχών σε cgMLST τύπους και τον καθορισμό των διαφόρων κλωνικών συμπλεγμάτων μας δίνεται η δυνατότητα να κάνουμε και υψηλής διακριτικής ικανότητας φυλογενετικές αναλύσεις. Το φυλογενετικό δέντρο που κατασκευάσαμε με τα στελέχη του γένους *Proteus* δείχνει ότι με τη χρήση του core genome είναι εφικτός ο διαχωρισμός των διαφόρων ειδών του *Proteus*. Στο δέντρο του γένους *Proteus* τα στελέχη του *P. mirabilis* και του *P. vulgaris* που δεν έχουν ταξινομηθεί μαζί με το είδος τους, πιθανότατα ανήκουν σε άλλο είδος ή αποτελούν κάποιο νέο είδος *Proteus*. Επίσης, με τη φυλογένεση με το core genome, μπορούμε να κατατάξουμε ορισμένα στελέχη *Proteus sp*., τα οποία προηγουμένως δεν έχουν ταυτοποιηθεί, στα διάφορα είδη *Proteus*. Επίσης από το φυλογενετικό δέντρο για την οικογένεια *Morganellaceae* βλέπουμε ότι με τη χρήση του core genome διαχωρίζονται σαφώς και τα γένη της οικογένειας *Morganellaceae*. Με τη χρήση των 357 κοινών γονιδίων της οικογένειας *Morganellaceae,* όλα τα στελέχη διαχωρίστηκαν και τοποθετήθηκαν στο γένος τους. Συνεπώς με τη χρήση του core genome είναι εφικτός ο διαχωρισμός σε γένη και είδη των διαφόρων στελεχών της οικογένειας *Morganellaceae.*

Με τη χρήση του core genome είναι εφικτή και η μεγάλης διακριτικής ικανότητας φυλογένεση των διαφόρων στελεχών του *P. mirabilis*. Οι εξελικτικές σχέσεις μεταξύ των διαφόρων στελεχών αποκαλύπτουν πολλά και χρήσιμα επιδημιολογικά στοιχεία σε σχέση με την εξέλιξη και τη μετάδοση τους. Για παράδειγμα, παρατηρούμε ότι τα στελέχη που φέρουν την CMY-4, δηλαδή τα στελέχη ESDY17, pmi p1, NO-051-03 και

AOUC-001 βρίσκονται εξελικτικά πολύ κοντά με τα στελέχη CYPV1 και CYPM1, τα οποία είναι ευαίσθητα στελέχη από την Ασία, και τα οποία πιθανώς να ήρθαν στην Ευρώπη όπου και απέκτησαν την CMY-4 και ως ανθεκτικά πλέον στελέχη μεταδόθηκαν στις διάφορες Ευρωπαϊκές χώρες. Επίσης, στην περίπτωση των στελεχών *P. mirabilis* που φέρουν την OXA-23 και OXA-58, η φυλογένεση με το core genome υποδεικνύει ότι τα στελέχη αυτά που εντοπίστηκαν στη Γαλλία και στην Ελβετία από το 1996 είναι εξελικτικά πολύ κοντά και συνιστούν ένα κλωνικό σύμπλεγμα, υποδεικνύοντας ότι όλα τα στελέχη *P. mirabilis* που φέρουν την OXA-23 και OXA-58 τόσο στον άνθρωπο όσο και σε ζώα στη Γαλλία και Ελβετία αποτελεί ένα βακτηριακό κλώνο που συνεχίζει να διασπείρεται. Το core genome με την υψηλής διακριτικής ικανότητα φυλογένεση που προσφέρει, μας δίνει τη δυνατότητα να αποκαλύψουμε σε βάθος τις εξελικτικές σχέσεις μεταξύ των κλώνων, καθιστώντας πιο εύκολη την παρακολούθηση και επιτήρηση διαφόρων επιδημιών.
Ένα ενδιαφέρον στοιχείο είναι ότι το φυλογενετικό δέντρο που προέκυψε χρησιμοποιώντας τα κοινά γονίδια όλων των στελεχών *P. mirabilis* που είναι κατατεθειμένα στη GenBank και όχι όλα τα γονίδια του core genome (952 γονίδια) δεν διέφερε ιδιαίτερα από το φυλογενετικό δέντρο που δημιουργήθηκε συγκρίνοντας τα στελέχη του *P. mirabilis* ως προς όλα τα γονίδια του core genome (1397 γονίδια). Η χρήση λιγότερων γονιδίων στη φυλογένεση θα μείωνε τον υπολογιστικό χρόνο της μεθόδου, αλλά θα μείωνε και την διακριτική ικανότητα της μεθόδου. Πιθανώς να μπορούμε να μειώσουμε τον αριθμό των γονιδίων που χρησιμοποιούμε στη φυλογένεση χωρίς να μειώσουμε αισθητά την διακριτική ικανότητα της μεθόδου, δημιουργώντας έτσι ένα πιο μικρό αλλά αξιόπιστο σύνολο γονιδίων τόσο για την φυλογενετική ανάλυση του είδους *P. mirabilis* όσο και του γένους *Proteus*.
Τέλος, το φυλογενετικό δέντρο που προέκυψε από τη σύγκριση των στελεχών του *P. mirabilis* ως προς όλα τα γονίδια του core genome δεν διέφερε σημαντικά από το φυλογενετικό δέντρο που προέκυψε με τη χρήση του εργαλείου του Harvest, το οποίο ακολουθεί ένα πολύ διαφορετικό αλγόριθμο για τη δημιουργία του φυλογενετικού δέντρου και του core genome. Βέβαια, θεωρούμε ότι με τη χρήση του core genome η φυλογένεση που προκύπτει διαθέτει τη μέγιστη διακριτική ικανότητα, καθώς τα στελέχη συγκρίνονται μεταξύ τους ως προς όλα τα κοινά γονίδια που φέρουν.

4.4. Βιοπληροφορικά εργαλεία

Μέσα από την παρούσα διπλωματική δημιουργήθηκαν 4 βιοπληροφορικά εργαλεία κάθε ένα από τα οποία ουσιαστικά επιτελεί ένα κομμάτι για μία πλήρη πανγονιδιωματική ανάλυση. Συγκεκριμένα το Genomes_Finder.sh αναζητά τα γενώματα της GenBank που είναι αξιόπιστα για την εύρεση του core genome και τα κατεβάζει. Το Pangenome_Finder.sh βρίσκει το core / dispensable / unique / pan genome, πραγματοποιεί την ανάλυση για το ανοικτό / κλειστό πανγονιδίωμα και δημιουργεί και το φυλογενετικό δέντρο με τα στελέχη που βρήκε το Genomes_Finder. Το cgMLST_prep.sh δημιουργεί όλα τα απαραίτητα αρχεία για το cg_finder, το οποίο ουσιαστικά για τα στελέχη που του εισάγουμε βρίσκει το cgMSLT τύπο τους και δημιουργεί και τα φυλογενετικά δέντρα των στελεχών που εισάγαμε και των στελεχών που βρήκε το Genomes_Finder. Ουσιαστικά τα 4 αυτά εργαλεία αποτελούν μία εύκολη και απλή πρόταση για μια πανγονιδιωματική ανάλυση.
Βέβαια στη βιβλιογραφία συναντάμε και άλλα προγράμματα που επιτελούν πανγονιδιωματικές αναλύσεις. Ένα βασικό πλεονέκτημα που παρουσιάζουν τα δικά μας προγράμματα είναι η αυτόματη εύρεση των γενωμάτων που είναι κατατεθειμένα στη Genbank και μπορούν να χρησιμοποιηθούν στην πανγονιδιωματική ανάλυση. Όπως αναφέραμε και παραπάνω, η χρήση πολλών, καλά αλληλουχημένων και από διάφορα γεωγραφικά μέρη στελεχών είναι ένα πολύ σημαντικό κομμάτι στην ανάλυση, ώστε το core genome να είναι αντιπροσωπευτικό του συνόλου των στελεχών του βακτηριακού είδους και όχι ενός περιορισμένου υποσυνόλου του.
Με τη χρήση των εργαλείων μας καθίσταται ιδιαίτερα εύκολη για ένα επιστήμονα, μία πλήρης πανγονιδιωματική ανάλυση ενός βακτηριακού είδους, δηλαδή η εύρεση των διαθέσιμων στελεχών της

Genbank για το βακτηριακό είδος που τον ενδιαφέρει, δημιουργία του core genome, και αξιολόγηση της φυλογένεσης των στελεχών που υπάρχουν στη διεθνή βιβλιογραφία. Επίσης με τη χρήση του cg_finder είναι πολύ εύκολη η εύρεση των cgMLST τύπων επιπλέον στελεχών και η δημιουργία του φυλογενετικού τους δέντρου ώστε να εξετάσει με ποια άλλα στελέχη είναι εξελικτικά κοντά τα στελέχη που διαθέτει. Τέλος, με τα εργαλεία που δημιουργήσαμε είναι δυνατή η περαιτέρω αξιολόγηση και σχολιασμός του core genome, είτε για την εύρεση των λειτουργικών κατηγοριών των γονιδίων που απαρτίζουν το core genome, είτε για την εύρεση γονιδίων ανθεκτικότητας σε αντιβιοτικά.
Τέτοια εργαλεία, όπως τα δικά μας, είναι ιδιαίτερα χρήσιμα σε μικροβιολογικά εργαστήρια που ασχολούνται με βακτηριακές επιδημίες, καθώς δίνουν τη δυνατότητα στους επιστήμονες να ελέγξουν εάν τα στελέχη που ευθύνονται για μία πιθανή επιδημία προέρχονται από το ίδιο στέλεχος, από πολύ κοντινά στελέχη ή απομακρυσμένα στελέχη. Στις περισσότερες περιπτώσεις επιδημιών για τη ταυτοποίηση των βακτηριακών στελεχών χρησιμοποιείται το 7-MLST, αλλά πλέον στην εποχή του NGS είναι εφικτό να χρησιμοποιούμε το cgMLST αυξάνοντας τη διακριτική ικανότητα και ανακαλύπτοντας μικρές εξελικτικές διαφορές μεταξύ των στελεχών. Η ύπαρξη απλών και εύχρηστων προγραμμάτων για την σύγκριση των στελεχών μέσω cgMLST, όπως τα δικά μας εργαλεία, διευκολύνει τους διάφορους επιστήμονες στη χρήση του core genome για φυλογενετικές αναλύσεις με μεγάλη διακριτική ικανότητα.
Ένα μειονέκτημα των εργαλείων που δημιουργήσαμε είναι αρχικά ότι για να τα χρησιμοποιήσει κάποιος πρέπει να διαθέτει linux λειτουργικό σύστημα στον υπολογιστή του. Επίσης, επειδή οι αναλύσεις για τη δημιουργία φυλογενετικών δέντρων με το core genome απαιτούν πολύ μεγάλη μνήμη, συνήθως δεν είναι εφικτή η ολοκλήρωση τους σε οικιακούς υπολογιστές με μικρή μνήμη. Βέβαια, τα εργαλεία μας δημιουργούν τα απαραίτητα αρχεία με τη στοίχιση των στελεχών και μπορεί ο χρήστης να χρησιμοποιήσει online servers που φιλοξενούν διάφορα φυλογενετικά εργαλεία, ώστε να πραγματοποιήσει τις φυλογενετικές αναλύσεις. Ένας ελεύθερα προσβάσιμος server για φυλογενετικές εργασίες είναι ο **CIPRES Science Gateway**, τον οποίο ο χρήστης μπορεί να χρησιμοποιήσει δωρεάν για να δημιουργήσει το φυλογενετικό δέντρο των στελεχών του.

4.5. Σύνοψη

Συμπερασματικά, στην παρούσα διπλωματική εργασία δημιουργήσαμε το core genome, το dispensable genome, το unique genome και το πανγονιδίωμα του βακτηρίου *Proteus mirabilis*. Είδαμε ότι μπορούμε να χρησιμοποιήσουμε το core genome για φυλογενετικές αναλύσεις, δημιουργώντας φυλογενετικά δέντρα με μεγάλη διακριτική ικανότητα, και έτσι δημιουργήσαμε ένα φυλογενετικό δέντρο με όλα τα στελέχη του *P. mirabilis* που είναι κατατεθειμένα στη GenBank, καθώς και ένα δέντρο για το γένος *Proteus* και για την οικογένεια *Morganellaceae*. Τέλος, στα πλαίσια της διπλωματικής δημιουργήσαμε και 4 προγράμματα, με τα οποία για οποιοδήποτε βακτηριακό είδος μπορούμε να κάνουμε μία πλήρη πανγονιδιωματική ανάλυση. Για την καθιέρωση του core genome και του cgMLST για την παρακολούθηση των βακτηριακών επιδημιών, απαιτούνται επιπλέον εργασίες, όπου θα συγκρίνονται τα αποτελέσματα από το MLST με τα αποτελέσματα του cgMLST, για την ομαδοποίηση των διαφόρων βακτηριακών στελεχών. Επίσης μία εκ βάθους σύγκριση μεταξύ των φυλογενετικών δέντρων που δημιουργούνται με το core genome και με άλλες μεθόδους απαιτείται ώστε να καθιερωθεί η άποψη ότι η φυλογένεση με τη χρήση του core genome παρέχει την μεγαλύτερη δυνατή διακριτική ικανότητα. Η καθιέρωση του core genome για τη φυλογένεση των βακτηριακών στελεχών και την επιδημιολογική παρακολούθηση ανθεκτικών στελεχών θα βελτιώσει σημαντικά την εικόνα που έχουμε για την εξέλιξη και τη μετάδοση των βακτηρίων.

5. Βιβλιογραφία

1. Adeolu, M., Alnajar, S., Naushad, S., & S. Gupta, R. (2016). Genome-based phylogeny and taxonomy of the ‘Enterobacteriales’: Proposal for Enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morganellaceae fam. nov., and Budviciaceae fam. nov. International Journal of Systematic and Evolutionary Microbiology, 66(12), 5575–5599.
2. Aghapour, Z., Gholizadeh, P., Ganbarov, K., Bialvaei, A. Z., Mahmood, S. S., Tanomand, A., Yousefi, M., Asgharzadeh, M., Yousefi, B., & Kafil, H. S. (2019). Molecular mechanisms related to colistin resistance in Enterobacteriaceae. Infection and Drug Resistance, 12, 965–975.
3. Aller, A. I., Castro, C., Medina, M. J., González, M. T., Sevilla, P., Morilla, M. D., Corzo, J. E., & Martín-Mazuelos, E. (2009). Isolation of Moellerella wisconsensis from blood culture from a patient with acute cholecystitis. Clinical Microbiology and Infection, 15(12), 1193–1194.
4. Arenas, M., & Posada, D. (2010). The Effect of Recombination on the Reconstruction of Ancestral Sequences. Genetics, 184(4), 1133–1139.
5. Auch, A. F., von Jan, M., Klenk, H.-P., & Göker, M. (2010). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Standards in Genomic Sciences, 2(1), 117–134.
6. Azarian, T., Huang, I.-T., & Hanage, W. P. (2020). Structure and Dynamics of Bacterial Populations: Pangenome Ecology. In H. Tettelin & D. Medini (Eds.), The Pangenome: Diversity, Dynamics and Evolution of Genomes (pp. 115–128). Springer International Publishing.
7. Belén, A., Pavón, I., & Maiden, M. C. J. (2009). Multilocus Sequence Typing. Methods in Molecular Biology (Clifton, N.J.), 551, 129–140.
8. Bialek-Davenet, S., Criscuolo, A., Ailloud, F., Passet, V., Jones, L., Delannoy-Vieillard, A.-S., Garin, B., Le Hello, S., Arlet, G., Nicolas-Chanoine, M.-H., Decré, D., & Brisse, S. (2014). Genomic Definition of Hypervirulent and Multidrug-Resistant Klebsiella pneumoniae Clonal Groups. Emerging Infectious Diseases, 20(11), 1812–1820.
9. Bonnin, R. A., Girlich, D., Jousset, A. B., Gauthier, L., Cuzon, G., Bogaerts, P., Haenni, M., Madec, J.-Y., Couvé-Deacon, E., Barraud, O., Fortineau, N., Glaser, P., Glupczynski, Y., Dortet, L., & Naas, T. (2020). A single Proteus mirabilis lineage from human and animal sources: A hidden reservoir of OXA-23 or OXA-58 carbapenemases in Enterobacterales. Scientific Reports, 10(1), 9160.
10. Bosi, E., Monk, J. M., Aziz, R. K., Fondi, M., Nizet, V., & Palsson, B. Ø. (2016). Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. Proceedings of the National Academy of Sciences, 113(26), E3801–E3809.
11. Bratcher, H. B., Corton, C., Jolley, K. A., Parkhill, J., & Maiden, M. C. (2014). A gene-by-gene population genomics platform: De novo assembly, annotation and genealogical analysis of 108 representative Neisseria meningitidis genomes. BMC Genomics, 15(1).
12. Cao, B., Wang, M., Liu, L., Zhou, Z., Wen, S., Rozalski, A., & Wang, L. (2009). 16S-23S rDNA internal transcribed spacer regions in four Proteus species. Journal of Microbiological Methods, 77(1), 109–118.
13. Chaudhari, N. M., Gupta, V. K., & Dutta, C. (2016). BPGA- an ultra-fast pan-genome analysis pipeline. Scientific Reports, 6(1), 24373.
14. Chen, Y., Gonzalez-Escalona, N., Hammack, T. S., Allard, M. W., Strain, E. A., & Brown, E. W. (2016). Core Genome Multilocus Sequence Typing for Identification of Globally Distributed Clonal Groups and Differentiation of Outbreak Strains of Listeria monocytogenes. Applied and Environmental Microbiology, 82(20), 6258–6272.
15. Dai, H., Chen, A., Wang, Y., Lu, B., Wang, Y., Chen, J., Huang, Y., Li, Z., Fang, Y., Xiao, T., Cai, H., Du, Z., Wei, Q., Kan, B., & Wang, D. (2019). Proteus faecis sp. Nov., and Proteus cibi sp. Nov., two new

species isolated from food and clinical samples in China. International Journal of Systematic and Evolutionary Microbiology, 69(3), 852–858.
16. D'Andrea, M. M., Literacka, E., Zioga, A., Giani, T., Baraniak, A., Fiett, J., Sadowy, E., Tassios, P. T., Rossolini, G. M., Gniadkowski, M., & Miriagou, V. (2011). Evolution and spread of a multidrug-resistant Proteus mirabilis clone with chromosomal AmpC-type cephalosporinases in Europe. Antimicrobial Agents and Chemotherapy, 55(6), 2735–2742.
17. de Been, M., Pinholt, M., Top, J., Bletz, S., Mellmann, A., van Schaik, W., Brouwer, E., Rogers, M., Kraat, Y., Bonten, M., Corander, J., Westh, H., Harmsen, D., & Willems, R. J. L. (2015). Core Genome Multilocus Sequence Typing Scheme for High-Resolution Typing of Enterococcus faecium. Journal of Clinical Microbiology, 53(12), 3788–3797.
18. Dingle, T. C., & MacCannell, D. R. (2015). Chapter 9—Molecular Strain Typing and Characterisation of Toxigenic Clostridium difficile. In A. Sails & Y.-W. Tang (Eds.), Methods in Microbiology (Vol. 42, pp. 329–357). Academic Press.
19. Eucast (2020). Intrinsic_Resistance_and_Unusual_Phenotypes_Tables_v3.2_20200225.pdf
20. Feldgarden, M., Brover, V., Haft, D. H., Prasad, A. B., Slotta, D. J., Tolstoy, I., Tyson, G. H., Zhao, S., Hsu, C.-H., McDermott, P. F., Tadesse, D. A., Morales, C., Simmons, M., Tillman, G., Wasilenko, J., Folster, J. P., & Klimke, W. (2019). Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. Antimicrobial Agents and Chemotherapy, 63(11).
21. Fouts, D. E., Brinkac, L., Beck, E., Inman, J., & Sutton, G. (2012). PanOCT: Automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Research, 40(22), e172.
22. Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics (Oxford, England), 28(23), 3150–3152.
23. Gajdács, M., & Urbán, E. (2019). Comparative Epidemiology and Resistance Trends of Proteae in Urinary Tract Infections of Inpatients and Outpatients: A 10-Year Retrospective Study. Antibiotics, 8(3).
24. Giammanco, G. M., Grimont, P. A. D., Grimont, F., Lefevre, M., Giammanco, G., & Pignato, S. (2011). Phylogenetic analysis of the genera Proteus, Morganella and Providencia by comparison of rpoB gene sequences of type and clinical strains suggests the reclassification of Proteus myxofaciens in a new genus, Cosenzaea gen. Nov., as Cosenzaea myxofaciens comb. Nov. International Journal of Systematic and Evolutionary Microbiology, 61(7), 1638–1644.
25. Girlich, D., Bonnin, R. A., Dortet, L., & Naas, T. (2020). Genetics of Acquired Antibiotic Resistance Genes in Proteus spp. Frontiers in Microbiology, 11.
26. Golemi-Kotra, D. (2008). Serratia, Edwardsiella and Morganella Infections. In S. J. Enna & D. B. Bylund (Eds.), xPharm: The Comprehensive Pharmacology Reference (pp. 1–6). Elsevier.
27. Harada, S., Ishii, Y., Saga, T., Tateda, K., & Yamaguchi, K. (2010). Chromosomally encoded blaCMY-2 located on a novel SXT/R391-related integrating conjugative element in a Proteus mirabilis clinical isolate. Antimicrobial Agents and Chemotherapy, 54(9), 3545–3550.
28. Hiller, N. L., Janto, B., Hogg, J. S., Boissy, R., Yu, S., Powell, E., Keefe, R., Ehrlich, N. E., Shen, K., Hayes, J., Barbadora, K., Klimke, W., Dernovoy, D., Tatusova, T., Parkhill, J., Bentley, S. D., Post, J. C., Ehrlich, G. D., & Hu, F. Z. (2007). Comparative Genomic Analyses of Seventeen Streptococcus pneumoniae Strains: Insights into the Pneumococcal Supragenome. Journal of Bacteriology, 189(22), 8186–8195.
29. Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by EggNOG-Mapper. Molecular Biology and Evolution, 34(8), 2115–2122.
30. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). EggNOG 5.0: A hierarchical,

functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research, 47(D1), D309–D314.
31. Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M., & Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics, 8(1), 460.
32. Huson, D. H., & Scornavacca, C. (2012). Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. Systematic Biology, 61(6), 1061–1067.
33. Innamorati, K. A., Earl, J. P., Aggarwal, S. D., Ehrlich, G. D., & Hiller, N. L. (2020). The Bacterial Guide to Designing a Diversified Gene Portfolio. In H. Tettelin & D. Medini (Eds.), The Pangenome: Diversity, Dynamics and Evolution of Genomes (pp. 51–87). Springer International Publishing.
34. Jaffuel, G., Imperiali, N., Shelby, K., Campos-Herrera, R., Geisert, R., Maurhofer, M., Loper, J., Keel, C., Turlings, T. C. J., & Hibbard, B. E. (2019). Protecting maize from rootworm damage with the combined application of arbuscular mycorrhizal fungi, Pseudomonas bacteria and entomopathogenic nematodes. Scientific Reports, 9.
35. Jain, S., Gaind, R., Kothari, C., Sehgal, R., Shamweel, A., Thukral, S. S., & Chellani, H. K. (2016). VEB-1 extended-spectrum β-lactamase-producing multidrug-resistant Proteus mirabilis sepsis outbreak in a neonatal intensive care unit in India: Clinical and diagnostic implications. JMM Case Reports, 3(4).
36. Jones, B. D., & Mobley, H. L. (1988). Proteus mirabilis urease: Genetic organization, regulation, and expression of structural genes. Journal of Bacteriology, 170(8), 3342.
37. Kim, Y., Koh, I., Young Lim, M., Chung, W.-H., & Rho, M. (2017). Pan-genome analysis of Bacillus for microbiome profiling. Scientific Reports, 7(1), 10984.
38. Laing, C., Buchanan, C., Taboada, E. N., Zhang, Y., Kropinski, A., Villegas, A., Thomas, J. E., & Gannon, V. P. (2010). Pan-genome sequence analysis using Panseq: An online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics, 11(1), 461.
39. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359.
40. Lukjancenko, O., Wassenaar, T. M., & Ussery, D. W. (2010). Comparison of 61 Sequenced Escherichia coli Genomes. Microbial Ecology, 60(4), 708–720.
41. Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., & Spratt, B. G. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences of the United States of America, 95(6), 3140–3145.
42. Maiden, M. C. J., Jansen van Rensburg, M. J., Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A., & McCarthy, N. D. (2013). MLST revisited: The gene-by-gene approach to bacterial genomics. Nature Reviews. Microbiology, 11(10), 728–736.
43. Manos, J., & Belas, R. (2006). The Genera Proteus, Providencia, and Morganella. In M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, & E. Stackebrandt (Eds.), The Prokaryotes: Volume 6: Proteobacteria: Gamma Subclass (pp. 245–269). Springer.
44. Martinson, V. G., Douglas, A. E., & Jaenike, J. (2017). Community structure of the gut microbiota in sympatric species of wild Drosophila. Ecology Letters, 20(5), 629–639.
45. McInerney, J. O., McNally, A., & O'Connell, M. J. (2017). Why prokaryotes have pangenomes. Nature Microbiology, 2(4), 1–5.
46. Medini, D., Donati, C., Rappuoli, R., & Tettelin, H. (2020). The Pangenome: A Data-Driven Discovery in Biology. In H. Tettelin & D. Medini (Eds.), The Pangenome: Diversity, Dynamics and Evolution of Genomes (pp. 3–20). Springer International Publishing.
47. Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. Current Opinion in Genetics & Development, 15(6), 589–594.
48. Méric, G., Yahara, K., Mageiros, L., Pascoe, B., Maiden, M. C. J., Jolley, K. A., & Sheppard, S. K. (2014). A Reference Pan-Genome Approach to Comparative Bacterial Genomics: Identification of Novel Epidemiological Markers in Pathogenic Campylobacter. PLOS ONE, 9(3), e92798.

49. Moura, A., Criscuolo, A., Pouseele, H., Maury, M. M., Leclercq, A., Tarr, C., Björkman, J. T., Dallman, T., Reimer, A., Enouf, V., Larsonneur, E., Carleton, H., Bracq-Dieye, H., Katz, L. S., Jones, L., Touchon, M., Tourdjman, M., Walker, M., Stroika, S., … Brisse, S. (2016). Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. Nature Microbiology, 2, 16185.
50. Μπάγκος, Π., 2015. *Βιοπληροφορική*. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: http://hdl.handle.net/11419/5016
51. Nagano, N., Shibata, N., Saitou, Y., Nagano, Y., & Arakawa, Y. (2003). Nosocomial Outbreak of Infections by Proteus mirabilis That Produces Extended-Spectrum CTX-M-2 Type β-Lactamase. Journal of Clinical Microbiology, 41(12), 5530–5536.
52. Nakano, R., Nakano, A., Abe, M., Inoue, M., & Okamoto, R. (2012). Regional outbreak of CTX-M-2 β-lactamase-producing Proteus mirabilis in Japan. Journal of Medical Microbiology, 61(12), 1727–1735.
53. Naum, M., Brown, E. W., & Mason-Gamer, R. J. (2008). Is 16S rDNA a Reliable Phylogenetic Marker to Characterize Relationships Below the Family Level in the Enterobacteriaceae? Journal of Molecular Evolution, 66(6), 630–642.
54. Neoh, H.-M., Tan, X.-E., Sapri, H. F., & Tan, T. L. (2019). Pulsed-field gel electrophoresis (PFGE): A review of the “gold standard” for bacteria typing and current alternatives. Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases, 74, 103935.
55. Nováková, E., Hypša, V., & Moran, N. A. (2009). Arsenophonus, an emerging clade of intracellular symbionts with a broad host distribution. BMC Microbiology, 9, 143.
56. Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics, 31(22), 3691–3693.
57. Pearce, M. E., Alikhan, N.-F., Dallman, T. J., Zhou, Z., Grant, K., & Maiden, M. C. J. (2018). Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak. International Journal of Food Microbiology, 274, 1–11.
58. Pearson, M. M., Sebaihia, M., Churcher, C., Quail, M. A., Seshasayee, A. S., Luscombe, N. M., Abdellah, Z., Arrosmith, C., Atkin, B., Chillingworth, T., Hauser, H., Jagels, K., Moule, S., Mungall, K., Norbertczak, H., Rabbinowitsch, E., Walker, D., Whithead, S., Thomson, N. R., … Mobley, H. L. T. (2008). Complete genome sequence of uropathogenic Proteus mirabilis, a master of both adherence and motility. Journal of Bacteriology, 190(11), 4027–4037.
59. Potron, A., Hocquet, D., Triponney, P., Plésiat, P., Bertrand, X., & Valot, B. (2019). Carbapenem-Susceptible OXA-23-Producing Proteus mirabilis in the French Community. Antimicrobial Agents and Chemotherapy, 63(6).
60. Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2—Approximately maximum-likelihood trees for large alignments. PloS One, 5(3), e9490.
61. Protonotariou, E., Poulou, A., Politi, L., Meletis, G., Chatzopoulou, F., Malousi, A., Metallidis, S., Tsakris, A., & Skoura, L. (2020). Clonal outbreak caused by VIM-4-producing Proteus mirabilis in a Greek tertiary-care hospital. International Journal of Antimicrobial Agents, 56(2), 106060.
62. Rasko, D. A., Rosovitz, M. J., Myers, G. S. A., Mongodin, E. F., Fricke, W. F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N. R., Chaudhuri, R., Henderson, I. R., Sperandio, V., & Ravel, J. (2008). The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates. Journal of Bacteriology, 190(20), 6881–6893.
63. Reimer, A. R., Van Domselaar, G., Stroika, S., Walker, M., Kent, H., Tarr, C., Talkington, D., Rowe, L., Olsen-Rasmussen, M., Frace, M., Sammons, S., Dahourou, G. A., Boncy, J., Smith, A. M., Mabon, P., Petkau, A., Graham, M., Gilmour, M. W., & Gerner-Smidt, P. (2011). Comparative Genomics of Vibrio cholerae from Haiti, Asia, and Africa. Emerging Infectious Diseases, 17(11), 2113–2121.
64. Rouli, L., Merhej, V., Fournier, P.-E., & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes and New Infections, 7, 72–85.

65. Roy, S. S., Dasgupta, R., & Bagchi, A. (2014). A Review on Phylogenetic Analysis: A Journey through Modern Era. Computational Molecular Bioscience, 4(3), 39–45.
66. Sahl, J. W., Caporaso, J. G., Rasko, D. A., & Keim, P. (2014). The large-scale blast score ratio (LS-BSR) pipeline: A method to rapidly compare genetic content between bacterial genomes. PeerJ, 2, e332.
67. Schaffer, J. N., & Pearson, M. M. (2015). Proteus mirabilis and Urinary Tract Infections. Microbiology Spectrum, 3(5).
68. Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. Bioinformatics (Oxford, England), 30(14), 2068–2069.
69. Sheppard, S. K., Colles, F. M., McCARTHY, N. D., Strachan, N. J. C., Ogden, I. D., Forbes, K. J., Dallas, J. F., & Maiden, M. C. J. (2011). Niche segregation and genetic structure of Campylobacter jejuni populations from wild and agricultural host species. Molecular Ecology, 20(16), 3484–3490.
70. Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. Protein Science: A Publication of the Protein Society, 27(1), 135–145.
71. Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England), 30(9), 1312–1313.
72. Stock, I. (2003). Natural Antibiotic Susceptibility of Proteus spp., with Special Reference to P. mirabilis and P. penneri Strains. Journal of Chemotherapy, 15(1), 12–26.
73. Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: A tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research, 28(1), 33–36.
74. Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Ros, I. M. y, Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome.” Proceedings of the National Academy of Sciences of the United States of America, 102(39), 13950.
75. Tettelin, H., & Medini, D. (Eds.). (2020). The Pangenome: Diversity, Dynamics and Evolution of Genomes. Springer International Publishing.
76. Tiessen, A., Pérez-Rodríguez, P., & Delaye-Arredondo, L. J. (2012). Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. BMC Research Notes, 5(1), 85.
77. Townsend, J. P., Su, Z., & Tekle, Y. I. (2012). Phylogenetic Signal and Noise: Predicting the Power of a Data Set to Resolve Phylogeny. Systematic Biology, 61(5), 835–835.
78. Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biology, 15.
79. van Vliet, A. H. M. (2017). Use of pan-genome analysis for the identification of lineage-specific genes of Helicobacter pylori. FEMS Microbiology Letters, 364(fnw296).
80. Venditti, C., Vulcano, A., D’Arezzo, S., Gruber, C. E. M., Selleri, M., Antonini, M., Lanini, S., Marani, A., Puro, V., Nisii, C., & Di Caro, A. (2019). Epidemiological investigation of an Acinetobacter baumannii outbreak using core genome multilocus sequence typing. Journal of Global Antimicrobial Resistance, 17, 245–249.
81. Visalli, M. A., Murphy, E., Projan, S. J., & Bradford, P. A. (2003). AcrAB Multidrug Efflux Pump Is Associated with Reduced Levels of Susceptibility to Tigecycline (GAR-936) in Proteus mirabilis. Antimicrobial Agents and Chemotherapy, 47(2), 665–669.
82. Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., Moore, L. H., Moore, W. E. C., Murray, R. G. E., Stackebrandt, E., Starr, M. P., & Truper, H. G. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. International Journal of Systematic and Evolutionary Microbiology, 37(4), 463–464.

83. Xiao, J., Zhang, Z., Wu, J., & Yu, J. (2015). A Brief Review of Software Tools for Pangenomics. Genomics, Proteomics & Bioinformatics, 13(1), 73–76.
84. Yang, X., Li, Y., Zang, J., Li, Y., Bie, P., Lu, Y., & Wu, Q. (2016). Analysis of pan-genome to identify the core genes and essential genes of Brucella spp. Molecular Genetics and Genomics: MGG, 291(2), 905–912.
85. Zhao, Y., Wu, J., Yang, J., Sun, S., Xiao, J., & Yu, J. (2012). PGAP: Pan-genomes analysis pipeline. Bioinformatics, 28(3), 416–418.