ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΧΗΜΙΚΩΝ ΜΗΧΑΝΙΚΩΝ

Τομέας IV : Σύνθεσης και Ανάπτυξης Βιομηχανικών Διαδικασιών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΔΗΜΙΟΥΡΓΙΑ ΡΟΗΣ ΔΙΕΡΓΑΣΙΩΝ ΣΤΗΝ ΥΠΟΛΟΓΙΣΤΙΚΗ ΠΛΑΤΦΟΡΜΑ GALAXY ΓΙΑ ΤΗΝ ΤΑΞΙΝΟΜΙΚΗ ΚΑΙ ΛΕΙΤΟΥΡΓΙΚΗ ΑΝΑΛΥΣΗ ΜΕΤΑΓΟΝΙΔΙΩΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

ΑΓΙΟΥΤΑΝΤΗΣ ΠΑΝΑΓΙΩΤΗΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:

ΦΡΑΓΚΙΣΚΟΣ ΚΟΛΙΣΗΣ

ΑΘΗΝΑ 2014

Ευχαριστίες

Η πραγματοποίηση και η παράδοση αυτής της διπλωματικής εργασίας αποτελεί το επιστέγασμα της φοιτητικής μου πορείας και προσπάθειας ως υποψήφιος Χημικός Μηχανικός σε προπτυχιακό επίπεδο.

Θα ήθελα να εκφράσω την ευγνωμοσύνη και τις θερμές μου ευχαριστίες προς τον Καθηγητή Φραγκίσκο Κολίση για την ευκαιρία που μου έδωσε να ασχοληθώ με το συγκεκριμένο θέμα, καθώς και για τις πολύτιμες συμβουλές που μοιράστηκε μαζί μου τόσο σε επίπεδο πραγματοποίησης αυτής της διπλωματικής εργασίας, όσο και σε γενικότερα πλαίσια ακαδημαϊκής και ερευνητικής φύσης.

Φυσικά δεν θα μπορούσα να παραλείψω να ευχαριστήσω τον υποψήφιο διδάκτορα Ευθύμιο Λαδουκάκη, δίχως την συμβολή, τις γνώσεις αλλά και την καθοδήγηση του οποίου η πραγματοποίηση αυτής της διπλωματικής εργασίας θα ήταν αδύνατη. Τον ευχαριστώ για την υπομονή που έδειξε αλλά και την στήριξη που μου παρείχε ώστε να μπορέσω να κάνω τα πρώτα μου βήματα στον - μέχρι πρότινος άγνωστο για μένα - τομέα της Βιοπληροφορικής.

Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου για την αγάπη και την στήριξη τους όλα αυτά τα χρόνια. Ελπίζω και εύχομαι να τους δικαιώσω στην πορεία της ζωής μου με όποιον τρόπο μπορώ.

Περίληψη

Η μεταγονιδιωματική είναι μια ανερχόμενη επιστήμη των μικροβιακών συστημάτων που βασίζεται σε μια μεγάλης κλίμακας ανάλυση του γενετικού υλικού των μικροβιακών κοινοτήτων στο φυσικό τους περιβάλλον. Περιλαμβάνει την ανάλυση του γονιδιώματος ολόκληρων κοινοτήτων μικροοργανισμών, που συνδέονται μεταξύ τους με περίπλοκες αλληλεπιδράσεις, σε διαφορετικές οικολογικές και περιβαλλοντικές συνθήκες και καταστάσεις. Οι εφαρμογές των δεδομένων που προκύπτουν από μεταγονιδιωματικές αναλύσεις είναι ποικίλες και καλύπτουν ένα ευρύ βιομηχανικό και ερευνητικό φάσμα. Τα μεταγονιδιωματικά προγράμματα και αναλύσεις έχουν αναπτυχθεί την τελευταία δεκαετία σε ιδιαίτερα σημαντικό βαθμό, ακολουθώντας τις εξελίξεις σε θέματα που άπτονται του τομέα της Βιοπληροφορικής και της ανάλυσης αλληλουχιών. Η ανάπτυξη νέων τεχνικών αλληλούχισης, συναρμολόγησης και επεξεργασίας δειγμάτων έχει οδηγήσει σε μια αλματώδη ανάπτυξη των γονιδιωματικών και κατά συνέπεια και των μεταγονιδιωματικών αναλύσεων, με απώτερο στόχο την απάντηση σε θεμελιώδη ερωτήματα που αφορούν τον επιστημονικό αυτό τομέα, όπως το ποιες ομάδες μικροοργανισμών εντοπίζονται σε δείγματα από συγκεκριμένα περιβάλλοντα και ταυτόχρονα ποιες λειτουργικές διεργασίες είναι δυνατό να λάβουν χώρα στα πλαίσια μιας συγκεκριμένης περιβαλλοντικής κοινότητας. Προς αυτή την κατεύθυνση και με σκοπό την ταξινομική κατηγοριοποίηση και τον λειτουργικό χαρακτηρισμό μεταγονιδιωματικών δεδομένων, στα πλαίσια της εργασίας αυτής, επιχειρήθηκε η δημιουργία υπολογιστικών εργαλείων ταξινομικής και λειτουργικής ανάλυσης στην πλατφόρμα GALAXY, μια πλατφόρμα που φιλοξενεί μια πληθώρα εργαλείων Βιοπληροφορικής, προσφέροντας άμεσα και εύχρηστα περιβάλλοντα διεπαφής ώστε ο κάθε ερευνητής να δύναται να πραγματοποιήσει έναν μεγάλο αριθμό αναλύσεων. Δημιουργήθηκαν τρία διαφορετικά υπολογιστικά εργαλεία που συνδέθηκαν μεταξύ τους σε μια ροή διεργασιών και τα οποία εκμεταλλεύονται τις λειτουργίες δύο βασικών εργαλείων Βιοπληροφορικής. Αρχικά επιχειρείται μια σύγκριση αλληλουχιών σε βάσεις δεδομένων μέσω της εφαρμογής του BLAST και στην συνέχεια παρέχεται μια πρώτη εικόνα σε ότι αφορά το ταξινομικό και λειτουργικό περιεχόμενο των μεταγονιδιωματικών δεδομένων, μέσω των δυνατοτήτων του εργαλείου MEGAN, ενός εργαλείου ταξινομικής, λειτουργικής και συγκριτικής ανάλυσης μεταγονιδιωματικών δειγμάτων. Τέλος πραγματοποιείται έλεγχος της λειτουργίας των υπολογιστικών αυτών εργαλείων μέσω της εφαρμογής τους σε πραγματικά μεταγονιδιωματικά δεδομένα που προέρχονται από δείγματα τριών διαφορετικών ειδών περιβάλλοντος, παρουσιάζονται τα αποτελέσματα που προέκυψαν και αξιολογείται η χρησιμότητα και η λειτουργικότητα τους στα πλαίσια μιας μεταγονιδιωματικής ανάλυσης.

Abstract

Metagenomics is an emerging science of microbial systems, based on a large-scale analysis of the genetic material of microbial communities in their natural environments. Metagenomics involve the genome analysis of entire microbial communities, which are connected through complex interactions, on different ecological and environmental conditions. The applications of data derived from a metagenomic analysis vary, and cover a wide range of industrial and research projects. The metagenomic analyzes have widely developed during the last decade, following the development on issues related to the fields of Bioinformatics and sequencing. The development of new sequencing and assembly techniques, as well as new approaches related to the processing of samples has led to a rapid development of genomic and consequently metagenomic analyzes, aiming to answer fundamental questions concerning this scientific field, such as which groups of microorganisms can be detected in samples from specific environments, while monitoring functional processes that may occur in the context of a specific environmental community. To this end, aiming to the taxonomic binning and functional annotation of metagenomic data, in the context of this diploma thesis, an attempt was made to create computational tools for taxonomic and functional analysis in the GALAXY platform, a platform that hosts a plethora of Bioinformatics tools, providing direct and easy interfaces, so every researcher is able to perform a large number of different analyzes. Three different computational tools were created and subsequently linked to produce a workflow that exploits main features of two basic Bioinformatics tools. Initially, we compared and aligned the metagenomic sequences against reference databases through the implementation of BLAST and then provided a first insight in terms of the taxonomic and functional content of specific metagenomic data through the implementation of the tool MEGAN, a tool for taxonomic, functional and comparative analysis of metagenomic samples. Finally, we tested the functionality of these computational tools through an application in real metagenomic data derived from samples extracted from three different types of environments and presented the results that we obtained. Also, we evaluated the usefulness and functionality of these tools within a metagenomic analysis.

Keywords : metagenomics, taxonomic binning, functional annotation, GALAXY, BLAST, MEGAN

ΠΕΡΙΕΧΟΜΕΝΑ

ΘΕΩΡΗΤΙΚΟ ΜΕΡΟΣ

1. ΜΕΤΑΓΟΝΙΔΙΩΜΑΤΙΚΗ

1.1 Ορισμός

Η μεταγονιδιωματική (metagenomics) είναι μια ανερχόμενη επιστήμη των μικροβιακών συστημάτων, που βασίζεται σε μια μεγάλης κλίμακας ανάλυση του DNA των μικροβιακών κοινοτήτων στο φυσικό τους περιβάλλον. Οι μελέτες των μεταγονιδιωμάτων αποκαλύπτουν το τεράστιο πεδίο της βιοποικιλότητας σε ένα ευρύ φάσμα ειδών περιβάλλοντος, καθώς επίσης και νέες λειτουργικές δυνατότητες μεμονωμένων κυττάρων και κοινοτήτων αλλά και τις πολύπλοκες εξελικτικές σχέσεις μεταξύ τους [1].

Η μεταγονιδιωματική - ή όπως αλλιώς ονομάζεται περιβαλλοντική γονιδιωματική, γονιδιωματική των κοινοτήτων, οικολογική γονιδιωματική ή γονιδιωματική των μικροβιακών πληθυσμών - περιλαμβάνει την ανάλυση του γονιδιώματος ολόκληρων κοινοτήτων μικροοργανισμών που συνδέονται μεταξύ τους με περίπλοκες αλληλεπιδράσεις, σε διαφορετικές οικολογικές και περιβαλλοντικές συνθήκες και καταστάσεις [1], δίχως την δημιουργία καλλιεργειών. Ουσιαστικά πρόκειται για γονιδιωματική ανάλυση των μικροβιακών πληθυσμών και περιλαμβάνει την κατευθείαν εξαγωγή DNA από ένα περιβαλλοντικό δείγμα - παραδείγματος χάριν από θαλάσσια ύδατα, το έδαφος, το ανθρώπινο πεπτικό σύστημα - και εν συνεχεία την μελέτη αυτού του δείγματος γενετικού υλικού[2]. Ο όρος μεταγονιδίωμα (metagenome) ορίστηκε για πρώτη φορά ως το γονιδίωμα του συνόλου των μικροοργανισμών ενός συγκεκριμένου περιβάλλοντος από τον Jo Handelsman και τους συνεργάτες του. Το μεταγονιδιωματικό DNA είναι ιδιαίτερα πολύπλοκο καθώς, ουσιαστικά, αποτελεί ένα πλήθος γονιδιωμάτων που προέρχονται από πολλούς διαφορετικούς οργανισμούς, γεγονός που καθιστά την ανάλυση του ιδιαίτερα περίπλοκη και απαιτητική.

1.2 Ιστορική αναδρομή

Η μικροβιολογία ήταν παραδοσιακά ένα πειθαρχικό όριο για την επικρατούσα κλασσική βιολογία των μακροοργανισμών. Τα ερευνητικά ενδιαφέροντα της, εισήλθαν στο κύριο ρεύμα της βιολογικής σκέψης μέσω της γενετικής και της μοριακής βιολογίας τη δεκαετία του 1940, όταν οι περισσότερες μοριακές και γενετικές αναλύσεις ήταν "δεμένες" σε μικροοργανισμούς για τεχνικούς λόγους. Η χρήση των μικροβίων και ειδικά των ιών και των βακτηρίων, ως εργαλεία για την κατανόηση της γενετικής κληρονομιάς, βασίστηκε στη πεποίθηση αλλά και την τελική απόδειξη ότι τα μικρόβια κατέχουν όχι μόνο το ίδιο γενετικό υλικό που κατέχουν και οι πολυκύτταροι οργανισμοί, αλλά εκτελούν επίσης και πολλές παρόμοιες βιολογικές διεργασίες, συμπεριλαμβανομένων και των

αναπαραγωγικών. Καθώς οι μελέτες ενός μόνο γονιδίου έδωσαν την θέση τους σε εκείνες που αφορούν ολόκληρο το γονιδίωμα στη δεκαετία του 1990, η σημασία των μικροβίων και της μικροβιακή γνώσης για τη μοριακή βιολογία ενισχύθηκε περαιτέρω, ιδιαίτερα καθώς η αλληλούχιση του γενετικού υλικού των προκαρυωτικών οργανισμών και των ιών αναπτύχθηκε ταχύτατα πέραν της αλληλούχισης των μεγάλων οργανισμών. Οι μελέτες του γονιδιώματος επιμέρους απομονωμένων μικροβίων, ενώ προσθέτουν τεράστιες ποσότητες νέων πληροφοριών και δυνατότητα κατανόησης, ιδίως στη συγκριτική εξελικτική βιολογία, παρέχουν μέχρι στιγμής περιορισμένες γνώσεις σχετικά με τη λειτουργία των πολλών εκατομμυρίων ακαλλιέργητων μικροβιακών ταξινομικών ομάδων.

Οι προσπάθειες για την αντιμετώπιση αυτών των ελλείψεων έχουν οδηγήσει στην διεύρυνση του πεδίου της μικροβιολογίας του μοριακού του περιβάλλοντος. Οι οικολογικές μικροβιακές μελέτες έχουν ιστορικά χαρακτηριστεί ως περιφερικές τόσο ως προς τη γενική οικολογία όσο και ως προς την κλασσική μικροβιολογία, ως προς την πρώτη κυρίως λόγω των γενικότερων τάσεων στη βιολογία και ως προς την δεύτερη, λόγω της επικράτησης του παραδείγματος της καθαρής καλλιέργειας . Αν και η μικροβιακή οικολογία είχε ρίζες στα τέλη του δέκατου ένατου αιώνα στο έργο του Ρώσου μικροβιολόγου εδάφους, S.Winogradsky και του ιδρυτή της διάσημης σχολής Delft της μικροβιολογίας, Beijerinck M W, χρειάστηκε να φτάσουν τα τέλη του 1960 ώστε η μικροβιακή οικολογία να αποκτήσει πραγματική αναγνώριση. Το κυρίαρχο ενοποιητικό θέμα της, ανεξαρτήτως των μεθόδων που χρησιμοποιούνται, είναι ότι οι μικροοργανισμοί πρέπει να γίνουν κατανοητοί μέσα στα οικολογικά τους πλαίσια (τα οποία συμπεριλαμβάνουν την ύπαρξη και άλλων οργανισμών), και όχι ως μεμονωμένες επιμέρους ομάδες μέσα σε τεχνητά περιβάλλοντα. Τα περισσότερα μικρόβια ζουν κατά προτίμηση σε πολύπλοκες, συχνά πολλών ειδών, κοινότητες, όπως τα βιοφίλμ, ενώ έως και το 99% των προκαρυωτικών ταξινομικών ομάδων δεν είναι σήμερα δυνατόν να καλλιεργηθούν σε τεχνητές εργαστηριακές συνθήκες [1].

Λαμβάνοντας υπ' όψιν μια περιβαλλοντική προσέγγιση, δόθηκε η δυνατότητα στη μικροβιακή γονιδιωματική να επεκταθεί πέραν της αλληλούχισης εργαστηριακών καλλιεργειών απομονωμένων μικροοργανισμών, στον προσδιορισμό της αλληλουχίας DNA που εξάγεται απευθείας από το φυσικό περιβάλλον. Αυτή η μετακίνηση έξω από το εργαστήριο οδήγησε στην σημαντική επέκταση του πεδίου των δεδομένων που συλλέγονται, καθώς και της κατανόησης της βιοποικιλότητας και των εξελικτικών σχέσεων. Χρειάστηκε το υπόλοιπο της δεκαετίας του 1990 ωστόσο, ώστε ξεπεραστεί μια άλλη συσχετιζόμενη προσεγγιστική δυσκολία, κατά την οποία, περιορισμένες DNA αλληλουχίες (συχνά μη-κωδικές ως προς τη δημιουργία πρωτεϊνών) χρησιμοποιούνταν ως δείκτες των ειδών αλλά και της βιοποικιλότητας. Η εστίαση σε συγκεκριμένα γονίδια και η υπόθεση ότι αυτά θα είναι ίδια σε κάθε είδος, ανεξάρτητα από το περιβάλλον, περιόριζε τις πληροφορίες

που λαμβάνονται σχετικά με τα φυσιολογικά ή οικολογικά χαρακτηριστικά των οργανισμών[1].

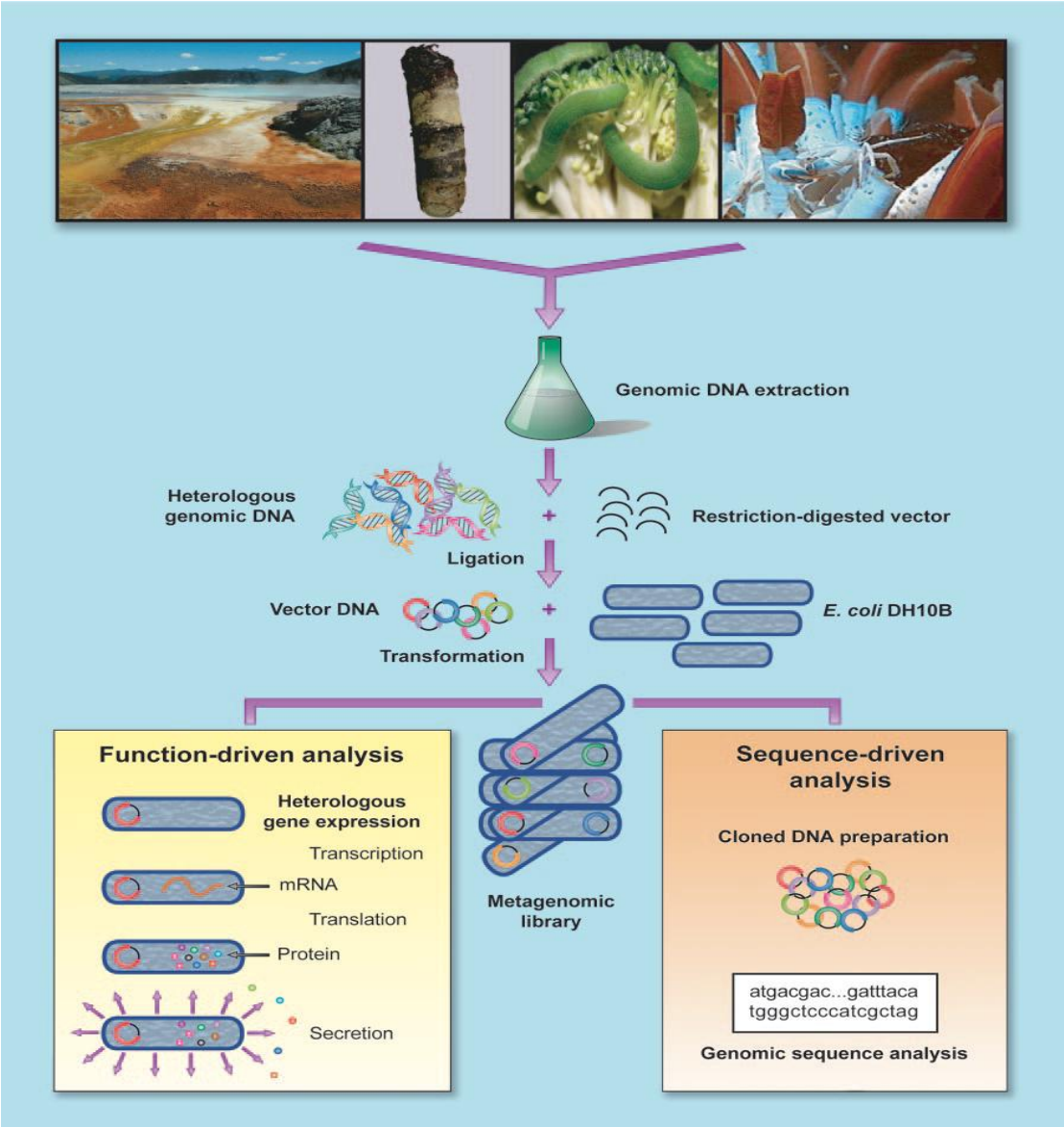
Η μεταγονιδιωματική γίνεται πια αντιληπτή ως μια προσέγγιση μέσω της οποίας είναι δυνατό να ξεπεραστούν τέτοιοι περιορισμοί. Αντί των μεμονωμένων γονιδιωμάτων (monogenomes) ή μονογονιδιακών δεικτών, η μεταγονιδιωματική ξεκινά με μεγάλες ποσότητες DNA που συλλέγονται από μικροβιακές κοινότητες στο φυσικό τους περιβάλλον, προκειμένου να διερευνηθεί η βιοποικιλότητα, οι λειτουργικές αλληλεπιδράσεις και οι εξελικτικές σχέσεις.

1.3 Προσεγγίσεις στην μεταγονιδιωματική ανάλυση

Η μεταγονιδιωματική εισήχθηκε στα επιστημονικά δρώμενα την δεκαετία του 1990 θέτοντας πολλαπλούς στόχους. Πάνω από όλα ήταν μια μέθοδος κατανόησης της συνεισφοράς στην βιόσφαιρα των στελεχών των κοινοτήτων για τα οποία δεν ήταν δυνατό να δημιουργηθούν καλλιέργειες στο εργαστήριο. Επιπλέον, σχεδιάστηκε με στόχο πρακτικά οφέλη όπως την ανακάλυψη νέων γονιδίων και παραγώγων αυτών που θα οδηγήσουν σε μια νέα μορφή φαρμακευτικής χημείας, σε αγροτικές καινοτομίες και νέες βιομηχανικές διεργασίες. Αυτές οι δύο κύριες κατηγοριοποιήσεις των στόχων οδήγησαν στην εξέλιξη δύο παράλληλων μεθοδολογικών προσεγγίσεων, της ανάλυσης που βασίζεται στην αλληλούχιση του DNA (sequence-based analysis) και της λειτουργικής ανάλυσης (function-based analysis)[2].

Η μεταγονιδιωματική ανάλυση περιλαμβάνει την απομόνωση του DNA από ένα περιβαλλοντικό δείγμα, την κλωνοποίηση του DNA σε έναν κατάλληλο φορέα, τον μετασχηματισμό των κλώνων σε ένα βακτήριο υποδοχέα και την παρακολούθηση και διαλογή των μετασχηματισμένων κλώνων[3], μια διαδικασία γνωστή ως δημιουργία μεταγονιδιωματικών βιβλιοθηκών (Εικόνα 1). Θεωρητικά, μία μεταγονιδιωματική βιβλιοθήκη περιέχει κλώνους που αντιπροσωπεύουν ολόκληρο το γενετικό συμπλήρωμα ενός απλού βιότοπου, αν και αυτό εξαρτάται και από την απόδοση της εξαγωγής του DNA και τις μεθόδους κλωνοποίησης. Οι πληροφορίες που περιέχονται μέσα σε μια μεταγονιδιωματική βιβλιοθήκη μπορούν να χρησιμοποιηθούν για τον προσδιορισμό της πολυμορφίας της κοινότητας και της μικροβιακής δραστηριότητας, την παρουσία συγκεκριμένων μικροοργανισμών ή βιοσυνθετικών μονοπατιών, καθώς και την αναζήτηση για την παρουσία μεμονωμένων γονιδίων[4]. Οι κλώνοι είναι δυνατό να ελεγχθούν για φυλογενετικούς δείκτες ή "anchors" όπως το 16SrRNA και το recA, ή για άλλα διατηρημένα γονίδια μέσω είτε υβριδισμού είτε πολλαπλής PCR, ή ακόμα ως προς την έκφραση συγκεκριμένων χαρακτηριστικών όπως ενζυμική δραστηριότητα ή παραγωγή αντιβιοτικών ουσιών. Επίσης είναι δυνατή η αλληλούχιση των κλώνων κατά τρόπο τυχαίο. Κάθε προσέγγιση έχει τα πλεονεκτήματα αλλά και τους περιορισμούς της ενώ από όλες αυτές τις προσεγγίσεις η κατανόηση του μη

καλλιεργήσιμου μικροβιακού κόσμου έχει εμπλουτιστεί, προσφέροντας μια εσωτερική ματιά σε ομάδες προκαρυωτικών οργανισμών που διαφορετικά θα ήταν τελείως άγνωστοι[3].



Εικόνα 1 Διάγραμμα ροής των βημάτων κατασκευής βιβλιοθηκών μεταγονιδιωματικού DNA. Το DNA απομονώνεται από τα κύτταρα του δείγματος και στη συνέχεια κατακερματίζεται, εισάγεται σε φορείς και κλωνοποιείται . Οι φορείς εισάγονται σε κύτταρα-ξενιστές. Μετά την δημιουργία της μεταγονιδιωματικής βιβλιοθήκης, τα γονιδιώματα υφίστανται είτε λειτουργική ανάλυση είτε ανάλυση αλληλούχισης[3].

Η **ανάλυση αλληλούχισης** βασίζεται στη χρήση συντηρημένων DNA αλληλουχιών για το σχεδιασμό ιχνηλατών υβριδισμού ή εκκινητών PCR για τη μελέτη των μεταγονιδιωματικών βιβλιοθηκών για κλώνους που περιέχουν αλληλουχίες που ενδιαφέρουν[5]. Μπορεί να περιλαμβάνει πλήρη αλληλούχιση κλώνων που

περιέχουν φυλογενετικούς δείκτες, που υποδηλώνουν την ταξινομική ομάδα από την οποία είναι πιθανό να προέρχεται ένα τμήμα DNA ή εναλλακτικά, μπορεί να διεξαχθεί τυχαία αλληλούχιση και μόλις εντοπιστεί ένα ζητούμενο ή ενδιαφέρον γονίδιο, θα αναζητηθούν φυλογενετικοί δείκτες (anchors) στο γειτονικό DNA ώστε να διαμορφωθεί ένας φυλογενετικός σύνδεσμος με το λειτουργικό γονίδιο. Η ανάλυση της αλληλουχίας που καθοδηγείται από τον προσδιορισμό των φυλογενετικών δεικτών είναι μια ισχυρή προσέγγιση η οποία προτάθηκε για πρώτη φορά από την ομάδα DeLong, η οποία παρήγαγε την πρώτη γονιδιωματική αλληλουχία που συνδέεται με ένα γονίδιο 16S rRNA ενός ακαλλιέργητου μικροοργανισμού της κατηγορίας των Αρχαίων. Στη συνέχεια, εντόπισαν ένα ένθετο τμήμα DNA από βακτήρια του θαλασσινού νερού που περιείχε ένα γονίδιο 16S rRNA που σχετίζεται με γ-πρωτεοβακτήρια . Η αλληλουχία του παρακείμενου DNA αποκάλυψε ένα γονίδιο όμοιο με αυτά που σχετίζονται με την παραγωγή βακτηριοροδοψίνης. Το γονιδιακό προϊόν του, αποδείχθηκε ότι είναι ένας αυθεντικός φωτοϋποδοχέας, γεγονός που οδήγησε στην κατανόηση του γεγονότος ότι η παρουσία των γονιδίων παραγωγής βακτηριοροδοψίνης δεν περιορίζεται στα Αρχαία αλλά είναι στην πραγματικότητα εν αφθονία, μεταξύ των πρωτεοβακτηρίων των ωκεανών[3]. Εν συνεχεία, η ετερόλογη έκφραση του γονιδίου παραγωγής της βακτηριοροδοψίνης σε E. coli έδωσε έναν λειτουργικό βιοχημικό χαρακτηρισμό της πρωτεΐνης, ολοκληρώνοντας το πλήρες φάσμα των μελετών που συνδέουν την φυλογένεση με την λειτουργία[5].

Μια πολλά υποσχόμενη εφαρμογή της φυλογενετικής αλληλούχισης που βασίζεται στους δείκτες, είναι η συλλογή και η αλληλούχιση πολλών γονιδιακών θραυσμάτων από μια ταξινομική ομάδα. Σε πιο πολύπλοκα περιβάλλοντα και ταξινομικές ομάδες η επανασυναρμολόγηση του γονιδιώματος μπορεί να μην είναι εφικτή, αλλά συμπεράσματα σχετικά με τη φυσιολογία και την οικολογία των μελών των ομάδων είναι δυνατό να αντληθούν από τα δεδομένα της αλληλούχισης. Αυτή η προσέγγιση έχει ξεκινήσει με κλώνους από διαφορετικά εδάφη που περιέχουν 16S rRNA γονίδια που συνδέονται με το φύλο των Acidobacteria, τα οποία είναι άφθονα στο έδαφος και ιδιαιτέρως πολυποίκιλα και για τα οποία οι γνώσεις είναι περιορισμένες. Η πλήρης αλληλούχιση των εκτιμώμενων περίπου 500 kb του Acidobacterium DNA στις μεταγονιδιωματικές βιβλιοθήκες μπορεί να παρέχει πληροφορίες για τις υποομάδες των βακτηρίων σε αυτό το φύλο, οι οποίες δεν έχουν καλλιεργηθεί ποτέ [3]. Μια παρόμοια ανάλυση έχει ξεκινήσει για τα Crenarchaeota του εδάφους, χρησιμοποιώντας 16S rRNA γονίδια για τον εντοπισμό κλώνων που προέρχονται από Αρχαία (Archaea) και εφαρμόζοντας αλληλούχιση του γονιδιώματος για την κατάρτιση πληροφοριών σχετικά με αυτά[5].

Η εναλλακτική λύση για μια φυλογενετική προσέγγιση βασιζόμενη σε δείκτες είναι η αλληλούχιση τυχαίων κλώνων, η οποία έχει δώσει αξιοσημείωτες ιδέες, ειδικά όταν διεξάγεται σε μαζική κλίμακα. Η κατανομή και η υπεραριθμία των

λειτουργιών σε μια κοινότητα, η σύνδεση διαφόρων χαρακτηριστικών, η οργάνωση του γονιδιώματος, και η οριζόντια μεταφορά γονιδίων ,μπορούν να συναχθούν από την ανάλυση αλληλούχισης. Οι πρόσφατες μνημειώδεις προσπάθειες αλληλούχισης οι οποίες περιλαμβάνουν ανασυγκρότηση του γονιδιώματος των ακαλλιέργητων οργανισμών μιας κοινότητας που προέρχεται από όξινες απορροές ορυχείων αλλά και το πρόγραμμα Sargasso Sea, απεικονίζουν τη δύναμη των προσπαθειών αλληλούχισης μεγάλης κλίμακας ώστε να διευρυνθεί η κατανόηση των ακαλλιέργητων κοινοτήτων οργανισμών. Οι μελέτες αυτές έχουν οδηγήσει σε νέες διασυνδέσεις μεταξύ της φυλογένεσης και της λειτουργίας, έχουν υποδείξει την εκπληκτική αφθονία ορισμένων τύπων γονιδίων και έχουν δώσει την ευκαιρία ανακατασκευής των γονιδιωμάτων των οργανισμών που δεν έχουν καλλιεργηθεί.

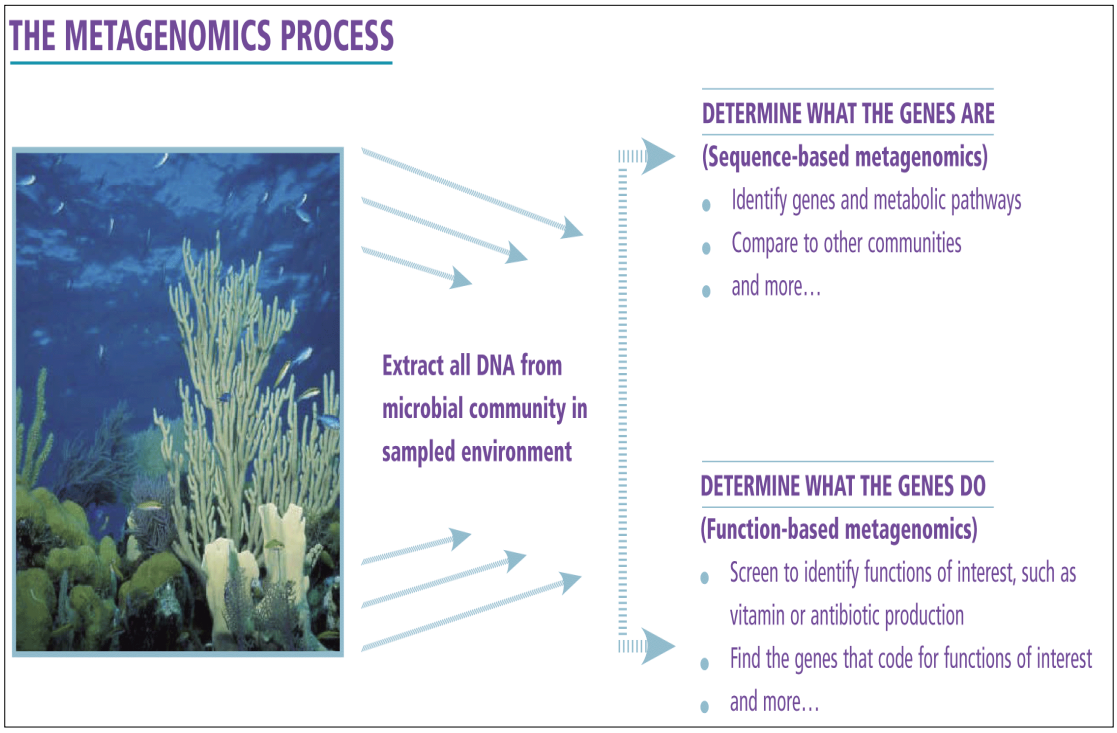
Η χρήση των φυλογενετικών δεικτών είτε ως αρχικά αναγνωριστικά εργαλεία των θραυσμάτων του DNA για τη μελέτη αυτών, είτε ως δείκτες της ταξινομικής υπαγωγής των τμημάτων DNA που φέρουν γονίδια που παρουσιάζουν ενδιαφέρον λόγω της λειτουργίας τους, περιορίζεται από το μικρό αριθμό των διαθέσιμων δεικτών που οδηγούν σε αξιόπιστη τοποθέτηση στο δέντρο της ζωής. Εάν ένα θραύσμα DNA που παρουσιάζει ενδιαφέρον για διάφορους λόγους δεν φέρει ένα αξιόπιστο δείκτη, ο οργανισμός της προέλευσης του παραμένει άγνωστος. Η συλλογή των φυλογενετικών δεικτών αυξάνεται συνεχώς και όσο η ποικιλία των δεικτών αυξάνει τόσο ενισχύεται και η δύναμη της εν λόγω προσέγγισης, καθιστώντας δυνατή την αντιστοίχιση περισσότερων τμημάτων άγνωστου DNA με τους οργανισμούς από τους οποίους έχουν απομονωθεί. Επιπλέον, καθώς ανασυναρμολογούνται όλο και περισσότερα γονιδιώματα, όλο και περισσότερα γονίδια συνδέονται με φυλογενετικούς δείκτες ακόμα και αν δεν είχαν κλωνοποιηθεί αρχικά στο ίδιο τμήμα DNA[3].

Η **λειτουργική ανάλυση** ξεκινά με ταυτοποίηση των κλώνων που εκφράζουν ένα επιθυμητό γνώρισμα, ακολουθούμενη από τον χαρακτηρισμό των ενεργών κλώνων μέσω ανάλυσης αλληλουχίας αλλά και βιοχημικής ανάλυσης. Αυτή η προσέγγιση εντοπίζει ταχύτατα κλώνους που έχουν πιθανές εφαρμογές στην ιατρική, τη γεωργία ή τη βιομηχανία, εστιάζοντας στα φυσικά προϊόντα ή τις πρωτεΐνες που παρουσιάζουν χρήσιμα χαρακτηριστικά. Οι περιορισμοί της προσέγγισης αυτής εντοπίζονται στο ότι απαιτεί την έκφραση της ζητούμενης λειτουργίας στο κύτταρο ξενιστή και την ομαδοποίηση όλων των γονιδίων που απαιτούνται για τη λειτουργία. Εξαρτάται επίσης από τη ύπαρξη μιας χημικής ανάλυσης εντοπισμού, για την λειτουργία που ενδιαφέρει, που να είναι δυνατό να εκτελεστεί αποτελεσματικά σε τεράστιες βιβλιοθήκες, καθώς η συχνότητα των ενεργών κλώνων είναι αρκετά χαμηλή[5]. Για παράδειγμα, σε μια αναζήτηση για λιπολυτικούς κλώνους που προέρχονται από γερμανικό έδαφος[3], μόνο 1 στους 730.000 κλώνους έδειξε δραστικότητα. Σε μια βιβλιοθήκη DNA από έδαφος της Βορείου Αμερικής, 29 από

τους συνολικά 25.000 κλώνους εξέφρασαν αιμολυτική δραστηριότητα. Ως εκ τούτου, η σπανιότητα των ενεργών κλώνων απαιτεί την ανάπτυξη αποτελεσματικών μεθόδων μελέτης, παρακολούθησης και επιλογής για την ανακάλυψη νέων δραστηριοτήτων ή μορίων. Ακριβώς όπως η γενετική των βακτηρίων βασίζεται σε μεθόδους επιλογής για την ανίχνευση συμβάντων χαμηλής συχνότητας, η μεταγονιδιωματική θα πρέπει να αναπτυχθεί με την ανίχνευση συγκεκριμένων φαινοτύπων που θα αυξήσουν τη συλλογή ενεργών κλώνων που μπορούν να συγκριθούν, να αναλυθούν και να χρησιμοποιηθούν για την κατασκευή ενός εννοιολογικού πλαισίου για τη λειτουργική ανάλυση[3]. Πολλές προσεγγίσεις αναπτύσσονται ώστε να μετριαστούν αυτοί οι περιορισμοί. Αναπτύσσονται βελτιωμένα συστήματα για την ετερόλογη έκφραση γονιδίων με φορείς μεταφοράς που διευκολύνουν την επιλογή του μεταγονιδιωματικού DNA σε ποικίλα είδη ξενιστών και με τροποποιήσεις του Escherichia coli ώστε να επεκταθεί το φάσμα της γονιδιακής έκφρασης[5]. Για παράδειγμα, η ομάδα Daniel σχεδίασε μια έξυπνη μέθοδο διαλογής για τους $Na^+(Li^+)/H^+$ αντιμεταφορείς που απαιτεί ολοκλήρωση μεταλλαγμένου E. coli με ανεπάρκεια στους τρεις Na^+/H^+ αντιμεταφορείς (nhaA, nhaB, και chaA) που επιτρέπει την ανάπτυξη σε μέσο που περιέχει 7.5 mM LiCl. Αυτή η ισχυρή μέθοδος επιλογής διευκόλυνε την ανακάλυψη δύο νέων αντιμεταφορικών πρωτεϊνών σε μία βιβλιοθήκη 1,480,000 κλώνων που περιείχαν DNA που απομονώθηκε από χώμα. Μία άλλη στρατηγική επιλογής εμπεριέχει την ολοκλήρωση (complementation) ενός μεταλλαγμένου E. coli με ανεπάρκεια στην παραγωγή βιοτίνης, η οποία οδήγησε στην απομόνωση επτά νέων οπερονίων για τη σύνθεση βιοτίνης από καλλιέργειες εμπλουτισμού που προέρχονται από δείγματα εδάφους ή περιττώματα αλόγων. Επίσης, μέσα επιλογής υψηλής απόδοσης μπορούν να αποτελέσουν εναλλακτική λύση όταν οι λειτουργίες που ενδιαφέρουν δεν παρέχουν τη βάση για την επιλογή. Για παράδειγμα, σε ορισμένα μέσα-δείκτες, οι ενεργοί κλώνοι παρουσιάζουν μια χαρακτηριστική, εύκολα διακριτή εμφάνιση, ακόμη και όταν βρίσκονται σε υψηλή πυκνότητα. Με την παρουσία του χλωριδίου του τετραζολίου, ως δείκτη σήμανσης, ο Henne και οι συνεργάτες του ανίχνευσαν κλώνους που χρησιμοποιούν 4-υδροξυβουτυρικό σε βιβλιοθήκες DNA προερχόμενες από αγροτικό ή παραποτάμιο έδαφος. Ιδιαίτερα σπάνιοι λιπολυτικοί κλώνοι ανιχνεύθηκαν στις ίδιες βιβλιοθήκες με παραγωγή καθαρών κρυστάλλων σε μέσα που περιέχουν ροδαμίνη και είτε τριολεΐνη είτε τριβουτυρινη.

Η επιτυχία της μεταγονιδιωματικής ανάλυσης που βασίζεται στην λειτουργία απαιτεί πιστή και ακριβής μεταγραφή και μετάφραση του γονιδίου ή των γονιδίων που ενδιαφέρουν και έκκριση του γονιδιακού προϊόντος, εάν απαιτείται αυτό να είναι εξωκυτταρικό, λόγω της μεθόδου μελέτης ή χημικής ανάλυσης. Η λειτουργική ανάλυση εντόπισε νέα αντιβιοτικά, γονίδια αντιβιοτικής αντίστασης, μεταφορείς $Na^+(Li^+)/H^+$ και αποικοδομητικά ένζυμα. Η ισχύς της προσέγγισης έγκειται στο ότι δεν απαιτεί τα γονίδια που ενδιαφέρουν να έχουν αναγνωριστεί από μια ανάλυση

αλληλούχισης, καθιστώντας την ως την μοναδική προσέγγιση μεταγονιδιωματικής που έχει τη δυνατότητα να προσδιορίσει καινούργιες κατηγορίες γονιδίων που σχετίζονται με νέες ή ήδη γνωστές λειτουργίες. Ο σημαντικός περιορισμός είναι ότι πολλά γονίδια, ίσως τα περισσότερα, δεν θα εκφράζονται στο οποιοδήποτε βακτήριο ξενιστή που επιλέγεται για κλωνοποίηση. Στην πραγματικότητα, υπάρχει μια εγγενής αντίφαση σε αυτή την προσέγγιση καθώς γονίδια κλωνοποιούνται από εξωτικούς οργανισμούς ώστε να ανακαλυφθούν νέα μοτίβα στη βιολογία και όμως αυτά τα γονίδια απαιτείται να εκφραστούν σε Escherichia coli ή σε άλλα "εξημερωμένα" βακτήρια, ώστε να μπορούν να ανιχνεύονται. Η ποικιλομορφία των οργανισμών των οποίων το DNA έχει επιτυχώς εκφραστεί σε E. coli είναι εκπληκτική, αλλά η ετερόλογη έκφραση παραμένει ένα εμπόδιο για την εξαγωγή της μέγιστης πληροφορίας από λειτουργικές αναλύσεις μεταγονιδιωματικής.



Εικόνα 2 Οι διαφορετικές προσεγγίσεις στην μεταγονιδιωματική ανάλυση. Ανάλυση αλληλούχισης και λειτουργική ανάλυση ως αποτέλεσμα της μεταγονιδιωματικής πορείας .

1.4 Οικολογικές και βιοτεχνολογικές εξελίξεις μέσω της μεταγονιδιωματικής

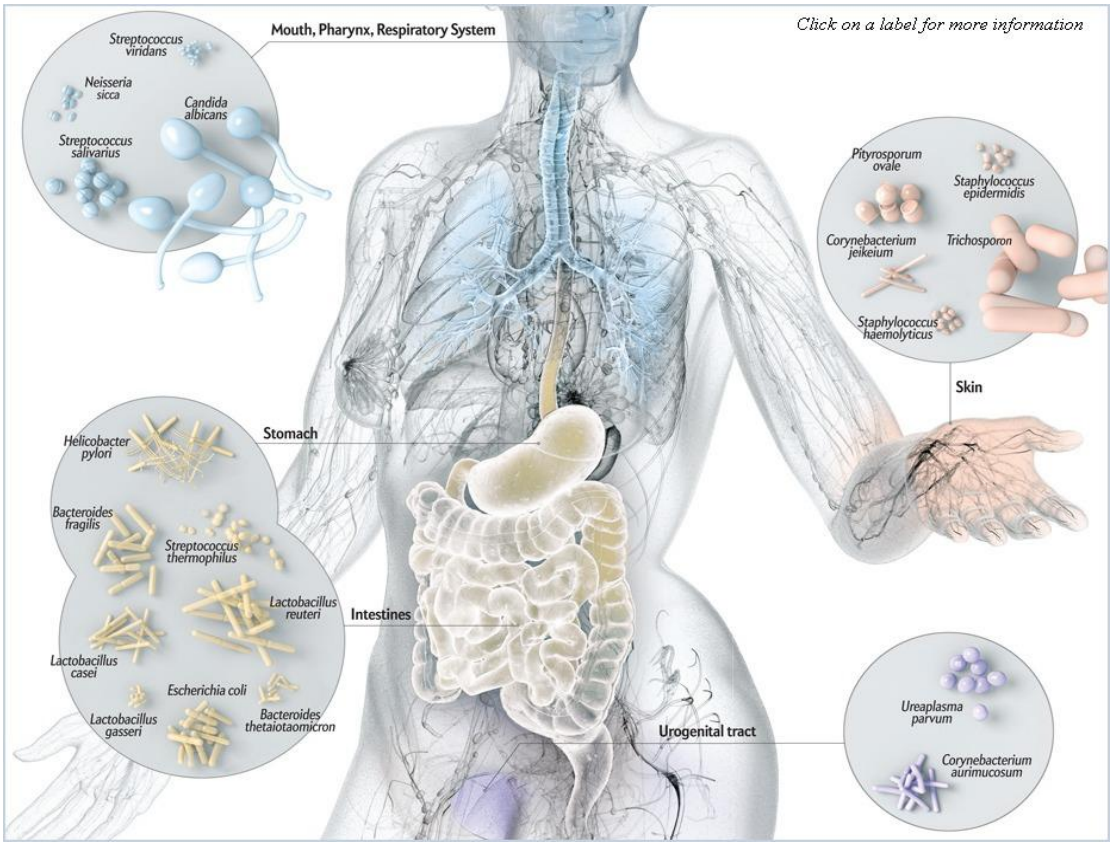
Συμβίωση : Πολλοί μικροοργανισμοί που παρουσιάζουν στενές συσχετίσεις ή έχουν συμβιωτικές σχέσεις με τους ξενιστές τους είναι δύσκολο να καλλιεργηθούν μακριά από τον κεντρικό οργανισμό ξενιστή και για αυτό, οι συμβιωτικοί οργανισμοί ήταν από τους πρώτους υποψήφιους για μεταγονιδιωματική ανάλυση. Σε μια συμβίωση μεταξύ ενός εντόμου και ενός βακτηρίου, μίας αφίδας και του

βακτηρίου Buchnera, η μεταγονιδιωματική ανάλυση έδειξε το λόγο για τον οποίο το βακτήριο Buchnera παρουσιάζει απειθαρχία στην καλλιέργεια του. Αν και σε αυτή την περίπτωση υπάρχει μια απλή μικροβιακή κοινότητα - η οποία περιλαμβάνει ένα μόνο είδος - λογίζεται ως μεταγονιδιωματική μελέτη, επειδή είναι το πρώτο παράδειγμα εφαρμογής της γονιδιωματικής για ένα μικροοργανισμό που δεν έχει καλλιεργηθεί in vitro. Η ανάλυση του γονιδιώματος του Buchnera έδωσε το εντυπωσιακό συμπέρασμα ότι το βακτήριο είχε χάσει σχεδόν 2000 γονίδια, αφότου άρχισε η συμβιωτική αυτή σχέση, 200-250 εκατομμύρια χρόνια πίσω. Ως αποτέλεσμα, περιέχει ένα ιδιαίτερα μικρό γονιδίωμα που αποτελείται από μόνο 564 γονίδια (περίπου το ένα δέκατο του μεγέθους του γονιδιώματος της E. coli) και δεν διεξάγει πολλές από τις λειτουργίες της ζωής που είναι απαραίτητες για την ανεξάρτητη ανάπτυξη του. Αντ 'αυτού, αυτές οι λειτουργίες παρέχονται από το έντομο-ξενιστή. Η γονιδιωματική μελέτη του Buchnera βοήθησε τους βιολόγους να βελτιώσουν την κατανόησή τους επί του θέματος της συμβίωσης[2].

Επιπροσθέτως, υπάρχουν στενές συσχετίσεις μεταξύ φυλογενετικά ποικίλων βακτηρίων και μιας ποικιλίας ευκαρυωτικών ξενιστών στο θαλάσσιο περιβάλλον. Ένας σημαντικός αριθμός φυσικών προϊόντων που σχετίζονται με την βιοϊατρική, όπως αντικαρκινικοί παράγοντες, αντι-μολυσματικοί παράγοντες ή παράγοντες που σχετίζονται με άλλες βιοδραστηριότητες λαμβάνονται από ασπόνδυλα ζώα, παραδείγματος χάριν σπόγγους, χιτωνόζωα και βρυόζωα. Ωστόσο, ένα σημαντικό μειονέκτημα είναι η χαμηλή ποσότητα αυτών των ενώσεων. Υπάρχουν αυξανόμενες ενδείξεις ότι πολλοί μεταβολίτες, κυρίως πολυκετίδια και μη ριβοσωμικά πεπτίδια δεν παράγονται από τα ίδια τα ζώα, αλλά από σχετιζόμενους βακτηριακούς συμβιωτικούς οργανισμούς[6].

Οι κοινότητες των οργανισμών που ζουν συμβιωτικά με τα ανθρώπινα όντα, γνωστές συλλογικά ως το "ανθρώπινο μικροβίωμα"(Εικόνα 3), παρουσιάζουν εξαιρετικά χαρακτηριστικά στην πολυπλοκότητά τους αλλά και ως προς την επίδραση στην υγεία. Μεταγονιδιωματικές μελέτες έχουν αποκαλύψει μοτίβα που ήταν αόρατα σε μελέτες που στηρίζονται σε καλλιέργειες. Για παράδειγμα, η μεταγονιδιωματική και άλλες μέθοδοι, ανεξάρτητες ύπαρξης καλλιεργειών, απέδειξαν ότι κάθε άτομο φέρει και μια μοναδική μικροβιακή κοινότητα στην γαστρεντερική του οδό και στην πραγματικότητα αυτές οι κοινότητες έχουν χαρακτηριστεί ως ένα "δεύτερο δακτυλικό αποτύπωμα", επειδή παρέχουν μια προσωπική υπογραφή για το κάθε άτομο. Σε αντίθεση με τα δακτυλικά αποτυπώματα ωστόσο, οι μικροβιακές κοινότητες είναι δυναμικές και αλλάζουν κατά τη διάρκεια της ζωής, γιατί η σύνθεσή τους επηρεάζεται από πολλούς παράγοντες. Οι άνθρωποι που ζουν μαζί αναπτύσσουν στο έντερο τους μικροβιακές κοινότητες που μοιράζονται ορισμένα χαρακτηριστικά και οι ομοζυγωτικοί δίδυμοι έχουν εντυπωσιακά παρόμοιες κοινότητες. Αυτό δείχνει ότι τόσο το περιβάλλον όσο και η γενετική, καθορίζουν την κοινότητα των μελών. Το ανθρώπινο μικροβίωμα

προσελκύει το ενδιαφέρον και την προσοχή επιστημόνων και των Εθνικών Ινστιτούτων Υγείας, εξ' αιτίας μιας εκπληκτικής έρευνας που δείχνει ότι η σύνθεση της μικροβιακής κοινότητας του ανθρώπινου εντέρου συνδέεται με την παχυσαρκία, τις καρδιακές παθήσεις, τον καρκίνο του παχέος εντέρου και το άσθμα. Η επόμενη δεκαετία θα είναι συναρπαστική καθώς η έρευνα θα δώσει έμφαση στη μελέτη των ρόλων των μικροβιακών "εταίρων" του ανθρώπου, σε περιπτώσεις υγείας αλλά και ασθένειας[2].



Εικόνα 3 Σχηματική ενδεικτική απεικόνιση του ανθρώπινου μικροβιώματος.

Παλαιογονιδιωματική: Το DNA αρχαίων ζώων δεν είναι μόνο κατεστραμμένο και κατακερματισμένο αλλά επίσης είναι και αναμεμιγμένο με γονιδιώματα των εν αφθονία ευκαιριακών μικροβίων. Μιτοχονδριακές αλληλουχίες ενισχυμένες μέσω PCR έχουν χρησιμοποιηθεί για τον προσδιορισμό φυλογενετικών σχέσεων μεταξύ ζώων που έχουν εξαφανιστεί και σύγχρονων ζώων. Μια μεταγονιδιωματική προσέγγιση μαζί με αλληλούχιση υψηλής απόδοσης παρέχει τα μέσα για την πρόσβαση σε πυρηνικά γονιδιώματα εξαφανισμένων οργανισμών χωρίς την ανάγκη ενίσχυσης. Χρησιμοποιώντας τη μεταγονιδιωματική στρατηγική, ελήφθησαν περίπου 27 kb υποτιθέμενης αλληλουχίας DNA από αρκούδα των σπηλαίων (Ursus spelaeus) και μέσω ενίσχυσης με PCR ορθόλογων αλληλουχιών από τις σύγχρονες μαύρες, καφέ και πολικές αρκούδες, επιβεβαιώθηκε η προέλευση και

ανακατασκευάστηκε το φυλογενετικό δέντρο. Αυτές οι προσεγγίσεις ανοίγουν τη δυνατότητα για την πραγματοποίηση προγραμμάτων γονιδιώματος που στοχεύουν σε εξαφανισμένα είδη και θα μπορούσαν να φέρουν επανάσταση στον κλάδο της Παλαιοβιολογίας[6].

Πηγή βιοκαταλυτών και ενζύμων: Η βιομηχανία της βιοτεχνολογίας χρησιμοποιεί ήδη εκατοντάδες μικροβιακά ένζυμα και σχετικά προϊόντα και η παγκόσμια βιομηχανική αγορά ενζύμων υπολογίζεται σήμερα σε πάνω από 2 δισ. δολάρια ετησίως, κυρίως σε ότι αφορά τεχνικές (συμπεριλαμβανομένων επιστημονικών, χαρτοπολτού και χαρτιού), γεωργικές, καθώς και σχετικές με τρόφιμα και ζωοτροφές, εφαρμογές. Η μεγάλη πλειονότητα τέτοιων ενζύμων είναι το αποτέλεσμα παραδοσιακών προσεγγίσεων: αρχικά εμπλουτισμός, στη συνέχεια δημιουργία καλλιεργειών, απομόνωση και τέλος καθαρισμός του ενζύμου. Συλλογικά, η μεταγονιδιωματικές βάσεις δεδομένων και η προσπάθεια, τώρα σε πλήρη εξέλιξη, να εκφραστούν, να κρυσταλλωθούν και να χαρακτηριστούν δομικά και λειτουργικά ολόκληρα πρωτεόματα πολλών οργανισμών μοντέλων, είναι πιθανόν να ενισχύσουν το ρυθμό ανακάλυψης των εν λόγω πολύτιμων καταλυτών κατά τουλάχιστον μία τάξη μεγέθους, φέροντας μια επανάσταση στην πράσινη χημεία. Κατά ειρωνικό τρόπο, ορισμένα από τα βασικά προϊόντα των δραστηριοτήτων αυτών μέχρι σήμερα διαδραματίζουν ζωτικό ρόλο στην ίδια τη διαδικασία της ανακάλυψης τους. Για παράδειγμα, η αλυσιδωτή αντίδραση πολυμεράσης-η οποία είναι η βάση της σύγχρονης μοριακής περιβαλλοντικής μικροβιολογίας, της εγκληματολογίας και της μοριακής διαγνωστικής, βασίζεται σε γονίδια που κλωνοποιούνται από θερμόφιλα βακτήρια και Αρχαία[7].

Η μεταγονιδιωματική προσέγγιση έχει λοιπόν χρησιμοποιηθεί με επιτυχία για να ληφθούν νέοι φυσικοί βιοκαταλύτες με ασυνήθιστες ιδιότητες, καθώς τα χαρακτηριστικά του ενζύμου ποικίλλουν ανάλογα με το περιβάλλον στο οποίο αναπτύσσονται οι οργανισμοί, διευρύνοντας έτσι το πεδίο της ενζυμολογίας[6]. Λαμβάνοντας υπόψη την μεγάλη ποικιλομορφία της προκαρυωτικής ζωής σε περιβάλλοντα εδάφους, οι μεταγονιδιωματικές βιβλιοθήκες του εδάφους προσφέρουν μια από τις καλύτερες πηγές κατά την αναζήτηση ενός ευρέος φάσματος βιοκαταλυτών. Αυτές οι βιβλιοθήκες μπορούν να αναζητηθούν μέσω της άμεσης αλληλούχισης των κλώνων και συγκρίνοντας τις αλληλουχίες με τις βάσεις δεδομένων, ή μέσω λειτουργικής ανάλυσης, αναζητώντας μια συγκεκριμένη δραστηριότητα. Ένα ποικιλόμορφο εύρος των βιοκαταλυτών έχουν ληφθεί από μεταγονιδιωματικές βιβλιοθήκες. Μερικά παραδείγματα βιοκαταλυτών που προέρχονται από μεταγονιδιώματα του εδάφους περιλαμβάνουν εστεράσες, υδρατάσες νιτρίλιου, αναγωγάσες αλκοολών, αμιδάσες, κυτταρινάσες, α-αμυλάσες, 1,4-α-γλυκάνες διακλαδωτικής αλυσίδας και πηκτικές λυάσες[4] .

Soil source (origin)	Enzyme	Feature of the enzyme
Alluvial soil (Seoho stream, Korea)	Amylase	Soluble starch and cyclodextrin hydrolysis, trans-glycosylation activity
Mountain soil (Kagil at Northwestern Himalayas, India)	Amylase	Cold-adapted amylolytic enzyme
Red soil (Yingtan, China)	Cellulase/xylanase	High activity at low temperature, pH and thermal stability, halotolerance, high stability in the presence of proteolytic enzymes
Compost soil (soil near hot spring, Japan)	Xylanase	Novel thermo-alkali-stable xylanase
Alluvial soil (Nakdong River, Korea)	Esterase/amidase	Chloramphenicol and florfenicol hydrolysis
Oil contaminated soil (Weitze, Germany)	Esterase/lipase	Highly enantioselective for (+)-menthylacetate
Forest soil (Gwangneung forest, Korea)	Esterase/lipase	First description of the GDSL family of serine esterases/lipases from metagenomic approach
Rhozosphere soil (Korea)	Esterase	A novel family of lipolytic enzyme
Arctic soil (the Dasan Station, Ny-Alesund)	Esterase	Two novel cold-active family VIII esterases showing lactamase activity
Forest soil (Parana' state, Brazil)	Lipase	Moderately thermostable lipase derived from a member of the phylum Acidobacteria
Alkaline polluted soil (Guangxi, China)	β-Glucosidase	First member of a novel family of the β-glucosidase gene
Mangrove soil (Shenzen, China)	β-Glucosidase	High hydrolysis ability for soybean isoflavone glycosides
Pasture soil (Toulouse, France)	Lactonase	Novel metallohydrolase-related enzyme with an N-acylhomoserine lactone hydrolysis activity
Field soil (Göttingen, Germany)	Lactonase	Novel lactonases to inhibit motility and biofilm formation in *Pseudomonas aeruginosa*
Desert sand soil (Gobi, Mongolia)	Protease	A heat resistant protease belonging to thermitase subfamily, and an alkaline protease belonging to sutilisin and protease K subfamily
Garden soil (Taichung, Taiwan)	Racemase	A lysine racemase (*lyr*) gene isolated by functional complementation of *Escherichia coli* BCRC 51734 cells as the host and D-lysine as the selection agent
Mangrove soil (Shenzen, China)	Oxidase	Novel multicopper oxidase with laccase activity
Field soil (Göttingen, Germany)	Reductase	NADP-dependent short-chain dehydrogenase/reductase (SDR) involved in inactivation of N-(3-oxo-dodecanoyl)-L-homoserine lactone (3-oxo-C12-HSL)

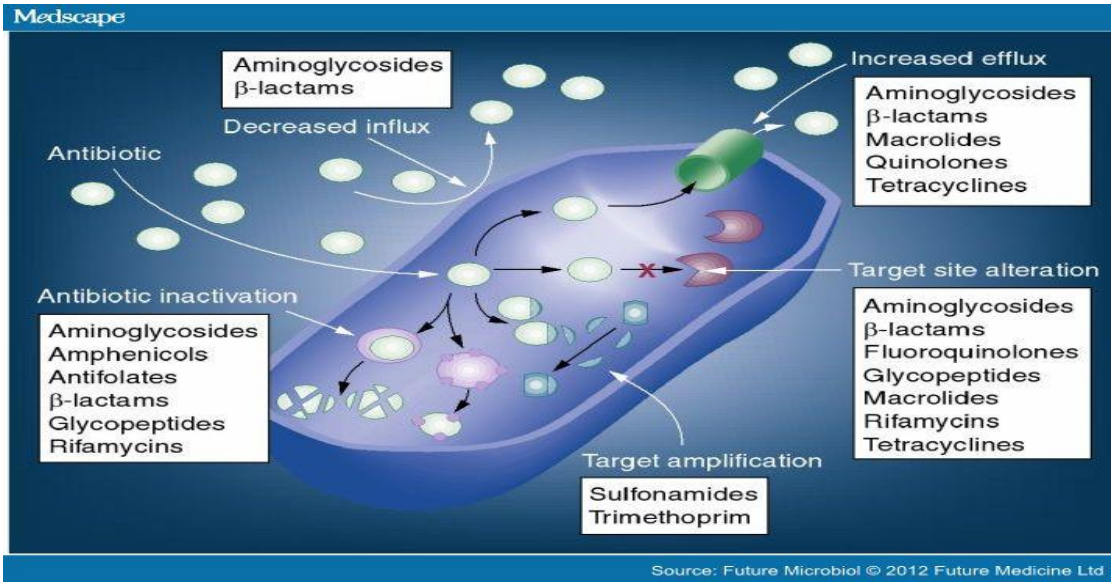
Πίνακας 1 Μικροβιακά ένζυμα και λειτουργικά τους χαρακτηριστικά, προερχόμενα από διάφορα είδη εδαφών μέσω της μεταγονιδιωματικής προσέγγισης[8].

Αντιβιοτικά: Η μεταγονιδιωματική ανάλυση που βασίζεται στην λειτουργική μελέτη προσφέρει έναν διαφορετικό δρόμο κατανόησης των μικροβιακών κοινοτήτων. Σε αυτή την εκδοχή της μεταγονιδιωματικής, οι κλώνοι που περιέχουν DNA από την κοινότητα, μελετώνται αρχικά ως προς μια λειτουργία ή μια διεργασία που μεταφέρουν σε ένα υποκατάστατο ξενιστή, όπως η E. coli. Το βασικό πλεονέκτημα της προσέγγισης της λειτουργικής ανάλυσης είναι ότι τα γονίδια που ανακαλύπτονται, δεν επιβάλλεται να είναι μέρος μιας οικογένειας γονιδίων ήδη γνωστής λειτουργίας για να παρέχουν πληροφορίες. Στην πραγματικότητα σημαντικότατο αποτέλεσμα αυτής της προσέγγισης είναι η προσδιορισμός λειτουργιών σε γονίδια άγνωστης λειτουργίας. Αυτό είναι σημαντικό για την εξέλιξη της μεταγονιδιωματικής αλλά και της γονιδιωματικής γενικά, επειδή πολλά γονίδια στις βάσεις δεδομένων δεν έχουν ομόλογα γνωστής λειτουργίας. Αυτά τα γονίδια μπορεί να αποτελούν έως και το 60% των γονιδίων στις μεταγονιδιωματικές βιβλιοθήκες. Εάν τα γονίδια που είναι υπεύθυνα για ενός είδους δραστηριότητα έχουν "συλληφθεί" σε ένα μεγάλο θραύσμα του DNA, στη συνέχεια, η αλληλουχία τους αλλά και οι γειτονικές αλληλουχίες είναι δυνατό να παρέχουν ενδείξεις για την ταυτότητα του οργανισμού από τον οποίο προέρχεται το DNA.

Η λειτουργική μεταγονιδιωματική ανάλυση οδήγησε στην ανακάλυψη νέων ενζύμων και μικρών μορίων, όπως τα αντιβιοτικά. Υπήρξε ιδιαίτερα παραγωγική ως προς την ανακάλυψη γονιδίων ανθεκτικότητας σε αντιβιοτικά από το έδαφος. Η εφαρμογή της μεταγονιδιωματικής στην παρακολούθηση γονιδίων ανθεκτικότητας στα αντιβιοτικά στο έδαφος, παρουσιάζει ιδιαίτερο ενδιαφέρον, διότι οι προσδιοριστές που σχετίζονται με την αντοχή στα αντιβιοτικά πιστεύεται ότι είναι άφθονοι στο περιβάλλον όπου παράγονται τα αντιβιοτικά. Πολλοί μικροοργανισμοί του εδάφους είναι δύσκολο να καλλιεργηθούν in vitro και παρόλο που η ανθεκτικότητα στα αντιβιοτικά είναι ένα θέμα αυξανόμενης σημασίας για την υγεία, λίγα είναι γνωστά για τις ρίζες της στο περιβάλλον και ακόμη λιγότερα είναι γνωστά σχετικά με το ρόλο των μη καλλιεργήσιμων μικροοργανισμών ως δεξαμενές γονιδίων ανθεκτικότητας. Το χαρακτηριστικό της ανθεκτικότητας στα αντιβιοτικά είναι ένας ιδανικός υποψήφιος για μεταγονιδιωματική ανάλυση επειδή οι κλώνοι που φέρουν τα γονίδια ανθεκτικότητας μπορούν να ανιχνευθούν εύκολα σε μεγάλες βιβλιοθήκες. Αν οι βιβλιοθήκες κατασκευαστούν σε ένα υποκατάστατο βακτήριο ξενιστή που είναι ευαίσθητος στο αντιβιοτικό που μελετάται, στη συνέχεια, όταν οι βιβλιοθήκες καλλιεργούνται σε μέσα που περιέχουν το κατάλληλο αντιβιοτικό, οι μόνοι κλώνοι που θα πρέπει να αναπτυχθούν είναι εκείνοι που περιέχουν τα γονίδια ανθεκτικότητας[2].

Η μεταγονιδιωματική αποκαλύπτει μια εκπληκτική πολυμορφία των γονιδίων ανθεκτικότητας. Η αντίσταση στις αμινογλυκοσίδες (μια ομάδα που περιλαμβάνει αντιβιοτικά όπως η στρεπτομυκίνη) προσδίδεται από γονίδια που κωδικοποιούν ένζυμα που τροποποιούν τα αντιβιοτικά με ακετυλίωση. Αυτά τα ένζυμα αποτελούν μια γνωστή κατηγορία προσδιοριστών αντίστασης στην αμινογλυκοσίδη αλλά εκείνα που προέρχονται από το μεταγονιδίωμα του εδάφους σχηματίζουν μια νέα ξεχωριστή υποομάδα. Ομοίως, το μεταγονιδίωμα του εδάφους περιέχει γονίδια για αντοχή σε β-λακτάμες (η ομάδα των αντιβιοτικών που περιλαμβάνει και την πενικιλίνη) που εμπίπτουν σε γνωστές κατηγορίες γονιδίων ανθεκτικότητας, αλλά αποκλίνουν σημαντικά από τα γονίδια που προέρχονται από καλλιεργημένους οργανισμούς. Επιπλέον, ορισμένοι προσδιοριστικοί παράγοντες αντίστασης δεν εμπίπτουν σε γνωστές κατηγορίες γονιδίων έκφρασης ανθεκτικότητας και συνεπώς μπορεί να αντιπροσωπεύουν νέους μηχανισμούς αντίστασης. Το εάν αυτά τα γονίδια θα βρουν το δρόμο τους προς τις κλινικές μελέτες είναι άγνωστο, αλλά η γνώση της ύπαρξής τους καθιστά δυνατή την έναρξη την παρακολούθησης της δραστηριότητας τους[2].

Οι εγκαταστάσεις επεξεργασίας λυμάτων αποτελούν πεδία διεπαφής μεταξύ διαφορετικών ειδών περιβάλλοντος και, ως εκ τούτου, προσφέρουν μια ευκαιρία σε κινητά στοιχεία (συμπεριλαμβανομένης της ανθεκτικότητας) να αναμειγνύονται μεταξύ παθογόνων οργανισμών, ευκαιριακών παθογόνων οργανισμών, και περιβαλλοντικών βακτηρίων .



Εικόνα 4 Μηχανισμοί ανθεκτικότητας των βακτηρίων στα αντιβιοτικά. Οι κατηγορίες αντιβιοτικών που επηρεάζονται από τον κάθε ένα μηχανισμό απαριθμούνται στα κουτιά[9].

Η παρουσία αντιβιοτικών στα λύματα (Πίνακας 2) χαρακτηρίζεται από ύπαρξη δεικτών αντίστασης που είναι σε θέση να εξαπλώνονται διαμέσου της μικροβιακής κοινότητας και αυτό έχει ως αποτέλεσμα, ανθεκτικά στα αντιβιοτικά βακτήρια να μπορούν δυνητικά να διαδώσουν ευρέως τα γονίδια ανθεκτικότητας τους, μεταξύ των μελών της ενδογενούς μικροβιακής κοινότητας. Τα προϊόντα της ιλύος των αστικών και αγροτικών εγκαταστάσεων επεξεργασίας λυμάτων χρησιμοποιούνται όλο και περισσότερο ως λιπάσματα σε γεωργικές καλλιέργειες, διασπείροντας άγνωστες ποσότητες γονιδίων ανθεκτικότητας αλλά και αντιβιοτικών που αντέχουν στις πρότυπες διεργασίες επεξεργασίας λυμάτων. Το επίπεδο της ανθεκτικότητας στα αντιβιοτικά σε δείγματα ενεργοποιημένης ιλύος δεν είναι πλήρως γνωστό και οι λίγες μελέτες μέχρι σήμερα δεν έδωσαν αξιόπιστα αποτελέσματα. Η ενεργοποιημένη ιλύς ως διαδικασία μπορεί να προάγει τις κυτταρικές αλληλεπιδράσεις μεταξύ των διαφορετικών μικροοργανισμών στα λύματα, αλλά είναι επίσης ένα ιδιαίτερα ανταγωνιστικό περιβάλλον όπου το κλάσμα των παθογόνων οργανισμών μειώνεται λόγω του ανταγωνισμού με άλλα μικρόβια ή μέσω της θήρευσης[9].

Κατά καιρούς, μεταγονιδιωματικές λειτουργικές αναλύσεις αλλά και αναλύσεις αλληλούχισης έχουν χρησιμοποιηθεί για την ταυτοποίηση προσδιοριστών ανθεκτικότητας σε αντιβιοτικά που εντοπίζονται σε βακτηριακά χρωμοσώματα, πλασμίδια ή ιούς σε ποσότητες ενεργού ιλύος ενώ υψηλά επίπεδα γονιδίων αντίστασης σε αντιβιοτικά έχουν επίσης βρεθεί, μέσω ανάλυσης αλληλούχισης, σε βακτήρια που ζουν στα ιζήματα ποταμών (Πίνακας 2) και προέρχονται από ρεύματα εγκαταστάσεων επεξεργασίας λυμάτων. [9].

Ανθεκτικά βακτήρια στα αντιβιοτικά έχουν βρεθεί ευρέως και στο υδάτινο περιβάλλον όπως αναφέρθηκε (Πίνακας 2). Ανθεκτικοί οργανισμοί σε θαλάσσιο περιβάλλον είναι δυνατό να αναπτυχθούν μετά από έκθεση τους σε αντιβιοτικά που προέρχονται από γεωργικές απορροές ή από παράγωγα εγκαταστάσεων επεξεργασίας λυμάτων. Οι υδατοκαλλιέργειες αποτελούν μία σημαντική πηγή αντιβιοτικών που μολύνουν τα θαλάσσια περιβάλλοντα, επειδή τα αντιβιοτικά προστίθενται ώστε να αντιμετωπιστούν οι ασθένειες που κυριαρχούν στα κλουβιά που περιέχουν υψηλό αριθμό ζώων. Υπάρχουν αποδείξεις ότι ανθεκτικότητα στα αντιβιοτικά μπορεί επίσης να εμφανιστεί στα θαλάσσια περιβάλλοντα χωρίς απαραίτητα την προσθήκη του αντιβιοτικού που θα προκαλέσει τη μόλυνση. Για παράδειγμα, τα ίδια γονίδια ανθεκτικότητας που βρέθηκαν στην κλινική μελέτη ανθρώπινων παθογόνων έχουν εντοπιστεί μεταξύ παρθένων οικοσυστημάτων, δίχως ιστορία μόλυνσης από αντιβιοτικά. Για παράδειγμα, μια μεταγονιδιωματική ανάλυση αλληλούχισης για την μελέτη μιας μικροβιακής κοινότητας που συνδέεται με τα κοράλλια Porites astreoides, έδειξε την παρουσία γονιδίων αντοχής στις φθοριοκινολόνες. Οι μελετητές πρότειναν την θεωρία ότι το κοράλλι φιλοξενεί εξειδικευμένους μικροβιοτικούς οργανισμούς που μπορούν να προστατεύσουν τον ύφαλο από παθογόνα με την παραγωγή αντιβιοτικών, ωστόσο, δεν υπήρξε ένδειξη παραγωγής φθοριοκινολόνης στα μεταγονιδιώματα και δεν υπήρχαν στοιχεία που να δείχνουν ύπαρξη φθοριοκινολόνης από κάποια πηγή ανθρώπινης προέλευσης στο περιβάλλον[9].

Antibiotic	Resistance mechanism	Host	Source
β-Lactams	PBP2 mutations	*Proteus mirabilis*	–[a]
	PBP5 mutations	*Enterococcus faecium*	–[a]
	*amp*C regulators mutations	Gram-negative species	–[a]
	*amp*C promoter region mutations	*Escherichia coli*	Recreational beaches, drinking water
	Acquired AmpC	*E. coli*	Recreational beaches, drinking water, river, biofilm of water supplies
	Acquired CTX-M	*E. coli*	River, sediment, birds
	Acquired KPC	*Klebsiella pneumoniae*	Hospital waste water effluent
	Acquired VIM	*Brevundimonas diminuta, Rhizobium radiobacter, Pseudomonas monteilii, Pseudomonas aeruginosa, Ochrobactrum anthropi, Enterobacter ludwigii, Pseudomonas pseudoalcaligenes*	Hospital waste water effluent
	Acquired IMP	*Pseudomonas fluorescens*	Waste water
	Acquired OXA-23	*Acinetobacter baumannii*	River, hospital waste water effluent
	Acquired OXA-48	*Serratia marcescens*	River
	Acquired NDM-1	*P. aeruginosa, Achromobacter* spp., *Kingella denitrificans*	Tap water
Fluoroquinolones	QRDR (quinolones resistance determining region) mutations	*P. aeruginosa*	Hospital and urban waste water effluent
		Aeromonas spp., *E. coli*	River and lake, urban effluent
	QnrS	*Aeromonas allosaccarophila*	Lake
		E. coli	River
	QnrS2	*Aeromonas punctata, Aeromonas media*	Lake
	QnrVC4	*A. punctata*	Waste water effluent
	QepA efflux	Metagenome	River sediment, water from farm environment
	OqxAB efflux	*E. coli*	Farm water
Vancomycin	modification of the peptidoglycan	*Enterococci* spp.	Waste water effluents, biofilm
Chloramphenicol and florfenicol	FloR efflux	Gram-negative species *Aeromonas bestiarium*	Aquacultures streams
Tetracyclines	Tet efflux	Several species	Farms, sediment
MDR[b]	Over-expression of RND efflux pumps	Gram-negative	–[a]

[a] *Observed in clinics but likely occurring in environmental and water habitats.*
[b] *MDR, multi-drug resistance.*

Πίνακας 2 Επισκόπηση ορισμένων μηχανισμών βακτηριακής ανθεκτικότητας σε αντιβιοτικά που παρουσιάζονται σε διάφορα υδατικά περιβάλλοντα[10].

Βιογεωχημικοί κύκλοι : Οι μικροοργανισμοί παίζουν σημαντικό ρόλο στους βιοχημικούς κύκλους που είναι υπεύθυνοι για το περιβάλλον του εδάφους και των ωκεανών. Οι περιβαλλοντικές μελέτες αλληλούχισης παρέχουν πρόσβαση σε μια πολύ μεγαλύτερη δεξαμενή γονιδιωματικών και μεταβολικών πληροφοριών και έχουν οδηγήσει σε νέες ανακαλύψεις που απεικονίζουν τις εξελικτικές σχέσεις μεταξύ των βακτηρίων. Η σχεδόν πλήρης συναρμολόγηση του γονιδιώματος ενός ακαλλιέργητου βακτηρίου, του Kuenenia stuttgartiensis, αποκάλυψε μοναδικές μεταβολικές προσαρμογές που σχετίζονται με την αναερόβια οξείδωση του αμμωνίου[2].

Μελέτη εγκαταστάσεων απορροής οξέων: Η μεταγονιδιωματική έχει οδηγήσει σε πολλές εκπλήξεις στην μελέτη δειγμάτων που προέρχονται από εγκαταστάσεις απορροής οξέων. Εγκαταλελειμμένα ορυχεία, εμφανίζουν συχνά απορροές ιδιαίτερα όξινες και πλούσιες σε τοξικά μεταλλικά ιόντα. Το pH του υλικού είναι δυνατό να πλησιάζει σε μηδενικές τιμές ή και ακόμα λιγότερο. Τα τοξικά απόβλητα θεωρούνται ένα από τα χειρότερα παραπροϊόντα των ορυχείων. Από τα ανασυγκροτημένα γονιδιώματα των κοινοτήτων προέκυψε ένα μοντέλο κύκλου του άνθρακα, του αζώτου και διαφόρων μετάλλων στο όξινο περιβάλλον των ορυχείων. Τα βιοφίλμ των μικροβιακών κοινοτήτων που προέρχονται από ορυχεία όξινης απορροής κυριαρχούνται από βακτήρια τύπου Leptospirillum και Sulfobacillus, αν και επίσης εντοπίζονται ποσότητες βακτηρίων Ferroplasma και Thermoplasmataleare. Τα θειούχα ορυκτά συμπεριλαμβανομένου και του σιδηροπυρίτη (FeS_2), που υπάρχουν στο ορυχείο, διαλύονται από αντιδράσεις οξείδωσης καταλυόμενες από τη μικροβιακή δραστηριότητα. Ο Tyson και συνεργάτες του εξήγαγαν DNA απευθείας από το βιοφίλμ και έλαβαν σχεδόν πλήρεις αλληλουχίες γονιδιώματος της ομάδας Leptospirillum II και του Ferroplasma τύπου II και μερικές (μη ολοκληρωμένες) αλληλουχίες γονιδιώματος για την ομάδα Leptospirillum III, του Ferroplasma τύπου I, καθώς και G-πλάσμα[6]. Ένα ιδιαίτερα ενδιαφέρον αποτέλεσμα ήταν η σύνδεση γονιδίων που σχετίζονται με την δέσμευση αζώτου, με την ομάδα Leptospirillum III που δεν βρίσκεται σε αφθονία και δεν θεωρούνταν μέχρι τότε ικανή να δεσμεύσει άζωτο. Μέλη αυτής της ομάδας, δεν είχαν καλλιεργηθεί ως τότε στο εργαστήριο αλλά μετά από αυτή την ανακάλυψη, προέκυψε ένας μηχανισμός καλλιέργειας τους. Το βιοφίλμ καλλιεργήθηκε με το αέριο άζωτο (N_2) ως μοναδική πηγή αζώτου και τα μέλη της ομάδας Leptospirillum III όντως αναπτύχθηκαν υπό αυτές τις συνθήκες[2].

Η shotgun αλληλούχιση (η αλληλούχιση που βασίζεται στον τυχαίο κατακερματισμό του DNA σε θραύσματα) από μόνη της δεν μπορεί εύκολα να χρησιμοποιηθεί για την ολοκλήρωση συνολικών μικροβιακών γονιδιωμάτων, ακόμη και σε κοινότητες που χαρακτηρίζονται μόνο ως μετρίως πολύπλοκες. Νεότερες μέθοδοι και προσεγγίσεις, όπως η μονοκυτταρική ενίσχυση του γονιδιώματος αναπτύσσονται πλέον και θα μπορούσαν ενδεχομένως να κατακερματίσουν και να συγκρίνουν τις

πολύπλοκες μικροβιακές κοινότητες που συναντώνται συχνά στη φύση. Χρησιμοποιώντας συνδυαστικά τις shotgun τεχνικές αλληλούχισης και την πρωτεομική που βασίζεται σε φασματομετρία μάζας με τη γονιδιωματική ανάλυση κοινοτήτων, παρατηρήθηκε υψηλή έκφραση πρωτεϊνών που σχετίζονται με το οξειδωτικό στρες, μαζί με ένα νέο πρωτεϊνικό κυτόχρωμα, το οποίο είναι ένα σημαντικό συστατικό της οξείδωσης του σιδήρου και του σχηματισμού των ορυχείων απορροής οξέων[6].

Το πρόγραμμα της Θάλασσας των Σαργασσών: Οι ωκεανοί του κόσμου φιλοξενούν τεράστιες ποσότητες μικροβιακών πληθυσμών που εν μέρει ρυθμίζουν τη ροή της ενέργειας, της ύλης και των αερίων του θερμοκηπίου στη θάλασσα. Οι βιολογικές ιδιότητες αυτών των κατανεμημένων μικροβιακών κοινοτήτων εξακολουθούν να έχουν περιγραφεί παρά μόνο ελάχιστα. Ως προσέγγιση σε αυτό το πρόβλημα, μια από τις μεγαλύτερες μεταγονιδιωματικές προσπάθειες αλληλούχισης που έχουν πραγματοποιηθεί μέχρι σήμερα χρησιμοποίησε μια shotgun έρευνα αλληλούχισης μικροβιακών συνόλων από τη Θάλασσα των Σαργασσών[7] (Εικόνα 5) -ένα ωκεάνιο περιβάλλον για το οποίο υπήρχε η πεποίθηση ότι χαρακτηρίζεται από χαμηλή πολυμορφία. Το έργο ξεκίνησε με τη συλλογή μικροβιακών κυττάρων και ιών σε διαφορετικά κλάσματα μεγέθους και εξαγωγή του DNA τους. Η ενιαία έρευνα ανέφερε 1.214.207 εντοπισμένα πιθανά γονίδια που κωδικοποιούν πρωτεΐνες, που αντιπροσώπευαν σχεδόν 10 φορές περισσότερες αλληλουχίες πρωτεϊνών από όσες ήταν παρούσες σε όλες τις επιμελημένες βάσεις δεδομένων πρωτεϊνών μέχρι εκείνη τη στιγμή. Το σετ δεδομένων από τη Θάλασσα των Σαργασσών είναι αξιοσημείωτο, όχι μόνο σε σχέση με τις νέες πληροφορίες που έδωσε, αλλά και λόγω του μεγάλου όγκου των δεδομένων που περιέχει. Η μελέτη της Θάλασσας των Σαργασσών ήταν μια από τις αρκετές πρόσφατες μελέτες που προαναγγέλλουν μια μαγική αλλαγή στις προσπάθειες της περιβαλλοντικής μικροβιολογίας και υπογραμμίζει τις σημαντικές προκλήσεις και ευκαιρίες που σχετίζονται με την αρχειοθέτηση, την ενσωμάτωση, και την ανάλυση μαζικών μεταγονιδιωματικών συνόλων δεδομένων. Γίνεται όλο και περισσότερο σαφές ότι οι μεταγονιδιωματικές αλληλουχίες DNA σύντομα θα ξεπερνούν όλους τους άλλους τύπους δεδομένων DNA-αλληλουχιών μαζί[7].

Το συμπλήρωμα του γονιδίου του συναρμολογημένου μικροβιακού πλαγκτόν της Θάλασσας των Σαργασσών περιλάμβανε 1412 ξεχωριστά γονίδια ριβοσωμικού RNA -ένα χρήσιμο μέτρο σύγκρισης για ταξινομική βαθμονόμηση. Τα χτυπήματα (counts) πρωτεϊνών που χρησιμεύουν ως ταξινομικοί δείκτες, χρησιμοποιήθηκαν για να εκτιμηθεί ο πλούτος των ειδών. Οι πρωτεΐνες έδειξαν ότι υπήρχαν περίπου 1800 είδη (όπως ορίζεται συνήθως) στο δείγμα, γεγονός το οποίο βρίσκεται σε συμφωνία με το σύνολο των rRNA χτυπημάτων. Οι τύποι των μικροβίων που εντοπίστηκαν ήταν γενικά σύμφωνοι με αυτούς που είναι γνωστό ότι είναι διαδεδομένοι στον ωκεανό (με ορισμένες εξαιρέσεις), πάνω στην βάση ερευνών ανεξάρτητων της

δημιουργίας καλλιεργειών, που διεξήχθησαν στη θάλασσα κατά την τελευταία δεκαετία. Ένα σοβαρό θέμα με την αρχική μεταγονιδιωματική προσπάθεια μελέτης της Θάλασσας των Σαργασσών ήταν η μικροβιακή μόλυνση που φαίνεται ότι είχε θέσει σε κίνδυνο ένα μεγάλο μέρος του δείγματος, περιορίζοντας σημαντικά τη χρησιμότητά της για οικολογικές ερμηνείες, αφού είχε προκληθεί μόλυνση με μη αυτόχθονα μικροβιακά γονίδια. Αυτό το ατυχές αποτέλεσμα δείχνει σαφώς ότι οι μελέτες μεταγονιδιωματικής απαιτούν πολύ πιο ολοκληρωμένες προσπάθειες, σε αντίθεση με την απλοϊκή κλωνοποίηση και αλληλούχιση τυχαίων περιβαλλοντικών δειγμάτων. Οι προσεκτικές δειγματοληψίες, οι διαδικασίες επαλήθευσης, ο συντονισμός με εμπειρογνώμονες του κάθε τομέα και η ανεξάρτητη επικύρωση του δείγματος είναι προαπαιτούμενα για μαζικές προσπάθειες μεταγονιδιωματικής αλληλούχισης. Η βαθιά γνώση του περιβάλλοντος της δειγματοληψίας, η εμπειρία σε σχέση με τους τύπους και τις κατανομές των αυτοχθόνων μικροβίων και οι ανεξάρτητες μέθοδοι επικύρωσης του δείγματος, θα διευκολύνουν, ιδιαίτερα αυστηρές επιστημονικά μεταγονιδιωματικές μελέτες. Δεξιότητες και γνώσεις από ένα ευρύ φάσμα επιστημονικών κλάδων συμπεριλαμβανομένης της περιβαλλοντικής επιστήμης, της μικροβιολογίας, της μοριακής βιολογίας, της βιοπληροφορικής, των μαθηματικών, της βιοχημείας, της φυσιολογίας και της οικολογίας, είναι όλες αναγκαίες για την ορθή τη συλλογή και ερμηνεία συνόλων δεδομένων μεταγονιδιωματικής.

Παρά την πολυπλοκότητα του γονιδιώματος των αυτοχθόνων οργανισμών, τα πιθανά προβλήματα δειγματοληψίας και τις αναλυτικές προκλήσεις, τα δεδομένα από τη Θάλασσα των Σαργασσών έχουν ήδη αποδειχθεί ένας χρήσιμος πόρος[7]. Ο Venter και οι συνεργάτες του δημιούργησαν 1.045 δισεκατομμύρια bps (ζεύγη βάσεων) από ακολουθίες περισσότερων από 1800 διαφορετικών ειδών βακτηρίων, συμπεριλαμβανομένων 148 πρωτότυπων βακτηριακών φυλοτύπων. Τα δείγματα αυτά αντιπροσωπεύουν πάνω από 1.2 εκατομμύρια νέες γονιδιακές αλληλουχίες με 782 βακτηριακά ομόλογα της πρωτεοροδοψίνης[6]. Οι ροδοψίνες, πρωτεΐνες που σχετίζονται με την δέσμευση του φωτός είχαν από καιρό βρεθεί σε καλλιεργημένα Αρχαία, μια από τις τρεις μεγάλες κατηγορίες μικροοργανισμών. Η ανάλυση μεταγονιδιωματικής, αποκάλυψε πως υπάρχει μια μεγάλη ποικιλομορφία γονιδίων παραγωγής ροδοψίνης που πολλά από αυτά προέρχονται ξεκάθαρα από βακτήρια[2], αν και όπως ειπώθηκε, ο Delong επεσήμανε ότι τα δύο γονιδιώματα που αναφέρθηκαν από τον Venter και τους συνεργάτες του ήταν πιθανώς προσμείξεις στο δείγμα θαλασσινού νερού (π.χ., contigs που περιέχουν τα βακτηριακά γονίδια 5S και 23S rRNA δίπλα σε ένα γονίδιο 16S rRNA από Αρχαία), υποδεικνύοντας έτσι ότι πρέπει να δίνεται τεράστια προσοχή κατά τη διεξαγωγή αναλύσεων μικροβιακής μεταγονιδιωματικής[6].

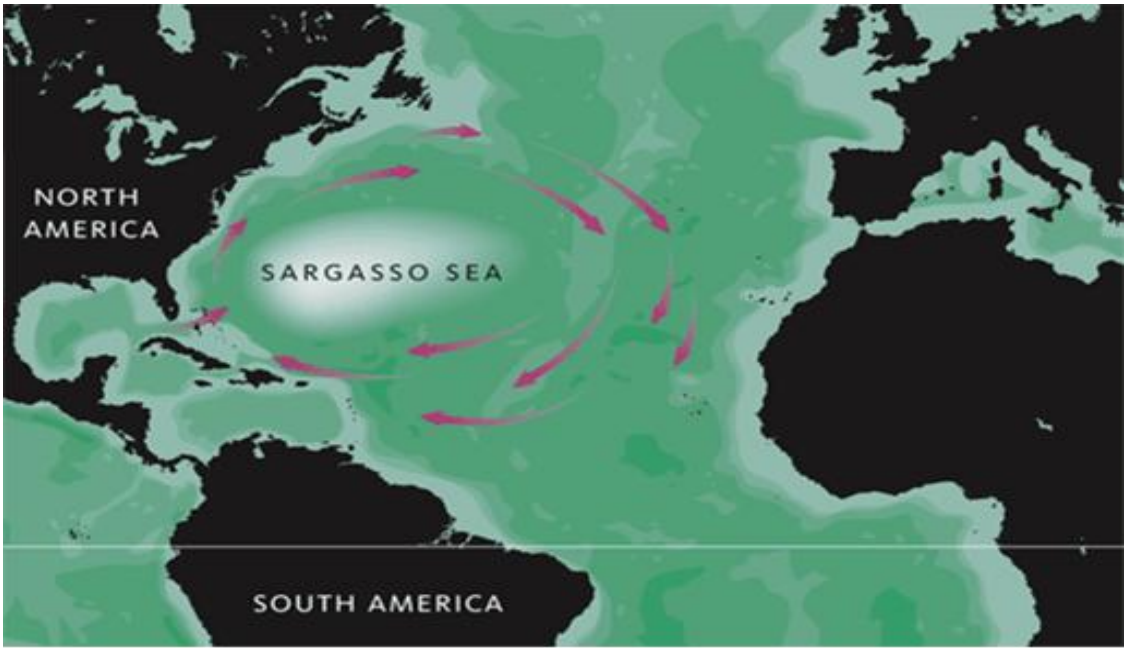
Μεταξύ των προηγουμένως ανακαλυφθέντων γονιδίων και πρωτεϊνών, όπως φωτοπρωτεΐνες σαν τις πρωτεοροδοψίνες, τα δεδομένα της Θάλασσας των

Σαργασσών αποκάλυψαν νέες παραλλαγές σε ένα γενικό μοτίβο. Η αξία του συνόλου δεδομένων αποδεικνύεται από τον μεγάλο αριθμό εγγράφων που έχουν εκμεταλλευτεί τις ταξινομικές πληροφορίες του, αναλύοντας αλληλουχίες νέων γονιδίων ή ακόμη και μέσω της συνθετικής παραγωγής και χαρακτηρισμού γονιδίων και γονιδιακών προϊόντων που δεν είχαν ποτέ πριν μελετηθεί. Οι λεπτομερείς μελέτες συνολικού γονιδιώματος ως προς τη γονιδιωματική μεταβλητότητα, τη δομική οργάνωση και την εξέλιξη ταξινομικών ομάδων, συμπεριλαμβανομένου παραδείγματος χάριν του Prochlorococcus, έχουν σε μεγάλο βαθμό ωφεληθεί από αυτά τα νέα δεδομένα. Οι θεωρητικές και πρακτικές ανακαλύψεις και εφαρμογές που ήδη προκύπτουν από αυτό το μοναδικό σύνολο δεδομένων παρέχουν άφθονες αποδείξεις για την αξία των μεγάλης κλίμακας γονιδιωματικών αναλύσεων της συνολικής μικροβιακής κοινότητας[7] .

Σε αντίθεση με την συγκρότηση του πλήρους γονιδιώματος, συγκριτικές αναλύσεις του συνόλου των δεδομένων της Θάλασσας των Σαργασσών και άλλων συνόλων δεδομένων έχουν αποδείξει τη χρησιμότητα των συγκρίσεων ενός γονιδίου εντός αλλά και μεταξύ των δειγμάτων. Μια νέα προσέγγιση στην μικροβιακή οικολογία, η συγκριτική γονιδιωματική κοινοτήτων, αναδύεται από τέτοιες μελέτες. Μια πρόσφατη μελέτη, για παράδειγμα, συνέκρινε, γονίδιο-προς-γονίδιο, ομοιότητες αλλά και διαφορές μεταξύ συνόλων δεδομένων γονιδιακής αλληλουχίας των κοινοτήτων από τη Θάλασσα των Σαργασσών, του κουφαριού μιας φάλαινας του βυθού και ορυχείων όξινης απορροής. Με τη λήψη μιας "γονιδιοκεντρικής" προσέγγισης, σε αντίθεση με μια προσπάθεια προσέγγισης γονιδιωματικής συναρμολόγησης, ήταν δυνατή η σύγκριση των μοτίβων εμφάνισης συγκεκριμένων κατηγοριών γονιδίων και η συναρμολόγηση προφίλ κοινοτήτων που περιέχουν λειτουργικά γονίδια. Η υπερεκπροσώπηση συγκεκριμένων κατηγοριών " περιβαλλοντικών ετικετών γονιδίων " (EGTs) σε διαφορετικά δείγματα (για παράδειγμα, η δυσανάλογη εκπροσώπηση των φωτοσυνθετικών γονιδίων και ροδοψινών στο δείγμα της Θάλασσας των Σαργασσών) επαλήθευσε τη χρησιμότητα της προσέγγισης για την κατανόηση μεταβολικών λειτουργιών που σχετίζονται με συγκεκριμένες μικροβιακές κοινότητες. Η συνετή δειγματοληψία μπορεί να διευκολύνει σε μεγάλο βαθμό τέτοιες συγκρίσεις, επιτρέποντας τη σύγκριση των κοινοτήτων κατά μήκος καλά επικυρωμένων περιβαλλοντικών επιπέδων. Μια άλλη πρόσφατη μελέτη στον Ειρηνικό Ωκεανό προχώρησε στη σύγκριση των μικροβιακών βιοκοινοτήτων κατά επίπεδα βάθους, από επιφανειακά ύδατα έως και βάθος 4 χιλιομέτρων. Γονιδιωματικά χαρακτηριστικά που σχετίζονταν με το περιβάλλον ήταν εμφανή και υποδείκνυαν ειδικά λειτουργικά και εξελικτικά μοτίβα εξαρτώμενα από το βάθος, στις μικροβιακές κοινότητες και τα γονιδιώματα. Αυτές οι πρόσφατες εργασίες αναδεικνύουν τη μελλοντική υπόσχεση και χρησιμότητα των συγκριτικών γονιδιωματικών μελετών των κοινοτήτων[7].

Λοιπές μελέτες: Η οξείδωση της αμμωνίας είναι το πρώτο βήμα στη νιτροποίηση, μια βασική διαδικασία στον παγκόσμιο κύκλο του αζώτου που έχει σαν αποτέλεσμα τον σχηματισμό του νιτρικού μέσω της μικροβιακής δραστηριότητας. Τα αυτότροφα βακτήρια οξείδωσης της αμμωνίας (AOB) των β-και γ-υποομάδων των Πρωτεοβακτηρίων είναι οι πιο σημαντικοί παράγοντες για την αερόβια οξείδωση της αμμωνίας. Γονίδια που κωδικοποιούν υπομονάδες μιας μονοοξυγενάσης της αμμωνίας (AMO), που είναι το ένζυμο-κλειδί της AOB, βρέθηκαν σε ένα μεταγονιδιωματικό κλώνο του εδάφους δίπλα σε οπερονίο rRNA από Αρχαία, που σχετίζεται με το φύλο Crenarchaeota .Η αφθονία του γονιδίου που κωδικοποιεί μια υπομονάδα του ενζύμου κλειδιού AMO (amoA) ερευνήθηκε σε 12 παρθένα γεωργικά εδάφη διαφορετικών κλιματολογικών ζωνών[6] .

Βακτηρίδια που είναι δυνατόν να προκαλέσουν ξενοβιοτική υποβάθμιση έχουν απομονωθεί από μικροβιακά στρώματα προερχόμενα από τη Μαύρη Θάλασσα μέσω μεταγονιδιωματικής προσέγγισης και ελήφθησαν όλα τα γονίδια που απαιτούνται για την πλήρη αποικοδόμηση του βενζοϊκού οξέος παρέχοντας την πρώτη ένδειξη της ικανότητας τους για την εν λόγω υποβάθμιση. Από μεταγονιδιωματικές βιβλιοθήκες εκροών επεξεργασίας ιλύος, ο Chauhan και οι συνεργάτες του έχουν προσδιορίσει γονίδια και ένα μηχανισμό αντίστασης σε αρσενικό που μπορεί να αποδειχθούν χρήσιμα στο σχεδιασμό στρατηγικών καλύτερης απομάκρυνσης του αρσενικού από το πόσιμο νερό[6].Τέλος, λεπτομερείς εξετάσεις υιικών μεταγονιδιωμάτων έχουν γίνει από διάφορους μελετητές, ώστε να παρατηρηθεί η γενετική ποικιλομορφία και να κερδιθεί η κατανόηση της οικολογικής και εξελικτικής σημασίας των ιογενών συναθροίσεων.

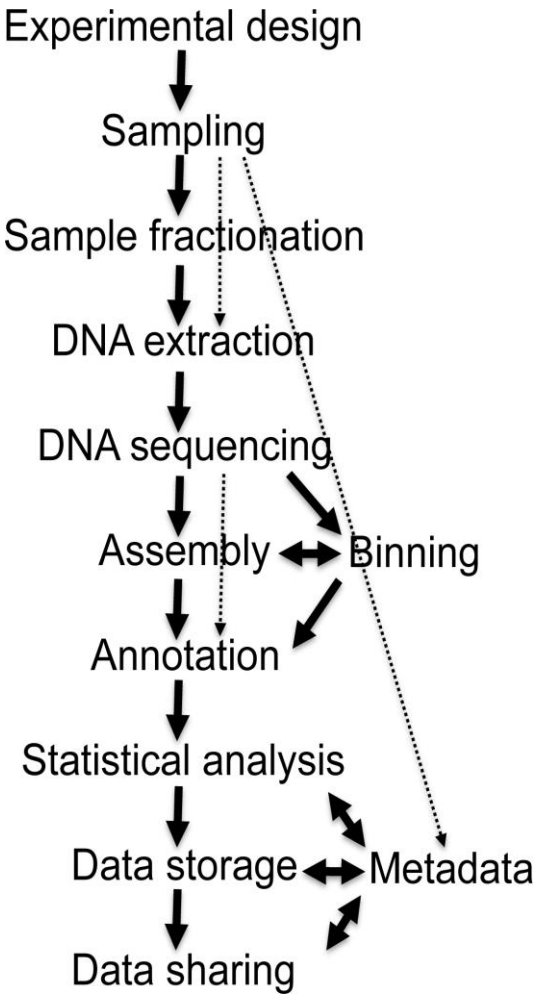


Εικόνα 5 Η Θάλασσα των Σαργασσών.

2. ΟΔΗΓΟΣ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕΤΑΓΟΝΙΔΙΩΜΑΤΙΚΩΝ ΑΝΑΛΥΣΕΩΝ

2.1 Μια τυπική μεταγονιδιωματική ανάλυση αλληλούχισης

Η μεταγονιδιωματική εφαρμόζει μια σειρά από τεχνολογίες γονιδιωματικής και εργαλείων βιοπληροφορικής για την άμεση πρόσβαση στο γενετικό περιεχόμενο ολόκληρων κοινοτήτων οργανισμών. Ο τομέας της μεταγονιδιωματικής ήταν υπεύθυνος για την ουσιαστική πρόοδο στην μικροβιακή οικολογία, την εξέλιξη και την ποικιλομορφία, κατά τα τελευταία 5 έως 10 χρόνια και πολλά ερευνητικά εργαστήρια συμμετέχουν πλέον ενεργά σε αυτόν. Με τον αυξανόμενο αριθμό των δραστηριοτήτων, έρχεται επίσης μια πληθώρα μεθοδολογικών γνώσεων και εμπειριών που πρέπει να διέπουν τις μελλοντικές εξελίξεις στον τομέα.



Εικόνα 6 Διάγραμμα ροής μιας τυπικής μεταγονιδιωματικής ανάλυσης. Τα διακεκομμένα βέλη δείχνουν βήματα που μπορούν να παραλειφθούν[11].

Η μεταγονιδιωματική ανάλυση αρχικά ξεκίνησε σαφέστατα με την κλωνοποίηση περιβαλλοντικού DNA και την εν συνεχεία μελέτη της λειτουργικής έκφρασης, ενώ γρήγορα συμπληρώθηκε με άμεση τυχαία shotgun αλληλούχιση περιβαλλοντικού DNA. Αυτά τα αρχικά πρότζεκτ, έδειξαν όχι μόνο την απόδειξη της αρχής της μεταγονιδιωματικής προσέγγισης, αλλά επίσης αποκάλυψαν μια τεράστια ποικιλομορφία λειτουργικών γονιδίων στο μικροβιακό κόσμο[11].

Η ταχύτατη αλλά και ουσιαστική μείωση του κόστους στις τεχνικές της επόμενης γενιάς αλληλούχισης έχει αυξήσει δραματικά τις δυνατότητες ανάπτυξης της μεταγονιδιωματικής ανάλυσης που βασίζεται στις αλληλουχίες. Στην πραγματικότητα, ο αριθμός των συνόλων δεδομένων μεταγονιδίωματος που προέρχεται από shotgun τεχνικές ανάλυσης αλληλουχίας έχει αυξηθεί υπερβολικά κατά τα τελευταία χρόνια. Στο μέλλον, η μεταγονιδιωματική θα χρησιμοποιηθεί με τον ίδιο τρόπο όπως οι μέθοδοι γονιδιακής εξακρίβωσης δακτυλικών αποτυπωμάτων 16S rRNA ,για την περιγραφή μικροβιακών προφίλ

κοινοτήτων. Ως εκ τούτου, θα αποτελεί πρότυπο εργαλείο για πολλά εργαστήρια και επιστήμονες που εργάζονται στον τομέα της μικροβιακής οικολογίας[11].

Είναι λοιπόν απαραίτητο να δοθεί μια επισκόπηση του πεδίου της μεταγονιδιωματικής, δίνοντας ιδιαίτερη έμφαση στα βήματα που εμπλέκονται σε μια τυπική μεταγονιδιωματική ανάλυση αλληλούχισης. Περιγράφεται και αναλύεται λεπτομερέστερα, όπου αυτό είναι απαραίτητο με βάση το περιεχόμενο της διπλωματικής αυτής εργασίας, η διαδικασία επιλογής του δείγματος επεξεργασίας, οι τεχνικές αλληλούχισης, η συναρμολόγηση, η αντιστοίχιση αλληλουχιών σε οργανισμούς, καθώς και ο χαρακτηρισμός των δεδομένων που προκύπτουν (Εικόνα 6). Σαφώς, κάθε είδους μεταγονιδιωματικά σύνολα δεδομένων θα επωφεληθούν από τις πλούσιες διαθέσιμες πληροφορίες από άλλα μεταγονιδιωματικά πρότζεκτ και υπάρχει η πίστη, ότι τα κοινά, αλλά και ευέλικτα πρότυπα και αλληλεπιδράσεις μεταξύ των επιστημόνων του τομέα αυτού, θα διευκολύνουν την ανταλλαγή πληροφοριών.

2.2 Δειγματοληψία και προεπεξεργασία (Sampling and Preprocessing)

Η επεξεργασία των δειγμάτων είναι το πρώτο και πιο κρίσιμο βήμα σε κάθε μεταγονιδιωματική ανάλυση. Το DNA που εξάγεται πρέπει να είναι αντιπροσωπευτικό όλων των κυττάρων που υπάρχουν στο δείγμα και θα πρέπει να λαμβάνονται επαρκείς ποσότητες από υψηλής ποιότητας νουκλεϊκά οξέα για την παραγωγή της μεταγονιδιωματικής βιβλιοθήκης και μετέπειτα την αλληλούχιση. Η επεξεργασία απαιτεί ειδικά πρωτόκολλα για κάθε είδος δείγματος και διάφορες ισχυρές μέθοδοι εξαγωγής DNA είναι διαθέσιμες[11]. Έχουν αναπτυχθεί διάφορα πρωτόκολλα για την εξαγωγή του DNA από το έδαφος και τις υδάτινες πηγές για την κατασκευή μεταγονιδιωματικών βιβλιοθηκών με στόχο την υψηλή ανάκτηση, την αποτελεσματικότητα και την καταλληλότητα για την περαιτέρω μοριακή ανάλυση[6]. Λαμβάνονται πλέον και πρωτοβουλίες ώστε να εξερευνάται η μικροβιακή βιοποικιλότητα από δεκάδες χιλιάδες οικοσυστήματα χρησιμοποιώντας μια μόνο είδους τεχνολογία εξαγωγής για να διασφαλιστεί η δυνατότητα σύγκρισης[11].

Η σύνθεση της εκάστοτε μικροβιακής κοινότητας επηρεάζει κατά πολύ τους τύπους αναλύσεων που μπορούν να εκτελεστούν σε ένα μεταγονιδιωματικό σύνολο δεδομένων. Οι μικροβιακές κοινότητες περιλαμβάνουν συνδυασμούς βακτηρίων, Αρχαίων, καθώς και μικροβιακά ευκαρυωτικά κύτταρα και οι ιούς, συχνά και με τις τέσσερις ομάδες να συμβιώνουν σε ένα ενιαίο περιβάλλον[12].

Με βάση την τρέχουσα ικανότητα αλληλούχισης και παρά τη ραγδαία εξέλιξη των τεχνικών αυτής, η μεταγονιδιωματική αλληλούχιση κοινοτήτων που περιέχουν ευκαρυωτικά κύτταρα, ιδίως πρώτιστα, είναι ως επί το πλείστον οικονομικά

απαγορευτική λόγω του τεράστιου μεγέθους του γονιδιώματος τους και τη χαμηλή πυκνότητα σε γονίδια που κωδικοποιούν πρωτεΐνες. Ως εκ τούτου, η επιλογή μιας κοινότητας που δεν περιέχει ευκαρυωτικά κύτταρα, ή μιας κοινότητας από την οποία μπορούν να αποκλειστούν οι ευκαρυώτες ή το DNA τους, αποτελεί σημαντική παράμετρο εξέτασης πριν από την έναρξη μιας μεταγονιδιωματικής ανάλυσης. Όταν επιχειρείται αλληλούχιση μικροβιακών κοινοτήτων που βρίσκονται σε στενές συμβιωτικές σχέσεις με ευκαρυωτικούς ξενιστές, η απομάκρυνση των κυττάρων-ξενιστών ή του εξαγόμενου DNA του ξενιστή, είναι σημαντική ώστε να αποφευχθεί ευκαρυωτική μόλυνση.

Το να απομακρύνονται απλά τα ευκαρυωτικά κύτταρα από μια μεταγονιδιωματική ανάλυση δεν είναι ιδανικό από οικολογικής άποψης, καθώς θέτει σε κίνδυνο την ικανότητα πρόσβασης και αξιολόγησης μιας μικροβιακής κοινότητας στο σύνολό της. Μια εναλλακτική ή συμπληρωματική στρατηγική θα μπορούσε να είναι η απόκτηση μοριακών δεδομένων σε RNA (metatranscriptomics) ή πρωτεϊνικό (metaproteomics) επίπεδο, παρακάμπτοντας έτσι το πρόβλημα των μεγάλων ποσοτήτων μη κωδικών ευκαρυωτικών δεδομένων αλληλουχίας[12].

Εάν η κοινότητα στόχος συνδέεται με κάποιον ξενιστή (π.χ. ένα ασπόνδυλο ή ένα φυτό), τότε είτε η κλασμάτωση είτε η επιλεκτική λύση θα μπορούσε να είναι κατάλληλη για να εξασφαλιστεί ελάχιστη λήψη DNA του ξενιστή. Αυτό είναι ιδιαίτερα σημαντικό, κυρίως όταν το γονιδίωμα του ξενιστή είναι μεγάλο και ως εκ τούτου θα μπορούσε να υπερκαλύψει τις ακολουθίες της μικροβιακής κοινότητας στην επακόλουθη προσπάθεια προσδιορισμού της αλληλουχίας. Η φυσική κλασμάτωση μπορεί επίσης να εφαρμοστεί όταν ο στόχος της ανάλυσης είναι μόνο ένα ορισμένο μέρος της κοινότητας, για παράδειγμα, κατά την μελέτη ιών σε δείγματα θαλασσινού νερού. Στην περίπτωση αυτή μια σειρά από επιλεκτικά φιλτραρίσματα ή στάδια φυγοκέντρησης, ή ακόμη και κυτταρομετρία ροής, μπορούν να χρησιμοποιηθούν για τον εμπλουτισμό του κλάσματος στόχου. Τα βήματα της κλασμάτωσης θα πρέπει να ελέγχονται για να διασφαλιστεί ότι επιτυγχάνεται επαρκής εμπλουτισμός του στόχου και ότι συμβαίνει ελάχιστη μόλυνση του υλικού από μη στοχευόμενα τμήματα[11].

Ο φυσικός διαχωρισμός και η απομόνωση των κυττάρων από το δείγμα θα μπορούσαν επίσης να είναι σημαντικές παράμετροι για τη μεγιστοποίηση απόδοσης DNA ή την αποφυγή συνεξαγωγής ενζυμικών αναστολέων (όπως χουμικά οξέα) που θα μπορούσαν να παρεμποδίσουν τη μετέπειτα επεξεργασία. Η κατάσταση αυτή είναι ιδιαίτερα σημαντική για τις μεταγονιδιωματικές αναλύσεις εδάφους και έχει γίνει σημαντικό έργο στον τομέα αυτό για την αντιμετώπιση του ζητήματος. Η άμεση λύση των κυττάρων στη μήτρα του εδάφους έναντι της έμμεσης λύσης (δηλαδή μετά τον διαχωρισμό των κυττάρων από το έδαφος)

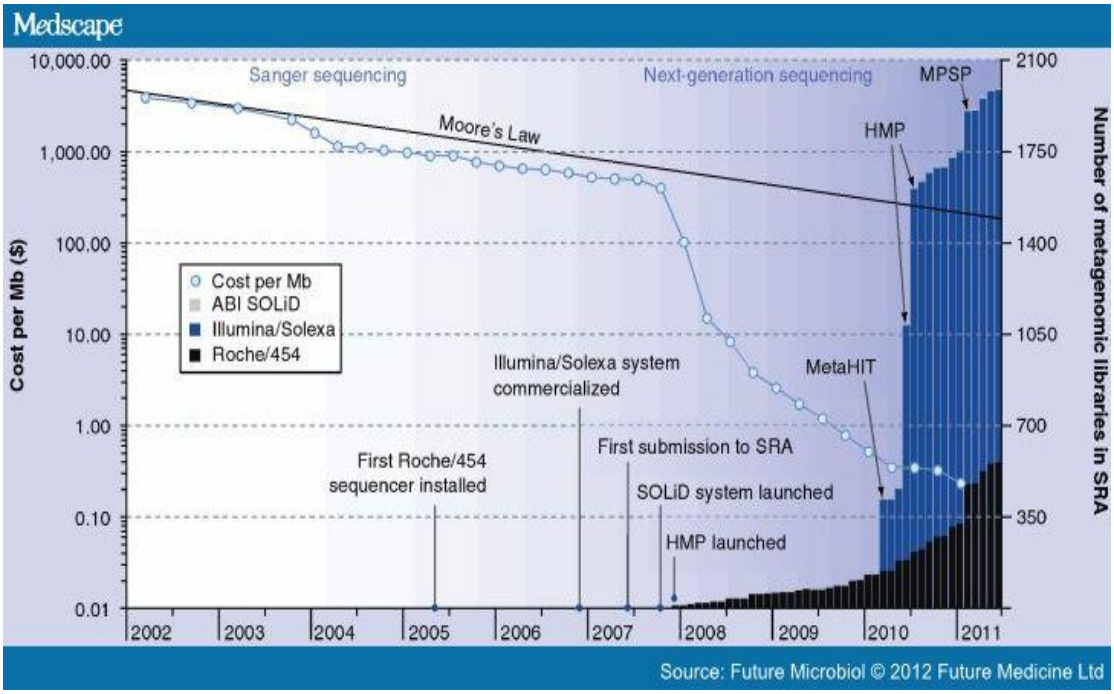
παρουσιάζει μια μετρήσιμη "προδιάθεση" από την άποψη της μικροβιακής ποικιλομορφίας, της απόδοσης DNA και του μήκους θραύσματος της προκύπτουσας ακολουθίας. Η εκτεταμένη μελέτη του εδάφους υπογραμμίζει την ανάγκη να εξασφαλιστεί η πρότυπη λειτουργία για τις διαδικασίες εξόρυξης και ότι απαιτείται σύγκριση πολλαπλών μεθόδων για να εξασφαλιστεί αντιπροσωπευτική εξαγωγή του DNA.

Ορισμένοι τύποι δειγμάτων (όπως βιοψίες ή υπόγεια ύδατα) αποδίδουν συχνά πολύ μικρές ποσότητες DNA. Η παραγωγή βιβλιοθηκών, για τις περισσότερες τεχνολογίες αλληλούχισης απαιτεί υψηλές ποσότητες νανογραμμαρίων ή μικρογραμμαρίων DNA και ως εκ τούτου ίσως απαιτηθεί ενίσχυση του αρχικού υλικού. Η πολλαπλή ενίσχυση δια της μεθόδου MDA (multiple displacement amplification) με τη χρήση τυχαίων εξαμερών και phi29 πολυμεράσης φάγων, είναι μία επιλογή που χρησιμοποιείται για να αυξήσει τις αποδόσεις DNA. Η μέθοδος αυτή μπορεί να ενισχύσει φεμτογραμμάρια του DNA για την παραγωγή μικρογραμμαρίων προϊόντος και ως εκ τούτου έχει χρησιμοποιηθεί ευρέως στον τομέα της γονιδιωματικής μονών κυττάρων και σε ένα ορισμένο βαθμό στη μεταγονιδιωματική. Όπως και με κάθε ενισχυτική μέθοδο, υπάρχουν πιθανά προβλήματα που σχετίζονται με μολύνσεις του αντιδραστηρίου, το σχηματισμό χιμαιρικών μορίων και υπερεντοπισμό συγκεκριμένων ακολουθιών στην ενίσχυση και ο αντίκτυπός τους θα εξαρτάται από την ποσότητα και τον τύπο του αρχικού υλικού και τον απαιτούμενο αριθμό κύκλων ενίσχυσης για την παραγωγή επαρκών ποσοτήτων νουκλεϊκών οξέων. Τα ζητήματα αυτά μπορούν να έχουν σημαντική επίπτωση στην ακόλουθη μεταγονιδιωματική ανάλυση της κοινότητας και έτσι είναι αναγκαίο να εξεταστεί κατά πόσον η ενίσχυση είναι επιτρεπτή.

Η ενίσχυση σε ολόκληρο το γονιδίωμα έχει επίσης χρησιμοποιηθεί σε περιπτώσεις μικρών αποδόσεων περιβαλλοντικού DNA ώστε να προκύψουν ποσότητες μικρογραμμαρίων αρκετές για αλληλούχιση. Ένα σημαντικό πλεονέκτημα αυτής της τεχνικής είναι ότι μπορεί να επεξεργαστεί και να διατηρήσει μονόκλωνο DNA, το οποίο είναι ανεκτίμητο χαρακτηριστικό για την μελέτη δειγμάτων ιων. Ωστόσο, η σχετική αντιπροσώπευση των γονιδιωματικών DNAs μπορεί να τίθεται σε κίνδυνο, ιδίως εάν η ποσότητα του αρχικού υλικού είναι μικρή. Αυτό είναι σημαντικό να λαμβάνεται υπ' όψιν σε συγκριτικές αναλύσεις, ιδίως μεταξύ δειγμάτων που χρησιμοποιήθηκε ενίσχυση ολόκληρου του γονιδιώματος και άλλων που δεν έγινε κάτι τέτοιο[12].

2.3 Αλληλούχιση (Sequencing)

Κατά τα τελευταία 10 χρόνια, η shotgun μεταγονιδιωματική ανάλυση αλληλούχισης μετατοπίστηκε σταδιακά από την κλασική αλληλούχιση τεχνολογίας Sanger προς τις τεχνολογίες επόμενης γενιάς αλληλούχισης (NGS). Από τις NGS τεχνολογίες, τόσο η 454/Roche όσο και το σύστημα Illumina / Solexa εμφανίζουν πλέον εκτεταμένη εφαρμογή σε μεταγονιδιωματικά δείγματα, ενώ όπως θα αναφερθεί παρακάτω όλο και περισσότερες τεχνολογίες κάνουν την εμφάνιση τους με την πάροδο των χρόνων.



Εικόνα 7 Επισκόπηση του κόστους αλληλούχισης και του αριθμού των μεταγονιδιωματικών βιβλιοθηκών που έχουν υποβληθεί στο Sequence Read Archive. Το κόστος της αλληλουχίας βασίζεται σε στοιχεία που παρέχονται από το αμερικανικό National Human Genome Research Institute, απεικονίζοντας την περισσότερο από λογαριθμική μείωση και ξαφνική αλλαγή, όταν τα κέντρα αλληλούχισης μετέβησαν στις επόμενης γενιάς τεχνολογίες αλληλούχισης[9].

Τεχνολογία αλληλούχισης Sanger : Ιστορικά, το 1975, ο Sanger εισήγαγε την έννοια της μεθόδου προσδιορισμού αλληλουχίας DNA και αργότερα, δημοσίευσε μία ταχεία μέθοδο για τον προσδιορισμό αλληλουχιών στο DNA μέσω σύνθεσης με DNA πολυμεράση. Κατά το έτος 1977, δημοσιεύτηκαν δύο άρθρα ορόσημα για την αλληλούχιση του DNA, δηλαδή, η ενζυματική διδεοξυ τεχνική αλληλούχισης DNA του Sanger με βάση τον τερματισμό της αλυσίδας από διδεοξυνουκλεοτιδικά ανάλογα και το άρθρο των Allan Maxam και Walter Gilbert, περί της τεχνικής αλληλούχισης του DNA μέσω χημικής αποδόμησης στην οποία τερματικώς επισημασμένα θραύσματα DNA διασπώνται χημικώς σε συγκεκριμένες βάσεις και διαχωρίζονται με ηλεκτροφόρηση σε πήκτωμα. Αυτά τα δύο εξέχοντα εργαστήρια ήταν υπεύθυνα για την πρόταση του πρώτου αυτοματοποιημένου DNA αναλυτή

αλληλούχισης, που στην συνέχεια εισήχθηκε στο εμπόριο από την Applied Biosystems (ABI), το Ευρωπαϊκό Εργαστήριο Μοριακής Βιολογίας (EMBL) και τη Pharmacia-Amersham, αργότερα γνωστή ως General Electric (GE). Αυτό το ραφινάρισμα και η δεδομένη εμπορευματοποίηση της μεθόδου προσδιορισμού αλληλουχίας οδήγησε στην ευρεία διάδοσή της σε όλη την παγκόσμια ερευνητική κοινότητα[13].

Από την πρώτη έκθεσή της το 1977, η μέθοδος προσδιορισμού αλληλουχίας Sanger έχει παραμείνει εννοιολογικά αμετάβλητη. Η μέθοδος βασίζεται στην εξαρτώμενη από την DNA πολυμεράση σύνθεση μιας συμπληρωματικής έλικας DNA υπό την παρουσία φυσικών 2'-δεοξυνουκλεοτιδίων (dNTPs) και 2 ', 3' διδεοξυνουκλεοτιδίων (ddNTPs), που χρησιμεύουν ως μη αναστρέψιμοι τερματιστές της σύνθεσης. Η αντίδραση σύνθεσης του DNA τερματίζεται τυχαία κάθε φορά που ένα ddNTP προστίθεται στην αλυσίδα του αναπτυσσόμενου ολιγονουκλεοτιδίου, καταλήγοντας σε περικομμένα προϊόντα ποικίλου μήκους με ένα κατάλληλο ddNTP στο 3 'άκρο τους. Τα προϊόντα διαχωρίζονται με βάση το μέγεθος χρησιμοποιώντας ηλεκτροφόρηση σε πήκτωμα πολυακρυλαμίδης και τα τερματικά ddNTPs χρησιμοποιούνται για να αποκαλυφθεί η DNA αλληλουχία του προτύπου κλώνου[14].

Αρχικά, τέσσερις διαφορετικές αντιδράσεις απαιτούνταν ανά πρότυπο, με κάθε αντίδραση να περιέχει ένα διαφορετικό τερματιστή ddNTP, ddATP, ddCTP, ddTTP ή ddGTP. Ωστόσο, η πρόοδος στην ανίχνευση φθορισμού επιτρέπει το συνδυασμό των τεσσάρων τερματιστών σε μία αντίδραση με το να έχει επισημανθεί ο καθένας με φθορίζουσες χρωστικές ουσίες διαφορετικών χρωμάτων. Μεταγενέστερες εξελίξεις έχουν αντικαταστήσει την αρχική ηλεκτροφόρηση πήγματος πλάκας με τριχοειδή ηλεκτροφόρηση πηκτώματος, επιτρέποντας έτσι σε πολύ ισχυρότερα ηλεκτρικά πεδία να εφαρμοστούν στην μήτρα διαχωρισμού. Ένα από τα αποτελέσματα αυτής της αλλαγής, ήταν να αυξηθεί ο ρυθμός με τον οποίο θα μπορούσε να δημιουργούνται τα θραύσματα διαχωρισμού. Η συνολική απόδοση της τεχνικής της τριχοειδούς ηλεκτροφόρησης αυξήθηκε περαιτέρω κατά την έλευση των τριχοειδών διατάξεων (arrays) όπου πολλά δείγματα θα μπορούσαν να αναλυθούν παράλληλα. Επιπλέον, σημαντικές ανακαλύψεις στη βιοχημεία πολυμερών, συμπεριλαμβανομένης της ανάπτυξης των γραμμικών πολυ-ακρυλαμιδίων αλλά και της πολυ-διμεθυλακρυλαμίδης επέτρεψαν την επαναχρησιμοποίηση των τριχοειδών αγγείων σε πολλαπλά ηλεκτροφορητικά τρεξίματα, προκαλώντας έτσι αύξηση της αποδοτικότητας της αλληλούχισης.

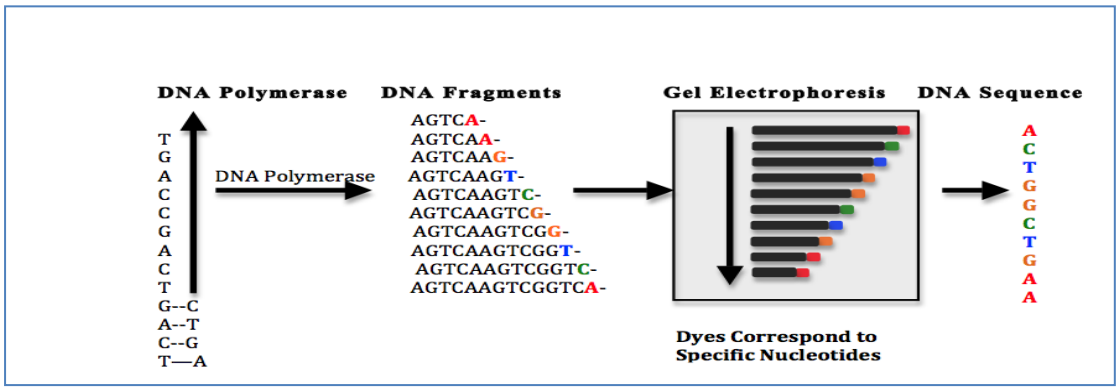
Αυτές και πολλές άλλες εξελίξεις στην τεχνολογία αλληλούχισης συνέβαλαν στο σχετικά χαμηλό ποσοστό σφάλματος, τη δυνατότητα επεξεργασίας αλληλουχιών μεγάλου μήκους και τα διάφορα ισχυρά χαρακτηριστικά των σύγχρονων αναλυτών αλληλούχισης Sanger. Για παράδειγμα, ένα συχνά χρησιμοποιούμενο μηχάνημα

αυτοματοποιημένης αλληλούχισης Sanger υψηλής απόδοσης από την Applied Biosystems, το 3730xl ABI, έχει 96 - τριχοειδούς μορφής συστοιχίες και είναι ικανό να παράγει 900 ή περισσότερα PHRED 20 bp (ζεύγη βάσεων) ανά read αλληλουχίας για ένα σύνολο έως και 96 kb, σε 3ωρο τρέξιμο. Ωστόσο, παρά τις μεγάλες προόδους στις χημικές βιομηχανίες και τις πολύ δυνατές επιδόσεις οργάνων, όπως το 3730xl, η εφαρμογή των σχετικά ακριβών τεχνικών ανάλυσης αλληλουχίας Sanger σε μεγάλα έργα μαζικής αλληλούχισης παραμένει έξω από τα προγράμματα τυπικής χρηματοδότησης μέσω επιχορήγησης[14].

Technology	Approach	Read length	Bp per run
Automated Sanger sequencer ABI3730xl	Synthesis in the presence of dye terminators	Up to 900 bp	96 kb
454/Roche FLX system	Pyrosequencing on solid support	200–300 bp	80–120 Mb
Illumina/ Solexa	Sequencing by synthesis with reversible terminators	30–40 bp	1 Gb
ABI/SOLiD	Massively parallel sequencing by ligation	35 bp	1–3 Gb

Πίνακας 3 Συγκριτικά χαρακτηριστικά τεχνολογιών αλληλούχισης. Sanger sequencing και NGS τεχνολογίες[14].

Η τεχνολογία αλληλούχισης Sanger πάντως, εξακολουθεί να θεωρείται το χρυσό πρότυπο για την αλληλούχιση, λόγω του χαμηλού ποσοστού σφάλματος της, καθώς και της δυνατότητας επεξεργασίας αλληλουχιών μεγάλου μήκους (> 700 bp) και την εισαγωγή μεγάλου μεγέθους ενθεμάτων (π.χ. > 30 Kb για φοσμίδια ή βακτηριακά τεχνητά χρωμοσώματα (BACs)). Όλα αυτά τα χαρακτηριστικά βελτιώνουν τα αποτελέσματα της συναρμολόγησης για shotgun δεδομένα, και ως εκ τούτου η αλληλούχιση Sanger μπορεί να εξακολουθεί να εφαρμόζεται, αν ο στόχος είναι η παραγωγή σχεδόν πλήρων γονιδιωμάτων σε περιβάλλοντα χαμηλής ποικιλομορφίας. Κάποια ακόμα μειονεκτήματα της τεχνολογίας Sanger είναι η απαιτητική διαδικασία κλωνοποίησης, η "μεροληπτική" στάση της έναντι των γονιδίων που είναι τοξικά για τον ξενιστή που χρησιμοποιείται κατά την κλωνοποίηση και όπως αναφέρθηκε και παραπάνω το συνολικό κόστος ανά gigabase [11].



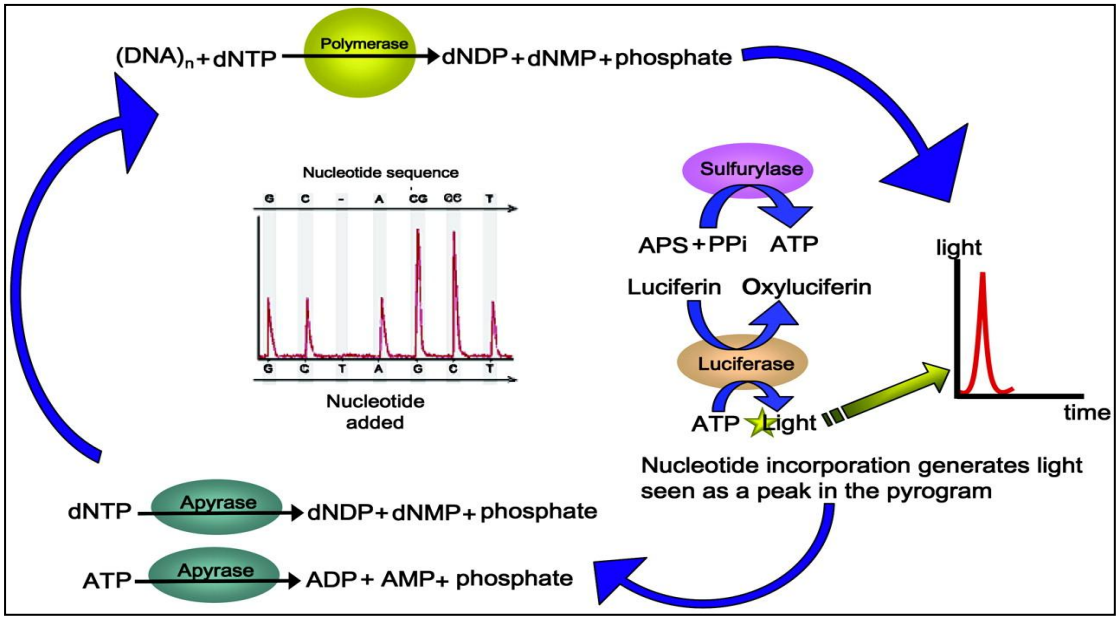
Εικόνα 8 Τυπική απεικόνιση της βασικής τεχνολογίας αλληλούχισης Sanger.

Τεχνολογίες αλληλούχισης νέας γενιάς: Πολύ πρόσφατα, η μέθοδος Sanger έχει μερικώς υποκατασταθεί από διάφορες τεχνολογίες αλληλούχισης «επόμενης γενιάς» που προσφέρουν δραματικές αυξήσεις στην απόδοση οικονομικώς συμφερόντων αλληλουχίσεων, θυσιάζοντας βέβαια τα μήκη των αλληλουχιών που είναι δυνατό να αναλυθούν. Μερικές από τις τεχνολογίες επόμενης γενιάς (2ης γενιάς) που είναι εμπορικά διαθέσιμες σήμερα περιλαμβάνουν τα μηχανήματα της 454 πυροαλληλούχισης της Roche Applied Science, τους αναλυτές Solexa /Illumina, το μηχάνημα SOLiD από την Applied Biosystems ενώ υπάρχουν πλέον και τεχνολογίες 3ης γενιάς όπως το Heliscope από την Helicos, Inc. και αρκετές άλλες.

Τεχνολογία 454/Roche : Το σύστημα 454/Roche εφαρμόζει αλυσιδωτή αντίδραση πολυμεράσης γαλακτωματοποίησης (ePCR) για να ενισχυθούν, με τη δημιουργία κλώνων, τυχαία θραύσματα DNA, τα οποία είναι συνδεμένα με μικροσκοπικά σφαιρίδια[11]. Στην ePCR, σφαιρίδια στρεπταβιδίνης που μεταφέρουν μεμονωμένα θραύσματα DNA που λαμβάνονται μέσω κατακερματισμού του DNA και της προσάρτησης των θραυσμάτων σε σφαιρίδια με χρήση προσαρμογέων, συλλαμβάνονται σε ξεχωριστά σταγονίδια.Τα σταγονίδια ενεργούν ως μεμονωμένοι αντιδραστήρες ενίσχυσης, παράγοντας ~ 10^7 κλωνικά αντίγραφα ενός μοναδικού προτύπου DNA ανά σφαιρίδιο[14]. Τα σφαιρίδια εναποτίθενται μέσα στα φρεάτια ειδικών πλακών με πηγάδια (Picotiter plates) και στη συνέχεια, πραγματοποιείται πυροαλληλούχιση τους μεμονωμένα και παράλληλα. Η πυροαλληλούχιση είναι μια τεχνική αλληλούχισης μέσω σύνθεσης (sequencing by synthesis) που μετρά την απελευθέρωση ανόργανου πυροφωσφορικού (PPi) μέσω μέτρησης της χημειοφωταύγειας[14]. Η διαδικασία της πυροαλληλούχισης περιλαμβάνει τη διαδοχική προσθήκη και των τεσσάρων τριφωσφορικών δεοξυνουκλεοσιδίων, τα οποία, αν είναι συμπληρωματικά προς τον πρότυπο κλώνο, ενσωματώνονται από μία DNA πολυμεράση. Αυτή η αντίδραση πολυμερισμού απελευθερώνει πυροφωσφορικό, το οποίο μετατρέπεται μέσω δύο ενζυμικών αντιδράσεων για την παραγωγή φωτός. Για την ακρίβεια, η ATP σουλφουρυλάση, παρουσία φωσφοθειικής αδενοσίνης (APS) μετατρέπει το πυροφοσφωρικό σε ATP και η παρουσία του ATP ενεργοποιεί μια αντίδραση κατά την οποία η λουσιφερίνη μετατρέπεται σε οξυλουσιφερίνη από την λουσιφεράση, με αποτέλεσμα την παραγωγή φωτός. Η παραγωγή φωτός ~ 1.2 εκατομμυρίων αντιδράσεων ανιχνεύεται εν παραλλήλω μέσω μιας συσκευής (κάμερας) συζευγμένου φορτίου (CCD) και μετατρέπεται στην πραγματική αλληλουχία του προτύπου[11]. Η αλληλουχία του πρότυπου DNA προσδιορίζεται λοιπόν από ένα «πυρόγραμμα" το οποίο συμφωνεί με τη σωστή σειρά των νουκλεοτιδίων που έχουν ενσωματωθεί. Δύο πτυχές είναι σημαντικές στη διαδικασία αυτή σε σχέση με τις εφαρμογές μεταγονιδιωματικής. Πρώτα, η ePCR έχει δειχθεί πως παράγει τεχνητές αλληλουχίες αντίγραφα, οι οποίες θα επηρεάσουν τυχόν εκτιμήσεις αφθονίας γονιδίων. Η κατανόηση του ποσού των αλληλουχιών αντιγράφων είναι ζωτικής σημασίας για

την ποιότητα των δεδομένων του κάθε τρεξίματος αλληλούχισης και τα αντίγραφα είναι δυνατό να εντοπιστούν και να φιλτράρονται με διάφορα εργαλεία βιοπληροφορικής. Δεύτερον, η ένταση του φωτός που παράγεται όταν η πολυμεράση διατρέχει ένα ομοπολυμερές είναι συχνά δύσκολο να συσχετιστεί με τον πραγματικό αριθμό των νουκλεοτιδικών θέσεων. Τυπικά, αυτό οδηγεί σε σφάλματα παρεμβολής ή διαγραφής στα ομοπολυμερή και ως εκ τούτου, μπορεί να προκαλέσει μετατοπίσεις του πλαισίου ανάγνωσης, εάν ακολουθίες κωδικοποίησης πρωτεΐνης (CDS) καλούνται σε μια ενιαία αλληλουχία ανάγνωσης (single read). Αυτού του είδος το σφάλμα είναι δυνατό, ωστόσο, να ενσωματωθεί στα μοντέλα πρόβλεψης των CDS με αποτέλεσμα την υψηλή, αν και όχι τέλεια, ακρίβεια.

Παρά αυτά τα μειονεκτήματα, το πολύ φθηνότερο κόστος, της τάξης των 20.000 δολαρίων ανά gigabase ζευγών βάσεων έχει κάνει την τεχνολογία της 454/Roche πυροαλληλούχισης μια δημοφιλή επιλογή για shotgun μεταγονιδιωματικές αναλύσεις αλληλούχισης. Επιπλέον, η τεχνολογία 454/Roche παράγει κατά μέσο όρο μήκη αλληλουχιών ανάγνωσης μεταξύ 600-800 bp (ανάλογα με την εξέλιξη των μηχανημάτων ανάλυσης), το οποίο είναι αρκετό για να προκαλέσει μόνο μια ασήμαντη απώλεια στον αριθμό των αλληλουχιών ανάγνωσης που μπορούν να σχολιαστούν. Η προετοιμασία του δείγματος έχει επίσης βελτιστοποιηθεί ώστε δεκάδες νανογραμμάρια DNA να είναι επαρκή για τον προσδιορισμό της αλληλουχίας σε single-end βιβλιοθήκες, αν και σε περιπτώσεις paired-end αλληλούχισης μπορεί να εξακολουθούν να απαιτούνται ποσότητες μικρογραμμαρίων. Επιπλέον, η πλατφόρμα ανάλυσης αλληλουχίας 454/Roche επιτρέπει να αναλυθούν έως και 12 δείγματα σε μια ενιαία λειτουργία της τάξης των 500 Mbp (οι τιμές είναι ενδεικτικές και μεταβάλλονται με την πρόοδο των τεχνολογιών)[11].

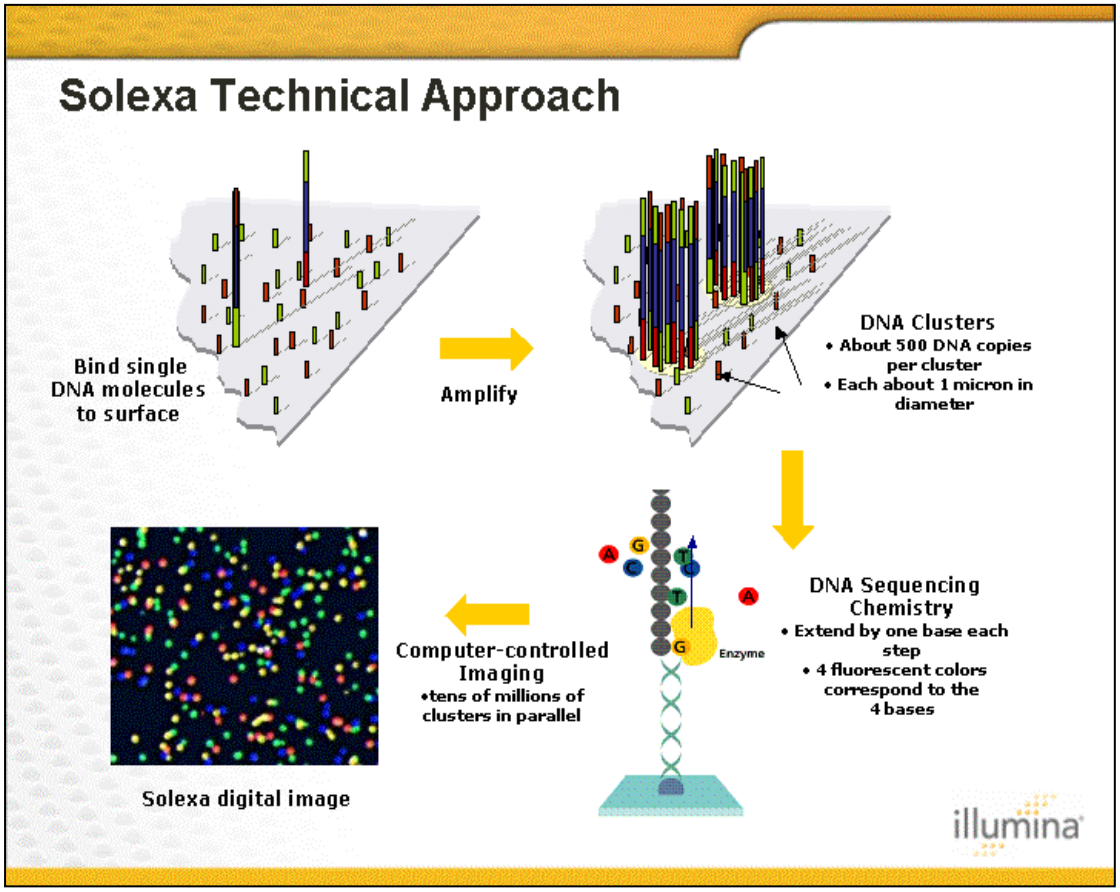


Εικόνα 9 Σχηματική αναπαράσταση της διαδικασίας της πυροαλληλούχισης.

Τεχνολογία Illumina / Solexa: Η τεχνολογία Illumina / Solexa ακινητοποιεί τυχαία θραύσματα DNA σε μια επιφάνεια και στη συνέχεια εκτελεί ενίσχυση PCR στερεάς επιφάνειας, καταλήγοντας σε συστάδες πανομοιότυπων θραυσμάτων DNA. Αυτές στη συνέχεια θα υποστούν μαζική παράλληλη αλληλούχιση χρησιμοποιώντας μια προσέγγιση αλληλούχισης DNA μέσω σύνθεσης που χρησιμοποιεί αναστρέψιμους τερματιστές με αφαιρούμενα τμήματα φθορισμού και ειδικές DNA πολυμεράσες που μπορούν να ενσωματώσουν αυτούς τους τερματιστές σε αυξανόμενες ολιγονουκλεοτιδικές αλυσίδες. Οι τερματιστές επισημαίνονται με φθορίζουσες ουσίες τεσσάρων διαφορετικών χρωμάτων για να γίνει διάκριση μεταξύ των διαφόρων βάσεων στη δεδομένη θέση της αλληλουχίας και η πρότυπη αλληλουχία του κάθε συμπλέγματος συνάγεται από την ανάγνωση του χρώματος σε κάθε διαδοχικό στάδιο προσθήκης νουκλεοτιδίων[14]. Η πυκνότητα των συμπλεγμάτων αυτών είναι τεράστια, με εκατοντάδες εκατομμύρια αλληλουχίες ανάγνωσης ανά κανάλι επιφάνειας και 16 κανάλια σε κάθε τρέξιμο στο όργανο HiSeq2000. Το μήκος των αλληλουχιών ανάγνωσης που είναι δυνατό να επεξεργαστούν προσεγγίζει πλέον τα 150 bp (σε αντίθεση με τα 30-40 bp που παρουσιάζονται στον Πίνακα 3 και αφορούν προγενέστερα μηχανήματα της ίδιας τεχνολογίας) και τα συγκεντρωμένα θραύσματα είναι δυνατό να υποστούν αλληλούχιση και από τα δύο άκρα. Συνεχόμενη πληροφορία αλληλουχίας περίπου 300 bp μπορεί να ληφθεί από δύο επικαλυπτόμενα paired-reads 150 bp από μια μόνο εισαγωγή. Συνεπώς αποδόσεις της τάξεως των 60 Gbp αναμένονται τυπικά από κάθε κανάλι. Ενώ η Illumina / Solexa παρουσιάζει περιορισμένα συστηματικά λάθη, μερικά σύνολα δεδομένων έχουν δείξει υψηλά ποσοστά σφάλματος στα άκρα-ουρές της αλληλουχίας ανάγνωσης. Σε γενικές γραμμές, το ψαλίδισμα των άκρων των αλληλουχιών ανάγνωσης έχει αποδειχθεί ότι είναι μια καλή στρατηγική για την εξάλειψη των σφαλμάτων σε "προβληματικά" σύνολα δεδομένων, ωστόσο, θα πρέπει επίσης να χρησιμοποιούνται τιμές χαρακτηριστικές της ποιότητας των αλληλουχιών για την ανίχνευση "προβληματικών" αλληλουχιών[11].

Το χαμηλότερο κόστος της τεχνολογίας αυτής (περίπου 50 USD ανά Gbp) και η πρόσφατη επιτυχία στην εφαρμογή της στη μεταγονιδιωματική κάνουν σήμερα την τεχνολογία Illumina μια όλο και πιο δημοφιλή επιλογή. Όπως και στην αλληλούχιση 454/Roche, το υλικό έναρξης μπορεί να είναι τόσο χαμηλό όσο 20 νανογραμμάρια, αλλά είναι δυνατό να απαιτούνται μεγαλύτερες ποσότητες (500-1000ng) σε κάποιες περιπτώσεις (παραδείγματος χάριν για την δημιουργία mate-pair βιβλιοθηκών). Η δυνατότητα επεξεργασίας περιορισμένου μήκους αλληλουχιών ανάγνωσης της τεχνολογίας Illumina / Solexa σημαίνει ότι μεγαλύτερο ποσοστό των μη συναρμολογημένων αλληλουχιών ανάγνωσης μπορεί να είναι πάρα πολύ μικρό για λειτουργική περιγραφή σε σχέση με την τεχνολογία 454/Roche. Ενώ η συναρμολόγηση μπορεί να προτείνεται σε μια τέτοια περίπτωση, μια πιθανή επιρροή, όπως η καταστολή των ειδών που βρίσκονται σε χαμηλή αφθονία (και τα

οποία δεν μπορούν να συναρμολογηθούν) πρέπει λαμβάνεται υπ' όψιν, όπως και το γεγονός ότι ορισμένα τρέχοντα λογισμικά πακέτα (π.χ. MG-RAST) είναι ικανά να αναλύουν ασυναρμολόγητες αλληλουχίες ανάγνωσης Illumina από 75 bp και πάνω . Ταυτόχρονη επεξεργασία δειγμάτων είναι επίσης διαθέσιμη για ξεχωριστά κανάλια αλληλούχισης, με περισσότερα από 500 δείγματα να αναλύονται ανά σειρά. Ένας άλλος σημαντικός παράγοντας προς εξέταση είναι ο χρόνος επεξεργασίας, με μια 2 × 100 bp paired-read ανάλυση αλληλούχισης να διαρκεί περίπου 10 ημέρες στον αναλυτή HiSeq2000, σε αντίθεση με την μια ημέρα για την τεχνολογία 454 / Roche. Ωστόσο, μπορεί να επιτευχθεί ταχύτερος χρόνος εκτέλεσης (αν και σε υψηλότερα κόστη, περίπου USD 600 ανά Gbp) με το νέο μηχάνημα Illumina MiSeq. Αυτή η μικρότερη έκδοση της τεχνολογίας Illumina / Solexa, μπορεί επίσης να χρησιμοποιηθεί σε δοκιμαστικές βιβλιοθήκες αλληλούχισης, πριν από την ανάλυση στον HiSeq αναλυτή για τη βαθύτερη ανάλυση αλληλουχίας[11].



Εικόνα 10 Η προσέγγιση Illumina / Solexa. Ενίσχυση στερεής επιφάνειας, ανάπτυξη των στοιβάδων DNA, προσθήκη τερματιστών με φθορίζοντα τμήματα και η τελική απεικόνιση.

Τεχνολογία ABI / SOLiD : Ο αναλυτής SOLiD της Applied Biosystems έχει χρησιμοποιηθεί εκτενώς στην αλληλούχιση γονιδιωμάτων. Ο αναλυτής αυτός παρέχει αναμφισβήτητα το χαμηλότερο ποσοστό σφάλματος σε σχέση με οποιαδήποτε τεχνολογία αλληλούχισης NGS δεύτερης γενιάς, όμως δεν καταφέρνει να παρέχει αξιόπιστα αποτελέσματα για μήκη αλληλουχιών ανάγνωσης πάνω από 50 νουκλεοτίδια. Αυτό περιορίζει την εφαρμογή του για άμεση περιγραφή γονιδίων που προέρχονται από αλληλουχίες ανάγνωσης που δεν έχουν συναρμολογηθεί ή για τη συναρμολόγηση μεγάλων αλληλεπικαλυπτόμενων τμημάτων DNA (contigs). Παρ 'όλα αυτά, για τη συναρμολόγηση ή τη χαρτογράφηση μεταγονιδιωματικών δεδομένων σε σχέση με ένα γονιδίωμα αναφοράς, πρόσφατες εργασίες έδειξαν ενθαρρυντικά αποτελέσματα[11].

Η μαζική παράλληλη αλληλούχιση με υβριδοποίηση-απολίνωση, που εφαρμόζεται στο υποστηριζόμενο σύστημα απολίνωσης (ligation) αλλά και ανίχνευσης ολιγονουκλεοτιδίων, SOLiD από την Applied Biosystems, έχει πρόσφατα καταστεί διαθέσιμη. Η χημεία της απολίνωσης που χρησιμοποιείται στον αναλυτή SOLiD βασίζεται στην polony τεχνική αλληλούχισης που δημοσιεύθηκε την ίδια χρονιά με τη μέθοδο πυροαλληλούχισης 454. Η κατασκευή βιβλιοθηκών για ανάλυση αλληλουχίας στην προσέγγιση αυτή ξεκινά με ένα μονομοριακό στάδιο ενίσχυσης μέσω ePCR, παρόμοιο με εκείνο που χρησιμοποιείται στην τεχνική 454. Τα προϊόντα ενίσχυσης μεταφέρονται πάνω σε μια γυάλινη επιφάνεια, όπου λαμβάνει χώρα η αλληλούχιση με διαδοχικούς γύρους υβριδισμού και απολίνωσης, με 16 συνδυασμούς δινουκλεοτιδίων να επισημαίνονται με τέσσερις διαφορετικές φθορίζουσες ουσίες (κάθε χρωστική χρησιμοποιείται για τη σήμανση τεσσάρων δινουκλεοτιδίων). Χρησιμοποιώντας το σχήμα κωδικοποίησης των τεσσάρων χρωστικών, κάθε θέση διερευνάται αποτελεσματικά δύο φορές και η ταυτότητα του κάθε νουκλεοτιδίου προσδιορίζεται με ανάλυση του χρώματος που προκύπτει ως αποτελέσμα από δύο διαδοχικές αντιδράσεις απολίνωσης. Αυτό το γεγονός είναι ιδιαίτερα σημαντικό καθώς, αυτό το σχήμα κωδικοποίησης των δυο βάσεων επιτρέπει τη διάκριση μεταξύ ενός σφάλματος αλληλούχισης και μιας πολυμορφικής αλληλουχίας: ένα σφάλμα θα ανιχνευθεί σε μόνο μία συγκεκριμένη αντίδραση σύνδεσης, ενώ ένας πολυμορφισμός θα ανιχνευθεί και στις δύο[14].

Τεχνολογίες αλληλούχισης 3ης γενιάς : Παρόλο που η ενίσχυση μέσω PCR έχει φέρει επαναστατικές αλλαγές στην ανάλυση του DNA, σε μερικές περιπτώσεις μπορεί να εισαγάγει σφάλματα αλληλουχίας βάσεων ή σφάλματα υπέρ ορισμένων ακολουθιών σε σχέση με άλλες, αλλάζοντας έτσι τη σχετική συχνότητα και την αφθονία των διαφόρων θραυσμάτων DNA που υπήρχαν πριν από την ενίσχυση. Για να ξεπεραστεί αυτό, η τελική μικρογράφηση εντός της νανοκλίμακας και η ελάχιστη χρήση των βιοχημικών, θα ήταν εφικτή εάν η αλληλουχία μπορούσε να προσδιοριστεί άμεσα από ένα μόνο μόριο DNA, χωρίς την ανάγκη για την εφαρμογή ενίσχυσης PCR και την ταυτόχρονη πιθανότητα στρέβλωσης των επιπέδων

αφθονίας. Αυτή η ανάλυση αλληλούχισης από ένα μόνο μόριο DNA είναι αυτό που σήμερα ονομάζεται ως "τρίτη γενιά των τεχνολογιών αλληλούχισης". Η έννοια της αλληλούχισης μέσω σύνθεσης χωρίς προηγουμένως το στάδιο ενίσχυσης, δηλαδή η μονομοριακή αλληλούχιση ερευνάται σήμερα από έναν μεγάλο αριθμό εταιρειών[13].

Μία από τις πρώτες τεχνικές για την ανάλυση αλληλουχίας από ένα ενιαίο μοναδικό μόριο DNA εισήχθη από Braslavsky και τους συνεργάτες του και έλαβε άδεια από την Helicos Biosciences, Inc ως το πρώτο εμπορικά διαθέσιμο σύστημα αλληλούχισης DNA από ένα μόριο, το 2007. Η αρχή λειτουργίας του αναλυτή **Heliscope** στηρίζεται στην τεχνολογία " πραγματικής μονομοριακής αλληλούχισης " (tSMS). Η τεχνολογία tSMS ξεκινά με την παρασκευή βιβλιοθήκης DNA μέσω διάτμησης του DNA, την προσθήκη πολυαδενινικής ουράς στα παραγόμενα DNA θραύσματα, ακολουθούμενη από υβριδισμό των θραυσμάτων DNA με πολυθυμινικά ολιγονουκλεοτίδια τα οποία συνδέονται με την κυψέλη ροής και την ταυτόχρονη αλληλούχιση σε παράλληλες αντιδράσεις. Ο κάθε κύκλος προσδιορισμού αλληλουχίας εμπεριέχει την επέκταση του DNA με ένα από τα τέσσερα σημασμένα με φθορισμό νουκλεοτίδια και ακολούθως την ανίχνευση του νουκλεοτιδίου με τον αναλυτή Heliscope. Η επακόλουθη χημική διάσπαση των φθοροφόρων νουκλεοτιδίων επιτρέπει στον επόμενο κύκλο επιμήκυνσης του DNA να αρχίσει με ένα άλλο επισημασμένο με φθορισμό νουκλεοτίδιο, γεγονός που επιτρέπει τον προσδιορισμό της αλληλουχίας του DNA. Ο αναλυτής Heliscope μπορεί να αλληλουχίσει έως 28 Gb σε ένα μόνο κύκλο αλληλούχισης που διαρκεί περίπου 8 ημέρες. Μπορεί να δημιουργήσει μικρές αλληλουχίες ανάγνωσης με ένα μέγιστο μήκος 55 βάσεων. Προσφάτως η Helicos ανέπτυξε μια νέα γενιά νουκλεοτιδίων "μιας-βάσης τη φορά" που επιτρέπουν περισσότερο ακριβείς αναλύσεις αλληλουχίας ομοπολυμερών αλλά και άμεσες αλληλουχίσεις RNA[13].

Η τεχνολογία Ion Torrent (και πιο πρόσφατα η Ion Proton) είναι μια άλλη αναδυόμενη τεχνολογία και βασίζεται επί της αρχής στο ότι τα πρωτόνια που απελευθερώνονται κατά τη διάρκεια του πολυμερισμού του DNA μπορούν να ανιχνεύσουν την ενσωμάτωση νουκλεοτιδίων. Το σύστημα αυτό υπόσχεται ανάλυση αλληλουχιών ανάγνωσης μήκους > 100 bp και ρυθμό απόδοσης της τάξεως μεγέθους των συστημάτων αλληλούχισης 454 / Roche[11].

Τέλος, η **Pacific Biosciences (PacBio)** έχει κυκλοφορήσει μια τεχνολογία αλληλούχισης με βάση τη μονομοριακή ανίχνευση σε πραγματικό χρόνο σε πηγάδια zero-mode κυματοδηγών. Θεωρητικά, αυτή η τεχνολογία με την πλατφόρμα της RS1, θα πρέπει να παρέχει πολύ μεγαλύτερα μήκη ανάγνωσης από τις άλλες τεχνολογίες που αναφέρονται, κάτι που θα διευκολύνει τον χαρακτηρισμό και την συναρμολόγηση. Επιπλέον, μια διαδικασία που ονομάζεται strobing θα μιμείται τις paired-end αλληλουχίες ανάγνωσης. Ωστόσο, η ακρίβεια της ανάλυσης μονών

αλληλουχιών ανάγνωσης με την μέθοδο της PacBio εντοπίζεται σήμερα μόνο στο 85% και χάνονται τυχαίες αλληλουχίες ανάγνωσης, καθιστώντας το μέσο άχρηστο στην τρέχουσα μορφή του για αλληλούχιση μεταγονιδιωματικών δεδομένων[11].

2.4 Συναρμολόγηση (Assembly)

Η συναρμολόγηση είναι η διαδικασία του συνδυασμού αλληλουχιών ανάγνωσης για τον σχηματισμό συνεχόμενων τμημάτων DNA που ονομάζονται contigs, με βάση την ομοιότητα αλληλουχίας μεταξύ των αλληλουχιών ανάγνωσης. Η συναινετική αλληλουχία (consensus sequence) για κάθε contig βασίζεται είτε στο νουκλεοτίδιο με την υψηλότερη ποιότητα σε κάθε αλληλουχία ανάγνωσης σε κάθε θέση είτε στην αρχή της πλειοψηφίας, δηλαδή, το πιο συχνά συναντώμενο νουκλεοτίδιο σε κάθε θέση. Ο αριθμός των υποκείμενων αλληλουχιών ανάγνωσης για κάθε συναινετική βάση ονομάζεται βάθος ή κάλυψη. Η αλληλούχιση τυπικά εκτελείται και από τις δύο πλευρές ενός ένθετου τμήματος γενετικού υλικού σε ένα πλασμίδιο φορέα και αυτά τα ζεύγη ονομάζονται paired-reads ή mate pairs. Η γνώση κατά προσέγγιση του μεγέθους του ενθέματος της βιβλιοθήκης διευκολύνει την παραγωγή μιας πιο ακριβούς συναρμολόγησης επειδή τα mate pairs παρέχουν έναν εξωτερικό περιορισμό ως προς την πορεία της συναρμολόγησης. Η παρουσία paired reads σε δύο διαφορετικά contigs επιτρέπει σε αυτά τα contigs να συνδεθούν σε μια μεγαλύτερη αλληλουχία DNA που δεν έχει τον επικαλυπτόμενο χαρακτήρα των contigs και που ονομάζεται ικρίωμα (scaffold), του οποίου το μέγεθος των κενών ανάμεσα στα contigs μπορεί να εκτιμηθεί με βάση το μέγεθος της ένθετης αλληλουχίας των ζευγών αλληλουχίας ανάγνωσης. Για το λόγο αυτό, μεγάλοι ένθετοι κλώνοι όπως φοσμίδια είναι ιδιαίτερα χρήσιμοι για τη βελτίωση της διαδικασίας της συναρμολόγησης[12]. Τα scaffolds, που μερικές φορές ονομάζονται και supercontigs ή metacontigs, ορίζουν λοιπόν την σειρά των contigs και τον προσανατολισμό τους, ενώ όπως αναφέρθηκε ήδη ορίζουν και τα μεγέθη των κενών μεταξύ των contigs. Η τοπολογία των scaffolds μπορεί να είναι μια απλή διαδρομή ή και ένα δίκτυο. Οι περισσότεροι συναρμολογητές βγάζουν ως έξοδο, επιπλέον, μια σειρά από μη συναρμολογημένες ή μερικώς συναρμολογημένες αλληλουχίες ανάγνωσης. Η πιο ευρέως αποδεκτή μορφή αρχείου δεδομένων για μια συναρμολόγηση είναι τα αρχεία FASTA, όπου η συναινετική αλληλουχία σε κάθε contig μπορεί να αναπαριστάται από χορδές των χαρακτήρων A, C, G, T, συν, ενδεχομένως, άλλους χαρακτήρες με ειδική σημασία. Οι παύλες, για παράδειγμα, μπορεί να αντιπροσωπεύουν επιπλέον βάσεις που παραλείπονται από τη συναινετική αλληλουχία, αλλά υπάρχουν σε μια μειονότητα υποκείμενων αλληλουχιών ανάγνωσης. Η συναινετική αλληλουχία στα scaffolds μπορεί να έχει μια σειρά από N για τα κενά μεταξύ των contigs. Ο αριθμός των διαδοχικών N μπορεί να υποδηλώνει την εκτίμηση του μήκους του κενού με βάση τα εκτεινόμενα paired ends [15].

Οι συναρμολογήσεις μετρώνται από το μέγεθος και την ακρίβεια των contigs και των scaffolds τους. Το μέγεθος της συναρμολόγησης δίνεται συνήθως από τις στατιστικές, συμπεριλαμβανομένου του μέγιστου μήκους, του μέσου μήκους, του συνδυασμένου συνολικού μήκους, και του N50. Το contig N50 είναι το μήκος του μικρότερου contig στο σύνολο που περιέχει τα λιγότερα (τα μεγαλύτερα δηλαδή) contigs των οποίων το συνδυασμένο μήκος αντιπροσωπεύει τουλάχιστον το 50% του συνόλου. Οι στατιστικές N50 για διαφορετικές συναρμολογήσεις δεν είναι συγκρίσιμες εκτός αν κάθε μια υπολογίζεται χρησιμοποιώντας την ίδια τιμή συνδυασμένου μήκους. Η ακρίβεια της συναρμολόγησης είναι δύσκολο να μετρηθεί. Η αντιστοίχιση σε ακολουθίες αναφοράς είναι χρήσιμη, όταν όμως υπάρχουν τέτοιες που να θεωρούνται αξιόπιστες[15].

Δύο στρατηγικές συναρμολόγησης μπορούν να χρησιμοποιηθούν για μεταγονιδιωματικά δείγματα: **συναρμολόγηση με βάση τις αναφορές (reference-based assembly)**και η **de novo συναρμολόγηση.**

Η συναρμολόγηση με βάση τις αναφορές μπορεί να γίνει με λογισμικά πακέτα όπως το Newbler (Roche), το AMOS ή το MIRA. Τα εν λόγω λογισμικά πακέτα περιλαμβάνουν αλγόριθμους που είναι γρήγοροι και αποδοτικοί ως προς θέματα μνήμης και ως εκ τούτου μπορεί συχνά να εκτελούνται σε μηχανές μεγέθους laptop σε μερικές ώρες. Η reference-based συναρμολόγηση λειτουργεί καλά, εάν το μεταγονιδιωματικό σύνολο δεδομένων περιέχει αλληλουχίες στενά συνδεδεμένες με διαθέσιμα γονιδιώματα αναφοράς. Ωστόσο, διαφορές ανάμεσα στο πραγματικό γονιδίωμα του δείγματος και το δείγμα αναφοράς, όπως ένα μεγάλο ένθετο κομμάτι, διαγραφές, ή πολυμορφισμοί, μπορεί να σημαίνουν ότι η συναρμολόγηση είναι κατακερματισμένη ή ότι οι αποκλίνουσες περιοχές δεν καλύπτονται[11].

Η συναρμολόγηση de novo συνήθως απαιτεί μεγαλύτερη υπολογιστική ισχύ. Έτσι, μια ολόκληρη κατηγορία εργαλείων συναρμολόγησης με βάση τα διαγράμματα de Bruijn έχει δημιουργηθεί ειδικά για να χειριστεί πολύ μεγάλες ποσότητες δεδομένων. Οι απαιτήσεις των μηχανημάτων των de Bruijn συναρμολογητών, όπως είναι ο Velvet ή ο SOAP εξακολουθούν να είναι σημαντικά υψηλότερες από ό, τι για τη συναρμολόγηση με βάση αλληλουχίες αναφοράς και συχνά απαιτούν μνήμη εκατοντάδων gigabytes σε ένα μοναδικό μηχάνημα και οι χρόνοι εκτέλεσης συχνά είναι ολόκληρες ημέρες[11].

Σε μια προσπάθεια καταγραφής διαδεδομένων αλγορίθμων συναρμολόγησης ,ένας από τους βασικούς αλγορίθμους συναρμολόγησης είναι εκείνος με βάση τις επικαλύψεις, ή όπως αναλυτικά ονομάζεται **Overlap-layout-consensus (OLC)** . Το πλαίσιο λειτουργίας του αλγορίθμου OLC είναι κατά βάση μια διαδικασία τριών

βημάτων. Αρχικά, υπολογίζονται οι επικαλύψεις μεταξύ των αλληλουχιών ανάγνωσης. Στη συνέχεια, προσδιορίζεται η διάταξη και ο προσανατολισμός των αλληλουχιών ανάγνωσης στη συναρμολόγηση. Τέλος, η συναίνεση επιτρέπει τον προσδιορισμό των νουκλεοτιδίων σε κάθε θέση στα contigs. Συναρμολογητές που εφαρμόζουν αυτήν την ιδέα είναι πολλοί και ήταν κατασκευασμένοι ώστε να ξεπεραστούν τα εμπόδια συναρμολόγησης σε δεδομένα που προέρχονται από αλληλούχιση τεχνολογίας Sanger, πριν εμφανιστούν τα συστήματα υψηλής απόδοσης νέας γενιάς. Τα εν λόγω λογισμικά είναι προσαρμοσμένα για την συναρμολόγηση μεγάλων αλληλουχιών ανάγνωσης, όπως αυτές που προέρχονται από αλληλούχιση τύπου Sanger. Περιλαμβάνουν συναρμολογητές όπως ο Celera αλλά και ο Arachne. Στη συνέχεια, το παράδειγμα αυτού του αλγορίθμου προσαρμόστηκε στο συστήμα Roche/454 και ο αναλυτής Roche/454 διανέμεται μαζί με τον συναρμολογητή Newbler. Επίσης, ο Edena είναι μια διάταξη συναρμολόγησης μέσω επικάλυψης συναίνεσης, η οποία μπορεί να επεξεργαστεί σύντομες αλληλουχίες ανάγνωσης μεγέθους περίπου 35 βάσεων[16].

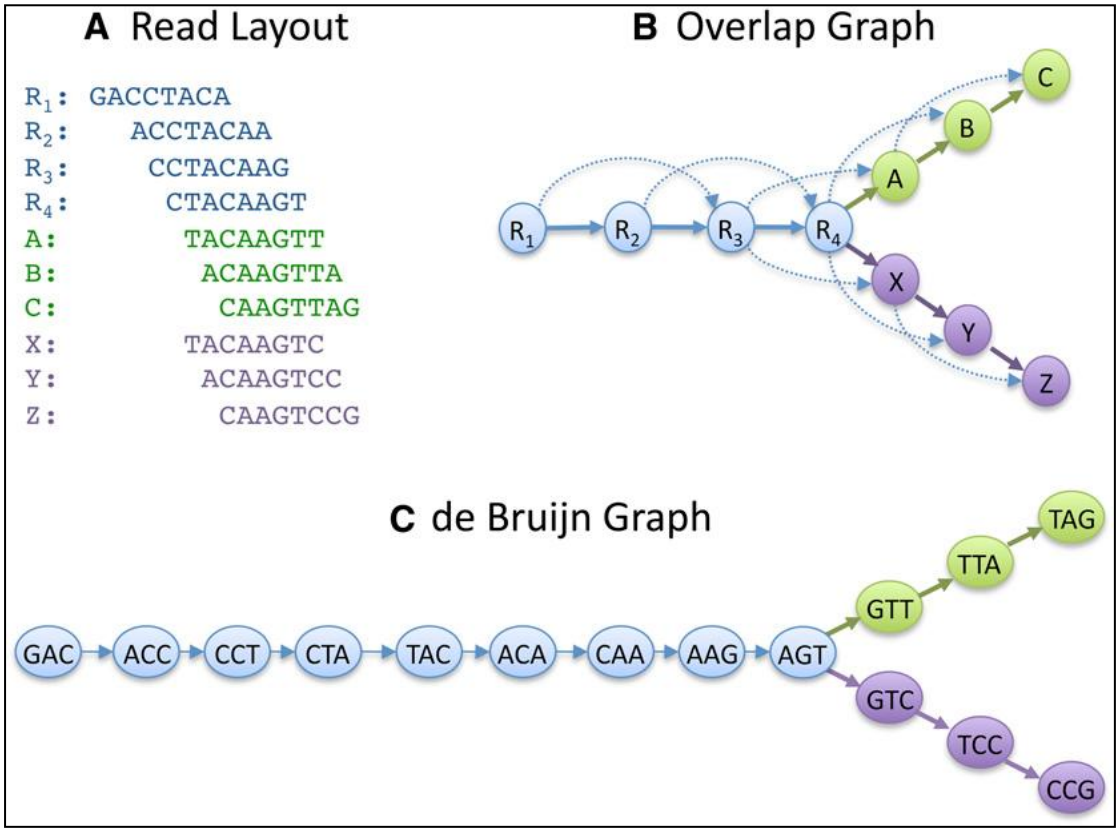
Αυτή την περίοδο, οι αποδεκτές διατάξεις που χαρακτηρίζονται ως πιο ικανές για τη συναρμολόγηση δεδομένων που προέρχονται από NGS τεχνικές, χρησιμοποιούν μεθόδους που βασίζονται σε **K-mer de Bruijn** γραφήματα, συμπεριλαμβανομένων προγραμμάτων όπως τα Velvet, SOAPdenovo, ALLPATHS, AbySS, CLC Bio, καθώς και ενός αριθμού νεότερων συναρμολογητών υπό ανάπτυξη. Η προσέγγιση Kmer αναπτύχθηκε σε μια προσπάθεια να ξεπεραστούν οι περιορισμοί του υπερβολικού χρόνου των παραδοσιακών στρατηγικών συναρμολόγησης με βάση τις επικαλύψεις. Ο κύριος περιορισμός των συναρμολογητών που βασίζονται σε τεχνικές επικάλυψης είναι ότι απαιτούν μια σειρά υπολογισμών ανάλογη προς N^2 (το τετράγωνο του αριθμού των αλληλουχιών ανάγνωσης της συναρμολόγησης). Καθώς τα αρχεία εξόδου που προέρχονται από NGS τεχνικές ξεπέρασαν το φράγμα των πολλών εκατομμυρίων αλληλουχιών ανάγνωσης, αυτή η διαδικασία έγινε απαγορευτικά αργή. Είναι σημαντικό να σημειωθούν δύο πράγματα περί των τεχνικών συναρμολόγησης με βάση τα Kmer : πρώτον, όπως ήδη αναφέρθηκε η μέθοδος αυτή μειώνει το χρόνο της συναρμολόγησης, αλλά με κόστος την σημαντική απαίτηση μνήμης RAM η οποία είναι ανάλογη με το μέγεθος του γονιδιώματος ή των γονιδιωμάτων που συναρμολογούνται ή / και της ποσότητας των δεδομένων, η οποία de facto θέτει όρια στο συνολικό μέγεθος του μεταγονιδιώματος που συναρμολογείται και δεύτερον, αυτή η μέθοδος είναι μη-ντετερμινιστική. Επειδή οι αλληλουχίες ανάγνωσης σπάνε σε μικρότερα κομμάτια καθορισμένου μήκους (Kmers), οι αλληλουχίες ανάγνωσης δεν είναι πλέον οι ίδιες ο στόχος της συναρμολόγησης, με αποτέλεσμα την πιθανή εισαγωγή σφαλμάτων κατά την διαδικασία αυτή[17].

Το γεγονός ότι οι περισσότερες (αν όχι όλες) μικροβιακές κοινότητες παρουσιάζουν σημαντικές διακυμάνσεις σε επίπεδο στελεχών και ειδών καθιστά τη χρήση των αλγορίθμων συναρμολόγησης που μελετούν κλωνικά γονιδιώματα λιγότερο κατάλληλη για μεταγονιδιωματική ανάλυση. Οι παραδοχές σε ότι αφορά τους κλώνους, στις οποίες στηρίζονται πολλοί συναρμολογητές θα μπορούσαν να οδηγήσουν στην υποβάθμιση της πληροφορίας των contigs ορισμένων ετερογενών ταξινομικών ομάδων υπό συγκεκριμένες ρυθμίσεις παραμέτρων. Πρόσφατα, δύο συναρμολογητές τύπου de Bruijn, ο MetaVelvet και ο Meta-IDBA έχουν κυκλοφορήσει και ασχολούνται αποκλειστικά με την "μη κλωνικότητα" φυσικών πληθυσμών. Και οι δύο συναρμολογητές έχουν στόχο τον εντοπισμό σε ολόκληρα γραφήματα de Bruijn ενός υπογραφήματος που αντιπροσωπεύει γονιδιώματα που σχετίζονται μεταξύ τους. Εναλλακτικά, το μεταγονιδιωματικό μείγμα ακολουθιών μπορεί να διχοτομηθεί σε "κάδους ειδών" μέσω του k-mer binning . Αυτά τα υπογραφήματα ή υποσύνολα, στη συνέχεια χρησιμοποιούνται να χτιστεί μια συναινετική αλληλουχία των γονιδιωμάτων[11].

Εκτός από τους συναρμολογητές που βασίζονται τις αναλύσεις τους στα διαγράμματα de Bruijn αλλά και τους OLC συναρμολογητές, υπάρχουν και άλλου τύπου μηχανήματα όπως οι συναρμολογητές Greedy, η λειτουργία των οποίων στηρίζεται σε έναν greedy αλγόριθμο ο οποίος αναπτύσσει επαναληπτικά contigs με πρώτη την επιλογή της βέλτιστης επικάλυψης. Η εφαρμογή του αλγόριθμου αυτού εισήχθηκε για πρώτη φορά στις τεχνολογίες επεξεργασίας σύντομων αλληλουχιών ανάγνωσης: SSAKE, VCAKE και SHARGCS [16].

Η κύρια αιτία της λανθασμένης συναρμολόγησης σε γονιδιωματικές αναλύσεις είναι οι επαναλαμβανόμενες περιοχές, κάτι που μπορεί να επιλυθεί κατά τον τερματισμό της διαδικασίας. Η συναρμολόγηση σε μεταγονιδιωματικές αναλύσεις επηρεάζεται επίσης από επαναλήψεις, αλλά δημιουργούνται και πρόσθετες προκλήσεις συναρμολόγησης, όπως το ανομοιόμορφο βάθος ανάγνωσης εξαιτίας ανομοιόμορφης κατανομής της αφθονίας των διάφορων ειδών και η πιθανότητα συναρμολόγησης μαζί, αλληλουχιών ανάγνωσης που προέρχονται από διαφορετικά είδη. Ως εκ τούτου, μπορεί όχι μόνο να διατηρηθούν λάθος συναρμολογημένες αλληλουχίες ανάγνωσης στα τελικά στοιχεία προς δημοσίευση λόγω της απουσίας της διαδικασίας του τερματισμού, αλλά είναι δυνατό αλληλουχίες ανάγνωσης από περισσότερα του ενός είδη να συναρμολογούνται μαζί, παράγοντας έτσι χιμαιρικά contigs. Αυτή η περίπτωση συν-συναρμολόγησης είναι πιο πιθανό να συμβεί με αλληλουχίες ανάγνωσης που προέρχονται από στενά συνδεόμενα γονιδιώματα όπου η ομοιότητα αλληλουχίας είναι υψηλότερη (παρατηρείται συν-συναρμολόγηση συνήθως σε ομόλογες περιοχές δύο ή περισσότερων στελεχών με έως και 4% νουκλεοτιδική απόκλιση αλληλουχίας), αλλά έχει παρατηρηθεί και μεταξύ αλληλουχιών ανάγνωσης που προέρχονται από φυλογενετικά μακρινές

ταξινομικές ομάδες, με τα συντηρημένα γονίδια να εξυπηρετούν ως βασικό σημείο της λανθασμένης συναρμολόγησης. Για παράδειγμα, ένα contig από μεταγονιδίωμα επιφάνειας θαλασσινού νερού μπορεί να αποτελείται από αλληλουχίες ανάγνωσης καταγωγής από βακτήρια και Αρχαία, όπως αποδεικνύεται από τη μελέτη των γονιδίων, με το γονίδιο 16S rRNA να χρησιμεύει ως σημείο εστίασης της λανθασμένης συναρμολόγησης σε αυτό το παράδειγμα. Μια πρόσφατη μελέτη προσομοίωσης διαπίστωσε ότι οι χίμαιρες είναι ιδιαίτερα διαδεδομένες μεταξύ των contigs που είναι μικρότερα από 10 kbp σε μέγεθος. Υψηλής πολυπλοκότητας μικροβιακές κοινότητες που στερούνται κυρίαρχων πληθυσμών σπάνια παράγουν contigs μεγαλύτερα από 10 kbp, με αποτέλεσμα να προτείνεται τα εν λόγω σύνολα δεδομένων να μην συναρμολογούνται καθόλου[12].



Εικόνα 11 Ενδεικτική σχηματική διαφοροποίηση της προσέγγισης OLC και του αλγορίθμου διαγραμμάτων de Bruijn σε ένα σετ 10 αλληλουχιών ανάγνωσης. Σημειώνεται πως στο παράδειγμα αυτό λήφθηκε υπ' όψιν ένας μόνο προσανατολισμός της κάθε αλληλουχίας χάριν απλοποίησης των διαγραμμάτων[18].

2.5 Κατηγοριοποίηση (Binning)

Η μεταγονιδιωματική πορεία μιας ανάλυσης αλληλουχιών παράγει μια συλλογή από reads, contigs αλλά και γονίδια. Η σύνδεση αυτών των δεδομένων με τους οργανισμούς από το οποίους προήλθαν είναι ιδιαίτερα επιθυμητή για την ερμηνεία και μελέτη του οικοσυστήματος[12]. Επιπροσθέτως, μια από τις μεγαλύτερες προκλήσεις για τους υπολογιστικούς βιολόγους δεν είναι μόνο απλά η καταγραφή των γνωστών οργανισμών, αλλά επίσης ο προσδιορισμός και ο χαρακτηρισμός νέων οργανισμών που ανήκουν σε γνωστές ή άγνωστες ταξινομικές ομάδες. Οι οργανισμοί αυτοί θα μπορούσαν να ανήκουν σε ένα εντελώς νέο γένος, είδος, οικογένεια, σειρά ή τάξη ή ακόμα και μια νέα συνομοταξία (φύλο)[19].

Για τους λόγους αυτούς, ένα από τα πρώτα βήματα σε μια μεταγονιδιωματική ανάλυση είναι η εκτίμηση της ταξινομικής ποικιλομορφίας του δεδομένου περιβαλλοντικού δείγματος. Το βήμα αυτό περιλαμβάνει την ταυτοποίηση των διαφορετικών ταξινομικών ομάδων στο δείγμα και την δημιουργία προφίλ αφθονίας τους. Ουσιαστικά επιχειρείται η κατάταξη reads ή και contigs που προέρχονται από μια μεγάλη ποικιλία μεθόδων αλληλούχισης όπως έχει ήδη αναφερθεί, στα σωστά "ταξινομικά δοχεία". Αυτή η διαδικασία ταξινομικής κατηγοριοποίησης και ομαδοποίησης αλληλουχιών αναφέρεται ως **binning**[20]. Ανάλογα με τις ανάγκες της κάθε έρευνας, αυτή η διαδικασία κατηγοριοποίησης μπορεί να εφαρμοστεί σε διάφορα ταξινομικά επίπεδα, αρχίζοντας από το επίπεδο του βασιλείου (Kingdom) ως το υψηλότερο επίπεδο και φτάνοντας μέχρι το είδος (Species) που θεωρείται το χαμηλότερο επίπεδο[21]. Η ακριβής κατηγοριοποίηση των μεταγονιδιωματικών αλληλουχιών είναι ένα πάρα πολύ σημαντικό στάδιο για κάθε μεταγονιδιωματική ανάλυση καθώς η λανθασμένη κατάταξη αλληλουχιών στα ταξινομικά δοχεία μπορεί να επηρεάσει την πορεία της ανάλυσης σε σχέση με οποιοδήποτε από τα στάδια που ακολουθούν[19].

Εφαρμόζονται δυο διαφορετικές μεθοδολογίες για την κατηγοριοποίηση μεταγονιδιωματικών αλληλουχιών. Η πρώτη μέθοδος βασίζεται στα συνθετικά χαρακτηριστικά των αλληλουχιών όπως το ποσοστό GC, η χρήση κωδικονίων και η συχνότητα κατανομής ολιγο-νουκλεοτιδίων ενώ η δεύτερη μέθοδος βασίζεται στην ομοιότητα των αλληλουχιών[20]. Οι αλγόριθμοι κατηγοριοποίησης που βασίζονται σε συνθετικά χαρακτηριστικά περιλαμβάνουν πακέτα όπως τα Phylopythia, S-GSOM, TETRA και TACOA, ενώ παραδείγματα λογισμικών binning που βασίζονται στην ομοιότητα αλληλουχιών αποτελούν κατά βάση τα IMG/M, MG-RAST, MEGAN, CARMA, SOrt-ITEMS και MetaPhyler. Υπάρχει επίσης αριθμός αλγορίθμων binning που λαμβάνει υπόψη τόσο τα συνθετικά χαρακτηριστικά όσο και την ομοιότητα, με χαρακτηριστικά παραδείγματα τα προγράμματα PhymmBL, SPHINX και MetaCluster. Όλα αυτά τα εργαλεία χρησιμοποιούν διαφορετικές μεθόδους ομαδοποίησης αλληλουχιών, συμπεριλαμβανομένων χαρτών αυτοοργάνωσης

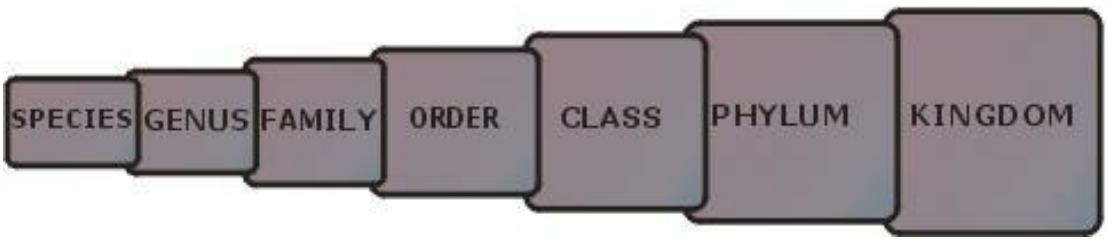
(SOMs) ή τεχνικές ιεραρχικής ομαδοποίησης και λειτουργούν είτε με έναν τρόπο δίχως επίβλεψη είτε με επιτήρηση από τον χρήστη για τον καθορισμό των δοχείων κατηγοριοποίησης[11].

Οι **μέθοδοι κατηγοριοποίησης που σχετίζονται με συνθετικά χαρακτηριστικά** ουσιαστικά ομαδοποιούν τμήματα DNA, με έναν επιτηρούμενο ή μη τρόπο, χρησιμοποιώντας κοινά χαρακτηριστικά όπως η δομή και η σύνθεση του γονιδιώματος. Κυτταρικές διεργασίες όπως η χρήση κωδικονίων, επιδιορθωτικοί μηχανισμοί, αμυντικοί μηχανισμοί περιορισμού και τροποποίησης, παρέχουν υπογραφές συνθετικών χαρακτηριστικών - κατά βάση ολιγονουκλεοτιδικές συχνότητες - ξεχωριστές για κάθε γονιδίωμα. Αυτή η ιδιότητα των γονιδιωμάτων έχει μελετηθεί με βάση μια μεγάλη ποικιλία μεθόδων, ώστε να εντοπίζονται ομάδες αλληλουχιών με παρόμοια συνθετικά χαρακτηριστικά όπως περιοχές κωδικοποίησης ταξινομικών δεικτών (16S rRNA, recA, rpoB και άλλα κοινώς αποδεκτά χαρακτηριστικά γονίδια) και να προσδιορίζονται οι φυλογενετικές τους ρίζες[12, 21]. Μια χαρακτηριστική μη επιτηρούμενη μέθοδος αυτής της κατηγορίας είναι το λογισμικό TETRA. Στις μη επιτηρούμενες προσεγγίσεις δεν απαιτείται η εξοικείωση των μοντέλων σε αλληλουχίες αναφοράς και συχνά χρησιμοποιούνται χάρτες αυτοοργάνωσης. Το λογισμικό Phylopythia είναι μια επιτηρούμενη μέθοδος που χρησιμοποιεί μοντέλα εκμάθησης SVM (support vector machines) για την κατηγοριοποίηση αλληλουχιών μεγαλύτερου μεγέθους από 1kb, βασιζόμενο σε ολιγονουκλεοτιδικές συχνότητες. Το S-GSOM χρησιμοποιεί και αυτό τεχνικές χαρτών αυτοοργάνωσης(SOMs)[22]. Το λογισμικό TACOA χρησιμοποιεί επίσης ολιγονουκλεοτιδικές συχνότητες βασιζόμενο στο ποσοστό GC (ποσοστό των βάσεων που είναι είτε γουανίνη είτε κυτοσίνη σε ένα τμημα DNA) για την κατασκευή μοντέλων συσχετιζόμενων άμεσα με συγκεκριμένους οργανισμούς[19]. Τέλος, ο αλγόριθμος Phymm χρησιμοποιεί ενσωματωμένα Μαρκοβιανά μοντέλα (interpolated Markov models) για τον χαρακτηρισμό αλληλουχιών ποικίλου μήκους, με βάση την φυλογενετική τους ομαδοποίηση. Ο εν λόγω αλγόριθμος, οδηγεί στη μη δημιουργία ομάδων αλληλουχιών που δεν έχουν καταταγεί σε κάποια ταξινομική ομάδα, αλλά θα πρέπει να λαμβάνεται υπ' όψιν η αξιοπιστία του μοντέλου σε περιπτώσεις αλληλουχιών που δεν είναι εκ των πραγμάτων δυνατό να αντιστοιχισθούν με ακρίβεια[22].

Στην περίπτωση των **προσεγγίσεων που βασίζονται στην ομοιότητα αλληλουχιών**, αρχικά ελέγχεται η ομοιότητα των αλληλουχιών ανάγνωσης σε σχέση με γνωστές αλληλουχίες που εντοπίζονται σε βιβλιοθήκες αναφοράς. Σε αυτό το πρώτο στάδιο, χρησιμοποιούνται λογισμικά ομοπαράθεσης αλληλουχιών (sequence alignment) όπως παραδείγματος χάριν το **BLAST** ώστε να συγκριθεί κάθε εξεταζόμενη αλληλουχία του μεταγονιδιωματικού δείγματος με όλες τις αλληλουχίες-στόχους που βρίσκονται στις προαναφερόμενες βιβλιοθήκες αναφοράς. Στην συνέχεια

ακολουθεί το δεύτερο στάδιο, όπου αναλύεται το μοτίβο και η ποιότητα των χτυπημάτων (hits) που έχουν προέλθει από το BLAST ώστε να προκύψουν πληροφορίες που θα οδηγήσουν τελικά στην αντιστοίχιση κάθε ακολουθίας σε μια ομάδα οργανισμών ή έναν κλάδο. Έτσι κάθε αλληλουχία ανάγνωσης που παρουσιάζει ομοιότητα με αλληλουχίες που ανήκουν σε έναν μόνο οργανισμό σύμφωνα με την βιβλιοθήκη αναφοράς, αντιστοιχίζεται στον οργανισμό αυτό. Η διαδικασία αυτή μπορεί να πραγματοποιηθεί μέσω πολλών διαφορετικών προσεγγίσεων ανάλογα με το λογισμικό που χρησιμοποιείται. Αξίζει να αναφερθούν παραδείγματα, όπως το MEGAN που αντιστοιχίζει κάθε αλληλουχία μέσω του αλγορίθμου του ελάχιστου κοινού προγόνου (Lowest Common Ancestor - LCA) πάνω σε ένα φυλογενετικό δέντρο. Το CARMA λειτουργεί με τρόπο παρόμοιο με το MEGAN αλλά χρησιμοποιεί την βιβλιοθήκη Pfam ως πηγή και βάση για την ταξινομική κατηγοριοποίηση. Το λογισμικό SOrt-ITEMS χρησιμοποιεί τον δικό του αλγόριθμο, ο οποίος αρχικά ελέγχει την ποιότητα της ομοπαράθεσης της κάθε αλληλουχίας με τις αλληλουχίες αναφοράς για τις οποίες έχουν προκύψει χτυπήματα έως ότου φτάσει στο κατάλληλο επίπεδο στο ταξινομικό δέντρο στο οποίο η αλληλουχία μπορεί να αντιστοιχισθεί. Στην συνέχεια χρησιμοποιεί την προσέγγιση ορθολογίας ώστε να ταυτοποιήσει εκείνα τα χτυπήματα που εμφανίζουν δεδομένη ορθολογία, παραδείγματος χάριν αμοιβαία ομοιότητα με την εξεταζόμενη αλληλουχία ανάγνωσης, με σκοπό την τελική κατάταξη της αλληλουχίας αυτής[19, 20, 22].

Τα υπάρχοντα λογισμικά κατηγοριοποίησης με βάση την ομοιότητα αλληλουχιών προσφέρουν γενικά μεγαλύτερη ακρίβεια σε σχέση με τις αντίστοιχες αναλύσεις με βάση τα συνθετικά χαρακτηριστικά, όμως παρουσιάζουν συγκεκριμένα μειονεκτήματα όπως η ανάγκη ύπαρξης αξιόπιστων βιβλιοθηκών αναφοράς, διάφορες δυσκολίες στην κατασκευή των ταξινομικών δέντρων, προβλήματα που σχετίζονται με την ύπαρξη αλληλουχιών που ανήκουν σε οργανισμούς που δεν περιλαμβάνονται στις βιβλιοθήκες αναφοράς, αλλά κυρίως η τεράστια υπολογιστική δύναμη αλλά και ο χρόνος που απαιτούνται ειδικά κατά την πρώτη φάση σύγκρισης των αλληλουχιών υπό μελέτη, με αυτές της βάσης δεδομένων[12, 19].



Εικόνα 12 Τα κυριότερα επίπεδα της ταξινομικής ιεραρχίας.

2.6 Χαρακτηρισμός (Annotation)

Οι μικροβιακές κοινότητες μπορούν να αντιμετωπιστούν όχι μόνο ως ξεχωριστές ομάδες μικροβίων αλλά και ως συλλογές βιοχημικών λειτουργιών που επηρεάζονται από το περιβάλλον ή τον οργανισμό ξενιστή στον οποίο εντοπίζονται. Έτσι λοιπόν η μεταγονιδιωματική μπορεί να οδηγήσει στην ταυτοποίηση γονιδίων αλλά και λειτουργικών μονοπατιών που συνδέονται με μικροβιακές κοινότητες.

Αρχικά, η ταυτοποίηση των γονιδίων, δηλαδή η διαδικασία πρόγνωσης γονιδίων (**gene prediction ή gene calling**), μπορεί να γίνει είτε σε συναρμολογημένα contigs μεγάλου μεγέθους, είτε να βασιστεί σε επεξεργασία δεδομένων που δεν έχουν υποστεί συναρμολόγηση και πρόκειται ουσιαστικά για απλές αλληλουχίες ανάγνωσης[23]. Στην περίπτωση ύπαρξης μεγάλων συναρμολογημένων contigs είναι προτιμότερο να χρησιμοποιούνται συγκεκριμένες ροές λογισμικών για γονιδιωματικό σχολιασμό όπως τα RAST και IMG. Εάν επιχειρείται χαρακτηρισμός που στηρίζεται σε μη συναρμολογημένες αλληλουχίες ανάγνωσης ή μικρά contigs τότε ειδικά σχεδιασμένα λογισμικά για μεταγονιδιωματική ανάλυση είναι καταλληλότερα[11].

Πριν από αυτή την διαδικασία της πρόγνωσης των γονιδίων, προηγείται ο προσδιορισμός των ανοικτών πλαισίων ανάγνωσης (Open Reading Frames - ORFs), δηλαδή τμημάτων αλληλουχίας DNA που ξεκινάνε από ένα κωδικόνιο έναρξης και φτάνουν έως ένα κωδικόνιο λήξης και ουσιαστικά αποτελούν έναν τρόπο υπόθεσης για την ύπαρξη υποψήφιων κωδικών περιοχών αλληλουχίας DNA που οδηγούν στον σχηματισμό πρωτεϊνών. Η διαδικασία αυτή είναι βοηθητική ως προς την πρόγνωση γονιδίων και η παρουσία των ORFs, δεν σημαίνει αναγκαστικά ότι το τμήμα αυτό της αλληλουχίας μεταφράζεται πραγματικά προς την παραγωγή πρωτεϊνών.

Υπάρχουν δύο προσεγγίσεις στην πρόγνωση γονιδίων. Η πρώτη, γνωστή ως "evidence-based" προσέγγιση, στηρίζεται σε αναζητήσεις ομολογίας ώστε να ταυτοποιήσει γονίδια παρόμοια με άλλα που έχουν ήδη παρατηρηθεί. Παραδείγματα τέτοιας προσέγγισης είναι η απλές συγκρίσεις με βιβλιοθήκες πρωτεϊνών (NCBI nr, KEGG Orthology, COGs) μέσω BLAST ή συγκεκριμένα εργαλεία όπως τα CRITICA και Orpheus[12, 23]. Στην δεύτερη περίπτωση, την λεγόμενη "ab initio" προσέγγιση γονιδιακής πρόγνωσης, τα μεταγονιδιωματικά contigs ελέγχονται για την ύπαρξη γονιδίων που κωδικοποιούν την παραγωγή πρωτεϊνών (CDSs) όπως επίσης και για επαναληπτικά CRISPRs, ncRNA και tRNAs[23]. Υπάρχει ένας μεγάλος αριθμός λογισμικών εργαλείων που είναι σχεδιασμένα αποκλειστικά για την μεταγονιδιωματική πρόβλεψη των CDSs όπως τα FragGeneScan, MetaGeneMark, MetaGeneAnnotator/Metagene και τέλος το Ophelia. Επιπλέον, υπάρχει και ένας σημαντικός αριθμός εργαλείων για την πρόγνωση γονιδίων που δεν κωδικοποιούν

πρωτεΐνες αλλά παραδείγματος χάριν tRNAs ,CRISPRs όπως ήδη αναφέρθηκε, αλλά τα εργαλεία αυτά είναι πολυ πιθανό στις περιπτώσεις μεγάλων συνεχόμενων αλληλουχιών να απαιτούν σημαντική υπολογιστική δύναμη[11, 23].

Εκτός όμως από την γονιδιακή πρόγνωση, στο επίπεδο του χαρακτηρισμού θα πρέπει να γίνει ιδιαίτερη αναφορά και στον λειτουργικό χαρακτηρισμό (**functional annotation**). Ο λειτουργικός χαρακτηρισμός μεταγονιδιωματικών δεδομένων είναι κατά έναν τρόπο σημαντικά όμοιος με τον γονιδιακό χαρακτηρισμό και βασίζεται σε συγκρίσεις προβλεπόμενων γονιδίων με γνωστά γονίδια που έχουν προέλθει από προηγούμενες μελέτες σχολιασμού[12]. Είναι λοιπόν σημαντικό να τονιστεί το γεγονός ότι ο λειτουργικός χαρακτηρισμός δεν πραγματοποιείται σε ένα "de novo" πλαίσιο αλλά μέσω χαρτογράφησης σε ήδη γνωστές βιβλιοθήκες γονιδίων ή πρωτεϊνών[11]. Η δημιουργία λοιπόν των λειτουργικών προφίλ σε ένα σετ μεταγονιδιωματικών δεδομένων μπορεί να βασίζεται σε συγκρίσεις χαρτογράφησης με δεδομένα αναφοράς, είτε στο νουκλεοτιδικό επίπεδο είτε ακόμα και με αναζητήσεις σε βιβλιοθήκες μεταφρασμένων πρωτεϊνών[23].

Η πιο κοινή προσέγγιση για λειτουργική πρόγνωση που υιοθετείται από την πλειονότητα των λογισμικών που χρησιμοποιούνται για τον λειτουργικό χαρακτηρισμό, είναι η σύγκριση των προβλεπόμενων από τις αλληλουχίες πρωτεϊνών με ήδη υπάρχουσες πρωτεϊνικές βιβλιοθήκες αναφοράς, όπως η NCBI-nr, η UniProt/UniRef και η SMART. Υπάρχουν παραδείγματα λογισμικών που παρέχουν την δυνατότητα λειτουργικού χαρακτηρισμού μέσω μεταφρασμένης χαρτογράφησης όπως είναι τα MG-RAST, IMG/M και MEGAN, λογισμικά τα οποία προσφέρουν τη δυνατότητα λειτουργικής ανάλυσης με βάση διαφορετικά χαρακτηριστικά και μια ποικιλία βιβλιοθηκών[24]. Αν και στις περισσότερες των περιπτώσεων η απαίτηση σε υπολογιστική δύναμη είναι πάρα πολύ μεγάλη, ουσιαστικά πραγματοποιείται μια αναζήτηση ομοιότητας μέσω εργαλείων όπως το BLASTX στις πρωτεϊνικές βιβλιοθήκες αναφοράς, ώστε από το επίπεδο του γονιδίου να γίνει η μετάβαση στο επίπεδο των πεπτιδίων.

Υπάρχει ένας μεγάλος αριθμός βιβλιοθηκών αναφοράς που είναι δυνατό να χρησιμοποιηθούν ώστε να μελετηθεί το λειτουργικό πλαίσιο σε μεταγονιδιωματικά σετ δεδομένων με χαρακτηριστικά παραδείγματα τις KEGG, eggNOG, COG/KOG, PFAM, TIGRFAM καθώς και SEED[11]. Κάθε μια από αυτές τις βιβλιοθήκες αναφοράς, καλύπτει μια συγκεκριμένη ομάδα λειτουργικών χαρακτηριστικών και όπως γίνεται εύκολα κατανοητό, δεν υπάρχει μια ενιαία βιβλιοθήκη που θα παρέχει πληροφορίες για το σύνολο των χαρακτηριστικών λειτουργιών μιας μικροβιακής κοινότητας.

3. ΕΡΓΑΛΕΙΑ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ

3.1 Η πλατφόρμα GALAXY

Η έρευνα που σχετίζεται με τις επιστήμες της ζωής συνεχίζει εδώ και πολλά χρόνια να χαρακτηρίζεται από μια ολοένα και αυξανόμενη εντατικοποίηση δεδομένων. Μέσω των νέων πειραματικών τεχνικών υψηλής απόδοσης, ένα μεμονωμένο εργαστήριο μπορεί να δημιουργήσει πρωτογενή δεδομένα αδιανόητης μέχρι πριν από λίγα χρόνια κλίμακας. Αυτές οι εξελίξεις παρέχουν μια τεράστια ευκαιρία για βασική καθώς και εφαρμοσμένη έρευνα. Ωστόσο, έχουν οδηγήσει, επίσης, στη δημιουργία ενός είδους κρίσης για πολλούς επιστήμονες, αφού το να γίνει πλήρως κατανοητός αυτός ο πλούτος των δεδομένων απαιτεί σημαντική υποδομή αναλυτικής δύναμης. Χωρίς την υποστήριξη της βιοπληροφορικής, οι πειραματικοί βιολόγοι, οι οποίοι κατέχουν βασικές βιολογικές γνώσεις και εμπειρία και ως εκ τούτου έχουν τη δυνατότητα νέων ανακαλύψεων, δεν θα μπορούσαν να χρησιμοποιήσουν αποτελεσματικά τα διαθέσιμα στοιχεία.

Η πλατφόρμα GALAXY, ένα open-source λογισμικό, μπορεί να δώσει την λύση σε μια τέτοια πρόκληση παρέχοντας την απαραίτητη πληροφορική υποδομή. Αποτελεί ένα λογισμικό σύστημα που παρέχει τέτοιου είδους υποστήριξη μέσω ενός πλαισίου που προσφέρει στους μελετητές απλές διεπαφές (interfaces) πολυ δυνατών εργαλείων, ενώ ταυτόχρονα ρυθμίζει τις υπολογιστικές λεπτομέρειες. Προσφέρει λοιπόν ένα περιβάλλον στο οποίο οι αναλύσεις μπορούν να εφαρμοστούν διαδραστικά. Το πλαίσιο της πλατφόρμας GALAXY περικλείει ποιοτικά και σημαντικά υπολογιστικά εργαλεία για τα οποία έχουν δημιουργηθεί εύχρηστα και αισθητικά ικανοποιητικά περιβάλλοντα διεπαφής, ενώ ταυτόχρονα αποκρύπτονται οι λεπτομέρειες της υπολογιστικής και αποθηκευτικής ρύθμισης. Με τον τρόπο αυτό, περιορίζεται σε μεγάλο βαθμό η ανάγκη εξειδικευμένης πληροφορικής γνώσης κατά την πραγματοποίηση συνηθισμένων τύπων αναλύσεων μεγάλης έκτασης [25].

Η πλατφόρμα GALAXY διατίθεται τόσο ως δημόσια διαθέσιμη υπηρεσία στο διαδίκτυο που παρέχει εργαλεία γονιδιωματικής ανάλυσης, γονιδιωματικών συγκρίσεων και λειτουργικών συγκρίσεων αλλά και ως πακέτο διατιθέμενο προς τοπική χρήση. Έτσι λοιπόν είναι δυνατό να δημιουργηθούν τοπικοί διακομιστές (servers), μέσω εγκατάστασης της εφαρμογής GALAXY, ενώ στη συνέχεια μπορεί να ακολουθήσει η προσαρμογή αυτών ώστε να καλύπτονται οι συγκεκριμένες ανάγκες των ερευνητών. Αρχικά το πρόγραμμα της πλατφόρμας αυτής αναπτύχθηκε για την υποστήριξη γονιδιωματικών ερευνών, αλλά πλέον θεωρείται και χρησιμοποιείται ως ένα γενικό σύστημα διαχείρισης βιοπληροφορικών εργαλείων σε αντίστοιχες αναλύσεις.

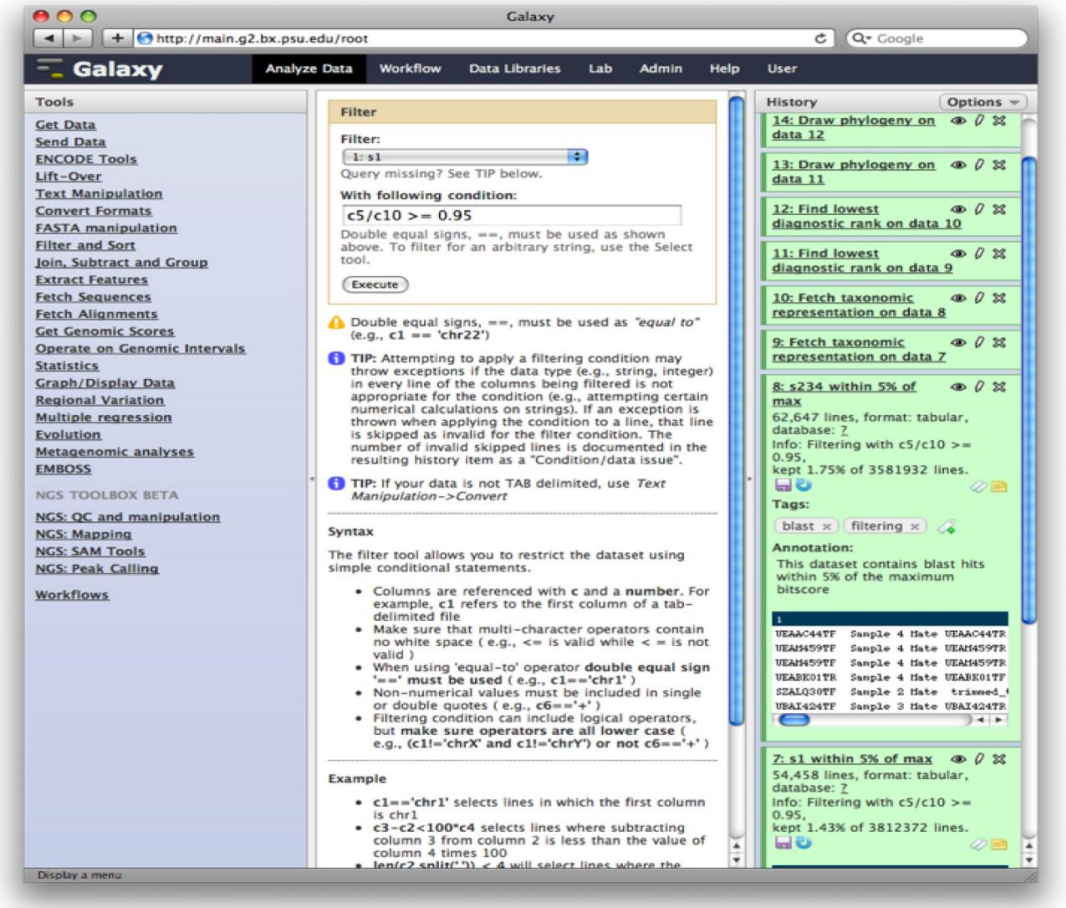
Όπως προαναφέρθηκε, στόχος του GALAXY είναι να δώσει την δυνατότητα σε απλούς χρήστες να πραγματοποιήσουν τις δικές τους υπολογιστικές αναλύσεις δίχως να είναι απαραίτητο να είναι ειδικοί στην υπολογιστική βιολογία ή γενικά στην επιστήμη της πληροφορικής. Η πλατφόρμα αυτή, προσθέτει τα απαραίτητα γραφικά περιβάλλοντα διεπαφής στην γονιδιωματική και μεταγονιδιωματική έρευνα, καθιστώντας την ανάλυση τέτοιων δεδομένων παγκοσμίως προσβάσιμη είτε μέσω του διαδικτύου είτε μέσω της τοπικής χρήσης, απελευθερώνοντας τους χρήστες από τις μικρολεπτομέρειες των ξεπερασμένων παραμέτρων μέσω γραμμής εντολών, της μορφοποίησης δεδομένων και των γλωσσών προγραμματισμού. Τα δεδομένα εισόδου και τα υπολογιστικά βήματα επιλέγονται από δυναμικά γραφικά μενού και τα αποτελέσματα παρουσιάζονται σε ενδιαφέροντα διαγράμματα και περιλήψεις που ενθαρρύνουν τη δημιουργία διαδραστικών ροών λογισμικών και εργαλείων. Τα εργαλεία αυτά μπορεί να αποτελούν κομμάτι σχεδόν οποιουδήποτε είδους λογισμικού και να έχουν γραφτεί σε οποιαδήποτε γλώσσα προγραμματισμού αλλά η πολυπλοκότητα τους κρύβεται τεχνηέντως μέσα στο GALAXY, επιτρέποντας στους χρήστες να ασχοληθούν κυρίως με επιστημονικούς παρά με τεχνικούς προβληματισμούς[26].

Ένας λοιπόν από τους στόχους της πλατφόρμας, είναι η προσβασιμότητα, που επιτυγχάνεται μέσω των παραπάνω χαρακτηριστικών, αλλά επιπλέον και από την δυνατότητα των χρηστών να προσθέσουν νέα εργαλεία, μέσω της διαμόρφωσης εργαλείων. Ο δημιουργός του εργαλείου δίνει πληροφορίες για το πώς θα πρέπει να τρέχει το εργαλείο συμπεριλαμβάνοντας και τις προδιαγραφές των δεδομένων και των παραμέτρων εισόδου αλλά και εξόδου. Μέσω αυτού του προσδιορισμού των προδιαγραφών, δημιουργείται το εκάστοτε περιβάλλον διεπαφής για το κάθε εργαλείο. Αν και ενδεχομένως αυτή η προσέγγιση να παρέχει μικρότερη ευελιξία από την απευθείας δημιουργία εργαλείων μέσω κάποιας γλώσσας προγραμματισμού, είναι αυτό το χαρακτηριστικό που λειτουργεί ως υπόστρωμα ώστε να εξασφαλιστεί η υπολογιστική προσβασιμότητα αλλά και οι δυο επιπλέον στόχοι της πλατφόρμας GALAXY, η επαναληψιμότητα και η διαφάνεια[27].

Το GALAXY επιτρέπει στους χρήστες να συγκεντρώσουν αλλά και να χειριστούν δεδομένα από τις υπάρχουσες πηγές, μέσω μιας ευρείας ποικιλίας επιλογών. Κάθε κίνηση του χρήστη καταγράφεται και αποθηκεύεται στο ιστορικό της πλατφόρμας, ένα χαρακτηριστικό κλειδί για την σημαντικότητα του GALAXY[28]. Το GALAXY καταγράφει αυτόματα τις εισόδους, τα εργαλεία, τις παραμέτρους και τις ρυθμίσεις που χρησιμοποιούνται για κάθε στάδιο σε μια ανάλυση και έτσι εξασφαλίζεται ότι κάθε αποτέλεσμα μπορεί να αναπαραχθεί με ακρίβεια και να αξιολογηθεί αργότερα. Αυτό το χαρακτηριστικό λειτουργίας έχει σημαντικές βραχυπρόθεσμες και μακροπρόθεσμες συνέπειες. Σε βραχυπρόθεσμο επίπεδο, μπορούν να εξερευνηθούν διαφορετικές παράμετροι και όρια και μόλις η ανάλυση ολοκληρωθεί, το αρχείο μνήμης του GALAXY θα εξαλείψει κάθε ασάφεια ως

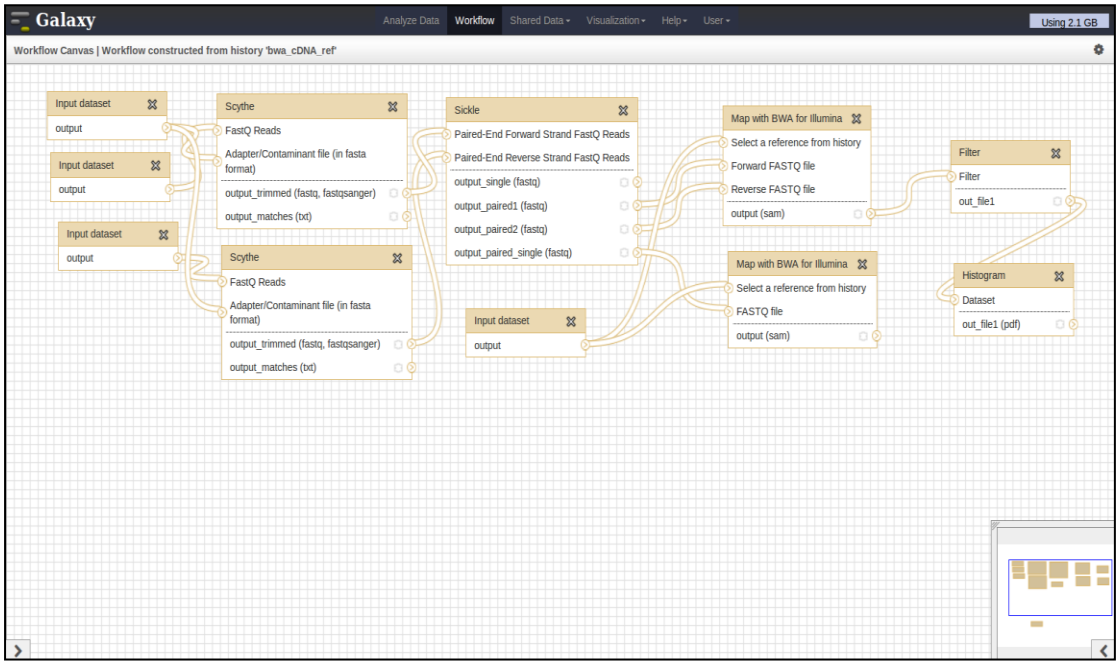
προς το ποιο αποτέλεσμα προέκυψε χρησιμοποιώντας τις συγκεκριμένες ρυθμίσεις. Σε μακροπρόθεσμη βάση, το ιστορικό του GALAXY είναι ανεκτίμητο, στην περίπτωση που μια μη προγραμματισμένη ανάλυση απαιτεί εκτέλεση[26].

Ουσιαστικά η πλατφόρμα, ομαδοποιεί μια σειρά από βήματα ανάλυσης σε ένα ιστορικό και οι χρήστες μπορούν να δημιουργήσουν, να αντιγράψουν και να μετατρέψουν τα ιστορικά. Όλα τα σετ δεδομένων σε ένα ιστορικό, αρχικά, ενδιάμεσα, τελικά, είναι προσβάσιμα και εμφανή και ο χρήστης μπορεί να ξανατρέξει κάθε βήμα της ανάλυσης. Επιπροσθέτως, ο χρήστης έχει την δυνατότητα σχολιασμού και χαρακτηρισμού κάθε βήματος ώστε πέρα από την αποθήκευση δεδομένων και παραμέτρων να είναι προφανές τι συμβαίνει σε κάθε βήμα και γιατί κάθε βήμα είναι σημαντικό. Η δημιουργία ετικετών για την περιγραφή και τον σχολιασμό τμημάτων του εκάστοτε εργαλείου συμβάλει στην ζητούμενη επαναληψιμότητα καθώς διευκολύνει την αναζήτηση εργαλείων αλλά και διεργασιών[27].



Εικόνα 13 Το βασικό παράθυρο εργασίας του GALAXY. Αριστερά διακρίνεται η στήλη των διαθέσιμων εργαλείων, στην μέση η στήλη των πληροφοριών και δεξιά η στήλη του ιστορικού. Το ιστορικό αποτελεί βασικό υπόστρωμα για την εξασφάλιση της επαναληψιμότητας, παρέχοντας πληροφορίες προέλευσης δεδομένων αλλά και δίνοντας στον χρήστη την δυνατότητα να επιλέξει υπάρχοντα workflows, να ξανατρέξει δουλειές, να τοποθετήσει επεξηγηματικές ετικέτες, να σχολιάσει βήματα και να εμφανίσει αποτελέσματα[27].

Τα παραπάνω χαρακτηριστικά αρκούν ώστε να εξασφαλιστεί η επαναληψιμότητα αλλά δεν αρκούν για να χαρακτηρίσουν ως εύκολη διαδικασία την επανάληψη μιας ανάλυσης. Αυτό διευκολύνεται μέσω του συστήματος του GALAXY για την δημιουργία ροών εργασίας (workflows), δηλαδή ροών που αποτελούν πρότυπα και εφαρμόζονται σε διαφορετικά δεδομένα. Η δημιουργία ενός workflow απαιτεί προσχεδιασμό και οργάνωση, ενώ η σωστή σύνδεση μεταξύ των εργαλείων είναι μεγάλης σημασίας. Κάθε φορά που τρέχει ένα workflow, χρησιμοποιούνται τα ίδια εργαλεία με τις ίδιες παραμέτρους.



Εικόνα 14 Παράθυρο δημιουργίας και επεξεργασίας ροών εργασίας στο GALAXY.

Τέλος, όπως ήδη αναφέρθηκε, η διαφάνεια είναι ένας από τους βασικούς στόχους αυτής της πλατφόρμας. Είναι επιθυμητό, οι χρήστες να έχουν την δυνατότητα να μοιράζονται και να επικοινωνούν τα πειραματικά τους αποτελέσματα και δεδομένα εξόδου με έναν τρόπο πρακτικό και ουσιώδη. Η διαφάνεια προωθείται στο GALAXY μέσω ενός μοντέλου διαμοιρασμού αντικειμένων που σχετίζονται με την πλατφόρμα (δεδομένα, ιστορικά, ροές εργασίας) και δημόσιων αποθηκών, ένα διαδικτυακό πλαίσιο παρουσίασης δημοσιευμένων ή μοιρασμένων δεδομένων καθώς και τις Σελίδες (Pages) του GALAXY, που αποτελούν ειδικά διαδικτυακά αρχεία που επιτρέπουν στους χρήστες να συζητούν και να παρουσιάζουν ολόκληρα υπολογιστικά πειράματα. Περιλαμβάνουν κείμενο και γραφήματα χαρακτηριστικά της ανάλυσης που πραγματοποιήθηκε, καθώς και συγκεντρωτικά στοιχεία όπως σετ δεδομένων, ιστορικά αλλά και workflows, που διευκολύνουν κάθε χρήστη να κατανοήσει ένα πείραμα σε κάθε λεπτομερειακό βάθος[27].

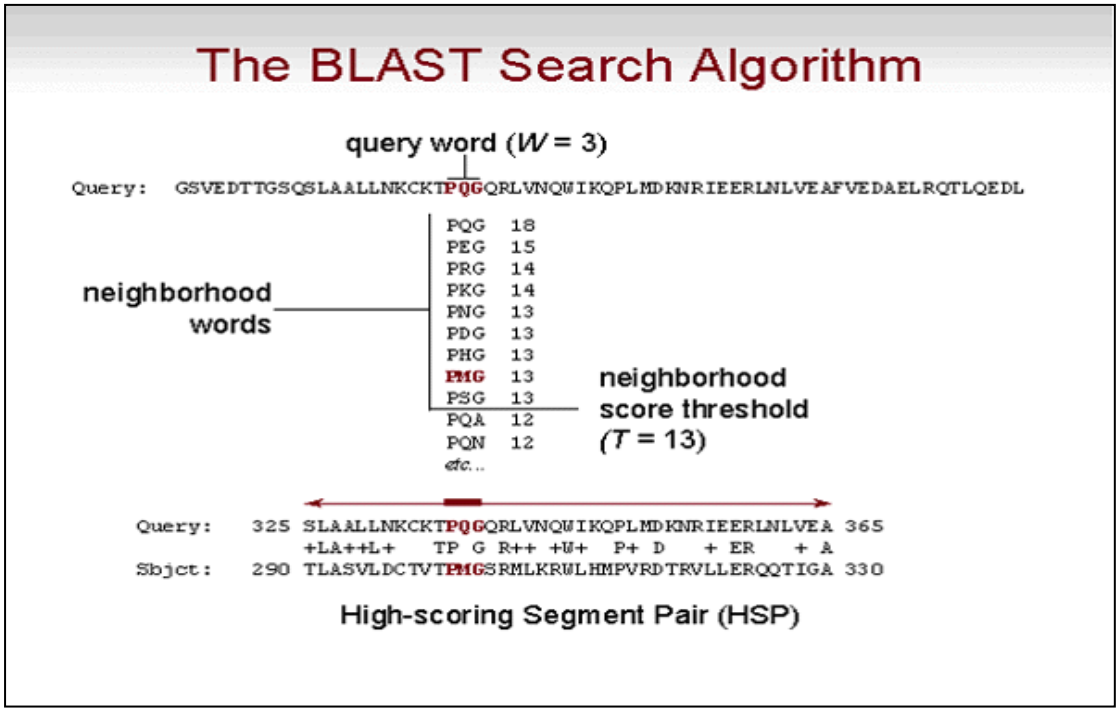
3.2 BLAST: Ένα εργαλείο σύγκρισης αλληλουχιών

Το Basic Local Alignment Search Tool (BLAST) [29] είναι ένα πρόγραμμα αναζήτησης ομοιότητας αλληλουχίας που μπορεί να χρησιμοποιηθεί μέσω ενός περιβάλλοντος διαδικτυακής διεπαφής ή ακόμα και ως αυτόνομο (stand-alone) εργαλείο για την σύγκριση ζητούμενων αλληλουχιών ενός χρήστη με σύνολα αλληλουχιών που βρίσκονται αποθηκευμένα σε βάσεις δεδομένων. Υπάρχουν εκδόσεις του BLAST που συγκρίνουν αλληλουχίες-ερωτήματα (query sequences) πεπτιδικής φύσης σε βάσεις δεδομένων πρωτεϊνών, αλληλουχίες-ερωτήματα νουκλεοτιδικών αλληλουχιών σε βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών, καθώς και εκδόσεις που μεταφράζουν αλληλουχίες-ερωτήματα ή βάσεις δεδομένων νουκλεοτιδίων σε όλα τα έξι πλαίσια ανάγνωσης και ακολουθεί σύγκριση σε σχέση με βάσεις δεδομένων πρωτεϊνών ή ερωτήματα.

Το BLAST λειτουργεί με βάση έναν ευρετικό αλγόριθμο που εντοπίζει ουσιαστικά μικρά όμοια τμήματα ανάμεσα σε δυο αλληλουχίες και επιχειρεί να ξεκινήσει την διαδικασία της ομοπαράθεσης, από τα τμήματα αυτά. Το εργαλείο αυτό λοιπόν εντοπίζει παρόμοιες αλληλουχίες όχι συγκρίνοντας τις στο σύνολο τους, αλλά μέσω του αρχικού εντοπισμού μικρών κοινών τμημάτων και αποτελεί ένα από τα πιο διαδεδομένα εργαλεία στον κόσμο της βιοπληροφορικής εδώ και αρκετά χρόνια. Εκτός από την εκτέλεση ομοπαραθέσεων (alignments), το BLAST παρέχει μια "προσδοκώμενη" αξία (expect value), δηλαδή στατιστικές πληροφορίες σχετικά με τη σημασία της κάθε ομοπαράθεσης[30].

Αξίζει βέβαια να γίνει μια σύντομη αναφορά στον βασικό αλγόριθμο λειτουργίας αυτού του εργαλείου. Ενώ γίνεται προσπάθεια για τον εντοπισμό ομοιοτήτων ανάμεσα στις αλληλουχίες, κάποια σετ κοινών γραμμάτων, γνωστά και ως λέξεις (words) παίζουν σημαντικό ρόλο. Η αρχική αναζήτηση γίνεται για μια λέξη μήκους "W" που θα πετύχει τουλάχιστον σκορ "T" σε σύγκριση με την αλληλουχία-ερώτημα χρησιμοποιώντας μια μήτρα υποκατάστασης. Τα hits της λέξης κατόπιν επεκτείνονται σε κάθε κατεύθυνση, σε μια προσπάθεια να δημιουργηθεί μια ομοπαράθεση με ένα σκορ που υπερβαίνει το όριο του "S". Η παράμετρος "T" υπαγορεύει την ταχύτητα και την ευαισθησία της αναζήτησης. Οι ακολουθίες εισόδου θα πρέπει να είναι σε μορφή FASTA ή Genbank ενώ τα δεδομένα εξόδου του BLAST είναι δυνατό να παραδοθούν σε μια ποικιλία μορφών. Αυτές οι μορφές περιλαμβάνουν HTML, απλό κείμενο και μορφοποίηση XML. Στην ιστοσελίδα του NCBI, η προεπιλεγμένη μορφή για την έξοδο είναι HTML. Κατά την εκτέλεση μιας αναζήτησης BLAST στην ιστοσελίδα του NCBI, τα αποτελέσματα δίνονται σε γραφική μορφή που δείχνει τα hits που βρέθηκαν, έναν πίνακα που δείχνει αναγνωριστικούς κωδικούς αλληλουχιών που αντιστοιχούν στα χτυπήματα αυτά συνοδευόμενα από δεδομένα που σχετίζονται με τα σκορ που επετεύχθησαν, καθώς και

ομοπαραθέσεις για την ακολουθία που ενδιαφέρει και τα χτυπήματα που λήφθησαν με τις αντίστοιχες βαθμολογίες BLAST για αυτά.



Εικόνα 15 Ο βασικός αλγόριθμος αναζήτησης του BLAST, όπως περιγράφεται και παραπάνω στο κείμενο, με ενδεικτικές τιμές παραμέτρων και παράδειγμα αλληλουχίας-ερωτήματος.

Το Εθνικό Κέντρο Πληροφοριών Βιοτεχνολογίας (NCBI) παρουσίασε για πρώτη φορά το BLAST το 1989. Το NCBI συνέχισε να διατηρεί και ενημερώνει το BLAST από την πρώτη αυτή έκδοση. Το 1997 εισήχθησαν και έγιναν διαθέσιμες δημόσια, βασικές εφαρμογές του εργαλείου αυτού (blastall, blastpgp) αλλά οι αρχικές αυτές εφαρμογές BLAST του 1997 δεν διέθεταν πολλά χαρακτηριστικά που σήμερα θεωρούνται δεδομένα. Εντός τριών ετών από την αρχική δημόσια διάθεση, το BLAST είχε τροποποιηθεί ώστε να μπορεί να χειριστεί βάσεις δεδομένων με περισσότερα από 2 δισεκατομμύρια γράμματα, να περιορίσει μια αναζήτηση μέσω μιας λίστας αναγνωριστικών GenInfo (GIs) και να πραγματοποιήσει αναζητήσεις ταυτόχρονα σε πολλαπλές βάσεις δεδομένων[31]. Το 2009, η NCBI εισήγαγε μια νέα έκδοση αυτόνομων (stand-alone) εφαρμογών BLAST (BLAST+). Οι BLAST+ εφαρμογές έχουν μια σειρά από βελτιώσεις που επιτρέπουν ταχύτερες αναζητήσεις, καθώς και μεγαλύτερη ευελιξία σε μορφές δεδομένων εξόδου αλλά και εισόδου, πάνω στα οποία βασίζεται και η αναζήτηση. Αυτές οι βελτιώσεις περιλαμβάνουν: τη διάσπαση των μεγαλύτερων αλληλουχιών-ερωτημάτων, έτσι ώστε να μειωθεί η απαίτηση για χρήση μνήμης και να επωφεληθούν από τις σύγχρονες αρχιτεκτονικές μεθόδους σχεδιασμού CPU, τη χρήση ενός πίνακα βάσης δεδομένων για να επιταχυνθεί δραματικά η αναζήτηση, την δυνατότητα να αποθηκευτεί μια "στρατηγική αναζήτησης" που μπορεί να χρησιμοποιηθεί

αργότερα για να ξεκινήσει μια νέα αναζήτηση καθώς και μεγαλύτερη ευελιξία στη διαμόρφωση των αποτελεσμάτων σε μορφή πίνακα[32].

Η λειτουργικότητα των BLAST+ εφαρμογών οργανώνεται από τον τύπο αναζήτησης που επιθυμείται. Παραδείγματος χάριν, υπάρχει μια εφαρμογή **blastp** που συγκρίνει πεπτιδικές αλληλουχίες σε βάσεις δεδομένων πρωτεϊνών. Επίσης, η εφαρμογή **blastn** που συγκρίνει νουκλεοτιδικές αλληλουχίες σε νουκλεοτιδικές βάσεις δεδομένων, ενώ υπάρχει και η εφαρμογή **blastx** που μεταφράζει μια νουκλεοτιδική αλληλουχία-ερώτημα σε έξι πλαίσια και την ψάχνει σε μια πρωτεϊνική βάση δεδομένων. Επιπροσθέτως, υπάρχουν επιλογές όπως οι εφαρμογή **tblastx** η οποία συγκρίνει μεταφρασμένες νουκλεοτιδικές αλληλουχίες σε σχέση με μεταφρασμένες αλληλουχίες στόχους ή μεταφρασμένες νουκλεοτιδικές βάσεις δεδομένων με σκοπό τον εντοπισμό πολυ μακρινών σχέσεων μεταξύ νουκλεοτιδικών αλληλουχιών, η εφαρμογή **tblastn** που αναζητά μια αλληλουχία-ερώτημα πρωτεϊνικής φύσης έναντι νουκλεοτιδικών αλληλουχιών ή νουκλεοτιδικών βάσεων δεδομένων ενώ η μετάφραση πραγματοποιείται κατά τη διάρκεια του χρόνου αναζήτησης, ενώ τέλος παρέχεται και η δυνατότητα αναζήτησης μιας πρωτεϊνικής αλληλουχίας σε σχέση με σετ πρωτεϊνικών προφίλ όπως είναι η βάση δεδομένων CDD (conserved domain database) μέσω της εφαρμογής **rpsblast**. Ανάλογα με τα δεδομένα εισόδου, τον στόχο της αναζήτησης και τις παραμέτρους που ρυθμίζουν κάθε αναζήτηση, ο χρήστης έχει την δυνατότητα να επιλέξει την κατάλληλη εφαρμογή, ενώ ακόμα, του δίνεται η δυνατότητα επιλογής υποεφαρμογών ή σωστότερα ειδικών εργασιών (tasks) της κάθε βασικής εφαρμογής με στόχο την ακριβέστερη αναζήτηση σε ειδικές περιπτώσεις δεδομένων. Χαρακτηριστικά παραδείγματα αποτελούν τα **blastp-short, blastn-short, megablast, dc-megablast**, ειδικές εργασίες που προσαρμόζουν αυτόματα τις παραμέτρους των βασικών εφαρμογών για συγκεκριμένες περιπτώσεις αναζητήσεων. Ο εκάστοτε ερευνητής μπορεί επίσης να δημιουργήσει ο ίδιος ειδικές BLAST βάσεις δεδομένων τις οποίες μπορεί να χρησιμοποιεί στις αναζητήσεις του (**Makeblastdb, Makeprofiledb, Makembindex**).

Αυτός ο τρόπος οργάνωσης είναι διαφορετικός από εκείνον των εφαρμογών που κυκλοφόρησαν για πρώτη φορά το 1997 (π.χ., blastall) τα οποία και υποστήριζαν όλα τα είδη των αναζητήσεων με μία μόνο εφαρμογή, αλλά μοιάζει με τον τρόπο οργάνωσης του BLAST στην ιστοσελίδα του NCBI. Ένα πλεονέκτημα αυτού του σχεδιασμού είναι ότι η κάθε εφαρμογή διαθέτει μόνο τις επιλογές που σχετίζονται με τις αναζητήσεις που εκτελεί. Επιπρόσθετα, κάθε εφαρμογή μπορεί να συγκρίνει μια αλληλουχία-ερώτημα σε σχέση με ένα σύνολο αλληλουχιών FASTA σε ένα μόνο αρχείο, παρακάμπτοντας την ανάγκη να δημιουργούνται ειδικές BLAST βάσεις δεδομένων σε περιπτώσεις μελέτης μικρών συνόλων ή και συνόλων που σπάνια αποτελούν μέρος μιας ανάλυσης. Τέλος, μια νέα «εξ'αποστάσεως» επιλογή

(remote) επιτρέπει σε κάθε εφαρμογή να στείλει προς εκτέλεση μια αναζήτηση στους διακομιστές του NCBI[32].

3.3 MEGAN: Ένα εργαλείο ταξινομικής, λειτουργικής και συγκριτικής ανάλυσης

Το MEGAN (MEtaGenome ANalyzer) είναι ένα εργαλείο για την ανάλυση αλληλουχιών μεταγονιδιωματικών δεδομένων, επιτρέποντας στο χρήστη να διερευνήσει με τρόπο διαδραστικό το ταξινομικό και λειτουργικό περιεχόμενο ενός δείγματος. Υποστηρίζει, επίσης, τη σύγκριση πολλαπλών δειγμάτων τόσο σε ταξινομικό όσο και σε λειτουργικό επίπεδο. Το πρόγραμμα παρουσιάστηκε για πρώτη φορά το 2007[33] και αρχικά αναπτύχθηκε ώστε να χρησιμοποιηθεί στην ανάλυση μικροβιακής κοινότητας που υπήρχε σε ένα δείγμα οστών μαμούθ[34] ενώ η τέταρτη έκδοση του δημοσιεύτηκε το 2011[35]. Αυτή τη στιγμή είναι ενεργή η πέμπτη έκδοση του MEGAN, ενώ πραγματοποιείται μια συνεχής προσπάθεια βελτίωσης, διόρθωσης και σταθεροποίησης της με αποτέλεσμα η ακριβής έκδοση λειτουργίας να θεωρείται προσωρινά η MEGAN V5.5.3.

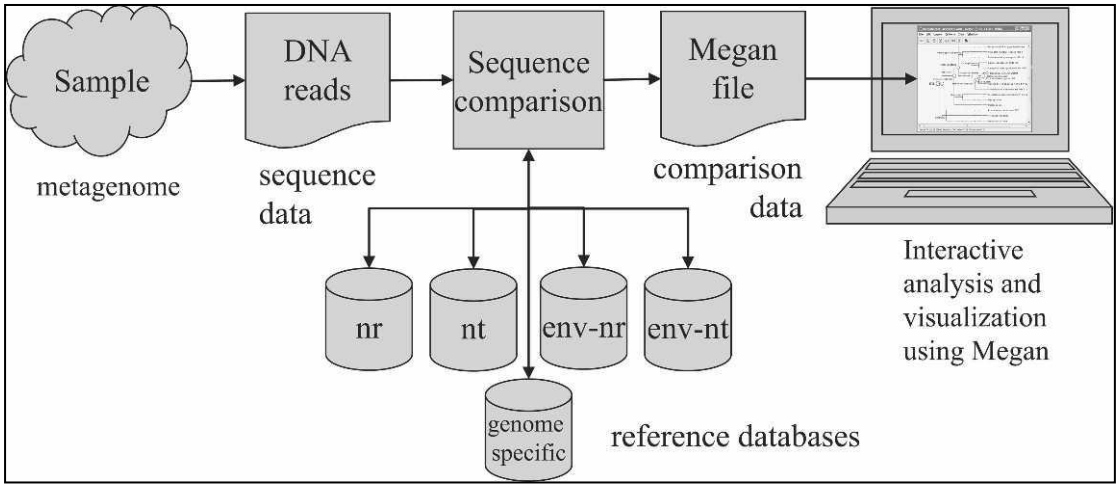
Σε ένα στάδιο προεπεξεργασίας, πρέπει να πραγματοποιηθεί μία σύγκριση αλληλουχίας όλων των αλληλουχιών ανάγνωσης με μία κατάλληλη βάση δεδομένων αλληλουχιών αναφοράς DNA ή και πρωτεϊνών για να παραχθεί ένα αρχείο εισόδου για το πρόγραμμα. Στο στάδιο αυτό, δημιουργούνται λοιπόν αρχεία εισόδου τα οποία προέρχονται από εργαλεία όπως το BLAST ή άλλα παρόμοια εργαλεία σύγκρισης αλληλουχιών. Το MEGAN είναι ένα εργαλείο κατάλληλο για ανάλυση αλληλουχιών DNA (μεταγονιδιωματικά δεδομένα), αλληλουχιών ανάγνωσης RNA (μεταμεταγραφικά δεδομένα), πεπτιδικών αλληλουχιών (μεταπρωτεομικά δεδομένα) και επιπροσθέτως, χρησιμοποιώντας ένα κατάλληλο αρχείο συνωνύμων που χαρτογραφεί και αντιστοιχεί SILVA ids σε ids ταξινομικών ομάδων, είναι δυνατό να χρησιμοποιηθεί στην ανάλυση 16S rRNA δεδομένων (αλληλούχιση αναδιπλασιασμού)[36]. Το MEGAN υποστηρίζει έναν αριθμό διαφορετικών μορφών αρχείων εισόδου, με κυριότερα τα αρχεία BLAST (απλό κείμενο, πίνακας, XML) καθώς και άλλους τύπους αρχείων όπως τα SAM, RapSearch2, RDP, NBC, QIIME ενώ τέλος υποστηρίζονται και αρκετές μορφές αρχείων CSV (comma-separated value). Τα δεδομένα που υποβάλλονται σε επεξεργασία αποθηκεύονται σε ένα RMA (read-match archive) αρχείο που περιέχει όλες τις αλληλουχίες ανάγνωσης (reads) και όλες τις αντιστοιχίες (matches) σε ένα συμπιεσμένο και μορφοποιημένο είδος αρχείου. Τα αποτελέσματα των αναλύσεων που παράγονται από το MEGAN μπορούν να εξαχθούν σε διάφορες μορφές CSV και όλες οι απεικονίσεις που παρέχονται από το πρόγραμμα μπορούν να εξαχθούν σε ένα ευρύ φάσμα γραφικών μορφών και επιλογών. Το πρόγραμμα παρέχει επίσης εργαλεία αναζήτησης για τον εντοπισμό ταξινομικών ομάδων και γονιδίων ενδιαφέροντος[37].

Κατά την εκκίνηση, το MEGAN διαβάζει και εμφανίζει υπό την μορφή δέντρου πρώτα την τρέχουσα ταξινόμηση κατά το NCBI (που αποτελείται από πάνω από ένα εκατομμύριο ταξινομικές ομάδες). Αυτό αποτελεί μια πρώτη εφαρμογή του προγράμματος καθώς διευκολύνει τη διαδραστική εξερεύνηση της NCBI ταξινόμησης. Η κύρια ωστόσο εφαρμογή του προγράμματος όπως ήδη αναφέρθηκε, είναι να αναλύσει το αποτέλεσμα μιας σύγκρισης μέσω BLAST ενός συνόλου αλληλουχιών ανάγνωσης έναντι μιας ή περισσότερων βάσεων δεδομένων αναφοράς, τυπικά χρησιμοποιώντας BLASTN, BLASTX ή BLASTP εφαρμογές για σύγκριση με βάσεις δεδομένων όπως οι NCBI-NT, NCBI-NR ή ακόμα και βάσεις δεδομένων που σχετίζονται με συγκεκριμένα είδη γονιδιώματος. Το αποτέλεσμα μιας τέτοιας ανάλυσης είναι η εκτίμηση του ταξινομικού περιεχομένου (προφίλ ταξινομικών ειδών) του δείγματος από το οποίο συλλέχθηκαν οι αλληλουχίες ανάγνωσης. Το πρόγραμμα χρησιμοποιεί μια σειρά από διαφορετικούς αλγόριθμους ώστε να τοποθετήσει τις αλληλουχίες ανάγνωσης μέσα στην ταξινόμηση αντιστοιχίζοντας καθεμία από αυτές σε μια ταξινομική ομάδα, σε κάποιο επίπεδο της NCBI ιεραρχίας, με βάση τα χτυπήματα (hits) αυτών σε ήδη γνωστές αλληλουχίες αναφοράς, όπως αυτά καταγράφονται στο αρχείο εισόδου BLAST. Φυσικά, η αναφορά γίνεται κυρίως για τα BLAST αρχεία καθώς αποτελούν τη συχνότερα χρησιμοποιούμενη φόρμα αρχείων εισόδου. Παρόμοιες διαδικασίες ισχύουν και για τα υπόλοιπα αρχεία εισόδου που έχουν αναφερθεί.

Η ταξινόμηση κατά το NCBI παρέχει μοναδικά ονόματα και κωδικούς-ταυτότητες για έναν τεράστιο αριθμό ταξινομικών ομάδων που περιέχουν μεταξύ άλλων προκαρυωτικούς οργανισμούς, ζώα, φυτά αλλά και ιούς. Οι ξεχωριστές διαφορετικές ταξινομικές ομάδες ομαδοποιούνται ιεραρχικά σε κλάδους και σε επίπεδα Βασιλείου, Φύλου, Τάξης, Σειράς, Οικογένειας, Γένους και Είδους (καθώς και κάποια ενδιάμεσα ανεπίσημα επίπεδα).

Αξίζει να γίνει μια αναφορά και στις βάσεις δεδομένων που χρησιμοποιούνται κατά το στάδιο της δημιουργίας των αρχείων εισόδου του MEGAN. Κυρίως, πέρα από ειδικά μορφοποιημένες βάσεις δεδομένων που δημιουργούνται από τους χρήστες ανάλογα με τις ανάγκες τους, δύο είναι οι ευρείες βάσεις δεδομένων που χρησιμοποιούνται κατά την BLAST σύγκριση αλληλουχιών με σκοπό την περαιτέρω ανάλυση μέσω MEGAN. Η NCBI-NR είναι μια (μη περιττή - non redundant) βάση δεδομένων αλληλουχιών πρωτεΐνης, η οποία είναι διαθέσιμη στην ιστοσελίδα του NCBI, ενημερώνεται συχνά, περιέχει πληροφορίες μεταβολικών μονοπατιών καθώς και λειτουργικά συνδεδεμένες ταξινομικές πληροφορίες[38] . Περιέχει εγγραφές από βάσεις δεδομένων όπως οι GenPept, Swissprot, PIR, PDF, PDB και RefSeq. Χαρακτηρίζεται ως μη-περιττή υπό την έννοια ότι οι ίδιες ακολουθίες συγχωνεύονται σε μία εγγραφή. Επίσης, η NBCI-NT βάση δεδομένων νουκλεοτιδικών αλληλουχιών, η οποία είναι διαθέσιμη επίσης στην ιστοσελίδα του NCBI και περιέχει εγγραφές από την GenBank και χαρακτηρίζεται ως μη - μη

περιττή. Περιέχει αμετάφραστες γονιδιακές κωδικοποιές αλληλουχίες καθώς και αλληλουχίες mRNA[36].

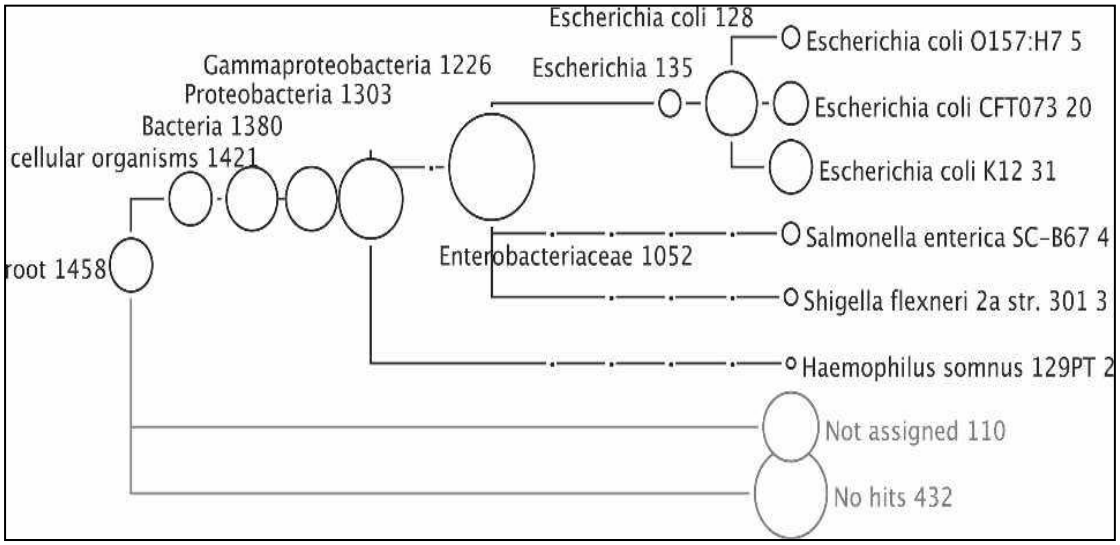


Εικόνα 16 Σχηματική παρουσίαση της πορείας της μεταγονιδιωματικής ανάλυσης. Από το δείγμα, στην παραγωγή των αλληλουχιών ανάγνωσης, την σύγκριση αυτών σε σχέση με βάσεις δεδομένων και την διαδραστική ανάλυση και παρουσίαση που ακολουθεί από το MEGAN[33].

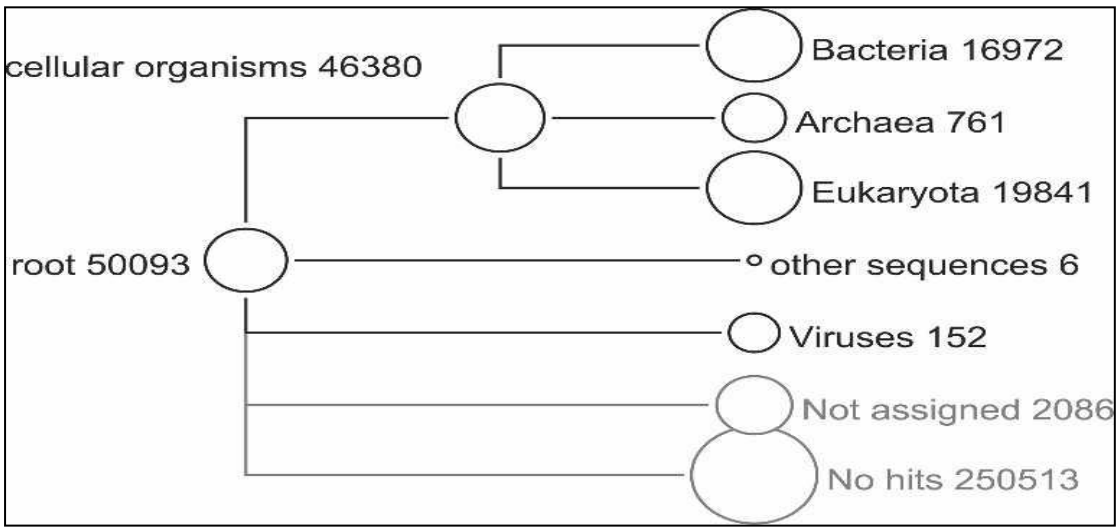
Έτσι λοιπόν, η κύρια εφαρμογή που βρίσκει το MEGAN και το βασικότερο πρόβλημα που καλείται να επιλύσει είναι η δημιουργία ταξινομικών προφίλ μέσω της αντιστοίχισης αλληλουχιών ανάγνωσης που προέρχονται από ένα μεταγονιδιωματικό πείραμα αλληλούχισης σε κατάλληλες ταξινομικές ομάδες που ανήκουν στην ταξινόμηση NCBI, μια διαδικασία που έχει αναφερθεί σε προηγούμενο κεφάλαιο ως ταξινομική κατηγοριοποίηση (taxonomic binning). Η προσέγγιση του εργαλείου MEGAN σε αυτό το πρόβλημα ονομάζεται αλγόριθμος **LCA** (**Lowest Common Ancestor algorithm**). Το πρώτο βήμα είναι γνωστό. Πραγματοποιείται σύγκριση των αλληλουχιών ανάγνωσης που ενδιαφέρουν με μια βάση δεδομένων αλληλουχιών αναφοράς μέσω ενός εργαλείου όπως το BLAST. Στη συνέχεια, γίνεται επεξεργασία των δεδομένων ώστε να προσδιοριστούν όλα τα πιθανά χτυπήματα των αλληλουχιών ανάγνωσης σε ταξινομικές ομάδες. Για κάθε αλληλουχία ανάγνωσης ***r***, ας θεωρηθεί ότι ***H*** είναι το σύνολο όλων των ταξινομικών ομάδων που χτυπάει το κάθε ***r***. Τέλος, εντοπίζεται ο χαμηλότερος κόμβος ***v*** στην ταξινόμηση του NCBI που περιλαμβάνει το σύνολο των ταξινομικών ομάδων ***H*** για το οποίο υπάρχουν hits και αντιστοιχίζεται η αλληλουχία ανάγνωσης ***r*** στην ταξινομική ομάδα που αντιπροσωπεύεται από τον κόμβο ***v***.

Κατά την προσέγγιση αυτή, κάθε αλληλουχία αντιστοιχίζεται σε μια ταξινομική ομάδα. Εάν η αλληλουχία ευθυγραμμίζεται πολυ συγκεκριμένα σε ένα μόνο ταξινομικό σύνολο τότε αντιστοιχίζεται αποκλειστικά και μόνο σε αυτό. Όσο μειώνεται η ακρίβεια με την οποία ένα read χτυπάει μια ταξινομική ομάδα, τόσο ψηλότερα τοποθετείται στην ταξινόμηση NCBI, ενώ σε ακραίες περιπτώσεις μπορεί να τοποθετηθεί και στον κόμβο που θεωρείται ως ρίζα (root) της ταξινόμησης.

Τέλος, εάν μια αλληλουχία ανάγνωσης ταιριάζει με δύο ταξινομικές ομάδες a και b όπου η a είναι πρόγονος της b στην ταξινόμηση NCBI, τότε κρατιέται μόνο το ταίριασμα με την ομάδα b ως πιο συγκεκριμένο και αυτό της ομάδας a απορρίπτεται. Φυσικά, η λειτουργία του αλγορίθμου αυτού, χαρακτηρίζεται και ρυθμίζεται από μια πληθώρα παραμέτρων και κατωφλιών που πρέπει να καθοριστούν ώστε η αναζήτηση να παρέχει αξιόπιστα και επιθυμητά αποτελέσματα και που θα παρουσιαστούν αναλυτικά σε επόμενο κεφάλαιο.



Εικόνα 17 Ταξινομική ανάλυση 2000 αλληλουχιών ανάγνωσης που προέρχονται από E. coli K12 χρησιμοποιώντας αλληλούχιση Roche GS20, μέσω σύγκρισης με την εφαρμογή BLASTX στην NCBI-NR βάση δεδομένων[33]. Μια ενδεικτική εικόνα ενός ταξινομικού δέντρου, ανάλυσης χαμηλού επιπέδου όπως παρουσιάζεται στο βασικό παράθυρο του MEGAN.



Εικόνα 18 Περίληψη υψηλού επιπέδου της ανάλυσης MEGAN συνόλου δεδομένων που σχετίζονται με μαμούθ, με βάση μια σύγκριση BLASTX των 302.692 αλληλουχιών ανάγνωσης σε σχέση με την NCBI-NR βάση δεδομένων[33]. Το υψηλό επίπεδο της ανάλυσης σχετίζεται με το βάθος του ταξινομικού δέντρου καθώς οι πιο απομακρυσμένοι κόμβοι στην περίπτωση αυτή τοποθετούνται κοντά στην ρίζα της ταξινόμησης.

Στις εικόνες 17 και 18, κάθε κύκλος (κόμβος) αντιπροσωπεύει μια ταξινομική ομάδα της ταξινόμησης NCBI και το μέγεθος του εξαρτάται από το πόσες αλληλουχίες ανάγνωσης έχουν αντιστοιχιστεί σε αυτόν. Εκτός από το όνομα της ταξινομικής ομάδας κάθε κόμβος χαρακτηρίζεται και από την ύπαρξη ενός αθροιστικού αριθμού που δείχνει πόσες αλληλουχίες ανάγνωσης έχουν αντιστοιχιστεί σε αυτόν τον κόμβο ή ακόμα και κάτω από αυτόν[39].

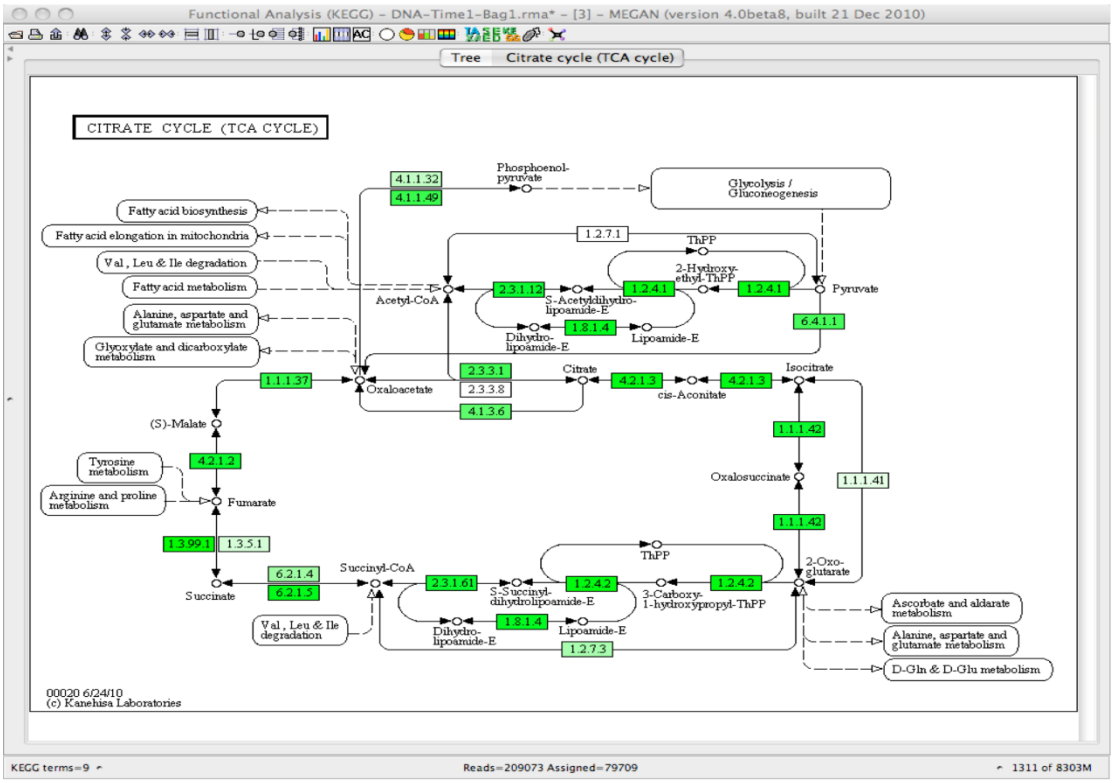
Το MEGAN παρέχει μια πληθώρα επιλογών που σχετίζονται με την παρουσίαση της ταξινομικής κατανομής, παραδείγματος χάριν μέσω διαγραμμάτων υπό την μορφή πίτας, είτε κλασσικών διαγραμμάτων με στήλες ή με γραμμές καθώς και πιο εξεζητημένες μορφές παρουσίασης. Παρόμοιες επιλογές παρουσίασης υπό την μορφή διαγραμμάτων προσφέρονται και κατά την επακόλουθη λειτουργική ανάλυση καθώς επίσης και κατά την περίπτωση που επιχειρείται συγκριτική ανάλυση μεταξύ δειγμάτων.

Σε ότι έχει να κάνει με το κομμάτι της λειτουργικής ανάλυσης μεταγονιδιωματικών δεδομένων, το MEGAN υποστηρίζει επί του παρόντος λειτουργική ανάλυση χρησιμοποιώντας τρία διαφορετικά λειτουργικά συστήματα ταξινόμησης και συγκεκριμένα τα SEED, KEGG, και COG (eggNog). Έτσι λοιπόν, το εργαλείο αυτό μπορεί να προσφέρει την δυνατότητα λειτουργικού σχολιασμού (functional annotation) μέσω τριών διαφορετικών προσεγγίσεων.

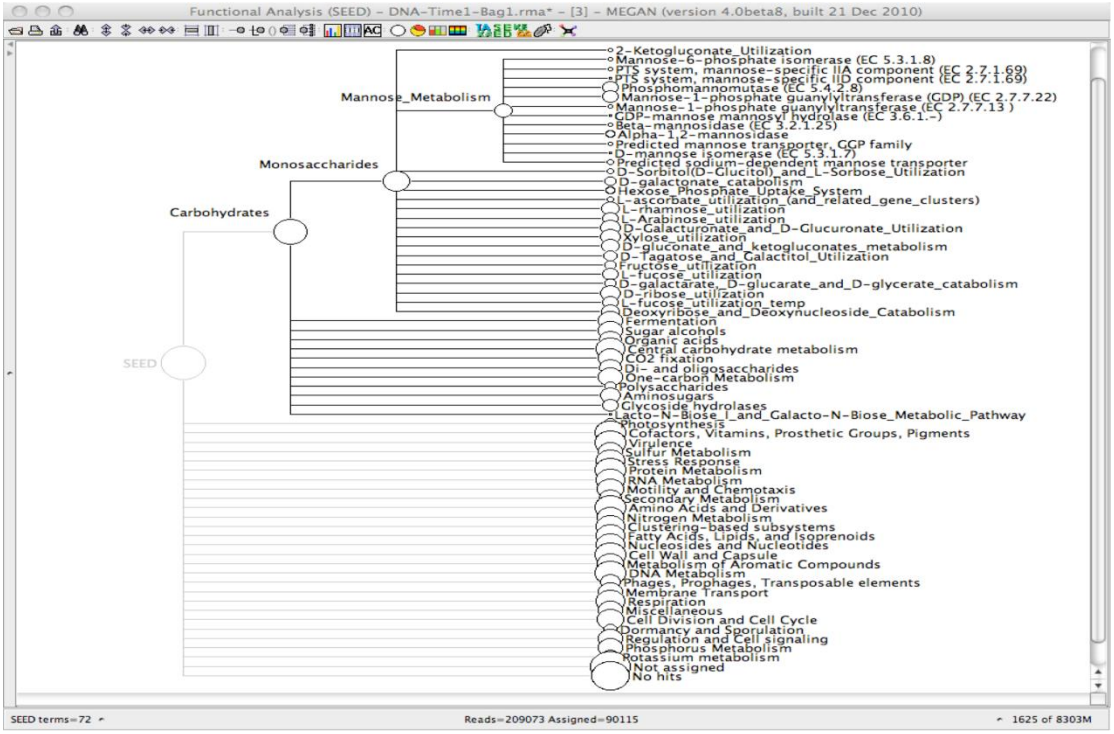
Η κατηγοριοποίηση SEED της γονιδιακής λειτουργίας αποτελείται από μια συλλογή από βιολογικά καθορισμένα υποσυστήματα. Σε αυτή την ταξινόμηση, γονίδια αντιστοιχίζονται σε λειτουργικούς ρόλους και οι διαφορετικοί λειτουργικοί ρόλοι ομαδοποιούνται σε υποσυστήματα. Η SEED κατηγοριοποίηση εκπροσωπείται από ένα δέντρο με μια αρχική ρίζα, όπου οι εσωτερικοί κόμβοι αντιπροσωπεύουν τα διάφορα υποσυστήματα και τα φύλλα (εξωτερικοί κύκλοι) αντιπροσωπεύουν τους λειτουργικούς ρόλους. Το δέντρο της κατηγοριοποίησης SEED έχει περίπου 10.000 κόμβους[40].Για να εκτελεστεί μια ανάλυση με βάση την κατηγοριοποίηση SEED, για κάθε αλληλουχία ανάγνωσης στα δεδομένα εισόδου, το MEGAN προσδιορίζει το χτύπημα με την υψηλότερη βαθμολογία σε μια ακολουθία αναφοράς, για το οποίο ο αντίστοιχος λειτουργικός ρόλος είναι γνωστός και στη συνέχεια αντιστοιχίζει την αλληλουχία ανάγνωσης σε αυτόν τον λειτουργικό ρόλο[37].

Η βάση δεδομένων KEGG (Kyoto Encyclopedia of Genes and Genomes) παρέχει μια συλλογή από μεταβολικά καθώς και άλλου είδους μονοπάτια. Η κατηγοριοποίηση KEGG μπορεί να παρουσιαστεί ως δέντρο, το οποίο αναφέρεται ως δέντρο KEGG. Για να πραγματοποιηθεί μια ανάλυση KEGG το MEGAN αντιστοιχίζει αλληλουχίες ανάγνωσης σε γονίδια, τα οποία με την σειρά τους χαρτογραφούνται σε KEGG ομάδες ορθολογίας οι οποίες εμφανίζονται σε διάφορα μεταβολικά ή μη μονοπάτια[37]. Λόγω αλλαγής στον τρόπο με τον οποίο λαμβάνονται οι άδειες

χρήσης της KEGG βάσης δεδομένων, το MEGAN χρησιμοποιεί παλαιότερη έκδοση αυτής[36].

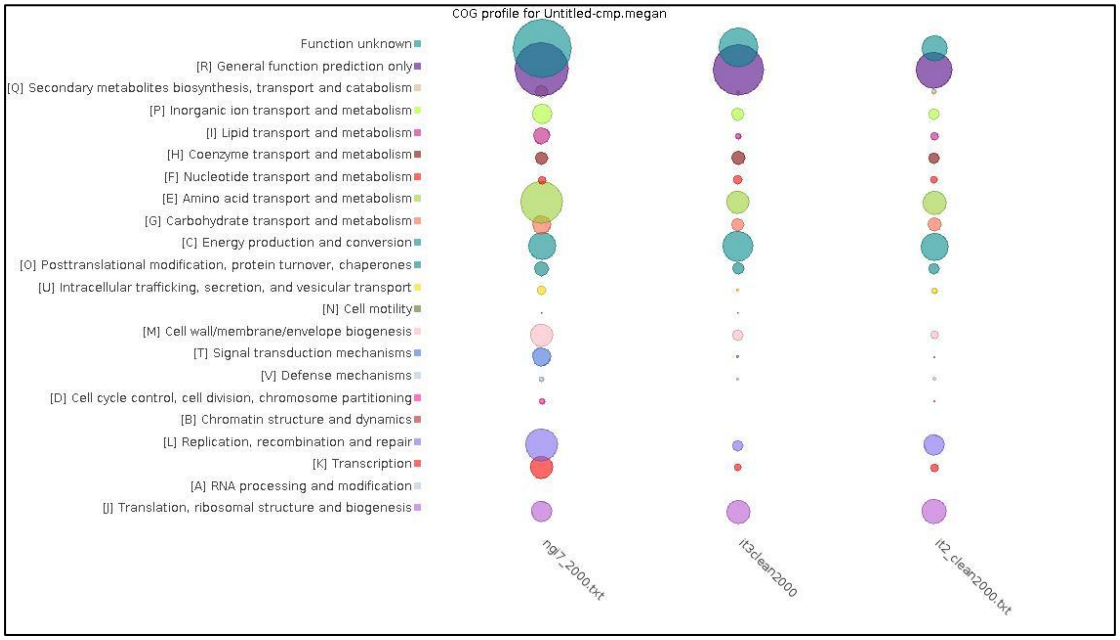


Εικόνα 19 Παράδειγμα μεταβολικού μονοπατιού που προέρχεται από ανάλυση KEGG μέσω του MEGAN.



Εικόνα 20 Λειτουργική ανάλυση μέσω MEGAN βασιζόμενη στην SEED κατηγοριοποίηση.

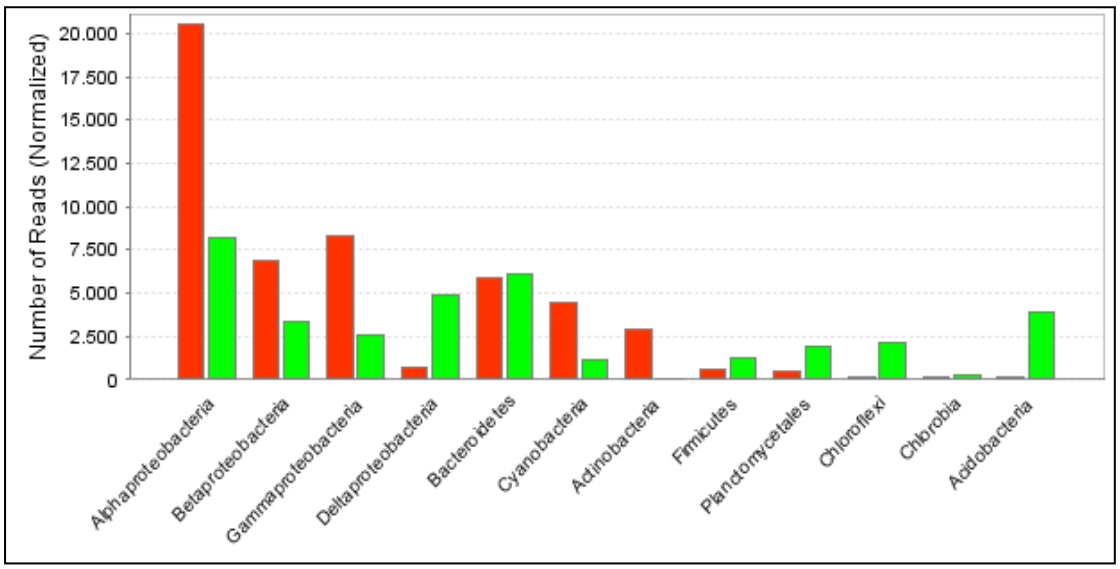
Η COG (Clusters of Orthologous Groups) κατηγοριοποίηση της γονιδιακής λειτουργίας, αποτελείται από μια συλλογή από βιολογικά καθορισμένα σύνολα ορθόλογων ομάδων. Η κατηγοριοποίηση κατά COG μπορεί να εμφανίζεται ως ένα δέντρο που περιέχει πολλούς κόμβους και ακμές. Τα γονίδια χαρτογραφούνται πάνω σε COGs και NOGs (Non-supervised Orthologous Groups , που αποτελούν ομάδες διευρυμένης ανάλυσης σε σχέση με την βάση των COGs και παρέχουν περισσότερες πληροφορίες αφού έχουν χαρακτηριστεί αναλυτικότερα). Το εργαλείο θα επιχειρήσει να χαρτογραφήσει και να αντιστοιχίσει κάθε αλληλουχία ανάγνωσης πάνω σε ένα γονίδιο που διαθέτει ένα γνωστό COG ή NOG[36].



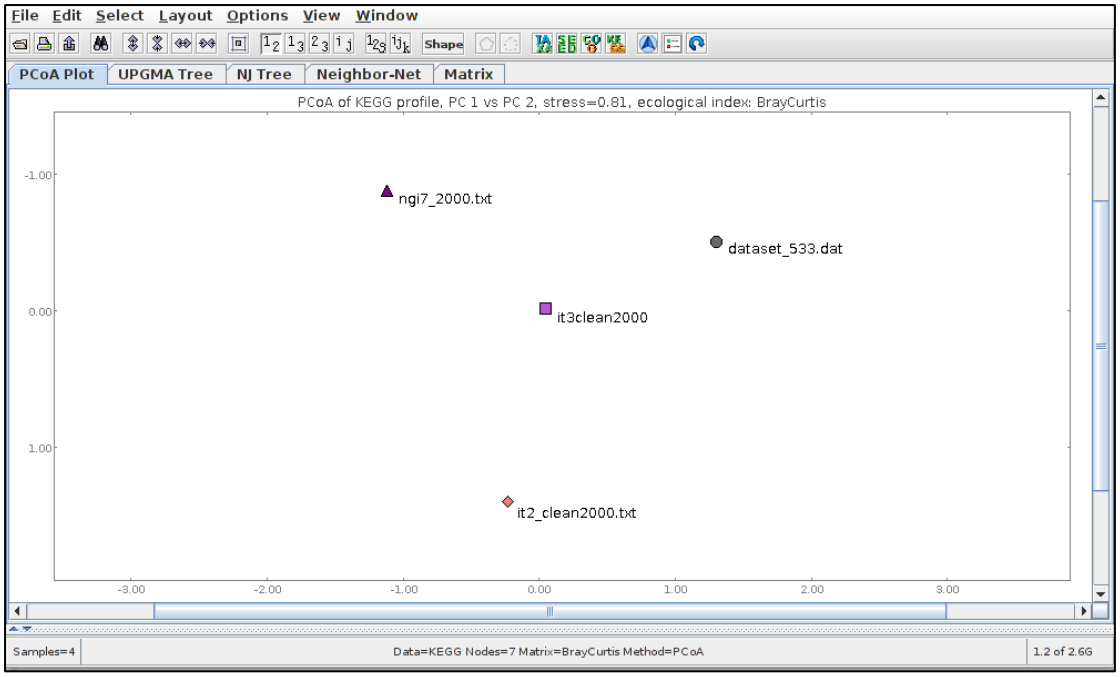
Εικόνα 21 Απεικόνιση των αποτελεσμάτων ανάλυσης με βάση την COG κατηγοριοποίηση. Η απεικόνιση στο παράδειγμα αυτό πραγματοποιείται υπό την μορφή διαγράμματος φυσαλίδας και το μέγεθος των κύκλων είναι ενδεικτικό του αριθμού των αλληλουχιών ανάγνωσης που έχουν αντιστοιχιστεί.

Το MEGAN χρησιμοποιεί ενσωματωμένες (built-in) RefSeq χαρτογραφήσεις και για τα τρία αυτά είδη κατηγοριοποίησης και λειτουργικής ανάλυσης. Ως εκ τούτου, εάν μια τέτοιου είδους ανάλυση είναι επιθυμητή, τότε η βάση δεδομένων που χρησιμοποιείται κατά τη σύγκριση BLAST πρέπει να περιέχει απαραιτήτως RefSeq ids. Σε αυτή την περίπτωση χρησιμοποιείται κυρίως η βάση δεδομένων NCBI-NR. Εναλλακτικά, οι χρήστες μπορούν να παρέχουν τα δικά τους αρχεία χαρτογράφησης σε έναν σημαντικό αριθμό διαφορετικών μορφών, γεγονός απαραίτητο κατά τη σύγκριση αλληλουχιών ανάγνωσης σε σχέση με μια βάση δεδομένων αναφοράς η οποία δεν περιέχει ταυτότητες RefSeq ή στην περίπτωση χρήσης μιας μεθόδου σύγκρισης η οποία δεν διατρέχει όλα τα στοιχεία της βάσης δεδομένων αναφοράς για την παραγωγή του αρχείου εξόδου (όπως στην περίπτωση του RAPSearch2).

Τα μεταγονιδιωματικά πρότζεκτ συνήθως περιλαμβάνουν πολλά δείγματα που λαμβάνονται από διαφορετικά περιβάλλοντα, προέρχονται από διαφορετικές πειραματικές ρυθμίσεις, χρονικές περιόδους, ή και θέσεις. Η σύγκριση των δειγμάτων αυτών είναι ένα δύσκολο έργο. Ανάλογα με το πρότζεκτ, ο στόχος της σύγκρισης μπορεί να ποικίλει από την ανίχνευση απλών αλλαγών στην ταξινομική σύνθεση έως και τον εντοπισμό πολύπλοκων λειτουργικών αλλαγών. Για να διευκολυνθεί η σύγκριση των δειγμάτων, το MEGAN επιτρέπει στους χρήστες να ανοίξουν πολλά δείγματα ταυτόχρονα, εμφανίζοντας κάθε δείγμα σε ένα ξεχωριστό παράθυρο. Ο χρήστης μπορεί στη συνέχεια να επιλέξει έναν αριθμό ανοιχτών δειγμάτων ώστε αυτά να συνδυαστούν σε ένα νέο ενιαίο αρχείο συγκριτικής ανάλυσης. Το νέο αρχείο σύγκρισης παρέχει όλες τις δυνατότητες ταξινομικής και λειτουργικής ανάλυσης, καθώς και οπτικής παρουσίασης, που παρέχονται και στα απλά αρχεία μεμονωμένων δειγμάτων, εκτός από την αποθήκευση, την μελέτη και την ομοπαράθεση των αλληλουχιών ανάγνωσης που έχουν αντιστοιχιστεί σε κάθε κόμβο. Τα διαφορετικά δείγματα είναι δυνατό να μελετηθούν και να συγκριθούν ως προς ομοιότητες και διαφορές τόσο σε σχέση με τα ταξινομικά προφίλ που τα χαρακτηρίζουν όσο και με βάση τις λειτουργικές διαφορές που παρουσιάζονται. Είναι ιδιαίτερα συχνό το φαινόμενο, δείγματα που είναι στενά συνδεδεμένα να παρουσιάζουν σημαντικές διαφορές ως προς το ταξινομικό περιεχόμενο ενώ ταυτόχρονα οι διαφορές στη λειτουργική συμπεριφορά να είναι σχετικά περιορισμένες. Το MEGAN παρέχει την δυνατότητα μελέτης της απόστασης που χωρίζει πολλαπλά δείγματα τόσο από ταξινομικής όσο και από λειτουργικής άποψης, χρησιμοποιώντας τεχνικές όπως η PCoA (principal coordinate analysis), καθώς και η ιεραρχική (δέντρο UPGMA, δέντρο NJ) και η μη-ιεραρχική ομαδοποίηση (μέθοδος Neighbor-net).



Εικόνα 22 Παράδειγμα ταξινομικής σύγκρισης δυο δειγμάτων εδάφους(πράσινο) και θαλάσσιου περιβάλλοντος(πορτοκαλί) σε συγκεκριμένο ταξινομικό επίπεδο, υπό μορφή γραφήματος με στήλες[41].



Εικόνα 23 Principal coordinate analysis (PCoA) για την μελέτη της λειτουργικής απόστασης πολλαπλών δειγμάτων με βάση την KEGG κατηγοριοποίηση.

ΥΠΟΛΟΓΙΣΤΙΚΟ ΜΕΡΟΣ

4. ΑΝΑΠΤΥΞΗ ΚΑΙ ΠΕΡΙΓΡΑΦΗ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΕΡΓΑΛΕΙΩΝ

4.1 Γενικές πληροφορίες

Ο κύριος στόχος κατά την πραγματοποίηση μιας υπολογιστικής ανάλυσης ενός μεταγονιδιωματικού συνόλου δεδομένων θα μπορούσε θεωρηθεί ότι είναι να δοθούν απαντήσεις σε δυο βασικά ερωτήματα[42]:

1. Ποιοί είναι εκεί έξω; Προσδιορισμός του ταξινομικού περιεχομένου ενός συνόλου δεδομένων μέσω της εκτίμησης των ταξινομικών ομάδων που είναι παρούσες και σε ποιες σχετικές αναλογίες εντοπίζονται. Ιδιαίτερα σημαντική η παρουσία ή η απουσία συγκεκριμένων ειδών ειδικής σημασίας.

2. Τι κάνουν; Καθορισμός του λειτουργικού περιεχομένου για ένα σύνολο δεδομένων μέσω του λειτουργικού χαρακτηρισμού, παραδείγματος χάριν με τον καθορισμό μεταβολικών μονοπατιών που ενδιαφέρουν και σχετίζονται με το δείγμα.

Η ανάπτυξη λοιπόν των υπολογιστικών εργαλείων που ακολουθούν, βασίστηκε στην ανάγκη απάντησης τέτοιου είδους ερωτημάτων στο βαθμό φυσικά που αυτό είναι δυνατό στα πλαίσια μιας διπλωματικής εργασίας. Για τον λόγο αυτό, τα εργαλεία που δημιουργήθηκαν καλύπτουν βασικές πτυχές μιας ταξινομικής και λειτουργικής ανάλυσης μεταγονιδιωματικών δειγμάτων.

Κύριος και πρωταρχικός στόχος ήταν η δημιουργία εργαλείων εύχρηστων και κατανοητών από χρήστες που δεν έχουν εξειδικευμένες γνώσεις πληροφορικής ή ακόμα πιο συγκεκριμένα βιοπληροφορικής, ενώ δεν είναι και εξοικειωμένοι με την χρήση βασικών εργαλείων της βιοπληροφορικής όπως θεωρείται το BLAST καθώς και χρήσιμων εργαλείων που προσφέρουν ποικίλες δυνατότητες ανάλυσης και μελέτης ενός μεταγονιδιωματικού δείγματος, όπως είναι το MEGAN. Αυτό επιχειρείται μέσω της χρήσης της πλατφόρμας του GALAXY για την δημιουργία απλών περιβαλλόντων διεπαφής, μέσα από τα οποία ο χρήστης μπορεί να προχωρήσει στην ανάλυση του δείγματος ή των δειγμάτων που επιθυμεί, ρυθμίζοντας μια σειρά από παραμέτρους που του περιγράφονται, χωρίς προαπαιτούμενη εμπειρία χρήσης αυτών των εργαλείων. Στην πραγματικότητα, μέσω των δυνατοτήτων που προσφέρει το GALAXY, ο ερευνητής, δεν έρχεται σε άμεση επαφή με τις εφαρμογές που απαιτούνται για την ανάλυση του δείγματος που μελετάται.

Επιπροσθέτως, σημαντικός παράγοντας στην επιλογή της πλατφόρμας του GALAXY ως βάση της δημιουργίας των εργαλείων αυτών, είναι η δυνατότητα αυτοματοποίησης της όλης διαδικασίας μέσω της επιλογής των ροών διεργασιών

(workflows). Μέσω της σύνδεσης των εργαλείων που δημιουργήθηκαν ο χρήστης μπορεί να επιλέξει το κατάλληλο workflow ώστε παρέχοντας τα δεδομένα εισόδου που επιθυμεί και ρυθμίζοντας εκ των προτέρων μια σειρά από παραμέτρους - εάν δεν καλύπτεται από τις προεπιλεγμένες τιμές τους - να θέτει σε λειτουργία μια σειρά από διεργασίες που εκτελούνται από τα εργαλεία, δίχως την ανάγκη ανθρώπινης παρέμβασης ενδιάμεσα των σταδίων της ανάλυσης.

Η ανάπτυξη των υπολογιστικών εργαλείων πραγματοποιήθηκε σε λειτουργικό σύστημα Linux Ubuntu και η γλώσσα προγραμματισμού που χρησιμοποιήθηκε για τον βασικό κώδικα ήταν η Python (Python Software Foundation. Python Language Reference, version 2.7.8. Available at http://www.python.org). Στην συνέχεια δημιουργήθηκαν τα κατάλληλα XML αρχεία ώστε να σχηματιστούν οι διεπαφές (interfaces) των εργαλείων στην πλατφόρμα του GALAXY, η οποία χρησιμοποιείται στην τοπική μορφή της και όχι στην διαδικτυακή της έκδοση (https://usegalaxy.org/). Τα εργαλεία που δημιουργήθηκαν εκμεταλλεύονται τις εφαρμογές και τις δυνατότητες δυο σημαντικών εργαλείων βιοπληροφορικής, του BLAST, ως εργαλείου σύγκρισης αλληλουχιών και κατά κύριο λόγο του MEGAN, με στόχο μια πρώτη, βασική και περιορισμένη, ταξινομική και λειτουργική ανάλυση μεταγονιδιωματικών δειγμάτων.

Το BLAST χρησιμοποιήθηκε στην τοπική έκδοση του, μέσω των stand-alone εφαρμογών του πακέτου BLAST+ και συγκεκριμένα της έκδοσης ncbi-blast-2.2.29+. Η online έκδοση του εργαλείου αυτού βρίσκεται στην διεύθυνση http://blast.ncbi.nlm.nih.gov/Blast.cgi. Είναι ίσως από τα πιο διαδεδομένα εργαλεία βιοπληροφορικής παγκοσμίως, για την αξιόπιστη σύγκριση αλληλουχιών.

Το MEGAN, χρησιμοποιήθηκε στην 5η έκδοση του και συγκεκριμένα στην έκδοση υπ' αριθμόν 5.5.3. Σημειώνεται για άλλη μια φορά πως η 5η έκδοση του MEGAN βελτιώνεται συνεχώς. Την περίοδο πραγματοποίησης της εν λόγω διπλωματικής η έκδοση που αναφέρεται ανωτέρω είναι η πιο πρόσφατη σταθερή. Για περισσότερες πληροφορίες http://ab.inf.uni-tuebingen.de/software/megan5/. Επιλέχθηκε ως βασικό εργαλείο μελέτης, καθώς προσφέρει αρκετές επιλογές αρχείων εισόδου που έχουν αναφερθεί στην θεωρητική παρουσίαση του, αλλά κυρίως επειδή προσφέρει την δυνατότητα ταξινομικής αλλά ταυτόχρονα και λειτουργικής ανάλυσης μεταγονιδιωματικών δειγμάτων.

4.1 RMA builder

Το πρώτο κατά σειρά εργαλείο που δημιουργήθηκε αποτελεί ουσιαστικά έναν συνδυασμό εφαρμογών. Είναι το εργαλείο **RMA builder (tool id = rma_builder)** για το οποίο δημιουργήθηκε αρχικά ο βασικός κώδικας σε γλώσσα Python που περιλαμβάνεται στο αρχείο **blastsub.py** και στην συνέχεια το αρχείο XML **rma_builder.xml** που αποτελεί το αρχείο δημιουργίας ενός περιβάλλοντος διεπαφής του εργαλείου στην πλατφόρμα GALAXY . Ο κώδικας αλλά και το αρχείο XML παρουσιάζονται στο παράρτημα αναλυτικά. Σε πρώτη φάση, πραγματοποιείται μια σύγκριση αλληλούχισης μέσω του BLAST σε σχέση με βάσεις δεδομένων νουκλεοτιδίων ή πρωτεϊνών. Ως δεδομένα εισόδου στο στάδιο αυτό, χρησιμοποιούνται αρχεία απλού κειμένου, τύπου FASTA που περιέχουν έναν αριθμό ζευγών από σειρές κειμένου. Σε κάθε ζεύγος από αυτά, η πρώτη σειρά αποτελεί το όνομα της κάθε αλληλουχίας ανάγνωσης, ενώ η δεύτερη που ακολουθεί αποτυπώνει την νουκλεοτιδική αλληλουχία αυτή καθεαυτή, δηλαδή μια σειρά από τέσσερις διαφορετικές αζωτούχες νουκλεοτιδικές βάσεις. Αδενίνη, θυμίνη, γουανίνη και κυτοσίνη. Σε ένα δεύτερο στάδιο, το αποτέλεσμα της BLAST σύγκρισης, χρησιμοποιείται ως είσοδος στο MEGAN, με σκοπό την δημιουργία του τελικού αρχείου εξόδου του εργαλείου, δηλαδή ενός αρχείου τύπου RMA, που αποτελεί την βασική μορφή παρουσίασης δεδομένων μέσω MEGAN και θα χρησιμεύσει ως βάση για τους επόμενους τύπους αναλύσεων.

```
>HOL3SHT01BU7DH
TAATTCGACATCTATAGTAATAGTTTCTCTGAGAGAAACTATTACTATAGATGTCGAATTA
>HOL3SHT01BBJXP
GTAGTTGTCGTAGGCGATGATATAATATAGGCGAAAACGAACAGAAGAAGGATTATTGCACC
>HOL3SHT01BXKLX
AATTGATGATTCACGTATTAAGCTTATAATGAACTAT
>HOL3SHT01E0STG
ATTCACGTATTAAGCTTATAATGAACTATAAGAGTTAGTAATAAGAGTAC
>HOL3SHT01EEU4B
ATATTGGAGAAGTTACAATAATGGTGAAATAATGTCACACATGTAGATTAACACC
>HOL3SHT01EK279
ATCTTCTCATCTCGTACAATTATAACTTAGATAGGCCTCG
>HOL3SHT01CB4PZ
ACATCATCATCCTGCTAATTAATATAATTTGTAGCTTTCGAATTATTCTTGGTGAATACTTCTGA
>HOL3SHT01BF7AI
GAATTATTAGACGAATGGTATTAAGACTAGTTCTATATCAGGAACGAACTTACGT
>HOL3SHT01DT13U
GTAAGCATTTTAGTCATCAACTATTACACACTGGAGTTATTATATGTTCTAGCTTCTTATTCCTCCTACTTACTC
>HOL3SHT01D41AN
AGTCTAGGCCTCAACACGAACTCAGTGTCATACGCATTCCACTGCCA
>HOL3SHT01AU4K2
TATCTCGTATGCGTGCATCGGTTATCGTTAAGCGC
>HOL3SHT01CW7SM
CTGCGTTGCTTGTGTATGACTGAGTAACGTATATTGAGN
>HOL3SHT01B31OJ
GCTATAGCAGAGGCGTCAACGACTAGCCGCATCTCTATCCTCTCTTATGAGGTCAACGCTGGTAG
>HOL3SHT01A3Z04
GACACAATTATGAATTATATTCCAGCAAGGCACACAGGA
>HOL3SHT01CQV3M
GACAATGCTAATGTCAATGCTTTGTCAGACGCCAT
```

Εικόνα 24 Μια τυπική μορφή ενός FASTA αρχείου που περιέχει νουκλεοτιδικές αλληλουχίες ανάγνωσης προερχόμενες από μεταγονιδιωματικά δείγματα.

Σημειώνεται ότι, κάθε εργαλείο που δημιουργείται πρέπει να δηλώνεται στην πλατφόρμα GALAXY. Η διαδικασία αυτή γίνεται μέσω επεξεργασίας του αρχείου **tool_conf.xml** του GALAXY, στο οποίο και γίνεται η δήλωση του μονοπατιού του κάθε εργαλείου (path - η τοποθεσία δηλαδή του αρχείου στον υπολογιστή) ώστε αυτό να αναγνωρίζεται από το GALAXY. Δημιουργήθηκε λοιπόν ένα ειδικό τμήμα στο αρχείο αυτό, το section **MyTools**, όπου δηλώνονται τα εργαλεία που κατασκευάζονται ώστε να εμφανίζονται ως επιλογές στην πλατφόρμα.

```
<section name="MyTools" id="mTools">
   <tool file="myTools/rma_builder.xml"/>
</section>
```

Αναλυτικότερα ως προς την περιγραφή του εργαλείου, ο χρήστης αρχικά έχει την δυνατότητα επιλογής μεταξύ της εφαρμογής BLASTX και της εφαρμογής BLASTN ώστε να πραγματοποιηθεί η σύγκριση αλληλουχίας που απαιτείται. Η εφαρμογή BLASTX, όπως έχει ήδη αναφερθεί, ουσιαστικά αποτελεί μια σύγκριση νουκλεοτιδικών αλληλουχιών σε σχέση με μια βάση δεδομένων πρωτεϊνών αφού πρώτα οι νουκλεοτιδικές αλληλουχίες μεταφραστούν στα έξι πιθανά πλαίσια ανάγνωσης, καθώς δεν είναι γνωστό το σωστό πλαίσιο ανάγνωσης. Αντίστοιχα, η εφαρμογή BLASTN είναι μια σύγκριση νουκλεοτιδικών αλληλουχιών σε σχέση με βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών. Εάν ο σκοπός της ανάλυσης του χρήστη είναι να λάβει μέσω της χρήσης των εργαλείων που ακολουθούν και σχετίζονται με το MEGAN, πληροφορίες τόσο σε ταξινομικό όσο και σε λειτουργικό επίπεδο, η επιλογή της προσέγγισης BLASTX αποτελεί μονόδρομο. Το MEGAN, παρέχει δυνατότητα λειτουργικού χαρακτηρισμού μέσω της SEED κατηγοριοποίησης, μέσω των KEGG μεταβολικών μονοπατιών καθώς και μέσω των COGs που αποτελούν σύνολα ορθόλογων ομάδων. Για κάθε μία από αυτές τις αναλύσεις απαιτείται η χαρτογράφηση και αντιστοίχιση κωδικών ταυτοτήτων που ονομάζονται RefSeq ids. Τα RefSeq ids εντοπίζονται σε βάσεις δεδομένων όπως η NCBI-NR, βάσεις λοιπόν δεδομένων πρωτεϊνικής φύσης και για τον λόγο αυτό , εάν επιθυμείται περαιτέρω λειτουργική ανάλυση εκτός της ταξινομικής, επιλέγεται η εφαρμογή BLASTX. Εάν, επιθυμείται μόνο ταξινομική κατηγοριοποίηση, ο χρήστης μπορεί να επιλέξει την εφαρμογή BLASTN, μια διαδικασία σαφώς λιγότερο απαιτητική και χρονοβόρα η οποία όμως πραγματοποιείται σε σχέση με βάσεις δεδομένων νουκλεοτιδίων όπως παραδείγματος χάριν η NCBI-NT, η οποίες δεν παρέχουν τα απαραίτητα RefSeq ids για την λειτουργική ανάλυση.

Εκτός από την εφαρμογή που θα επιλεγεί για την σύγκριση αλληλουχιών, ο ερευνητής έχει την δυνατότητα επιλογής και της βάσης δεδομένων σε σχέση με την οποία γίνεται αυτή η σύγκριση. Στην περίπτωση του BLASTX, υπάρχει η δυνατότητα επιλογής της βάσης δεδομένων NCBI-NR τόσο μέσω της online έκδοσης της στον

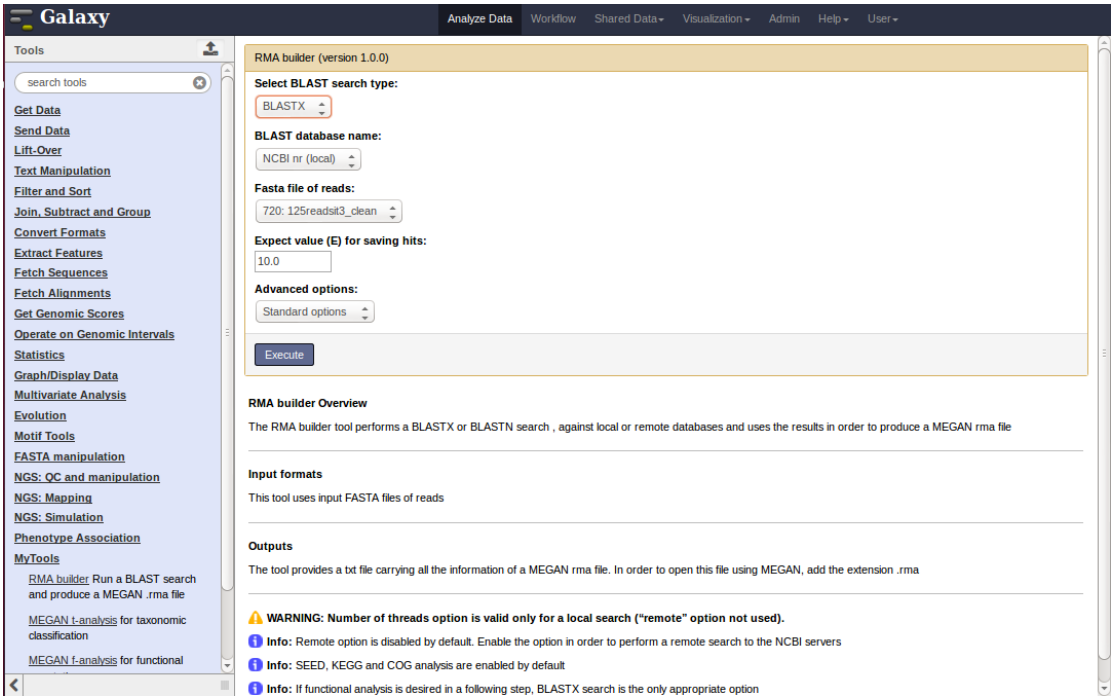
server του NCBI με την βοήθεια της επιλογής remote, όσο και σε τοπική μορφή μέσω σύγκρισης σε τοπικά διαθέσιμη έκδοση της βάσης δεδομένων στο υπολογιστικό σύστημα του χρήστη, δίχως την ανάγκη απομακρυσμένης σύνδεσης. Κατά τρόπο παρόμοιο, προσφέρονται και οι επιλογές στην περίπτωση του BLASTN, με την διαφορά ότι εδώ γίνεται αναφορά στην NCBI-NT και τις δυο εκδοχές χρήσης της - την απομακρυσμένη και την τοπική.

Στην συνέχεια ο χρήστης, επιλέγει το αρχείο εισόδου FASTA το οποίο αποτελεί την βάση της διαδικασίας σύγκρισης αλληλουχιών μέσω του BLAST. Πρέπει λοιπόν να έχει προηγηθεί εισαγωγή του επιθυμητού αρχείου στην πλατφόρμα GALAXY με ευθύνη του εκάστοτε χρήστη, ώστε να προσφέρεται η δυνατότητα επιλογής και επεξεργασίας του από το υπολογιστικό εργαλείο. Το αρχείο FASTA επιπροσθέτως και πέραν της χρήσης του ως αρχείο εισόδου για την διεργασία της BLAST σύγκρισης, λειτουργεί και ως μια επιπλέον βοήθεια για το MEGAN ώστε να αποτυπωθούν όσο το δυνατόν ακριβέστερα τα αποτελέσματα της σύγκρισης μέσω του BLAST, στο ταξινομικό διάγραμμα αλλά και την λειτουργική ανάλυση του MEGAN. Στην συνέχεια, είναι απαραίτητο να καθοριστεί ένας αριθμός παραμέτρων που αφορούν τόσο το BLAST όσο και το MEGAN ώστε τα αποτελέσματα να καλύπτουν της ανάγκες και τους στόχους του εκάστοτε ερευνητή στην κάθε ανάλυση.

Η πρώτη παράμετρος που ζητείται να ρυθμιστεί είναι η Expect Value (E) δηλαδή μια παράμετρος που περιγράφει τον αριθμό των hits που μπορεί κανείς να περιμένει να δει κατά τύχη κατά την αναζήτηση σε μια βάση δεδομένων ενός συγκεκριμένου μεγέθους. Η παράμετρος αυτή μειώνεται εκθετικά με την αύξηση του σκορ (S) που χαρακτηρίζει κάθε match. Ουσιαστικά, η τιμή Ε περιγράφει τον τυχαίο θόρυβο. Για παράδειγμα, μια τιμή Ε της τάξης του 1 που ανατίθεται σε ένα χτύπημα (hit) μπορεί να ερμηνευθεί υπό την προσέγγιση, ότι σε μια βάση δεδομένων ενός δεδομένου μεγέθους, μπορεί κανείς να περιμένει να δει 1 match με παρόμοιο score απλώς κατά τύχη. Όσο χαμηλότερη είναι η τιμή Ε, ή όσο πιο κοντά είναι στο μηδέν, τόσο πιο "συγκεκριμένο" είναι το κάθε match, η κάθε αντιστοιχία δηλαδή, το κάθε ταίριασμα. Ωστόσο, πρέπει κανείς να έχει κατά νου ότι σχεδόν πανομοιότυπες ομοπαραθέσεις (alignments) μικρού μεγέθους έχουν σχετικά υψηλές τιμές Ε. Αυτό συμβαίνει επειδή ο υπολογισμός της τιμής Ε λαμβάνει υπόψη και το μήκος της αλληλουχίας-ερωτήματος. Αυτές οι υψηλές τιμές Ε έχουν νόημα, διότι οι μικρότερες αλληλουχίες παρουσιάζουν υψηλότερη πιθανότητα να εμφανιστούν στη βάση δεδομένων καθαρά από τύχη. Τέλος, η τιμή της παραμέτρου Ε μπορεί να χρησιμοποιηθεί ως ένας απλός και βολικός τρόπος για την δημιουργία ορίων (κατωφλιών) σημαντικότητας για τα αποτελέσματα. Όταν παραδείγματος χάριν, η τιμή της παραμέτρου αυξηθεί πάνω από την προεπιλεγμένη (default) τιμή της που ισούται με 10.0, μπορεί να δημιουργηθεί μια μεγαλύτερη λίστα αποτελεσμάτων

που θα περιέχει περισσότερα χτυπήματα, τα οποία όμως χαρακτηρίζονται και από χαμηλά σκορ.

Εκτός από την παράμετρο E, που αποτελεί παράμετρο ρύθμισης του BLAST παρέχεται η δυνατότητα ρύθμισης ενός αριθμού πρόσθετων παραμέτρων για τις οποίες έχει γίνει προεπιλογή συγκεκριμένων τιμών με βάση τις επίσημες προεπιλεγμένες τιμές των εργαλείων BLAST και MEGAN. Στην περίπτωση που ο χρήστης θέλει να τρέξει την ανάλυση του με βάση της προεπιλεγμένες τιμές, η επιλογή των Standard Options του δίνει άμεσα αυτή την δυνατότητα χωρίς να είναι ανάγκη να ρυθμίσει ο ίδιος τις παραμέτρους αυτές. Κάπου εδώ, είναι σημαντικό να τονιστεί πως το MEGAN αναγνωρίζει και επεξεργάζεται συγκεκριμένους τύπους αρχείων εξόδου που προέρχονται από μια BLAST έξοδο, συνεπώς, εκ των πραγμάτων δεν ήταν δυνατό να παρέχονται όλες οι δυνατές επιλογές παραμέτρων που παρέχονται από τους δημιουργούς του BLAST, ενώ κάποιες άλλες δεν παρουσίαζαν ιδιαίτερο ενδιαφέρον για την μελέτη των μεταγονιδιωματικών δειγμάτων και παραλήφθηκαν από το τελικό αποτέλεσμα. Σε ότι έχει να κάνει με τις παραμέτρους του MEGAN, αυτές στην πρώτη αυτή φάση της ανάλυσης, στην δημιουργία δηλαδή του αρχείου RMA, είναι περιορισμένες καθώς το κύριο ενδιαφέρον εντοπίζεται απλώς στην δημιουργία του RMA αρχείου. Η προσαρμογή των παραμέτρων θα γίνει στα επόμενα εργαλεία.



Εικόνα 25 Το εργαλείο RMA builder, στο επίπεδο των Standard Options. Επιλογή BLASTX.

Μια από τις παραμέτρους λοιπόν του BLAST που χαρακτηρίζουν την σύγκριση αλληλουχιών και είναι δυνατό να χρησιμοποιήσει ο χρήστης είτε στις default τιμές μέσω των Standard Options είτε να τις ρυθμίσει ο ίδιος μέσω της επιλογής

Advanced Options, είναι η παράμετρος Word_size. Όπως είναι γνωστό από την θεωρητική περιγραφή, το BLAST ξεκινάει την διαδικασία της ομοπαράθεσης και της σύγκρισης από ορισμένα όμοια σημεία των αλληλουχιών, που ονομάζονται words και με βάση αυτά συνεχίζει την διαδικασία ψάχνοντας για αλληλουχίες που μπορεί να οδηγήσουν σε πλήρεις ομοπαραθέσεις. Στην περίπτωση νουκλεοτιδικών συγκρίσεων όπως το BLASTN απαιτείται να βρεθεί ακριβής αντιστοιχία (match) με τις words ώστε να συνεχιστεί η διαδικασία της ομοπαράθεσης, συνεπώς το Word_size αντιπροσωπεύει το μήκος της ακριβούς αρχικής αντιστοιχίας. Η ρύθμιση του λοιπόν, καθορίζει και σε μεγάλο βαθμό την ευαισθησία και τον απαιτούμενο χρόνο της ανάλυσης. Στην περίπτωση του BLASTN η προεπιλεγμένη τιμή της παραμέτρου ισούται με 11 (ζεύγη βάσεων, bp) και ρυθμίζεται αυτόματα μόλις ο χρήστης επιλέξει αυτού του τύπου την σύγκριση. Για άλλους τύπους BLAST, όπως και το BLASTX, λαμβάνονται υπ' όψιν μη ακριβείς αντιστοιχίες με βάση την ομοιότητα μεταξύ των words και ουσιαστικά το Word_size ρυθμίζει το ποσοστό αυτής της ομοιότητας που απαιτείται. Στην περίπτωση του BLASTX η προεπιλεγμένη τιμή της παραμέτρου ισούται με 3 (bp) και ρυθμίζεται επίσης αυτόματα μόλις ο χρήστης επιλέξει αυτού του τύπου την σύγκριση.

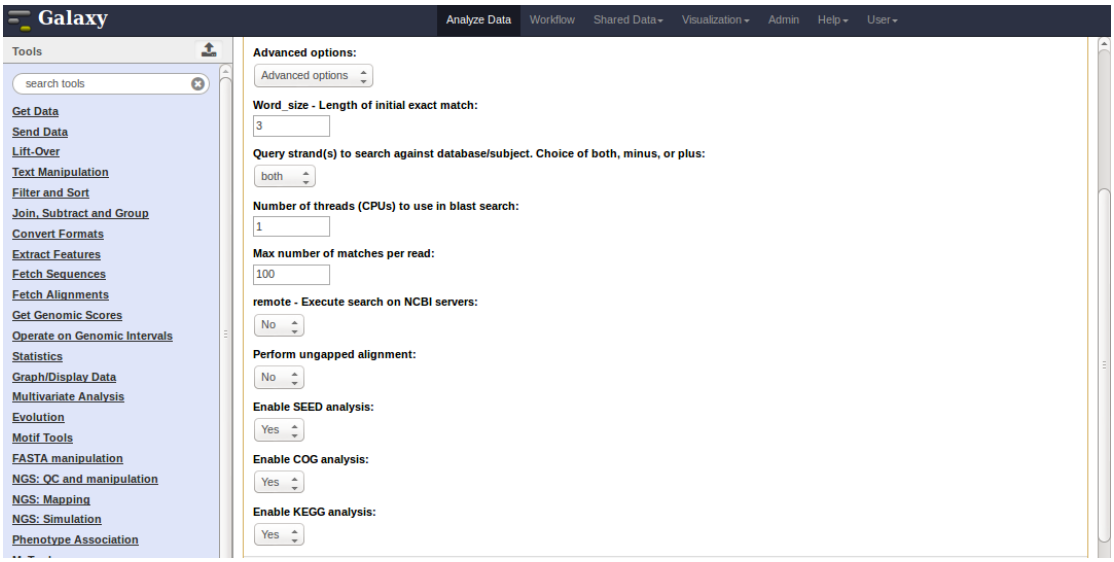
Στην συνέχεια ρυθμίζεται η παράμετρος Strand του BLAST. Οι πιθανές τιμές της είναι Both, Plus ή Minus, με προεπιλεγμένη την πρώτη. Αναφέρεται φυσικά στους δύο κλώνους του DNA και στο εάν η ομοπαράθεση θα γίνει ως προς και τους δύο, ως προς τον θετικό μόνο - δηλαδή τον κλώνο από τον οποίο μπορεί να προέλθει αλληλουχία RNA που εν τέλει θα οδηγήσει στην παραγωγή πρωτεΐνης - ή μόνο ως προς τον συμπληρωματικό αυτού, τον αρνητικό. Επιπροσθέτως, προσφέρεται η επιλογή ή όχι της παραμέτρου, ungapped, για την πραγματοποίηση ungapped ομοπαράθεσης. Ως προεπιλογή έχει ρυθμιστεί η απενεργοποίηση αυτής της παραμέτρου καθώς φυσιολογικά το BLAST εκτελεί gapped ομοπαραθέσεις, όπου συμπεριλαμβάνονται στα αποτελέσματα, πιθανές ελλείψεις και προσθήκες βάσεων στις ομοπαραθέσεις.

Τέλος, σε ότι αφορά το BLAST, εμφανίζονται και δύο ακόμα παράμετροι, που σχετίζονται άμεσα μεταξύ τους. Η πρώτη είναι η επιλογή remote, δηλαδή η επιλογή της πραγματοποίησης της σύγκρισης σε σχέση με τις βάσεις δεδομένων στον server του NCBI. Η παράμετρος είναι απενεργοποιημένη στις standard default επιλογές και ο χρήστης χρειάζεται να διαθέτει τοπικά τις βάσεις δεδομένων που απαιτούνται για την σύγκριση. Η δεύτερη παράμετρος είναι η παράμετρος επιλογής του αριθμού των CPU's που θα χρησιμοποιηθούν για την ανάλυση (number of threads) και αρχικά είναι προεπιλεγμένη η τιμή 1. Οποιαδήποτε αλλαγή στην τιμή αυτή είναι ασύμβατη στην περίπτωση που είναι ενεργοποιημένη η παράμετρος remote.

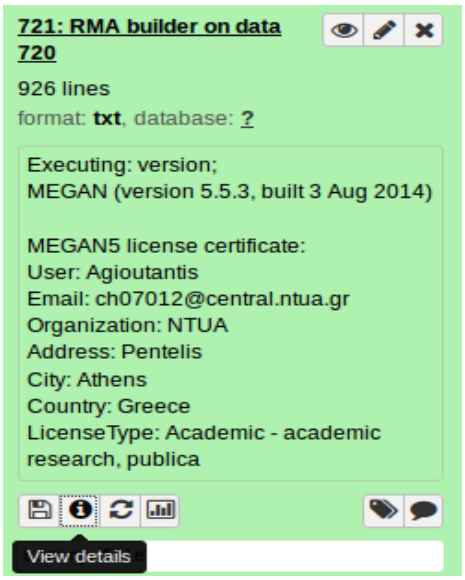
Τέλος, σε ότι αφορά τις παραμέτρους του MEGAN, είναι αρχικά περιορισμένες σε αυτό το εργαλείο. Ζητείται η ενεργοποίηση (έχει προεπιλεγεί) ή όχι των επιλογών

SEED, COG και KEGG για την λειτουργική ανάλυση (διαδικασία απαραίτητη για τα περαιτέρω εργαλεία εάν επιζητείται ο λειτουργικός χαρακτηρισμός) και ο καθορισμός του αριθμού των αντιστοιχιών που θα κρατηθούν για κάθε αλληλουχία ανάγνωσης. Η προεπιλεγμένη τιμή από το MEGAN, ισούται με 100.

Όλες οι παραπάνω παράμετροι είναι διαθέσιμες στην περίπτωση του BLASTX ενώ στην περίπτωση επιλογής του BLASTN, παραλείπονται οι επιλογές που σχετίζονται με την ενεργοποίηση των SEED, KEGG και COG καθώς εκ των πραγμάτων, μια τέτοια σύγκριση δεν μπορεί να δώσει αποτελέσματα στην λειτουργική ανάλυση.



Εικόνα 26 **Οι Advanced Options που παρέχονται στο εργαλείο RMA builder, στην περίπτωση του BLASTX.**



Εικόνα 27 **Αρχείο εξόδου του εργαλείου RMA builder. Δυνατότητα άμεσης εμφάνισης αποτελεσμάτων, αποθήκευσης τους, επανάληψης της διεργασίας ή μελέτης των λεπτομερειών της.**

Ως αποτέλεσμα εξόδου του εργαλείου, δημιουργείται ένα αρχείο txt που περιέχει όλες τις πληροφορίες που θα περιείχε ένα αρχείο RMA του MEGAN. Ο λόγος που γίνεται αυτό είναι πως το GALAXY δεν αναγνωρίζει ως τύπους αρχείων τα αρχεία RMA. Το γεγονός όμως αυτό δεν αποτελεί τροχοπέδη για τις περαιτέρω αναλύσεις, καθώς μέσω μιας απλής μετονομασίας του αρχείου με προσθήκη κατάληξης .rma που γίνεται αυτόματα από τα εργαλεία που ακολουθούν, το MEGAN το αναγνωρίζει δίχως πρόβλημα. Το αρχείο αυτό θα χρησιμοποιηθεί ως αρχείο εισόδου για τα υπόλοιπα εργαλεία, ενώ επίσης μπορεί να χρησιμοποιηθεί και ως κύριο RMA αρχείο τοπικά στο σύστημα του χρήστη, μέσω άμεσης χρήσης του MEGAN, εάν αυτό επιθυμείται.

Input Parameter	Value	Note for rerun
Select BLAST search type	blastx	
BLAST database name	NCBI nr	
Fasta file of reads	720: 125readsit3_clean	
Expect value (E) for saving hits	0.0001	
Advanced options	advanced	
Word_size - Length of initial exact match	3	
Query strand(s) to search against database/subject. Choice of both, minus, or plus	both	
Number of threads (CPUs) to use in blast search	1	
Max number of matches per read	100	
remote - Execute search on NCBI servers	Yes	
Perform ungapped alignment	No	
Enable SEED analysis	Yes	
Enable COG analysis	Yes	
Enable KEGG analysis	Yes	

Εικόνα 28 Οι πληροφορίες των παραμέτρων εισόδου με βάση τις οποίες προέκυψε το αρχείο εξόδου. Η αποθήκευση και η δυνατότητα ελέγχου αυτών των πληροφοριών, αποτελεί σημαντικό χαρακτηριστικό του GALAXY, καθώς δεν αφήνει ερωτηματικά ως προς τις λεπτομέρειες μιας ανάλυσης.

4.2 MEGAN t - analysis

Το δεύτερο κατά σειρά εργαλείο που δημιουργήθηκε, αποτελεί ένα εργαλείο που αποσκοπεί σε μια πρώτη προσέγγιση στην προσπάθεια ταξινομικής ανάλυσης ενός μεταγονιδιωματικού δείγματος. Πρόκειται για το εργαλείο **MEGAN t-analysis (tool id = megan_analysis)**, του οποίου ο κώδικας βρίσκεται στο αρχείο **megansub.py** ενώ το αντίστοιχο XML αρχείο που χρησιμοποιείται για την δημιουργία διεπαφής στο GALAXY, ονομάζεται **megan_analysis.xml**. Ο κώδικας και το XML αρχείο παρουσιάζονται στο παράρτημα. Το δεύτερο εργαλείο με την σειρά του, δηλώνεται στο GALAXY

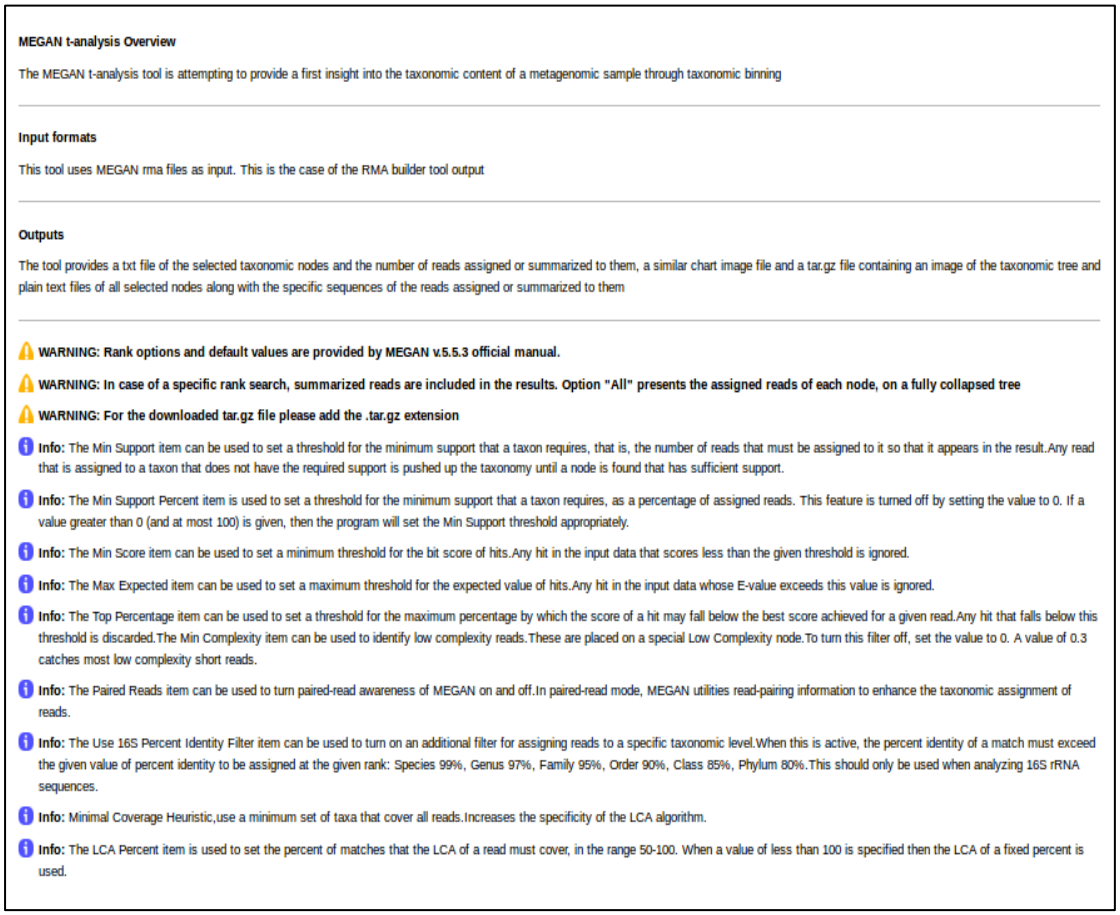
```
<section name="MyTools" id="mTools">
    <tool file="myTools/rma_builder.xml"/>
    <tool file="myTools/megan_analysis.xml"/>
</section>
```

Η διασύνδεση του πρώτου εργαλείου με αυτό το εργαλείο ταξινομικής ανάλυσης εντοπίζεται στο γεγονός ότι το rma αρχείο εξόδου του πρώτου χρησιμοποιείται ως είσοδος στο δεύτερο με σκοπό την παραγωγή αρχείων κειμένου, διαγραμμάτων αλλά και εικόνων που σχετίζονται με το ταξινομικό προφίλ του δείγματος που μελετάται. Έτσι λοιπόν, ως πρώτο βήμα, ο χρήστης καλείται να διαλέξει το αρχείο RMA, πάνω στο οποίο θα βασιστεί η ταξινομική κατηγοριοποίηση.

Στην συνέχεια, ουσιαστικά ακολουθεί μια διαδικασία ρύθμισης παραμέτρων του MEGAN, ανάλογα με τους στόχους της ανάλυσης. Όπως, ήδη αναφέρθηκε στο πρώτο εργαλείο, ο αριθμός των παραμέτρων του MEGAN που ζητήθηκε να καθοριστούν ήταν περιορισμένος. Πλέον, στο δεύτερο αυτό εργαλείο μέσω μιας διαδικασίας που στο MEGAN ονομάζεται επανυπολογισμός των παραμέτρων του LCA αλγορίθμου (recompute LCA parameters), ο χρήστης ρυθμίζει τις παραμέτρους του βασικού αλγορίθμου του προγράμματος ώστε να ελέγχει πλήρως την πορεία της ανάλυσης. Η επιλογή αυτής της διαδικασίας στην αρχή του δεύτερου προγράμματος αποσκοπεί στο να δοθεί στους χρήστες η δυνατότητα να επαναλαμβάνουν γρήγορα και απλά αναλύσεις στο ίδιο δείγμα, πειραματιζόμενοι με τις τιμές των παραμέτρων. Η διαδικασία της ταξινομικής ανάλυσης (αλλά και της λειτουργικής) μέσω του MEGAN είναι μια διαδικασία που δεν μπορεί να συγκριθεί από πλευράς απαιτούμενου χρόνου με αυτή της σύγκρισης μέσω BLAST. Συνεπώς, στο πρώτο εργαλείο δημιουργείται το αρχείο RMA, με βάση τις προεπιλεγμένες τιμές των παραμέτρων του MEGAN και έπειτα αυτές ρυθμίζονται κατά βούληση του χρήστη κάθε φορά που είναι να τρέξει το δεύτερο αυτό εργαλείο για την πραγματοποίηση της ταξινομικής ανάλυσης. Επίσης, για καθαρά τεχνικούς λόγους που σχετίζονται με την κατασκευή του MEGAN, ακόμα και αν ρυθμιζόταν εξ αρχής η κατάσταση των παραμέτρων κατά την κατασκευή του RMA, θα έπρεπε να ρυθμιστούν ξανά και στο δεύτερο εργαλείο, οπότε αποφεύγονται έτσι περιττές ενέργειες. Φυσικά και σε αυτή την περίπτωση υπάρχει η επιλογή των Standard Options, όπου διατηρούνται οι προεπιλεγμένες τιμές του LCA αλγορίθμου. Εναλλακτικά στις Advanced Options ο χρήστες μπορεί να ορίσει τις τιμές που εκείνος επιθυμεί, δίνοντας όμως προσοχή στις ακραίες τιμές που είναι δυνατόν να λάβουν αυτές.

Διαλέγοντας τις προχωρημένες επιλογές, ο χρήστης καλείται να ορίσει για το πρόγραμμα τις παραμέτρους του βασικού αλγορίθμου. Στο κάτω μέρος του εργαλείου στην μορφή που έχει στην πλατφόρμα GALAXY, παρέχονται πληροφορίες σχετικά με τις παραμέτρους αυτές καθώς και για τις οριακές τιμές τους. Ο χρήστης ορίζει την παράμετρο Minscore, που καθορίζει το κατώτερο όριο στα σκορ των χτυπημάτων που γίνονται αποδεκτά, με προεπιλεγμένη τιμή ίση με 5.0 . Κάθε hit στο αρχείο εισόδου (δηλαδή στο αρχείο BLAST καθώς το recompute είναι μια επαναδημιουργία του αρχείου rma) με σκορ κάτω από το όριο αυτό, απορρίπτεται. Η παράμετρος Max Expected ορίζει το ανώτερο όριο της expected value των hits στο αρχείο εισόδου που γίνονται αποδεκτά και το 0.01 έχει οριστεί ως default τιμή. Η παράμετρος Top Percent ορίζει το μέγιστο ποσοστό από το οποίο το σκορ ενός hit μπορεί να πέσει κάτω σε σχέση με το καλύτερο σκορ που έχει επιτευχθεί για μια αλληλουχία ανάγνωσης, ώστε ένα χτύπημα να είναι αποδεκτό και έχει προεπιλεγμένη τιμή ίση με 10.0 . Η παράμετρος Minsupport, ορίζει τον ελάχιστο αριθμό αλληλουχιών ανάγνωσης που απαιτείται να αντιστοιχιστούν σε μια

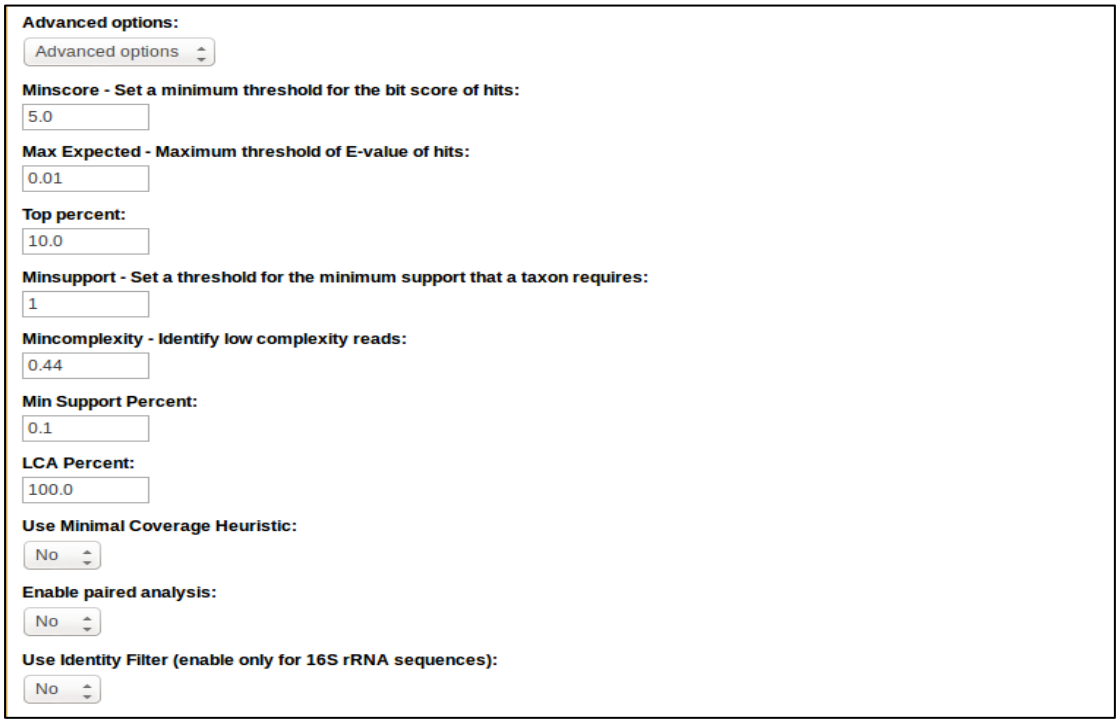
ταξινομική ομάδα ώστε αυτή να εμφανιστεί στα αποτελέσματα, έχει default τιμή ίση με 1, ενώ κάθε αλληλουχία που αντιστοιχίζεται σε μια ομάδα που δεν έχει τον απαραίτητο αριθμό support, προωθείται υψηλότερα στην ταξινόμηση μέχρι να βρει έναν κόμβο με τον απαραίτητο αριθμό support. Η παράμετρος Min Support Percent, ορίζει τον ελάχιστο αριθμό support που απαιτεί μια ταξινομική ομάδα, ως ποσοστό όμως των αντιστοιχισμένων reads. Εχει προεπιλεγμένη τιμή ίση με 0.1 ενώ λαμβάνει τιμές από 0 έως 100 που καθορίζουν καταλλήλως το κατώφλι του Min Support, με την τιμή μηδέν να την απενεργοποιεί.



Εικόνα 29 Τα βοηθητικά μηνύματα του εργαλείου MEGAN t-analysis. Περιγραφή του εργαλείου, πληροφορίες σχετικά με τις παραμέτρους και τις τιμές τους, καθώς και οδηγίες ως προς την επιλογή των στόχων της ταξινομικής ανάλυσης.

Η επιλογή Min Complexity μπορεί να ενεργοποιηθεί στην περίπτωση που επιθυμείται η ταυτοποίηση αλληλουχιών χαμηλής πολυπλοκότητας, οι οποίες και τοποθετούνται σε έναν ειδικό ξεχωριστό κόμβο στο ταξινομικό δέντρο. Έχει προεπιλεγμένη τιμή ίση με 0.44 και απενεργοποιείται θέτοντας την τιμή μηδέν. Μια τιμή ίση με 0.3 πιάνει τις περισσότερες μικρού μήκους αλληλουχίες χαμηλής πολυπλοκότητας. Η παράμετρος LCA Percent χρησιμοποιείται για να ρυθμίσει το ποσοστό των αντιστοιχιών που πρέπει να καλύπτει ο LCA μιας αλληλουχίας ανάγνωσης και κυμαίνεται σε τιμές μεταξύ 50.0 - 100.0, με προεπιλεγμένη την τιμή 100.0 . Μέσω της ενεργοποίησης της Paired Reads παραμέτρου, ειδοποιείται το

MEGAN πως τα δεδομένα προέρχονται από paired-read αναλύσεις, ώστε να βελτιστοποιήσει την διαδικασία αντιστοίχισης ενώ μέσω της ενεργοποίησης της παραμέτρου Minimal Read Heuristic, γίνεται χρήση ενός ελάχιστου αριθμού ταξινομικών ομάδων που καλύπτουν όλες τις αλληλουχίες ανάγνωσης, με αποτέλεσμα την αύξηση της ειδικότητας του LCA αλγορίθμου. Τέλος, το 16 S Percent Identity filter είναι ένα επιπλέον φίλτρο για την αντιστοίχιση αλληλουχιών ανάγνωσης σε ταξινομικές ομάδες και πρέπει να ενεργοποιείται μόνο σε περίπτωση ανάλυσης αλληλουχιών 16S rRNA. Οι τρεις τελευταίες παράμετροι που μόλις αναφέρθηκαν είναι απενεργοποιημένες στην προεπιλεγμένη μορφή τους.



Εικόνα 30 Advanced Options του εργαλείου MEGAN t-analysis.

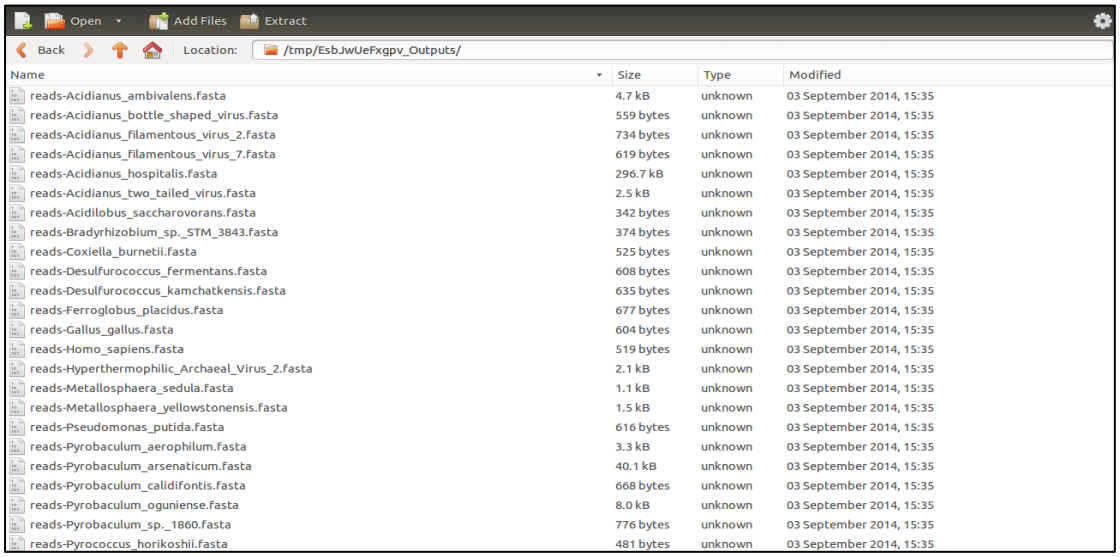
Πέρα από την διαδικασία ρύθμισης των παραμέτρων του βασικού αλγορίθμου, η λειτουργία του εργαλείου αυτού, βασίζεται ουσιαστικά στην επιλογή του χρήστη ως προς τον στόχο της ανάλυσης. Του δίνεται η δυνατότητα επιλογής του ταξινομικού επιπέδου που επιθυμεί να μελετήσει. Μπορεί είτε να επιλέξει να μελετήσει όλες μαζί τις ταξινομικές ομάδες που εντοπίζονται στο ταξινομικό δέντρο, είτε να εστιάσει σε ένα συγκεκριμένο ταξινομικό επίπεδο.

Στην πρώτη περίπτωση, διαλέγοντας την επιλογή All στο κατάλληλο σημείο του εργαλείου, πραγματοποιείται ένα πλήρες ξεδίπλωμα του ταξινομικού δέντρου και η περαιτέρω ανάλυση αφορά όλες τις ταξινομικές ομάδες που εντοπίζονται σε αυτό συμπεριλαμβανομένων των κόμβων της χαμηλής πολυπλοκότητας, των μη αντιστοιχισμένων αλληλουχιών, των αλληλουχιών που δεν παρουσίασαν καθόλου hits αλλά και της αρχικής ρίζας του συνόλου των αλληλουχιών. Σε κάθε κόμβο εμφανίζονται μόνο τα reads που έχουν αντιστοιχιστεί και ανήκουν επακριβώς σε

αυτόν. Τα αρχεία λοιπόν εξόδου του εργαλείου σε αυτή την περίπτωση είναι ένα διάγραμμα όπου στον οριζόντιο άξονα βρίσκονται τα ονόματα όλων των κόμβων και στον κάθετο ο αριθμός των αλληλουχιών ανάγνωσης που έχουν αντιστοιχιστεί σε κάθε κόμβο και μπορεί να χρησιμοποιηθεί ως μια πρώτη εκτίμηση του ταξινομικού προφίλ και της αφθονίας των ειδών, καθώς και ένα αρχείο διαμορφωμένο σε μορφή με στήλες όπου παρουσιάζονται τα ίδια δεδομένα με το προηγούμενο διάγραμμα απλά σε μορφή κειμένου. Επιπροσθέτως, δημιουργείται και ένας συμπιεσμένος φάκελος tar.gz ο οποίος περιέχει ξεχωριστά για τον κάθε κόμβο αρχεία FASTA που καταγράφουν ποίες επακριβώς αλληλουχίες ανάγνωσης αντιστοιχούν σε κάθε κόμβο (όνομα αλληλουχίας και αλληλουχία), ενώ περιέχει και μια εικόνα του ταξινομικού δέντρου σε πλήρη ξεδιπλωμένη μορφή με όλες τις ταξινομικές ομάδες όλων των επιπέδων.

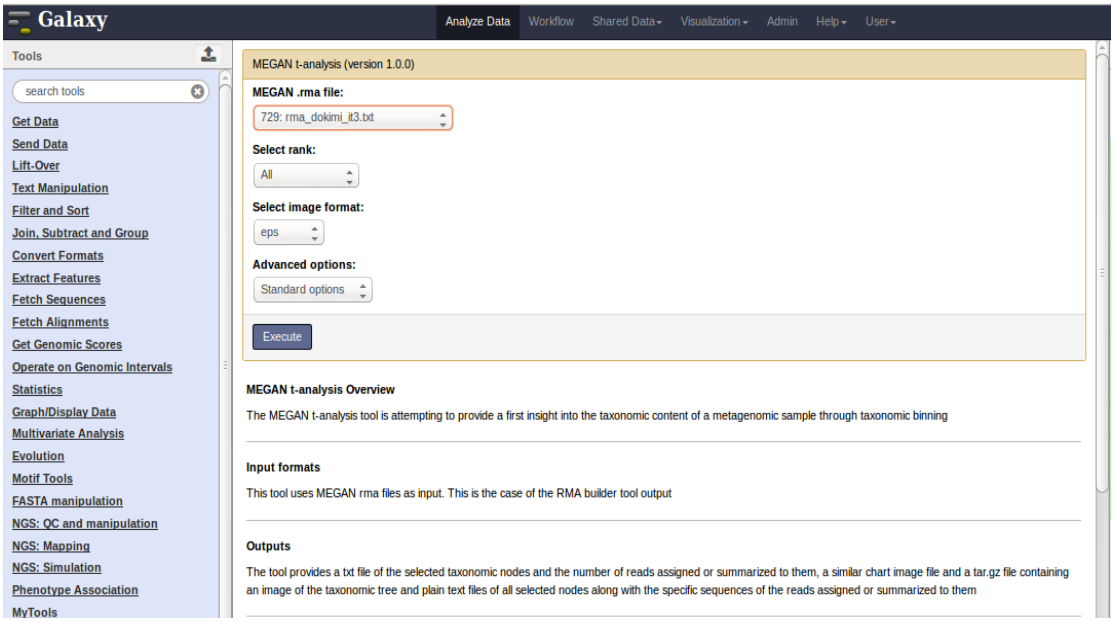
Στην δεύτερη περίπτωση ο χρήστης επιλέγει το επίπεδο της ταξινομικής ιεραρχίας που επιθυμεί συγκεκριμένα να μελετήσει. Οι επιλογές του είναι αυτές που προσφέρονται από το MEGAN (SuperKingdom, Kingdom, Phylum, Class, Order, Family, Varietas, Genus, Species_group, Subspecies, Species) και το ταξινομικό δέντρο ξεδιπλώνεται μέχρι το επίπεδο αυτό. Τα αρχεία εξόδου που δημιουργούνται είναι παρόμοια με τα προηγούμενα αλλά παρουσιάζουν κάποιες μικρές διαφοροποιήσεις. Αρχικά δημιουργείται ένα παρόμοιο διάγραμμα με το διάγραμμα της πρώτης περίπτωσης, με την διαφορά όμως ότι επειδή το ταξινομικό δέντρο έχει ξεδιπλωθεί μέχρι ένα συγκεκριμένο επίπεδο, κάποιες από τις αλληλουχίες που εμφανίζονται σε κάθε κόμβο, μπορεί να αντιστοιχίζονται στην πραγματικότητα σε επίπεδα χαμηλότερα στο ταξινομικό δέντρο. Επειδή όμως ο χρήστης ενδιαφέρεται για κάποιο συγκεκριμένο επίπεδο, είναι προφανές ότι αλληλουχίες που θα αντιστοιχίζονταν παραδείγματος χάριν επακριβώς στο επίπεδο των Species, ανήκουν επιπροσθέτως και στην αντίστοιχη κατηγορία Family από την οποία προέρχονται αυτά τα Species. Ουσιαστικά, δεν μιλάμε για assigned reads σε κάθε κόμβο αλλά για summarized reads όπως χαρακτηριστικά διαφοροποιούνται στο MEGAN, εννοώντας και τα reads που αντιστοιχίζονται στον κόμβο αυτό επακριβώς αλλά και όσα άλλα αντιστοιχίζονται σε αυτόν επειδή το δέντρο ξεδιπλώνεται μέχρι αυτό το σημείο, ενώ στην πραγματικότητα αντιστοιχίζονται επακριβώς σε κόμβους χαμηλότερου επιπέδου. Επίσης δημιουργείται ένα αντίστοιχο αρχείο κειμένου, με τα ονόματα των κόμβων και τα reads που συναντώνται, αποκλειστικά και μόνο για τους κόμβους του επιπέδου που έχει επιλεχθεί και χωρίς να παρουσιάζονται οι κόμβοι των no hits, των not assigned ή των low complexity. Στην περίπτωση αυτή λοιπόν έχει νόημα να μιλήσει κανείς για ποσοστά οργανισμών, σχετικές δηλαδή αναλογίες διαφορετικών ταξινομικών ομάδων για ένα συγκεκριμένο ταξινομικό επίπεδο, και μάλιστα αυτά παρουσιάζονται ως μια στήλη στο αρχείο κειμένου. Τέλος, δημιουργείται ένας παρόμοιος με την άλλη περίπτωση συμπιεσμένος φάκελος που περιέχει αρχεία για τον κάθε κόμβο με τις αλληλουχίες που έχουν

γίνει summarized σε αυτόν, καθώς και μια εικόνα του ταξινομικού δέντρου ξεδιπλωμένου μέχρι το επίπεδο που έχει επιλεχθεί.

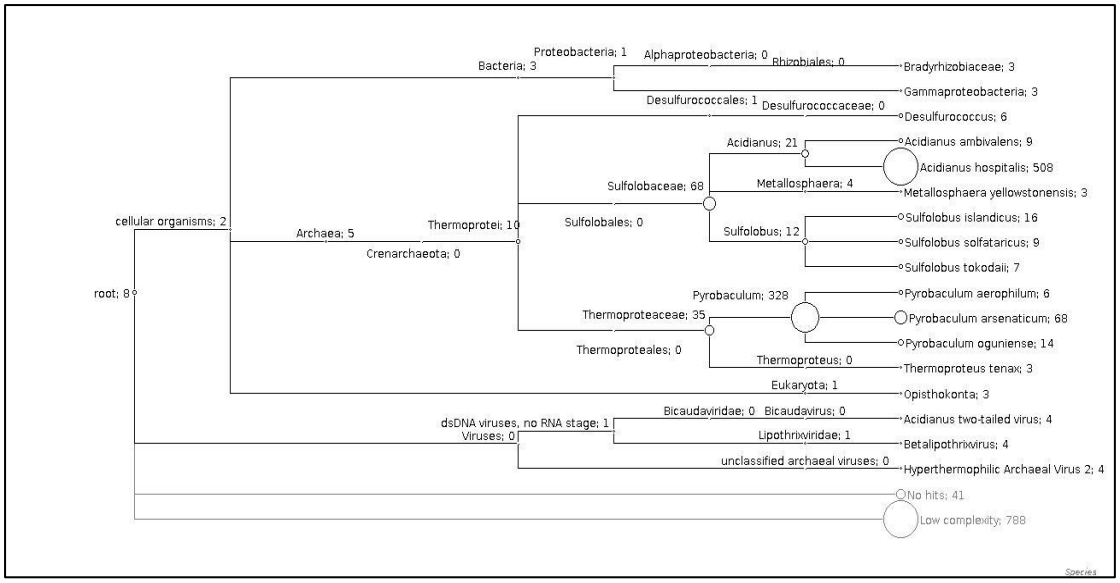


Εικόνα 31 Τα FASTA αρχεία για κάθε ταξινομική ομάδα που περιέχουν τις αλληλουχίες ανάγνωσης που έχουν αντιστοιχιστεί σε κάθε μια από αυτές.

Ο χρήστης έχει την δυνατότητα επιλογής του είδους της εικόνας του ταξινομικού δέντρου με βάση τις επιλογές που προσφέρει το ίδιο το MEGAN (bmp, eps, gif, jpg, pdf, png, svg), ενώ λόγω περιορισμού στην αναγνώριση αρχείων εικόνας από το GALAXY η εικόνα του διαγράμματος είναι σε συγκεκριμένη jpg μορφή. Για τον ίδιο λόγο, ο συμπιεσμένος φάκελος tar.gz είναι σε μορφή txt και απαιτεί μετονομασία με προσθήκη της κατάληξης tar.gz ώστε ο χρήστης να αποκτήσει πρόσβαση στα περιεχόμενα του.



Εικόνα 32 Το εργαλείο t-analysis στο επίπεδο των Standard Options. Επιλογή του αρχείου rma, του ταξινομικού επιπέδου ανάλυσης και του format της εικόνας του ταξινομικού δέντρου.



Εικόνα 33 Παράδειγμα ταξινομικού δέντρου, όπως αυτό δημιουργείται από το εργαλείο MEGAN t-analysis.

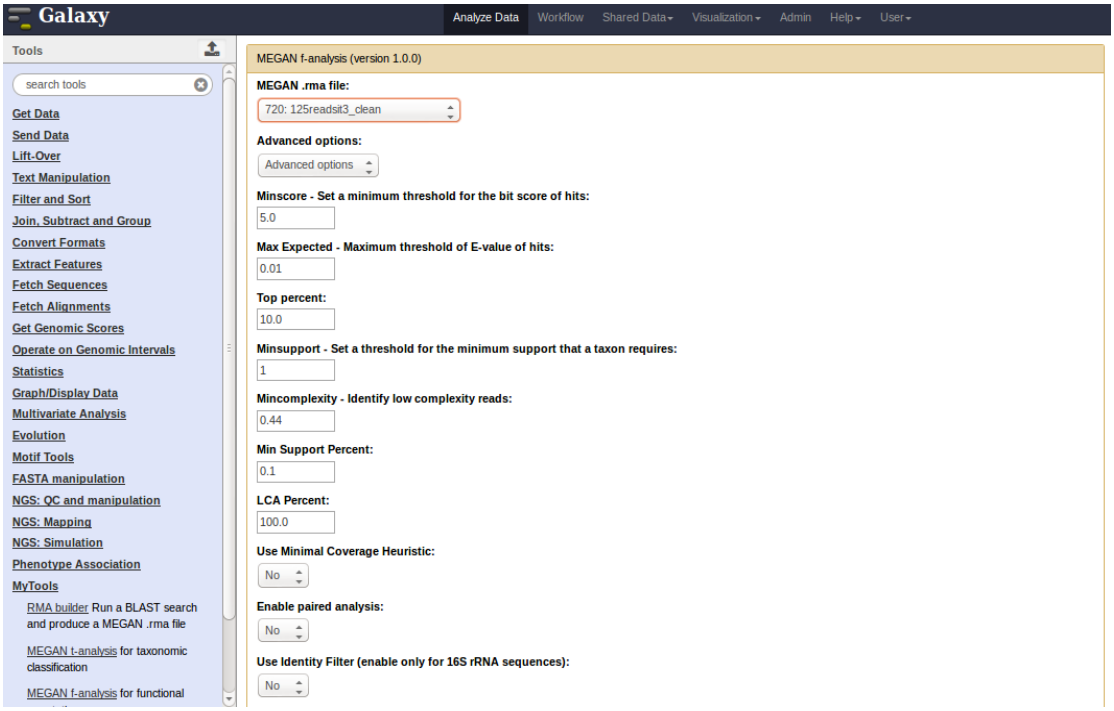
4.3 MEGAN f - analysis

Το τρίτο κατά σειρά εργαλείο παρέχει την δυνατότητα στον ερευνητή να αποκτήσει μια πρώτη εικόνα ως προς το λειτουργικό προφίλ του μεταγονιδιωματικού δείγματος που μελετάται. Το εργαλείο ονομάζεται **MEGAN f-analysis (tool id = megan_analysisf)**, αντλεί τον βασικό κώδικα του από το αρχείο **functionalsub.py** και δημιουργείται ως διεπαφή στην πλατφόρμα από το XML αρχείο **megan_analysisf.xml** που δηλώνεται στο GALAXY. Ο κώδικας και το XML αρχείο παρουσιάζονται στο παράρτημα.

```
<section name="MyTools" id="mTools">
    <tool file="myTools/rma_builder.xml"/>
    <tool file="myTools/megan_analysis.xml"/>
    <tool file="myTools/megan_analysisf.xml"/>
</section>
```

Γίνεται προσπάθεια, πάντα μέσα στα πλαίσια που επιτρέπει το ίδιο το MEGAN, που αποτελεί το εργαλείο της ανάλυσης, να πραγματοποιηθεί ένας λειτουργικός χαρακτηρισμός πάνω στο δείγμα, μέσω τριών διαφορετικών προσεγγίσεων. Αξίζει να σημειωθεί πως, ως αρχείο εισόδου στο εργαλείο αυτό χρησιμοποιείται το αρχείο εξόδου του εργαλείου RMA builder δηλαδή το αρχείο RMA που δημιουργείται. Για την πραγματοποίηση των τριών ειδών λειτουργικής το MEGAN απαιτεί συγκεκριμένες πληροφορίες, όπως τα RefSeq ids, που προκύπτουν από βάσεις δεδομένων όπως η NCBI-NR, συνεπώς η χρήση του εργαλείου MEGAN f-analysis αποκτά νόημα μόνο στην περίπτωση που στο πρώτο εργαλείο έχει γίνει επιλογή της BLASTX σύγκρισης.

Αρχικά, όπως και στην περίπτωση του εργαλείου MEGAN t-analysis, επιχειρείται μια ρύθμιση των παραμέτρων του LCA αλγορίθμου. Υπάρχει φυσικά η δυνατότητα χρήσης των προεπιλεγμένων τιμών, δίχως ο χρήστης να έρθει σε επαφή με τις παραμέτρους μέσω των Standard Options αλλά και αλλαγή των default τιμών μέσω της επιλογής Advanced Options. Οι παράμετροι που ρυθμίζονται είναι ακριβώς ίδιες με εκείνες που ρυθμίστηκαν στο δεύτερο εργαλείο και για τον λόγο αυτό κρίνεται σκόπιμο να μην γίνει ξανά αναλυτική παρουσίαση τους.



Εικόνα 34 Το εργαλείο MEGAN f-analysis στο επίπεδο των Advanced Options. Οι τιμές των παραμέτρων είναι οι προεπιλεγμένες.

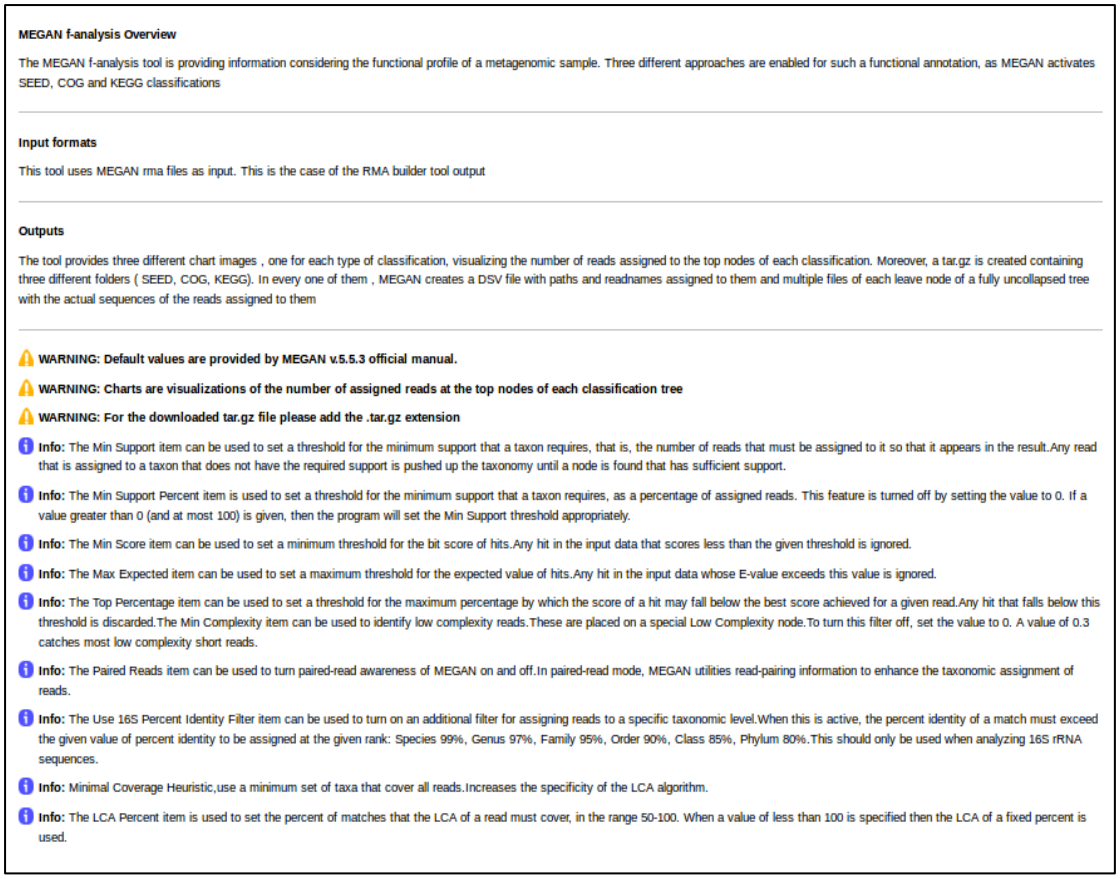
Μέσω του ιστορικού του GALAXY αλλά και της δυνατότητας έλεγχου των παραμέτρων που χαρακτηρίζουν μια ήδη ολοκληρωμένη ανάλυση, ο χρήστης μπορεί να θυμηθεί τις τιμές των παραμέτρων που χρησιμοποίησε κατά την ταξινομική ανάλυση και είτε να κάνει τις ίδιες επιλογές, είτε εάν κρίνει σκόπιμο ,να μελετήσει το λειτουργικό περιεχόμενο υπό διαφορετικές συνθήκες και παραμέτρους.

Ουσιαστικά, μέσω του εργαλείου αυτού πραγματοποιούνται οι τρεις βασικές προσεγγίσεις λειτουργικής ανάλυσης του MEGAN, η SEED, η COG και τέλος η KEGG αναλύσεις. Το περιβάλλον των αναλύσεων αυτών κατά την άμεση επαφή του χρήστη με το MEGAN, χαρακτηρίζεται από έναν ιδιαίτερα σημαντικό βαθμό διαδραστικότητας και αλληλεπίδρασης χρήστη και προγράμματος, συνεπώς εκ των πραγμάτων κάποιες από τις επιλογές είναι δύσκολο να αποδοθούν μέσω ενός εύχρηστου και γρήγορου εργαλείου σε μια πλατφόρμα όπως το GALAXY. Το εργαλείο λοιπόν παράγει τέσσερα βασικά αρχεία εξόδου. Αρχικά, για κάθε μια από

τις αναλύσεις, δημιουργείται ένα διάγραμμα που περιέχει τον αριθμό των αλληλουχιών που αντιστοιχίζονται στον πρώτο κόμβο (top) του δέντρου κάθε ανάλυσης και παρέχει μια πρώτη επαφή με το λειτουργικό περιεχόμενο του δείγματος, υπό την οπτική γωνία της κάθε προσέγγισης.

Τα γονίδια μπορούν να αντιστοιχιστούν σε λειτουργικούς ρόλους, τους οποίους επιτελούν στα πλαίσια συμμετοχής τους σε βιολογικά υποσυστήματα[43]. Κατά την SEED ανάλυση ,το πρόγραμμα επιχειρεί να αντιστοιχίσει τις αλληλουχίες ανάγνωσης με γονίδια που παρουσιάζουν γνωστούς λειτουργικούς ρόλους και στην συνέχεια να τα κατατάξει σε ένα ή περισσότερα βιολογικά υποσυστήματα. Έτσι, στο διάγραμμα που εμφανίζεται ως αρχείο εξόδου για την SEED ανάλυση, παρουσιάζεται ο αριθμός των αλληλουχιών που είναι δυνατό να αντιστοιχιστούν σε συγκεκριμένα βιολογικά συστήματα και διεργασίες - στην γενικότερη και ευρύτερη μορφή τους, δηλαδή το πρώτο (top) επίπεδο ανάλυσης. Αντίστοιχα, στην περίπτωση της ανάλυσης COG[44], οι αλληλουχίες ανάγνωσης αντιστοιχίζονται σε COGs και NOGs, όπου κάθε ένα από αυτά περιέχει ένα γκρουπ πρωτεϊνών που θεωρούνται ορθόλογες για τουλάχιστον τρεις γενεές οργανισμών και πιθανότατα προέρχονται από έναν αρχαίο διατηρημένο κοινό τομέα. Αυτά τα σύνολα πρωτεϊνών, υπό το όνομα των COGs και NOGs, παρουσιάζουν συγκεκριμένους λειτουργικούς ρόλους αλλά είναι δυνατό να αντιστοιχιστούν και σε γενικότερα λειτουργικά σύνολα. Στο διάγραμμα που προκύπτει λοιπόν για τα σύνολα αυτά, παρουσιάζεται ο αριθμός των αλληλουχιών που αντιστοιχούν σε κάθε ένα τέτοιο γενικό λειτουργικό σύνολο στο τοπ επίπεδο του δέντρου (Cellular Processes and Signaling, Information Storage and Processing, Metabolism, Poorly Characterized). Τέλος, στην περίπτωση της ανάλυσης KEGG[45] οι αλληλουχίες ανάγνωσης τοποθετούνται στην πορεία μεταβολικών μονοπατιών, ανάλογα με τον λειτουργικό ρόλο των πρωτεϊνών που είναι πιθανό να προκύψουν από αυτές. Το διάγραμμα που προκύπτει ως έξοδος, παρουσιάζει τον αριθμό των αλληλουχιών που αντιστοιχίζονται στις κορυφές (σημεία έναρξης) αυτών των μονοπατιών.

Εκτός όμως από τα τρία αυτά διαγράμματα που προκύπτουν, δημιουργείται και ένα τέταρτο αρχείο εξόδου, ένας φάκελος tar.gz που περιέχει τρεις υποφακέλους, έναν για κάθε είδος ανάλυσης. Στους υποφακέλους αυτούς περιέχεται υπό την μορφή κειμένου ένα αρχείο DSV (delimiter separated values), για το οποίο χρησιμοποιούνται tabs για τον ορισμό των στηλών και περιέχει το πλήρως ξεδιπλωμένο δέντρο που περιγράφει κάθε ανάλυση και τα ονόματα των αλληλουχιών ανάγνωσης που αντιστοιχίζονται στους ακραίους κόμβους αυτού του δέντρου. Επίσης, σε κάθε υποφάκελο, παρουσιάζονται αναλυτικά, με το όνομα αλλά και την ίδια την αλληλουχία βάσεων, ως συμπλήρωμα του ανωτέρω DSV αρχείου, οι αλληλουχίες ανάγνωσης που αντιστοιχίζονται στους ακραίους κόμβους του δέντρου της ανάλυσης, δηλαδή στο πιο αναλυτικό επίπεδο αντιστοίχισης αλληλουχιών, υπό την μορφή αρχείων FASTA.



MEGAN f-analysis Overview

The MEGAN f-analysis tool is providing information considering the functional profile of a metagenomic sample. Three different approaches are enabled for such a functional annotation, as MEGAN activates SEED, COG and KEGG classifications

Input formats

This tool uses MEGAN rma files as input. This is the case of the RMA builder tool output

Outputs

The tool provides three different chart images , one for each type of classification, visualizing the number of reads assigned to the top nodes of each classification. Moreover, a tar.gz is created containing three different folders (SEED, COG, KEGG). In every one of them , MEGAN creates a DSV file with paths and readnames assigned to them and multiple files of each leave node of a fully uncollapsed tree with the actual sequences of the reads assigned to them

WARNING: Default values are provided by MEGAN v.5.5.3 official manual.

WARNING: Charts are visualizations of the number of assigned reads at the top nodes of each classification tree

WARNING: For the downloaded tar.gz file please add the .tar.gz extension

Info: The Min Support item can be used to set a threshold for the minimum support that a taxon requires, that is, the number of reads that must be assigned to it so that it appears in the result.Any read that is assigned to a taxon that does not have the required support is pushed up the taxonomy until a node is found that has sufficient support.

Info: The Min Support Percent item is used to set a threshold for the minimum support that a taxon requires, as a percentage of assigned reads. This feature is turned off by setting the value to 0. If a value greater than 0 (and at most 100) is given, then the program will set the Min Support threshold appropriately.

Info: The Min Score item can be used to set a minimum threshold for the bit score of hits.Any hit in the input data that scores less than the given threshold is ignored.

Info: The Max Expected item can be used to set a maximum threshold for the expected value of hits.Any hit in the input data whose E-value exceeds this value is ignored.

Info: The Top Percentage item can be used to set a threshold for the maximum percentage by which the score of a hit may fall below the best score achieved for a given read.Any hit that falls below this threshold is discarded.The Min Complexity item can be used to identify low complexity reads.These are placed on a special Low Complexity node.To turn this filter off, set the value to 0. A value of 0.3 catches most low complexity short reads.

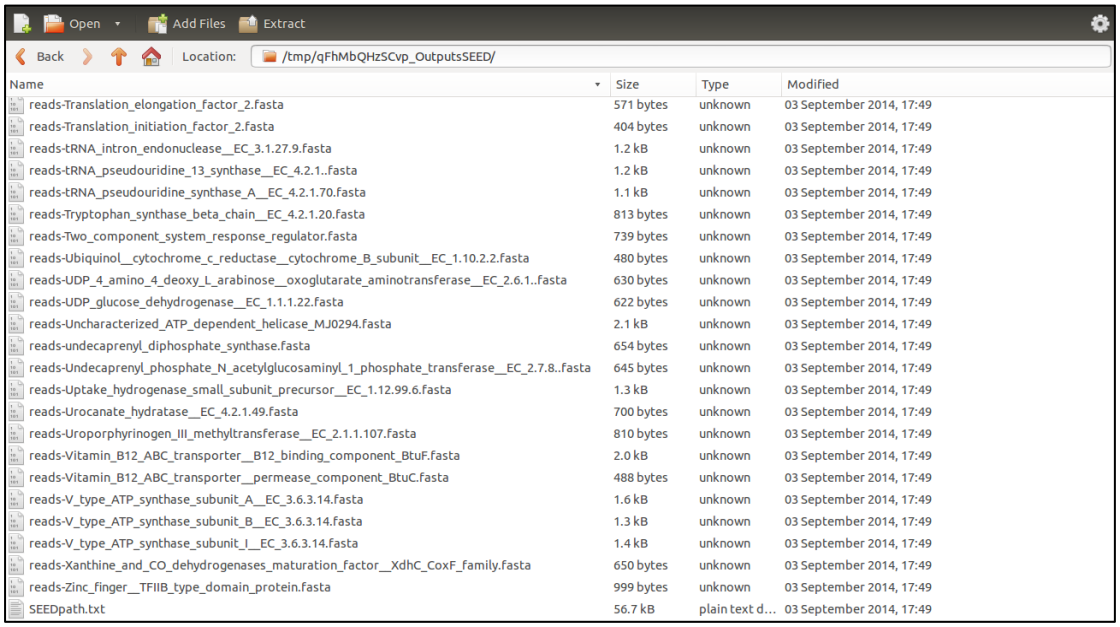
Info: The Paired Reads item can be used to turn paired-read awareness of MEGAN on and off.In paired-read mode, MEGAN utilities read-pairing information to enhance the taxonomic assignment of reads.

Info: The Use 16S Percent Identity Filter item can be used to turn on an additional filter for assigning reads to a specific taxonomic level.When this is active, the percent identity of a match must exceed the given value of percent identity to be assigned at the given rank: Species 99%, Genus 97%, Family 95%, Order 90%, Class 85%, Phylum 80%.This should only be used when analyzing 16S rRNA sequences.

Info: Minimal Coverage Heuristic,use a minimum set of taxa that cover all reads.Increases the specificity of the LCA algorithm.

Info: The LCA Percent item is used to set the percent of matches that the LCA of a read must cover, in the range 50-100. When a value of less than 100 is specified then the LCA of a fixed percent is used.

Εικόνα 35 Τα βοηθητικά μηνύματα του εργαλείου MEGAN f-analysis. Περιγραφή της λειτουργίας, οδηγίες χρήσης καθώς και παρουσίαση των παραμέτρων που διατίθενται ως επιλογή στις Advanced Options.

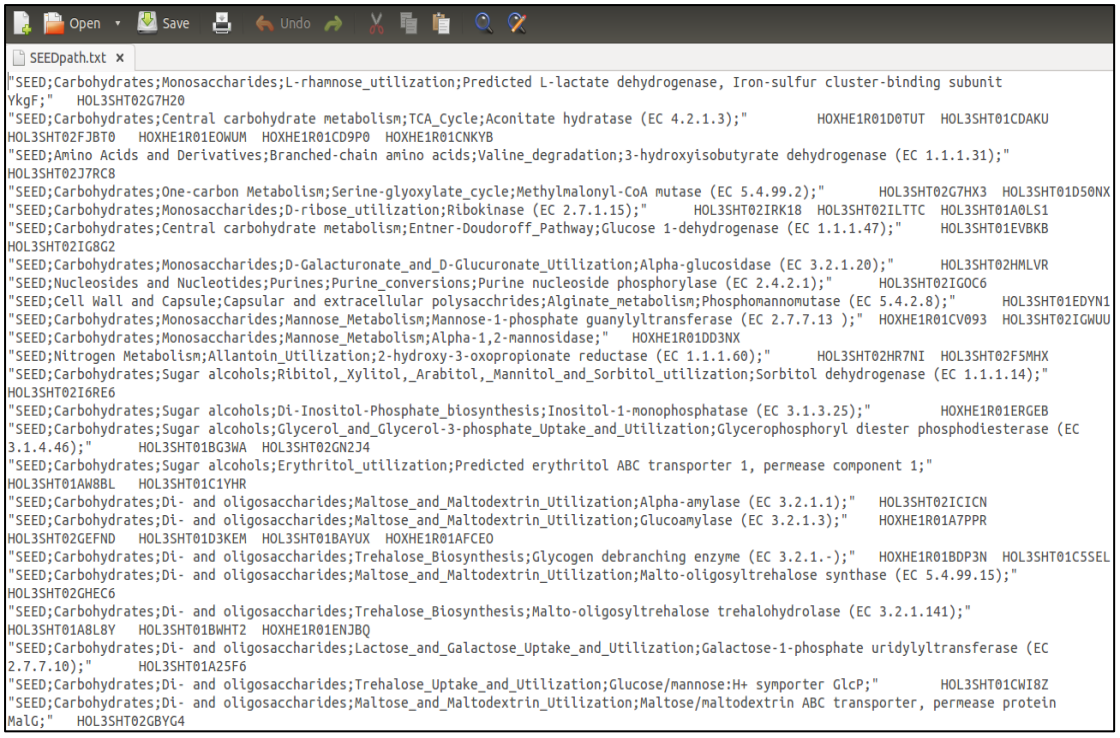


Open ▾ Add Files Extract

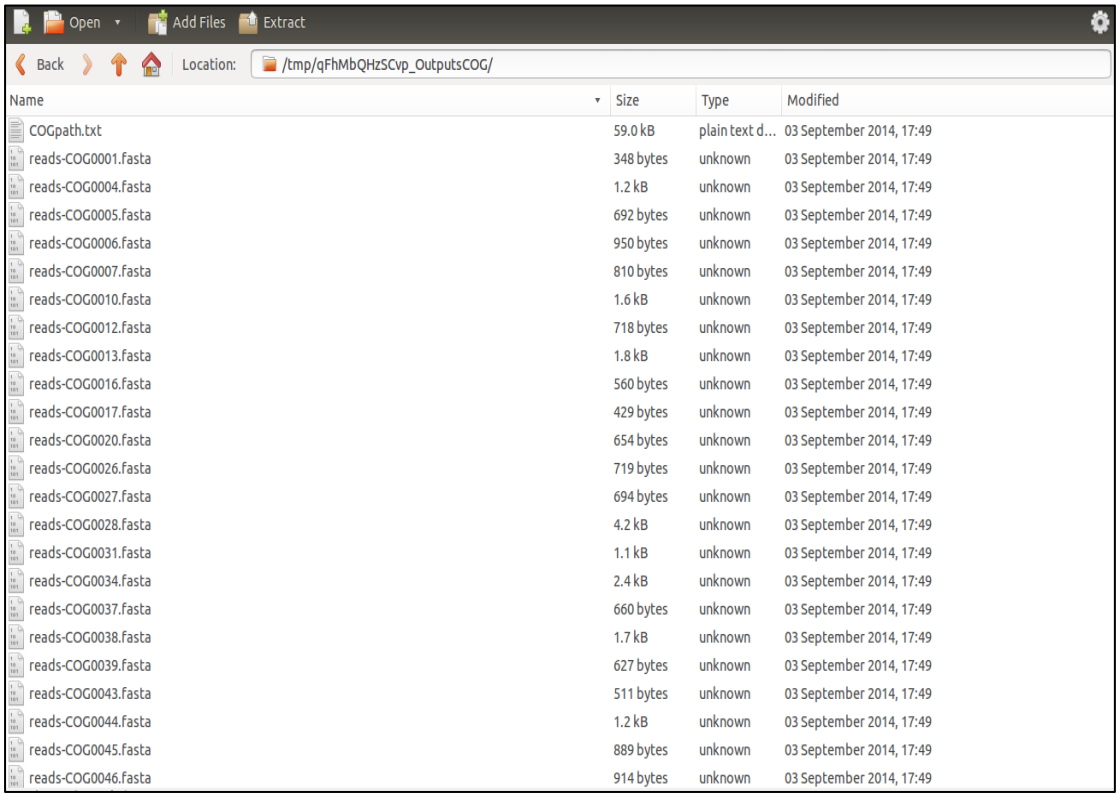
Back Location: /tmp/qFhMbQHzSCvp_OutputsSEED/

Name	Size	Type	Modified
reads-Translation_elongation_factor_2.fasta	571 bytes	unknown	03 September 2014, 17:49
reads-Translation_initiation_factor_2.fasta	404 bytes	unknown	03 September 2014, 17:49
reads-tRNA_intron_endonuclease__EC_3.1.27.9.fasta	1.2 kB	unknown	03 September 2014, 17:49
reads-tRNA_pseudouridine_13_synthase__EC_4.2.1..fasta	1.2 kB	unknown	03 September 2014, 17:49
reads-tRNA_pseudouridine_synthase_A__EC_4.2.1.70.fasta	1.1 kB	unknown	03 September 2014, 17:49
reads-Tryptophan_synthase_beta_chain__EC_4.2.1.20.fasta	813 bytes	unknown	03 September 2014, 17:49
reads-Two_component_system_response_regulator.fasta	739 bytes	unknown	03 September 2014, 17:49
reads-Ubiquinol__cytochrome_c_reductase__cytochrome_B_subunit__EC_1.10.2.2.fasta	480 bytes	unknown	03 September 2014, 17:49
reads-UDP_4_amino_4_deoxy_L_arabinose__oxoglutarate_aminotransferase__EC_2.6.1..fasta	630 bytes	unknown	03 September 2014, 17:49
reads-UDP_glucose_dehydrogenase__EC_1.1.1.22.fasta	622 bytes	unknown	03 September 2014, 17:49
reads-Uncharacterized_ATP_dependent_helicase_MJ0294.fasta	2.1 kB	unknown	03 September 2014, 17:49
reads-undecaprenyl_diphosphate_synthase.fasta	654 bytes	unknown	03 September 2014, 17:49
reads-Undecaprenyl_phosphate_N_acetylglucosaminyl_1_phosphate_transferase__EC_2.7.8..fasta	645 bytes	unknown	03 September 2014, 17:49
reads-Uptake_hydrogenase_small_subunit_precursor__EC_1.12.99.6.fasta	1.3 kB	unknown	03 September 2014, 17:49
reads-Urocanate_hydratase__EC_4.2.1.49.fasta	700 bytes	unknown	03 September 2014, 17:49
reads-Uroporphyrinogen_III_methyltransferase__EC_2.1.1.107.fasta	810 bytes	unknown	03 September 2014, 17:49
reads-Vitamin_B12_ABC_transporter__B12_binding_component_BtuF.fasta	2.0 kB	unknown	03 September 2014, 17:49
reads-Vitamin_B12_ABC_transporter__permease_component_BtuC.fasta	488 bytes	unknown	03 September 2014, 17:49
reads-V_type_ATP_synthase_subunit_A__EC_3.6.3.14.fasta	1.6 kB	unknown	03 September 2014, 17:49
reads-V_type_ATP_synthase_subunit_B__EC_3.6.3.14.fasta	1.3 kB	unknown	03 September 2014, 17:49
reads-V_type_ATP_synthase_subunit_I__EC_3.6.3.14.fasta	1.4 kB	unknown	03 September 2014, 17:49
reads-Xanthine_and_CO_dehydrogenases_maturation_factor__XdhC_CoxF_family.fasta	650 bytes	unknown	03 September 2014, 17:49
reads-Zinc_finger__TFIIB_type_domain_protein.fasta	999 bytes	unknown	03 September 2014, 17:49
SEEDpath.txt	56.7 kB	plain text d…	03 September 2014, 17:49

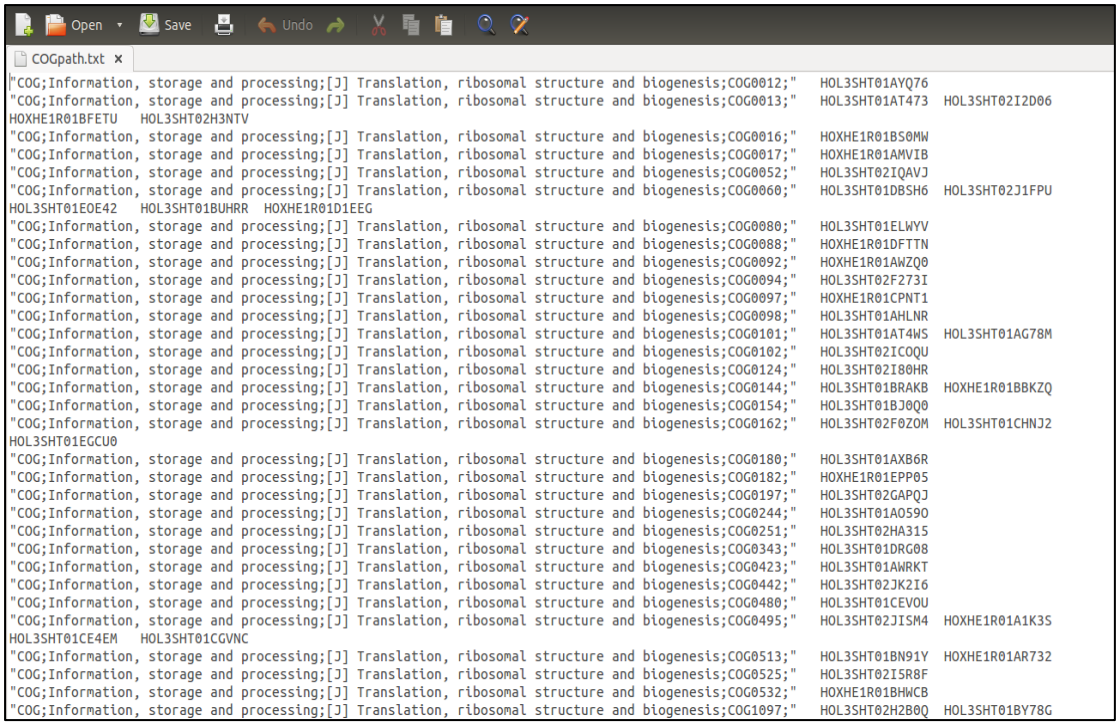
Εικόνα 36 Τα αρχεία που περιέχει ο φάκελος εξόδου tar.gz του εργαλείου MEGAN f-analysis για την ανάλυση SEED. Διακρίνεται το αρχείο κειμένου SEEDpath καθώς και μια σείρα αρχείων FASTA.



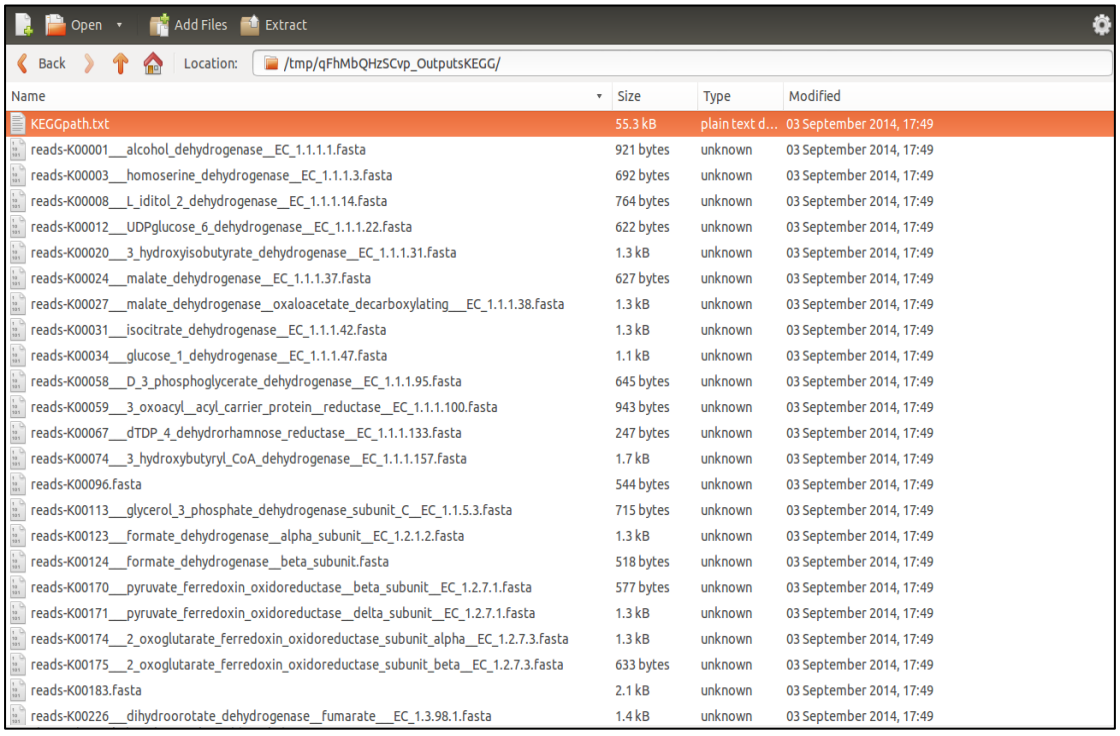
Εικόνα 37 **Τμήμα του εγγράφου που περιέχει τα μονοπάτια των υποσυστημάτων της SEED κατηγοριοποίησης και τις αλληλουχίες ανάγνωσης που αντιστοιχίζονται σε αυτά.**



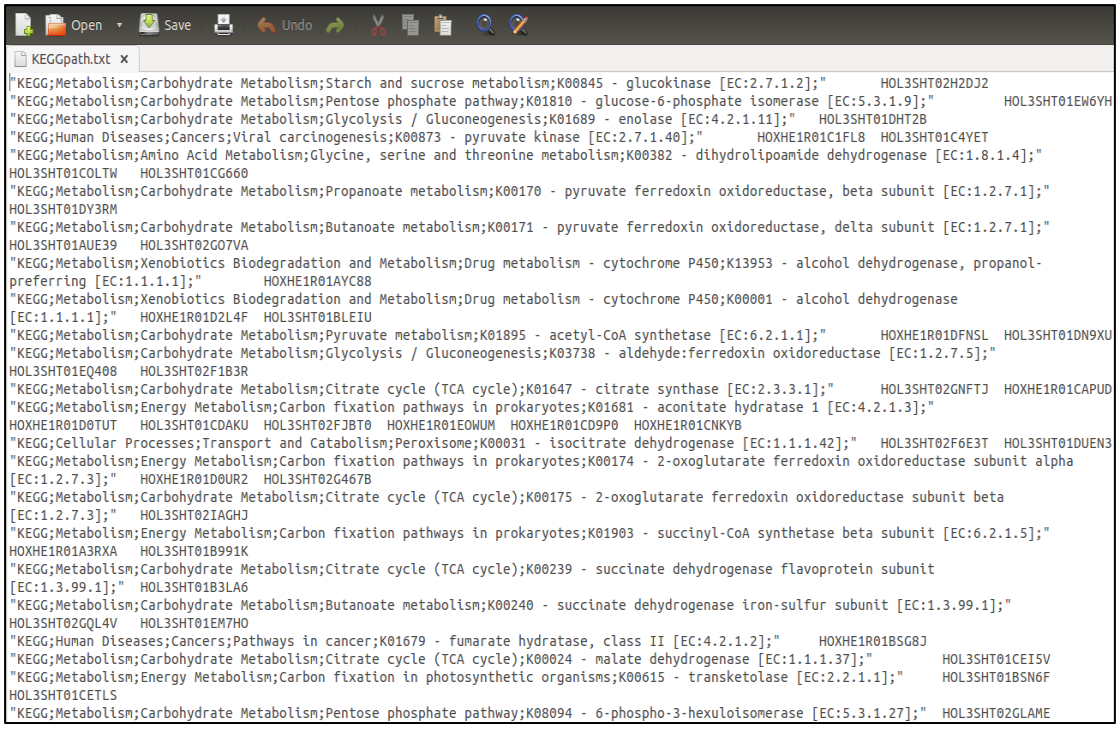
Εικόνα 38 **Τα αρχεία που περιέχει ο φάκελος εξόδου tar.gz του εργαλείου MEGAN f-analysis για την ανάλυση COG. Διακρίνεται το αρχείο κειμένου COGpath καθώς και μια σείρα αρχείων FASTA.**



Εικόνα 39 Τμήμα του εγγράφου που περιέχει τα μονοπάτια της COG κατηγοριοποίησης και τις αλληλουχίες ανάγνωσης που αντιστοιχίζονται σε αυτά.



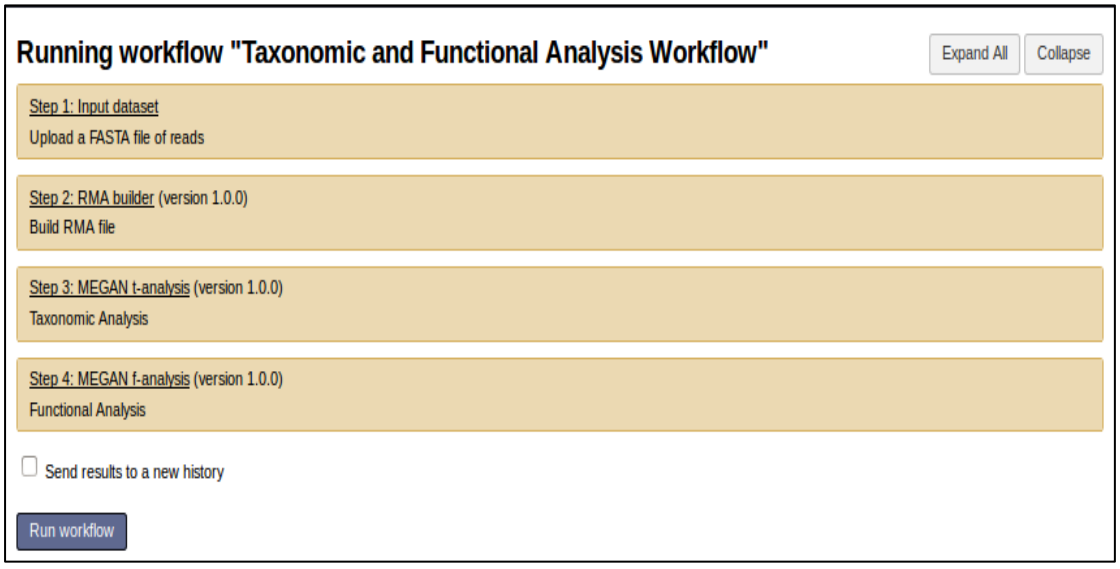
Εικόνα 40Τα αρχεία που περιέχει ο φάκελος εξόδου tar.gz του εργαλείου MEGAN f-analysis για την ανάλυση KEGG. Διακρίνεται το αρχείο κειμένου KEGGpath καθώς και μια σείρα αρχείων FASTA.



Εικόνα 41 Τμήμα του εγγράφου που περιέχει τα μεταβολικά μονοπάτια της KEGG κατηγοριοποίησης και τις αλληλουχίες ανάγνωσης που αντιστοιχίζονται σε αυτά.

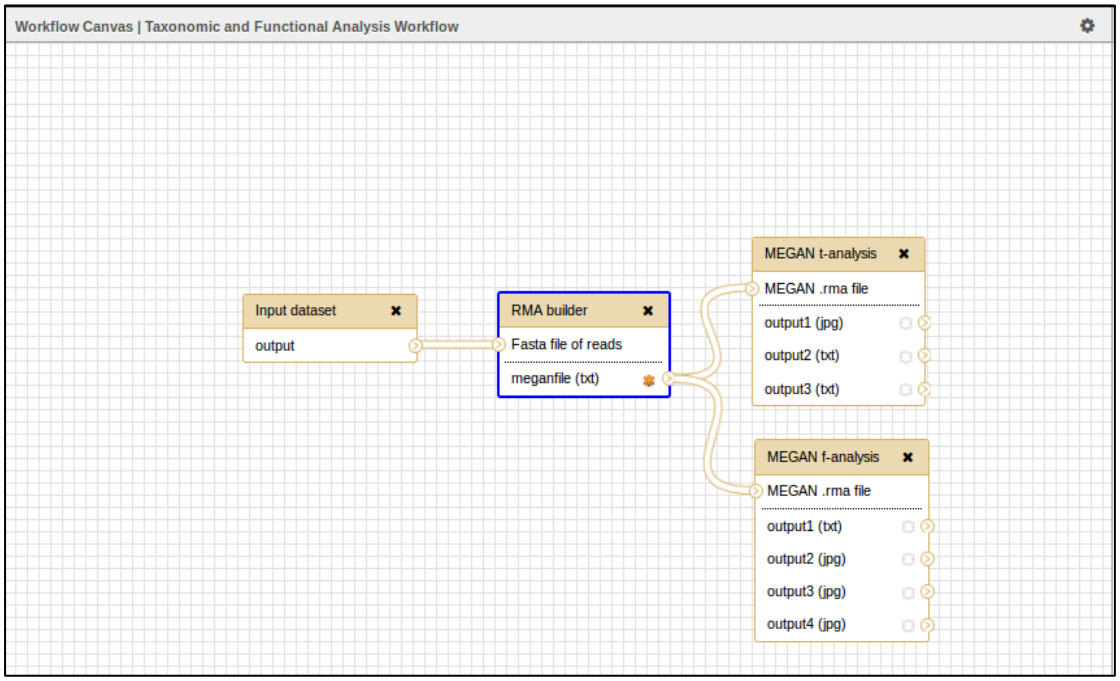
4.4 Taxonomic and Functional Analysis Workflow

Αφού ολοκληρώθηκε η κατασκευή των υπολογιστικών εργαλείων, δημιουργήθηκε μια ροη διεργασιών (workflow) που συμπεριλαμβάνει όλα τα υπολογιστικά εργαλεία ανάλογα με τον τρόπο διασύνδεσης των αρχείων εισόδου και εξόδου τους.



Εικόνα 42 Τα βήματα της ροής διεργασιών που δημιουργήθηκε με βάση τα υπολογιστικά εργαλεία που κατασκευάστηκαν.

Με τον τρόπο αυτό, δημιουργείται μια αυτοματοποιημένη διαδικασία κατά την οποία ο χρήστης παρέχει το αρχικό αρχείο FASTA που περιέχει τις αλληλουχίες ανάγνωσης του μεταγονιδιωματικού δείγματος και ρυθμίζει - εάν το επιθυμεί - τις παραμέτρους των υπολογιστικών εργαλείων. Η αυτοματοποιημένη διαδικασία οδηγεί αρχικά στην δημιουργία του αρχείου RMA μέσω του εργαλείου RMA builder, το οποίο χρησιμοποιείται ως είσοδος στο εργαλείο MEGAN t-analysis ώστε να παραχθούν τα αρχεία εξόδου της ταξινομικής ανάλυσης, καθώς και ως είσοδος για το εργαλείο MEGAN f-analysis που παρέχει αποτελέσματα λειτουργικής ανάλυσης υπό την μορφή συγκεκριμένων αρχείων εξόδου, όπως αυτά περιγράφονται ανωτέρω.



Εικόνα 43 Η σχηματική παρουσίαση της ροής διεργασιών "Taxonomic and Functional Analysis Workflow".

Η μελέτη του εξεταζόμενου μεταγονιδιωματικού δείγματος μέσω της προσέγγισης αυτής της ροής διεργασιών, παρέχει μια πρώτη εικόνα σε ότι αφορά το ταξινομικό και λειτουργικό περιεχόμενο του δείγματος καθώς και ένα αρχείο RMA που μπορεί να χρησιμοποιηθεί για περαιτέρω αναλύσεις και αναζητήσεις. Ο χρήστης δεν απαιτείται να ελέγχει ανά πάσα στιγμή τις διεργασίες που πραγματοποιούνται, ούτε να βρίσκεται πάνω από την πλατφόρμα ώστε να ρυθμίζει τα επόμενα βήματα. Εξασφαλίζεται η επαναληψιμότητα της ανάλυσης, ενώ οι παράμετροι είναι δυνατό να προσαρμοστούν εκ των προτέρων ώστε να επιτευχθεί ο επιθυμητός στόχος.

5. ΕΦΑΡΜΟΓΗ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΕΡΓΑΛΕΙΩΝ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

5.1 Απόκτηση δεδομένων και επεξεργασία

Τα δεδομένα που χρησιμοποιήθηκαν κατά την δοκιμή εφαρμογής των υπολογιστικών εργαλείων είναι δεδομένα που προέρχονται από το πρόγραμμα HotZyme. Το HotZyme (GA: 265933) είναι ένα σχέδιο που υποστηρίζεται από το 7ο πρόγραμμα πλαίσιο για την έρευνα και την τεχνολογική ανάπτυξη (FP7) της Ευρωπαϊκής Ένωσης. Το ερευνητικό αυτό έργο, είναι ένα μεγάλης κλίμακας συλλογικό έργο για τη συστηματική εξέταση και μελέτη για νέα γονίδια ή ενζυματικά είδη λειτουργικότητας που προέρχονται από ζεστά περιβάλλοντα.

Η επεξεργασία των δεδομένων, έγινε στον server Helios του Πανεπιστημίου της Κοπεγχάγης με σκοπό την αύξηση της υπολογιστικής δύναμης κατά την πραγματοποίηση των αναλύσεων. Στην πραγματικότητα, οι αναλύσεις δειγμάτων πολλών χιλιάδων ή ακόμα και εκατομμυρίων αλληλουχιών ανάγνωσης που προέρχονται από μεταγονιδιωματικά δείγματα, ειδικά στο επίπεδο των BLAST συγκρίσεων, απαιτούν τεράστια υπολογιστική δύναμη αλλά και πολυ μεγάλο καθαρό υπολογιστικό χρόνο. Κάτι τέτοιο, με τα διαθέσιμα μέσα δεν ήταν δυνατό να πραγματοποιηθεί. Για τον λόγο αυτό, μέσω ενός αλγορίθμου επιλογής τυχαίων γραμμών (μη επαναλαμβανόμενων) από τα εξεταζόμενα αρχεία FASTA των δειγμάτων, επιλέχθηκαν 2000 τυχαίες αλληλουχίες ανάγνωσης προς εξέταση. Οι αλληλουχίες αυτές μπορεί να αποτελούν ένα τυχαίο υποσύνολο του κάθε αρχείου, αλλά προφανέστατα ο αριθμός τους σε σχέση με τον αρχικό αριθμό δεν μπορεί να οδηγήσει σε αναλύσεις που να είναι ενδεικτικές της πραγματικής κατάστασης που περιγράφει το δείγμα. Χρησιμοποιούνται λοιπόν περισσότερο ως πραγματικά μεταγονιδιωματικά δεδομένα δοκιμής της ορθής λειτουργίας των υπολογιστικών εργαλείων, τα οποία μπορούν να βρουν εφαρμογή και σε πλήρη σετ μεταγονιδιωματικών δεδομένων εφόσον υπάρχει η κατάλληλη διαθέσιμη υπολογιστική δύναμη.

5.2 Εφαρμογή και αποτελέσματα

Ακολουθεί ενδεικτική παρουσίαση των αποτελεσμάτων και των αρχείων εξόδου για μια σειρά από μεταγονιδιωματικά δείγματα. Η ταξινομική ανάλυση επιλέχθηκε να παρουσιαστεί στο επίπεδο των ειδών (Species), αλλά φυσικά θα ήταν δυνατό να παρουσιαστεί σε οποιοδήποτε άλλο ταξινομικό επίπεδο διατίθεται μέσω του MEGAN. Γίνεται προσπάθεια να παρουσιαστούν όσο το δυνατόν, περισσότερες πληροφορίες που προκύπτουν από τα αποτελέσματα των αναλύσεων γνωρίζοντας εκ των πραγμάτων μια σειρά από περιορισμούς που τίθενται λόγω της φύσης και του μεγέθους ορισμένων αρχείων εξόδου που προκύπτουν από τα εργαλεία. Για τον λόγο αυτό, παρουσιάζονται ως αποτελέσματα τα διαγράμματα της ταξινομικής και

της λειτουργικής ανάλυσης, για κάθε ένα από τα δείγματα που επιλέχθηκαν προς ανάλυση, καθώς και τα αντίστοιχα αρχεία κειμένου που παρέχουν τις ίδιες πληροφορίες. Η παρουσίαση των αρχείων εξόδου που εκ των πραγμάτων δεν είναι δυνατό να παρουσιαστούν σε ένα αρχείο κειμένου, όπως διάφορων αρχείων DSV και FASTA, περιορίζεται στην παραδειγματικής φύσης εμφάνιση τους στο κεφάλαιο της Ανάπτυξης και Εφαρμογής των Υπολογιστικών Εργαλείων.

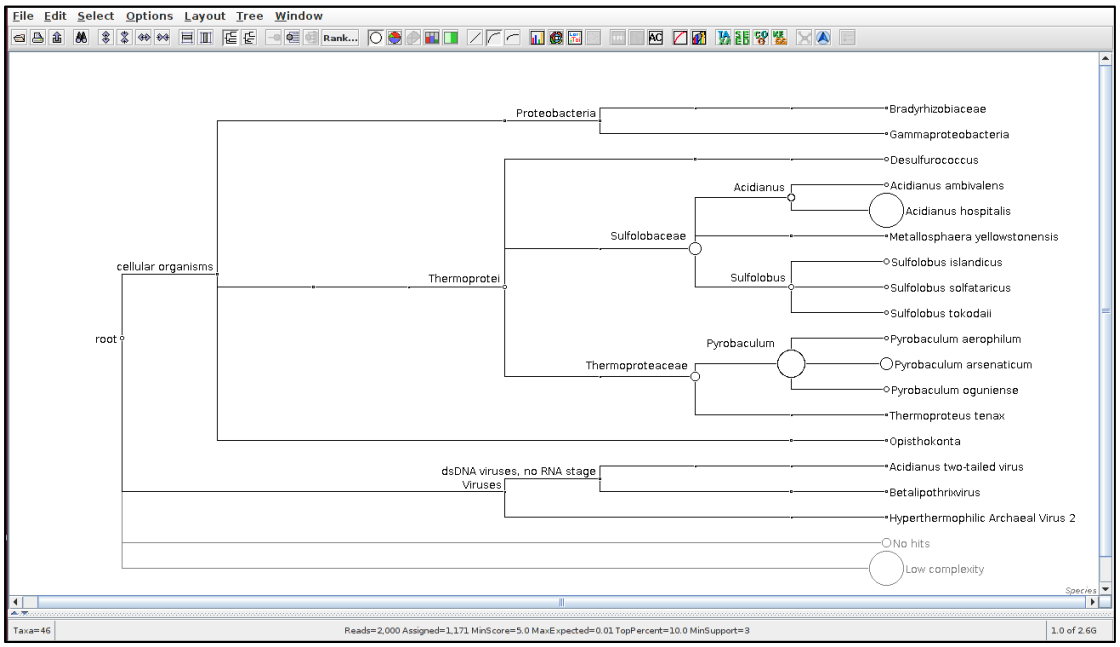
Για τις συγκρίσεις μέσω BLAST, θα γίνει παρουσίαση των περιπτώσεων σύγκρισης BLASTX καθώς αυτές εμφανίζουν ενδιαφέρον δίνοντας την επιπρόσθετη δυνατότητα λειτουργικής ανάλυσης. Οι παράμετροι του BLAST αλλά και του MEGAN, χρησιμοποιήθηκαν με τις default τιμές τους, γεγονός που επιτρέπει την εμφάνιση μιας γενικά αξιόπιστης ανάλυσης, καλύπτοντας μια ευρύτερη γκάμα δεδομένων, δίχως όμως να παραγνωρίζεται ότι σε δεδομένες περιπτώσεις η προσαρμογή των παραμέτρων στο είδος των δεδομένων αλλά και τους στόχους της ανάλυσης μπορεί να αποδώσει ακριβέστερα αποτελέσματα. Η μόνη παράμετρος που δεν χρησιμοποιήθηκε στην προεπιλεγμένη τιμής της είναι η E value, όπου τέθηκε ίση με 1.0 με στόχο να περιοριστεί ο αριθμός των hits που χαρακτηρίζονται από ιδιαίτερα μικρά σκορ και να επιταχυνθεί η σύγκριση μέσω BLAST. Επίσης, η παράμετρος του MEGAN Min Support χρησιμοποιήθηκε με τιμή ίση με 3, ώστε να απορρίπτονται ταξινομικές ομάδες στις οποίες αντιστοιχίζονται λιγότερες από τρείς αλληλουχίες ανάγνωσης, ώστε η ανάλυση να επικεντρωθεί στις εντονότερα παρούσες ταξινομικές ομάδες και να απορριφθεί ένας μεγάλος αριθμός ταξινομικών ομάδων με μια μόνο αντιστοιχισμένη αλληλουχία.

Δείγμα it3_clean2000

Το πρώτο δείγμα που μελετήθηκε ήταν το μεταγονιδιωματικό δείγμα **It-3** από το οποίο επιλέχθηκαν 2000 τυχαίες αλληλουχίες και δημιουργήθηκε το προς ανάλυση αρχείο FASTA **it3_clean2000**. Πρόκειται για αλληλουχίες ανάγνωσης που έχουν προκύψει από τη μέθοδο της 454 Titanium πυροαλληλούχισης και στο αρχικό αρχείο περιέχονται 674.766 αλληλουχίες. Το δείγμα προέρχεται από μια θερμή πηγή της περιοχής Pisciarelli της Ιταλίας, ενώ σε ότι έχει να κάνει με τις συνθήκες που επικρατούν στην πηγή η θερμοκρασία μετρήθηκε ίση με 85^{O}C και η τιμή του pH βρέθηκε ίση με 3.5 .

Αρχικά τα δεδομένα αναλύθηκαν από το εργαλείο RMA builder που δημιουργεί ως αρχείο εξόδου ένα txt αρχείο που χρησιμοποιείται ως είσοδος στα επόμενα εργαλεία. Εάν όμως σε αυτό το αρχείο, όπως έχει ήδη αναφερθεί, προστεθεί η κατάληξη .rma, το αρχείο αυτό μετατρέπεται σε έναν τύπο αρχείου read-match που είναι δυνατό να επεξεργαστεί άμεσα μέσω MEGAN όπως φαίνεται και στην εικόνα που ακολουθεί. Ο χρήστης μπορεί να αλλάξει τις παραμέτρους, να επιλέξει άλλα επίπεδα ανάλυσης και γενικά να επεξεργαστεί διαδραστικά την έξοδο του

εργαλείου αυτού εκμεταλλευόμενος πλήρως τις δυνατότητες του MEGAN. Από τις 2000 αλληλουχίες ,οι 1171 αντιστοιχίστηκαν σε κάποιον κόμβο που συμβολίζει μια ταξινομική ομάδα, ανεξαρτήτως ταξινομικού επιπέδου.

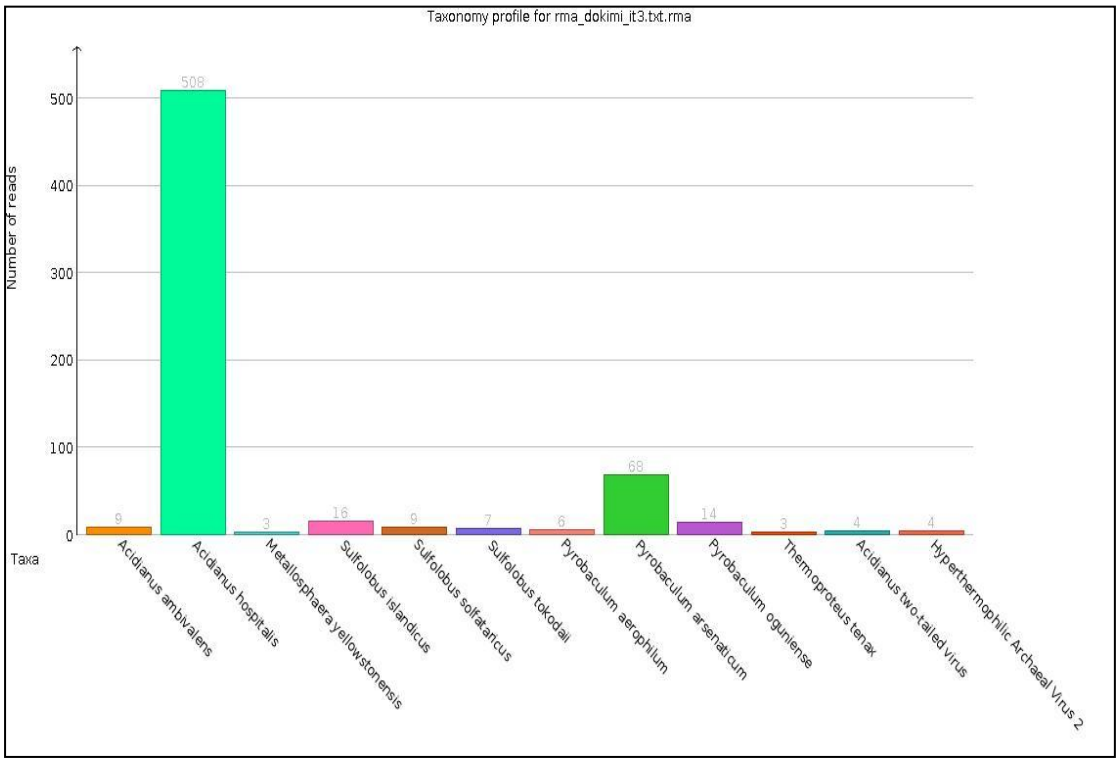


Εικόνα 44 Το βασικό παράθυρο (mainviewer) του αρχείου rma για τα δεδομένα του αρχείου it3_clean2000 στο ταξινομικό επίπεδο των Species.

Η έξοδος του πρώτου εργαλείου χρησιμοποιήθηκε ως είσοδος στο εργαλείο MEGAN t-analysis και προέκυψαν τα αρχεία εξόδου, όπως αυτά έχουν περιγραφεί κατά το κεφάλαιο της Ανάπτυξης και Περιγραφής των Υπολογιστικών Εργαλείων. Παρουσιάζεται, το διάγραμμα, υπό μορφή στηλών, των ταξινομικών ομάδων που ανήκουν στο επίπεδο των Species με τον αριθμό των αλληλουχιών ανάγνωσης που αντιστοιχίζονται σε κάθε μια από αυτές(Εικόνα 45).

Όπως φαίνεται από το παρακάτω διάγραμμα, το συγκεκριμένο δείγμα αλληλουχιών ανάγνωσης, υπό τις προεπιλεγμένες παραμέτρους του MEGAN και στο επίπεδο των Species, εμφανίζει κατά πρώτο λόγο έντονη παρουσία του είδους μικροοργανισμού **Acidianus hospitalis.** Οι συνθήκες της πηγής από την οποία προέρχεται το δείγμα, είναι γεγονός ότι πιθανότατα επιτρέπουν την ανάπτυξη του συγκεκριμένου είδους μικροοργανισμού, καθώς οι ιδανικές συνθήκες ανάπτυξης γενικότερα του γένους Acidianus κυμαίνονται από 65–95°C και σε συνθήκες τιμών pH από 2 έως 4[46]. Κατά δεύτερο λόγο ,σε σχετικά μικρότερο ποσοστό εμφανίζεται ο μικροοργανισμός **Pyrobaculum arsenaticum,** που πράγματι εντοπίζεται σε θερμές πηγές της περιοχής Pisciarelli[47], ενώ φυσικά υπάρχουν αντιστοιχίσεις ορισμένων αλληλουχιών και με είδη μικροοργανισμών όπως οι **Acidianus ambivalens, Sulfolobus islandicus, Sulfolobus solfataricus, Sulfolobus tokodaii, Pyrobaculum oguniense, Pyrobaculum aerophilum** σε σημαντικά μικρότερα ποσοστά, κάθε

ένας από τους οποίους ευνοείται από τις συνθήκες οξύτητας που επικρατούν σε τέτοιες θερμές πηγές.



Εικόνα 45 Διάγραμμα ταξινομικής ανάλυσης στο επίπεδο των Species για το δείγμα it3_clean2000. Στον οριζόντιο άξονα παρουσιάζονται τα διαφορετικά είδη που εντοπίστηκαν στο δείγμα ενώ στον κάθετο αντιστοιχίζεται ο αριθμός των αλληλουχιών ανάγνωσης που ανήκουν σε κάθε έναν από αυτούς τους κόμβους

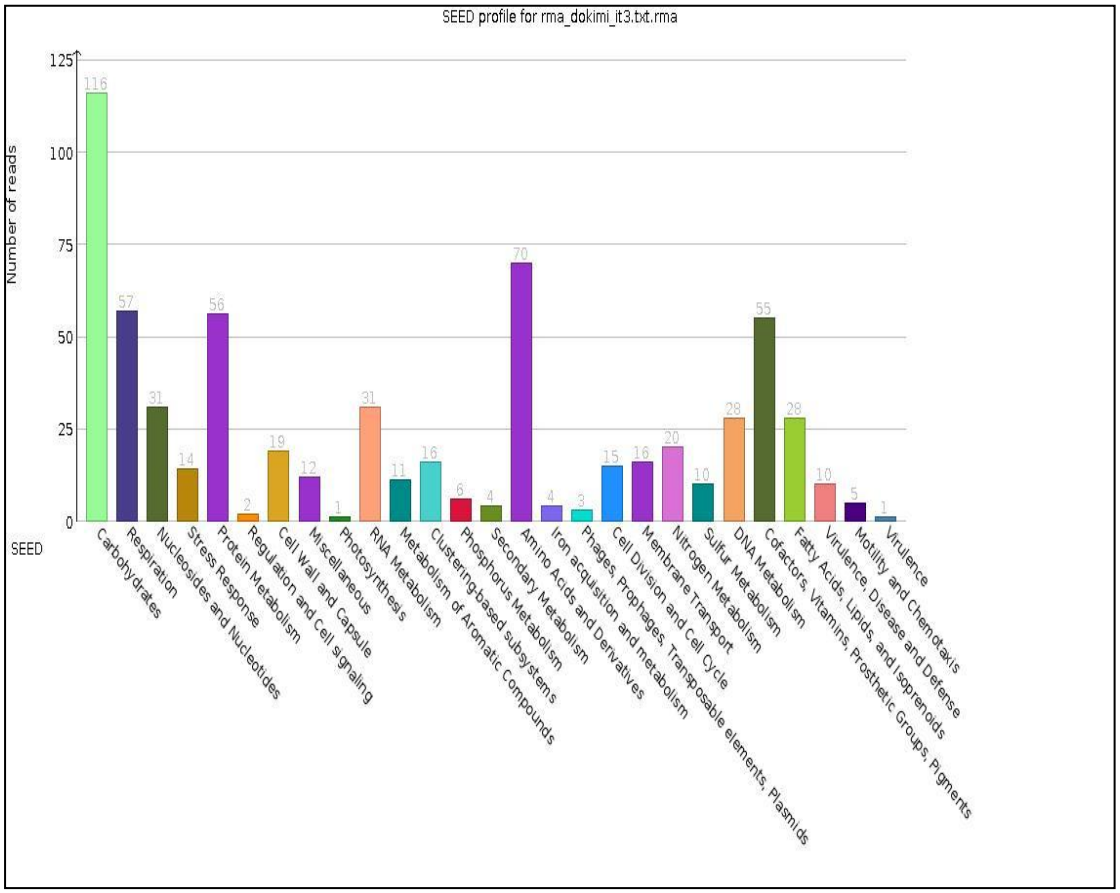
dataset_785.dat

```
#Series                    Reads   Percentages
Acidianus ambivalens       9.0     1.38248847926
Acidianus hospitalis       508.0   78.0337941628
Metallosphaera yellowstonensis  3.0     0.460829493088
Sulfolobus islandicus      16.0    2.45775729647
Sulfolobus solfataricus    9.0     1.38248847926
Sulfolobus tokodaii        7.0     1.0752688172
Pyrobaculum aerophilum     6.0     0.921658986175
Pyrobaculum arsenaticum    68.0    10.44546851
Pyrobaculum oguniense      14.0    2.15053763441
Thermoproteus tenax        3.0     0.460829493088
Acidianus two-tailed virus      4.0     0.614439324117
Hyperthermophilic Archaeal Virus 2      4.0     0.614439324117
```

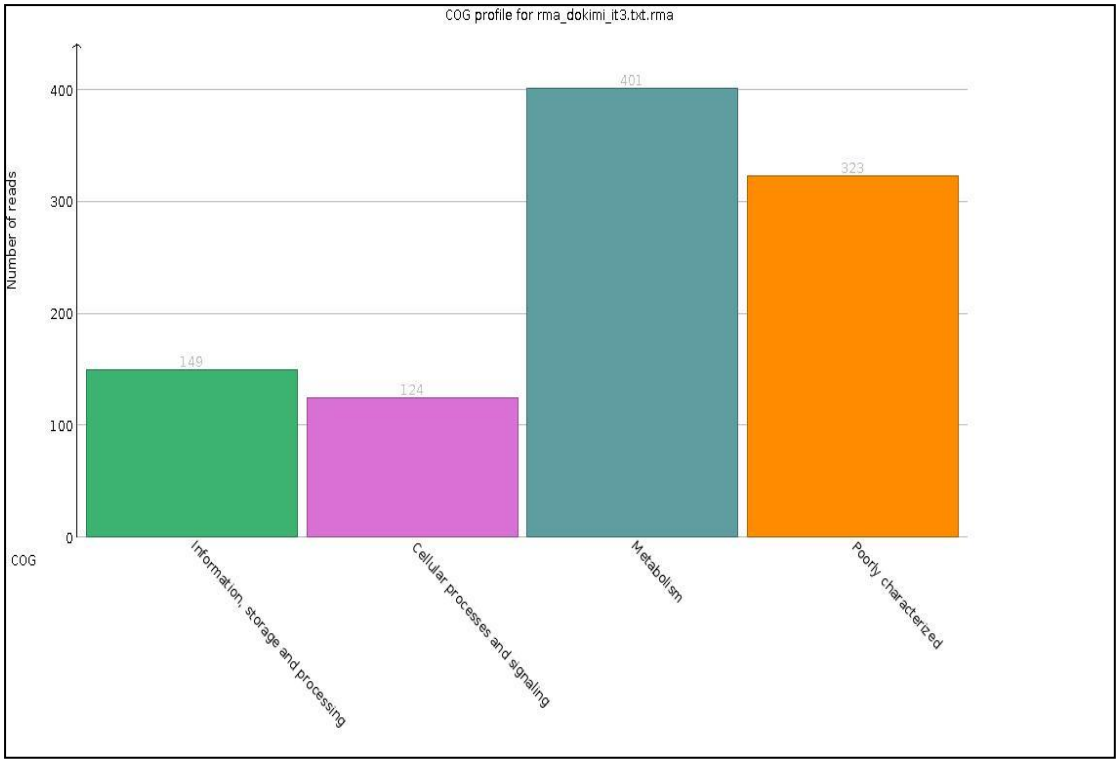
Εικόνα 46 Το αρχείο κειμένου που περιγράφει το ταξινομικό περιεχόμενο του δείγματος it3_clean2000 στο επίπεδο των Species. Στην πρώτη στήλη αναφέρεται η ταξινομική ομάδα, στην δεύτερη ο αριθμός των reads που αντιστοιχίζονται στην κάθε μια και στην τρίτη τα μεταξύ τους ποσοστά. Ο διαχωρισμός των στηλών, γίνεται με χρήση tabs.

Η ιδιαίτερα σημαντική παρουσία του Acidianus hospitalis φαίνεται και από το ποσοστό των reads που συμπεριλαμβάνει σε σχέση με τον συνολικό αριθμό των αλληλουχιών ανάγνωσης, το οποίο φτάνει το 78%. Αξίζει να αναφερθεί και το ποσοστό του Pyrobaculum arsenaticum που φτάνει περίπου στο 10% των συνολικών αλληλουχιών ανάγνωσης που έχουν αντιστοιχιστεί σε κόμβους στο επίπεδο των Species(Εικόνα 46).

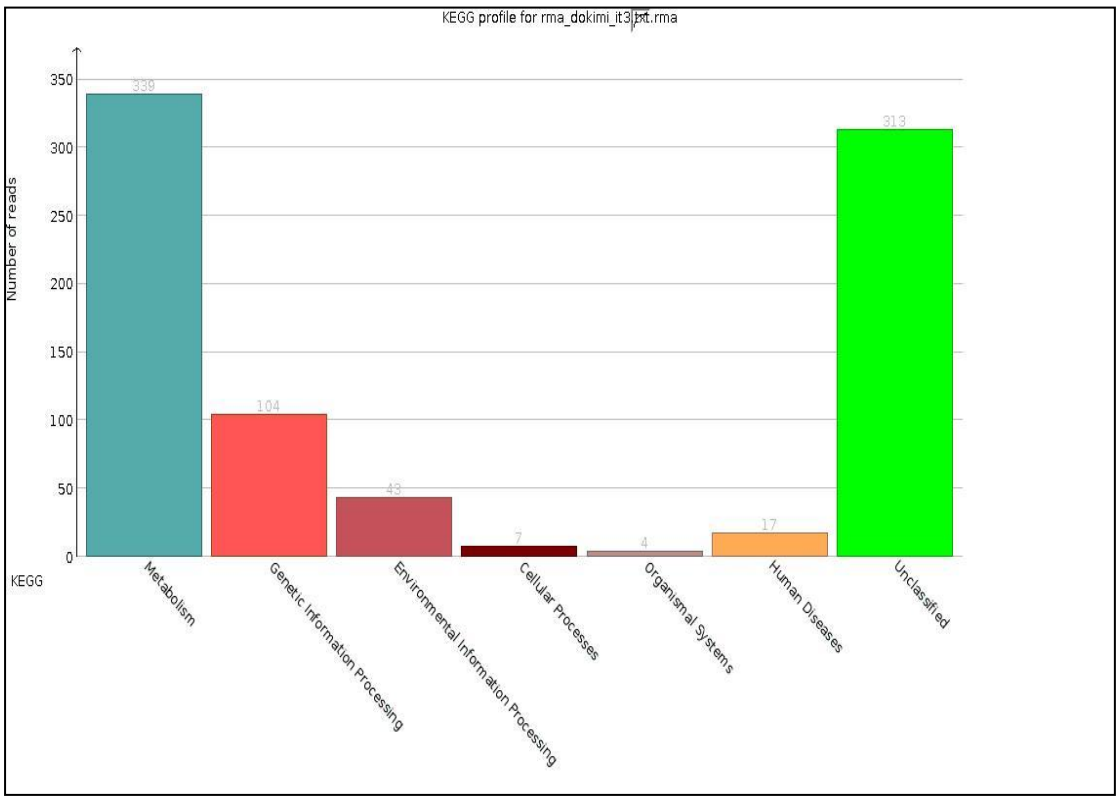
Με τα αποτελέσματα αυτά καθώς και τα αρχεία εξόδου τα οποία δεν είναι δυνατό να παρατεθούν στο δεδομένο κεφάλαιο αποτελεσμάτων λόγω της φύσης και του μεγέθους τους όπως το ταξινομικό δέντρο μέχρι το επίπεδο των Species αλλά και τα αρχεία FASTA με τις αλληλουχίες που αντιστοιχίζονται σε κάθε ταξινομική ομάδα στους κόμβους των Species, ολοκληρώνεται η ταξινομική ανάλυση για το δείγμα it3_clean2000 που επιχειρείται μέσω του εργαλείου MEGAN t-analysis και πλέον απομένουν τα αποτελέσματα της λειτουργικής ανάλυσης μέσω του εργαλείου MEGAN f-analysis. Δημιουργήθηκαν λοιπόν τα τρία διαγράμματα που αντιστοιχούν σε κάθε είδος ανάλυσης για το δείγμα αυτό και παρέχουν μια πρώτη εικόνα για το λειτουργικό περιεχόμενο του δείγματος υπό την σκοπιά της κάθε κατηγοριοποίησης.



Εικόνα 47 Διάγραμμα SEED ανάλυσης για το δείγμα it3_clean2000 στο πρώτο επίπεδο του δέντρου της SEED κατηγοριοποίησης.



Εικόνα 48 Διάγραμμα COG ανάλυσης για το δείγμα it3_clean2000 στο πρώτο επίπεδο του δέντρου της COG κατηγοριοποίησης.



Εικόνα 49 Διάγραμμα KEGG ανάλυσης για το δείγμα it3_clean2000 στο πρώτο επίπεδο του δέντρου της KEGG κατηγοριοποίησης.

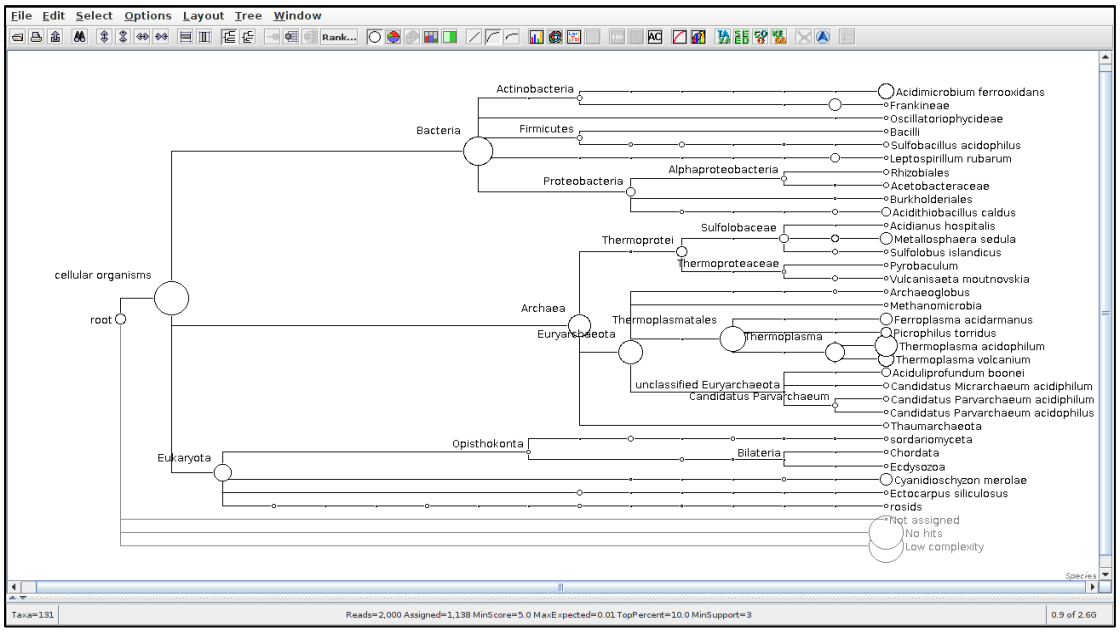
Τα διαγράμματα αυτά αποτελούν μια πρώτη επαφή του χρήστη με το λειτουργικό περιεχόμενο του δείγματος it3_clean2000 υπό την σκοπιά της κάθε ανάλυσης. Προφανώς, λεπτομερής λειτουργική ανάλυση δεν μπορεί να πραγματοποιηθεί μέσω των διαγραμμάτων αυτών όμως ο χρήστης μπορεί να μελετήσει εις βάθος τα λειτουργικά μονοπάτια και υποσυστήματα που προκύπτουν μέσω των λοιπών αρχείων εξόδου του εργαλείου. Στο διάγραμμα της SEED ανάλυσης (Εικόνα 47) παρουσιάζονται οι κόμβοι που αποτελούν λειτουργικά υποσυστήματα για το υψηλότερο επίπεδο της SEED κατηγοριοποίησης (οριζόντιος άξονας) καθώς και ο αριθμός των αλληλουχιών ανάγνωσης που αντιστοιχίζονται σε κάθε ένα από αυτούς τους κόμβους (κάθετος άξονας). Εντοπίζεται έντονη λειτουργική παρουσία σε υποσυστήματα που σχετίζονται κυρίως με υδατάνθρακες, αλλά επιπλέον και με διεργασίες της αναπνοής, του μεταβολισμού πρωτεϊνών, των αμινοξέων και των παραγώγων τους καθώς και μια πληθώρα άλλων λειτουργικών υποσυστημάτων. Αντίστοιχα, στο διάγραμμα της COG ανάλυσης (Εικόνα 48) παρουσιάζονται οι κόμβοι που αντιστοιχούν στο πρώτο επίπεδο της COG κατηγοριοποίησης και αντιπροσωπεύουν ομάδες λειτουργικών διεργασιών όπως αυτές έχουν οριστεί στην COG βάση δεδομένων, καθώς και ο αριθμός των reads που αντιστοιχούν σε κάθε έναν από αυτούς. Για το συγκεκριμένο δείγμα, φαίνεται ιδιαίτερα έντονη η πιθανή παρουσία ομάδων πρωτεϊνών που σχετίζονται με διεργασίες μεταβολισμού και σε μικρότερο ποσοστό με κυτταρικές διεργασίες και σηματοδότηση, καθώς και με διεργασίες αποθήκευσης και επεξεργασίας πληροφοριών. Υπάρχει επίσης αντιστοίχιση σημαντικού ποσοστού των αλληλουχιών ανάγνωσης στην κατηγορία των poorly characterized για τις οποίες δεν μπορεί να προσδιοριστεί με βεβαιότητα και ακρίβεια ο λειτουργικός ρόλος. Τέλος, το διάγραμμα της KEGG ανάλυσης (Εικόνα 49) παρουσιάζει ως κόμβους τις κορυφές των - μεταβολικών ή μη - μονοπατιών στις οποίες συμμετέχουν οι πρωτεΐνες που πιθανώς προέρχονται από τις αλληλουχίες ανάγνωσης του δείγματος καθώς και τον αριθμό των αλληλουχιών ανάγνωσης που αντιστοιχίζονται σε κάθε κόμβο. Ιδιαίτερα έντονη φαίνεται να είναι η παρουσία πρωτεϊνών που σχετίζονται με μεταβολικά μονοπάτια, ενώ εξίσου σημαντικός είναι ο αριθμός των αλληλουχιών που αντιστοιχίστηκαν σε μια κατηγορία που ονομάζεται unclassified και ουσιαστικά υποδεικνύει την ύπαρξη μιας λειτουργικής διεργασίας δίχως όμως να είναι δυνατός ο επακριβής προσδιορισμός της φύσης της.

Επιπροσθέτως, παρέχονται για κάθε είδος ανάλυσης, αρχεία FASTA που περιέχουν τα ονόματα των ακραίων κόμβων του δέντρου κάθε ανάλυσης και τις αλληλουχίες ανάγνωσης που αντιστοιχίζονται σε αυτά καθώς και ένα αρχείο κειμένου στο οποίο σημειώνονται όλα τα μονοπάτια του δέντρου της κατηγοριοποίησης με αρχή των πρώτο (top) κόμβο και κατάληξη τους ακραίους κόμβους, στους οποίους έχει γίνει η αντιστοίχηση, ώστε ο χρήστης να μπορεί να μελετήσει αναλυτικά σε ποια βιολογικά υποσυστήματα, σε ποια μεταβολικά μονοπάτια και σε ποιες λειτουργικές

ομάδες διεργασιών μπορεί να ανήκουν οι προβλεπόμενες πρωτεΐνες που προέρχονται από τις αλληλουχίες ανάγνωσης του δείγματος. Τα αρχεία αυτά δεν είναι δυνατό να παρατεθούν πλήρη για το κάθε δείγμα και παρουσιάζονται μόνο ως παράδειγμα στο προηγούμενο κεφάλαιο.

Δείγμα it2_clean 2000

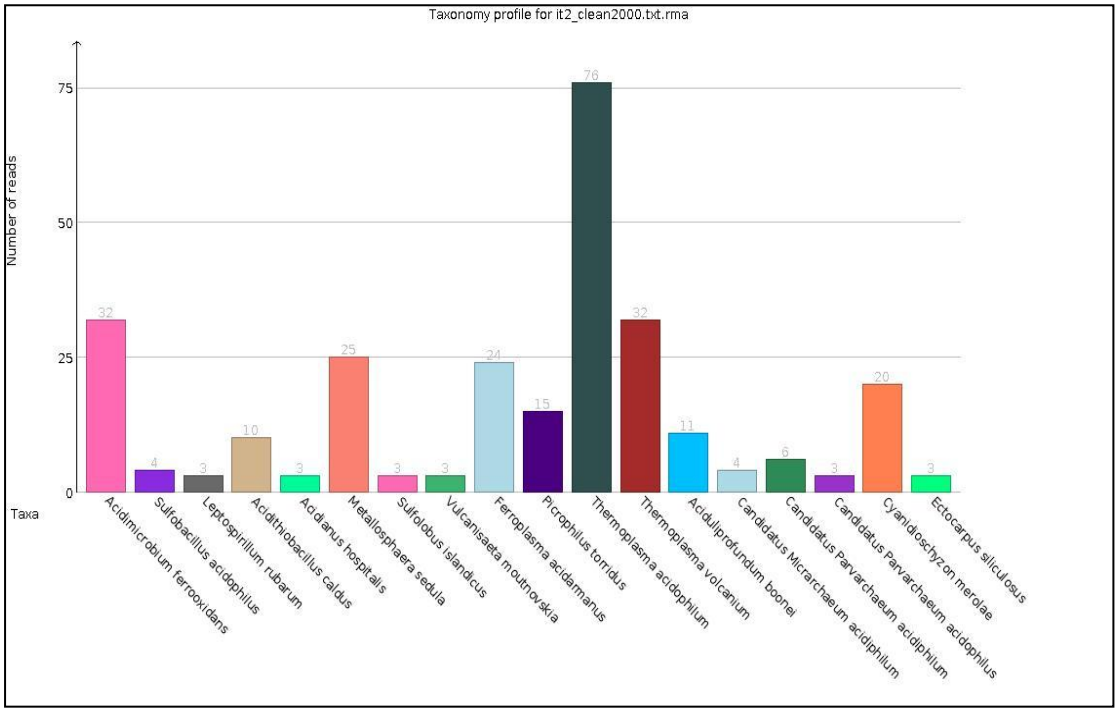
Το δεύτερο δείγμα που μελετήθηκε ήταν το μεταγονιδιωματικό δείγμα **It-2** από το οποίο επιλέχθηκαν 2000 τυχαίες αλληλουχίες και δημιουργήθηκε το προς ανάλυση αρχείο FASTA **it2_clean2000**. Πρόκειται για αλληλουχίες ανάγνωσης που έχουν προκύψει από τη μέθοδο της 454 Titanium πυροαλληλούχισης και στο αρχικό αρχείο περιέχονται 707.206 αλληλουχίες. Το δείγμα προέρχεται από θερμό έδαφος της περιοχής Pisciarelli της Ιταλίας κοντά στις θερμές πηγές από τις οποίες προέρχεται το δείγμα **It-3.** Η θερμοκρασία του εδάφους μετρήθηκε ίση με 49^{o}C. Από τις 2000 αλληλουχίες, οι 1138 αντιστοιχίστηκαν σε κάποιον κόμβο που συμβολίζει μια ταξινομική ομάδα, ανεξαρτήτως ταξινομικού επιπέδου .



Εικόνα 50 Η ταξινομική εικόνα στο επίπεδο των Species όπως παρουσιάζεται στο αρχείο rma για το δείγμα it2_clean2000.

Από τα παρακάτω στοιχεία φαίνεται η έντονη παρουσία του είδους μικροοργανισμού **Thermoplasma acidophilum** σε ένα ποσοστό περίπου 27% επί των συνολικών αλληλουχιών βάσης που αντιστοιχίστηκαν σε ταξινομικές ομάδες στο επίπεδο των Species. Σημαντικό ποσοστό της τάξης του 12% αποτελούν και οι αλληλουχίες που έχουν αντιστοιχιστεί σε ένα άλλο είδος που ανήκει στο γένος Thermoplasma, το **Thermoplasma volcanium**. Τα δυο αυτά ήδη μικροοργανισμών είναι επιβεβαιωμένο πως εντοπίζονται σε περιοχές όπως η περιοχή Pisciarelli[48]. Στο ίδιο ποσοστό κυμαίνεται και η παρουσία του **Acidimicrobium ferrooxidans**,

ενός είδους που απομονώνεται από απορροές θερμών πηγών. Αξιοσημείωτη είναι και η παρουσία του **Metallosphaera sedula**, το οποίο είναι γνωστό ότι έχει απομονωθεί από δείγματα της συγκεκριμένης περιοχής[49], ενώ δεν προκαλεί εντύπωση και η παρουσία του είδους **Ferroplasma acidarmanus**, που απομονώνεται συχνά σε δείγματα από περιοχές θερμών πηγών και οι δεδομένες θερμοκρασιακές συνθήκες επιτρέπουν την ανάπτυξη του. Τέλος, αξίζει να αναφερθεί και η παρουσία ενός ακόμα είδους που επιβεβαιωμένα έχει απομονωθεί από δείγματα της περιοχής, του **Cyanidioschyzon merolae**[50].



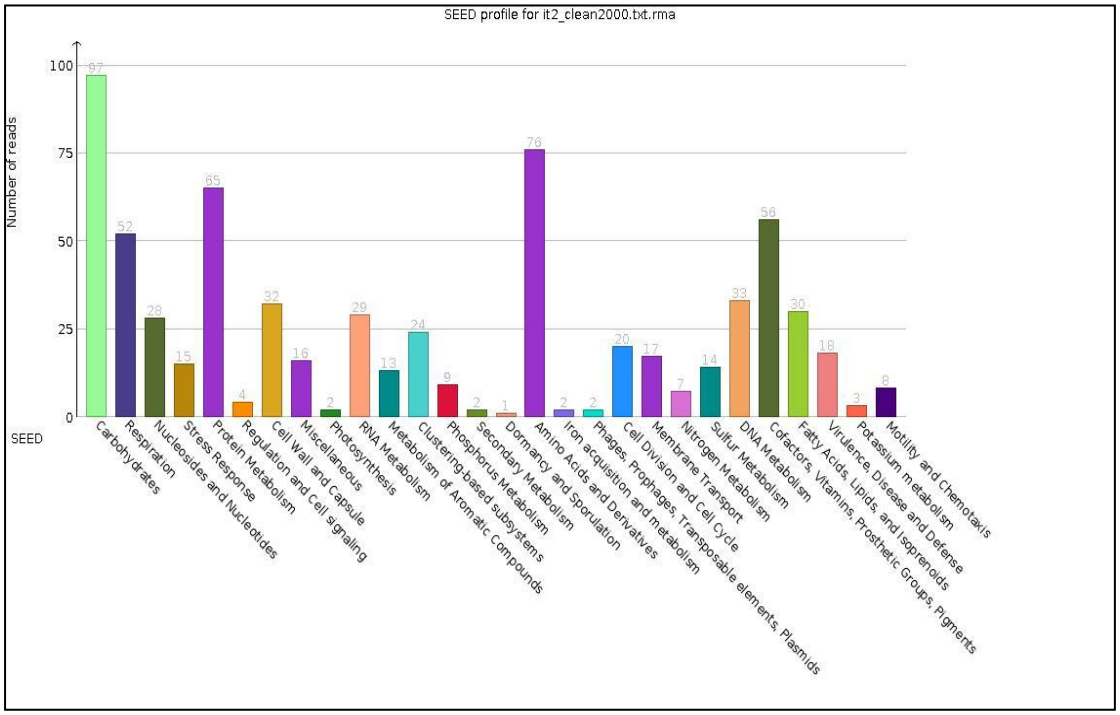
Εικόνα 51 Ταξινομικό προφίλ του δείγματος it2_clean2000 στο επίπεδο των Species.

dataset_789.dat

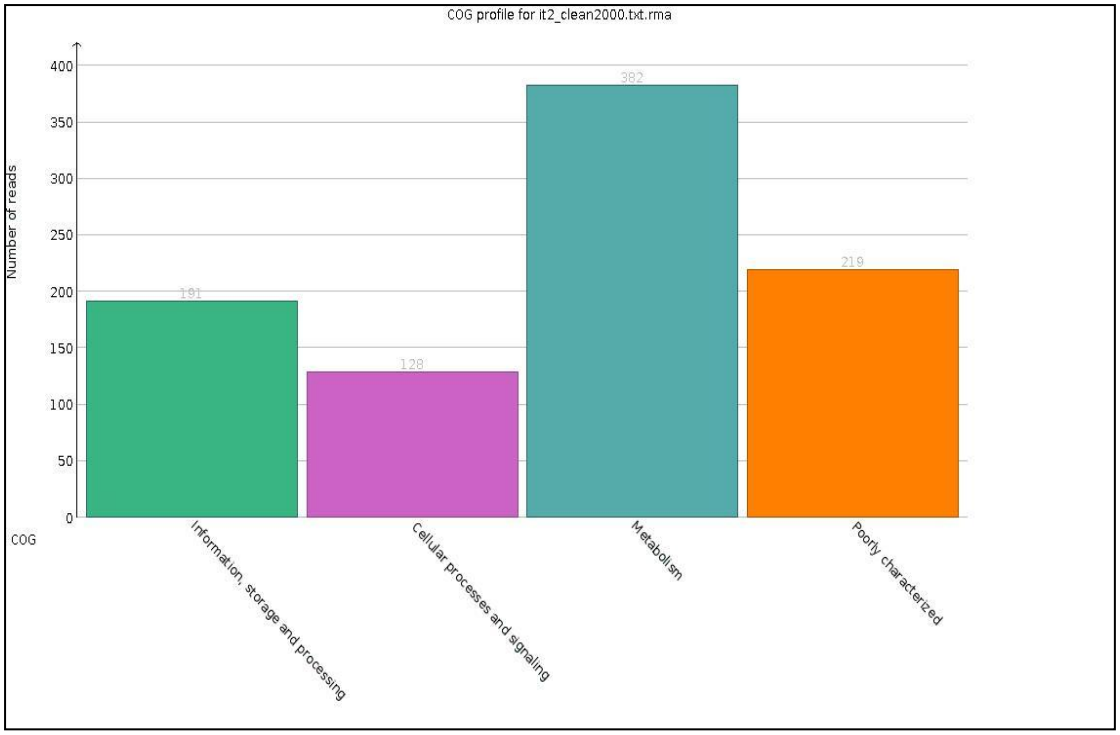
#Series	Reads	Percentages
Acidimicrobium ferrooxidans	32.0	11.5523465704
Sulfobacillus acidophilus	4.0	1.4440433213
Leptospirillum rubarum	3.0	1.08303249097
Acidithiobacillus caldus	10.0	3.61010830325
Acidianus hospitalis	3.0	1.08303249097
Metallosphaera sedula	25.0	9.02527075812
Sulfolobus islandicus	3.0	1.08303249097
Vulcanisaeta moutnovskia	3.0	1.08303249097
Ferroplasma acidarmanus	24.0	8.6642599278
Picrophilus torridus	15.0	5.41516245487
Thermoplasma acidophilum	76.0	27.4368231047
Thermoplasma volcanium	32.0	11.5523465704
Aciduliprofundum boonei	11.0	3.97111913357
Candidatus Micrarchaeum acidiphilum	4.0	1.4440433213
Candidatus Parvarchaeum acidiphilum	6.0	2.16606498195
Candidatus Parvarchaeum acidophilus	3.0	1.08303249097
Cyanidioschyzon merolae	20.0	7.2202166065
Ectocarpus siliculosus	3.0	1.08303249097

Εικόνα 52 Αρχείο κειμένου με την ταξινομική εικόνα του δείγματος it2_clean2000 και τα ποσοστά για κάθε ταξινομική ομάδα για το επίπεδο των Species.

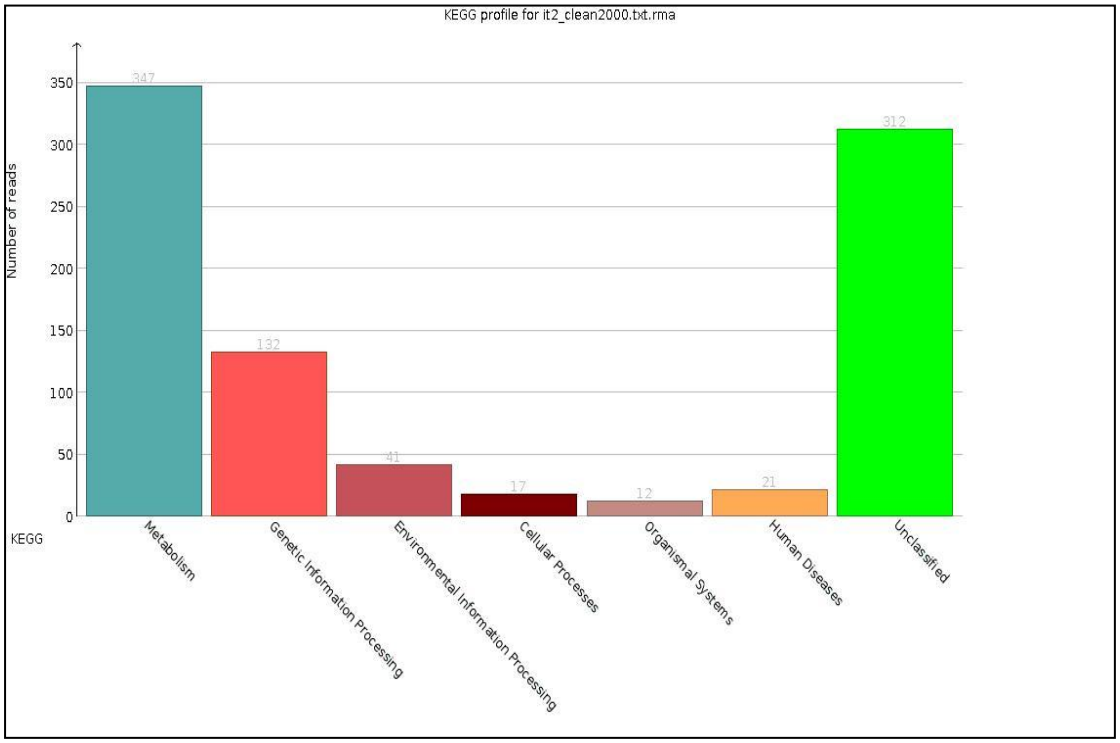
Σε ότι αφορά την λειτουργική ανάλυση που επιχειρήθηκε μέσω του εργαλείου MEGAN f-analysis για το δείγμα it2_clean2000, εφόσον τα αναλυτικά μονοπάτια που φανερώνουν πλήρως τις διασυνδέσεις των λειτουργικών ρόλων στα πλαίσια των υποσυστημάτων ή των μεταβολικών μονοπατιών είναι αδύνατο να παρουσιαστούν λόγω μεγέθους των αντίστοιχων αρχείων, παρουσιάζονται τα διαγράμματα για κάθε είδος ανάλυσης. Στο διάγραμμα της Εικόνας 53, παρουσιάζεται το προφίλ της SEED κατηγοριοποίησης στο υψηλότερο επίπεδο του δέντρου της όπως περιγράφεται και στο δείγμα it3_clean2000 , με τις περισσότερες αλληλουχίες να έχουν αντιστοιχιστεί σε λειτουργικά υποσυστήματα σχετικά με υδατάνθρακες, αμινοξέα και παράγωγα τους, συμπαράγοντες, βιταμίνες, προσθετικές ομάδες, χρωστικές ουσίες αλλά και διεργασίες μεταβολισμού πρωτεϊνών, μεταβολισμού DNA και αναπνοής. Στο διάγραμμα της COG κατηγοριοποίησης (Εικόνα 54) έχει αντιστοιχιστεί ένα μεγάλο μέρος των reads κυρίως στον κόμβο του μεταβολισμού ενώ οι υπόλοιπες αλληλουχίες ανάγνωσης που αντιστοιχίστηκαν σε κάποιο λειτουργικό ρόλο διαμοιράστηκαν στους υπόλοιπους τρεις βασικούς κόμβους του υψηλότερου επιπέδου του δέντρου της COG ανάλυσης. Τέλος, το διάγραμμα της KEGG ανάλυσης (Εικόνα 55) παρουσιάζει γενικά όμοια εικόνα με αυτό της αντίστοιχης ανάλυσης του προηγούμενου δείγματος με τα διάφορα μονοπάτια που τοποθετούνται γενικότερα στις διεργασίες μεταβολισμού να συγκεντρώνουν το μεγαλύτερο ποσοστό των αλληλουχιών ανάγνωσης, ενώ ιδιαίτερα έντονη είναι και η παρουσία unclassified αντιστοιχίσεων.



Εικόνα 53 Διάγραμμα SEED ανάλυσης για το δείγμα it2_clean2000 στο πρώτο επίπεδο του δέντρου της SEED κατηγοριοποίησης.



Εικόνα 54 Διάγραμμα COG ανάλυσης για το δείγμα it2_clean2000 στο πρώτο επίπεδο του δέντρου της COG κατηγοριοποίησης.

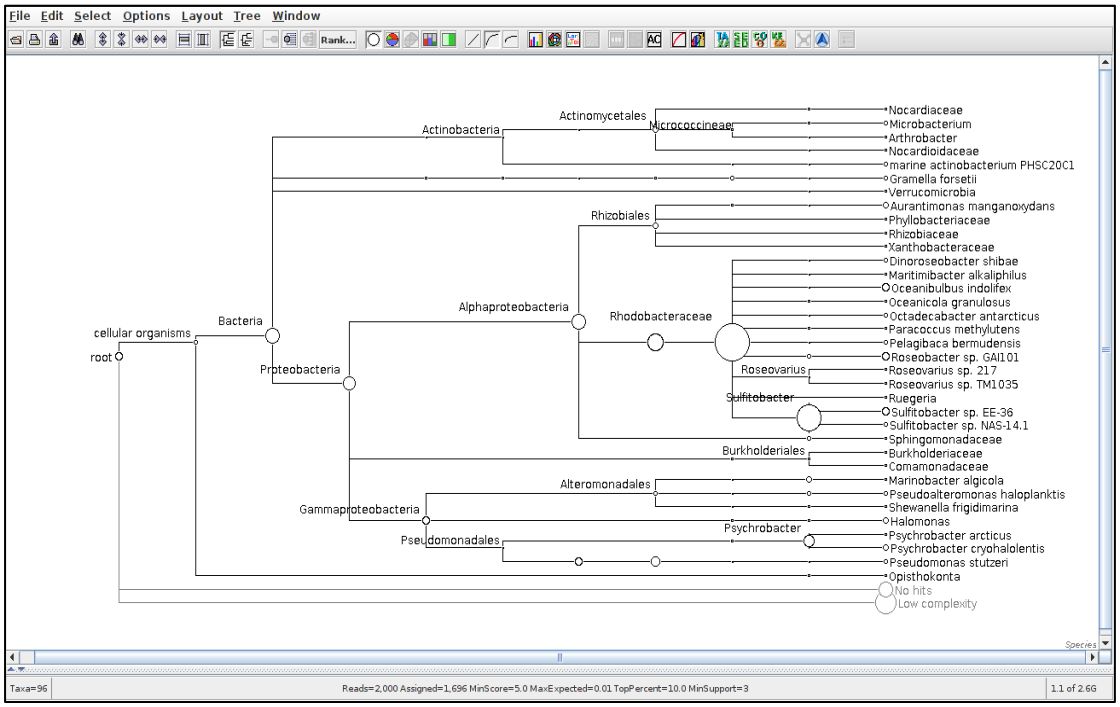


Εικόνα 55 Διάγραμμα KEGG ανάλυσης για το δείγμα it2_clean2000 στο πρώτο επίπεδο του δέντρου της KEGG κατηγοριοποίησης.

Δείγμα ngi7_2000

Το τρίτο δείγμα που μελετήθηκε ήταν το μεταγονιδιωματικό δείγμα **NGI-7** από το οποίο επιλέχθηκαν 2000 τυχαίες αλληλουχίες και δημιουργήθηκε το προς ανάλυση αρχείο FASTA **ngi7_2000**. Πρόκειται για αλληλουχίες ανάγνωσης που έχουν προκύψει από τη μέθοδο της 454 Titanium πυροαλληλούχισης και στο αρχικό αρχείο περιέχονται 600.474 αλληλουχίες. Το δείγμα προέρχεται από μια γεώτρηση σε βάθος 404 μέτρων κάτω από την επιφάνεια της γης, στο νησί Spitsbergen στον Αρκτικό Ωκεανό και η θερμοκρασία του δείγματος μετρήθηκε ίση με 17^{O}C.

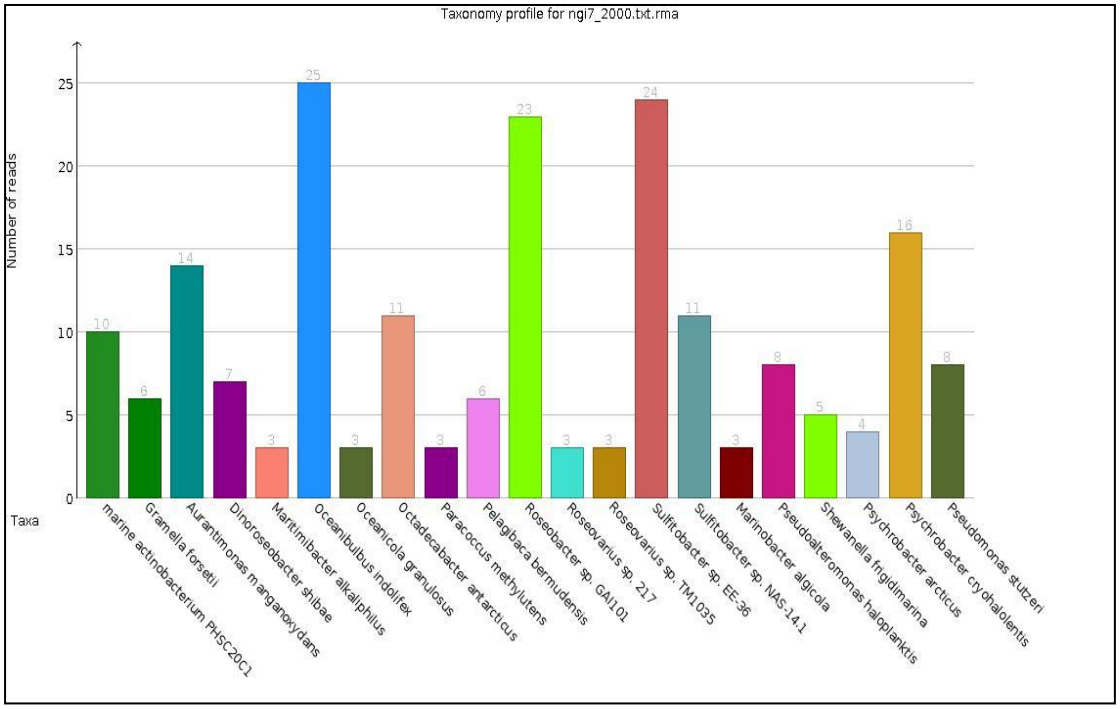
Από το πρώτο εργαλείο προέκυψε το παρακάτω αρχείο RMA, μέσα από το οποίο παρουσιάζεται και η εικόνα του ταξινομικού δέντρου για το δείγμα αυτό στο επίπεδο των Species. Σε σχέση με τα δυο προηγούμενα δείγματα είναι σημαντικά αυξημένος ο αριθμός των αλληλουχιών ανάγνωσης που αντιστοιχίστηκαν σε κάποια ταξινομική ομάδα, ανεξάρτητα από το επίπεδο που εντοπίζεται αυτή στην ταξινομική ιεραρχία. Από τις 2000 αλληλουχίες, οι 1696 αντιστοιχίστηκαν σε κάποιον κόμβο που συμβολίζει μια ταξινομική ομάδα ανεξαρτήτως ταξινομικού επιπέδου, ενώ οι υπόλοιπες ανήκουν στον κόμβους χωρίς χτυπήματα ή στον κόμβο χαμηλής πολυπλοκότητας.



Εικόνα 56 Η ταξινομική εικόνα στο επίπεδο των Species όπως παρουσιάζεται στο αρχείο rma για το δείγμα ngi7_2000. Από το μικρό μέγεθος των κόμβων στο επίπεδο, αλλά και το μεγάλο μέγεθος κόμβων που προηγούνται, φαίνεται πως η πληθώρα των reads, αντιστοιχίστηκε σε υψηλότερο επίπεδο από τα Species.

Σε ότι αφορά την ταξινομική κατηγοριοποίηση των αλληλουχιών ανάγνωσης που απαρτίζουν αυτό το μεταγονιδιωματικό δείγμα δημιουργήθηκε το διάγραμμα του ταξινομικού προφίλ στο επίπεδο των Species, μαζί με το αρχείο κειμένου που

παρέχει τις αντίστοιχες πληροφορίες καθώς και τα ποσοστά της κάθε ταξινομικής ομάδας στο ταξινομικό επίπεδο που γίνεται η ανάλυση.



Εικόνα 57 Διάγραμμα ταξινομικής ανάλυσης στο επίπεδο των Species για το δείγμα ngi7_2000.

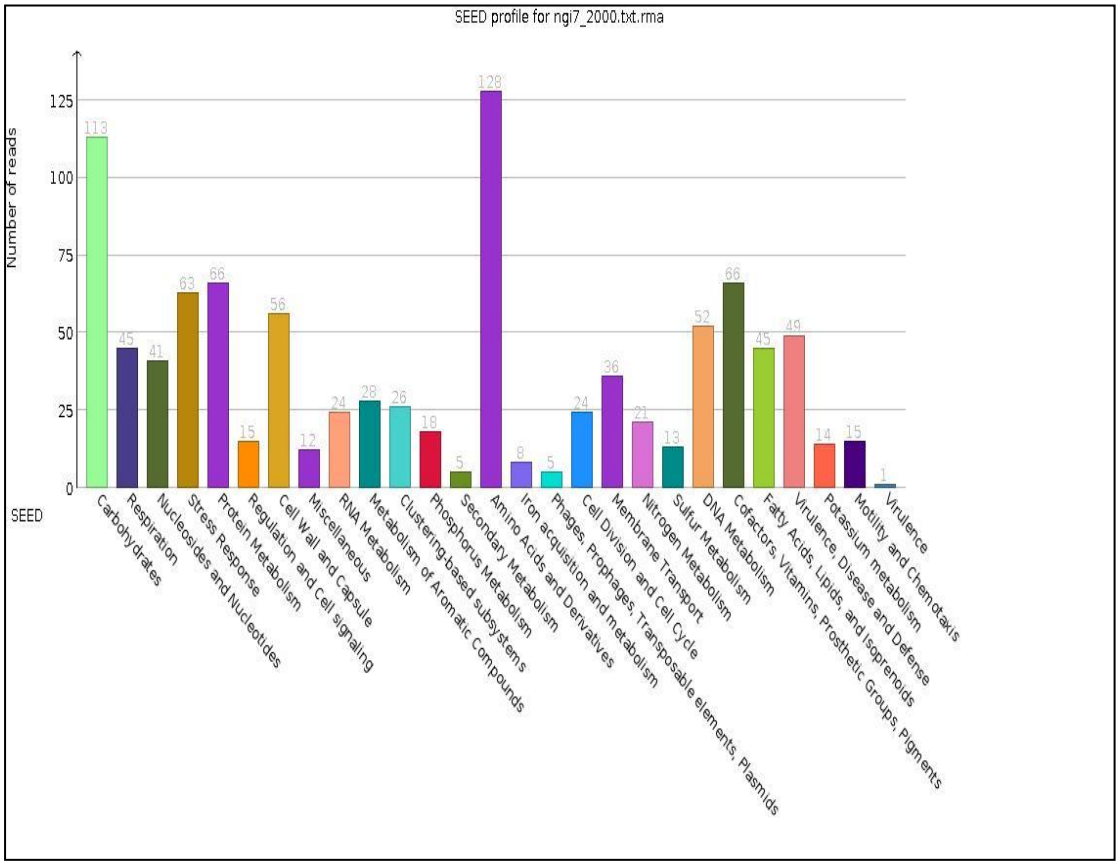
dataset_769.dat

#Series	Reads	Percentages
marine actinobacterium PHSC20C1	10.0	5.10204081633
Gramella forsetii	6.0	3.0612244898
Aurantimonas manganoxydans	14.0	7.14285714286
Dinoroseobacter shibae	7.0	3.57142857143
Maritimibacter alkaliphilus	3.0	1.5306122449
Oceanibulbus indolifex	25.0	12.7551020408
Oceanicola granulosus	3.0	1.5306122449
Octadecabacter antarcticus	11.0	5.61224489796
Paracoccus methylutens	3.0	1.5306122449
Pelagibaca bermudensis	6.0	3.0612244898
Roseobacter sp. GAI101	23.0	11.7346938776
Roseovarius sp. 217	3.0	1.5306122449
Roseovarius sp. TM1035	3.0	1.5306122449
Sulfitobacter sp. EE-36	24.0	12.2448979592
Sulfitobacter sp. NAS-14.1	11.0	5.61224489796
Marinobacter algicola	3.0	1.5306122449
Pseudoalteromonas haloplanktis	8.0	4.08163265306
Shewanella frigidimarina	5.0	2.55102040816
Psychrobacter arcticus	4.0	2.04081632653
Psychrobacter cryohalolentis	16.0	8.16326530612
Pseudomonas stutzeri	8.0	4.08163265306

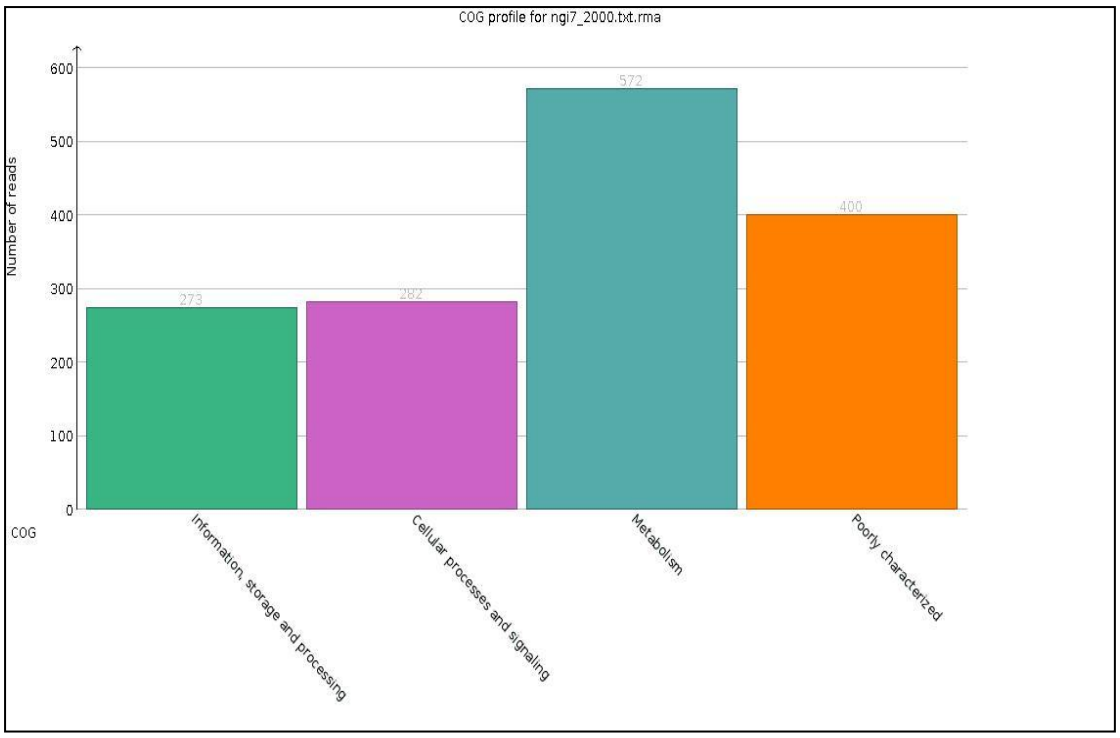
Εικόνα 58 Αρχείο κειμένου με την ταξινομική εικόνα του δείγματος ngi7_2000 και τα ποσοστά για κάθε ταξινομική ομάδα για το επίπεδο των Species.

Από τα παραπάνω δεδομένα γίνεται εύκολα αντιληπτό ότι κατά την ανάλυση έχει προκύψει μια ποικιλία οργανισμών που σχετίζεται με υδατικά περιβάλλοντα. Ξεχωρίζει η παρουσία του **Oceanibulbus indolifex** (13% των αλληλουχιών που έχουν αντιστοιχιστεί), ενός είδους μικροοργανισμού που είναι γνωστό ότι αναπτύσσεται ικανοποιητικά και συχνά σε εντυπωσιακό βαθμό σε βάθη πολύ κάτω από την επιφάνεια της θάλασσας[51]. Στα ίδια ποσοστά κυμαίνεται και η παρουσία δυο άλλων ειδών, του **Roseobacter sp. GAI101** και του **Sulfitobacter sp. EE-36** που αναπτύσσονται σε σχετικά χαμηλές θερμοκρασίες, ενώ αξιοσημείωτη είναι και η παρουσία του είδους **Psychrobacter cryohalolentis**, ενός είδους βακτηρίου που συχνά εντοπίζεται σε περιοχές χαμηλών θερμοκρασιών ή και σε υδατικά περιβάλλοντα μεγάλου βάθους , ενώ στελέχη του γένους **Psychrobacter** έχουν απομονωθεί από την συγκεκριμένη περιοχή από την οποία προέρχεται το δείγμα[52]. Τέλος, αξίζει να αναφερθεί και η παρουσία του μικροοργανισμού **Aurantimonas manganoxydans**, που σύμφωνα και με την ονομασία του παρουσιάζει την ικανότητα να οξειδώνει το μαγνήσιο.

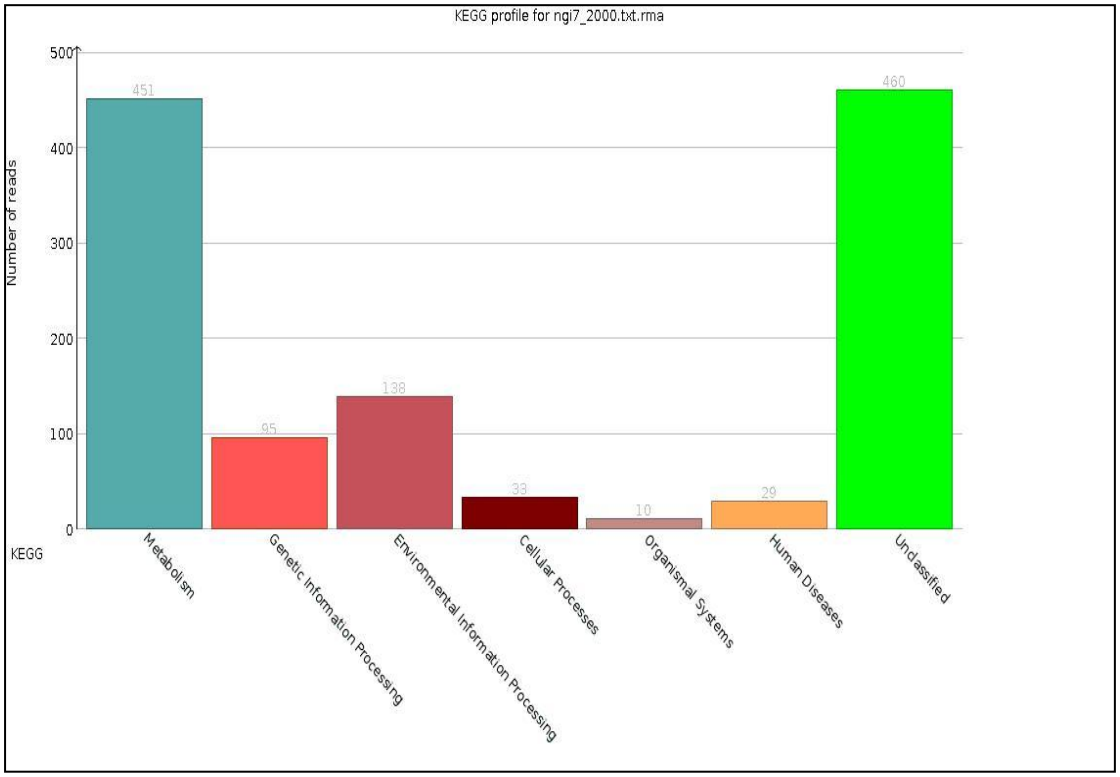
Σε ότι αφορά την λειτουργική ανάλυση που προέκυψε από το εργαλείο MEGAN f-analysis προέκυψαν και παρουσιάζονται τα τρία διαγράμματα που αφορούν το κάθε είδος ανάλυσης που πραγματοποιείται.



Εικόνα 59 Διάγραμμα SEED ανάλυσης για το δείγμα ngi7_2000 στο πρώτο επίπεδο του δέντρου της SEED κατηγοριοποίησης.



Εικόνα 60 Διάγραμμα COG ανάλυσης για το δείγμα ngi7_2000 στο πρώτο επίπεδο του δέντρου της COG κατηγοριοποίησης.



Εικόνα 61 Διάγραμμα KEGG ανάλυσης για το δείγμα στο πρώτο επίπεδο του δέντρου της KEGG κατηγοριοποίησης.

Στην προσπάθεια να αποδοθεί μια πρώτη εικόνα του λειτουργικού περιεχομένου του δείγματος **ngi7_2000**, αρχικά θα πρέπει να τονιστεί το ιδιαίτερα αυξημένο ποσοστό αλληλουχιών που σχετίζονται με την παραγωγή πρωτεϊνών στα πλαίσια υποσυστημάτων όπως αυτά των υδατανθράκων και των αμινοξέων και των παραγώγων τους όπως γίνεται φανερό από το διάγραμμα της SEED κατηγοριοποίησης (Εικόνα 59). Σημαντική είναι και η αντιστοίχιση αλληλουχιών και σε υποσυστήματα όπως του μεταβολισμού πρωτεϊνών, της αντίδρασης στο στρες (Stress Response) καθώς και σε διεργασίες που σχετίζονται με το υποσύστημα του κυτταρικού τοιχώματος και της κάψας (Cell Wall and Capsule). Στο διάγραμμα της COG ανάλυσης (Εικόνα 60), οι αλληλουχίες έχουν διαμοιραστεί στους τέσσερις κόμβους της COG κατηγοριοποίησης με εντονότερη την παρουσία τους στις διεργασίες μεταβολισμού, ενώ μέσω του διαγράμματος της KEGG ανάλυσης για το δείγμα αυτό (Εικόνα 61) διαφαίνεται ένας μεγάλος αριθμός αλληλουχιών που έχουν αντιστοιχιστεί σε απροσδιόριστους ρόλους , πέρα από τους ρόλους που σχετίζονται με μεταβολικά μονοπάτια. Επίσης, σε αντίθεση με τα προηγούμενα δείγματα, ένας σημαντικός αριθμός αλληλουχιών ανάγνωσης έχει αντιστοιχιστεί σε μονοπάτια επεξεργασίας περιβαλλοντικών πληροφοριών (Environmental Information Processing). Για περαιτέρω αναλυτική μελέτη του λειτουργικού προφίλ , απαιτείται η επεξεργασία των υπόλοιπων αρχείων εξόδου του εργαλείου MEGAN f-analysis.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Όπως γίνεται φανερό στο προηγούμενο κεφάλαιο, τα υπολογιστικά εργαλεία δημιουργήθηκαν και εφαρμόστηκαν με επιτυχία σε τρία διαφορετικά μεταγονιδιωματικά δείγματα, ακόμα και αν, στα πλαίσια αυτού του κειμένου είναι εκ των πραγμάτων αδύνατο να παρουσιαστούν με ακρίβεια όλα τα αρχεία εξόδου που δημιουργούνται. Ένα πάρα πολύ μεγάλο ποσοστό της πληροφορίας που παρέχεται δεν ήταν δυνατόν να παρουσιαστεί στα αποτελέσματα της εφαρμογής των εργαλείων υπό μορφή κειμένου, όμως μπορεί να αποδώσει με μεγαλύτερη σαφήνεια το ταξινομικό αλλά - κυρίως - και το λειτουργικό προφίλ, που σχηματίζεται από τις αλληλουχίες ανάγνωσης που μελετήθηκαν.

Σε ταξινομικό επίπεδο μέσω των διαγραμμάτων που παρουσιάστηκαν αλλά και των αντίστοιχων αρχείων κειμένου, αντικατοπτρίζεται η παρουσία αλλά και οι αναλογίες των ταξινομικών ομάδων οργανισμών σε ένα συγκεκριμένο δείγμα, είτε σε ένα συγκεκριμένο επίπεδο ταξινόμησης είτε σε σχέση με ένα πλήρως ξεδιπλωμένο ταξινομικό δέντρο. Ο χρήστης δύναται να αναζητήσει την παρουσία ή την απουσία συγκεκριμένων ταξινομικών ομάδων που παρουσιάζουν ιδιαίτερο ενδιαφέρον για την ανάλυση του, τόσο σε επίπεδο ειδών μικροοργανισμών όσο και σε γενικότερο επίπεδο ταξινόμησης, όπως παραδείγματος χάριν μια συγκεκριμένη οικογένεια, κλάση ή γένος μικροβίων. Αφού λοιπόν αποκτήσει αυτή την πρώτη, έστω και περιορισμένη, εικόνα σε σχέση με το ταξινομικό περιεχόμενο, μπορεί μελετώντας ταυτόχρονα το ταξινομικό δέντρο που δημιουργείται ως αρχείο εξόδου, να εστιάσει στις ταξινομικές ομάδες που επιθυμεί και να λάβει συγκεκριμένες πληροφορίες για αυτές, όπως παραδείγματος χάριν τον αριθμό αλλά και την ίδια την νουκλεοτιδική αλληλουχία των αλληλουχιών ανάγνωσης που αντιστοιχίζονται σε αυτές, μέσω των αρχείων FASTA που δημιουργούνται για κάθε ταξινομική ομάδα. Ο διαχωρισμός αυτός των αλληλουχιών ανάγνωσης σε αρχεία FASTA, με βάση τις ταξινομικές ομάδες που αυτές ανήκουν, διευκολύνει κατά έναν τεράστιο βαθμό την περαιτέρω επεξεργασία των δεδομένων του δείγματος, αφού πλέον έχει απομονωθεί το γενετικό υλικό μιας ταξινομικής ομάδας που ενδιαφέρει και οι επακόλουθες αναλύσεις είναι στοχευμένες, καθώς πλέον δεν πραγματοποιούνται σε ένα τυχαίο σύνολο μεταγονιδιώματος μιας μικροβιακής κοινότητας.

Σε λειτουργικό επίπεδο τα δεδομένα και οι πληροφορίες που παρέχονται από τα διαγράμματα - λόγω του ιδιαίτερα διαδραστικού χαρακτήρα αυτών των αναλύσεων - είναι σχετικά περιορισμένες και απλώς παρέχουν μια γενικότερη εικόνα για το λειτουργικό περιεχόμενο. Ο κύριος όγκος της πληροφορίας, εντοπίζεται στα αναλυτικά μονοπάτια της κάθε προσέγγισης στην λειτουργική ανάλυση (SEED, COG και KEGG), που δημιουργούνται υπό μορφή κειμένου και ανάλογα με το είδος της ανάλυσης παρουσιάζουν μια διαφορετική οπτική γωνία στο λειτουργικό προφίλ και συγκεκριμένα τους λειτουργικούς ρόλους που μπορεί να επιτελούνται από

πρωτεΐνες και πρωτεϊνικές ομάδες που είναι δυνατό να προέρχονται από τα μελετώμενα μεταγονιδιωματικά δεδομένα.

Θα πρέπει κάπου εδώ να τονιστεί πως το MEGAN, ένα από τα κύρια εργαλεία ανάλυσης που σχετίζονται με τα υπολογιστικά εργαλεία που δημιουργήθηκαν, παρέχει μια πληθώρα δυνατοτήτων και αναλύσεων στους χρήστες του που εκ των πραγμάτων δεν είναι δυνατόν να αξιοποιηθεί στα πλαίσια δημιουργίας υπολογιστικών εργαλείων σε πλατφόρμες όπως το GALAXY. Σε ένα βαθμό, ότι κερδίζεται με την δημιουργία άμεσων και εύχρηστων εργαλείων στο GALAXY, ώστε ο χρήστης να μην έρχεται σε άμεση επαφή με εργαλεία όπως το MEGAN, χάνεται σε πλήθος επιλογών και δυνατοτήτων. Στην περίπτωση αυτή, ο χρήστης έχει πάντα την δυνατότητα να αξιοποιήσει πλήρως τις επιλογές του MEGAN μέσω της χρήσης του αρχείου εξόδου του πρώτου κατά σειρά εργαλείου, δηλαδή του αρχείου RMA που δημιουργείται.

Επιπροσθέτως, λόγω του προβλήματος κυρίως της υπολογιστικής δύναμης και δευτερευόντως του υπολογιστικού χρόνου που απαιτείται για το καθοριστικότερο βήμα όλης της αναλυτικής διαδικασίας, δηλαδή την πραγματοποίηση της σύγκρισης και ομοπαράθεσης αλληλουχιών-ερωτημάτων σε σχέση με βάσεις δεδομένων μέσω BLAST, ο περιορισμός που τέθηκε ως προς το μέγεθος των μεταγονιδιωματικών δεδομένων προς μελέτη, δεν είναι δυνατό να οδηγήσει σε αποτελέσματα αξιόπιστα και ενδεικτικά του μεταγονιδιωματικού περιεχομένου για πλήρη μεταγονιδιωματικά δείγματα μεγάλου μεγέθους. Φυσικά, αυτός είναι ένας περιορισμός που εντοπίζεται λόγω των διαθέσιμων υπολογιστικών μηχανημάτων. Στην περίπτωση που υπάρχει η κατάλληλη υπολογιστική δύναμη για τέτοιες μεγάλης κλίμακας αναλύσεις, προφανώς τα εργαλεία αυτά είναι δυνατό να αποδώσουν σε ικανοποιητικό βαθμό, με τις κατάλληλες ρυθμίσεις των παραμέτρων που τα χαρακτηρίζουν, παρέχοντας έτσι δεδομένες πληροφορίες για πλήρη μεταγονιδιωματικά δείγματα και όχι μόνο για περιορισμένα τυχαία υποσύνολα αυτών. Θα πρέπει επίσης να αναφερθούν και κάποιες εναλλακτικές επιλογές σε σχέση με το BLAST που προσφέρουν κατά ένα πολυ σημαντικό βαθμό γρηγορότερες αναλύσεις στο κομμάτι της σύγκρισης αλληλουχιών. Στην περίπτωση του BLASTX θα μπορούσε να χρησιμοποιηθεί η εναλλακτική του συνδυασμού του λογισμικού PAUDA[53] και του αναλυτή Bowtie2 που προσφέρει έως και 10.000 φορές γρηγορότερη ανάλυση, έχοντας όμως πάντα υπ' όψιν τις δεδομένες διαφορές στην ακρίβεια σύγκρισης και ομοπαράθεσης. Επιπροσθέτως, ένα νέο λογισμικό που θα μπορούσε να αντικαταστήσει τις διεργασίες του BLASTX αλλά βρίσκεται αυτή την εποχή σε στάδιο κατασκευής και βελτίωσης είναι το DIAMOND[54] που υπόσχεται έως και 16.000 φορές γρηγορότερη ανάλυση παρόμοιου τύπου με την ανάλυση BLASTX και ελάχιστη απώλεια ευαισθησίας. Στην παρούσα διπλωματική όπως και στην πλειοψηφία επιστημονικών άρθρων και αναλύσεων χρησιμοποιείται το BLAST παρά τους μεγάλους χρόνους και την υπολογιστική δύναμη που απαιτεί. Αποτελεί

ένα από τα πλέον διαδεδομένα και αξιόπιστα εργαλεία σύγκρισης και ομοπαράθεσης αλληλουχιών και για αυτό προτιμήθηκε η παρουσίαση αυτού.

Παρόλα αυτά, ο στόχος για την δημιουργία εύχρηστων υπολογιστικών εργαλείων που θα επιτρέπουν άμεσες αναλύσεις μεταγονιδιωματικών δεδομένων σε ταξινομικό και λειτουργικό επίπεδο επετεύχθη και όπως φαίνεται και στα αποτελέσματα των αναλύσεων των δειγμάτων δοκιμής, η εικόνα που παρουσιάζεται, αν και προφανώς δεν μπορεί να είναι σε καμία περίπτωση ενδεικτική των συνολικών δειγμάτων από τα οποία προέρχονται αυτά, μπορεί να οδηγήσει σε συγκρατημένα αισιόδοξες προβλέψεις, ως προς την λειτουργικότητα και την αξιοπιστία των αναλύσεων. Οι ταξινομικές ομάδες που παρουσιάστηκαν στα αποτελέσματα, μπορούν να συσχετιστούν σε ένα σημαντικό ποσοστό με πραγματικά καταγεγραμμένα δεδομένα μικροοργανισμών που έχουν εντοπιστεί και απομονωθεί στα συγκεκριμένα ή σε παρόμοια έστω περιβάλλοντα. Η εφαρμογή λοιπόν των υπολογιστικών αυτών εργαλείων σε μεγαλύτερης κλίμακας δεδομένα υπό την σκέπη της κατάλληλης υπολογιστικής δύναμης, είναι δυνατό να παρέχει πληροφορίες σχετικά με την παρουσία συγκεκριμένων μικροοργανισμών αλλά και τον λειτουργικό ρόλο πρωτεϊνών που εντοπίζονται σε συγκεκριμένα περιβάλλοντα με δεδομένες συνθήκες θερμοκρασίας, pH, πίεσης, επιπέδων οξυγόνου ή και θρεπτικών υλικών. Τα αποτελέσματα τέτοιων ερευνών μπορεί να αποκαλύψουν εκπληκτικά δεδομένα σε σχέση με τη δυνατότητα ανάπτυξης, συνύπαρξης αλλά και λειτουργίας μικροοργανισμών σε συγκεκριμένες συνθήκες, στοιχεία άγνωστα μέχρι πρότινος, που μπορεί να οδηγήσουν σε ερευνητικό και μετέπειτα σε βιομηχανικό ενδιαφέρον και εφαρμογή.

Τέλος αξίζει να τονιστεί πως μια επιπλέον λειτουργία του MEGAN που θα μπορούσε να βρει εφαρμογή μέσω της δημιουργίας τέτοιων υπολογιστικών εργαλείων θα ήταν η άμεση σύγκριση πολλαπλών δειγμάτων τόσο σε ταξινομικό όσο και σε λειτουργικό επίπεδο. Η πραγματοποίηση μιας τέτοιας διαδικασίας θα εμφάνιζε ιδιαίτερο ερευνητικό αλλά και πρακτικό ενδιαφέρον, όμως βασικά προβλήματα που εντοπίζονται στη λειτουργία της πιο πρόσφατης έκδοσης του MEGAN μέσω γραμμής εντολών και σχετίζονται με την σύγκριση δειγμάτων, ήταν αποτρεπτικά για την επιχείρηση μιας τέτοιας προσέγγισης, καθώς δεν ήταν δυνατό να εγγυηθεί η ορθή και αξιόπιστη ανάλυση αυτών. Σε κάθε περίπτωση, όταν οι συνθήκες το επιτρέψουν, μια τέτοια προσπάθεια, θα αποτελούσε σημαντικό συμπλήρωμα της παρούσας προσέγγισης για την ταξινομική και λειτουργική ανάλυση μεταγονιδιωματικών δεδομένων.

ΠΑΡΑΡΤΗΜΑ

blastsub.py

```
#!/usr/bin/python
import sys
import subprocess
import random
import string
import os

unique_id= "".join(random.sample(string.letters, 12))

print 'Number of arguments:', len(sys.argv), 'arguments.'
print 'Argument List:', str(sys.argv)


blast_type = sys.argv[1] #choose type of BLAST search. BLASTX
or BLASTN

if blast_type == "blastx":


 db = sys.argv[2] #Database choice for the BLAST search
 query = sys.argv[3] #FASTA file of nucleotide sequences.
Input data
 evalue = sys.argv[4] #BLAST parameter - Expect value (E) for
saving hits
 meganfile = sys.argv[5] #MEGAN .rma file. Output data
 word_size = sys.argv[6] #BLAST parameter - Length of initial
exact match
 strand = sys.argv[7] #BLAST parameter - Query strand(s) to
search against database/subject. Choice of both, minus, or
plus
 num_threads = sys.argv[8] #BLAST parameter - Number of
threads (CPUs) to use in blast search. Available only if
remote option is disabled.
 useseed = sys.argv[9] #MEGAN parameter - Enable SEED analysis
 usecog = sys.argv[10] #MEGAN parameter - Enable COG analysis
 usekegg = sys.argv[11] #MEGAN parameter - Enable KEGG
analysis
 maxmatches = sys.argv[12] #MEGAN parameter - The Max number
of matches per read item specifies how many matches per read
to save in the RMA file


 flag_remote = False #BLAST parameter - Execute search on NCBI
servers
 flag_ung = False    #BLAST parameter - Perform ungapped
alignment
 for grp in (sys.argv[1:]):
     if 'remote' == grp:
         flag_remote = True
     if 'ungapped' == grp:
```

```
      flag_ung = True


 if (flag_remote): #if remote option is enabled
      print "remote on"
      if (flag_ung): #if ungapped option is enabled
        print "ungapped on"
        subprocess.call('/home/panagiout/ncbi-blast-
2.2.29+/bin/blastx -query %s -out %s_blastfile -db %s -remote
-evalue %s -word_size %s -ungapped -strand %s '
%(query,unique_id,db,evalue,word_size,strand), shell=True)
#call BLAST in order to perform BLASTX search, ungapped on
      else: #if ungapped option is disabled
        print "ungapped off"
        subprocess.call('/home/panagiout/ncbi-blast-
2.2.29+/bin/blastx -query %s -out %s_blastfile -db %s -remote
-evalue %s -word_size %s -strand %s'
%(query,unique_id,db,evalue,word_size,strand), shell=True)
#call BLAST in order to perform BLASTX search, ungapped off

 else: #if remote option is enabled
      print "remote off"
      if (flag_ung): #if ungapped option is enabled
        print "ungapped on"
        subprocess.call('/home/panagiout/ncbi-blast-
2.2.29+/bin/blastx -query %s -out %s_blastfile -db %s -evalue
%s -word_size %s -ungapped -strand %s -num_threads %s'
%(query,unique_id,db,evalue,word_size,strand,num_threads),
shell=True) #call BLASTX,  ungapped on
      else: #if ungapped option is disabled
        print "ungapped off"
        subprocess.call('/home/panagiout/ncbi-blast-
2.2.29+/bin/blastx -query %s -out %s_blastfile -db %s -evalue
%s -word_size %s -strand %s -num_threads %s '
%(query,unique_id,db,evalue,word_size,strand,num_threads),
shell=True) #call BLASTX,  ungapped off

 textfile=open('/tmp/%s_commandsall.txt' % (unique_id),'w')
 textfile.write("load keggRefSeqFile='ref2kegg.map'\nload
seedRefSeqFile='ref2seed.map'\nload
cogRefSeqFile='ref2cog.map'\nimport blastFile=" + (unique_id)
+'_blastfile' " fastaFile=" + (query) + " meganFile=" +
(meganfile) + " maxMatches="+ (maxmatches) +"
useSeed="+(useseed)+" useCOG="+(usecog)+"
useKegg="+(usekegg)+"
mapping='Taxonomy:BUILT_IN=true,SEED:REFSEQ_MAP=true,KEGG:REFS
EQ_MAP=true,COG:REFSEQ_MAP=true'\nquit;\n")
 textfile.close() #create a command file of MEGAN commands

 subprocess.call ('xvfb-run --auto-servernum --server-num=1
/home/panagiout/megan5/./MEGAN -g -E <
/tmp/'+unique_id+'_commandsall.txt ',  shell=True ) #call
MEGAN and read commands from the file created

 os.remove ('/tmp/%s_commandsall.txt' % (unique_id)) # remove
the file of MEGAN commands from tmp
```

```
elif blast_type == "blastn":  #the following script lines
describe the same procedures as the above considering BLASTN

 db = sys.argv[2]
 query = sys.argv[3]
 evalue = sys.argv[4]
 meganfile = sys.argv[5]
 word_size = sys.argv[6]
 strand = sys.argv[7]
 num_threads = sys.argv[8]
 maxmatches = sys.argv[9]

 flag_remote = False
 flag_ung = False
 for grp in (sys.argv[1:]):
     if 'remote' == grp:
         flag_remote = True
     if 'ungapped' == grp:
      flag_ung = True

 if (flag_remote):
      print "remote on"
      if (flag_ung):
        print "ungapped on"
        subprocess.call('/home/panagiout/ncbi-blast-
2.2.29+/bin/blastn -query %s -out %s_blastfile -db %s -remote
-evalue %s -word_size %s -ungapped -strand %s '
%(query,unique_id,db,evalue,word_size,strand), shell=True)
      else:
        print "ungapped off"
        subprocess.call('/home/panagiout/ncbi-blast-
2.2.29+/bin/blastn -query %s -out %s_blastfile -db %s -remote
-evalue %s -word_size %s -strand %s'
%(query,unique_id,db,evalue,word_size,strand), shell=True)

 else:
      print "remote off"
      if (flag_ung):
        print "ungapped on"
        subprocess.call('/home/panagiout/ncbi-blast-
2.2.29+/bin/blastn -query %s -out %s_blastfile -db %s -evalue
%s -word_size %s -ungapped -strand %s -num_threads %s'
%(query,unique_id,db,evalue,word_size,strand,num_threads),
shell=True)
      else:
        print "ungapped off"
        subprocess.call('/home/panagiout/ncbi-blast-
2.2.29+/bin/blastn -query %s -out %s_blastfile -db %s -evalue
%s -word_size %s -strand %s -num_threads %s '
%(query,unique_id,db,evalue,word_size,strand,num_threads),
shell=True)

 textfile=open('/tmp/%s_commandsall.txt' % (unique_id),'w')
```

```
 textfile.write("import blastFile=" + (unique_id)
+'_blastfile' " fastaFile=" + (query) + " meganFile=" +
(meganfile) + " maxMatches="+ (maxmatches) +";\nquit;\n")
 textfile.close()

 subprocess.call ('xvfb-run --auto-servernum --server-num=1
/home/panagiout/megan5/./MEGAN -g -E <
/tmp/'+unique_id+'_commandsall.txt ',  shell=True )

 os.remove ('/tmp/%s_commandsall.txt' % (unique_id))

else:
 print "blast_type error. blastx or blastn"
```

rma_builder.xml

```
<tool id="rma_builder" name="RMA builder">
 <description> Run a BLAST search and produce a MEGAN .rma
file</description>
  <command interpreter="python">blastsub.py
$blast_type_option.blast_type $blast_type_option.db
$blast_type_option.query $blast_type_option.evalue $meganfile
  #if $blast_type_option.blast_type == "blastx":
   #if $blast_type_option.more_options.opts_selector ==
"advanced":
    ${blast_type_option.more_options.word_size}
    ${blast_type_option.more_options.strand}
    ${blast_type_option.more_options.num_threads}
    ${blast_type_option.more_options.useseed}
    ${blast_type_option.more_options.usecog}
    ${blast_type_option.more_options.usekegg}
    ${blast_type_option.more_options.maxmatches}
    ${blast_type_option.more_options.remote}
    ${blast_type_option.more_options.ungapped}
   #else:
    3
    both
    1
    true
    true
    true
    100
   #end if
   2> /home/panagiout/warnings.txt
  #else:
   #if $blast_type_option.more_options.opts_selector ==
"advanced":
    ${blast_type_option.more_options.word_size}
    ${blast_type_option.more_options.strand}
    ${blast_type_option.more_options.num_threads}
    ${blast_type_option.more_options.maxmatches}
    ${blast_type_option.more_options.remote}
    ${blast_type_option.more_options.ungapped}
```

```
   #else:
    11
    both
    1
    100
   #end if
   2> /home/panagiout/warnings.txt
  #end if
  </command>

  <inputs>
   <conditional name="blast_type_option">
    <param name="blast_type" type="select" label="Select BLAST
search type">
      <option value="blastx">BLASTX</option>
      <option value="blastn">BLASTN</option>
    </param>
    <when value="blastx">
     <param name="db" type="select" label="BLAST database
name">
      <option value="/hotzyme/ncbi/nr">NCBI nr (local)
</option>
      <option value="nr">NCBI nr </option>
     </param>
     <param format="fasta" name="query" type="data"
label="Fasta file of reads"/>
     <param name="evalue" type="float" value="10.0"
label="Expect value (E) for saving hits"/>
      <conditional name="more_options">
       <param name="opts_selector" type="select"
label="Advanced options">
        <option value="standard">Standard options</option>
        <option value="advanced">Advanced options</option>
       </param>
       <when value="standard" />
       <when value="advanced">
       <param name="word_size" type="integer" value="3"
label=" Word_size - Length of initial exact match"/>
       <param name="strand" type="select" label="Query
strand(s) to search against database/subject. Choice of both,
minus, or plus">
        <option value="both">both</option>
        <option value="plus">plus</option>
        <option value="minus">minus</option>
       </param>
       <param name="num_threads" type="integer" value="1"
label="Number of threads (CPUs) to use in blast search"/>
       <param name="maxmatches" type="integer" value="100"
label="Max number of matches per read"/>
       <param name="remote" type="select" label=" remote -
Execute search on NCBI servers">
        <option value="">No</option>
        <option value="remote">Yes</option>
       </param>
       <param name="ungapped" type="select" label="Perform
ungapped alignment">
```

```
        <option value="">No</option>
        <option value="ungapped">Yes</option>
       </param>
       <param name="useseed" type="select" label="Enable SEED
analysis">
        <option value="true">Yes</option>
        <option value="false">No</option>
       </param>
       <param name="usecog" type="select" label="Enable COG
analysis">
        <option value="true">Yes</option>
        <option value="false">No</option>
       </param>
       <param name="usekegg" type="select" label="Enable KEGG
analysis">
        <option value="true">Yes</option>
        <option value="false">No</option>
       </param>
       </when>
      </conditional>
    </when>
    <when value="blastn">
     <param name="db" type="select" label="BLAST database
name">
      <option value="/hotzyme/ncbi/nt">NCBI nt (local)
</option>
      <option value="nt">NCBI nt</option>
     </param>
     <param format="fasta" name="query" type="data"
label="Fasta file of reads"/>
     <param name="evalue" type="float" value="10.0"
label="Expect value (E) for saving hits"/>
      <conditional name="more_options">
       <param name="opts_selector" type="select"
label="Advanced options">
        <option value="standard">Standard options</option>
        <option value="advanced">Advanced options</option>
       </param>
       <when value="standard" />
       <when value="advanced">
       <param name="word_size" type="integer" value="11"
label=" Word_size - Length of initial exact match"/>
       <param name="strand" type="select" label="Query
strand(s) to search against database/subject. Choice of both,
minus, or plus">
        <option value="both">both</option>
        <option value="plus">plus</option>
        <option value="minus">minus</option>
       </param>
       <param name="num_threads" type="integer" value="1"
label="Number of threads (CPUs) to use in blast search"/>
       <param name="maxmatches" type="integer" value="100"
label="Max number of matches per read"/>
       <param name="remote" type="select" label=" remote -
Execute search on NCBI servers">
        <option value="">No</option>
```

```
        <option value="remote">Yes</option>
       </param>
       <param name="ungapped" type="select" label="Perform
ungapped alignment">
        <option value="">No</option>
        <option value="ungapped">Yes</option>
       </param>
       </when>
      </conditional>
    </when>
   </conditional>
  </inputs>
  <outputs>
  <data format="txt" name="meganfile"/>
  </outputs>
  <help>

**RMA builder Overview**

The RMA builder tool performs a BLASTX or BLASTN search,
against local or remote databases and uses the results in
order to produce a MEGAN rma file

-----

**Input formats**

This tool uses input FASTA files of reads

-----

**Outputs**

The tool provides a txt file carrying all the information of a
MEGAN rma file. In order to open this file using MEGAN, add
the extension .rma

-----

.. class:: warningmark

**WARNING: Number of threads option is valid only for a local
search (β€□remoteβ€□ option not used).**

.. class:: infomark

**Info:**
Remote option is disabled by default. Enable the option in
order to perform a remote search to the NCBI servers

.. class:: infomark

**Info:**
SEED, KEGG and COG analysis are enabled by default

.. class:: infomark
```

```
**Info:**
If functional analysis is desired in a following step, BLASTX
search is the only appropriate option


  </help>
 </tool>
```

megansub.py

```
#!/usr/bin/python
import sys
import subprocess
import string
import os
import csv
import tarfile
import random
import shutil

unique_id= "".join(random.sample(string.letters, 12))

print 'Number of arguments:', len(sys.argv), 'arguments.'
print 'Argument List:', str(sys.argv)

meganfile= sys.argv[1] # tool input
rank= sys.argv[2] # taxonomic rank of analysis
chartdata = sys.argv[3] # chart output
txtdata = sys.argv[4] # txt output
tardata = sys.argv[5] # tar.gz output
pick_format = sys.argv[6] # format of the taxonomic tree image
minscore =sys.argv[7] # MEGAN parameter
maxexpected = sys.argv[8] # MEGAN parameter
toppercent = sys.argv[9] # MEGAN parameter
minsupport = sys.argv[10] # MEGAN parameter
mincomplexity = sys.argv[11] # MEGAN parameter
use_minimal_coverage_heuristic = sys.argv[12] # MEGAN
parameter
use_identity_filter = sys.argv[13] # MEGAN parameter
minsupportpercent = sys.argv[14] # MEGAN parameter
lcapercent = sys.argv[15] # MEGAN parameter
paired_reads = sys.argv[16] # MEGAN parameter

a=os.path.realpath(meganfile)
b=a+'.rma'
subprocess.call('cp %s %s' % (a,b),  shell=True)
meganfile=b # add the .rma extension in order to read the
input file

flag_all = False
for grp in (sys.argv[1:]):
    if 'all' == grp:
```

```
        flag_all = True

if (flag_all): #  all ranks are selected

 newpath = ('/tmp/%s_Outputs' % (unique_id))
 os.makedirs(newpath) #create a directory for output files

 textfile=open('/tmp/%s_commandsall.txt' % (unique_id),'w')
 textfile.write("open file=" + (meganfile) + ";\nrecompute
minSupportPercent="+(minsupportpercent)+"
minSupport="+(minsupport)+" minScore="+(minscore)+"
maxExpected="+(maxexpected)+" topPercent="+(toppercent)+"
lcaPercent="+(lcapercent)+" minComplexity="+(mincomplexity)+"
useMinimalCoverageHeuristic="+(use_minimal_coverage_heuristic)
+" pairedReads="+(paired_reads)+"
useIdentityFilter="+(use_identity_filter)+" ;\n collapse
level=100;\n select nodes=all;\n extract what=reads
outDir=/tmp/"+(unique_id)+"_Outputs outFile=reads-%t.fasta
data=Taxonomy ids=SELECTED allBelow=false;\n select
nodes=none;\n nodeLabels assigned=true;\n show
intermediate=true;\n set scaleBy=Assigned;\n zoom full;\n
expand direction=vertical;\n expand direction=vertical;\n
expand direction=horizontal;\n expand direction=horizontal;\n
expand direction=horizontal;\n expand direction=horizontal;\n
expand direction=horizontal;\n exportImage file=/tmp/"
+(unique_id)+ "_tree format=" +(pick_format)+ "
replace=true;\n select nodes=all;\n show chart
data=taxonomy;\n set context=TaxaChart;\n  show values=true;\n
export what=chartData file=" +(txtdata)+ ";\n exportImage
file= "+(chartdata)+ " format=jpg replace=true;\n quit;\n " )
 textfile.close() # create a command file for MEGAN


 subprocess.call ('xvfb-run --auto-servernum --server-num=1
/home/panagiout/megan5/./MEGAN -g -E <
/tmp/'+unique_id+'_commandsall.txt ',  shell=True) # call
MEGAN

 os.remove ('/tmp/%s_commandsall.txt' % (unique_id)) #remove
command file

 tar = tarfile.open("%s" % (tardata), "w:gz")
 for name in ["/tmp/%s_tree" % (unique_id),  "/tmp/%s_Outputs"
% (unique_id)]:
    tar.add(name)
 tar.close() # create the tar.gz file

 os.remove ("/tmp/%s_tree" % (unique_id)) #remove tmp tree
file
 shutil.rmtree("/tmp/%s_Outputs" % (unique_id)) #remove tmp
Outputs directory

else: # a speficic rank is selected

 newpath = ('/tmp/%s_Outputs' % (unique_id) )
 os.makedirs(newpath) #create a directory for output files
```

```
 textfile=open('/tmp/%s_commandsrank.txt' % (unique_id),'w')
 textfile.write("open file=" + (meganfile) + ";\nrecompute
minSupportPercent="+(minsupportpercent)+"
minSupport="+(minsupport)+" minScore="+(minscore)+"
maxExpected="+(maxexpected)+" topPercent="+(toppercent)+"
lcaPercent="+(lcapercent)+" minComplexity="+(mincomplexity)+"
useMinimalCoverageHeuristic="+(use_minimal_coverage_heuristic)
+" pairedReads="+(paired_reads)+"
useIdentityFilter="+(use_identity_filter)+" ;\n\n update;\n
collapse rank=" + (rank) + ";\n nodeLabels assigned=true;\n
show intermediate=true;\n set scaleBy=Assigned;\n zoom full;\n
expand direction=vertical;\n expand direction=vertical;\n
expand direction=horizontal;\n expand direction=horizontal;\n
expand direction=horizontal;\n expand direction=horizontal;\n
expand direction=horizontal;\n exportImage file=/tmp/"+
(unique_id) +"_tree format="+(pick_format)+" replace=true;\n
select rank=" + (rank) + ";\n extract what=reads
outDir=/tmp/"+(unique_id)+"_Outputs outFile=reads-%t.fasta
data=Taxonomy ids=SELECTED allBelow=true;\n show chart
data=taxonomy;\n set context=TaxaChart;\n show values=true;\n
export what=chartData file=/tmp/"+ (unique_id) +".txt;\n
exportImage file="+(chartdata)+" format=jpg replace=true;\n
quit;\n " )
 textfile.close() # create a command file for MEGAN

 subprocess.call ('xvfb-run --auto-servernum --server-num=1
/home/panagiout/megan5/./MEGAN -g -E <
/tmp/'+unique_id+'_commandsrank.txt',  shell=True) # call
MEGAN

 os.remove ('/tmp/%s_commandsrank.txt' % (unique_id)) #remove
command file

 found=-1
 names = []
 values = []

 with open('/tmp/%s.txt' % (unique_id), 'r') as csv:

    for line in csv.readlines():
        if found==-1:
          elements = line.strip().split('\t')
          found=1
        else:
          elements = line.strip().split('\t')
          values.append(float(elements[1]))
          names.append(elements[0])
 csum = sum(values)

 percentage = []
 for x in values:
   percentage.append((x/csum)*100) # calculate percentages


 file=open('%s' % (txtdata),'w')
```

```
 file.write(" #Series " +"\t"+ "Reads" + "\t" + "Percentages"
+"\n")
 for i,j,k in zip (names,values,percentage):
  file.write(str(i) + "\t" +str(j)+ "\t" +str(k) +"\n")
 file.close() # create a new txt file containing percentages

 os.remove('/tmp/%s.txt' % (unique_id)) # remove tmp txt file

 tar = tarfile.open("%s" % (tardata), "w:gz")
 for name in ["/tmp/%s_tree" % (unique_id),  "/tmp/%s_Outputs"
% (unique_id)]:
    tar.add(name)
 tar.close() #create the tar.gz

 os.remove ("/tmp/%s_tree" % (unique_id)) # remove tmp tree
file
 shutil.rmtree("/tmp/%s_Outputs" % (unique_id)) # remove tmp
Outputs directory


os.remove(b)
```

megan_analysis.xml

```
<tool id="megan_analysis" name="MEGAN t-analysis">
 <description>for taxonomic classification</description>
  <command interpreter="python">megansub.py $meganfile $rank
$output1 $output2 $output3 $pick_format
  #if $more_options.opts_selector == "advanced":
   ${more_options.minscore}
   ${more_options.maxexpected}
   ${more_options.toppercent}
   ${more_options.minsupport}
   ${more_options.mincomplexity}
   ${more_options.use_minimal_coverage_heuristic}
   ${more_options.use_identity_filter}
   ${more_options.minsupportpercent}
   ${more_options.lcapercent}
   ${more_options.paired}
  #else:
  5.0
  0.01
  10.0
  1
  0.44
  false
  false
  0.1
  100.0
  false
  #end if
  </command>
  <inputs>
```

```
   <param format="txt" name="meganfile" type="data"
label="MEGAN .rma file"/>
   <param name="rank" type="select" label="Select rank">
    <option value="all">All</option>
    <option value="SuperKingdom">SuperKingdom</option>
    <option value="Kingdom">Kingdom</option>
    <option value="Phylum">Phylum</option>
    <option value="Class">Class</option>
    <option value="Order">Order</option>
    <option value="Family">Family</option>
    <option value="Varietas">Varietas</option>
    <option value="Genus">Genus</option>
    <option value="Species_group">Species_group</option>
    <option value="Subspecies">Subspecies</option>
    <option value="Species">Species</option>
   </param>
   <param name="pick_format" type="select" label="Select image
format">
    <option value="eps">eps</option>
    <option value="svg">svg</option>
    <option value="gif">gif</option>
    <option value="png">png</option>
    <option value="jpg">jpg,jpeg</option>
    <option value="pdf">pdf</option>
    <option value="bmp">bmp</option>
   </param>
   <conditional name="more_options">
      <param name="opts_selector" type="select"
label="Advanced options">
         <option value="standard">Standard options</option>
         <option value="advanced">Advanced options</option>
      </param>
      <when value="standard" />
      <when value="advanced">
       <param name="minscore" type="float" value="5.0" label="
Minscore - Set a minimum threshold for the bit score of
hits"/>
       <param name="maxexpected" type="float" value="0.01"
label="Max Expected - Maximum threshold of E-value of hits"/>
       <param name="toppercent" type="float" value="10.0"
label="Top percent"/>
       <param name="minsupport" type="integer" value="1"
label=" Minsupport - Set a threshold for the minimum support
that a taxon
requires"/>
       <param name="mincomplexity" type="float" value="0.44"
label="Mincomplexity - Identify low complexity
reads"/>
       <param name="minsupportpercent" type="float"
value="0.1" label="Min Support Percent"/>
       <param name="lcapercent" type="float" value="100.0"
label="LCA Percent"/>
       <param name="use_minimal_coverage_heuristic"
type="select" label="Use Minimal Coverage Heuristic">
        <option value="false">No</option>
        <option value="true">Yes</option>
```

```
       </param>
       <param name="paired" type="select" label="Enable paired
analysis">
        <option value="false">No</option>
        <option value="true">Yes</option>
       </param>
       <param name="use_identity_filter" type="select"
label="Use Identity Filter (enable only for 16S rRNA
sequences)">
        <option value="false">No</option>
        <option value="true">Yes</option>
       </param>
      </when>
   </conditional>
  </inputs>
  <outputs>
  <data format="jpg" name="output1" label="Chart image"/>
  <data format="txt" name="output2" label=" Chart data txt "/>
  <data format="txt" name="output3" label=" tar.gz file "/>
  </outputs>
  <help>

**MEGAN t-analysis Overview**

The MEGAN t-analysis tool is attempting to provide a first
insight into the taxonomic content of a metagenomic sample
through taxonomic binning

-----

**Input formats**

This tool uses MEGAN rma files as input. This is the case of
the RMA builder tool output

-----

**Outputs**

The tool provides a txt file of the selected taxonomic nodes
and the number of reads assigned or summarized to them, a
similar chart image file and a tar.gz file containing an image
of the taxonomic tree and plain text files of all selected
nodes along with the specific sequences of the reads assigned
or summarized to them

-----

.. class:: warningmark

**WARNING: Rank options and default values are provided by
MEGAN v.5.5.3 official manual.**

.. class:: warningmark
```

WARNING: In case of a specific rank search, summarized reads are included in the results. Option "All" presents the assigned reads of each node, on a fully collapsed tree

.. class:: warningmark

WARNING: For the downloaded tar.gz file please add the .tar.gz extension

.. class:: infomark

Info:
The Min Support item can be used to set a threshold for the minimum support that a taxon requires, that is, the number of reads that must be assigned to it so that it appears in the result.Any read that is assigned to a taxon that does not have the required support is pushed up the taxonomy until a node is found that has sufficient support.

.. class:: infomark

Info:
The Min Support Percent item is used to set a threshold for the minimum support that a taxon requires, as a percentage of assigned reads. This feature is turned off by setting the value to 0. If a value greater than 0 (and at most 100) is given, then the program will set the Min Support threshold appropriately.

.. class:: infomark

Info:
The Min Score item can be used to set a minimum threshold for the bit score of hits.Any hit in the input data that scores less than the given threshold is ignored.

.. class:: infomark

Info:
The Max Expected item can be used to set a maximum threshold for the expected value of hits.Any hit in the input data whose E-value exceeds this value is ignored.

.. class:: infomark

Info:
The Top Percentage item can be used to set a threshold for the maximum percentage by which the score of a hit may fall below the best score achieved for a given read.Any hit that falls below this threshold is discarded.The Min Complexity item can be used to identify low complexity reads.These are placed on a special Low Complexity node.To turn this filter off, set the value to 0. A value of 0.3 catches most low complexity short reads.

.. class:: infomark

```
**Info:**
The Paired Reads item can be used to turn paired-read
awareness of MEGAN on and off.In paired-read mode, MEGAN
utilities read-pairing information to enhance the taxonomic
assignment of reads.

.. class:: infomark

**Info:**
The Use 16S Percent Identity Filter item can be used to turn
on an additional filter for assigning reads to a specific
taxonomic level.When this is active, the percent identity of a
match must exceed the given value of percent identity to be
assigned at the given rank:
Species 99%, Genus 97%, Family 95%, Order 90%, Class 85%,
Phylum 80%.This should only be used when analyzing 16S rRNA
sequences.

.. class:: infomark

**Info:**
Minimal Coverage Heuristic,use a minimum set of taxa that
cover all reads.Increases the specificity of the LCA
algorithm.

.. class:: infomark

**Info:**
The LCA Percent item is used to set the percent of matches
that the LCA of a read must cover, in the range 50-100. When a
value of less than 100 is specified then the LCA of a fixed
percent is used.

  </help>

 </tool>
```

functionalsub.py

```
#!/usr/bin/python
import sys
import subprocess
import string
import os
import tarfile
import random
import shutil

unique_id= "".join(random.sample(string.letters, 12))

print 'Number of arguments:', len(sys.argv), 'arguments.'
print 'Argument List:', str(sys.argv)
```

```
meganfile = sys.argv[1] # tool input
tardata = sys.argv[2] # tar.gz output
chartdata1 = sys.argv[3] # chart image SEED output
chartdata2 = sys.argv[4] # chart image COG output
chartdata3 = sys.argv[5] # chart image KEGG output
minscore =sys.argv[6] # MEGAN parameter
maxexpected = sys.argv[7] # MEGAN parameter
toppercent = sys.argv[8] # MEGAN parameter
minsupport = sys.argv[9] # MEGAN parameter
mincomplexity = sys.argv[10] # MEGAN parameter
use_minimal_coverage_heuristic = sys.argv[11] # MEGAN
parameter
use_identity_filter = sys.argv[12] # MEGAN parameter
minsupportpercent = sys.argv[13] # MEGAN parameter
lcapercent = sys.argv[14] # MEGAN parameter
paired_reads = sys.argv[15] # MEGAN parameter


a=os.path.realpath(meganfile)
b=a+'.rma'
subprocess.call('cp %s %s' % (a,b),  shell=True)
meganfile=b # add the extension .rma to the input file

newpath = ('/tmp/%s_OutputsSEED' % (unique_id) )
os.makedirs(newpath) # create a directory for SEED outputs

newpath = ('/tmp/%s_OutputsCOG' % (unique_id) )
os.makedirs(newpath) # create a directory for COG outputs

newpath = ('/tmp/%s_OutputsKEGG' % (unique_id) )
os.makedirs(newpath) # create a directory for KEGG outputs

textfile=open('/tmp/%s_commandsall.txt' % (unique_id),'w')
textfile.write("open file=" + (meganfile) + ";\nrecompute
minSupportPercent="+(minsupportpercent)+"
minSupport="+(minsupport)+" minScore="+(minscore)+"
maxExpected="+(maxexpected)+" topPercent="+(toppercent)+"
lcaPercent="+(lcapercent)+" minComplexity="+(mincomplexity)+"
useMinimalCoverageHeuristic="+(use_minimal_coverage_heuristic)
+" pairedReads="+(paired_reads)+"
useIdentityFilter="+(use_identity_filter)+" useSeed=true
useCOG=true useKegg=true;\nshow window=seedViewer;\nshow
window=seedViewer;\nset context=seedViewer;\nuncollapse
nodes=all;\nset context=seedViewer;\nselect
nodes=leaves;\nextract what=reads
outDir=/tmp/"+(unique_id)+"_OutputsSEED outFile=reads-%t.fasta
data=SEED ids=SELECTED allBelow=true;\nexport what=DSV
format=seedpath_readname separator=tab
file='/tmp/"+(unique_id)+"_OutputsSEED/SEEDpath.txt';\ncollaps
e nodes=top;\nselect nodes=leaves;\nshow chart data=seed;\nset
context=SeedChart;\nshow values='true';\nexportImage file="+
(chartdata1) +" format=jpg replace=true;\nshow
window=cogViewer;\nshow window=cogViewer;\nset
context=cogViewer;\nuncollapse nodes=all;\nset
context=cogViewer;\nselect nodes=leaves;\nextract what=reads
```

```
outDir=/tmp/"+(unique_id)+"_OutputsCOG outFile=reads-%t.fasta
data=COG ids=SELECTED allBelow=true;\nexport what=DSV
format=cogpath_readname separator=tab
file='/tmp/"+(unique_id)+"_OutputsCOG/COGpath.txt';\ncollapse
nodes=top;\nselect nodes=leaves;\nshow chart data=cog;\nset
context=CogChart;\nshow values='true';\nexportImage file="+
(chartdata2) +" format=jpg replace=true;\nshow
window=keggViewer;\nshow window=keggViewer;\nset
context=keggViewer;\nuncollapse nodes=all;\nset
context=keggViewer;\nselect nodes=leaves;\nextract what=reads
outDir=/tmp/"+(unique_id)+"_OutputsKEGG outFile=reads-%t.fasta
data=KEGG ids=SELECTED allBelow=true;\nexport what=DSV
format=keggpath_readname separator=tab
file='/tmp/"+(unique_id)+"_OutputsKEGG/KEGGpath.txt';\ncollaps
e nodes=top;\nselect nodes=leaves;\nshow chart data=kegg;\nset
context=KeggChart;\nshow values='true';\nexportImage file="+
(chartdata3) +" format=jpg replace=true;\n quit;\n ")
textfile.close() # command file for MEGAN


subprocess.call ('xvfb-run --auto-servernum --server-num=1
/home/panagiout/megan5/./MEGAN -g -E <
/tmp/'+unique_id+'_commandsall.txt ',  shell=True) # call
MEGAN

os.remove ('/tmp/%s_commandsall.txt' % (unique_id)) # remove
command file

tar = tarfile.open("%s" % (tardata), "w:gz")
for name in ["/tmp/%s_OutputsSEED" % (unique_id),
"/tmp/%s_OutputsCOG" % (unique_id),  "/tmp/%s_OutputsKEGG" %
(unique_id)]:
    tar.add(name)
tar.close() # create the tar.gz file


shutil.rmtree("/tmp/%s_OutputsSEED" % (unique_id)) #remove all
tmp output directories
shutil.rmtree("/tmp/%s_OutputsCOG" % (unique_id))
shutil.rmtree("/tmp/%s_OutputsKEGG" % (unique_id))

os.remove(b)
```

megan_analysisf.xml

```
<tool id="megan_analysisf" name="MEGAN f-analysis">
 <description>for functional annotation</description>
  <command interpreter="python">functionalsub.py $meganfile
$output1 $output2 $output3 $output4
  #if $more_options.opts_selector == "advanced":
   ${more_options.minscore}
   ${more_options.maxexpected}
   ${more_options.toppercent}
   ${more_options.minsupport}
```

```
   ${more_options.mincomplexity}
   ${more_options.use_minimal_coverage_heuristic}
   ${more_options.use_identity_filter}
   ${more_options.minsupportpercent}
   ${more_options.lcapercent}
   ${more_options.paired}
  #else:
  5.0
  0.01
  10.0
  1
  0.44
  false
  false
  0.1
  100.0
  false
  #end if
  </command>
  <inputs>
   <param format="txt" name="meganfile" type="data"
label="MEGAN .rma file"/>
   <conditional name="more_options">
      <param name="opts_selector" type="select"
label="Advanced options">
         <option value="standard">Standard options</option>
         <option value="advanced">Advanced options</option>
      </param>
      <when value="standard" />
      <when value="advanced">
       <param name="minscore" type="float" value="5.0" label="
Minscore - Set a minimum threshold for the bit score of
hits"/>
       <param name="maxexpected" type="float" value="0.01"
label="Max Expected - Maximum threshold of E-value of hits"/>
       <param name="toppercent" type="float" value="10.0"
label="Top percent"/>
       <param name="minsupport" type="integer" value="1"
label=" Minsupport - Set a threshold for the minimum support
that a taxon
requires"/>
       <param name="mincomplexity" type="float" value="0.44"
label="Mincomplexity - Identify low complexity
reads"/>
       <param name="minsupportpercent" type="float"
value="0.1" label="Min Support Percent"/>
       <param name="lcapercent" type="float" value="100.0"
label="LCA Percent"/>
       <param name="use_minimal_coverage_heuristic"
type="select" label="Use Minimal Coverage Heuristic">
        <option value="false">No</option>
        <option value="true">Yes</option>
       </param>
       <param name="paired" type="select" label="Enable paired
analysis">
        <option value="false">No</option>
```

```
        <option value="true">Yes</option>
       </param>
       <param name="use_identity_filter" type="select"
label="Use Identity Filter (enable only for 16S rRNA
sequences)">
        <option value="false">No</option>
        <option value="true">Yes</option>
       </param>
      </when>
   </conditional>
  </inputs>
  <outputs>
  <data format="txt" name="output1" label=" tar.gz file
(functional analysis)"/>
  <data format="jpg" name="output2" label=" SEED Chart image
"/>
  <data format="jpg" name="output3" label="COG Chart image "/>
  <data format="jpg" name="output4" label="KEGG Chart image
"/>
  </outputs>
  <help>

**MEGAN f-analysis Overview**

The MEGAN f-analysis tool is providing information considering
the functional profile of a metagenomic sample. Three
different approaches are enabled for such a functional
annotation, as MEGAN activates SEED, COG and KEGG
classifications

-----

**Input formats**

This tool uses MEGAN rma files as input. This is the case of
the RMA builder tool output

-----

**Outputs**

The tool provides three different chart images,  one for each
type of classification, visualizing the number of reads
assigned to the top nodes of each classification. Moreover, a
tar.gz is created containing three different folders ( SEED,
COG, KEGG). In every one of them,  MEGAN creates a DSV file
with paths and readnames assigned to them and multiple files
of each leave node of a fully uncollapsed tree with the actual
sequences of the reads assigned to them

-----

.. class:: warningmark

**WARNING: Default values are provided by MEGAN v.5.5.3
official manual.**
```

.. class:: warningmark

WARNING: Charts are visualizations of the number of assigned reads at the top nodes of each classification tree

.. class:: warningmark

WARNING: For the downloaded tar.gz file please add the .tar.gz extension

.. class:: infomark

Info:
The Min Support item can be used to set a threshold for the minimum support that a taxon requires, that is, the number of reads that must be assigned to it so that it appears in the result.Any read that is assigned to a taxon that does not have the required support is pushed up the taxonomy until a node is found that has sufficient support.

.. class:: infomark

Info:
The Min Support Percent item is used to set a threshold for the minimum support that a taxon requires, as a percentage of assigned reads. This feature is turned off by setting the value to 0. If a value greater than 0 (and at most 100) is given, then the program will set the Min Support threshold appropriately.

.. class:: infomark

Info:
The Min Score item can be used to set a minimum threshold for the bit score of hits.Any hit in the input data that scores less than the given threshold is ignored.

.. class:: infomark

Info:
The Max Expected item can be used to set a maximum threshold for the expected value of hits.Any hit in the input data whose E-value exceeds this value is ignored.

.. class:: infomark

Info:
The Top Percentage item can be used to set a threshold for the maximum percentage by which the score of a hit may fall below the best score achieved for a given read.Any hit that falls below this threshold is discarded.The Min Complexity item can be used to identify low complexity reads.These are placed on a special Low Complexity node.To turn this filter off, set the value to 0. A value of 0.3 catches most low complexity short reads.

```
.. class:: infomark

**Info:**
The Paired Reads item can be used to turn paired-read
awareness of MEGAN on and off.In paired-read mode, MEGAN
utilities read-pairing information to enhance the taxonomic
assignment of reads.

.. class:: infomark

**Info:**
The Use 16S Percent Identity Filter item can be used to turn
on an additional filter for assigning reads to a specific
taxonomic level.When this is active, the percent identity of a
match must exceed the given value of percent identity to be
assigned at the given rank:
Species 99%, Genus 97%, Family 95%, Order 90%, Class 85%,
Phylum 80%.This should only be used when analyzing 16S rRNA
sequences.

.. class:: infomark

**Info:**
Minimal Coverage Heuristic,use a minimum set of taxa that
cover all reads.Increases the specificity of the LCA
algorithm.

.. class:: infomark

**Info:**
The LCA Percent item is used to set the percent of matches
that the LCA of a read must cover, in the range 50-100. When a
value of less than 100 is specified then the LCA of a fixed
percent is used.

  </help>

 </tool>
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Dupre J, O'Malley MA: **Metagenomics and biological ontology**. *Studies in history and philosophy of biological and biomedical sciences* 2007, **38**(4):834-846.

2. Handelsman J: **Metagenomics and Microbial Communities**. 2007.

3. Handlesman J: **Metagenomics: Application of Genomics to Uncultured Microorganisms**. *MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS,* Dec. 2004.

4. Steele HL, Streit WR: **Metagenomics: advances in ecology and biotechnology**. *FEMS microbiology letters* 2005, **247**(2):105-111.

5. Schloss PD, Handelsman J: **Biotechnological prospects from metagenomics**. *Current Opinion in Biotechnology* 2003, **14**(3):303-310.

6. Singh J, Behal A, Singla N, Joshi A, Birbian N, Singh S, Bali V, Batra N: **Metagenomics: Concept, methodology, ecological inference and recent advances**. *Biotechnology journal* 2009, **4**(4):480-494.

7. Committee on Metagenomics : Challenges and Functional Applications **The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet**; 2007.

8. Lee MH, Lee S-W: **Bioprospecting Potential of the Soil Metagenome: Novel Enzymes and Bioactivities**. *Genomics Inform* 2013, **11**(3):114-120.

9. Schmieder R, Edwards R: **Insights into antibiotic resistance through metagenomic approaches**. *Future Microbiol* 2012, **7**(1), 73–89.

10. Lupo A, Coyne S, Berendonk TU: **Origin and evolution of antibiotic resistance: the common mechanisms of emergence and spread in water bodies**. *Frontiers in microbiology* 2012, **3**:18.

11. Thomas T, Gilbert J, Meyer F: **Metagenomics - a guide from sampling to data analysis**. *Microbial informatics and experimentation* 2012, **2**(1):3.

12. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P: **A bioinformatician's guide to metagenomics**. *Microbiology and molecular biology reviews : MMBR* 2008, **72**(4):557-578, Table of Contents.

13. Pareek CS, Smoczynski R, Tretyn A: **Sequencing technologies and genome sequencing**. *Journal of applied genetics* 2011, **52**(4):413-435.

14. Morozova O, Marra MA: **Applications of next-generation sequencing technologies in functional genomics**. *Genomics* 2008, **92**(5):255-264.

15. Miller JR, Koren S, Sutton G: **Assembly algorithms for next-generation sequencing data**. *Genomics* 2010, **95**(6):315-327.

16. Boisvert S, Laviolette F, Corbeil J: **Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies**. *Journal of computational biology : a journal of computational molecular cell biology* 2010, **17**(11):1519-1533.

17. Scholz MB, Lo CC, Chain PS: **Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis**. *Curr Opin Biotechnol* 2012, **23**(1):9-15.

18. Schatz MC, Delcher AL, Salzberg SL: **Assembly of large genomes using second-generation sequencing**. *Genome research* 2010, **20**(9):1165-1173.

19. Mohammed MH, Ghosh TS, Singh NK, Mande SS: **SPHINX--an algorithm for taxonomic binning of metagenomic sequences**. *Bioinformatics* 2011, **27**(1):22-30.

20. Monzoorul Haque M, Ghosh TS, Komanduri D, Mande SS: **SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences**. *Bioinformatics* 2009, **25**(14):1722-1730.

21. Leung HC, Yiu SM, Yang B, Peng Y, Wang Y, Liu Z, Chen J, Qin J, Li R, Chin FY: **A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio**. *Bioinformatics* 2011, **27**(11):1489-1495.

22. Wooley JC, Godzik A, Friedberg I: **A Primer on Metagenomics**. *PLoS Comput Biol* 2010, **6(2):** e1000667.

23. Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C: **Computational meta'omics for microbial community studies**. *Molecular systems biology* 2013, **9**:666.

24. Prakash T, Taylor TD: **Functional assignment of metagenomic data: challenges and applications**. *Briefings in bioinformatics* 2012, **13**(6):711-727.

25. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M Nekrutenko A, Taylor J. **"Galaxy: a web-based genome analysis tool for experimentalists".** *Current Protocols in Molecular Biology*. 2010 Jan, Chapter 19:Unit 19.10.1-21.

26. Schartz MC:**The missing graphical user interface for genomics**. *Genome Biology* 2010,**11**:128.

27. Goecks J, Nekrutenko A, Taylor J and The Galaxy Team. **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences**. *Genome Biol.* 2010 Aug 25, **11**(8):R86.

28. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. **"Galaxy: a platform for interactive large-scale genome analysis."** *Genome Research*. 2005 Oct, **15**(10):1451-5.

29. Altschul S, Gish W, Miller W, Myers E, Lipman D: **Basic local alignment search tool**. *J Mol Biol*. 1990, **215**(3):403–410

30. McGinnis S, Madden TL: **BLAST: at the core of a powerful and diverse set of sequence analysis tools.** *Nucleic Acids Research* 2004, Vol. 32, Web Server issue

31. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC Bioinformatics* 2009, **10**: 421

32. Camacho C, Madden TL, Ma N, Tao T, Agarwala R, Morgulis A: **BLAST Command Line Applications User Manual.** *National Center for Biotechnology Information(US).* http://www.ncbi.nlm.nih.gov/books/NBK1763/ Accessed July 2013

33. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data.** *Genome Research* 2007, **17**(3): 377–386.

34. Mitra S, Klar B ,Huson DH: **Visual and statistical comparison of metagenomes.** *Genome analysis 2009,* Vol. 25 no. 15, 1849–1855.

35. Huson DH, Mitra S, Weber N, Ruscheweyh H, Schuster SC: **Integrative analysis of environmental sequences using megan4.** *Genome Research* 2011, **21**: 1552–1560.

36. Huson DH: **User Manual for MEGAN V5.5.3.** August 3, 2014

37. Huson DH, Weber N: **Microbial Community Analysis Using MEGAN.** *Methods in Enzymology* 2013, Volume 531, Chapter 21, 465-485

38. Yu K, Zhang T: **Construction of Customized Sub-Databases from NCBI-nr Database for Rapid Annotation of Huge Metagenomic Datasets Using a Combined BLAST and MEGAN Approach.** *PLoS ONE 2013*, **8**(4): e59831. doi:10.1371/journal.pone.0059831

39. El Hadidi M, Ruscheweyh HJ, Huson DH: **Improved Metagenome Analysis using MEGAN5.** *F1000Posters* 2013, **4**: 887

40. Mitra S, Rupek P, Richter DC, Urich T, Gilbert A J, Meyer F, Wilke A, Huson DH: **Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG.** *BMC Bioinformatics* 2011, **12**(Suppl 1):S21

41. Huson DH, Richter DC, Mitra S, Auch AF, Schuster SC: **Methods for comparative metagenomics.** *BMC Bioinformatics* 2009, **10**(Suppl 1):S12

42. Huson DH, Mitra S: **Comparative Metagenome Analysis Using MEGAN.** *Handbook of Molecular Microbial Ecology, Volume I: Metagenomics and Complementary Approaches 2011,* Chapter 39, 343-352

43. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, et al: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes**. *Nucleic Acids Res.* 2005, **33**:5691–5702.

44. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome scale analysis of protein functions and evolution.** *Nucleic Acids Res*. 2000, **28**: 33–36

45. Kanehisa M, Goto S: **KEGG: Kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* Jan 2000, **28**(1):27-30

46. You XY, Liu C, Wang SY, Jiang CY, Shah SA, Prangishvili D, She Q, Liu S J, Garrett RA**: Genomic analysis of *Acidianus hospitalis* W1 a host for studying crenarchaeal virus and plasmid life cycles.** *Extremophiles* 2011, **15**:487–497

47. Huber H, Huber R, Stetter KO: **"Thermoproteales." The Prokaryotes: Vol. 3: Archaea. Bacteria: Firmicutes, Actinomycetes.** Springer 2006, Chapter 2, **3**:10-22

48. Huber H, Stetter KO: **"Thermoplasmatales." The Prokaryotes: Vol. 3: Archaea. Bacteria: Firmicutes, Actinomycetes.** Springer 2006, Chapter 7, **3**:101-112

49. Mukherjee A, Wheaton GH, Blum PH, Kelly RM: **Uranium extremophily is an adaptive, rather than intrinsic, feature for extremely thermoacidophilic Metallosphaera species.** *Proc Natl Acad Sci U S A* 2012, **109**(41):16702-7

50. Gross W, Heilmann I, Lenze D,Schnarrenberger C: **Biogeography of the Cyanidiaceae (Rhodophyta) based on 18S ribosomal RNA sequence data.** *European Journal of Phycology* 2001, 36:3, 275-280

51. Wagner-Döbler I, Rheims H, Felske A, El-Ghezal A, Flade-Schröder D, Laatsch H, Lang S, Pukall R, Tindall BJ: **Oceanibulbus indolifex gen. nov., sp. nov., a North Sea alphaproteobacterium that produces bioactive metabolites.** *International Journal of Systematic and Evolutionary Microbiology* 2004, **54**(4):1177-84.

52. Dziewit L, Cegielski A, Romaniuk K, Uhrynowski W, Szych A, Niesiobedzki P, Zmuda-Baranowska MJ, Zdanowski MK, Bartosik D : **Plasmid diversity in Arctic strains of *Psychrobacter* spp.** Extremophiles 2013, **17**:433–444.

53. Huson DH, Xie C: **A poor man's BLASTX - high-throughput metagenomic protein database search using PAUDA.** Bioinformatics 2014, **30**(1): 38–39.

54. Buchfink B, Xie C, Huson DH: **Fast and Sensitive Protein Alignment using DIAMOND,** under review.