ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Σχολή Επιστημών Υγείας

Τμήμα Βιοχημείας και Βιοτεχνολογίας

Πρόγραμμα Μεταπτυχιακών Σπουδών

ΤΟΞΙΚΟΛΟΓΙΑ

« Ανάπτυξη βιοπληροφορικού πρωτοκόλλου για την ανάλυση μικροβιακών γονιδιωμάτων τοξικολογικού και εγκληματολογικού ενδιαφέροντος με τις τεχνολογίες αλληλούχισης νέας γενιάς Illumina και Pacific Biosciences »

Δρ. ΣΤΑΥΡΟΣ ΣΠΕΤΣΑΡΙΑΣ

Επιβλέπων:

Επίκουρος Καθηγητής
ΓΡΗΓΟΡΙΟΣ ΑΜΟΥΤΖΙΑΣ

ΛΑΡΙΣΑ 2018

Ανάπτυξη βιοπληροφορικού πρωτοκόλλου για την ανάλυση μικροβιακών γονιδιωμάτων τοξικολογικού και εγκληματολογικού ενδιαφέροντος με τις τεχνολογίες αλληλούχισης νέας γενιάς Illumina και Pacific Biosciences

Development of a Bioinformatics protocol for genome analysis of microbes with toxicological and forensic interest, using Illumina and Pacific Bioscience technologies

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

κ. Γρηγόριος Αμούτζιας (Επιβλέπων)
Επίκουρος Καθηγητής Βιοπληροφορικής στη Γενωμική
Τμήμα Βιοχημείας και Βιοτεχνολογίας
Πανεπιστήμιο Θεσσαλίας

κ. Παναγιώτης Μαρκουλάτος
Καθηγητής Εφαρμοσμένης Μικροβιολογίας με έμφαση στη Βιοτεχνολογία
Τμήμα Βιοχημείας και Βιοτεχνολογίας
Πανεπιστήμιο Θεσσαλίας

κ. Δημήτριος Μόσιαλος
Επίκουρος Καθηγητής Βιοτεχνολογίας Μικροβίων
Τμήμα Βιοχημείας και Βιοτεχνολογίας
Πανεπιστήμιο Θεσσαλίας

ΕΥΧΑΡΙΣΤΙΕΣ

Ευχαριστώ τη σύζυγό μου Ιωάννα που συνετέλεσε πολύπλευρα για την υλοποίηση της φοίτησής μου στο παρόν Μεταπτυχιακό Πρόγραμμα. **Αφιερώνω, επομένως, αυτή την εργασία στην Ιωάννα και στη μικρή μου κόρη Χριστίνα**.

Ευχαριστώ, επίσης, τον Επίκουρο Καθηγητή και πλέον, φίλο μου, κύριο Γρηγόριο Αμούτζια για την εμπιστοσύνη, το ενδιαφέρον, τις πολύτιμες επισημάνσεις του, την καλή του διάθεση, τον χρόνο και τον κόπο που διέθεσε για να υλοποιήσω αυτή την εργασία, όπως επίσης τον Καθηγητή κύριο Παναγιώτη Μαρκουλάτο και τον Επίκουρο Καθηγητή κύριο Δημήτριο Μόσιαλο για το ενδιαφέρον και τις επισημάνσεις τους επί του περιεχομένου της παρούσας.

Θέλω οπωσδήποτε να ευχαριστήσω τους: Μάριο Νικολαΐδη, Χρύσα Ντουντούμη και Βίκη Φλιάτουρα, φοιτητές του Τμήματος και του Εργαστηρίου Βιοπληροφορικής για την καλή τους διάθεση, το ενδιαφέρον και την πολύτιμη βοήθειά τους ως προς την πρακτική λειτουργία των βιοπληροφορικών εφαρμογών που αναφέρονται σε αυτή τη μελέτη.

Τέλος, ευχαριστώ τον Καθηγητή κύριο Δημήτριο Κουρέτα για το ενδιαφέρον και τη συνολική του εποπτεία επί των θεμάτων που άπτονταν του Μεταπτυχιακού Προγράμματος και την εύρυθμη λειτουργία του, καθώς και τον κύριο Αλέξανδρο Τουλουμτζίδη για το συνολικό του ενδιαφέρον και τη βοήθειά του επί των θεμάτων της Γραμματειακής υποστήριξης.

Δρ. Σταύρος Σπετσαρίας

ΠΕΡΙΛΗΨΗ

Η δυναμική των τεχνολογιών αλληλούχισης νέας γενιάς προμηνύει το μέλλον των τοξικολογικών, εγκληματολογικών, κλινικών και των ερευνητικών αναλύσεων. Οι τεχνολογίες που εφαρμόζουν οι αναλυτές των εταιριών Illumina και Pacific Biosciences (PacBio), αν και διαφορετικές μεταξύ τους, αποτελούν τον ακρογωνιαίο λίθο για την αξιόπιστη και ταχεία ανάλυση μικροβιακών γονιδιωμάτων τοξικολογικού και εγκληματολογικού ενδιαφέροντος.

Η υβριδική συναρμολόγηση των δεδομένων αλληλούχισης των εν λόγω μικροβιακών γονιδιωμάτων, μέσω της συνδυαστικής αξιοποίησης των δεδομένων της Illumina και της PacBio, οδηγεί σε συναρμολογήσεις εξαιρετικά υψηλής ακρίβειας, γεγονός ιδιαίτερα σημαντικό για την τελική αποτίμηση των αποτελεσμάτων των τοξικολογικών, καθώς και των εγκληματολογικών αναλύσεων.

Επειδή, ο όγκος των δεδομένων είναι τεράστιος, απαιτείται η εφαρμογή βιοπληροφορικών αναλύσεων για τη διαχείριση και αξιοποίησή τους. Σε αυτή την εργασία περιγράφεται η ανάπτυξη ενός βιοπληροφορικού πρωτοκόλλου ανάλυσης μικροβιακών γονιδιωμάτων τοξικολογικού και εγκληματολογικού ενδιαφέροντος.

Σε βάση δεδομένων αναζητήθηκαν γονιδιωματικά δεδομένα των τεχνολογιών αλληλούχισης Illumina και PacBio. Ακολούθησε το ποιοτικό φιλτράρισμα και η ανάλυση της ποιότητάς τους με ειδικά λογισμικά προγράμματα. Στη συνέχεια συναρμολογήθηκαν, ξεχωριστά, τα δεδομένα των δύο τεχνολογιών. Τα αποτελέσματα των συναρμολογήσεων συνδυάστηκαν σε μία υβριδική συναρμολόγηση.

Η νουκλεοτιδική αλληλουχία των συναρμολογημένων μικροβιακών γονιδιωμάτων εξετάστηκε για την εύρεση συγγενικών γονιδιωμάτων σε ειδική βάση δεδομένων, ενώ διενεργήθηκε, επιπλέον, η σύγκριση της αλληλουχίας του κάθε εξεταζόμενου γονιδιώματος με την αλληλουχία του πιο συγγενικού μικροβιακού γονιδιώματος αναφοράς. Τέλος, εξετάστηκε ο εντοπισμός γονιδίων παθογονικότητας σε ειδική βάση δεδομένων.

ABSTRACT

The dynamics of next-generation sequencing technologies outline the future of toxicological, forensic, clinical and research analyses. Technologies implemented by Illumina and Pacific Biosciences (PacBio) sequencers, although different from each other, are at the forefront. Hybrid assembly of Illumina and PacBio data, leads to high-precision genome assemblies, which is particularly important for the final assessment of the toxicological and forensic results. Because the volume of data is enormous, bioinformatics analyses are required to manage and exploit them. This thesis describes the development of a bioinformatics protocol for the analysis of microbial genomes of toxicological and forensic interest.

Genomic data from the Illumina and PacBio sequencing technologies were searched and retrieved from a public database. Quality control and filtering were performed by special software. The data of the two technologies were then separately assembled. The results of the assemblies were combined in a hybrid assembly.

The nucleotide sequence of the assembled microbial genomes was used for finding related genomes in a specific database, and a pairwise comparison was performed. Finally, the detection of pathogenicity genes in a specific database was examined.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ

1.1 Σκοπός της εργασίας

1.1.1 Βιοτρομοκρατία και παράγοντες τοξικολογικού και εγκληματολογικού ενδιαφέροντος

Σύμφωνα με τα Αμερικανικά Κέντρα Ελέγχου και Πρόληψης Νοσημάτων (Centers for Disease Control and Prevention, CDC) των Η.Π.Α., η βιοτρομοκρατία είναι η σκόπιμη απελευθέρωση ιών, βακτηρίων, τοξινών ή άλλων επιβλαβών παραγόντων που προκαλούν ασθένεια ή θάνατο σε ανθρώπους, ζώα ή φυτά (Bioterrorism|Anthrax). Αυτοί οι παράγοντες βρίσκονται συνήθως στη φύση, αλλά μπορούν να μεταλλαχθούν ή να τροποποιηθούν για να αυξήσουν την ικανότητά τους να προκαλέσουν ασθένειες, να καταστούν ανθεκτικοί σε φάρμακα ή να αυξήσουν την ικανότητά τους να εξαπλωθούν στο περιβάλλον. Γενικότερα, οι βιολογικοί παράγοντες μπορούν να εξαπλωθούν στον αέρα, στο νερό και στα τρόφιμα.

Οι τρομοκράτες τείνουν να χρησιμοποιούν τους βιολογικούς παράγοντες επειδή είναι εξαιρετικά δύσκολο να εντοπιστούν και να μην προκαλέσουν ασθένεια για αρκετές ώρες έως αρκετές ημέρες. Ορισμένοι παράγοντες βιοτρομοκρατίας, όπως ο ιός της ευλογιάς (smallpox virus), μπορούν να εξαπλωθούν από άτομο σε άτομο, ενώ κάποιοι άλλοι, όπως το βακτήριο του στελέχους *Bacillus anthracis* που προκαλεί τη λοιμώδη νόσο του άνθρακα (anthrax), δε μπορούν (Bioterrorism Overview / Preston, 2002).

Η βιοτρομοκρατία αποτελεί μια «ελκυστική» παράνομη δράση, επειδή οι βιολογικοί παράγοντες αποκτώνται εύκολα, κοστίζουν φθηνά, μπορούν εύκολα να διαδοθούν και μπορούν να προκαλέσουν εκτεταμένο φόβο και πανικό, πέρα από τις προκληθείσες σωματικές βλάβες (Advantages of Biologics as Weapons Bioterrorism, 2008). Ωστόσο, η βιοτρομοκρατία υπόκειται στον εξής σημαντικό περιορισμό: Είναι δύσκολο να χρησιμοποιηθεί

ένα βιολογικό όπλο με τρόπο που να επηρεάζει μόνο τον εχθρό και όχι τις φίλιες δυνάμεις. Βέβαια, ένα βιολογικό όπλο είναι χρήσιμο για τους τρομοκράτες, κυρίως, ως μια μέθοδος δημιουργίας μαζικού πανικού και διατάραξης της κοινωνικής ηρεμίας. Επιστήμονες, όπως ο Bill Joy, έχουν προειδοποιήσει για τη δυνητική δύναμη που μπορεί να θέσει η γενετική μηχανική στα χέρια των μελλοντικών βιοτρομοκρατών (Joy, 2007).

Ο όρος «βιοτρομοκρατία» δύναται να χρησιμοποιηθεί και για τη χρήση βιολογικών παραγόντων που δεν προκαλούν βλάβη στους ανθρώπους, αλλά μπορούν να προξενήσουν σημαντικές ζημίες στην οικονομία (Ray, 2002). Σε αυτή την κατηγορία ανήκει και ο ιός του αφθώδους πυρετού (Foot-and-Mouth Disease, FMD), ο οποίος είναι ικανός να προκαλέσει εκτεταμένες οικονομικές ζημίες και ανησυχίες για το κοινό, όπως διαπιστώθηκε στις εστίες του αφθώδους πυρετού κατά τα έτη 2001 και 2007 στο Ηνωμένο Βασίλειο, ενώ δεν έχει σχεδόν καμία ικανότητα να μολύνει τον άνθρωπο.

1.1.2 Κατηγορίες βιολογικών παραγόντων τοξικολογικού ενδιαφέροντος

Τα Κέντρα Ελέγχου και Πρόληψης Ασθενειών των Η.Π.Α. κατηγοριοποιούν τους βιολογικούς παράγοντες που έχουν τη δυνατότητα να δημιουργήσουν σοβαρή απειλή για τη δημόσια υγεία και ασφάλεια σε τρεις κατηγορίες: Α, Β και Γ. Ωστόσο, κάποιοι επιβλαβής βιολογικοί παράγοντες, όπως ο ιός H5N1 των πτηνών, δεν έχουν ακόμη κατηγοριοποιηθεί.

- **Κατηγορία Α**

Οι βιολογικοί παράγοντες αυτής της κατηγορίας θεωρούνται υψίστης σημασίας διότι δύνανται να θέσουν σε κίνδυνο την εθνική ασφάλεια, καθώς μπορούν εύκολα να μεταδοθούν και να διαδοθούν, να οδηγήσουν σε υψηλή θνησιμότητα, να έχουν σημαντικές επιπτώσεις στη δημόσια υγεία, να προκαλέσουν δημόσιο πανικό ή να απαιτήσουν ειδική δράση ετοιμότητας. Στην κατηγορία Α ανήκουν οι εξής παράγοντες:

- Τουλαραιμία (Tularemia)

Η Τουλαραιμία έχει πολύ χαμηλό ποσοστό θνησιμότητας εάν ο προσβεβλημένος ασθενής υποβληθεί, εγκαίρως, σε θεραπεία, αλλά μπορεί και να υποστεί σοβαρή ανικανότητα (Tularemia - Emergency Preparedness & Response). Προκαλείται από το βακτήριο *Francisella tularensis* και είναι ασθένεια των ζωών που προσβάλλει, κυρίως, τον λαγό και σπανιότερα το ζαρκάδι, τη γάτα, τον σκύλο, τα πρόβατα, τους χοίρους, τα βοοειδή κ.α. Από τα προσβεβλημένα ζώα μεταδίδεται στον άνθρωπο, κυρίως με ενοφθαλμισμό του βακτηρίου στο δέρμα των κυνηγών, των κτηνοτρόφων των σφαγέων, και των κτηνιάτρων. Μπορεί να μεταδοθεί και από την κατανάλωση κρέατος που δε βράστηκε ή δε ψήθηκε καλά, με την εισπνοή, με το πόσιμο νερό (Τουλαραιμία, Βικιπαίδεια), καθώς και με τα τσιμπήματα εντόμων. Το βακτήριο *Francisella tularensis* είναι πολύ μολυσματικό. Ένας πολύ μικρός αριθμός κυττάρων (10-50 περίπου) μπορεί να προκαλέσει ασθένεια. Εάν το εν λόγω βακτήριο χρησιμοποιηθεί ως όπλο, τα βακτήρια πιθανότατα θα είναι αερομεταφερόμενα, με έκθεση μέσω της εισπνοής. Οι άνθρωποι που εισπνέουν ένα μολυσματικό αεροζόλ, εάν δεν υποβληθούν σε θεραπεία, θα υποστούν σοβαρές αναπνευστικές νόσους, συμπεριλαμβανομένης της πνευμονίας και των συστημικών λοιμώξεων. Το βακτήριο της Τουλαραιμίας εμφανίζεται ευρέως στη φύση και μπορεί να απομονωθεί και να αναπτυχθεί σε ποσότητα σε ένα εργαστήριο, αν και η κατασκευή ενός αποτελεσματικού όπλου αεροζόλ θα ήταν αρκετά πολύπλοκη διαδικασία και θα απαιτούσε εξαιρετικά εξειδικευμένες γνώσεις (Tularemia - Key Facts About Tularemia).

- Άνθρακας (Anthrax)

Η ασθένεια του Άνθρακα είναι μη μεταδοτική και προκαλείται από το βακτήριο *Bacillus anthracis* το οποίο σχηματίζει σπόρια. Η ικανότητα του Άνθρακα να αναπαράγεται με μικρά σπόρια το καθιστά εύκολα διαπερατό από το πορώδες δέρμα και μπορεί να προκαλέσει άμεσα συμπτώματα, εντός 24 ωρών, από την έκθεση. Η διασπορά αυτού του παθογόνου παράγοντα σε πυκνοκατοικημένες περιοχές θεωρείται ότι θα μπορούσε να επιφέρει ποσοστό θνησιμότητας μικρότερο από το ένα τοις εκατό (Adalja et al., 2015). Υπάρχει εμβόλιο για τον Άνθρακα, αλλά απαιτούνται πολλαπλές ενέσεις και σε

σταθερές δόσεις. Όταν εντοπιστεί νωρίς, ο Άνθρακας μπορεί να θεραπευθεί με τη χορήγηση αντιβιοτικών, όπως η σιπροφλοξασίνη (Vietri et al., 2009).

Ο Άνθρακας είναι ένας από τους λίγους βιολογικούς παράγοντες στους οποίους έχουν εμβολιαστεί οι ομοσπονδιακοί υπάλληλοι των Ηνωμένων Πολιτειών. Στις Η.Π.Α. υπάρχει ένα εμβόλιο για τον Άνθρακα με ονομασία «Anthrax Vaccine Adsorbed» (AVA) και απαιτεί πέντε σειρές ενέσεων σε σταθερές δόσεις. Εκτός από το AVA, υπάρχουν και άλλα εμβόλια κατά του Άνθρακα.

- Ευλογιά (Smallpox)

Η Ευλογιά είναι ένας εξαιρετικά μεταδοτικός ιός. Μεταδίδεται εύκολα μέσω της ατμόσφαιρας και έχει υψηλό ποσοστό θνησιμότητας που ανέρχεται σε 20-40% (Smallpox Home). Εξαφανίστηκε τη δεκαετία του 1970 χάρη σε ένα παγκόσμιο πρόγραμμα εμβολιασμού (Smallpox - What CDC Is Doing to Protect the Public From Smallpox). Ωστόσο, ορισμένα δείγματα ιών εξακολουθούν να είναι διαθέσιμα στα ρωσικά και στα αμερικανικά εργαστήρια. Ορισμένοι πιστεύουν ότι μετά την κατάρρευση της Σοβιετικής Ένωσης, οι καλλιέργειες της Ευλογιάς έγιναν διαθέσιμες και σε άλλες χώρες. Αν και οι άνθρωποι που γεννήθηκαν πριν από το 1970 θα είχαν την ευκαιρία να εμβολιαστούν για την Ευλογιά στο πλαίσιο του προγράμματος του Παγκόσμιου Οργανισμού Υγείας, η αποτελεσματικότητα του εμβολιασμού είναι περιορισμένη δεδομένου ότι το εμβόλιο παρέχει υψηλό επίπεδο ανοσίας μόνο για 3 έως 5 χρόνια. Η προστασία με επανεμβολιασμό διαρκεί περισσότερο (IHB - The DoD Immunization Information and Training Portal). Ως βιολογικό όπλο η Ευλογιά είναι επικίνδυνη λόγω της εξαιρετικά μεταδοτικής φύσης της. Επίσης, η σπανιότητα με την οποία χορηγούνται εμβόλια στον γενικό πληθυσμό μετά την εκρίζωση της νόσου, σίγουρα, έχει αφήσει απροστάτευτους τους περισσότερους ανθρώπους σε περίπτωση εμφάνισης κάποιας εστίας. Η ευλογιά εμφανίζεται μόνο στους ανθρώπους και δεν έχει εξωτερικούς ξενιστές ή φορείς.

- Αλλαντική τοξίνη (Botulinum toxin)

Η Αλλαντική νευροτοξίνη είναι μία από τις πιο θανατηφόρες τοξίνες που παράγεται από το βακτηρίδιο *Clostridium botulinum* (Botulism - Emergency Preparedness & Response / Nigam et al., 2010). Η Αλλαντίαση προκαλεί τον

θάνατο από αναπνευστική ανεπάρκεια και παράλυση (Facts About Botulism). Ωστόσο, η τοξίνη είναι διαθέσιμη για καλλυντικές και, σε ορισμένες ειδικές περιπτώσεις, για θεραπευτικές εφαρμογές, όπως στην ουροδόχο κύστη των παραπληγικών ατόμων.

- Πανώλη (Plague)

Η Πανώλη είναι μια ασθένεια που προκαλείται από το βακτήριο *Yersinia pestis* (Plague Information). Τα τρωκτικά είναι ο φυσιολογικός ξενιστής της πανώλης και η ασθένεια μεταδίδεται στον άνθρωπο από δαγκώματα ψύλλων και, περιστασιακά, με αεροζόλ με τη μορφή πνευμονικής πανώλης (Plague Home Page). Η ασθένεια έχει ιστορικό χρήσης σε βιολογικό πόλεμο που χρονολογείται πολλούς αιώνες πίσω και θεωρείται απειλή λόγω της ευκολίας της καλλιέργειας και της ικανότητάς της να παραμένει στην κυκλοφορία μεταξύ των τοπικών πλυθησμών τρωκτικών για μεγάλο χρονικό διάστημα. Η χρήση της ως βιολογικό όπλο θα μπορούσε να γίνει με μόλυνση δια της εισπνοής, κυρίως, με τη μορφή της πνευμονικής Πανώλης (Plague - Frequently Asked Questions). Ήταν η ασθένεια που προκάλεσε τον «Μαύρο Θάνατο» στη Μεσαιωνική Ευρώπη. Με τον όρο «Μαύρη Πανώλη» ή «Μαύρος Θάνατος» αναφέρεται η πανδημία των ετών 1348-1353, η οποία ήταν από τις πλέον καταστροφικές στην παγκόσμια ιστορία. Ο συνολικός ανθρώπινος απολογισμός της, υπολογίζεται σε 75 έως 100 εκατομμύρια νεκρούς στην Ευρώπη και στην Ασία (Μαύρη πανώλη, Βικιπαίδεια).

- Ιογενείς αιμορραγικοί πυρετοί (Viral hemorrhagic fevers)

Στους Ιογενείς αιμορραγικούς πυρετούς περιλαμβάνονται αιμορραγικοί πυρετοί που προκαλούνται από μέλη της οικογένειας Filoviridae (ιός Marburg και ιός Ebola) και από την οικογένεια Arenaviridae (για παράδειγμα, οι ιοί Lassa και Machupo). Ειδικότερα, η νόσος του ιού Ebola προκάλεσε υψηλά ποσοστά θνησιμότητας που ανέρχονταν σε 25-90% με μέσο όρο 50%. Δεν υπάρχει θεραπεία, αν και η έρευνα για την ανακάλυψη εμβολίου βρίσκεται σε εξέλιξη. Η Σοβιετική Ένωση διερεύνησε τη χρήση των φιλοϊών για βιολογικό πόλεμο και η ομάδα του Aum Shinrikyo προσπάθησε ανεπιτυχώς να αποκτήσει καλλιέργειες ιού Ebola. Ο θάνατος από την ασθένεια του ιού Ebola οφείλεται συνήθως σε πολλαπλή οργανική ανεπάρκεια και στην υποογκαιμία (hypovolemic shock), η οποία είναι μια κρίσιμη κατάσταση του σώματος που

συμβαίνει στην περίπτωση μιας απότομης μείωσης στον πραγματικό όγκο του κυκλοφορούντος αίματος. Ένας άλλος επικίνδυνος ιός της οικογένειας των Filoviridae, ο ιός Marburg, ανακαλύφθηκε για πρώτη φορά στο Marburg της Γερμανίας. Καμιά θεραπεία δεν υπάρχει εκτός από την υποστηρικτική φροντίδα. Οι αρενοϊοί έχουν μειωμένο ποσοστό θνησιμότητας σε σχέση με τις ασθένειες που προκαλούνται από τους φιλοϊούς, αλλά είναι ευρύτερα διανεμημένοι, κυρίως στην κεντρική Αφρική και στη Νότια Αμερική (Viral Hemorrhagic Fevers).

- **Κατηγορία Β**

Οι βιολογικοί παράγοντες της κατηγορίας Β περιλαμβάνουν εκείνους που είναι σχετικά εύκολο να διαδοθούν, να οδηγήσουν σε μέτρια ποσοστά νοσηρότητας και σε χαμηλά ποσοστά θνησιμότητας (Bioterrorism Agents/Diseases)

Τα Βακτήρια της κατηγορίας Β και οι ασθένειες που προκαλούν είναι τα εξής:

Τα βακτήρια του γένους Brucella προκαλούν βρουκέλλωση (Brucellosis - Emergency Preparedness & Response). Το νόσημα παρατηρείται τόσο στα ζώα όσο και στον άνθρωπο και μεταδίδεται και από τα ζώα στον άνθρωπο. Το στέλεχος το οποίο προσβάλλει κυρίως τον άνθρωπο είναι η *Brucella melitensis* που συναντάται κατά βάση στα αιγοπρόβατα. Το μικρόβιο μεταδίδεται κυρίως με κατανάλωση μολυσμένων ζωικών προϊόντων, με άμεση επαφή με μολυσμένα ζώα και σπανιότερα από άνθρωπο σε άνθρωπο. Στις επιπλοκές της βρουκέλλωσης συμπεριλαμβάνονται η ενδοκαρδίτιδα, η μηνιγγοεγκεφαλίτιδα, η κερατίτιδα, η χρόνια ιριδοκυκλίτιδα και η ορχίτιδα. Κυρίως όμως μπορεί να εμφανιστεί αρθρίτιδα, οστεοαρθρίτιδα, οστεομυελίτιδα και δισκοσπονδυλίτιδα (Βρουκέλλωση, Βικιπαίδεια).

Το βακτήριο *Clostridium perfringens* και η τοξίνη «Epsilon» που παράγει, προκαλεί έντονες κράμπες στην κοιλιά και διάρροια που αρχίζει 8-22 ώρες μετά την κατανάλωση τροφίμων που περιέχουν μεγάλο αριθμό αυτών των βακτηρίων (epsilon toxin).

Βακτήρια του γένους Salmonella, Shigella, του είδους *Staphylococcus aureus* και το βακτηριακό στέλεχος *Escherichia coli* O157:H7 είναι επικίνδυνα για την ασφάλεια των τροφίμων, καθώς μπορούν να προκαλέσουν σοβαρή τροφική δηλητηρίαση. Επίσης, τα βακτήρια *Vibrio cholerae* (WebMD, Cholera)

και *Cryptosporidium parvum* μπορούν να προκαλέσουν μόλυνση της παροχής νερού με άμεσες και πολύ σοβαρές συνέπειες για την υγεία.

Το βακτήριο *Burkholderia mallei* προκαλεί την ασθένεια μάλη (Glanders) (Glanders, CDC). Το *Burkholderia mallei* είναι ικανό να μολύνει τους ανθρώπους. Η μετάδοση πραγματοποιείται μέσω της άμεσης επαφής με μολυσμένα ζώα και η είσοδος του μικροοργανισμού γίνεται μέσω εκδορών του δέρματος, των επιφανειών των ρινικών και των στοματικών βλεννογόνων, καθώς και με την εισπνοή. Λόγω του υψηλού ποσοστού θνησιμότητας στους ανθρώπους και του μικρού αριθμού των βακτηρίων που απαιτούνται για την εδραίωση της λοίμωξης, το *Burkholderia mallei* θεωρείται πιθανός βιολογικός παράγοντας βιοτρομοκρατίας (*Burkholderia mallei*, Wikipedia).

Το βακτήριο *Burkholderia pseudomallei* προκαλεί τη μελιοείδωση (Melioidosis, CDC / Why has melioidosis become a current issue?, CDC) Η θνησιμότητα από μελιοείδωση ανέρχεται σε ποσοστό 20-50%, ακόμη και κατόπιν θεραπείας (Wuthiekanun και Peacock, 2006).

Το βακτήριο *Chlamydia psittaci* ευθύνεται για την πρόκληση της ψιττάκωσης. Η ψιττάκωση μεταδίδεται από μολυσμένα πτηνά και δύναται να μεταδοθεί και στον άνθρωπο. Πιο ευαίσθητα να νοσήσουν είναι τα παιδιά, οι έγκυες, οι καρδιοπαθείς, οι νεφροπαθείς, οι διαβητικοί και οι εργαζόμενοι σε χώρους όπου υπάρχουν πτηνά (Ψιττάκωση, https://blog.doctoranytime.gr/glossary/psittakwsi/).

Το βακτήριο *Coxiella burnetii* προκαλεί τον πυρετό Q (Q Fever - Emergency Preparedness and Response). Οι άνθρωποι είναι ευάλωτοι στον πυρετό Q και η λοίμωξη μπορεί να προκληθεί και από μικρό αριθμό βακτηρίων (Q fever, https://www.cdc.gov/qfever/).

Το βακτήριο *Rickettsia typhi* ή *Rickettsia prowazekii* προκαλεί τον τύφο. Μεταδίδεται στον άνθρωπο μέσω ενδιάμεσων μολυσμένων ξενιστών όπως διάφορα είδη αρθρόποδων, ψύλλοι, ψείρες, τσιμπούρια και η μόλυνση πραγματοποιείται είτε μέσω κάποιου τσιμπήματος ή μέσω των περιττωμάτων (Τύφος, Βικιπαίδεια).

Στην κατηγορία Β περιλαμβάνονται επίσης οι Αλφαϊοί (Alphaviruses), οι οποίοι προκαλούν ιογενής εγκεφαλίτιδα.

- **Κατηγορία C**

Οι παράγοντες της κατηγορίας C είναι παθογόνοι ιοί, οι οποίοι μπορούν να προκαλέσουν σημαντικές επιπτώσεις στην υγεία. Στην κατηγορία αυτή ανήκουν οι εξής ιοί:

Ιός Νίπα (Nipahvirus)

Χανταϊός (Hantavirus)

Κορωναϊός σοβαρού οξέος αναπνευστικού συνδρόμου (Severe Acute Respiratory Syndrome coronavirus, SARS coronavirus)

Ιός της γρίπης H1N1 (Influenza A virus subtype H1N1)

Ιός ανθρώπινης ανοσοανεπάρκειας (Human Immunodeficiency Virus, HIV)

1.2 Ιστορική Αναδρομή

Κατά την περίοδο του Πρώτου Παγκοσμίου Πολέμου, οι προσπάθειες να χρησιμοποιηθεί η νόσος του άνθρακα κατευθύνονταν στους πληθυσμούς των ζώων. Ωστόσο, αυτό απεδείχθη ότι είναι αναποτελεσματικό. Επίσης, λίγο μετά την έναρξη του Πρώτου Παγκοσμίου Πολέμου, η Γερμανία ξεκίνησε μια εκστρατεία βιολογικού σαμποτάζ στις Ηνωμένες Πολιτείες, στη Ρωσία, στη Ρουμανία και στη Γαλλία (Gregory και Waag, 1997). Το 1915 ο Anton Dilger, ο οποίος ζούσε στη Γερμανία, στάλθηκε στις Ηνωμένες Πολιτείες φέρνοντας μαζί του καλλιέργειες μολυσματικών ασθενειών για άλογα και μουλάρια. Ο Dilger έστησε εργαστήριο στο σπίτι του στο Chevy Chase του Maryland και χρησιμοποίησε αχθοφόρους που εργάζονταν στις αποβάθρες στη Βαλτιμόρη για να μολύνουν άλογα τα οποία θα αποστέλλονταν στη Βρετανία. Ο Dilger ήταν ύποπτος ως γερμανός πράκτορας, αλλά δε συνελήφθη ποτέ. ‘Εφυγε τελικά στη Μαδρίτη της Ισπανίας, όπου πέθανε κατά τη διάρκεια της πανδημίας του ιού της γρίπης το 1918 (Experts Q & A, Public Broadcasting Service, 2006). Το 1916, οι Ρώσοι είχαν συλλάβει έναν Γερμανό πράκτορα με παρόμοιες προθέσεις. Η Γερμανία και οι σύμμαχοί της χρησιμοποίησαν μεθόδους για να μολύνουν τα άλογα του γαλλικού ιππικού και πολλά ρωσικά μουλάρια και άλογα στο Ανατολικό Μέτωπο. Οι ενέργειες αυτές

παρεμπόδισαν την κίνηση του πυροβολικού και των στρατευμάτων, καθώς και τις συνοδείες των εφοδίων και προμηθειών (Gregory και Waag, 1997).

Το 1972, η αστυνομία του Σικάγο συνέλαβε δύο φοιτητές, τους Allen Schwander και Stephen Pera, οι οποίοι είχαν προγραμματίσει να δηλητηριάσουν την υδροδότηση της πόλης με διάφορα παθογόνα βακτήρια και με τυφοειδή πυρετό, μια ασθένεια που προκαλείται από το βακτήριο *Salmonella typhi*. Ο Schwander ίδρυσε μια τρομοκρατική ομάδα, την "R.S.S.E.", ενώ ο Πέρα συγκέντρωσε και ανέπτυξε καλλιέργειες παθογόνων βακτηρίων στο νοσοκομείο όπου εργαζόταν.

Το 1980, ο Παγκόσμιος Οργανισμός Υγείας (WHO) ανακοίνωσε την εξάλειψη της ευλογιάς (smallpox), μια εξαιρετικά μεταδοτική ασθένεια. Αν και η ασθένεια έχει εξαλειφθεί στο φυσικό περιβάλλον, κατεψυγμένα αποθέματα ιού ευλογιάς εξακολουθούν να διατηρούνται από τις κυβερνήσεις των Ηνωμένων Πολιτειών και της Ρωσίας. Υπάρχει γενική ανησυχία για τις καταστροφικές συνέπειες στην περίπτωση που επικίνδυνοι πολιτικοί ή τρομοκράτες χρησιμοποιούσαν τα στελέχη της ευλογιάς. Δεδομένου ότι τα προγράμματα εμβολιασμού έχουν πλέον τερματιστεί, ο παγκόσμιος πληθυσμός είναι πολύ ευαίσθητος πλέον στην ευλογιά.

Στην πόλη The Dalles του Όρεγκον το 1984, οι οπαδοί του Bhagwan Shree Rajneesh προσπάθησαν να ελέγξουν τις τοπικές εκλογές καθιστώντας «ανίκανο» τον τοπικό πληθυσμό. Αυτό επιδιώχθηκε με τη μόλυνση στελέχους *Salmonella typhimurium* σε μπουφέ σαλάτας σε 11 εστιατόρια, σε προϊόντα παντοπωλείων και μανάβικων, σε πόμολα από πόρτες και άλλων σημείων επαφής σε δημόσιους χώρους. Η βιολογική επίθεση προκάλεσε σοβαρή τροφική δηλητηρίαση σε 751 άτομα, χωρίς θύματα όμως. Το περιστατικό αυτό αποτέλεσε την πρώτη γνωστή βιοτρομοκρατική επίθεση στις Ηνωμένες Πολιτείες τον 20ο αιώνα (Past U.S. Incidents of Food Bioterrorism, 2008). Ήταν, επίσης, η μεγαλύτερη τρομοκρατική επίθεση στο έδαφος των Η.Π.Α. (Novak, 2016).

Τον Ιούνιο του 1993, η θρησκευτική ομάδα Aum Shinrikyo απελευθέρωσε το βακτήριο του άνθρακα στο Τόκιο. Αυτόπτες μάρτυρες ανέφεραν μια άσχημη οσμή. Η επίθεση, ωστόσο, απέτυχε, καθώς η ομάδα χρησιμοποίησε εμβολιακό στέλεχος του βακτηρίου. Τα σπόρια που

ανακτήθηκαν από την επίθεση, έδειξαν ότι ήταν ταυτόσημα με ένα εμβολιακό στέλεχος άνθρακα που είχε δοθεί σε ζώα εκείνη την εποχή. Αυτά τα στελέχη του εμβολίου στερούνταν των γονιδίων που προκαλούν συμπτωματική απόκριση (Bacillus anthracis Incident).

Τον Σεπτέμβριο και τον Οκτώβριο του 2001, διάφορα περιστατικά άνθρακα συνέβησαν στις Ηνωμένες Πολιτείες, τα οποία προκλήθηκαν σκόπιμα. Συγκεκριμένα, επιστολές μολυσμένες με άνθρακα παραδόθηκαν ταυτόχρονα σε γραφεία μέσων μαζικής ενημέρωσης και στο Κογκρέσο. Οι επιστολές αυτές προκάλεσαν τον θανατο πέντε ανθρώπων (CNN, http://edition.cnn.com/2008/CRIME/08/06/anthrax.case/index.html).

Στην Ελλάδα, ευτυχώς για την ώρα, δεν έχουν καταγραφεί, επίσημα, περιστατικά που να παραπέμπουν σε απόπειρες ή σε τετελεσμένες ενέργειες βιοτρομοκρατίας.

1.3 Γονιδιωματική και Κλινική/Εγκληματολογική Μικροβιολογία

Η κλινική μικροβιολογία, καθώς και ο κλάδος της εγκληματολογικής μικροβιολογίας, κινούνται προς την κατεύθυνση των μοριακών διαγνωστικών προσεγγίσεων με τις οποίες θα αντιμετωπιστούν οι περιορισμοί των τρεχουσών διαγνωστικών μεθόδων που αφορούν στις μολυσματικές νόσους και στους παράγοντες που τις προκαλούν. Η αλληλούχιση νέας γενιάς έχει τη δυνατότητα χρήσης ως εξαιρετικού διαγνωστικού εργαλείου για τα λοιμώδη νοσήματα, περιλαμβάνοντας τη χρήση είτε για μεμονωμένους ασθενείς ή για τη δημόσια υγεία και την επιτήρηση των νόσων. Η εφαρμογή της αλληλούχισης νέας γενιάς για τη διάγνωση των μολυσματικών ασθενειών θα ενισχύσει την ακριβή και ταχεία αναγνώριση των παθογόνων παραγόντων, έτσι ώστε να μπορεί να εφαρμοστεί η πλέον ενδεδειγμένη και αποτελεσματική θεραπεία σε πολύ σύντομο χρόνο και με οικονομικά χαμηλό κόστος.

Αναφορικά με τις παραδοσιακές προσεγγίσεις για τη διάγνωση και επιτήρηση των λοιμωδών νοσημάτων, η τυπική μέθοδος της καλλιέργειας και της δοκιμής ευαισθησίας στα αντιβιοτικά εξακολουθεί να χρησιμοποιείται.

Ωστόσο, προϋποθέτει ότι μια ασθένεια προκαλείται από έναν μικροοργανισμό που μπορεί να καλλιεργηθεί. Συνεπώς, παθογόνα που δεν είναι καλλιεργήσιμα δεν εντοπίζονται με αποτέλεσμα να αποτυγχάνει η σωστή διάγνωση. Επιπλέον, είναι συχνά τα σφάλματα στην ταυτοποίηση των παθογόνων, ενώ ο χρόνος για τα αποτελέσματα μπορεί να διαρκέσει για κάποιες ημέρες. Το γεγονός αυτό αυξάνει τον κίνδυνο της απειλητικής για τη ζωή εξέλιξης και της εξάπλωσης της ασθένειας.

Οι ανοσοενζυμικές διαγνωστικές δοκιμές (Enzyme immunoassays) ανιχνεύουν τις σχετιζόμενες με τα παθογόνα πρωτεΐνες χρησιμοποιώντας αντισώματα. Αυτή η μεθοδολογία παρέχει ταχύτερα αποτελέσματα, αλλά εξακολουθούν να υπάρχουν ζητήματα αναφορικά με τις διακυμάνσεις στην ευαισθησία των εν λόγω δοκιμών. Επίσης, η διάγνωση μπορεί να επηρεαστεί, ανάλογα με τον χρόνο της εξέτασης σε σχέση με την έναρξη των συμπτωμάτων.

Η PCR πραγματικού χρόνου (Real-time PCR) χρησιμοποιείται, επίσης, στα κλινικά εργαστήρια μικροβιολογίας. Έχει αποδειχθεί ότι παρέχει ταχεία και έγκαιρη ανίχνευση παθογόνων παραγόντων. Δεδομένης της ευαισθησίας της, ωστόσο, τα ψευδή θετικά αποτελέσματα είναι συνηθισμένα, λόγω της ενίσχυσης του DNA από μικροοργανισμούς που δεν είναι στην πραγματικότητα η αιτία της νόσου σε ένα συγκεκριμένο άτομο ή πληθυσμό. Ψευδή αρνητικά αποτελέσματα συμβαίνουν επίσης για διάφορους λόγους, συμπεριλαμβανομένης της παρουσίας αναστολέων της PCR στα δείγματα.

Η ηλεκτροφόρηση παλμικού πεδίου (Pulsed-field gel electrophoresis) δημιουργεί πολύ συγκεκριμένα γονιδιωματικά πρότυπα DNA για διάφορους μικροοργανισμούς και τα στελέχη τους. Η χρονοβόρα πτυχή της συγκεκριμένης τεχνικής και η πολύπλοκη προετοιμασία των δειγμάτων αποτελούν τα μειονεκτήματά της.

Η γονοτύπηση της αλληλουχίας πολλαπλών γονιδιακών τόπων (Multilocus sequence typing, MLST) βασίζεται στην ανάλυση της αλληλουχίας του DNA των νουκλεοτιδικών πολυμορφισμών των διαχειριστικών γονιδίων (housekeeping genes), τα οποία είναι γονίδια που απαιτούνται για τη διατήρηση της βασικής κυτταρικής λειτουργίας. Παρόλο που η τεχνική MLST έχει αποδειχθεί εξαιρετικά σαφής και εύχρηστη, μπορεί συχνά να αποτύχει

στη διάκριση μεταξύ στελεχών, καθιστώντας περιορισμένη την εφαρμογή της για την επιδημιολογική επιτήρηση (Using Next-Generation Sequencing in Infectious Disease Diagnosis).

Με τη χρήση των τεχνολογιών αλληλούχισης νέας γενιάς, δεν υφίσταται η ανάγκη διεξαγωγής της χρονοβόρας διαδικασίας της καλλιέργειας, επειδή οι αναλύσεις μπορούν να ανιχνεύσουν μικροοργανισμούς απευθείας από τα δείγματα, συμπεριλαμβανομένων του αίματος και του εγκεφαλονωτιαίου υγρού. Επιπλέον, τα πλεονεκτήματα έναντι των συμβατικών μεθόδων γονοτύπησης για τη διερεύνηση των επιδημιών έχουν αποδειχθεί (Roetzer et al., 2013 / Köser et al., 2012), με κύριο εκείνο της ανίχνευσης στελεχών μικροοργανισμών και ιών τα οποία, σε πολλές περιπτώσεις, δε θα ήταν ανιχνεύσιμα με την εφαρμογή των συμβατικών μεθόδων γονοτύπησης, λόγω της περιορισμένης ευαισθησίας που τις διέπει. Επιπλέον, δεδομένου ότι οι σχετικές πληροφορίες με τη γονιδιωματική αλληλούχιση μπορούν να ληφθούν σε μερικές ημέρες, η ταυτοποίηση των παθογόνων μπορεί να επιτευχθεί σε πρώιμες καταστάσεις από την εκδήλωση της νόσου που προκαλούν.

Επιπροσθέτως, με την εφαρμογή των τεχνολογιών αλληλούχισης νέας γενιάς δύναται να ταυτοποιηθούν παθογόνα τα οποία θα χάνονταν στις συνήθεις καλλιέργειες και με την εφαρμογή των παραδοσιακών μεθόδων ταυτοποίησης. Με τη χρήση των τεχνολογιών αλληλούχισης νέας γενιάς διακρίνονται όλα τα στελέχη των διαφόρων παθογόνων, ανιχνεύονται μικτές λοιμώξεις και ανακαλύπτονται νέοι παθογόνοι παράγοντες.

Τα δεδομένα της πλήρους αλληλούχισης ενός γονιδιώματος, τα οποία αποκτώνται μέσω της αλληλούχισης νέας γενιάς, έχουν αποκαλύψει νέες πληροφορίες για αλληλουχίες που έχουν χρησιμοποιηθεί επιτυχώς για τον εντοπισμό των καθοριστικών παραγόντων της ανθεκτικότητας έναντι των φαρμάκων, όπως για παράδειγμα τον εντοπισμό του γονιδίου *mecC* στο βακτήριο *Staphylococcus aureus* (Jorgensen και Ferraro, 2009) και του *ampC* στο *Escherichia coli*. Εκτός από τον εντοπισμό μικροοργανισμών που είναι ανθεκτικοί σε φάρμακα, η εφαρμογή της αλληλούχισης ολόκληρου του γονιδιώματος βοήθησε στην αποκάλυψη μεταλλάξεων και άλλων γενετικών παραγόντων που σχετίζονται με την εξάπλωση των ανθεκτικών σε φάρμακα παθογόνων (Poirel et al., 2015).

1.4 Το Μέγεθος των βακτηριακών γονιδιωμάτων

Τα βακτηριακά γονιδιώματα ποικίλουν λιγότερο σε μέγεθος μεταξύ των διαφόρων ειδών των βακτηρίων και είναι μικρότερα σε σύγκριση με τα γονιδιώματα των ζώων και των ευκαρυωτικών μονοκύτταρων οργανισμών. Αυτό το γεγονός τα κάνει σχετικά εύκολη την πλήρη αλληλούχιση του γονιδιώματός τους.

Τα βακτηριακά γονιδιώματα κυμαίνονται σε μέγεθος από περίπου 130 kbp (McCutcheon και von Dohlen, 2011 / Van Leuven et al., 2014) έως και πάνω από 14 Mbp (Han et al., 2013). Μια μελέτη που διεξήχθη σε 478 βακτηριακά γονιδιώματα, κατέληξε στο συμπέρασμα ότι καθώς αυξάνεται το μέγεθος του γονιδιώματος, ο αριθμός των γονιδίων αυξάνεται με δυσανάλογα βραδύτερο ρυθμό στους ευκαρυώτες από ότι στους προκαρυωτικούς οργανισμούς. Έτσι, στους ευκαρυώτες, η αναλογία του μη κωδικοποιητικού DNA (non-coding DNA) αυξάνεται με το μέγεθος του γονιδιώματος πιο πολύ από ότι στα βακτήρια. Αυτό συμβαδίζει με το γεγονός ότι το περισσότερο ευκαρυωτικό πυρηνικό DNA δεν κωδικοποιεί γονίδια (Hou και Lin, 2009).

Μέχρι σήμερα έχουν προσδιοριστεί τουλάχιστον οι γονιδιωματικές αλληλουχίες 50 διαφορετικών βακτηριακών φύλων, καθώς και 11 διαφορετικών φύλων αρχαίων. Με τη μεθοδολογία αλληλούχισης δεύτερης γενεάς έχουν προσδιοριστεί αλληλουχίες μεγάλου αριθμού γονιδιωμάτων, ωστόσο σχεδόν το 90% των βακτηριακών γονιδιωμάτων που έχει κατατεθεί στη βάση δεδομένων γενετικών αλληλουχιών GenBank δεν είναι πλήρες. Με την εφαρμογή των τεχνολογιών αλληλούχισης τρίτης γενιάς, ενδεχομένως, δύναται να ταυτοποιηθεί η πλήρης αλληλουχία ενός γονιδιώματος μέσα σε χρονικό διάστημα λίγων ωρών. Η ανάλυση των γονιδιωματικών αλληλουχιών αποκαλύπτει μεγάλη γενετική ποικιλομορφία στα βακτήρια. Συγκεκριμένα, η ανάλυση σε περισσότερα από 2000 γονιδιώματα του βακτηρίου *Escherichia coli* αποκαλύπτει έναν αριθμό περίπου 3100 γονιδιακών οικογενειών, ενώ συνολικά στα βακτήρια υπολογίζονται συνολικά περίπου 89.000 διαφορετικές οικογένειες γονιδίων (Land et al., 2015).

Επιπλέον, από την ανάλυση των γονιδιωματικών αλληλουχιών αποδεικνύεται ότι τα παρασιτικά βακτήρια έχουν 500-1200 γονίδια, τα

ελεύθερα στη φύση βακτήρια διαθέτουν 1500-7500 γονίδια και τα αρχαία έχουν 1500-2700 γονίδια (Gregory, 2005).

Μια εντυπωσιακή ανακάλυψη, δείχνει τεράστιου μεγέθους γονιδιακή αλλοίωση όταν συγκρίνεται το μέγεθος του γονιδιώματος του βάκιλου της λέπρας με τα προγονικά του βακτήρια (Cole, et al., 2001). Μελέτες έχουν δείξει ότι πολλά βακτήρια έχουν μικρότερο μέγέθος γονιδιώματος από ότι οι πρόγονοί τους (Ochman, 2005). Κατά καιρούς έχουν προταθεί διάφορες θεωρίες με σκοπό να εξηγηθεί η γενική τάση της αποσύνθεσης του βακτηριακού γονιδιώματος και το σχετικά μικρό μέγεθος των βακτηριακών γονιδιωμάτων. Πειστικές αποδείξεις δείχνουν ότι η αποικοδόμηση των βακτηριακών γονιδιωμάτων οφείλεται στο φαινόμενο της μεροληπτικής εξάλειψης (deletional bias).

Ειδικότερα, παρόλο που τα βακτήρια αυξάνουν το περιεχόμενο του DNA τους μέσω της οριζόντιας μεταφοράς και του αναδιπλασιασμού των γονιδίων τους, τα γονιδιώματά τους παραμένουν μικρά και στερούνται μη λειτουργικών αλληλουχιών. Αυτό το μοτίβο εξηγείται ευκολότερα από μια διάχυτη μεροληψία προς μεγαλύτερους αριθμούς διαγραφών σε σχέση με τις προσθήκες. Κατ’ αυτόν τον τρόπο, όταν η φυσική επιλογή δεν είναι αρκετά δυνατή για να διατηρήσει την ισορροπία των προσθηκών σε σχέση με τις διαγραφές, τα γονίδια χάνονται σε μεγάλες διαγραφές ή απενεργοποιούνται και στη συνέχεια διαβρώνονται. Η απενεργοποίηση και η απώλεια γονιδίων είναι ιδιαίτερα εμφανείς στα υποχρεωτικά και στα συμβιωτικά παράσιτα, στα οποία οι δραματικές μειώσεις στο μέγεθος του γονιδιώματός τους μπορεί να προκύψουν όχι από την επιλογή για να χάσουν το DNA τους, αλλά από τη μειωμένη επιλογή για τη διατήρηση της γονιδιακής λειτουργικότητας. Έτσι λοιπόν, η μεροληπτική εξάλειψη του γονιδιωματικού DNA είναι μια σημαντική δύναμη που διαμορφώνει τα βακτηριακά γονιδιώματα (Mira et al., 2001).

1.5 Η αλληλούχιση Shotgun

Η τεχνολογία αλληλούχισης που είναι γνωστή ως «Sanger» μπορεί να χρησιμοποιηθεί μόνο για αρκετά βραχείς κλώνους DNA, από 100 έως 1000

ζεύγη βάσεων. Αντίθετα, η μεθοδολογία αλληλούχισης «Shotgun» (Shotgun sequencing) είναι μια μέθοδος που χρησιμοποιείται για την αλληλούχιση μεγάλων κλώνων DNA. Ονομάζεται «shotgun» κατ' αναλογία με το σχεδόν τυχαίο πρότυπο που προκύπτει από την πυροδότηση ενός κυνηγετικού όπλου σε έναν στόχο.

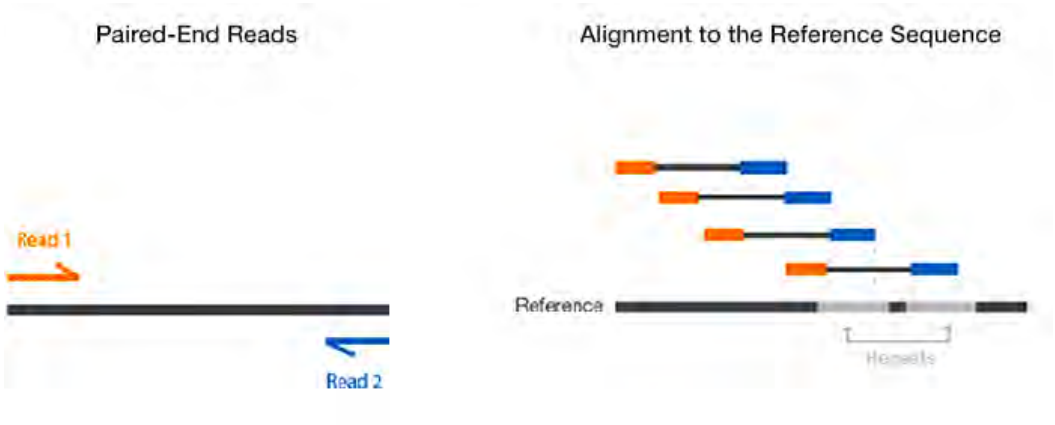
Γενικά, οι μακρύτερες αλληλουχίες σπάνε σε μικρότερα θραύσματα τα οποία μπορούν να υποβληθούν σε αλληλούχιση χωριστά και στη συνέχεια συναρμολογούνται για να δώσουν τη συνολική αλληλουχία. Δύο βασικές μέθοδοι χρησιμοποιούνται για αυτό τον σκοπό: το Χρωμοσωμικό περπάτημα (Chromosome walking), το οποίο προχωρά σε ολόκληρο τον κλώνο, τμηματικά, κομμάτι προς κομμάτι και η αλληλούχιση Shotgun, η οποία είναι μια ταχύτερη αλλά πιο περίπλοκη διαδικασία που χρησιμοποιεί τυχαία θραύσματα.

Στην αλληλούχιση Shotgun το DNA διασπάται τυχαία σε πολυάριθμα μικρά τμήματα-θραύσματα και, στη συνέχεια, προσδιορίζεται η αλληλουχία τους με κάποια μέθοδο. Από την αλληλούχιση των θραυσμάτων λαμβάνονται τα λεγόμενα «reads», τα οποία είναι οι ταυτοποιημένες αλληλουχίες των θραυσμάτων που προέκυψαν από τη διάσπαση του DNA.

Με την εκτέλεση αρκετών κύκλων αυτής της θραυσματοποίησης και της αλληούχισης των τμημάτων DNA που προκύπτουν, λαμβάνονται πολλαπλά αλληλεπικαλυπτόμενα reads του DNA στόχου. Έπειτα, ειδικά προγράμματα υπολογιστών χρησιμοποιούν τα αλληλεπικαλυπτόμενα άκρα των reads για να τα συναρμολογήσουν σε μια συνεχή αλληλουχία (Staden, 1979 / Anderson, 1981). Καμία από τις αλληλουχίες των reads που προκύπτουν δεν καλύπτει το πλήρες μήκος της αρχικής αλληλουχίας, αλλά τα reads μπορούν να συναρμολογηθούν στην αρχική αλληλουχία, χρησιμοποιώντας την επικάλυψη των άκρων τους. Στην πραγματικότητα, αυτή η διαδικασία χρησιμοποιεί τεράστιο όγκο πληροφοριών, ο οποίος είναι γεμάτος νουκλεοτιδικές αμφισημίες και σφάλματα. Η συναρμολόγηση σύνθετων γονιδιωμάτων περιπλέκεται, επιπρόσθετα, από τη μεγάλη αφθονία επαναλαμβανόμενων αλληλουχιών, γεγονός που σημαίνει ότι παρόμοια σύντομα reads θα μπορούσαν να προέρχονται από τελείως διαφορετικά μέρη της εξεταζόμενης αλληλουχίας. Έτσι λοιπόν, πολλά αλληλοεπικαλυπτόμενα

reads για κάθε τμήμα του αρχικού DNA είναι απαραίτητα για να ξεπεραστούν αυτές οι δυσκολίες και για να συναρμολογηθεί με μεγάλη ακρίβεια η αρχική αλληλουχία.

Συνήθως, κατά την αλληλούχιση Shotgun, πραγματοποιείται η αλληλούχιση ενός θραύσματος DNA σε ένα από τα δύο άκρα του, με αποτέλεσμα η νουκλεοτιδική αλληλουχία του read για το εν λόγω θραύσμα να προκύπτει από το ένα άκρο (single-read sequencing). Η βελτιωμένη παραλλαγή του single-read sequencing είναι γνωστή ως «Paired-end sequencing» (αλληλούχιση ζυγαρωμένων άκρων), ή ως double-barrel (διπλής κάνης) shotgun sequencing. Με τη μέθοδο Paired-end sequencing διεξάγεται η αλληλούχιση και στα δύο άκρα ενός θραύσματος DNA και με αυτόν τον τρόπο παράγονται δεδομένα αλληλουχιών reads υψηλής ποιότητας, τα οποία μπορούν να στοιχηθούν με μεγαλύτερη ευκολία σε μια αλληλουχία αναφοράς. Η αλληλούχιση Paired-end διευκολύνει την ανίχνευση γονιδιωματικών αναδιατάξεων, επαναλαμβανόμενων στοιχείων και γονιδιακών συντήξεων.

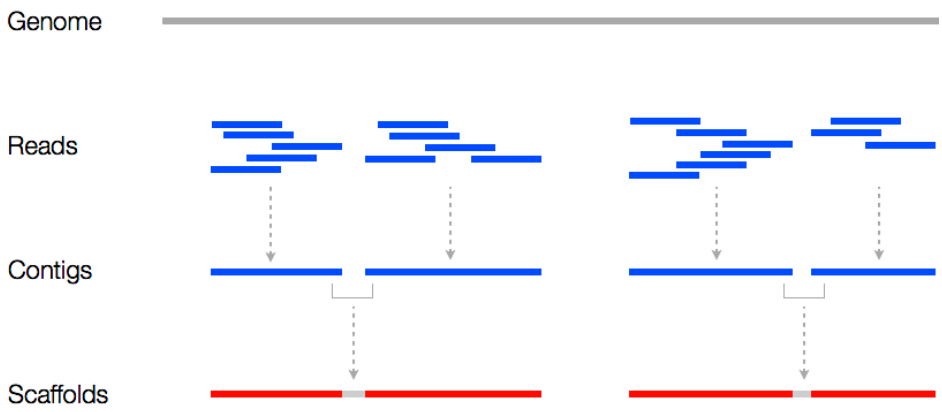


Εικόνα 1 Με την αλληλούχιση Paired-end είναι εφικτή η αλληλούχιση και των δύο άκρων ενός τμήματος DNA. Έτσι, είναι γνωστή η απόσταση μεταξύ του κάθε ζευγαρωμένου άκρου και οι αλγόριθμοι στοίχισης μπορούν να χρησιμοποιήσουν αυτές τις πληροφορίες για να χαρτογραφήσουν με μεγαλύτερη ακρίβεια τα reads.

Από: Advantages of paired-end and single-read sequencing, https://www.illumina.com/science/technology/next-generation-sequencing/paired-end-vs-single-read-sequencing.html

Όπως απεικονίζεται και στην Εικόνα 1, με την αλληλούχιση Paired-end καθίσταται εφικτή η αλληλούχιση και των δύο άκρων ενός τμήματος DNA. Επειδή, με αυτόν τρόπο είναι γνωστή η απόσταση μεταξύ του κάθε ζευγαρωμένου άκρου, οι αλγόριθμοι στοίχισης μπορούν να χρησιμοποιήσουν

αυτές τις πληροφορίες για να χαρτογραφήσουν με μεγαλύτερη ακρίβεια τα reads σε επαναλαμβανόμενες περιοχές των οποίων είναι δύσκολο να προσδιοριστεί η αλληλουχία (Advantages of paired-end and single-read sequencing). Η αρχική γονιδιωματική αλληλουχία ανακατασκευάζεται από τα reads χρησιμοποιώντας ειδικό λογισμικό συναρμολόγησης αλληλουχιών (sequence assembly software). Κατ 'αρχάς, τα αλληλοεπικαλυπτόμενα reads συναρμολογούνται σε μακρύτερες σύνθετες αλληλουχίες που είναι γνωστές ως «contigs». Τα contigs μπορούν να συνδεθούν μεταξύ τους σχηματίζοντας τις δομές «scaffolds», όπως απεικονίζεται σχηματικά στην Εικόνα 2.



Εικόνα 2 Τα αλληλοεπικαλυπτόμενα reads συναρμολογούνται σε μακρύτερες αλληλουχίες γνωστές ως contigs. Τα contigs συνδέονται μεταξύ τους στις δομές scaffolds.

Από: Biostars, «How to assemble contigs?», https://www.biostars.org/p/253222/

Ανάλογα με το μέγεθος του κενού που σχηματίζεται μεταξύ των contigs, μπορούν να χρησιμοποιηθούν διάφορες τεχνικές για να βρεθεί η αλληλουχία αυτών των κενών. Εάν το κενό είναι μικρό, της τάξης των 5-20 kb, τότε απαιτείται η χρήση της αλυσιδωτής αντίδρασης της πολυμεράσης (Polymerase Chain Reaction, PCR) για την ενίσχυση της εν λόγω περιοχής και στη συνέχει θα λάβει χώρα η αλληλούχιση. Εάν το κενό είναι μεγάλο, μεγαλύτερο από 20 kb, τότε το μεγάλο θραύσμα DNA κλωνοποιείται, ως ένθεμα, σε ειδικούς φορείς, τα βακτηριακά τεχνητά χρωμοσώματα (Bacterial artificial chromosomes, BAC) και έπειτα ακολουθεί η αλληλούχιση του ενθέματος του φορέα BAC.

Πολύ βασική έννοια στην αλληλούχιση shotgun είναι η κάλυψη ή βάθος (coverage ή depth) που αποδίδει τον αριθμό των reads που περιλαμβάνουν

ένα δεδομένο νουκλεοτίδιο στην ανακατασκευασμένη αλληλουχία (Sequencing Coverage / Sims et al., 2014). Συνεπώς, αλληλούχιση υψηλής κάλυψης, ή βαθιά αλληλούχιση σημαίνει δημιουργία μεγάλου αριθμού reads για κάθε περιοχή μιας ανακατασκευασμένης αλληλουχίας (Mardis, 2008). Στην Εικόνα 3 απεικονίζεται σχηματικά η αλληλοεπικάλυψη τεσσάρων reads, η συνολική αλληλουχία των οποίων θα χρησιμοποιηθεί για την ανακατασκευή μιας συναινετικής αλληλουχίας (consensus) συνολικού μήκους 37 βάσεων που θα προκύψει από τα δεδομένα της αλληλοεπικάλυψης των reads που τη συνθέτουν. Το μέγεθος του βάθους (κάλυψης) της αλληλούχισης σε κάθε σημείο υποδεικνύεται από τον αριθμό των αλληλοεπικαλυπτόμενων reads.

Εικόνα 3 Αλληλοεπικάλυψη τεσσάρων reads. Το μέγεθος της κάλυψης της αλληλούχισης σε κάθε σημείο υποδεικνύεται από τον αριθμό των αλληλοεπικαλυπτόμενων reads.

Από: Coverage (genetics) From Wikipedia, the free encyclopedia, https://en.wikipedia.org/wiki/Coverage_(genetics)

Γενικά, η μέση κάλυψη για ένα ολόκληρο γονιδίωμα μπορεί να υπολογιστεί από το μήκος του αρχικού γονιδιώματος G, τον αριθμό N των reads και το μέσο μήκος τους L ως: **N** x **L** / **G**. Για παράδειγμα, ένα υποθετικό γονιδίωμα μήκους 2.000 ζευγών βάσεων ανακατασκευασμένο από 8 reads που έχουν μέσο μήκος 500 νουκλεοτιδίων, θα έχει κάλυψη ίση με 2. Αυτή η παράμετρος επιτρέπει, επίσης, την εκτίμηση άλλων ποσοτήτων, όπως το ποσοστό του γονιδιώματος που καλύπτεται από reads, μέγεθος το οποίο μερικές φορές ονομάζεται, επίσης, ως κάλυψη. Γενικά, η υψηλή κάλυψη είναι επιθυμητή στην αλληλούχιση shotgun, επειδή έτσι μπορούν να μειωθούν τα σφάλματα κατά τη διαδικασία της συναρμολόγησης (Sims et al., 2014).

Πολλές θέσεις σε ένα γονιδίωμα περιέχουν σπάνιους μονονουκλεοτιδικούς πολυμορφισμούς (Single Nucleotide Polymorphisms, SNPs). Ως εκ τούτου, για να γίνει διάκριση μεταξύ των σφαλμάτων

αλληλούχισης και των πραγματικών SNPs, είναι απαραίτητο να αυξηθεί η ακρίβεια της αλληλούχισης ακόμη περισσότερο μέσω της διεξαγωγής της αλληλούχισης των μεμονωμένων γονιδιωμάτων πολλές φορές (Coverage, Wikipedia).

Παρόλο που η αλληλούχιση shotgun μπορεί θεωρητικά να εφαρμοστεί σε ένα γονιδίωμα οποιουδήποτε μεγέθους, η άμεση εφαρμογή της στην αλληλούχιση μεγάλων γονιδιωμάτων, για παράδειγμα στο ανθρώπινο γονιδίωμα, ήταν περιορισμένη μέχρι τα τέλη της δεκαετίας του 1990, όταν η τεχνολογική πρόοδος κατέστησε υπολογιστκά πρακτικό τον χειρισμό των τεράστιων ποσοτήτων δεδομένων (Dunham, 2005).

Ο πλήρης προσδιορισμός της αλληλουχίας ενός γονιδιώματος με αλληλούχιση shotgun μέχρι πρόσφατα περιοριζόταν τόσο από το μέγεθος των μεγάλων γονιδιωμάτων όσο και από την πολυπλοκότητα του υψηλού ποσοστού επαναλαμβανόμενου DNA, μεγαλύτερο από 50% για το ανθρώπινο γονιδίωμα, που υπάρχει στα μεγάλα γονιδιώματα. Δεν ήταν ευρέως αποδεκτό ότι η αλληλουχία ενός μεγάλου γονιδιώματος που έχει προκύψει με αλληλούχιση shotgun θα παρείχε αξιόπιστα δεδομένα. Για αυτούς τους λόγους, άλλες μεθοδολογίες που μείωσαν το υπολογιστικό φορτίο της συναρμολόγησης των αλληλουχιών έπρεπε να χρησιμοποιηθούν πριν από την εκτέλεση της αλληλούχισης shotgun (Venter, 2006). Στην ιεραρχική αλληλούχιση (hierarchical sequencing), γνωστή και ως αλληλούχιση «από πάνω προς τα κάτω» (top-down sequencing), δημιουργείται ένας φυσικός χάρτης με χαμηλή ανάλυση του γονιδιώματος πριν από την πραγματική αλληλούχιση. Από αυτόν τον χάρτη, ένας ελάχιστος αριθμός θραυσμάτων που καλύπτουν ολόκληρο το χρωμόσωμα επιλέγονται για αλληλούχιση. (Gibson G. and Muse SV. A Primer of Genome Science). Με τον τρόπο αυτό, απαιτείται το ελάχιστο της διεξαγωγής των διαδικασιών της αλληλούχισης υψηλής απόδοσης (high-throughput sequencing) και της συναρμολόγησης.

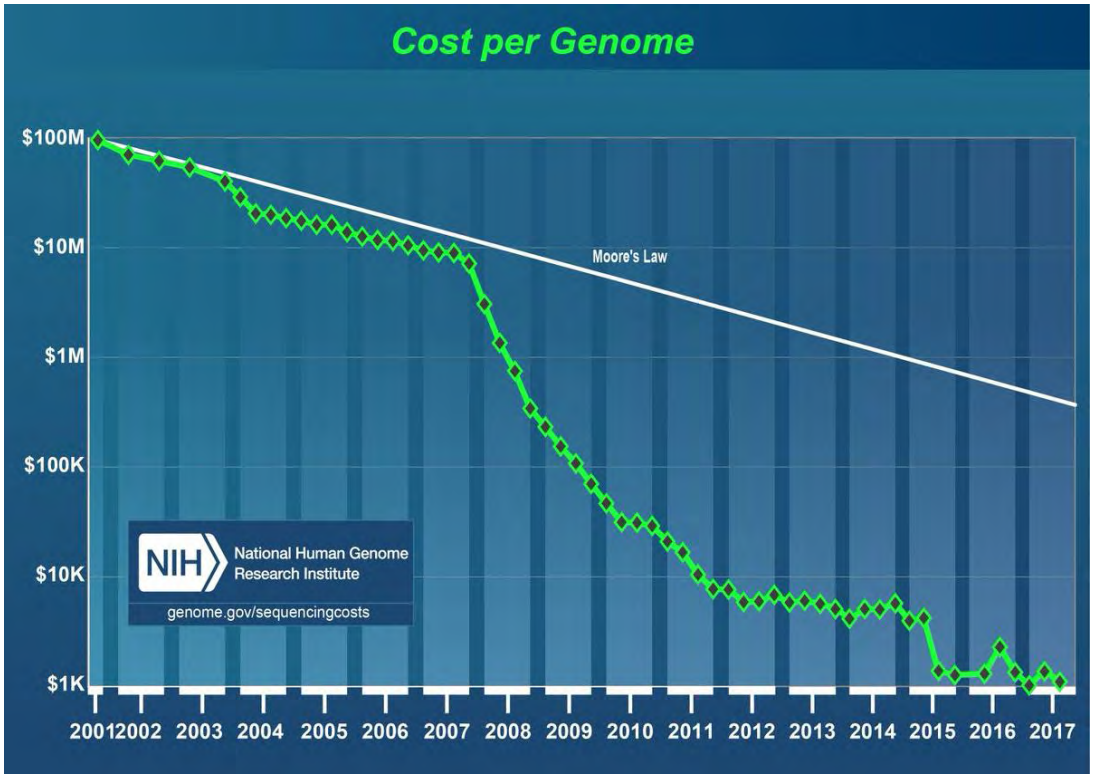
1.6 Τεχνολογίες αλληλούχισης νέας γενιάς

Η κλασική αλληλούχιση Shotgun βασίστηκε στη μέθοδο αλληλούχισης Sanger και αποτέλεσε την πιο προηγμένη τεχνική αλληλούχισης γονιδιωμάτων κατά τη χρονική περίοδο 1995-2005. Η μεθοδολογία αλληλούχισης Shotgun εξακολουθεί να εφαρμόζεται σήμερα, χρησιμοποιώντας, ωστόσο, τις τεχνολογίες αλληλούχισης «νέας γενιάς» (Next-generation sequencing, NGS) ή, όπως αλλιώς ονομάζονται, «υψηλής απόδοσης» (High-throughput sequencing). Με κάποιες από αυτές τις τεχνολογίες παράγονται reads μικρότερου μήκους, και συγκεκριμένα από 25 έως 500 bp, αλλά δημιουργούνται πολλές εκατοντάδες χιλιάδες ή εκατομμύρια reads σε σχετικά σύντομο χρονικό διάστημα, ακόμη και μέσα σε πολύ λίγο χρόνο (Voelkerding et al., 2009). Αυτό έχει ως αποτέλεσμα, από τη μια πλευρά, την επίτευξη υψηλής κάλυψης, αλλά από την άλλη, η διαδικασία της συναρμολόγησης είναι, σε μεγάλο βαθμό, πολύ πιο υπολογιστική. Οι τεχνολογίες αλληλούχισης νέας γενιάς είναι εξαιρετικά ανώτερες από την αλληλούχιση Sanger, λόγω του μεγάλου όγκου των παραγόμενων δεδομένων και του σχετικά μικρού χρόνου που απαιτείται για την αλληλούχιση ενός ολόκληρου γονιδιώματος (Metzker, 2010).

Σε σχέση με τις παραδοσιακές μεθόδους, η τεχνολογία αλληλούχισης νέας γενιάς είναι, πλέον, πολύ πιο φθηνή και πολύ πιο ταχεία,. Για παράδειγμα, η πρώτη αλληλούχιση του ανθρώπινου γονιδιώματος που δημιουργήθηκε από το «Human Genome Project» χρησιμοποιώντας την αλληλούχιση Sanger πήρε μία δεκαετία να διεκπεραιωθεί και κόστιζε δισεκατομμύρια δολάρια. Πλέον, η αλληλούχιση ενός ανθρώπινου γονιδιώματος με μεθόδους αλληλούχισης νέας γενιάς μπορεί να κοστίσει μόλις 1.000 δολάρια (High-Throughput Sequencing of DNA).

Στην Εικόνα 4 απεικονίζεται ένα γράφημα σχετικό με την ραγδαία πτώση του οικονομικού κόστους της αλληλούχισης του ανθρώπινου γονιδιώματος στη χρονική περίοδο από το 2001 έως σήμερα. Έπειτα από την ανακοίνωση της τετελεσμένης αλληλούχισης του ανθρώπινου γονιδιώματος το έτος 2004 από το «International Human Genome Sequencing Consortium 2004», το Εθνικό Ινστιτούτο Έρευνας Ανθρώπινου Γονιδιώματος (National

Human Genome Research Institute, NGHRI) δημιούργησε μια τεχνολογική πρωτοβουλία αλληλούχισης του DNA και διέθεσε για τον σκοπό αυτό 70 εκατομμύρια δολάρια με στόχο την επίτευξη της αλληλούχισης ενός ανθρώπινου γονιδιώματος στο ποσό των 1000 δολάρια σε βάθος χρόνου δέκα ετών (Schloss, 2008). Με αφορμή το γεγονός αυτό, από το 2004 και έπειτα, ανακαλύφθηκε μια σειρά τεχνολογιών αλληλούχισης νέας γενιάς. Όπως φαίνεται στην Εικόνα 4, από το 2006, όταν τέθηκε σε κυκλοφορία ο αναλυτής «Genome Analyzer II» της εταιρίας Illumina (Solexa), το κόστος της αλληλούχισης ξεκίνησε να πέφτει δραματικά, με αποτέλεσμα, σήμερα, να έχει ικανοποιηθεί ο στόχος του NGHRI.



Εικόνα 4 Γράφημα πτώσης του οικονομικού κόστους της αλληλούχισης του ανθρώπινου γονιδιώματος στη χρονική περίοδο από το 2001 έως σήμερα.

Από: Genome Atlantic, Here's the @NIH Cost per Genome graph updated for 2017. Wow., https://www.genome.gov/images/content/costpergenome_2017.jpg

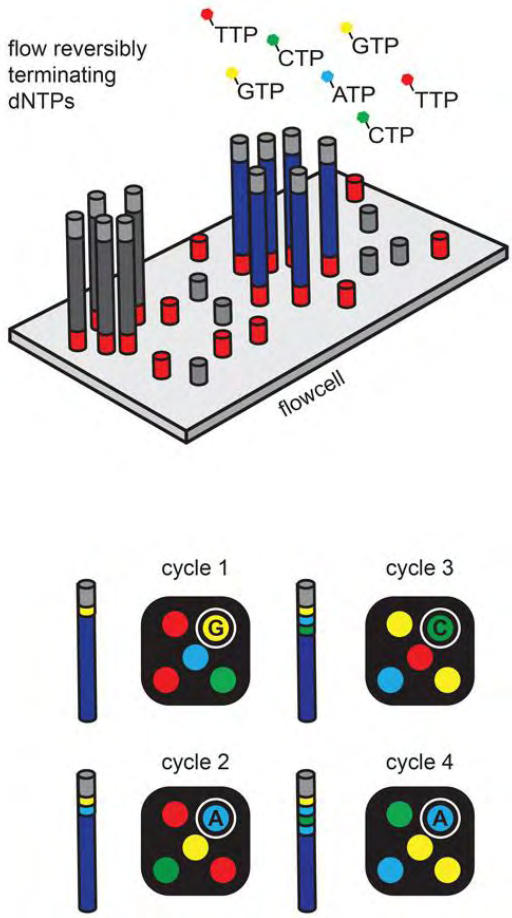
Ο δρόμος προς αυτό το ορόσημο περιλάμβανε πολλές εμπορικές πλατφόρμες τεχνολογιών αλληλούχισης νέας γενιάς, οι οποίες διαφέρουν στις λεπτομέρειες τους, αλλά συνήθως ακολουθούν ένα παρόμοιο γενικό παράδειγμα: προετοιμασία του γονιδιωματικού υλικού, κλωνική ενίσχυσή του,

ακολουθούμενη από κύκλους μαζικών και παράλληλων διαδικασιών αλληλούχισης. Η συγκεκριμένη στρατηγική που χρησιμοποιείται σε κάθε τεχνολογία καθορίζει την ποιότητα, την ποσότητα και τη μεροληψία των δεδομένων των αλληλουχιών που προκύπτουν, καθώς και τη χρησιμότητα της κάθε τεχνολογίας για συγκεκριμένες εφαρμογές.

Σήμερα, οι περισσότερο χρησιμοποιούμενες τεχνολογίες αλληλούχισης νέας γενιάς και ο τρόπος με τον οποίο λειτουργούν είναι οι εξής:

- **Illumina**

Η Illumina/Solexa κυκλοφόρησε τον αναλυτή «Genome Analyzer II» το 2006 ενώ οι πρόοδοι στην τεχνολογία της εν λόγω εταιρίας κατά τα παρελθόντα έτη την έχουν κάνει να κυριαρχήσει, επί του παρόντος, στην αγορά αλληλούχισης νέας γενιάς. Όπως φαίνεται στην Εικόνα 5, η διαδικασία της αλληλούχισης βασίζεται στην κλωνική ενίσχυση τμημάτων DNA που συνδέονται μέσω ειδικού προσαρμογέα στην επιφάνεια μιας γυάλινης κυψελίδας ροής (Bentley et al., 2008).



Εικόνα 5 Η αλληλούχιση με την τεχνολογία της Illumina βασίζεται στην κλωνική ενίσχυση τμημάτων DNA που συνδέονται μέσω προσαρμογέα στην επιφάνεια μιας γυάλινης κυψελίδας ροής. Οι βάσεις ταυτοποιούνται χρησιμοποιώντας μια μέθοδο κυκλικού αναστρέψιμου τερματισμού, κατά την οποία διεξάγεται ο προσδιορισμός της αλληλουχίας του υπό αλληλούχιση κλώνου DNA, ένα νουκλεοτίδιο κάθε φορά, μέσω προοδευτικών κύκλων που περιλαμβάνουν την ενσωμάτωση της βάσης, έκπλυση, απεικόνιση και διάσπαση.

Από: Reuter JA, Spacek D, Snyder MP. (2015). High-Throughput Sequencing Technologies. Molecular Cell 21; 58(4): 586-597. doi:10.1016/j.molcel.2015.05.004
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4494749/

Οι βάσεις ταυτοποιούνται χρησιμοποιώντας μια μέθοδο κυκλικού αναστρέψιμου τερματισμού, κατά την οποία διεξάγεται ο προσδιορισμός της αλληλουχίας του υπό αλληλούχιση κλώνου DNA (DNA template), ένα νουκλεοτίδιο κάθε φορά, μέσω προοδευτικών κύκλων που περιλαμβάνουν την ενσωμάτωση της βάσης, έκπλυση, απεικόνιση και διάσπαση. Σε αυτή τη μέθοδο χρησιμοποιούνται ελεύθερα δεοξυριβονουκλεοτίδια του τύπου «3΄-O-azidomethyldNTPs» για την παύση της αντίδρασης πολυμερισμού, τα οποία σημαίνονται με φθορισμό. Με την ενσωμάτωση ενός 3΄-O-azidomethyldNTP

σταματά ο πολυμερισμός του DNA, επιτρέποντας την απομάκρυνση των μη ενσωματωμένων δεοξυριβονουκλεοτιδίων, με έκπλυση, και την φθορίζουσα απεικόνιση για τον προσδιορισμό του προστιθέμενου νουκλεοτιδίου, μέσω μίας συσκευής συζευγμένου φορτίου (coupled-charge device, CCD) (Guo ey al., 2008). Μετά την απεικόνιση, αφαιρείται η φθορίζουσα μονάδα του ενσωματωμένου 3΄-O-azidomethyldNTP και έτσι απελευθερώνεται το 3΄ άκρο του υπό αλληλούχιση DNA, ώστε να επαναληφθεί η διαδικασία. Σε όλα τα μοντέλα Illumina, τα συνολικό ποσοστό σφάλματος είναι κάτω από 1% και ο συνηθέστερος τύπος σφάλματος είναι το λάθος διάβασμα βάσεων (Dohm et al., 2008).

Η εταιρία Illumina κατασκευάζει σήμερα μια ευρεία σειρά συσκευών αλληλούχισης (sequencers) βελτιστοποιημένων για ποικίλες διεργασίες, οι οποίες απεικονίζονται στην Εικόνα 6. Οι συσκευές MiSeq και HiSeq είναι οι πιο διαδεδομένες. Με τη συσκευή MiSeq η εκτέλεση μιας ανάλυσης πραγματοποιείται σε μόλις 4 ώρες και προορίζεται, μεταξύ άλλων, για αλληλούχιση μικρών γονιδιωμάτων. Η συσκευή HiSeq 2500, από την άλλη πλευρά, είναι κατασκευασμένη για εφαρμογές υψηλής απόδοσης. Η HiSeq 2500 μπορεί, επίσης, να τεθεί σε γρήγορη λειτουργία, η οποία ωστόσο είναι λιγότερο αποδοτική, αλλά μπορεί να παράγει 30 φορές το ανθρώπινο γονιδίωμα μέσα σε 27 ώρες.

Στις αρχές του 2014, η Illumina παρουσίασε τη συσκευή NextSeq 500, καθώς και τη HiSeq X Ten. Ομοίως με τη MiSeq, η NextSeq 500 έχει σχεδιαστεί ως ένας γρήγορος αναλυτής αλληλουχιών. Ωστόσο, η NextSeq 500 είναι σε θέση να παράγει 120 Gb δεδομένων, ή ένα μόνο γονιδίωμα 30 φορές, σε λιγότερο από 30 ώρες. Η συσκευή NextSeq 500 χρησιμοποιεί, επίσης, μια νέα μεθοδολογία αλληλούχισης δύο καναλιών. Σε αυτή τη μέθοδο, η κυτοσίνη είναι επισημασμένη με κόκκινο χρώμα, η θυμίνη με πράσινο, η αδενίνη με κίτρινο, ενώ η γουανίνη δεν είναι επισημασμένη. Σε αντίθεση με τη μεθοδολογία των τεσσάρων καναλιών που χρησιμοποιείται στις συσκευές MiSeq και HiSeq, η αλληλούχιση σε δύο κανάλια απαιτεί μόνο δύο εικόνες για την ανίχνευση των νουκλεοτιδίων, μείωση των χρόνων επεξεργασίας των δεδομένων και αύξηση της απόδοσης. Παρά τη μειωμένη πολυπλοκότητα, τα

συνολικά ποσοστά σφάλματος, τα οποία αντιστοιχούν σε λιγότερο από 1%, είναι παρόμοια με τις διαδεδομένες συσκευές HiSeq.

Η HiSeq X Ten είναι μια συσκευή κατάλληλη για την πλήρη αλληλούχιση του γονιδιώματος, σε πληθυσμιακή κλίμακα, που κυκλοφόρησε, επίσης, το 2014. Είναι ικανή να παράγει 18.000 γονιδιώματα με κάλυψη 30Χ ανά έτος. Τέλος, η συσκευή NovaSeq 6000 της Illumina, αποτελεί, σήμερα, την τελευταία λέξη της τεχνολογίας στη γονιδιωματική αλληλούχιση, συνδυάζοντας όλα τα παραπάνω προτερήματα σε μια συσκευή αλληλούχισης εξαιρετικά υψηλής απόδοσης για όλους τους τύπους των γονιδιωμάτων.

Εικόνα 6 Μοντέλα αναλυτών (sequencers) της Illumina.

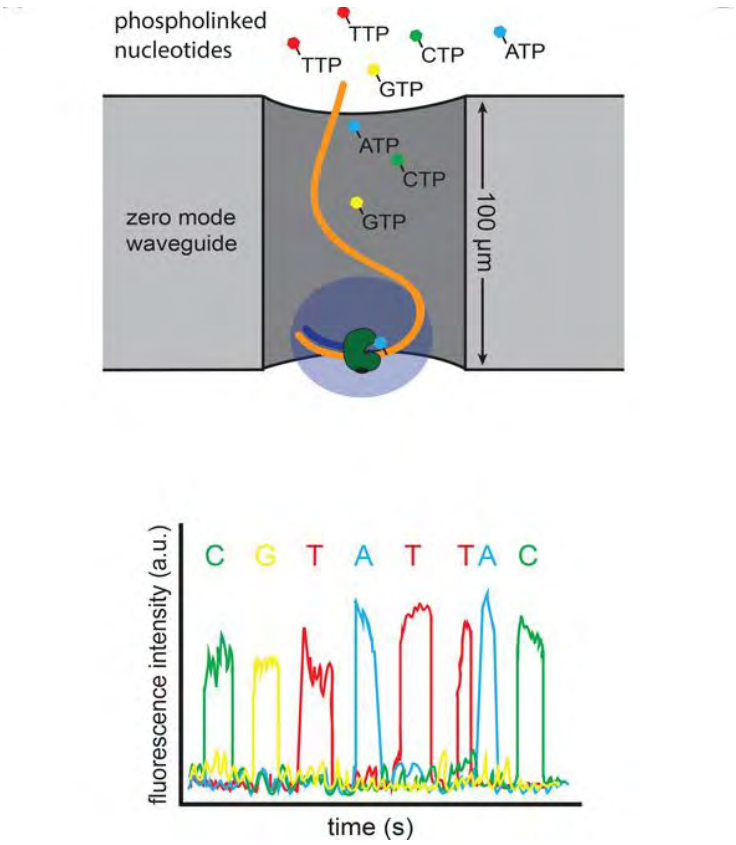
Από: Illumina ALL SYSTEMS, https://www.illumina.com/systems.html

- **Pacific Biosciences (PacBio)**

Η μέθοδος αλληλούχισης ενός μορίου DNA σε πραγματικό χρόνο (Single-molecule real-time, SMRT) ανακαλύφθηκε από την εταιρία Nanofluidics Inc. και διατέθηκε στο εμπόριο από την Pacific Biosciences. Στη μεθοδολογία της

Pacific Biosciences χρησιμοποιείται η τεχνική της ενίσχυσης πολλαπλής μετατόπισης (Multiple Displacement Amplification, MDA) του DNA, η οποία δε βασίζεται σε PCR. Αυτή η μέθοδος μπορεί να ενισχύσει, γρήγορα, μικρές ποσότητες δειγμάτων DNA, αποδίδοντας επαρκή ποσότητα για γονιδιωματική ανάλυση. Η αντίδραση ξεκινά με τον υβριδισμό εξανουκλεοτιδικών εκκινητών (random hexamer primers) που προσδένονται σε τυχαίες θέσεις στο υπό ενίσχυση DNA. Η σύνθεση διεξάγεται με τη χρήση ειδικής DNA πολυμεράσης που διαθέτει δράση μετατόπισης κλώνου (DNA Polymerase Strand Displacement Activity), σε σταθερή θερμοκρασία. Σε σύγκριση με τις συμβατικές τεχνικές ενίσχυσης PCR, η MDA παράγει προϊόντα μεγαλύτερου μεγέθους, με σχετικά μεγάλη όμως συχνότητα σφάλματος (Multiple displacement amplification, Wikipedia). Με τη χρήση της DNA πολυμεράσης μετατόπισης κλώνου (strand displacing DNA polymerase), δύναται να διεξαχθεί πολλές φορές η διαδικασία της αλληλούχισης στο αρχικό μόριο DNA, αυξάνοντας την ακρίβεια της μεθοδολογίας Pac Bio (Travers et al., 2010).

Όπως απεικονίζεται σχηματικά και στην Εικόνα 7, χρησιμοποιούνται δεοξυριβονουκλεοτίδια και των τεσσάρων βάσεων, των οποίων οι φωσφορικές ομάδες είναι κατάλληλα επισημασμένες με φθορίζοντα υλικά που δεν εμποδίζουν τον συνεχή πολυμερισμό του DNA. Η σύνθεση του DNA πραγματοποιείται σε ειδικούς θαλάμους, μεγέθους της τάξης του ζεπτολίτρου (1 ζεπτόλιτρο = 1.0×10^{-21} λίτρα), που ονομάζονται zero-mode waveguide, (ZMW), στον πυθμένα των οποίων ακινητοποιείται μία μόνο DNA πολυμεράση (Levene et al., 2003).



Εικόνα 7 Μέθοδος αλληλούχισης ενός μορίου DNA σε πραγματικό χρόνο που χρησιμοποιεί η τεχνολογία PacBio. Εφαρμόζεται η τεχνική της ενίσχυσης πολλαπλής μετατόπισης με τη χρήση DNA πολυμεράσης που διαθέτει δράση μετατόπισης κλώνου, σε σταθερή θερμοκρασία. Χρησιμοποιούνται δεοξυριβονουκλεοτίδια και των τεσσάρων βάσεων, των οποίων οι φωσφορικές ομάδες είναι κατάλληλα επισημασμένες με φθορίζοντα υλικά που δεν εμποδίζουν τον συνεχή πολυμερισμό του DNA.

Από: Reuter JA, Spacek D, Snyder MP. (2015). High-Throughput Sequencing Technologies. Molecular Cell 21; 58(4): 586-597. doi:10.1016/j.molcel.2015.05.004
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4494749/

Η φυσική αυτών των θαλάμων μειώνει τον θόρυβο του υποβάθρου και δε δημιουργείται πρόβλημα με την καταγραφή των σημάτων και των τεσσάρων ειδών των επισημασμένων δεοξυριβονουκλεοτιδίων, τα οποία μπορούν να συνυπάρχουν ταυτόχρονα. Έτσι, ο πολυμερισμός λαμβάνει χώρα χωρίς διακοπή και η αλληλουχία του DNA μπορεί να προσδιοριστεί σε πραγματικό χρόνο από την καταγραφή των φθοριζόντων σημάτων (Eid et al., 2009).

Στην Εικόνα 8 απεικονίζεται η συσκευή αλληλούχισης PacBio RS II της Pacific Biosciences με την οποία επιτυγχάνεται, μεταξύ άλλων, η αλληλούχιση ολόκληρων γονιδιωμάτων μικρού μεγέθους. Η συσκευή PacBio RS II

κυκλοφόρησε το 2010. Ωστόσο, έχουν γίνει σημαντικές τεχνολογικές παρεμβάσεις που έχουν βελτιώσει σημαντικά την απόδοση.

Εικόνα 8 Αναλυτής PacBio RS II της PacBio.

Από: PRODUCTS + SERVICES PACBIO SYSTEMS,
http://www.pacb.com/products-and-services/pacbio-systems/

Όπως συμβαίνει με τις περισσότερες τεχνολογίες αλληλούχισης, οι εξερχόμενες αλληλουχίες των reads που προκύπτουν με μια μοναδική αλληλούχιση, παρουσιάζουν υψηλά ποσοστά σφάλματος, τα οποία κυμαίνονται γύρω στο 11%. Αυτά τα σφάλματα περιλαμβάνουν, κυρίως, μικρές προσθήκες και διαγραφές (small **in**sertions and **del**etions, INDELs). Τα σφάλματα αλληλουχίας, ωστόσο, κατανέμονται τυχαία, με αποτέλεσμα η τελική αλληλουχία των reads που προκύπτει από τη συναινετική αλληλουχία ενός συνόλου από reads με αυξημένη κάλυψη, ή η αλληλουχία των reads που προκύπτει από πολλαπλά περάσματα αλληλούχισης της ίδιας πρότυπης αλληλουχίας, να χαρακτηρίζεται από ικανοποιητική ακρίβεια (Carneiro te al., 2012 / Koren et al., 2012).

Η αλληλούχιση SMRT είναι, επίσης, πολύ λιγότερο ευαίσθητη σε αλληλουχίες πλούσιες σε GC σε σχέση με άλλες τεχνολογίες (Loomis et al., 2013). Αυτά τα χαρακτηριστικά την καθιστούν ιδιαίτερα χρήσιμη για εργασίες που αφορούν στην εκ νέου/*de novo* συναρμολόγηση των βακτηριακών και ιικών γονιδιωμάτων, καθώς και μεγαλύτερων γονιδιωμάτων (English et al., 2012). Ωστόσο, η χαμηλότερη απόδοση σε σχέση με ορισμένες άλλες

τεχνολογίες και το υψηλότερο κόστος αλληλούχισης ανά βάση, περιορίζουν, επί του παρόντος, την εφαρμογή της συγκεκριμένης τεχνολογίας σε ευρείες γονιδιωματικές μελέτες.

Εκτός από την αλληλούχιση reads που είναι μακριές σε μήκος και αμερόληπτες σε ποιότητα ανάλυσης, ένα άλλο διακριτικό χαρακτηριστικό της αλληλούχισης SMRT είναι ότι γίνεται σε πραγματικό χρόνο, επιτρέποντας τη συλλογή δεδομένων που αφορούν τόσο στη σύνθεση των βάσεων, όσο και στην ενζυματική κινητική. Συγκεκριμένα, ξεχωριστά κινητικά προφίλ παράγονται καθώς η πολυμεράση συναντά διάφορους τύπους μεθυλιωμένου DNA (Flusberg et al., 2010). Αυτές οι κινητικές υπογραφές έχουν χρησιμοποιηθεί για τη χαρτογράφηση πιθανών θέσεων μεθυλίωσης με τη μορφή 6-μεθυλαδενίνης στους προκαρυώτες και 5-μεθυλοκυτοσίνης στους ευκαρυώτες (Fang et al., 2012). Είναι πολύ πιθανό ότι αυτές οι προσεγγίσεις θα επεκταθούν για να χαρτογραφήσουν και άλλους τύπους τροποποιήσεων στο DNA, συμπεριλαμβανομένης της βλάβης του DNA που προκαλείται στα καρκινικά κύτταρα. Επιπλέον, οι συσκευές αλληλούχισης που χρησιμοποιούν τη μεθοδολογία SMRT δεν περιορίζονται μόνο για τη μελέτη του DNA, καθώς άλλα μόρια, όπως τα ριβοσώματα, μπορούν να προσδεθούν στον πυθμένα ενός ZMW και να παρακολουθούνται σε μονομοριακή ανάλυση (Uemura et al., 2010).

- **Oxford Nanopore Technologies**

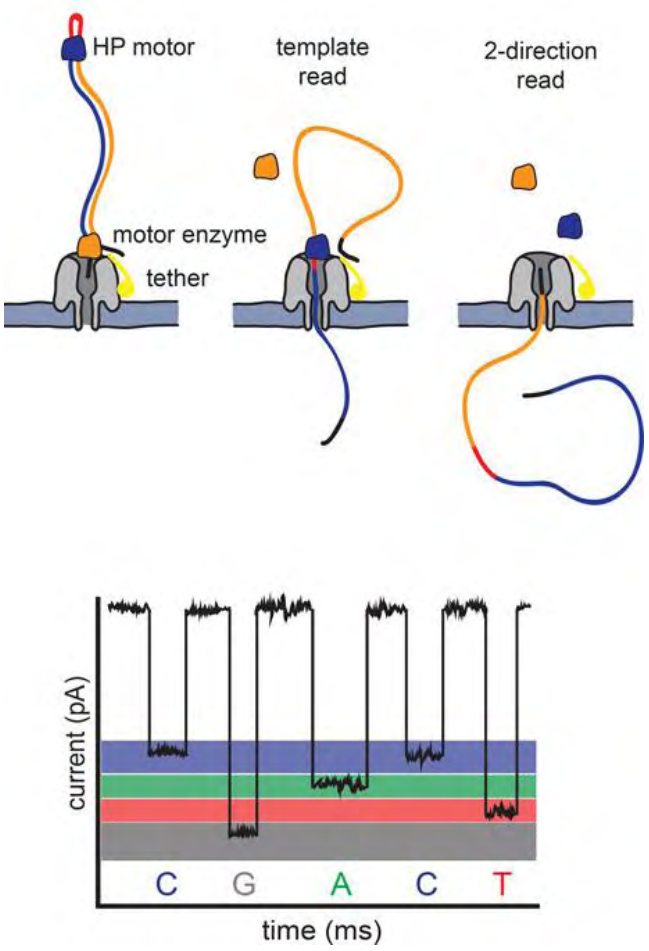
Η εταιρία Oxford Nanopore Technologies ανέπτυξε και εμπορεύεται τη μέθοδο αλληλούχισης «Nanopore». Η αλληλούχιση Nanopore βασίζεται στη μετάβαση του DNA μέσω ενός «νανοπόρου», ο οποίος είναι μια μικροσκοπική οπή, με μορφή διαύλου, της τάξης μεγέθους του νανομέτρου (nm) (Wang et al., 2015). Η εν λόγω μεθοδολογία αποτελεί μια εξελισσόμενη τεχνική αλληλούχισης που έχει σημειώσει σημαντική πρόοδο τα τελευταία χρόνια.

Στις συσκευές της εταιρίας Oxford Nanopore περνά ιοντικό ρεύμα μέσω των νανοπόρων και καταγράφονται οι μεταβολές του, καθώς τα βιολογικά μόρια περνούν δια μέσω αυτών των μικροσκοπικών πόρων. Οι πληροφορίες σχετικά με τη μεταβολή του ρεύματος χρησιμοποιούνται για την αναγνώριση των μορίων που διαπερνούν τον νανοπόρο. Οι οπές των νανοπόρων

μπορούν να δημιουργηθούν σε μεμβράνες που διαπερνώνται είτε από φυσικές πρωτεΐνες (βιολογικοί νανοπόροι), ή από στερεά υλικά (νανοπόροι στερεάς κατάστασης) (Oxford Nanopore Technologies, https://nanoporetech.com/how-it-works).

Η διαδικασία της αλληλούχισης Nanopore λαμβάνει χώρα σε κυψελίδα ροής, κατάλληλα διαμορφωμένη ώστε να περιλαμβάνει εκατοντάδες ανεξάρτητα μικροφρεάτια, που το καθένα περιέχει μια συνθετική μεμβρανική διπλοστοιβάδα διάτρητη από βιολογικούς νανοπόπους. Η αλληλούχιση επιτυγχάνεται μετρώντας τις χαρακτηριστικές μεταβολές στο ρεύμα που προκαλούνται καθώς οι βάσεις περνούν δια μέσω του πόρου, με τη βοήθεια ενός φυσικού ενζύμου που λειτουργεί ως μοριακός «κινητήρας» (motor enzyme). Στη συγκεκριμένη τεχνολογία το υπό αλληλούχιση γονιδιωματικό DNA κατακερματίζεται και στα θραύσματα που προκύπτουν, και που πρόκειται να διαπεράσουν τους νανοπόρους, συνδέονται ειδικοί προσαρμογείς (adapters).

Όπως απεικονίζεται και στο σχήμα της Εικόνας 9, αναφορικά με τους ειδικούς προσαρμογείς, ο πρώτος συνδέεται με το ένζυμο «κινητήρα», καθώς και με ένα μοριακό πρόσδεμα (tether), ενώ ο δεύτερος προσαρμογέας είναι ένα ολιγονουκλεοτίδιο σε σχήμα φουρκέτας που συνδέεται στο DNA από μια δεύτερη πρωτεΐνη «κινητήρα», τη λεγόμενη «HP motor protein» (Quick et al., 2014). Με την εφαρμογή της τεχνολογίας αλληλούχισης Nanopore επιτρέπεται ο προσδιορισμός της αλληλουχίας και των δύο κλώνων ενός μορίου DNA (2-direction reads), γεγονός που αυξάνει την ακρίβεια (Ashton et al., 2014 / Quick et al., 2014).



Εικόνα 9 Τεχνολογία αλληλούχισης Oxford Nanopore. Η αλληλούχιση επιτυγχάνεται μετρώντας τις μεταβολές ρεύματος που προκαλούνται καθώς οι βάσεις περνούν δια μέσω ενός νανοπόρου, με τη βοήθεια ενός φυσικού ενζύμου που λειτουργεί ως μοριακός κινητήρας (motor enzyme). Το υπό αλληλούχιση γονιδιωματικό DNA κατακερματίζεται και στα θραύσματα που πρόκειται να διαπεράσουν τους νανοπόρους συνδέονται ειδικοί προσαρμογείς (adapters). Ένας πρώτος συνδέεται με το ένζυμο κινητήρα, καθώς και με ένα μοριακό πρόσδεμα (tether), ενώ ο δεύτερος προσαρμογέας είναι ένα ολιγονουκλεοτίδιο σε σχήμα φουρκέτας που συνδέεται στο DNA από μια δεύτερη πρωτεΐνη κινητήρα, την HP motor protein.

Από: Reuter JA, Spacek D, Snyder MP. (2015). High-Throughput Sequencing Technologies. Molecular Cell 21; 58(4): 586-597. doi:10.1016/j.molcel.2015.05.004
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4494749/

Στην Εικόνα 10, και συγκεκριμένα στη φωτογραφία α), απεικονίζεται το MinION, η πρώτη εμπορικά διαθέσιμη συσκευή για Nanopore sequencing, η οποία κυκλοφόρησε στις αρχές του 2014 από την Oxford Nanopore Technologies. Είναι η μοναδική φορητή συσκευή αλληλούχισης DNA και RNA. Διαθέτει θήρα σύνδεσης USB για την επικοινωνία του με ηλεκτρονικό υπολογιστή και για να τροφοδοτείται με ενέργεια. Σε μια συνεχή περίοδο

λειτουργίας 18 ωρών μπορεί να παράγει δεδομένα άνω των 90 Mbp από περίπου 16.000 συνολικά reads, με μέσο μήκος των reads περίπου στα 6 kb και με μέγιστο άνω των 60 kb (Ashton et al., 2014). Ωστόσο σε ό,τι αφορά στον αναλυτή MinION, το καταγεγραμμένο ποσοστό σφάλματος από την εφαρμογή του είναι αρκετά υψηλό. Σε μελέτη τους, το 2015, ο Jain και οι συνεργάτες του ανέφεραν ποσοστά προσθηκών, διαγραφών και αντικαταστάσεων σε βάσεις που ανερχόταν σε 4,9%, 7,8% και 5,1%, αντίστοιχα (Jain et al., 2015). Επίσης, παρουσιάζει πολύ υψηλό ποσοστό αποτυχημένων εκτελέσεων. Παρά τα υψηλά ποσοστά σφάλματος, τα δεδομένα των reads που εξήχθησαν με το MinION έχουν χρησιμοποιηθεί επιτυχώς, σε αντιπαραβολή με τα αντίστοιχα reads που λήφθηκαν από την εφαρμογή της τεχνολογίας Illumina, για τον προσδιορισμό της θέσης και της δομής μιας νησίδας βακτηριδιακής ανθεκτικότητας (bacterial resistance island), δηλαδή, ενός γενετικού τόπου που περιέχει συγκεντρωμένα πολλαπλά γονίδια ανθεκτικότητας, (Ashton et al., 2014).

Οι υπόλοιπες συσκευές της εταιρίας Oxford Nanopore Technologies, απεικονίζονται, επίσης, στην Εικόνα 10. Η συσκευή GridIONx5, της φωτογραφίας β), περιλαμβάνει πέντε κυψελίδες ροής για αλληλούχιση Nanopore και, συνεπώς, η χρήση της είναι κατάλληλη για πολλαπλή αλληλούχιση. Στη φωτογραφία γ) απεικονίζεται η συσκευή PromethION, η οποία χαρακτηρίζεται από την εταιρία ως υψηλής απόδοσης περιλαμβάνοντας 48 κυψελίδες ροής. Τέλος, η συσκευή SmidgION, που φαίνεται στη φωτογραφία δ), αναμένεται να είναι η μικρότερη φορητή συσκευή για αλληλούχιση Nanopore, αλλά και για ανάλυση άλλων βιολογικών μορίων. Το εντυπωσιακό είναι ότι η το SmidgION θα συνδέεται απευθείας στο κινητό τηλέφωνο, καθιστώντας εντελώς «φορητές» τις αναλύσεις DNA και των λοιπών βιομορίων για τις οποίες έχει κατασκευαστεί.

Εικόνα 10 Οι αναλυτές της Oxford Nanonopore: α) MinION, β) GridIONx5, γ) PromethION και δ) SmidgION.

Από: Oxford NANOPORE Technologies, Products, https://nanoporetech.com/index.php/products

Θα πρέπει, ωστόσο, να αναφερθεί ότι, τουλάχιστον για το εγγύς μέλλον, δεδομένων των σχετικά υψηλών ποσοστών σφάλματος και της σχετικά χαμηλής της απόδοσης, η αλληλούχιση Nanopore είναι δύσκολο να ξεπεράσει τις σημερινές τεχνολογίες αλληλούχισης. Από την άλλη πλευρά, ο συνδυασμός των προτερημάτων, που περιλαμβάνουν την αλληλούχιση σε πραγματικό χρόνο, την ταχύτητα, το μήκος των reads και το κόστος των συσκευών, υπόσχεται πολλά για το μέλλον (Reuter, et al., 2015). Πάντως, πιο αισιόδοξοι ερευνητές, όπως ο Feng και οι συνεργάτες του, ανέφεραν το 2015, σε δημοσιευμένη τους αναφορά, ότι οι συσκευές αλληλούχισης που βασίζονται στην τεχνολογία των νανοπόρων, θα έχουν τη δυνατότητα να διεξάγουν την αλληλούχιση, γρήγορα και αξιόπιστα, ολόκληρου του

ανθρώπινου γονιδιώματος για λιγότερο από 1000 δολάρια και, ενδεχομένως κάποια στιγμή, για λιγότερο από 100 δολάρια (Feng et al., 2015).

1.7 Συναρμολόγηση δεδομένων αλληλούχισης νέας γενιάς

Κάθε τεχνολογία αλληλούχισης νέας γενιάς διαθέτει ένα χαρακτηριστικό προφίλ σφαλμάτων που περιλαμβάνει σφάλματα στα 3΄ άκρα των reads, μεροληψία υπέρ ή κατά της σύνθεσης των αλληλουχιών που είναι πλούσιες σε GC και τον ανακριβή προσδιορισμό των απλών επαναλήψεων (Huse et al., 2007 / Dohm et al., 2008 / Harismendy et al., 2009).

Γενικά, η διαδικασία της συναρμολόγησης ομαδοποιεί τα reads σε contigs και τα contigs σε scaffolds. Τα contigs παρέχουν μια στοίχιση πολλαπλών αλληλουχιών των reads. Τα scaffolds, τα οποία ονομάζονται και supercontigs ή metacontigs, ορίζουν τη διευθέτηση, τον προσανατολισμό και τα μεγέθη των κενών μεταξύ των contigs. Τα περισσότερα προγράμματα συναρμολόγησης (assemblers) παράγουν, επιπλέον, ένα σύνολο μη συναρμολογημένων ή μερικώς συναρμολογημένων reads. Η πιο αποδεκτή μορφή αρχείου δεδομένων για μια συναρμολόγηση είναι η FASTA, όπου η συναινετική (consensus) αλληλουχία των contigs μπορεί να αναπαρασταθεί ως μια σειρά των χαρακτήρων A, C, G, T, συμπεριλαμβανομένων, ενδεχομένως, και άλλων χαρακτήρων με ιδιαίτερο νόημα. Για παράδειγμα, οι παύλες, μπορούν να αντιπροσωπεύουν επιπλέον βάσεις που παραλείπονται από τη συναινετική αλληλουχία, αλλά εμφανίζονται σε μια μειοψηφία των reads. Η συναινετική αλληλουχία του scaffold μπορεί να περιλαμβάνει το σύμβολο «N» στα κενά μεταξύ των contigs. Ο αριθμός των διαδοχικών N υποδεικνύει την εκτίμηση του μήκους των κενών με βάση τη σύνδεση των ζευγαρωμένων άκρων.

Οι συναρμολογήσεις μετρώνται από το μέγεθος και από την ακρίβεια των contigs και των scaffolds. Το μέγεθος της συναρμολόγησης δίδεται, συνήθως, από στατιστικά στοιχεία που περιλαμβάνουν το μέγιστο μήκος, το μέσο μήκος, το συνδυασμένο συνολικό μήκος και το N50. Το N50 είναι μια μέτρηση αξιολόγησης της ποιότητας της συναρμολόγησης, καθώς ένα

υπερβολικά μικρό N50 υποδηλώνει ότι δε μπορούν να δημιουργηθούν πολλά contigs που να έχουν μέγεθος με κάποια βιολογική σημασία. Είναι, δηλαδή, πιθανό να υπάρχουν πολλά ψευδή μικρά contigs στη συναρμολόγηση. Γενικά, το N50 είναι το διάμεσο μέγεθος της συναρμολόγησης του εξεταζόμενου γονιδιώματος (Dan Spiegelman, Université de Montréal).

Ειδικότερα, δεδομένου ενός συνόλου από contigs, καθένα με το δικό του μήκος, το μήκος N50 ορίζεται ως το βραχύτερο μήκος αλληλουχίας στο 50% του γονιδιώματος. Για παράδειγμα, εξετάζονται 9 contigs με μήκη 2, 3, 4, 5, 6, 7, 8, 9 και 10 βάσεις, αντίστοιχα. Το άθροισμά τους είναι ίσο με 54, το ήμισυ του αθροίσματος ισούται με 27 και το μέγεθος του γονιδιώματος υποτίθεται ότι έχει μήκος 54 βάσεις. Το 50% αυτής της συναρμολόγησης θα είναι 10 + 9 + 8 = 27 (το μισό μήκος της αλληλουχίας). Έτσι, το N50 είναι ίσο με 8, το οποίο είναι το μέγεθος του contig που, μαζί με τα μεγαλύτερα contigs, περιέχει το ήμισυ της αλληλουχίας ενός συγκεκριμένου γονιδιώματος. Σημειώνεται ότι όταν συγκρίνονται οι τιμές N50 από διαφορετικές συναρμολογήσεις, τα μεγέθη συναρμολόγησης πρέπει να έχουν το ίδιο μέγεθος για να έχει νόημα το N50 (N50, L50, and related statistics, Wikipedia). Το N50 μπορεί να αυξηθεί εξαλείφοντας τις αλληλουχίες που είναι πιθανό να προκαλέσουν προβλήματα, π.χ. τα μικρά επαναλαμβανόμενα τμήματα. Σημειώνεται ότι αυτή η μέτρηση ισχύει μόνον όταν διεξάγεται *de novo* συναρμολόγηση. Αν γίνεται στοίχιση με ένα γονιδίωμα αναφοράς, αυτή η μέτρηση δεν ισχύει.

Η συναρμολόγηση της πλήρους αλληλουχίας ενός γονιδιώματος διαταράσσεται από τη μη ομοιόμορφη κάλυψη της αλληλουχίας στόχου. Η μεταβλητότητα της κάλυψης εισάγεται τυχαία, από τη διακύμανση του κυτταρικού αριθμού των αντιγράφων μεταξύ των μορίων του DNA της πηγής και από τη μεροληψία των τεχνολογιών αλληλούχισης. Πολύ χαμηλή κάλυψη προκαλεί κενά στις συναρμολογήσεις. Επιπλέον, η μεταβλητότητα στην κάλυψη ακυρώνει τις στατιστικές δοκιμές και υπονομεύει τα διαγνωστικά που βασίζονται στην κάλυψη.

Έτσι λοιπόν, η διαδικασία της συναρμολόγησης της πλήρους αλληλουχίας ενός γονιδιώματος περιπλέκεται από την υπολογιστική πολυπλοκότητα της επεξεργασίας του μεγάλου όγκου των δεδομένων που

παράγονται ως αποτέλεσμα της υψηλής απόδοσης των τεχνολογιών αλληλούχισης νέας γενιάς. Για να αντιμετωπιστεί αυτό το πρόβλημα, καθώς και το πρόβλημα που ανακύπτει από το διαφορετικό μήκος που έχουν μεταξύ τους τα reads που προκύπτουν από τις τεχνολογίες αλληλούχισης νέας γενιάς, στην ανάλυση που διεξάγεται με τις αλγοριθμικές μεθοδολογίες των προγραμμάτων συναρμολόγησης, εισήχθη η έννοια του k-μερούς (k-mer), καθώς στη διαδικασία της *de novo* συναρμολόγησης τα k-μερή χρησιμοποιούνται για την κατασκευή γραφημάτων αλληλοεπικάλυψης (overlap graphs). Γενικά, ως k-μερή αναφέρονται όλες οι πιθανές «υποαλληλουχίες» (subsequences), μήκους k, που προκύπτουν από μια αλληλουχία read που λαμβάνεται μέσω αλληλούχισης του DNA.

Τα περισσότερα αλγοριθμικά προγράμματα της *de novo* συναρμολόγησης των δεδομένων αλληλούχισης νέας γενιάς βασίζονται στη θεωρία των γραφημάτων, στα οποία κόμβοι (nodes) αναπαριστούν τα reads, ή τα σχηματιζόμενα από τα reads k-μερή και ακμές (edges) αναπαριστούν τις αλληλοεπικαλύψεις τους. Τα contigs αποτελούν τις διαδρομές που σχηματίζονται στο γράφημα αλληλοεπικαλύψης (Myers, 1995). Οι αλληλοεπικαλύψεις υπολογίζονται με μια σειρά από, υπολογιστικά δαπανηρές, στοιχίσεις αλληλουχιών ανά ζεύγη. Οι κόμβοι καλούνται και ως κορυφές, ενώ οι ακμές αναφέρονται και ως γραμμές. Το γράφημα δύναται να έχει ξεχωριστά στοιχεία για να διακρίνονται τα 5΄ και 3΄ άκρα των reads ή των αντίστοιχων k-μερών τους, οι εμπρόσθιες (forward) και ανάστροφες (reverse) συμπληρωματικές τους αλληλουχίες, το μήκος τους, το μήκος και ο τύπος της αλληλοεπικάλυψής τους (k-mer, Wikipedia).

Προκειμένου να δημιουργηθεί ένα γράφημα αλληλοεπικάλυψης, όπως για παράδειγμα ένα γράφημα «de Bruijn», οι αλληλουχίες, μήκους L, που αποθηκεύονται σε κάθε ακμή, πρέπει να αλληλοεπικαλύπτουν μία άλλη αλληλουχία σε μία άλλη ακμή, με αλληλοεπικάλυψη L-1, ώστε να δημιουργηθεί ένας κόμβος. Ο όρος «L-1» αναφέρεται στη διαδοχική αλληλοεπικάλυψη που λαμβάνει χώρα κατά μία βάση λιγότερο.

΄Ενα κλασικό πρόβλημα με τη *de novo* συναρμολόγηση είναι η εύρεση μιας διαδρομής στο γράφημα αλληλοεπικαλύψεων που περνάει από κάθε έναν από τους κόμβους μόνο μία φορά (Χαμιλτονιανή διαδρομή), ή από κάθε

ακμή μόνο μία φορά (διαδρομή Euler). Αυτό, συχνά, έχει ως αποτέλεσμα την απώλεια της σύνδεσης μεταξύ πολύ απομακρυσμένων αλληλουχιών, υποδεικνύοντας ότι οι μεθοδολογίες που βασίζονται σε γραφήματα, ειδικά η μέθοδος de Bruijn, είναι εξαιρετικά ευαίσθητες σε σφάλματα αλληλούχισης (Pevzner και Tang, 2001). Μελέτες δείχνουν ότι μέχρι το 96,29% ενός γονιδίου μπορεί να ανασκευαστεί χρησιμοποιώντας βραχείες αλληλουχίες μήκους από 25 νουκλεοτίδια (Kingsford et al., 2010).

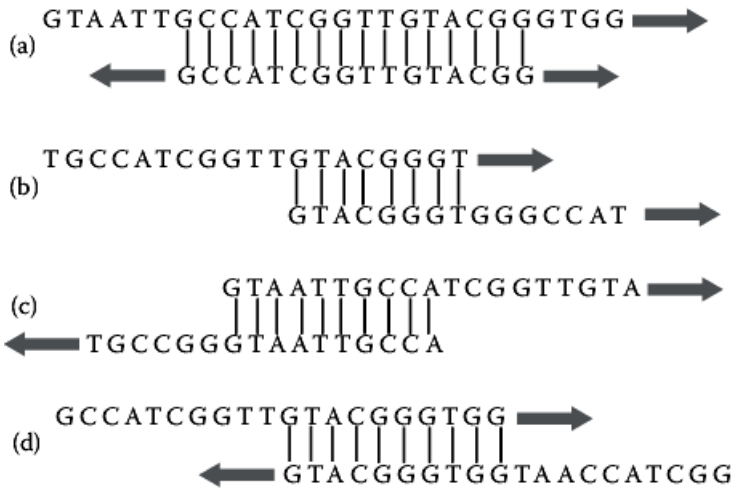
1.8 Κατηγορίες αλγορίθμων συναρμολόγησης

Οι αλγόριθμοι συναρμολόγησης και οι εφαρμογές τους που αφορούν στην πλήρη αλληλούχιση ενός γονιδιώματος είναι, συνήθως, πολύπλοκοι. Η αλγοριθμική επιτυχία εξαρτάται από ευρετικές μεθόδους, δηλαδή, από εμπειρικούς κανόνες. Οι ευρετικές μέθοδοι βοηθούν να ξεπεραστούν τα περίπλοκα μοτίβα επαναλήψεων, το τυχαίο και συστηματικό σφάλμα στα δεδομένα αλληλούχισης και οι φυσικοί περιορισμοί των υπολογιστών. Τα προγράμματα για τη *de novo* συναρμολόγηση μπορούν να κατηγοριοποιηθούν ανάλογα με τους αλγορίθμους που χρησιμοποιούν στις εξής τρεις κατηγορίες:

- **Overlap/Layout/Consensus (OLC)**,
- **de Bruijn Graph (DBG)** και
- **Greedy Graph**.

Η κατηγορία των αλγορίθμων που χρησιμοποιούν τη μεθοδολογία OLC είναι η πιο χρησιμοποιούμενη για μεγάλες αλληλουχίες. Ωστόσο, υπάρχουν και εφαρμογές που βασίζονται σε αυτή τη μέθοδο για σύντομα reads, όπως η Edena (Hernandez et al., 2008). Η μέθοδος OLC μπορεί να χωριστεί σε τρεις φάσεις: αλληλοεπικάλυψη (overlap), διάταξη (layout) και συναίνεση (consensus). Στη φάση της αλληλοεπικάλυψης, κάθε read, ή το αντίστοιχο k-μερές, συγκρίνεται με όλα τα άλλα για να προσδιοριστούν οι αλληλοεπικαλύψεις, λαμβάνοντας υπόψη το ελάχιστο μέγεθος αλληλοεπικάλυψης, η οποία θα επηρεάσει την ακρίβεια των contigs. Καταγράφονται τέσσερις τύποι αλληλοεπικαλύψεων οι οποίοι απεικονίζονται

στην Εικόνα 11 και είναι οι εξής: (a) συνοχής (containment), (b) μερικής αλληλοεπικάλυψης (partial overlap), (c) προθέματος (prefix overlap) και (d) κατάληξης (suffix overlap).



Εικόνα 11 Οι τύποι αλληλοεπικαλύψεων των reads: (a) συνοχής (containment), (b) μερικής αλληλοεπικάλυψης (partial overlap), (c) προθέματος (prefix overlap) και (d) κατάληξης (suffix overlap).

Από: Next-Generation Sequencing and Assembly of Bacterial Genomes
https://www.researchgate.net/publication/236610137_Next-Generation_Sequencing_and_Assembly_of_Bacterial_Genomes

1.9 Η επιδημία χολέρας στην Αϊτή. Αλληλούχιση νέας γενιάς και ανάλυση μικροβιακών γονιδιωμάτων τοξικολογικού ενδιαφέροντος.

Στις 12 Ιανουαρίου 2010, ένας καταστροφικός σεισμός μεγέθους 7,0 βαθμών της κλίμακας Ρίχτερ χτύπησε την Αϊτή, επηρεάζοντας 3.500.000 ανθρώπους (United States Geological Survey, 2010) / Farmer, 2013). Το γεγονός αυτό έπληξε σοβαρά ένα ήδη υποβαθμισμένο δημόσιο σύστημα υδροδότησης και αποχέτευσης, δημιουργώντας ιδανικές συνθήκες για την εκδήλωση σοβαρών μολυσματικών ασθενειών. Τον Οκτώβριο του 2010, εννέα μήνες μετά τον σεισμό, εκδηλώθηκε η λοιμώδης νόσος της χολέρας, η οποία εξαπλώθηκε γρήγορα σε όλη τη χώρα (Delva, 2010). Μέχρι και τις 7 Ιανουαρίου του 2014,

είχαν αναφερθεί από το Υπουργείο Δημόσιας Υγείας και Πληθυσμού της Αϊτής 8.534 θάνατοι και 697.256 μολυσματικές περιπτώσεις (Ministère de la Santé Publique et de la Population, 2014) (Εικόνα 12).

Εικόνα 12 Απολογισμός από την επιδημία της χολέρας στην Αϊτή: 8.534 θάνατοι και 697.256 μολυσματικές περιπτώσεις.

Από: Haiti: The Importance Of Communicating About Cholera, http://www.doctorswithoutborders.org/news-stories/field-news/haiti-importance-communicating-about-cholera

Πριν από το 2010, δεν είχε αναφερθεί ιστορικό χολέρας στην Αϊτή, παρά τις καταστροφικές συνέπειες της νόσου στην περιοχή της Καραϊβικής τον 19ο αιώνα (Jenson et al., 2011). Συνεπώς, πολλοί είχαν αναρωτηθεί από πού θα μπορούσε να είχε προέλθει η χολέρα στην Αϊτή και έτσι προέκυψαν δύο υποθέσεις σχετικά με την προέλευσή της. Η κλιματολογική υπόθεση υποστήριξε ότι το μη παθογόνο βακτήριο *Vibrio cholerae*, που ενδημεί στα παράκτια ύδατα της Αϊτής, υπό τις κατάλληλες περιβαλλοντικές συνθήκες, εξελίχθηκε σε ένα παθογόνο στέλεχος (Parker, 2010). Από την άλλη πλευρά, η υπόθεση της ανθρώπινης μετάδοσης ανέφερε ότι η χολέρα εισήχθη στην Αϊτή από ανθρώπους που είχαν μολυνθεί σε ξένη χώρα.

Η χολέρα είναι μια οξεία διαρροϊκή νόσος που οφείλεται στην προσβολή του εντέρου από την εντεροτοξίνη που παράγει το βακτήριο *Vibrio*

cholerae και μπορεί να οδηγήσει σε θάνατο σε λιγότερο από 48 ώρες εάν αφεθεί χωρίς θεραπεία. Η κατανάλωση μολυσμένου νερού είναι η κύρια αιτία για την ανθρώπινη μόλυνση. Η ισχυρή τοξίνη της χολέρας κωδικοποιείται από τα γονίδια *ctxAB* (Waldor et al., 1996) που βρίσκεται στο γονιδίωμα βακτηριοτοξικών στελεχών του *Vibrio cholerae*. Η τοξίνη, μαζί με άλλους λοιμογόνους παράγοντες, οδηγεί στις επιβλαβείς επιδράσεις της μόλυνσης. Αυτοί οι βοηθητικοί παράγοντες μολυσματικότητας κωδικοποιούνται σε γονιδιωματικές νησίδες, οι οποίες αποκτώνται μέσω οριζόντιας γονιδιακής μεταφοράς (de la Cruz και Davies, 2000). Το βακτήριο *Vibrio cholerae* αποτελεί σοβαρό υγειονομικό πρόβλημα εξαιτίας της πρόκλησης εκτεταμένων πανδημιών. Η χολέρα εξαλείφθηκε από τις βιομηχανοποιημένες χώρες, μέσω των αποτελεσματικών επεξεργασιών του πόσιμου νερού και των αστικών λυμάτων. Ωστόσο, αποτελεί πολύ σοβαρή απειλή για τις λιγότερο αναπτυγμένες χώρες που διαθέτουν κακό δίκτυο υδροδότησης και αποχέτευσης.

Στις 27 Οκτωβρίου 2010, η ευθύνη για την επιδημία της χολέρας αποδόθηκε στο προσωπικό της αποστολής των Ηνωμένων Εθνών στην Αϊτή που είχε καταφθάσει από το Νεπάλ και η οποία είχε πρόσφατα εγκαταστήσει στρατόπεδο στο Meille, ένα μικρό χωριό 2 χλμ. νότια της τοποθεσίας Mirebalais της Αϊτής. Η αποκάλυψη του γεγονότος έγινε από δημοσιογράφους που επέδειξαν τις ακατάλληλες συνθήκες απορροής των λυμάτων του στρατοπέδου (Katz, 2010 / Al Jazeera English, 2010).

Τα δείγματα κοπράνων που συλλέχθηκαν από ασθενείς με χολέρα στην αρχή της εκδήλωσης της επιδημίας, απεστάλησαν στα Κέντρα Ελέγχου και Πρόληψης Νοσημάτων (CDC) των Η.Π.Α. για ανάλυση. Στις 13 Νοεμβρίου, τα CDC ανέφεραν ότι το στέλεχος του βακτηρίου *Vibrio cholerae* El Tor O1 απομονώθηκε από τα δείγματα υποδηλώνοντας ότι ένα μόνο στέλεχος προκάλεσε την εκδήλωση και το οποίο, πιθανότατα, εισήχθη στην Αϊτή με αφορμή ενός γεγονότος (Centers for Disease Control and Prevention, 2010).

Σε μελέτη τους ο Piarroux και οι συνεργάτες του χρησιμοποίησαν όλα τα διαθέσιμα επιδημιολογικά δεδομένα, ελέγχοντας τα νοσοκομειακά αρχεία, διεξάγοντας επιτόπιες έρευνες και εφαρμόζοντας στατιστική έρευνα

χωροχρονικής ανάλυσης, με σκοπό να εντοπίσουν την πηγή και την εξάπλωση της εστίας (Piarroux et al., 2011). Τα ευρήματα της έρευνάς τους επιβεβαίωσαν τους ισχυρισμούς των δημοσιογράφων. Με βάση όλα τα συγκεντρωμένα στοιχεία, συνέταξαν το εξής πιθανό σενάριο: το στρατόπεδο της αποστολής των Ηνωμένων Εθνών μόλυνε έναν παραπόταμο του Meille με περιττωματική ύλη προερχόμενη από τις ανθυγιεινές πρακτικές απορροής των λυμάτων (Εικόνα 13).

Εικόνα 13 Απόδοση ευθυνών για την έξαρση της επιδημίας της χολέρας στην Αϊτή. «Το στρατόπεδο της αποστολής των Ηνωμένων Εθνών μόλυνε έναν παραπόταμο του Meille με περιττωματική ύλη προερχόμενη από τις ανθυγιεινές πρακτικές απορροής των λυμάτων.» (Piarroux et al., 2011).

Από: Coordinating Committee for International Staff Unions and Associations, LEAKED UN REPORT ON HAITI'S CHOLERA OUTBREAK SLAMMED SANITATION AT ITS BASES, http://www.ccisua.org/2016/04/07/leaked-un-report-on-haitis-cholera-outbreak-slammed-sanitation-at-its-bases/

Ο παραπόταμος του Meille συνδέεται με τον ποταμό Latem που διέρχεται από την πόλη του Mirebalais, την τοποθεσία του πρώτου αναφερθέντος περιστατικού της χολέρας (Ivers και Walton, 2012). Ο ποταμός Latem συνδέεται με τη σειρά του με τον ποταμό Artibonite, τον μακρύτερο και τον σημαντικότερο ποταμό που εκτείνεται στην Αϊτή. Η μετακίνηση και η εξάπλωση της χολέρας στην πρώιμη έναρξη της επιδημίας ήταν στενά συνδεδεμένες με την εγγύτητα με τον ποταμό Artibonite.

Επιπλέον, είχε αναφερθεί ότι το Κατμαντού, η πρωτεύουσα του Νεπάλ, όπου τα στρατεύματα εκπαιδεύτηκαν για μικρό χρονικό διάστημα πριν εγκατασταθούν στην Αϊτή, αντιμετώπισε επιδημία χολέρας στις 23 Σεπτεμβρίου (Maharjan, 2010). Τα πρώτα στρατεύματα έφτασαν στην Αϊτή στις 8 Οκτωβρίου (Lantagne et al., 2014) και το πρώτο περιστατικό χολέρας αναφέρθηκε στις 12 Οκτωβρίου (Ivers και Walton, 2012). Επειδή, κανένα από τα στρατεύματα δεν εμφάνισε συμπτώματα χολέρας κατά τη διάρκεια της ιατρικής εξέτασης πριν από την εγκατάσταση στην Αϊτή, ο επικεφαλής ιατρός της αποστολής των Ηνωμένων Εθνών, αργότερα, αποκάλυψε ότι δε διενεργήθηκαν ειδικές δοκιμές παρακολούθησης (BBC News, 2010). Ωστόσο, η απουσία των συμπτωμάτων δεν αποδείκνυε ότι τα στρατεύματα δεν είχαν μολυνθεί από το βακτήριο *Vibrio cholerae*, καθώς θα μπορούσαν να είχαν μολυνθεί τις ημέρες μετά την ιατρική εξέταση και πριν από την εγκατάστασή τους στην Αϊτή, ή θα μπορούσαν να ήταν ασυμπτωματικοί φορείς (Benčić και Sinha, 1972 / Piarroux et al., 2011). Δυστυχώς, εκτός από αυτά που συνέβησαν από την αποστολή των Ηνωμένων Εθνών, δεν έγινε ανεξάρτητος έλεγχος των στρατευμάτων για επιβεβαίωση της παρουσίας ή απουσίας του βακτηρίου.

Η πρώτη μοριακή μελέτη σχετικά με την προέλευση του *Vibrio cholerae* στην Αϊτή δημοσιεύθηκε στις 9 Δεκεμβρίου 2010 (Chin et al., 2011). Ο Chin και οι συνεργάτες του διενήργησαν την πλήρη γονιδιωματική αλληλούχιση των βακτηριακών προϊόντων απομόνωσης H1 και H2 της Αϊτής, τα οποία προέκυψαν από την έξαρση της νόσου, καθώς και των επιδημικών στελεχών των βακτηριακών προϊόντων απομόνωσης: C6 από την επιδημία του 1991 στο Περού, M4 από την επιδημία του 2008 στο Μπαγκλαντές και N5 από την επιδημία του 1971 στο Μπαγκλαντές. Από τη σύγκριση των μονο-νουκλεοτιδικών παραλλαγών, καθώς και των υπερμεταβλητών χρωμοσωμικών στοιχείων στα αναλυθέντα γονιδιώματα, αποδείχθηκε ότι και τα δύο βακτηριακά προϊόντα απομόνωσης της Αϊτής ήταν γενετικά πανομοιότυπα και ήταν περισσότερο όμοια με το στέλεχος M4 από το Μπαγκλαντές παρά με το στέλεχος C6 από το Περού. Παρόλο που τα αποτελέσματα προέκυψαν από ένα μικρό μέγεθος δειγμάτων, συμφωνούσαν με μια κλωνική πηγή για την εκδήλωση της νόσου. Η μελέτη του Chin και των

συνεργατών του απέδειξε ότι η χολέρα εισήχθη στην Αϊτή με ανθρώπινη μετάδοση από μια μακρινή γεωγραφική πηγή, πιθανότατα από τη Νότια Ασία και μάλλον από το Μπαγκλαντές, αν και τα συμπεράσματά τους βασίστηκαν σε μια πολύ περιορισμένη ανάλυση στελεχών από τον Αϊτινό και από τον παγκόσμιο πληθυσμό.

Δύο μεταγενέστερες και μεγαλύτερες γονιδιωματικές μελέτες χρησιμοποίησαν 23 (Reimer et al., 2011) και 154 (Mutreja et al., 2011) αλληλουχίες ολόκληρου του γονιδιώματος για να τεκμηριώσουν την επαναλαμβανόμενη ιστορική εξάπλωση του στελέχους *Vibrio cholerae* O1 από τη Νότια Ασία. Αυτές οι μελέτες χρησιμοποίησαν μέχρι και εννέα, επιπλέον, βακτηριακά προϊόντα απομόνωσης της Αϊτής και τα εξέτασαν στο πλαίσιο της διευρυμένης γονιδιωματικής συλλογής των βακτηριακών στελεχών του *Vibrio cholerae*. Βρέθηκε φυλογενετική συγγένεια μεταξύ των στελεχών της Αϊτής του 2010 και αυτών που παρατηρήθηκαν τα προηγούμενα χρόνια από το Καμερούν, το Μπαγκλαντές, την Ινδία και το Πακιστάν. Τα προϊόντα απομόνωσης της Αϊτής ήταν σχεδόν πανομοιότυπα και συμφωνούσαν και πάλι με μία μεμονωμένη κλωνική εστία. Ωστόσο, βακτηριακά στελέχη του έτους 2010 από το Νεπάλ δε συμπεριλήφθηκαν σε αυτές τις μελέτες και η γονιδιωματική σχέση μεταξύ των στελεχών της Αϊτής και του Νεπάλ δε διαφοροποιήθηκε από άλλες περιοχές της Νότιας Ασίας ή ακόμη και από την Αφρική.

Η πρώτη μελέτη που περιλάμβανε στελέχη *Vibrio cholerae* από το Νεπάλ δημοσιεύθηκε από τον Hendriksen και τους συνεργάτες του στις 23 Αυγούστου 2011 (Hendriksen et al., 2011). Συγκρίθηκαν τα γονιδιώματα 24 στελεχών που απομονώθηκαν από πέντε γεωγραφικές περιοχές του Νεπάλ, κατά τη χρονική περίοδο μεταξύ 30 Ιουλίου και 1 Νοεμβρίου 2010, με δέκα γονιδιώματα του βακτηρίου, συμπεριλαμβανομένων τριών από την Αϊτή. Όλα τα στελέχη από το Νεπάλ, την Αϊτή και το Μπαγκλαντές συγκεντρώθηκαν μαζί σε μία μονοφυλετική ομάδα. Μοιράζονταν, δηλαδή, έναν κοινό πρόγονο. Ωστόσο, το πιο σημαντικό ήταν ότι τα τρία στελέχη της Αϊτής και τα τρία του Νεπάλ σχημάτισαν μια πολύ στενή υποομάδα και αυτά ήταν σχεδόν πανομοιότυπα μεταξύ τους, με μόνο μία ή δύο νουκλεοτιδικές διαφορές στο γονιδίωμα του πυρήνα τους. Η μελέτη αυτή, σε συνδυασμό με την κλασική

επιδημιολογία (Piarroux et al., 2011 / Lantagne et al., 2013) έδειξε πειστικά στοιχεία ότι η χολέρα εισήχθη στην Αϊτή από μια εξωτερική πηγή, με το Νεπάλ να είναι η πιο πιθανή προέλευση.

Οι έρευνες για την εκδήλωση της επιδημίας της χολέρας στην Αϊτή έδειξαν πώς οι παραδοσιακές επιδημιολογικές έρευνες μπορούν να ενισχυθούν σημαντικά με τη γονιδιωματική αλληλούχιση και τη φυλογενετική ανάλυση. Οι γονιδιωματικές αναλύσεις εξασφάλισαν πολύ ισχυρές αποδείξεις για τη στήριξη της υπόθεσης της ανθρώπινης μετάδοσης. Αυτές οι αναλύσεις, λόγω της ικανότητάς τους να ανιχνεύουν ελάχιστες διαφορές μεταξύ των διαφορετικών στελεχών, επέτρεψαν, επίσης, τον προσδιορισμό της ακριβούς πηγής της εστίας. Αυτή η διαπίστωση μπορεί να βοηθήσει πάρα πολύ στην πρόληψη μελλοντικών εστιών μόλυνσης και στις αντίστοιχες νομικές επιπτώσεις από την εκδήλωση αυτών των εστιών σε σχέση με τα αίτια πρόκλησής τους. Σε μια πρόσφατη εξέλιξη, έχει ασκηθεί αγωγή κατά των Ηνωμένων Εθνών στο ομοσπονδιακό δικαστήριο των Ηνωμένων Πολιτειών για ζημίες που προκλήθηκαν από την εκδήλωση της χολέρας στην Αϊτή (Gladstone, 2013). Βέβαια, με τις δυνατότητες της νομικής δράσης, οι μέθοδοι γονιδιωματικής ανάλυσης πρέπει να είναι πολύ αυστηρές, καθώς το δικαστικό σύστημα θα απαιτήσει αυστηρά πρότυπα για να τα δεχτεί ως αποδεικτικά στοιχεία.

Ο προσδιορισμός της προέλευσης της επιδημίας της χολέρας στην Αϊτή περιορίστηκε από δύο βασικούς παράγοντες. Πρώτον, η αλληλούχιση του γονιδιώματος χρησιμοποιήθηκε μόνο στα μεταγενέστερα στάδια της έρευνας και όχι ως μέθοδος «πρώτης γραμμής» για τον εντοπισμό των μολυσματικών παραγόντων. Εισερχόμαστε σε μια εποχή όπου ο γονιδιωματικός, ή ακόμη και ο μεταγονιδιωματικός έλεγχος, θα γίνει μέρος της ιατρικής διάγνωσης. Ο δεύτερος περιορισμός ήταν η απουσία δημόσιας βάσης δεδομένων που να περιείχε αρκετές γονιδιωματικές αλληλουχίες παθογόνων από διάφορες γεωγραφικές τοποθεσίες. Για αυτές τις δύο διαπιστώσεις καταβάλλονται σημαντικές προσπάθειες με σκοπό την εξάλειψη και των δύο αναφερόμενων περιορισμών (Aarestrup et al., 2012). Πάντως, στις μέρες μας είναι διαθέσιμες οι τεχνολογίες για την πλήρη αλληλούχιση ενός γονιδιώματος και για την εκτέλεση των βασικών αναλύσεων

γονοτύπησης και ταυτοποίησης εντός 24 ωρών (Heger, 2013). Με τη μείωση του κόστους της αλληλούχισης, οι ολοκληρωμένες και γεωγραφικά ενημερωμένες βάσεις δεδομένων των αλληλουχιών του γονιδιώματος των παθογόνων μπορούν σύντομα να είναι πραγματικότητα.

Όπως συμπεραίνεται από το περιγραφόμενο παράδειγμα της επιδημίας της χολέρας στην Αϊτή το 2010, η ανάγκη για ακριβή και ταχεία ταυτοποίηση των μολυσματικών παραγόντων τόσο σε μεμονωμένες ασθένειες όσο και σε παγκόσμιες απειλές για τη δημόσια υγεία φαίνεται να υπόκειται σε περιορισμούς στον χρόνο και στην ακρίβεια της ταυτοποίησης, περιορισμοί οι οποίοι δύνανται να οδηγήσουν σε καταστροφικές συνέπειες για τη δημόσια υγεία. Οι τρέχουσες μέθοδοι παρέχουν, συνήθως, δεδομένα σε χρονικό διάστημα που δεν επιτρέπει τις πιο αποτελεσματικές, ειδικές και άμεσες επιλογές θεραπείας. Αυτό οδηγεί σε αντιβιοτικές θεραπείες που είναι, συχνά, ευρείες και βασισμένες σε συμπτώματα ή σε μη επιβεβαιωμένες δοκιμαστικές διαγνώσεις. Αυτή η προσέγγιση αυξάνει, γενικότερα, τον κίνδυνο ανθεκτικότητας στα αντιβιοτικά.

ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

2.1 Το Linux και το Ubuntu

Το Linux είναι ένα λειτουργικό σύστημα που αποτελείται από ελεύθερο λογισμικό. Η χρήση του είναι παρόμοια με αυτή του Unix, αλλά ο πηγαίος κώδικάς του έχει γραφτεί από την αρχή ως ελεύθερο λογισμικό υπό την ελεύθερη άδεια χρήσης GNU General Public License.

Το Linux μπορεί να εγκατασταθεί και να λειτουργήσει σε μεγάλη ποικιλία υπολογιστικών συστημάτων, από μικρές συσκευές, όπως κινητά τηλέφωνα, μέχρι μεγάλα υπολογιστικά συστήματα και υπερυπολογιστές. Η κύρια διαφορά μεταξύ του Linux και άλλων δημοφιλών λειτουργικών είναι ότι ο πυρήνας του Linux, καθώς και οι σημαντικότερες εφαρμογές του, αναπτύσσονται μαζικά και συμμετοχικά από τις κοινότητες των χρηστών, μιας και πρόκειται για ελεύθερο και ανοικτού κώδικα λογισμικό (Linux, Βικιπαίδεια).

Το Ubuntu είναι ένα ανοικτού κώδικα, ελεύθερο και δωρεάν λειτουργικό σύστημα βασισμένο στον πυρήνα Linux. Στόχος του Ubuntu είναι η παροχή ενός διαρκώς ενημερωμένου, σταθερού λειτουργικού συστήματος για τον μέσο χρήστη, με ενισχυμένη έμφαση στην ευκολία χρήσης και εγκατάστασης. Το Ubuntu έχει χαρακτηριστεί ως η πιο δημοφιλής διανομή Linux για επιτραπέζιους υπολογιστές. Η εφαρμογή εγκατάστασης Ubiquity επιτρέπει την εγκατάσταση του Ubuntu στο σκληρό δίσκο από ένα περιβάλλον Live DVD ή Live USB, χωρίς να χρειάζεται η επανεκκίνηση του συστήματος για την εγκατάσταση.

Το Ubuntu διατίθεται ελεύθερα στον σύνδεσμο:

https://www.ubuntu.com/download/desktop

καθώς και στον σύνδεσμο:

http://www.ubuntu-gr.org/download/

Στην παρούσα εργασία χρησιμοποιήθηκε η έκδοση Ubuntu 14.04.5 LTS (Trusty Tahr), η οποία διατίθεται ελεύθερα από τον σύνδεσμο:

http://releases.ubuntu.com/14.04/

2.2 Λήψη δεδομένων των reads από τη βάση δεδομένων SRA

Για τις δοκιμές της παρούσας μελέτης χρησιμοποιήθηκαν τα δεδομένα των αλληλουχιών reads ολόκληρης της γονιδιωματικής αλληλουχίας των μικροοργανισμών *Bacillus anthracis*, *Bacillus subtilis* και *Vibrio cholerae*, με σκοπό την ανάπτυξη ενός βιοπληροφορικού πρωτοκόλλου για την ανάλυση μικροβιακών γονιδιωμάτων τοξικολογικού και εγκληματολογικού ενδιαφέροντος με τα δεδομένα των τεχνολογιών Illumina και PacBio.

Χρησιμοποιήθηκε η βάση δεδομένων «Sequencing Read Archive (SRA)» όπου αναζητήθηκαν και λήφθηκαν τα δεδομένα των reads των υπό μελέτη μικροβιακών γονιδιωματικών αλληλουχιών, ανάλογα με την επιθυμητή μέθοδο αλληλούχισης με την οποία προέκυψαν. Τα δεδομένα των reads κατέβηκαν και σώθηκαν στη μορφή FASTQ.

2.3 Φιλτράρισμα των δεδομένων με το πρόγραμμα ConDeTri

Χρησιμοποιήθηκε η έκδοση v2.2 του προγράμματος ConDeTri για το ποιοτικό φιλτράρισμα των reads που προέρχονται από την τεχνολογία Illumina. Το ConDeTri αποκόπτει και αφαιρεί (trimming) ολόκληρα reads ή τμήματά τους, χρησιμοποιώντας τις βαθμολογίες ποιότητάς τους (Q-scores).

2.4 Ανάλυση ποιότητας δεδομένων με το πρόγραμμα FastQC

Χρησιμοποιήθηκε η έκδοση v0.11.5 του προγράμματος FastQC για τον ποιοτικό έλεγχο που μπορεί να εντοπίσει προβλήματα τα οποία προέρχονται είτε από τον αναλυτή, είτε από το αρχικό υλικό των υπό ανάλυση δεδομένων. Με το FastQC ελέγχεται η ποιότητα των reads πριν ή μετά το trimming. Το FastQC παρέχει μια σειρά αναλύσεων με σκοπό να εξασφαλιστεί μια γρήγορη εικόνα για το αν τα ελεγχόμενα δεδομένα παρουσιάζουν προβλήματα πριν διεξαχθεί οποιαδήποτε περαιτέρω ανάλυση.

2.5 Συναρμολόγηση δεδομένων της Illumina με το πρόγραμμα SPAdes και υβριδική συναρμολόγηση με τα δεδομένα της Illumina και PacBio

Χρησιμοποιήθηκε, η έκδοση 3.11.1 του προγράμματος SPAdes για τη συναρμολόγηση της πλήρους γονιδιωματικής αλληλουχίας των εξεταζόμενων μικροοργανισμών από τα επεξεργασμένα δεδομένα των reads που προέκυψαν με την τεχνολογία Illumina. Με το ίδιο πρόγραμμα διενεργήθηκε και η υβριδική συναρμολόγηση των δεδομένων της Illumina με τα αντίστοιχα δεδομένα που αποκτήθηκαν με την τεχνολογία PacBio. Τα δεδομένα της PacBio αποτελούνται από reads μεγάλου μήκους, η ευκολότερη συναρμολόγηση των οποίων βοηθά την ορθή αλληλοεπικάλυψη και, άρα, τη συναρμολόγηση των δεδομένων της Illumina που αποτελούνται από reads ιδιαίτερα μικρού μήκους. Αυτό συνεπάγεται σε μια τελική συναρμολόγηση μεγαλύτερης κάλυψης και ακρίβειας.

2.6 Συναρμολόγηση δεδομένων της PacBio με το πρόγραμμα Canu

Χρησιμοποιήθηκε η έκδοση v1.6 του προγράμματος Canu για τη συναρμολόγηση των reads της τεχνολογίας PacBio. Το Canu μπορεί να συναρμολογήσει, με αξιοπιστία, πλήρη μικροβιακά γονιδιώματα χρησιμοποιώντας είτε την τεχνολογία PacBio ή την Oxford Nanopore και λειτουργεί διαδοχικά σε τρία στάδια τα οποία είναι τα εξής: διόρθωση, trimming και συναρμολόγηση.

2.7 Απεικόνιση της στοίχισης των reads με τα προγράμματα BLASR και IGV

Χρησιμοποιήθηκε το πρόγραμμα BLASR, το οποίο μπορεί να στοιχίσει, με μεγάλη ακρίβεια και ταχύτητα, τα μεγάλα σε μήκος reads της PacBio σε αντίστοιχα μικροβιακά γονιδιώματα αναφοράς.

Επίσης, χρησιμοποιήθηκε η έκδοση 2.4.3 του προγράμματος IGV για την απεικόνιση των αλληλοεπικαλυπτόμενων reads, όπως έχουν συναρμολογηθεί με βάση ένα συγκεκριμένο μικροβιακό γονιδίωμα αναφοράς.

2.8 Αναζήτηση του περισσότερο συγγενικού γονιδιώματος με το πρόγραμμα BLAST

Χρησιμοποιήθηκε το πρόγραμμα BLAST για την αναζήτηση στη βάση δεδομένων του Εθνικού Κέντρου Βιοτεχνολογικών Πληροφοριών των Ηνωμένων Πολιτειών (NCBI) ομόλογων γονιδιωματικών αλληλουχιών των περισσότερο συγγενικών μικροβιακών γονιδιωμάτων και των πληροφοριών της νουκλεοτιδικής τους αλληλουχίας και, άρα, των πιο στενά συγγενικών μικροοργανισμών με εκείνον που τελεί υπό εξέταση.

2.9 Σύγκριση με άλλο γονιδίωμα αναφοράς με το πρόγραμμα Blast2seq και με στιγμοπίνακα

Χρησιμοποιήθηκε το πρόγραμμα Blast2seq (BLAST 2 sequences), για τη σύγκριση του υπό μελέτη συναρμολογημένου μικροβιακού γονιδιώματος με τη νουκλεοτιδική αλληλουχία του γονιδιώματος του πιο στενά συγγενικού μικροοργανισμού.

Μέσω της εφαρμογής του προγράμματος Blast2seq δίνεται η δυνατότητα της εξέτασης της απόστασης της γενετικής συγγένειας των δύο συγκρινόμενων γονιδιωμάτων με απεικόνιση σε στιγμοπίνακα (Dot Plot).

2.10 Εντοπισμός γονιδίων παθογονικότητας με αναζήτηση στη βάση δεδομένων VFDB

Χρησιμοποιήθηκε η βάση δεδομένων VFDB (Virulence Factors Database) για τον εντοπισμό γονιδίων παθογονικότητας στο υπό εξέταση συναρμολογημένο μικροβιακό γονιδίωμα και του είδους των πρωτεϊνικών προϊόντων που κωδικοποιούνται από αυτά τα γονίδια. Η VFDB παρέχει επικαιροποιημένη γνώση σχετικά με τα χαρακτηριστικά της δομής των σημαντικών βακτηριακών μικροοργανισμών και τους μηχανισμούς που χρησιμοποιούνται από αυτούς για να παρακάμψουν τους αμυντικούς μηχανισμούς του ξενιστή και να προκαλέσουν ασθένειες.

ΑΠΟΤΕΛΕΣΜΑΤΑ

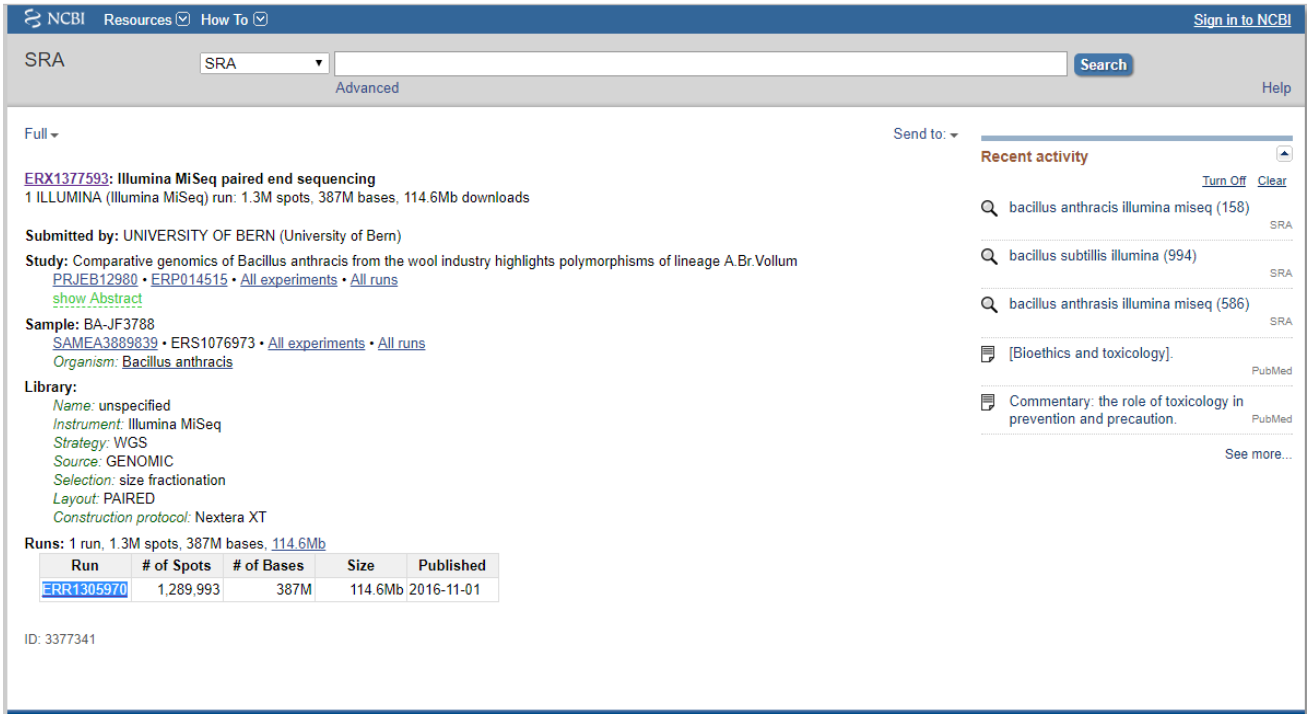
3.1 Αναζήτηση των reads στο SRA και λήψη τους σε μορφή FASTQ

Τα δεδομένα των νουκλεοτιδικών αλληλουχιών των reads αποθηκεύονται σε αρχεία με τη μορφή «FASTQ». Στα αρχεία FASTQ, εκτός από τα δεδομένα της αλληλουχίας, αποτυπώνεται και η ποιότητα των αλληλουχιών των reads. Ο έλεγχος της ποιότητας των reads είναι ένα από τα πιο σημαντικά βήματα στην προεπεξεργασία των αλληλουχιών τους, καθώς οι ακατέργαστες αλληλουχίες των reads περιέχουν, αναπόφευκτα, λάθη. Η πιθανότητα ενός σφάλματος για κάθε νουκλεοτίδιο σε κάθε read είναι πάντα αποτυπωμένη σε ένα αρχείο FASTQ.

Η αναζήτηση πληροφοριών σχετικά με τα reads μιας γονιδιωματικής αλληλουχίας, καθώς και της μεθόδου αλληλούχισης από την οποία προέκυψαν, λαμβάνει χώρα στη βάση δεδομένων «Sequencing Read Archive (SRA)» που διαχειρίζεται το Εθνικό Κέντρο Βιοτεχνολογικών Πληροφοριών των Η.Π.Α. (National Center for Biotechnology Information, NCBI). Η πρόσβαση στην εν λόγω βάση δεδομένων γίνεται μέσω του συνδέσμου:
https://www.ncbi.nlm.nih.gov/sra

Το SRA δημιουργήθηκε το 2007 από το NCBI και αποτελεί μια βάση δεδομένων βιοπληροφορικής που παρέχει ένα δημόσιο αρχείο αποθήκευσης για τα δεδομένα νουκλεοτιδικών αλληλουχιών DNA, ειδικά για τα reads που προκύπτουν από τις τεχνολογίες αλληλούχισης νέας γενιάς (Wheeler et al. 2008). Η εν λόγω βάση δεδομένων αποτελεί μέρος της Διεθνούς Συνεργασίας Βάσεων Δεδομένων Νουκλεοτιδικών Αλληλουχιών (International Nucleotide Sequence Database Collaboration, INSDC) και λειτουργεί ως συνεργασία μεταξύ του NCBI, του Ευρωπαϊκού Ινστιτούτου Βιοπληροφορικής (European Bioinformatics Institute, EBI) και της Ιαπωνικής Τράπεζας Δεδομένων DNA (DNA Data Bank of Japan, DDBJ) (Sequence Read Archive, Wikipedia).

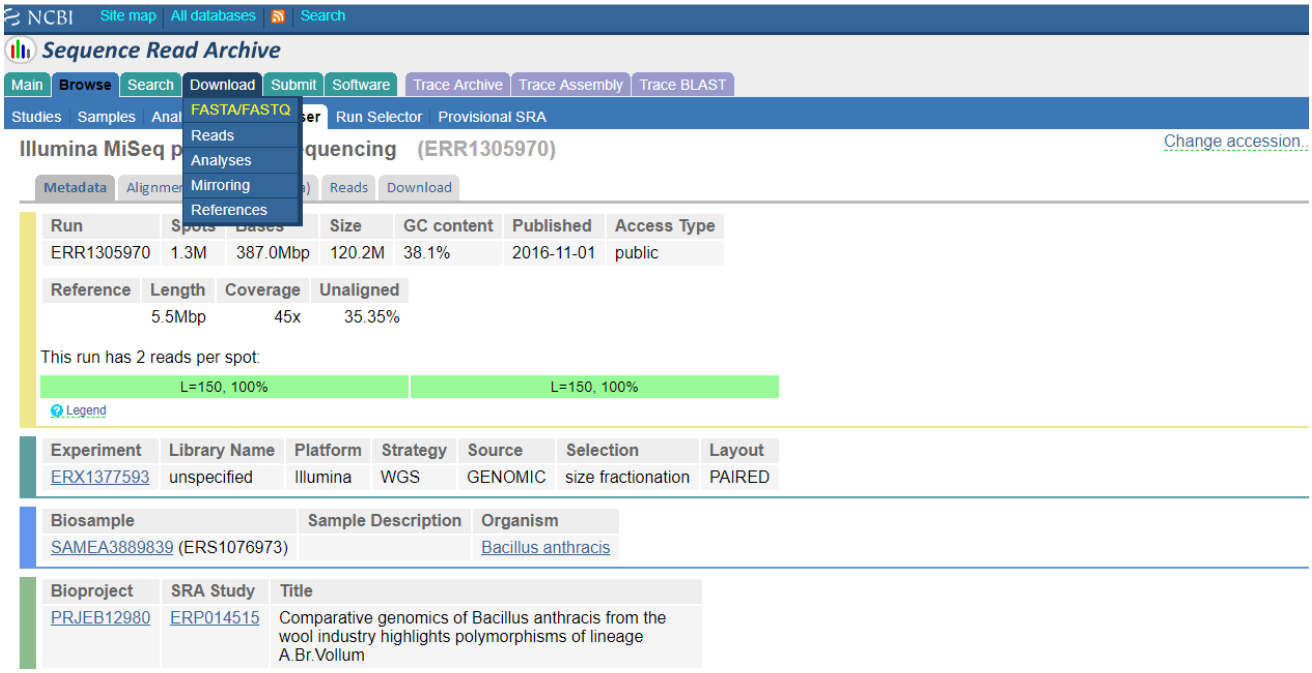
Από την αναζήτηση στο SRA για δεδομένα reads ενός μικροοργανισμού, π.χ. του μικροοργανισμού *Bacillus anthracis*, τα οποία αποκτήθηκαν μέσω κάποιας συσκευής αλληλούχισης νέας γενιάς, π.χ. με τον αναλυτή MiSeq® της εταιρίας Illumina, προκύπτει το περιεχόμενο της Εικόνας 14.



Εικόνα 14 Αναζήτηση δεδομένων reads στο SRA. Στην ενότητα Runs επιλέγουμε τον αριθμό πρόσβασης που βρίσκεται αριστερά στον πίνακα κάτω από το πλαίσιο Run.

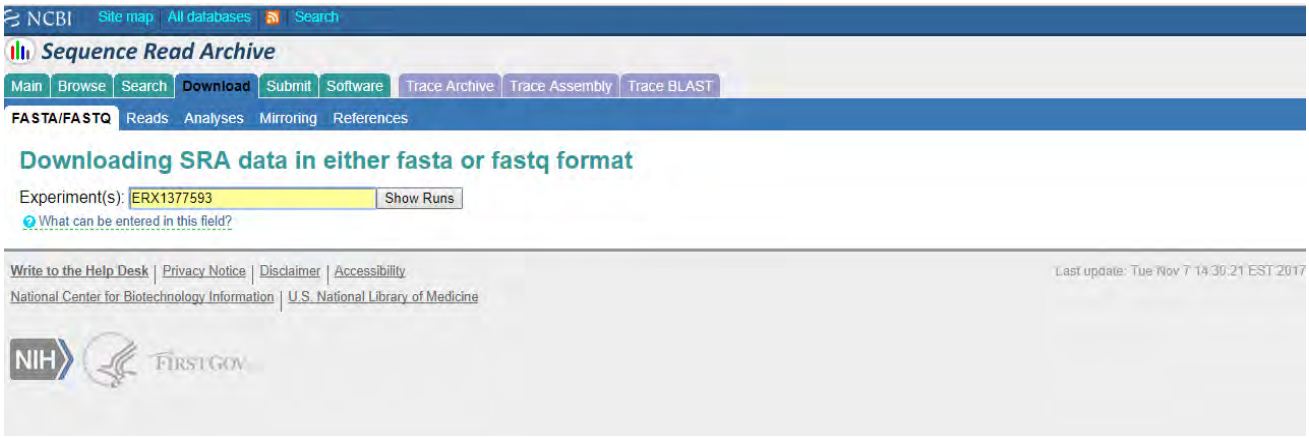
Στο πεδίο της Εικόνας 14, το οποίο είναι αντίστοιχο και για σχετικά δεδομένα άλλων μικροοργανισμών, από την ενότητα «Runs» επιλέγουμε τον αριθμό πρόσβασης (accession number) που βρίσκεται αριστερά στον πίνακα κάτω από το πλαίσιο «Run». Με αυτόν τον τρόπο ανοίγει το παράθυρο της Εικόνας 15.

Στην Εικόνα 15, στην ενότητα «Browse - Run Browser - Metadata» αντιγράφουμε τον αριθμό του «Experiment» και στη συνέχεια από την ενότητα «Download» επιλέγουμε «FASTA/FASTQ».



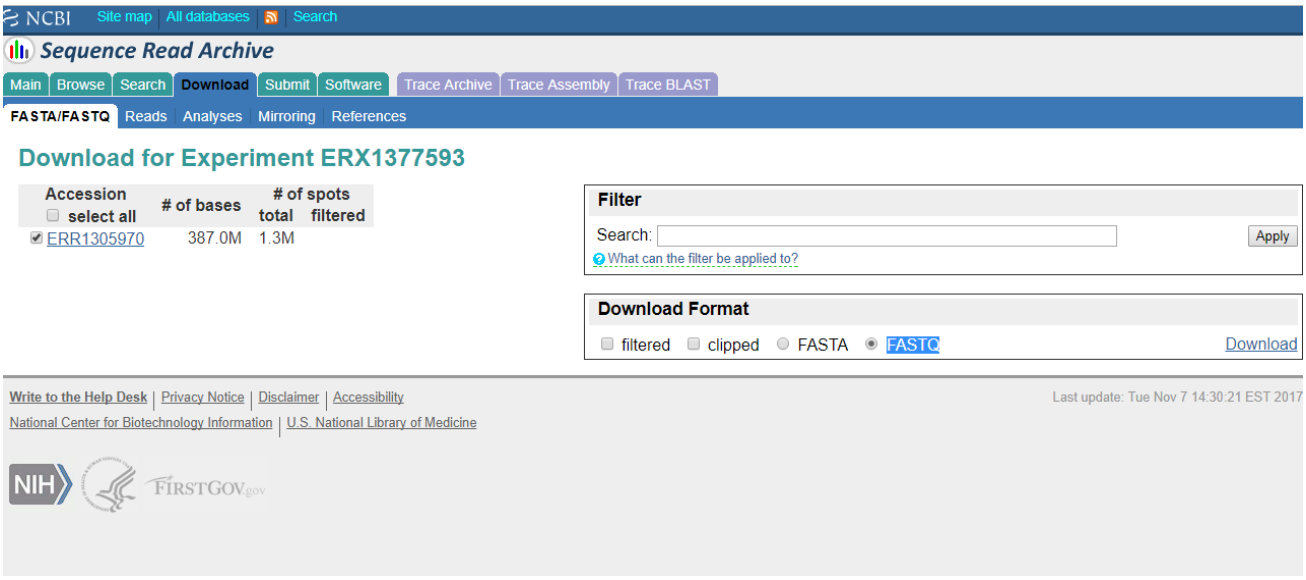
Εικόνα 15 Στην ενότητα Browse - Run Browser - Metadata αντιγράφουμε τον αριθμό του Experiment και στη συνέχεια από την ενότητα Download επιλέγουμε FASTA/FASTQ.

Τον αριθμό του «Experiment» τον εισάγουμε στο κίτρινο πλαίσιο του παραθύρου που απεικονίζεται στην Εικόνα 16 και εκτελούμε «Show Runs».



Εικόνα 16 Τον αριθμό του Experiment τον εισάγουμε στο κίτρινο πλαίσιο του παραθύρου και εκτελούμε Show Runs.

Με την ενέργεια «Show Runs» ανοίγει το παράθυρο της Εικόνας 17, όπου επιλέγουμε να κατεβάσουμε και να σώσουμε στον υπολογιστή μας τα δεδομένα των reads σε μορφή FASTQ.



Εικόνα 17 Επιλέγουμε να κατεβάσουμε και να σώσουμε στον υπολογιστή μας τα δεδομένα των reads σε μορφή FASTQ.

Με αυτό τον τρόπο κατεβαίνει ένα συμπιεσμένο αρχείο τύπου «sra_data.fastq.gz», το οποίο αποσυμπιέζουμε. Από την αποσυμπίεση του αρχείου sra_data.fastq.gz προκύπτει ένα αρχείο «sra_data.fastq». Για πρακτικούς λόγους περιγραφής, το αρχείο αυτό το μετονομάζουμε με τα στοιχεία του υπό μελέτη μικροοργανισμού, π.χ. «b_anthracis.fastq» και το αποθηκεύουμε σε έναν φάκελο με την αντίστοιχη ονομασία, π.χ. «bacillus_anthracis».

Στο σημείο αυτό αξίζει να αναφερθεί η δομή ενός αρχείου FASTQ και να επισημανθεί πώς κωδικοποιούνται σε αυτό οι πληροφορίες για την ποιότητα των reads. Όπως φαίνεται και στην Εικόνα 18, τα αρχεία FASTQ έχουν πάντα 4 γραμμές ανά αλληλουχία. Στην πρώτη γραμμή εμφανίζεται το αναγνωριστικό της αλληλουχίας και μια προαιρετική περιγραφή. Η δεύτερη γραμμή περιέχει μια αλληλουχία νουκλεοτιδίων. Η τρίτη γραμμή διατηρεί, γενικά, μόνο το σύμβολο «+» και, περιστασιακά, το ίδιο αναγνωριστικό και περιγραφή αλληλουχίας με την πρώτη γραμμή. Η τέταρτη γραμμή αναφέρει τη βαθμολογία ποιότητας (Quality score, Q-score) κάθε νουκλεοτιδίου που εμφανίζεται στη δεύτερη γραμμή.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Εικόνα 18 Το αρχείο FASTQ έχει 4 γραμμές ανά αλληλουχία. Στην πρώτη γραμμή εμφανίζεται το αναγνωριστικό της αλληλουχίας και μια προαιρετική περιγραφή. Η δεύτερη γραμμή περιέχει μια αλληλουχία νουκλεοτιδίων. Η τρίτη γραμμή διατηρεί μόνο το σύμβολο «+» και, περιστασιακά, το ίδιο αναγνωριστικό και περιγραφή αλληλουχίας με την πρώτη γραμμή. Η τέταρτη γραμμή αναφέρει τη βαθμολογία ποιότητας κάθε νουκλεοτιδίου που εμφανίζεται στη δεύτερη γραμμή. Η τέταρτη γραμμή δεν περιέχει διψήφιους αριθμούς, αλλά μόνο τους ειδικούς χαρακτήρες ASCII που χρησιμοποιούνται για την κωδικοποίηση των τιμών ποιότητας με ένα μόνο σύμβολο.

Από: FASTQ format From Wikipedia, the free encyclopedia, https://en.wikipedia.org/wiki/FASTQ_format

Οι βαθμολογίες ποιότητας στην τέταρτη γραμμή αντιπροσωπεύουν, για κάθε νουκλεοτίδιο, την πρόβλεψη της πιθανότητας να έχει συμβεί σφάλμα σε μια βάση κατά τη διαδικασία της αλληλούχισης. Σε επίπεδο βάσης, υψηλή βαθμολογία ποιότητας υποδηλώνει πιο αξιόπιστη αλληλούχιση μιας βάσης και μικρότερη πιθανότητα να είναι εσφαλμένη. Έτσι, όταν αναφέρεται τιμή Q-score ίση με 40, θα υπάρχει πιθανότητα σφάλματος ίση με 0,0001, δηλαδή, 1 σφάλμα στις 10.000 βάσεις. Αντίστοιχα, για Q-score ίσο με 30 θα έχουμε πιθανότητα σφάλματος ίση με 0,001, δηλαδή, 1 σφάλμα στις 1000 βάσεις (Illumina, Technical Note: Informatics, Understanding Illumina Quality Scores). Κατά τον υπολογισμό αυτό, σε τιμή Q-score=10 αντιστοιχεί ακρίβεια της τάξης του 90%, σε Q-score=20 αντιστοιχεί ακρίβεια 99% και σε τιμή Q-score=30 αντιστοιχεί ακρίβεια 99,9% (Illumina, Introduction to Sequencing Quality Scores). Επομένως, αξιόπιστη βαθμολογία ποιότητας θεωρείται εκείνη που έχει τιμή Q-score από 30 και άνω (Illumina, Technical Note: Informatics, Quality Scores for Next-Generation Sequencing).

Εάν γνωρίζουμε την πιθανότητα p ενός τέτοιου σφάλματος, τότε είναι εύκολο να αποκτήσουμε τη βαθμολογία ποιότητας χρησιμοποιώντας την εξίσωση:

$$\text{Q-score} = -10 \log_{10} p$$

Για παράδειγμα, εάν η πιθανότητα ενός σφάλματος p ισούται με 0,01, τότε η αντίστοιχη βαθμολογία ποιότητας θα είναι 20. Αν p=0,001, τότε Q-score=30. Όμως, όπως βλέπουμε στην τέταρτη γραμμή του παραδείγματος, το αρχείο

FASTQ δεν περιέχει διψήφιους αριθμούς, αλλά μόνο χαρακτήρες. Αυτοί είναι ειδικοί χαρακτήρες ASCII (American Standard Code for Information Interchange) που χρησιμοποιούνται για την κωδικοποίηση τιμών ποιότητας με ένα μόνο σύμβολο, αντί για ένα διπλό ή τριπλό ψηφίο (ASCII, Wikipedia). Αυτή η κωδικοποίηση ενός χαρακτήρα είναι απαραίτητη, επειδή θέλουμε να έχουμε αντιστοιχία ένα προς ένα μεταξύ κάθε νουκλεοτιδίου στη γραμμή 2 και κάθε βαθμολογίας στη γραμμή 4. Αυτό επιτυγχάνεται με τη χρήση ενός μόνο χαρακτήρα και στις δύο γραμμές (InsideDNA, 2016). Η μορφή FASTQ αποτελεί το πρότυπο για την αποθήκευση των παραγόμενων δεδομένων των τεχνολογιών αλληλούχισης νέας γενιάς (Cock et al., 2009).

3.2 Trimming των reads με το πρόγραμμα ConDeTri

Το ConDeTri, είναι μια εφαρμογή βιοπληροφορικής για το φιλτράρισμα των δεδομένων που προέρχονται από την τεχνολογία Illumina.

Το πρόγραμμα ConDeTri λειτουργεί ως ένα εργαλείο διαλογής που εφαρμόζει την επεξεργασία αποκοπής και αφαίρεσης (trimming) ολόκληρων των reads, ή τμημάτων τους, χρησιμοποιώντας τις βαθμολογίες ποιότητας για κάθε βάση ξεχωριστά. Το ConDeTri μπορεί να επεξεργάζεται δεδομένα των reads είτε είναι μεμονωμένων (single-end) ή ζευγαρωμένων άκρων (paired-end). Διατίθεται ελεύθερα από τον σύνδεσμο:

https://github.com/linneas/condetri

Για πρακτικούς λόγους, τα χρησιμοποιούμενα προγράμματα τα αποθηκεύουμε σε έναν φάκελο με την ονομασία, π.χ. «Programs». Στον φάκελο Programs κατεβάζουμε και αποθηκεύουμε το πρόγραμμα ConDeTri, το οποίο φαίνεται ως ένας φάκελος που ονομάζεται «condetri-master». Από τον εν λόγω φάκελο αντιγράφουμε το αρχείο «condetri.pl» και το τοποθετούμε στον φάκελο bacillus_anthracis που περιέχει το αρχείο b_anthracis.fastq με τα δεδομένα των reads σε μορφή FASTQ, τα οποία αφορούν στην πλήρη νουκλεοτιδική αλληλουχία του γονιδιώματος του υπό μελέτη μικροοργανισμού.

Ανοίγουμε το τερματικό (terminal) και με την κατάλληλη εντολή εισερχόμαστε στον φάκελο «bacillus_anthracis» όπου βρίσκονται τα προς ανάλυση αρχεία «b_anthracis.fastq» και «condetri.pl». Στη συνέχεια δίνουμε την παρακάτω εντολή για να εκτελεστεί το πρόγραμμα ConDeTri:

perl condetri.pl -fastq1=b_anthracis.fastq -hq=30 -lq=15 -minlen=100 -sc=33 -rmN

Όπου:

-hq, η ελάχιστη καλύτερη ποιότητα που θέλουμε να έχουν τα reads.

-lq, η ελάχιστη χειρότερη ποιότητα που θέλουμε να έχουν τα reads.

-minlen, το ελάχιστο μήκος που θα έχουν τα reads που θα κρατήσουμε (GitHub linneas/condetri).

Με την εκτέλεση της παραπάνω εντολής λαμβάνουμε, στον φάκελο «bacillus_anthracis», ένα αρχείο με όνομα «b_anthracis_trim.fastq», το οποίο περιέχει τμήματα, ή ολόκληρα reads με ελάχιστο μήκος τα 100 bp, ελάχιστη καλύτερη ποιότητα βαθμολογίας ίση με 30, ενώ απορρίπτονται εκείνα τα reads, ή τμήματά τους, με ελάχιστη χειρότερη ποιότητα που είναι ίση με 15. Σε ό,τι αφορά στην τιμή της παραμέτρου –sc, τίθεται στον αριθμό 33, καθώς αναφέρεται στα δεδομένα που προκύπτουν από την τεχνολογία Illumina και συγκεκριμένα σε εκείνα που προκύπτουν από τις νεότερες εκδόσεις της.

3.3 Έλεγχος της ποιότητας των reads με το πρόγραμμα FastQC

Οι σύγχρονοι αναλυτές αλληλούχισης νέας γενιάς μπορούν να παράγουν εκατομμύρια αλληλουχίες σε μία μόνο αναλυτική διαδικασία. Ωστόσο, πριν αναλυθεί μία αλληλουχία με σκοπό να εξαχθούν βιολογικά συμπεράσματα, θα πρέπει πάντα να εκτελεστούν ορισμένοι απλοί ποιοτικοί έλεγχοι, ώστε να διασφαλιστεί ότι τα ανεπεξέργαστα δεδομένα φαίνονται καλά και ότι δεν υπάρχουν προβλήματα ή μεροληψία που μπορεί να επηρεάσουν τον τρόπο με τον οποίο μπορεί κάποιος να τα εκμεταλλευτεί. Οι περισσότεροι αναλυτές δημιουργούν μια αναφορά ποιοτικού ελέγχου, ως μέρος της ανάλυσής τους,

αλλά αυτό, συνήθως, επικεντρώνεται μόνο στην αναγνώριση των προβλημάτων που δημιουργήθηκαν από τον ίδιο τον αναλυτή.

Το πρόγραμμα «FastQC» στοχεύει στην παροχή μιας αναφοράς ποιοτικού ελέγχου που μπορεί να εντοπίσει προβλήματα τα οποία προέρχονται είτε από τον αναλυτή, είτε από το αρχικό υλικό των υπό ανάλυση δεδομένων. Επιπλέον, το FastQC μπορεί να εκτελεστεί με έναν από τους δύο τρόπους:

- Μπορεί είτε να λειτουργήσει ως μια αυτόνομη διαδραστική εφαρμογή για την άμεση ανάλυση μικρού αριθμών αρχείων FASTQ, ή
- μπορεί να εκτελεστεί σε μη διαδραστική λειτουργία, η οποία θα ήταν κατάλληλη για την ενσωμάτωση σε μια ευρύτερη αναλυτική διαδικασία για τη συστηματική επεξεργασία μεγάλου αριθμού αρχείων (What is FastQC, https://www.bioinformatics.babraham.ac.uk)

Ειδικότερα, με την εφαρμογή του προγράμματος FastQC δύναται να ελεγχθεί η ποιότητα των reads τόσο πριν από το trimming με το ConDeTri, όσο και μετά το trimming. Το FastQC στοχεύει στην παροχή ενός απλού τρόπου για να γίνει ένας ποιοτικός έλεγχος σε ακατέργαστα δεδομένα αλληλουχιών reads που προέρχονται από τις τεχνολογίες αλληλούχισης νέας γενιάς. Παρέχει μια διαρθρωμένη σειρά αναλύσεων που μπορεί κάποιος να χρησιμοποιήσει με σκοπό να εξασφαλίσει μια γρήγορη εικόνα για το αν τα ελεγχόμενα δεδομένα παρουσιάζουν προβλήματα, τα οποία θα πρέπει να γνωρίζει πριν προβεί σε οποιαδήποτε περαιτέρω ανάλυση. Οι κύριες λειτουργίες του εν λόγω προγράμματος είναι οι εξής:

- Εισάγει δεδομένα από αρχεία BAM, SAM ή FASTQ.
- Παρέχει μια γρήγορη επισκόπηση, ώστε να δηλωθεί σε ποιες περιοχές μπορεί να υπάρχουν προβλήματα.
- Δημιουργεί γραφήματα και πίνακες για να αξιολογηθούν γρήγορα τα προς ανάλυση δεδομένα.
- Εξάγει αποτελέσματα σε αναφορά βασισμένη σε HTML.
- Λειτουργεί και εκτός σύνδεσης για την αυτόματη δημιουργία αναφορών, χωρίς εκτέλεση της διαδραστικής εφαρμογής (FastQC A quality control tool for high throughput sequence data).

Η έκδοση v0.11.5 του προγράμματος FastQC διατίθεται ελεύθερα από το Babraham Bioinformatics (Babraham Institute) στον σύνδεσμο:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

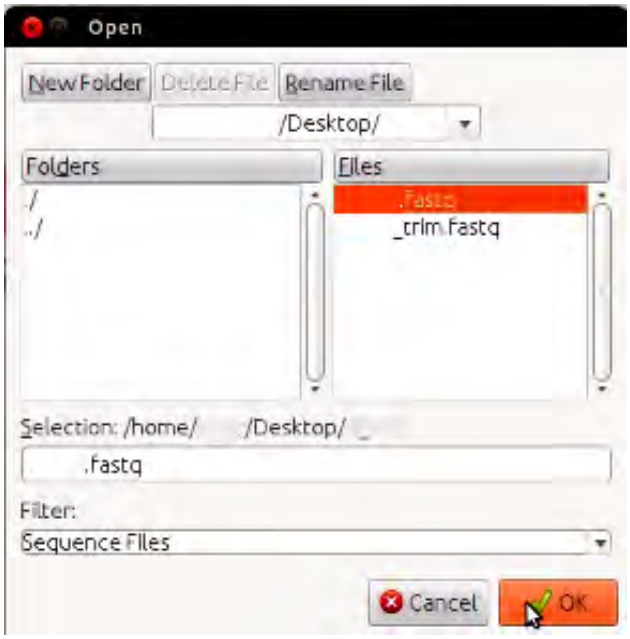
Το πρόγραμμα FastQC το κατεβάζουμε και το αποθηκεύουμε στον φάκελο Programs. Στη συνέχεια, από το τερματικό, με την κατάλληλη εντολή, εισερχόμαστε στον φάκελο Programs και από εκεί στον φάκελο όπου εμπεριέχεται το πρόγραμμα FastQC. Δίνουμε την εξής εντολή για να εκτελεστεί το πρόγραμμα:

perl fastqc

Το FastQC ανοίγει με την προβολή του παραθύρου της Εικόνας 19.

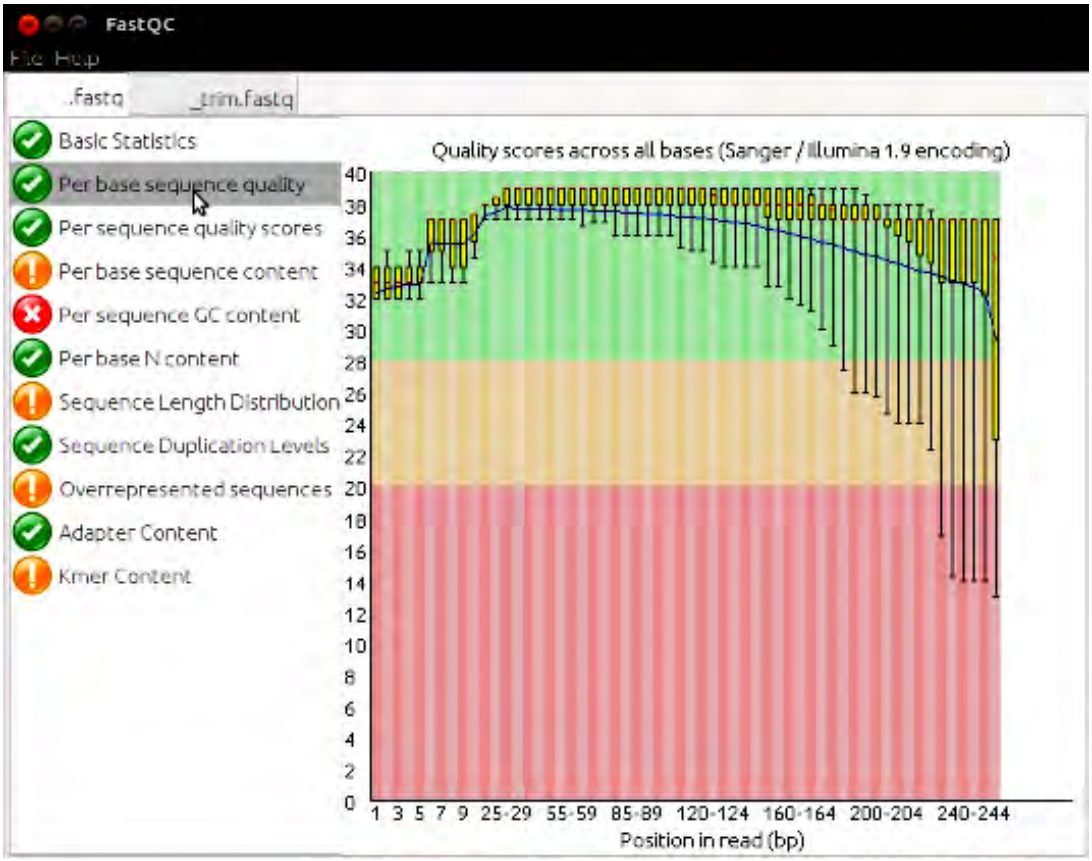
Εικόνα 19 Παράθυρο έναρξης της έκδοσης 0.11.5 του προγράμματος FastQC.

Ανοίγουμε την ενότητα «File» και επιλέγοτας «Open» ανοίγει ένα νέο παράθυρο, παράδειγμα του οποίου απεικονίζεται στην Εικόνα 20. Διαμέσου αυτού του παραθύρου μεταφερόμαστε στον φάκελο bacillus_anthracis που περιέχει το αρχείο b_anthracis.fastq με τα δεδομένα των reads πριν γίνει το trimming και το αρχείο b_anthracis_trim.fastq που περιέχει τα δεδομένα μετά το trimming.



Εικόνα 20 Διαμέσου αυτού του παραθύρου μεταφερόμαστε στον φάκελο, π.χ., bacillus_anthracis που περιέχει το αρχείο b_anthracis.fastq με τα δεδομένα των reads πριν γίνει το trimming.

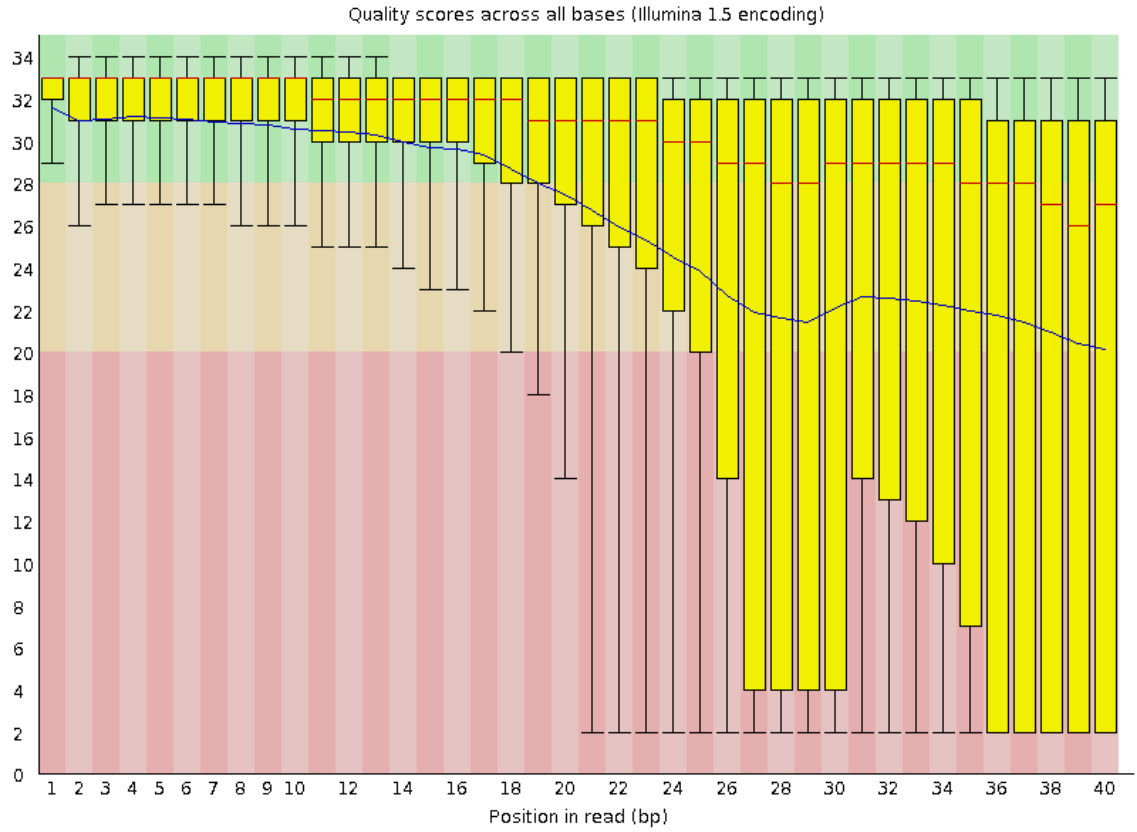
Επιλέγοντας ένα από τα δύο αρχεία κάθε φορά και πατώντας το «√OK» ανοίγει, αντίστοιχα για κάθε αρχείο, ένα παράθυρο. Ειδικότερα, επιλέγοντας το αρχείο «.fastq» ανοίγει ένα παράθυρο, όπως αυτό της Εικόνας 21.



Εικόνα 21 Από το παράθυρο της Εικόνας 20, επιλέγοντας το αρχείο «.fastq» ανοίγει το παράθυρο με τα δεδομένα των reads πριν γίνει το trimming.

Στο παράθυρο της Εικόνας 21 απεικονίζεται μια επισκόπηση του εύρους των τιμών ποιότητας για όλες τις βάσεις, σε κάθε θέση, στο αρχείο FASTQ. Για κάθε θέση σχεδιάζεται μια γραφική παράσταση τύπου θηκογράμματος (BoxWhisker).

Η Εικόνα 22 παρατίθεται για να εστιάσουμε στα στοιχεία αυτού του διαγράμματος, τα οποία είναι τα εξής:



Εικόνα 22 Επισκόπηση του εύρους των τιμών ποιότητας για όλες τις βάσεις, σε κάθε θέση, σε ένα αρχείο FASTQ. Για κάθε θέση σχεδιάζεται μια γραφική παράσταση τύπου θηκογράμματος (BoxWhisker).

Από: FastQC Report
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html#M0

- Το κίτρινο πλαίσιο (Box) αντιπροσωπεύει το εύρος μεταξύ τεταρτημορίων (25-75%).
- Η κόκκινη γραμμή, που βρίσκεται στο κίτρινο πλαίσιο, είναι η μέση τιμή.
- Οι άνω και οι κάτω μαύρες γραμμές (Whiskers) αντιπροσωπεύουν τα σημεία 10% και 90%, αντίστοιχα.
- Η μπλε γραμμή, η οποία διέρχεται, αρχικά, προς τη βάση των κίτρινων πλαισίων, αντιπροσωπεύει τη μέση ποιότητα.

Σε ένα γράφημα FastQC ο άξονας y δείχνει τις βαθμολογίες ποιότητας. Όσο υψηλότερη είναι η βαθμολογία ποιότητας, τόσο περισσότερο έχει αποφευχθεί η πιθανότητα να έχει γίνει λάθος στον προσδιορισμό της βάσης κατά τη διαδικασία της αλληλούχισης στο συγκεκριμένο σημείο. Το φόντο του γραφήματος διαιρεί τον άξονα y σε αλληλουχίσεις πολύ καλής ποιότητας

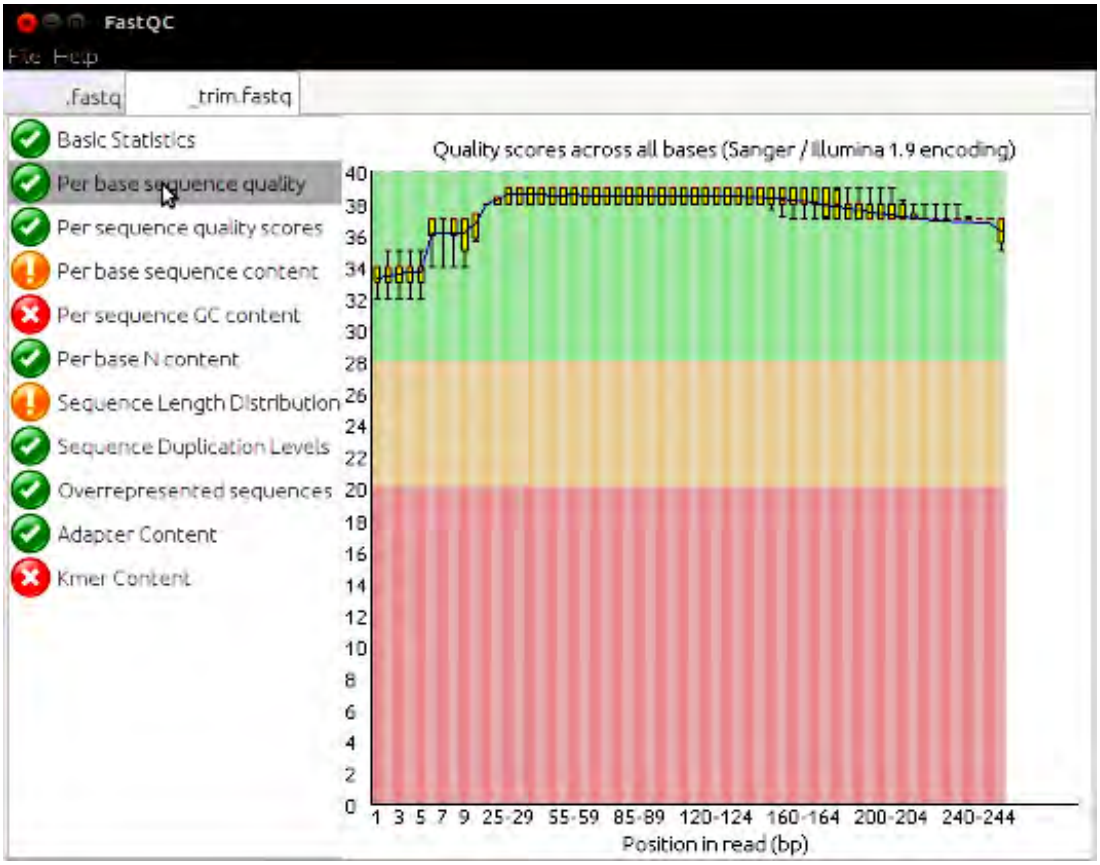
(πράσινο), ικανοποιητικής ποιότητας (πορτοκαλί) και χαμηλής ποιότητας (ροζ).

Επιπλέον, θα πρέπει να αναφερθεί ότι υπάρχουν διάφοροι τρόποι για να κωδικοποιηθεί μια βαθμολογία ποιότητας σε ένα αρχείο FASTQ. Το πρόγραμμα FastQC επιχειρεί να προσδιορίσει αυτόματα ποια μέθοδο κωδικοποίησης χρησιμοποιήθηκε, αλλά ορισμένες φορές είναι πιθανό να το μαντέψει εσφαλμένα. Ο τίτλος του γραφήματος θα αναγράφει την κωδικοποίηση την οποία το πρόγραμμα διαπιστώνει ότι χρησιμοποιήθηκε στο εισαχθέν προς επεξεργασία αρχείο. Τα αποτελέσματα της επεξεργασίας δε θα εμφανίζονται σε περίπτωση που το εισαχθέν αρχείο είναι σε μορφή BAM / SAM, καθώς σε αυτή τη μορφή δεν καταγράφονται οι βαθμολογίες ποιότητας.

Αναφορικά με τις προειδοποιήσεις του προγράμματος, προειδοποίηση θα παρουσιαστεί εάν το κατώτερο τεταρτημόριο για οποιαδήποτε βάση είναι μικρότερο από 10, ή εάν η μέση τιμή για οποιαδήποτε βάση είναι μικρότερη από 25. Αποτυχία θα υφίσταται εάν το κατώτερο τεταρτημόριο για οποιαδήποτε βάση είναι μικρότερο από 5 ή εάν η μέση τιμή για οποιαδήποτε βάση είναι μικρότερη από 20. Ο συνηθέστερος λόγος για προειδοποιήσεις και αποτυχίες είναι η γενική υποβάθμιση της ποιότητας κατά τη διάρκεια των μεγάλων διαδικασιών αλληλούχισης. Επομένως, για μακρές διαδικασίες, δύναται να διαπιστωθεί ότι η γενική ποιότητα της αλληλούχισης πέφτει σε ένα επίπεδο όπου ενεργοποιείται μια προειδοποίηση ή ένα σφάλμα.

Εάν η ποιότητα πέσει σε χαμηλό επίπεδο, τότε η πιο κοινή λύση είναι η εκτέλεση της ποιοτικής αποκοπής και αφαίρεσης (trimming), όπου τμήματα, ή ολόκληρα, reads κόβονται με βάση τη μέση ποιότητά τους.

Επιλέγοντας το αρχείο «_trim.fastq» ανοίγει το παράθυρο της Εικόνας 23, το οποίο απεικονίζει την ποιότητα των reads κατόπιν του trimming με το πρόγραμμα ConDeTri. Επομένως, απεικονίζονται τα ποιοτικά δεδομένα των reads που βρίσκονται στο αρχείο «_trim.fastq» και τα οποία προέκυψαν από το αρχείο «.fastq» έπειτα από την ποιοτική επεξεργασία τους.



Εικόνα 23 Από το παράθυρο της Εικόνας 20, επιλέγοντας το αρχείο «_trim.fastq» ανοίγει το παράθυρο με τα δεδομένα των reads μετά το trimming.

Έτσι λοιπόν, στο παράθυρο της Εικόνας 21 απεικονίζονται τα δεδομένα των reads πριν από την επεξεργασία με trimming, ενώ στο παράθυρο της Εικόνας 23 τα δεδομένα μετά το trimming. Στην Εικόνα 23, διαπιστώνουμε ότι τα δεδομένα είναι πολύ καλύτερα από εκείνα που φαίνονται στην Εικόνα 21 και τούτο διότι είναι η ίδια η διαδικασία του trimming που, ουσιαστικά, αποκόπτει και αφαιρεί ολόκληρα ή τμήματα από τα λιγότερο ποιοτικά reads, κρατώντας μόνον εκείνα που έχουν τιμή βαθμολογικής ποιότητας πάνω από 30. Για τον λόγο αυτό τα εν λόγω δεδομένα απεικονίζονται, αποκλειστικά, στην πράσινη περιοχή του διαγράμματος.

3.4 Συναρμολόγηση μικροβιακών γονιδιωμάτων με το πρόγραμμα SPAdes

Ωστόσο, η συναρμολόγηση του γονιδιώματος είναι δύσκολη λόγω της μη ομοιόμορφης κάλυψής του σε reads, της ποικιλίας του μήκους των αλληλουχιών στις οποίες λαμβάνει χώρα η αλληλούχιση, των σφαλμάτων που απορρέουν από την ίδια τη μεθοδολογία της αλληλούχισης και της ύπαρξης των χιμαιρικών reads. Η αλγοριθμική προσέγγιση SPAdes (St. Petersburg genome assembler), σχεδιάστηκε για να αντιμετωπίσει αυτά τα ζητήματα. (Ishoey et al., 2008 / Rodrigue et al., 2009 / Medvedev et al. 2011).

Το SPAdes είναι ένας αλγόριθμος συναρμολόγησης μικρών γονιδιωμάτων. Έχει δοκιμαστεί σε βακτηριακά, μυκητιακά, καθώς και σε άλλα μικρά γονιδιώματα και προϊόντα απομόνωσης μικρού μεγέθους. Συνεπώς, δεν προορίζεται για μεγαλύτερα γονιδιώματα, π.χ. του μεγέθους των θηλαστικών και δε θεωρείται κατάλληλο για γονιδιωματικές μελέτες μεγάλου εύρους (SPAdes 3.11.1 Manual / Bankevich et al., 2012).

Η μέθοδος συναρμολόγησης με το πρόγραμμα SPAdes χρησιμοποιεί ως εισερχόμενα δεδομένα τις νουκλεοτιδικές αλληλουχίες των reads και βασίζεται στα γραφήματα αλληλοεπικάλυψης de Bruijn Graph (DBG). Τα εισερχόμενα δεδομένα των αλληλουχιών των reads πρέπει να είναι σε μορφή FASTA ή FASTQ. Ωστόσο, για να εκτελεστεί η διόρθωση των σφαλμάτων, θα πρέπει τα δεδομένα να είναι σε μορφή FASTQ ή BAM. Σε ό,τι αφορά στα δεδομένα που προέρχονται από τις τεχνολογίες αλληλούχισης Sanger, Oxford Nanopore και PacBio CLR, δύναται να παρέχονται σε αμφότερες τις μορφές, δεδομένου ότι το SPAdes δεν εκτελεί διόρθωση σφαλμάτων για αυτούς τους τύπους δεδομένων.

Το πρόγραμμα SPAdes δημιουργήθηκε από τον Bankevich και τους συνεργάτες του το 2012 (Bankevich et al., 2012). Η έκδοση 3.11.1 του προγράμματος SPAdes διατίθεται ελεύθερα από τον ιστότοπο του Κέντρου Αλγοριθμικής Βιοτεχνολογίας του Κρατικού Πανεπιστημίου της Αγίας Πετρούπολης (Center for Algorithmic Biotechnology, St. Petersburg State University), μέσω του συνδέσμου:

http://cab.spbu.ru/software/spades/

Η εν λόγω έκδοση είναι η πιο πρόσφατη και λειτουργεί με αλληλουχίες reads που προέρχονται από τις τεχνολογίες Illumina και IonTorrent, αλλά είναι σε θέση να παρέχει και υβριδικές (συνδυαστικές) συναρμολογήσεις, χρησιμοποιώντας τα reads των τεχνολογιών PacBio, Oxford Nanopore και Sanger.

Για να εκτελεστεί το πρόγραμμα SPAdes 3.11.1 απαιτούνται δεδομένα από τουλάχιστον έναν από τους τύπους των τεχνολογιών Illumina, IonTorrent και PacBio CCS. Τα δεδομένα των reads από τις τεχνολογίες Illumina και IonTorrent δε θα πρέπει να συναρμολογούνται μαζί, ενώ όλοι οι άλλοι τύποι των εισερχόμενων δεδομένων είναι συμβατοί. Επίσης, το SPAdes δε θα πρέπει να χρησιμοποιηθεί με:

- δεδομένα μόνο από τις τεχνολογίες PacBio CLR, Oxford Nanopore και Sanger,
- με reads που προέρχονται από την τεχνολογία PacBio και παρουσιάζουν χαμηλή κάλυψη και συγκεκριμένα με τιμή μικρότερη από 5 και
- με reads προερχόμενα από την τεχνολογία PacBio για μεγάλα γονιδιώματα (SPAdes 3.11.1 Manual).

Για τη συναρμολόγηση των reads που αφορούν στο γονιδίωμα ενός μικροοργανισμού χρησιμοποιείται ένα παράδειγμα για πρακτικούς λόγους, όπως και στις λοιπές παραγράφους.

Ανοίγουμε τον φάκελο όπου περιέχονται τα δεδομένα του υπό μελέτη μικροοργανισμού, π.χ. τον φάκελο bacillus_anthracis και αντιγράφουμε το αρχείο b_anthracis_trim.fastq, το οποίο περιέχει τα δεδομένα των reads μετά το trimming, στον φάκελο «SPAdes-3.11.1-Linux» με το πρόγραμμα SPAdes. Έπειτα, ανοίγουμε το τερματικό και με την κατάλληλη εντολή εισερχόμαστε στον φάκελο SPAdes-3.11.1-Linux. Κατόπιν, εισάγουμε την εξής εντολή:

./bin/spades.py -s b_anthracis_trim.fastq -o b_anthracis_illum

Όπου:

-s, το αρχείο με τα δεδομένα των reads μετά το trimming.

-o, ο φάκελος όπου θα αποθηκευτούν τα αποτελέσματα από την εκτέλεση του προγράμματος (SPAdes 3.11.1 Manual).

Με την εκτέλεση της παραπάνω εντολής, δημιουργήθηκε εντός του φακέλου SPAdes-3.11.1-Linux ένας φάκελος με ονομασία

«b_anthracic_illum», ο οποίος, εκτός των λοιπών αρχείων του, περιέχει και δύο αρχεία το ένα με όνομα «contigs.fasta» και το άλλο «scaffolds.fasta», όπου υπάρχουν συναρμολογημένα τα reads σε contigs και τα contigs σε scaffolds, αντίστοιχα.

3.5 Υβριδική συναρμολόγηση με το πρόγραμμα SPAdes

Αναφορικά με την περιγραφόμενη μεθοδολογία για την ανάλυση των μικροβιακών γονιδιωμάτων, όταν χρησιμοποιούνται τα δεδομένα της τεχνολογίας Illumina ή της PacBio, θα ήταν πιο καλό να διεξαχθεί και μία συνδυαστική συναρμολόγηση των δεδομένων που προέρχονται και από τις δύο τεχνολογίες αλληλούχισης. Ειδικότερα, για να εξασφαλιστεί μια ακριβέστερη και πιο ποιοτική συναρμολόγηση ενός μικροβιακού γονιδιώματος προτείνεται η εφαρμογή της υβριδικής χρήσης των δεδομένων των reads που προέρχονται από τις τεχνολογίες Illumina και PacBio, με σκοπό να σχηματιστεί μια καλύτερη εικόνα για τη θέση στο γονιδίωμα του κάθε contig που δημιουργείται.

Γενικά, τα reads που προέρχονται από την τεχνολογία Illumina, ενώ χαρακτηρίζονται από υψηλή ποιότητα αλληλούχισης, έχουν μικρό μέγεθος και δεν είναι απλό να προσδιοριστεί η σωστή τους θέση επάνω στο γονιδίωμα, με αποτέλεσμα να καθίσταται δυσχερής η υπολογιστική εργασία που αποσκοπεί στην πλήρη συναρμολόγηση ενός γονιδιώματος. Τα προερχόμενα από την PacBio reads είναι μεγαλύτερα και έτσι διαθέτουν μεγαλύτερη επιφάνεια αλληλοεπικάλυψης μεταξύ τους.

Στο παράδειγμα της παραγράφου 3.1 αναζητήθηκαν στο SRA τα δεδομένα των reads για το βακτήριο *Bacillus anthracis*, τα οποία δημιουργήθηκαν με τη χρήση του αναλυτή MiSeq® της Illumina. Στη διαδικασία της συναρμολόγησης που περιεγράφηκε στο παράδειγμα της παραγράφου 3.4 χρησιμοποιήθηκαν, κατόπιν της κατάλληλης επεξεργασίας, τα εν λόγω δεδομένα.

Τα δεδομένα των reads από την αλληλούχιση με την τεχνολογία PacBio των αντίστοιχων θραυσμάτων DNA του γονιδιώματος ενός

μικροοργανισμού, π.χ. του *Bacillus anthracis*, αναζητούνται στο SRA, όπως ακριβώς γίνεται και με τα δεδομένα των λοιπών τεχνολογιών αλληλούχισης. Τα δεδομένα των reads τα λαμβάνουμε στη μορφή FASTQ. Συγκεκριμένα, λαμβάνουμε, αρχικά, έναν συμπιεσμένο φάκελο με την ονομασία «sra_data.fastq.gz». Τον αποσυμπιέζουμε και παίρνουμε το αρχείο «sra.data.fastq», το οποίο μετονομάζουμε με το όνομα του υπό μελέτη μικροοργανισμού, π.χ. «b_anthracis_pacbio.fastq». Το αρχείο b_anthracis_pacbio.fastq το μεταφέρουμε στον φάκελο SPAdes-3.11.1-Linux όπου βρίσκεται εγκατεστημένο το πρόγραμμα SPAdes.

Έπειτα, ανοίγουμε το τερματικό και εισάγουμε την παρακάτω εντολή για να εκτελεστεί η υβριδική συναρμολόγηση των δεδομένων της Illumina με εκείνα της PacBio:

./bin/spades.py -s b_anthracis_trim.fastq -o b_anthracis_pacbio_illum --pacbio b_anthracis _pacbio.fastq

Τα επεξεργασμένα, με trimming, δεδομένα των reads που προέρχονται από την τεχνολογία Illumina περιέχονται στον φάκελο b_anthracis_trim.fastq, ενώ τα δεδομένα των reads της PacBio στον φάκελο b_anthracis_pacbio.fastq. Με την παράμετρο της εντολής -o δημιουργείται ένα φάκελος με ονομασία «b_anthracis_pacbio_illum» που περιέχει, μεταξύ άλλων, το αρχείο «scaffolds.fasta», όπου τα reads των τεχνολογιών Illumina και PacBio βρίσκονται συναρμολογημένα, συνδυαστικά, σε scaffolds.

3.6 Συναρμολόγηση μικροβιακών γονιδιωμάτων με το πρόγραμμα Canu

Το πρόγραμμα Canu χρησιμοποιεί τον αλγόριθμο Overlap/Layout/Consensus (OLC) και προέρχεται από το λογισμικό Celera Assembler της εταιρείας Celera Genomics. Η έκδοση Canu v1.6 διατίθεται ελεύθερα από τον σύνδεσμο:

https://github.com/marbl/canu/releases/tag/v1.6

Το Canu εξειδικεύεται στη συναρμολόγηση των μεγάλων σε μήκος και χαμηλής ποιότητας reads που χαρακτηρίζονται από σχετικά υψηλό ποσοστό

σφαλμάτων. Δεδομένου του σχετικά υψηλού ποσοστού σφάλματος, η αποτελεσματική και ακριβής συναρμολόγηση μεγάλων επαναλήψεων και στενά συγγενικών απλοτύπων παραμένει μια επιστημονική πρόκληση. Τα ζητήματα αυτά αντιμετωπίζονται με την εφαρμογή του προγράμματος Canu, το οποίο έχει σχεδιαστεί ειδικά για χαμηλής ποιότητας reads. Ο ερευνητής Koren και οι συνεργάτες του απέδειξαν ότι το Canu μπορεί να συναρμολογήσει, αξιόπιστα, ολοκληρωμένα μικροβιακά γονιδιώματα και σχεδόν πλήρη ευκαρυωτικά χρωμοσώματα, χρησιμοποιώντας είτε την τεχνολογία PacBio ή την Oxford Nanopore (Koren et al., 2017).

Το Canu λειτουργεί σε τρεις διαδοχικές φάσεις: διόρθωση, trimming και συναρμολόγηση. Στη φάση της διόρθωσης θα βελτιωθεί η ακρίβεια των βάσεων στα reads. Στη φάση του trimming θα αποκοπούν και θα αφαιρεθούν ολόκληρα ή τμήματα των reads με σκοπό να παραμείνουν μόνον τα reads με αλληλουχία υψηλής ποιότητας. Στη φάση της συναρμολόγησης τα reads θα συναρμολογηθούν σε αλληλουχίες contigs, θα δημιουργηθούν συναινετικές αλληλουχίες και γραφήματα εναλλακτικών διαδρομών. Επιπλέον, το Canu μπορεί να επαναλάβει ατελείς συναρμολογήσεις, επιτρέποντας την αποκατάσταση από διακοπές λειτουργίας του συστήματος ή από άλλους μη φυσιολογικούς τερματισμούς (Canu Quick Start - canu 1.6 documentation).

Στην παράγραφο 3.4, χρησιμοποιήθηκαν τα δεδομένα των reads του βακτηρίου *Bacillus anthracis*, τα οποία δημιουργήθηκαν με την τεχνολογία Illumina, για τη συναρμολόγηση του γονιδιώματος του εν λόγω μικροοργανισμού με το πρόγραμμα SPAdes. Στην παράγραφο 3.5, το SPAdes χρησιμοποιήθηκε και για την υβριδική συναρμολόγηση των δεδομένων των reads που προέκυψαν από τις τεχνολογίες Illumina και PacBio, με σκοπό την πιο έγκυρη συναρμολόγηση του γονιδιώματος του *Bacillus anthracis*. Στην περίπτωση που εξετάζουμε τη συναρμολόγηση των δεδομένων των reads που προέρχονται μόνο από την τεχνολογία PacBio, το πρόγραμμα Canu αποτελεί το κατάλληλο λογισμικό εργαλείο.

Στο εξεταζόμενο παράδειγμα του βακτηρίου *Bacillus anthracis*, μεταφέρουμε το αρχείο b_anthracis_pacbio.fastq εντός του φακέλου «canu-1.6», όπου περιέχεται το πρόγραμμα Canu. Στο αρχείο b_anthracis_pacbio.fastq βρίσκονται τα δεδομένα των reads, σε μορφή

FASTQ, που προέκυψαν από την αλληλούχιση με την τεχνολογία PacBio. Έπειτα, ανοίγουμε το τερματικό και με την κατάλληλη εντολή μεταφερόμαστε στον φάκελο «canu-1.6». Από αυτό το σημείο εισάγουμε την εξής εντολή:

./bin/canu -p b_anthracis -d b_anthracis_pacbio genomeSize=4.8m -pacbio-raw b_anthracis _pacbio.fastq gnuplotTested=true stopOnReadQuality=false

Όπου:

-p, το όνομα που θέλουμε να δώσουμε στο αρχείο όπου θα αποθηκευτεί η συναρμολόγηση.

-d, το όνομα που θέλουμε να δώσουμε στον φάκελο όπου θα αποθηκευτούν όλα τα αποτελέσματα.

genomeSize=, το κατά προσέγγιση μέγεθος του γονιδιώματος του μικροοργανισμού που χρησιμοποιούμε.

-pacbio-raw, το όνομα του αρχείου με τα ακατέργαστα reads.

Με την εκτέλεση της ανωτέρω εντολής δημιουργείται ένας φάκελος με όνομα «b_anthracis_pacbio» εντός του φακέλου «canu-1.6», ο οποίος περιέχει, μεταξύ άλλων, ένα αρχείο με όνομα «b_anthracis.contigs.fasta» με τα contigs που δημιουργήθηκαν και ένα αρχείο με όνομα «b_anthracis.unassembled.fasta» που περιέχει τα reads που δε συναρμολογήθηκαν σε contigs. Επιπλέον, υπάρχει ένας συμπιεσμένος φάκελος με όνομα «b_anthracis.correctedReads.fasta.gz» από τον οποίο εξάγουμε, μεταξύ άλλων, το αρχείο «b_anthracis.correctedReads.fasta» και ένας ακόμα συμπιεσμένος φάκελος με όνομα «b_anthracis.trimmedReads.fasta.gz» από τον οποίο εξάγουμε, μεταξύ άλλων, το αρχείο «b_anthracis.trimmedReads.fasta».

3.7 Απεικόνιση της στοίχισης των reads πάνω στο γονιδίωμα με τα προγράμματα BLASR και IGV

Το BLASR (Basic Local Alignment with Successive Refinement) είναι ένα εκτελέσιμο αρχείο, το οποίο είχε καταρτιστεί, αρχικά, στην έκδοση του Ubuntu 12.04. Μπορεί, με μεγάλη ακρίβεια και ταχύτητα, να στοιχίσει τα μεγάλα σε

μήκος reads που προέρχονται από την τεχνολογία PacBio με γονιδιώματα αναφοράς (Chaisson και Tesler, 2012). Διατίθεται ελεύθερα από το GitHub μέσω του συνδέσμου:

https://github.com/ylipacbio/blasrbinary/raw/master/blasr

Συνεχίζοντας στη χρήση του παραδείγματος για το βακτήριο *Bacillus anthracis*, το πρόγραμμα BLASR χρησιμοποιεί τα αρχεία b_anthracis.trimmedReads.fasta και b_anthracis.contigs.fasta, τα οποία δημιουργήθηκαν από την εκτέλεση του προγράμματος Canu, με σκοπό να ευθυγραμμίσει τα επεξεργασμένα reads που περιέχονται στο αρχείο b_anthracis.trimmedReads.fasta με τα contigs που βρίσκονται στο αρχείο b_anthracis.contigs.fasta.

Ειδικότερα, αφού εκτελεστεί η *de novo* συναρμολόγηση του γονιδιώματος με το πρόγραμμα Canu, μεταφέρουμε τα αρχεία b_anthracis.contigs.fasta και b_anthracis.trimmedReads.fasta στον φάκελο «blasrbinary-master» που περιέχει το πρόγραμμα BLASR και από το τερματικό μεταβαίνουμε με την κατάλληλη εντολή εντός του φακέλου blasrbinary-master. Έπειτα, εισάγουμε την παρακάτω εντολή:

./blasr b_anthracis.trimmedReads.fasta b_anthracis.contigs.fasta -sam -header -nproc 6 -minMatch 15 -bestn 1 - b_anthracis_out.sam -unaligned unaligned_reads.fa

Όπου:

-sam, για να μας δώσει το αποτέλεσμα σε αρχείο μορφής SAM.

-out, δίνουμε το όνομα του αρχείου που θα συγκεντρώνει όλα τα αποτελέσματα.

-unaligned, δίνουμε το όνομα του αρχείου που θα αποθηκευτούν τα reads που δεν έχουν ευθυγραμμιστεί πάνω στα contigs.

Με την εκτέλεση της ως άνω εντολής δημιουργούνται τα αρχεία «b_anthracis_out.sam» και «unaligned_reads.fa».

Η μορφή αρχείου SAM (Sequence Alignment Map) αναπτύχθηκε από τον Heng Li και τους συνεργάτες του το 2009 και είναι μια μορφή βασισμένη σε κείμενο (text-based format) για την αποθήκευση των νουκλεοτιδικών αλληλουχιών που είναι στοιχισμένες με μια αλληλουχία αναφοράς. Χρησιμοποιείται ευρέως για την αποθήκευση δεδομένων, όπως αλληλουχίες

νουκλεοτιδίων, που παράγονται από τις τεχνολογίες αλληλούχισης νέας γενιάς (Li et al., 2009). Η μορφή SAM υποστηρίζει μικρές και μεγάλες αλληλουχίες reads, έως 128 Mbp, που παράγονται από διάφορες τεχνολογίες αλληλούχισης.

Το πρόγραμμα IGV (Integrative Genomics Viewer) είναι ένα λογισμικό εργαλείο οπτικοποίησης υψηλής απόδοσης για τη διαδραστική εξερεύνηση μεγάλων και ολοκληρωμένων γονιδιωματικών δεδομένων. Υποστηρίζει ευρεία ποικιλία τύπων δεδομένων, συμπεριλαμβανομένων των δεδομένων reads που προέρχονται από την αλληλούχιση νέας γενιάς (Robinson et al., 2011 / Thorvaldsdóttir et al., 2013). Διατίθεται ελεύθερα από τον σύνδεσμο:

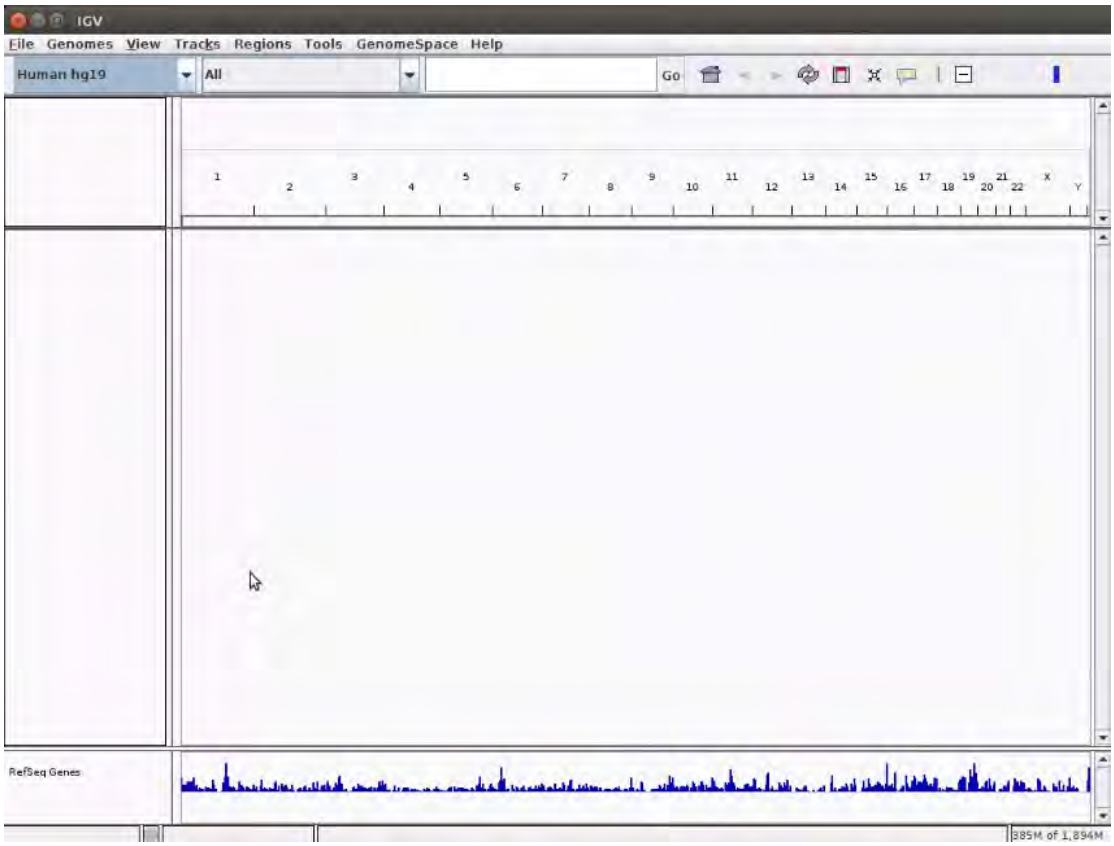
http://software.broadinstitute.org/software/igv/download

Γενικά, με το πρόγραμμα IGV γίνεται η απεικόνιση των reads σε σύγκριση με ένα γονιδίωμα αναφοράς, με ένα scaffold, ή με ένα contig.

Για τον σκοπό της περιγραφής χρησιμοποιήθηκε η έκδοση «IGV 2.4.3». Συγκεκριμένα, από το τερματικό, με την κατάλληλη εντολή μεταφερόμαστε στον φάκελο «IGV_2.4.3», στον οποίο περιέχεται το πρόγραμμα IGV. Έπειτα, εισάγουμε την εντολή:

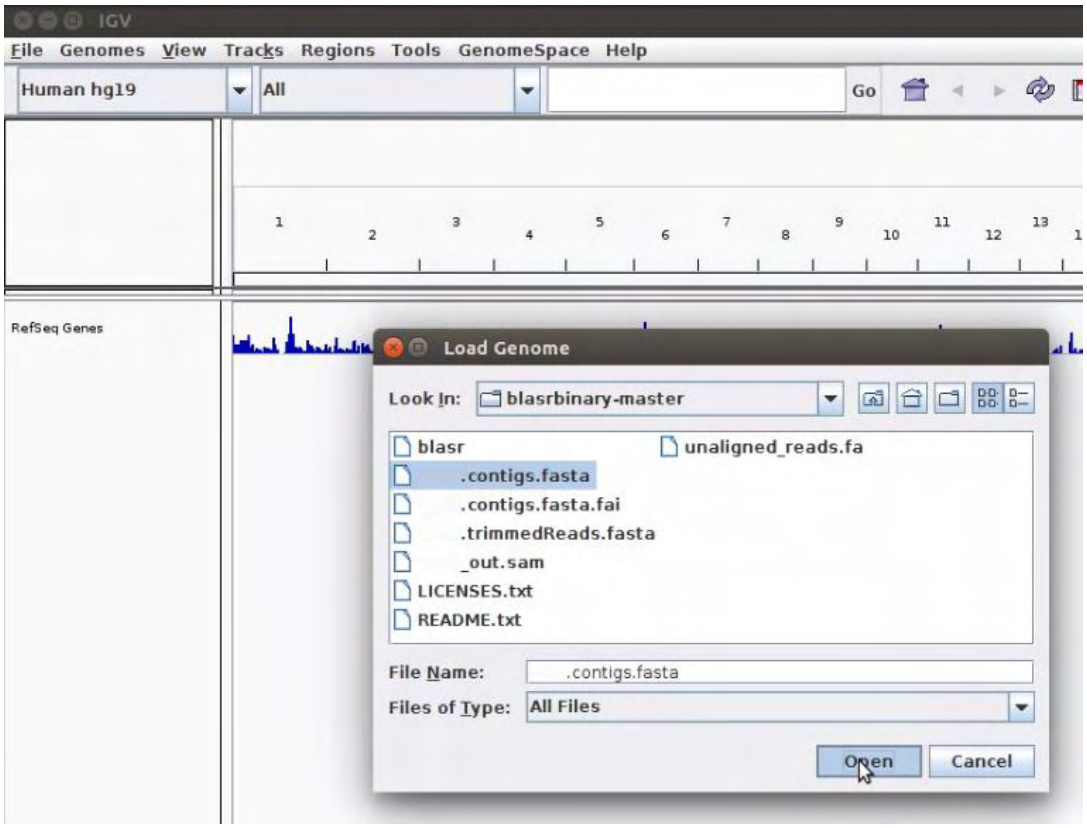
./igv.sh

με την οποία ανοίγει το πρόγραμμα σε ένα παράθυρο με την ονομασία «IGV» στην οθόνη του υπολογιστή, όπως το παράθυρο που απεικονίζεται στην Εικόνα 24.



Εικόνα 24 Παράθυρο έναρξης της έκδοσης 2.4.3 του προγράμματος IGV.

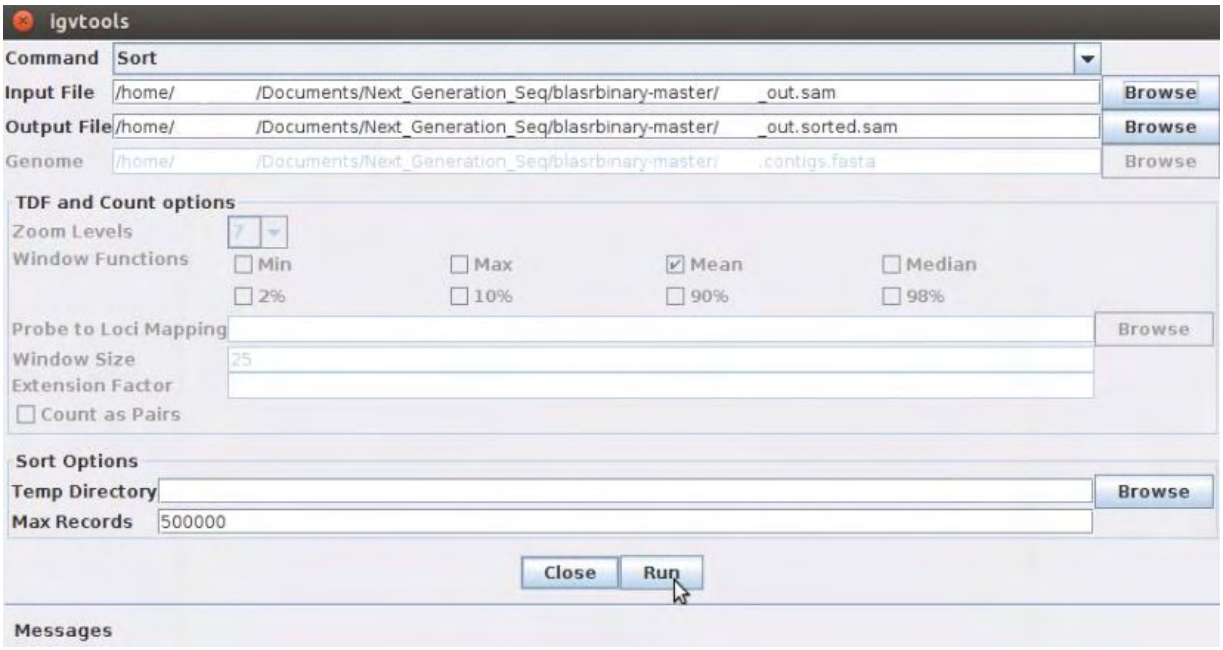
Σε αυτό το παράθυρο επιλέγουμε την ενότητα «Genomes» και από εκεί επιλέγουμε «Load Genome from file». Με αυτόν τον τρόπο ανοίγει ένα δεύτερο, μικρότερο, παράθυρο που ονομάζεται «Load Genome», το οποίο φαίνεται εντός του παραθύρου IGV στην Εικόνα 25.



Εικόνα 25 Από το παράθυρο έναρξης της Εικόνας 24, επιλέγοντας την ενότητα Genomes και κατόπιν Load Genome from file ανοίγει το μικρό παράθυρο Load Genome. Από την ενότητα «Look In:» του παραθύρου Load Genome ανοίγουμε τον φάκελο blasrbinary-master του προγράμματος BLASR, από όπου επιλέγουμε να ανοίξει ένα αρχείο «.contigs.fasta».

Στο παράθυρο «Load Genome» και από την ενότητα «Look In:» ανοίγουμε τον φάκελο blasrbinary-master του προγράμματος BLASR, από όπου επιλέγουμε να ανοίξει ένα αρχείο «.contigs.fasta», π.χ. το b_anthracis.contigs.fasta.

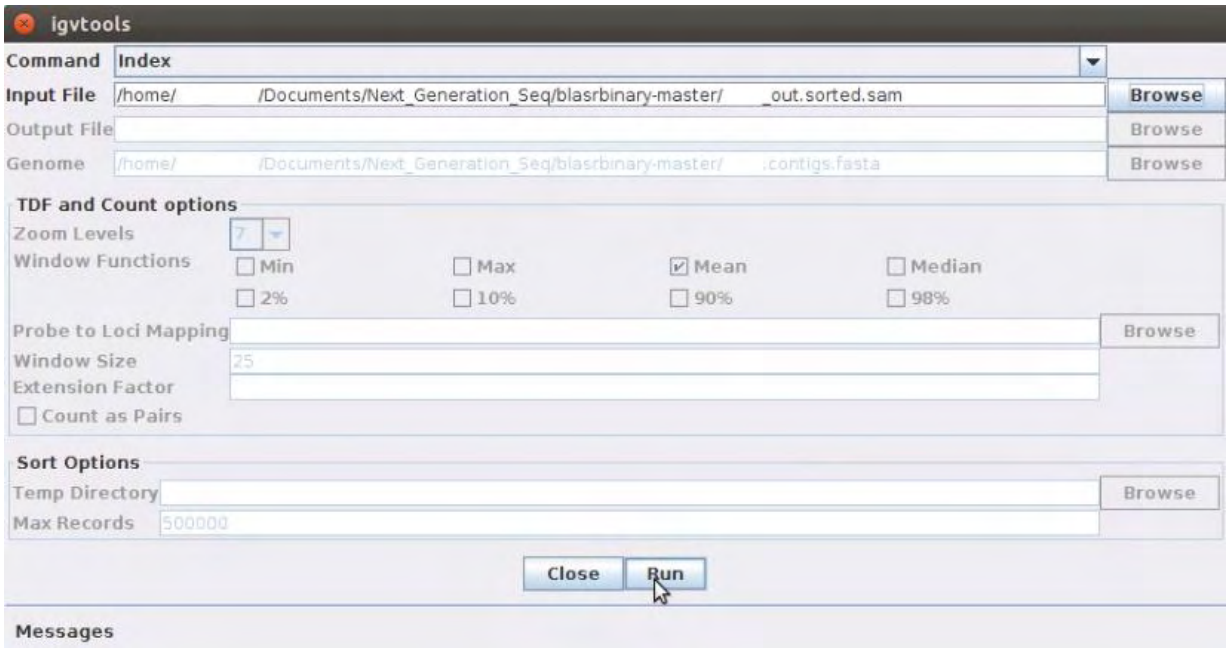
Στη συνέχεια, στο (μεγάλο) παράθυρο IGV, επιλέγουμε «Tools» και από εκεί επιλέγουμε «Run igvtools» και ανοίγει το παράθυρο «Igvtools» που απεικονίζεται στην Εικόνα 26.



Εικόνα 26 Στο (μεγάλο) παράθυρο IGV της Εικόνας 25, επιλέγουμε Tools και έπειτα Run igvtools. Έτσι ανοίγει το παράθυρο Igvtools. Από την ενότητα Command επιλέγουμε Sort και στην ενότητα Input File ανεβάζουμε, από το Browse, ένα αρχείο «_out.sam», το οποίο προέκυψε από την εφαρμογή του προγράμματος BLASR και περιέχει εκείνα τα reads που έχουν στοιχηθεί επάνω στα contigs. Εκτελώντας Run, στην ενότητα Output File θα φανεί ότι δημιουργήθηκε, εντός του φακέλου blasrbinary-master, ένα αρχείο με ονομασία «_out.sorted.sam».

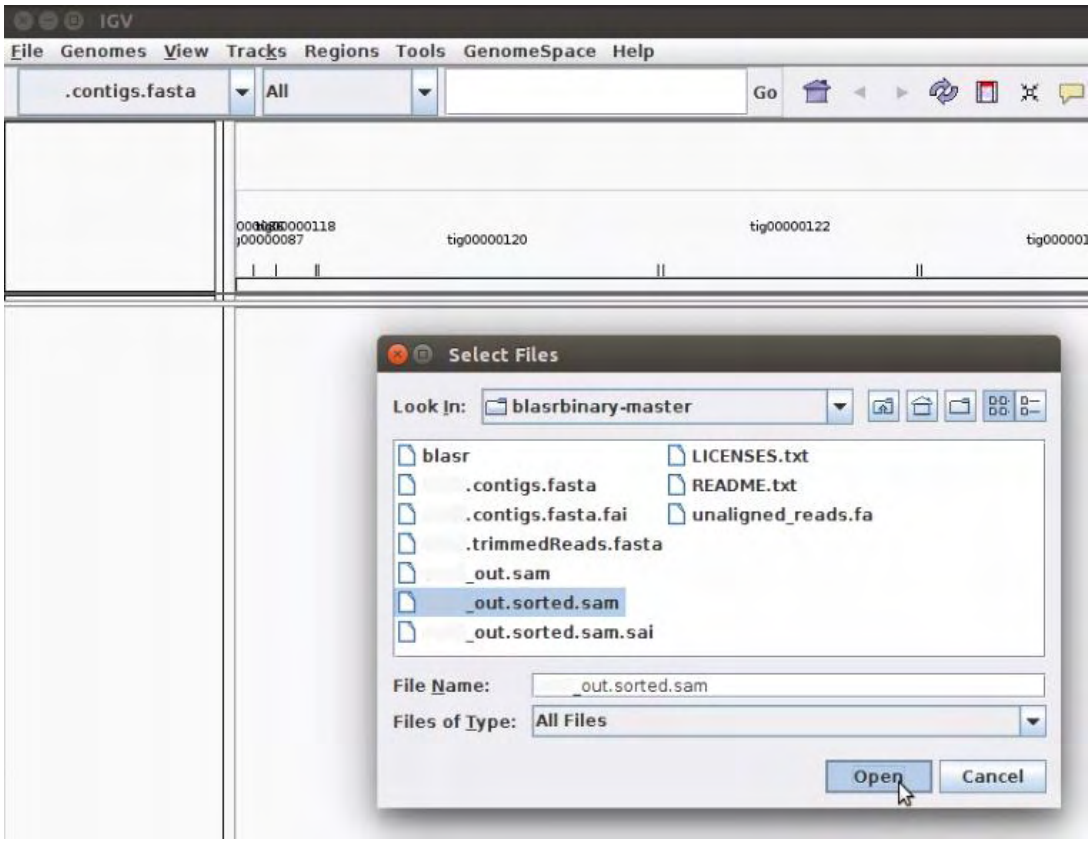
Από την ενότητα «Command» επιλέγουμε «Sort» και στην ενότητα «Input File» ανεβάζουμε, από το Browse, ένα αρχείο «_out.sam», π.χ. το b_anthracis_out.sam, το οποίο προέκυψε από την εφαρμογή του προγράμματος BLASR και περιέχει εκείνα τα reads που έχουν στοιχηθεί επάνω στα contigs του αρχείου b_anthracis.contigs.fasta. Κατόπιν των παραπάνω ενεργειών εκτελούμε «Run». Στην ενότητα «Output File» θα φανεί ότι δημιουργήθηκε, εντός του φακέλου blasrbinary-master, ένα αρχείο με ονομασία «_out.sorted.sam», π.χ. το «b_anthracis_out.sorted.sam».

Έπειτα, στο παράθυρο Igvtools, όπως απεικονίζεται στην Εικόνα 27, από την ενότητα «Command» επιλέγουμε «Index», ενώ ως «Input File» θα ανεβάσουμε το αρχείο _out.sorted.sam που δημιουργήθηκε στο προηγούμενο βήμα και θα εκτελέσουμε «Run».



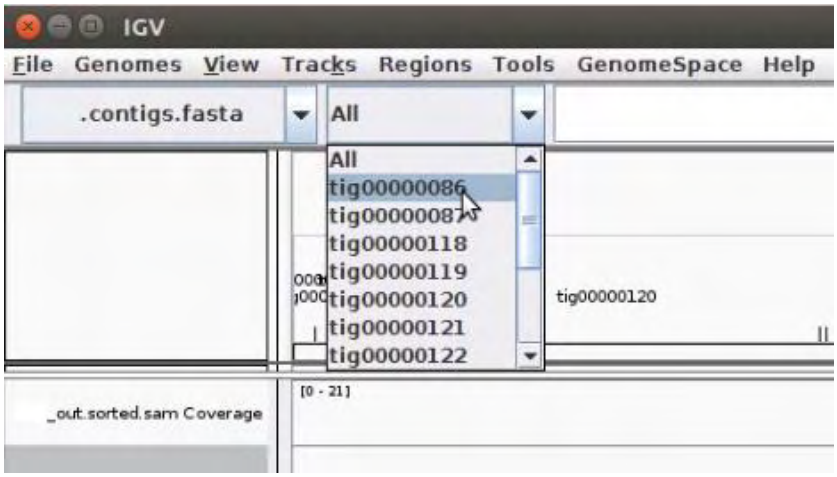
Εικόνα 27 Από την ενότητα Command επιλέγουμε Index, ενώ ως Input File ανεβάζουμε το αρχείο «_out.sorted.sam» και εκτελούμε Run.

Θα ανοίξει ένα νέο παράθυρο IGV, όπως αυτό της Εικόνας 28 και από την ενότητα «File» θα επιλέξουμε «Load from File». Στο (μικρό) αναδυόμενο παράθυρο με την ονομασία «Select Files» ανοίγουμε το αρχείο _out.sorted.sam από τον φάκελο blasrbinary-master.



Εικόνα 28 Εκτελώντας Run στο παράθυρο της Εικόνας 27 ανοίγει το νέο παράθυρο IGV. Από την ενότητα File θα επιλέξουμε Load from File. Στο μικρό αναδυόμενο παράθυρο Select Files ανοίγουμε το αρχείο _out.sorted.sam από τον φάκελο blasrbinary-master.

Κατόπιν, από την ενότητα «All» του παραθύρου IGV επιλέγουμε «tig00000086», όπως απεικονίζεται στην Εικόνα 29.



Εικόνα 29 Από την ενότητα All του παραθύρου IGV επιλέγουμε «tig00000086».

Όπως φαίνεται στην Εικόνα 30, μόλις επιλέξουμε «tig00000086», στο παράθυρο IGV απεικονίζεται το σύνολο των reads ως γκρι ραβδόμορφα και στενόμακρα σχήματα που προβάλλονται σε οριζόντια διάταξη και τα οποία δύνανται να επικαλύπτονται, είτε ολόκληρα ή σε τμήματά τους, από άλλα reads που βρίσκονται από πάνω ή από κάτω. Επιπροσθέτως, μπορούμε να έχουμε και άποψη του οπτικοποιημένου μήκους των αλληλουχιών τους, καθώς και των διαφορών στις βάσεις των νουκλεοτιδικών αλληλουχιών τους, όπως π.χ. η προσθήκη βάσεων σε μια συγκεκριμένη θέση ενός read, η οποία συμβολίζεται με μια κάθετη μωβ γραμμή. Η οπτικοποίηση βασίζεται στα δεδομένα του αρχείου «_out.sorted.sam» σε σχέση με τα contigs του αρχείου «.contigs.fasta».

Εικόνα 30 Επιλέγοντας «tig00000086», απεικονίζεται το σύνολο των reads ως γκρι ραβδόμορφα και στενόμακρα σχήματα που προβάλλονται σε οριζόντια διάταξη και τα οποία δύνανται να επικαλύπτονται, είτε ολόκληρα ή σε τμήματά τους, από άλλα reads που βρίσκονται από πάνω ή από κάτω. Μπορούμε να έχουμε και άποψη του οπτικοποιημένου μήκους των αλληλουχιών τους, καθώς και των διαφορών στις βάσεις των νουκλεοτιδικών αλληλουχιών τους. Η οπτικοποίηση βασίζεται στα δεδομένα ενός αρχείου «_out.sorted.sam» σε σχέση με τα contigs του αντίστοιχου αρχείου «.contigs.fasta».

Έχουμε, λοιπόν, την πλήρη απεικόνιση της αλληλοεπικάλυψης (coverage) μεταξύ των αλληλουχιών των reads και τη σχηματική αποτύπωση του μεγέθους της στην υπό ένδειξη «_out.sorted.sam Coverage» περιοχή, καθώς και την απεικόνιση της ευθυγράμμισης των reads σε σχέση με τις αλληλουχίες των contigs. Με τα σύμβολα «+» και «-», μπορούμε να εστιάσουμε, αντίστοιχα, εντός και εκτός του πεδίου της εικονιζόμενης προβολής, ενώ με το κόκκινο στενόμακρο πλαίσιο μπορούμε να εξετάσουμε την ευθυγράμμιση των reads σε διάφορα σημεία επί του συνόλου των αλληλουχιών των contigs.

3.8 Αναζήτηση συγγενικού γονιδιώματος με το NCBI BLAST

Το BLAST (Basic Local Alignment Search Tool) είναι ένα από τα πιο ευρέως χρησιμοποιούμενα προγράμματα βιοπληροφορικής (Casey, 2005). Πρόκειται για έναν αλγόριθμο αναζήτησης ομόλογων αλληλουχιών. Η αναζήτηση με το BLAST δίνει τη δυνατότητα στον χρήστη να συγκρίνει μία αλληλουχία επερώτησης (query sequence) με αλληλουχίες μιας βάσης δεδομένων και να αναγνωρίσει αλληλουχίες της βάσης αυτής που μοιάζουν, λόγω ομολογίας, με την αλληλουχία επερώτησης πάνω από ένα συγκεκριμένο όριο. Ο αλγόριθμος BLAST και το πρόγραμμα που τον υλοποιεί σχεδιάστηκαν το 1990 από τον Altschul και τους συνεργάτες του (Altschul et al., 1990). Το πρόγραμμα BLAST λειτουργεί ως on line εφαρμογή και διατίθεται ελεύθερα στον σύνδεσμο:

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch

τον οποίο διαχειρίζεται το Εθνικό Κέντρο Βιοτεχνολογικών Πληροφοριών των Ηνωμένων Πολιτειών (NCBI).

Βασικό προτέρημα του αλγόριθμου που χρησιμοποιεί το BLAST είναι η ταχύτητα, γεγονός που τον καθιστά ιδιαίτερα πρακτικό σε βάσεις δεδομένων τεραστίων διαστάσεων. Πριν από το BLAST, είχε αναπτυχθεί το λογισμικό πρόγραμμα FASTA από τους Lipman και Pearson το 1985 (Lipman και Pearson, 1985) και, γενικότερα, πριν αναπτυχθούν οι αλγόριθμοι FASTA και BLAST, η διεξαγωγή των αναζητήσεων για αλληλουχίες νουκλεοτιδίων ή

πρωτεϊνών σε βάσεις δεδομένων ήταν πολύ χρονοβόρα, επειδή χρησιμοποιούνταν αλγόριθμοι δυναμικού προγραμματισμού, όπως ο αλγόριθμος Smith-Waterman. Το BLAST, ωστόσο, είναι πιο γρήγορο.

Υπάρχουν διαθέσιμα διάφορα προγράμματα BLAST, τα οποία διαφέρουν στο είδος της εισαχθείσας αλληλουχίας επερώτησης και της χρησιμοποιούμενης βάσης δεδομένων και των δεδομένων που τίθενται προς σύγκριση. Στην πραγματικότητα, το BLAST είναι μια ομάδα προγραμμάτων, τα οποία περιλαμβάνονται στο εκτελέσιμο αρχείο «blastall» (BLAST Basic Local Alignment Search Tool, Blast Program Selection Guide) και είναι τα εξής:

- **Nucleotide-nucleotide BLAST (blastn)**

Χρησιμοποιεί μια αλληλουχία DNA, ως αλληλουχία επερώτησης και προβάλλει τις πιο όμοιες αλληλουχίες DNA από τη βάση δεδομένων που καθορίζει ο χρήστης.

- **Protein-protein BLAST (blastp)**

Μια πρωτεϊνική αλληλουχία αμινοξέων χρησιμοποιείται ως αλληλουχία επερώτησης και προβάλλονται οι πιο όμοιες πρωτεϊνικές αλληλουχίες από την πρωτεϊνική βάση δεδομένων που καθορίζει ο χρήστης.

- **Position-Specific Iterative BLAST (PSI-BLAST, blastpgp)**

Χρησιμοποιείται για την εύρεση απομακρυσμένων ομολόγων μιας πρωτεΐνης. Αρχικά, δημιουργείται ένας κατάλογος όλων των στενά συγγενικών πρωτεϊνών. Αυτές οι πρωτεΐνες δημιουργούν ένα «προφίλ», το οποίο χρησιμοποιείται για να αναζητηθούν ομόλογες ακολουθίες στη βάση δεδομένων. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να μην βρεθούν νέες ακολουθίες.

- **Blastx**

Με αυτό το πρόγραμμα γίνεται αναζήτηση προϊόντων μετάφρασης στα 6 αναγνωστικά πλαίσια μιας νουκλεοτιδικής ακολουθίας σε βάση δεδομένων πρωτεϊνικών αλληλουχιών.

- **Tblastn**

Χρησιμοποιείται για την αναζήτηση σε βάσεις δεδομένων μεταφρασμένων νουκλεοτιδίων (στα 6 αναγνωστικά πλαίσια), χρησιμοποιώντας ως ακολουθία επερώτησης μια πρωτεΐνη.

- **Nucleotide 6-frame translation-nucleotide 6-frame translation (tblastx)**

Είναι το πιο αργό πρόγραμμα της ομάδας BLAST. Η ακολυθία επερώτησης είναι νουκλεοτιδική που μεταφράζεται στα 6 αναγνωστικά πλαίσια και οι ακολουθίες στη βάση δεδομένων είναι νουκλεοτιδικές που μεταφράζονται στα 6 αναγνωστικά πλαίσια. Ουσιαστικά, γίνονται 6Χ6 Blastp.

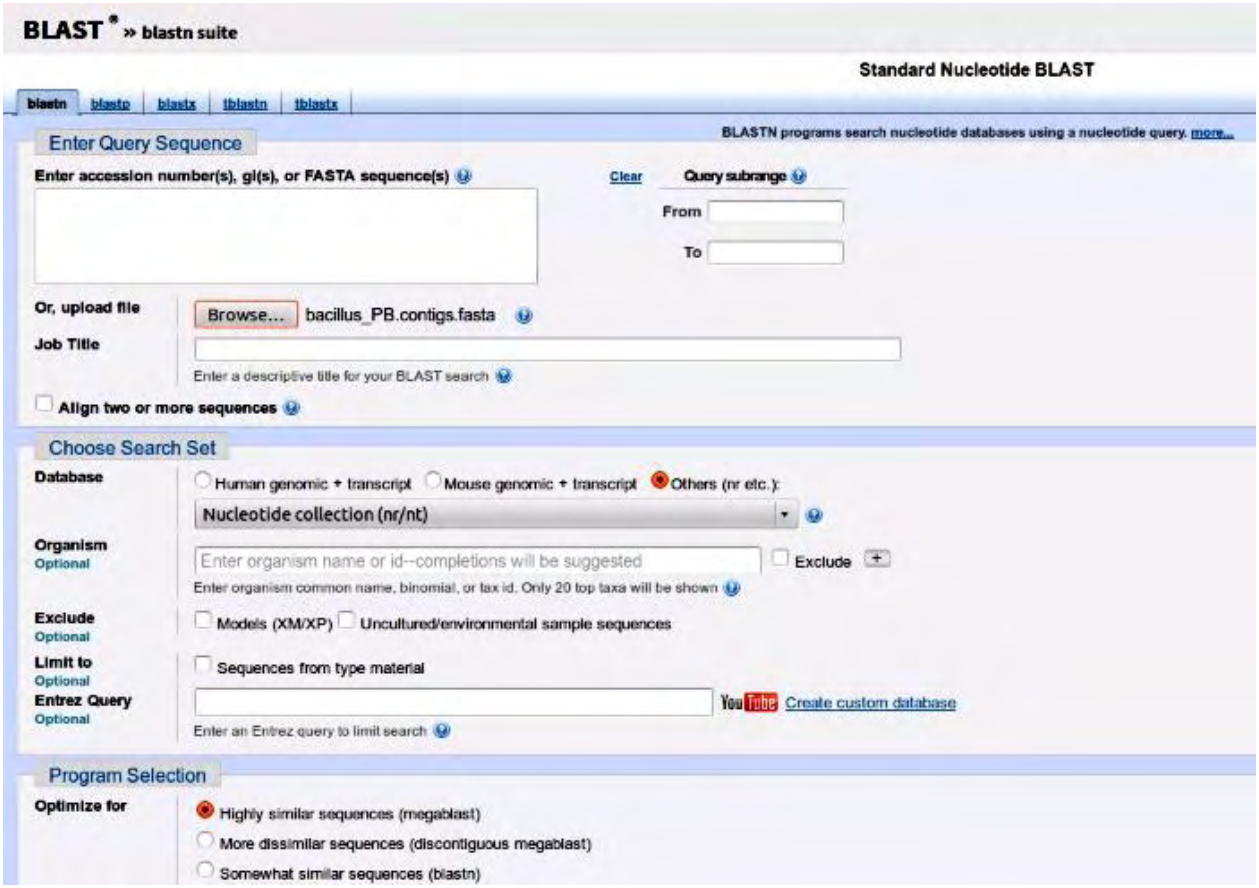
- **Megablast**

Το MegaBlast αναζητεί πολύ γρήγορα νουκελοτιδικές ακολουθίες που έχουν όμως μεγάλο ποσοστό ταύτισης με την νουκλεοτιδική ακολουθία επερώτησης.

Σε συνέχεια της περιγραφόμενης μεθοδολογίας με σκοπό την ανάλυση ενός μικροβιακού γονιδιώματος, κατόπιν της περαίωσης των διαδικασιών της συναρμολόγησής του, θα ήταν ιδιαίτερα χρήσιμο να βρεθεί το πιο συγγενικό γονιδίωμα που είναι αποθηκευμένο στη βάση δεδομένων του BLAST.

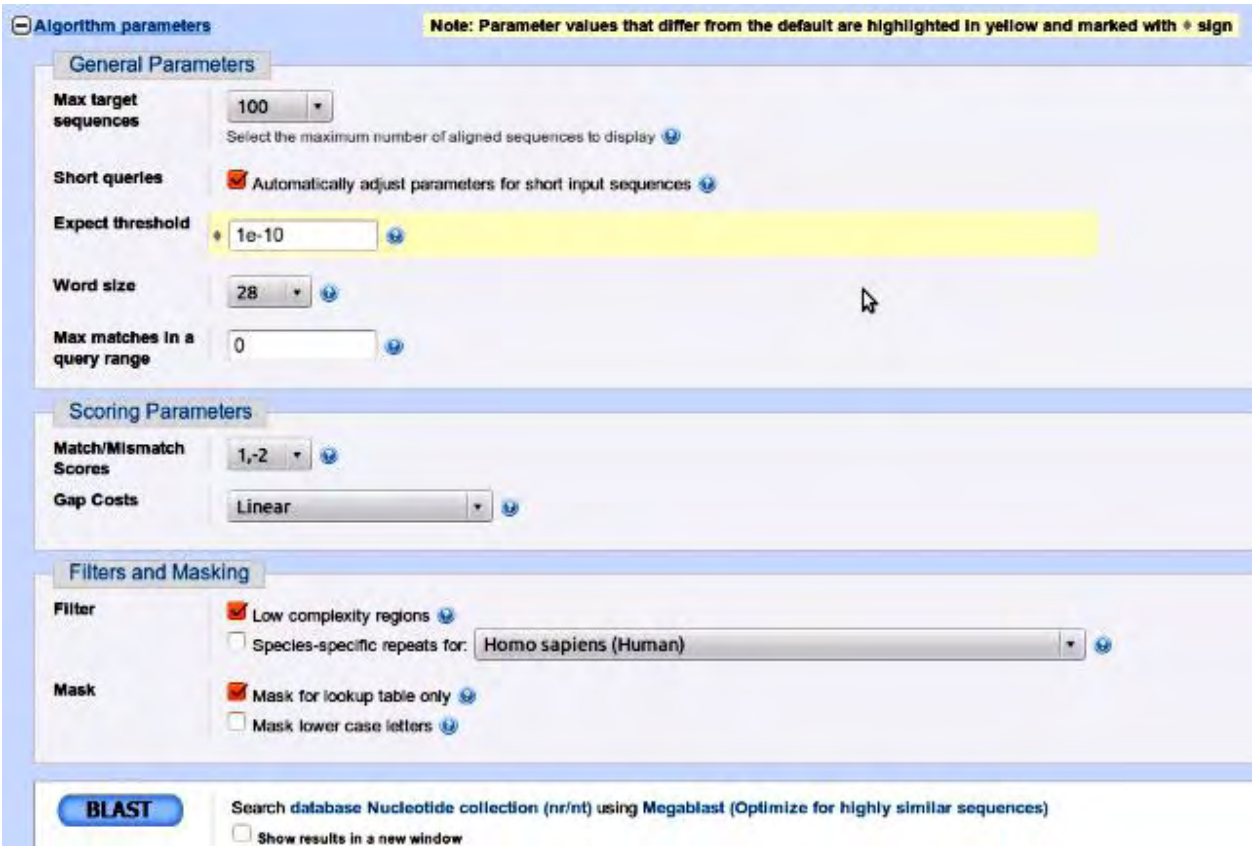
Όπως φαίνεται στο παράδειγμα της Εικόνας 31, στο BLAST μεταφορτώνεται το αρχείο του τύπου «.contigs.fasta», το οποίο περιέχει τα δεδομένα των αλληλουχιών reads του υπό μελέτη μικροοργανισμού που είχαν συναρμολογηθεί σε contigs, μέσω της εφαρμογής του προγράμματος Canu.

Συγκεκριμένα, στη μηχανή αναζήτησης του διαδικτύου πληκτρολογούμε «blast» και επιλέγουμε την τοποθεσία «BLAST: Basic Local Alignment Search Tool - NCBI - NIH». Ακολούθως, επιλέγουμε «Nucleotide BLAST» και αναδύεται το παράθυρο της Εικόνας 31 με ονομασία «BLAST® » blastn suite». Από το «Browse», στην ενότητα «Enter Query Sequence», μεταφορτώνεται το επιθυμητό αρχείο του τύπου «.contigs.fasta». Στην ενότητα «Choose Search Set - Database» επιλέγουμε «Others (nr etc.)», ενώ στην ενότητα «Program Selection - Optimize for» επιλέγουμε « Highly similar sequences (megablast)».



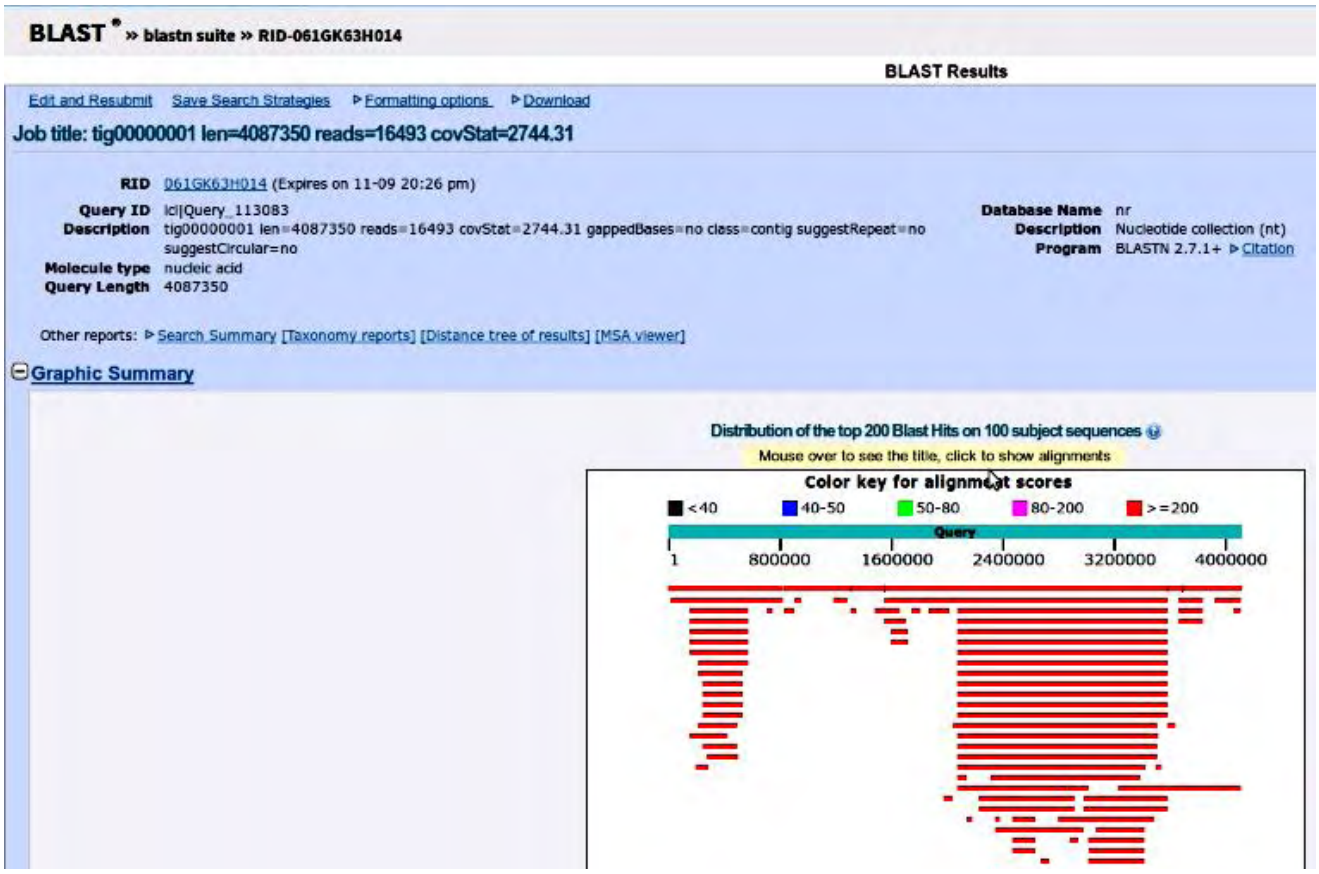
Εικόνα 31 Στο BLAST μεταφορτώνεται το αρχείο του τύπου «.contigs.fasta», το οποίο περιέχει τα δεδομένα των αλληλουχιών reads του υπό μελέτη μικροοργανισμού που είχαν συναρμολογηθεί σε contigs, μέσω της εφαρμογής του προγράμματος Canu.

Στο ίδιο παράθυρο (BLAST® » blastn suite), και όπως απεικονίζεται στην Εικόνα 32, μεταβαίνουμε από την ενότητα «Algorithm parameters» στην υποενότητα «General Parameters - Expect threshold», όπου εισάγουμε την τιμή «1e-10». Αφού ολοκληρώσουμε αυτά τα βήματα, εκτελούμε «BLAST».



Εικόνα 32 Από την ενότητα Algorithm parameters μεταβαίνουμε στην υποενότητα General Parameters - Expect threshold, όπου εισάγουμε την τιμή «1e-10». Αφού ολοκληρώσουμε αυτά τα βήματα, εκτελούμε BLAST.

Μόλις εκτελεστεί το BLAST, θα ανοίξει ένα παράθυρο με ονομασία «BLAST Results», σαν αυτό που φαίνεται στην Εικόνα 33. Στο σχήμα της ενότητας «Graphic Summary» παρουσιάζεται με γαλάζιο χρώμα η συναρμολογημένη αλληλουχία του υπό μελέτη μικροοργανισμού. Η αλληλουχία αυτή είναι η αλληλουχία επερώτησης, η οποία απεικονίζεται ως «Query». Με παράλληλες κόκκινες γραμμές, κάτω από τη γαλάζια γραμμή, εμφανίζονται τα συγγενικά γονιδιώματα μικροοργανισμών που βρίσκονται στη βάση δεδομένων.



Εικόνα 33 Μόλις εκτελεστεί το BLAST από το παράθυρο της Εικόνας 32, θα ανοίξει το παράθυρο BLAST Results. Στο σχήμα της ενότητας Graphic Summary παρουσιάζεται με γαλάζιο χρώμα η συναρμολογημένη αλληλουχία του υπό μελέτη μικροοργανισμού. Η αλληλουχία αυτή είναι η αλληλουχία επερώτησης (Query). Με παράλληλες κόκκινες γραμμές, κάτω από τη γαλάζια γραμμή, εμφανίζονται τα συγγενικά γονιδιώματα μικροοργανισμών. Η κόκκινη γραμμή που βρίσκεται πιο ψηλά είναι εκείνη που παρουσιάζει τη μεγαλύτερη συγγένεια με την αλληλουχία επερώτησης.

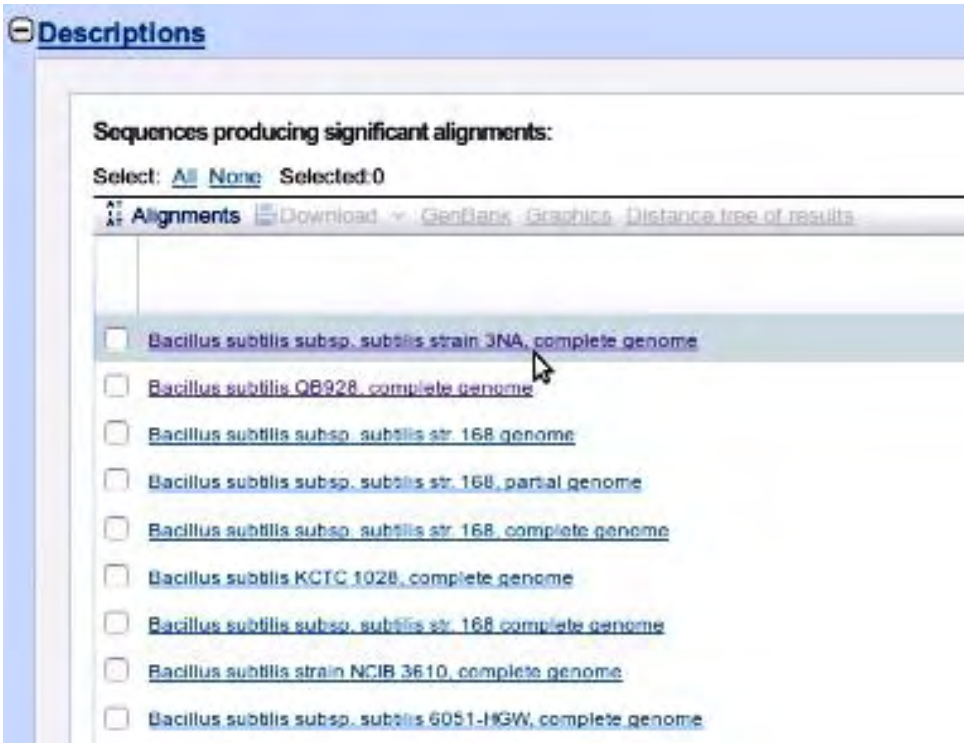
Έτσι λοιπόν στο σχήμα της ενότητας «Graphic Summary», η κόκκινη γραμμή που βρίσκεται πιο ψηλά, και αμέσως κάτω από τη γαλάζια, είναι εκείνη που παρουσιάζει τη μεγαλύτερη συγγένεια με την αλληλουχία επερώτησης. Σε ορισμένα σημεία της παρουσιάζει μικρές κάθετες μαύρες γραμμές, οι οποίες συμβολίζουν τα σημεία των νουκλεοτιδικών διαφορών με την αλληλουχία επερώτησης (γαλάζια γραμμή-Query).

Συμπερασματικά, η πρώτη κόκκινη γραμμή αντιστοιχεί στον πιο στενά συγγενικό μικροοργανισμό με τον υπό μελέτη μικροοργανισμό. Για να διαπιστώσουμε το όνομά του θέτουμε τον κέρσορα επάνω στην κόκκινη γραμμή και με αυτόν τον τρόπο απεικονίζονται τα στοιχεία του, όπως

αποτυπώνεται και στο παράδειγμα της Εικόνας 34. Τις περισσότερες φορές πρόκειται για κάποιο άλλο πολύ στενά συγγενικό στέλεχος του ιδίου είδους.

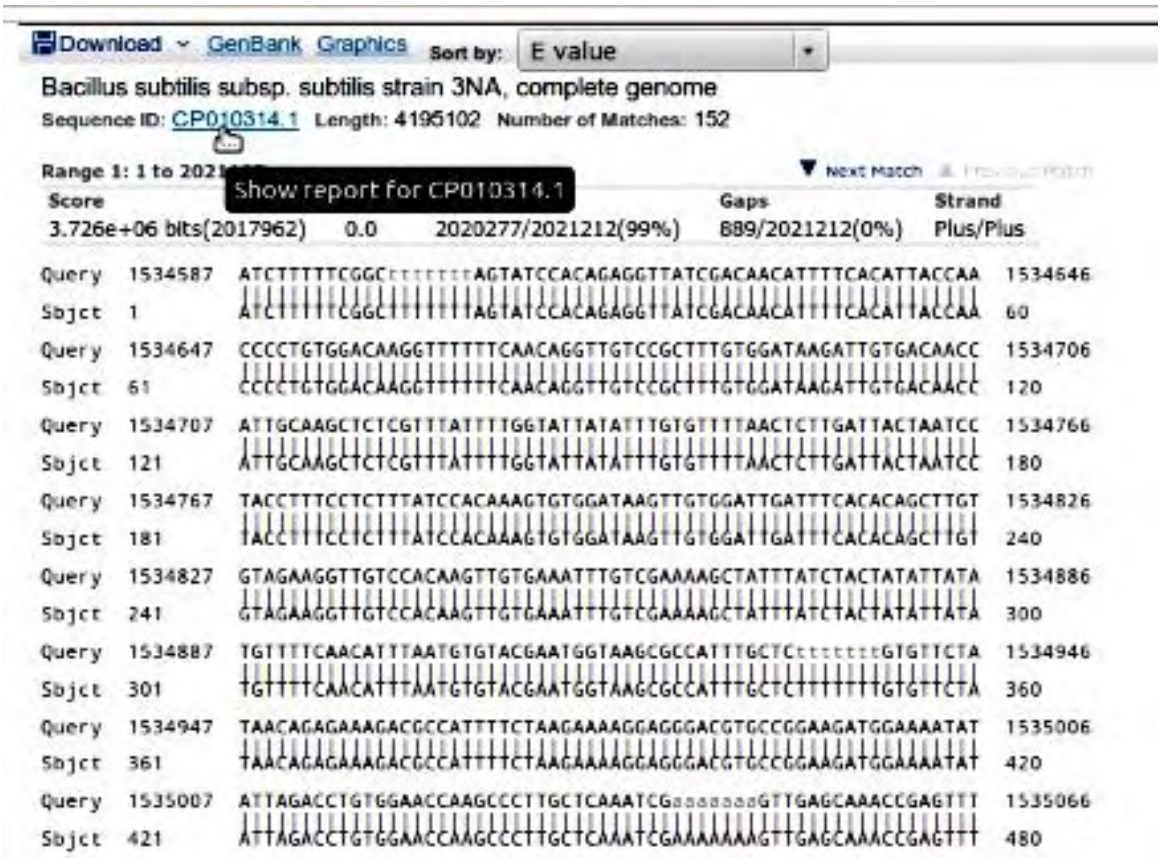
Εικόνα 34 Η πρώτη κόκκινη γραμμή αντιστοιχεί στον πιο στενά συγγενικό μικροοργανισμό με τον υπό μελέτη μικροοργανισμό. Για να διαπιστώσουμε το όνομά του θέτουμε τον κέρσορα επάνω στην κόκκινη γραμμή και με αυτόν τον τρόπο απεικονίζονται τα στοιχεία του.

Στην ενότητα «Descriptions» του παραθύρου «BLAST Results» που απεικονίζεται στην Εικόνα 35, αναγράφονται, σε αντιστοιχία με τις κόκκινες γραμμές του σχήματος της ενότητας «Graphic Summary», τα ονόματα των πιο συγγενικών μικροοργανισμών, με πρώτο το όνομα εκείνου που είναι περισσότερο συγγενικός με τον υπό μελέτη μικροοργανισμό και με τελευταίο του μικροοργανισμού που είναι λιγότερο συγγενικός.



Εικόνα 35 Στην ενότητα Descriptions του παραθύρου BLAST Results αναγράφονται, σε αντιστοιχία με τις κόκκινες γραμμές του σχήματος της ενότητας Graphic Summary, τα ονόματα των πιο συγγενικών μικροοργανισμών, με πρώτο το όνομα εκείνου που είναι περισσότερο συγγενικός.

Επιλέγοντας, για παράδειγμα, το πρώτο όνομα της ενότητας «Descriptions» ανοίγει ένα παράθυρο, όπως εκείνο της Εικόνας 36, στο οποίο εμφανίζεται σε ευθυγράμμιση η νουκλεοτιδική αλληλουχία του πιο συγγενικού μικροοργανισμού (Sbjct) με την αλληλουχία επερώτησης (Query) του υπό μελέτη μικροοργανισμού.

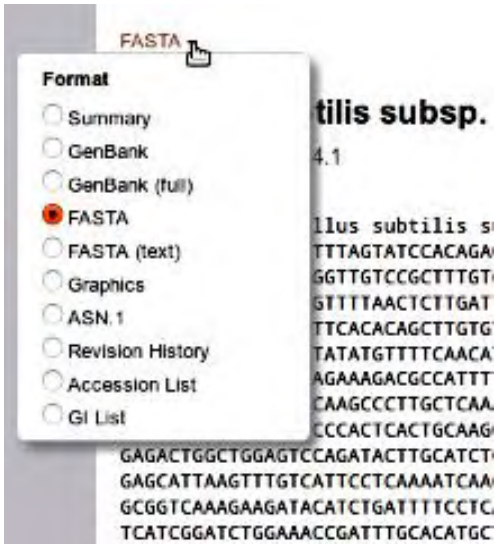


Εικόνα 36 Επιλέγοντας το πρώτο όνομα του μικροοργανισμού από την ενότητα Descriptions, εμφανίζεται, σε ευθυγράμμιση, η νουκλεοτιδική αλληλουχία του πιο συγγενικού μικροοργανισμού (Sbjct) με την αλληλουχία επερώτησης (Query) του υπό μελέτη μικροοργανισμού.

Ακολούθως, επιλέγοντας τον κωδικό «Sequence ID» του πιο συγγενικού μικροοργανισμού, ανοίγει το παράθυρο αναφοράς του στη «GenBank», παράδειγμα του οποίου απεικονίζεται στην Εικόνα 37. Επάνω και αριστερά στο παράθυρο της Εικόνας 37, κάτω από την ονομασία του μικροοργανισμού και τον κωδικό GenBank, επιλέγουμε «FASTA» και αναδύεται το μικρό παράθυρο της Εικόνας 38.

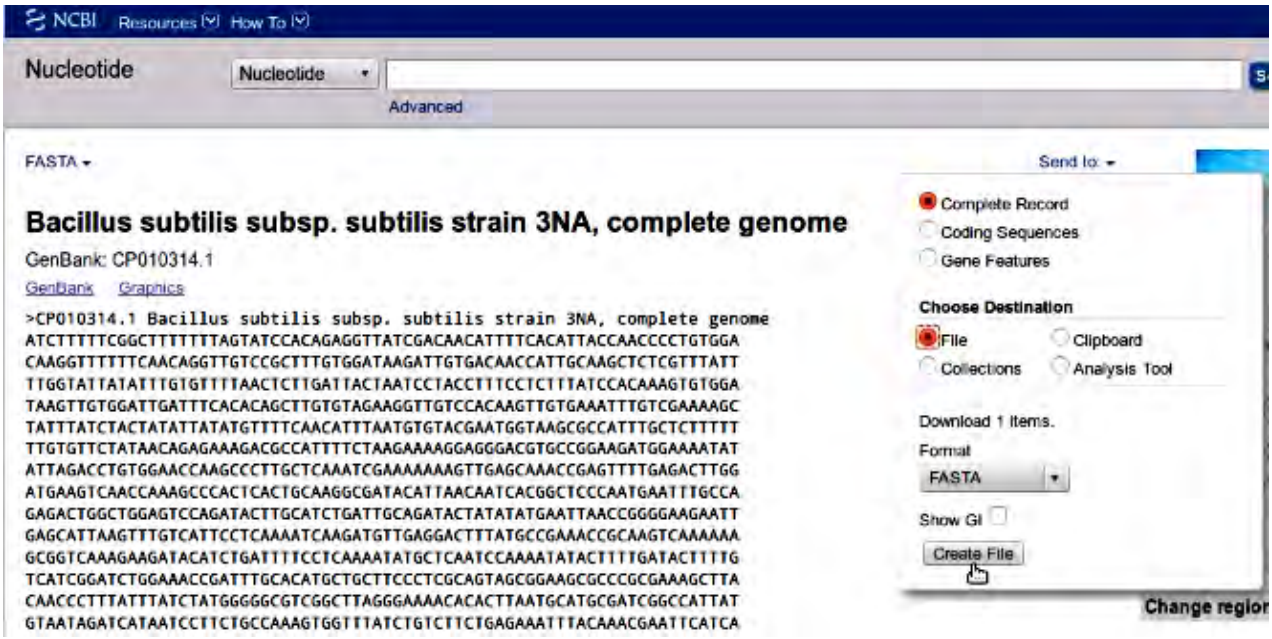
Εικόνα 37 Επιλέγοντας τον κωδικό Sequence ID του πιο συγγενικού μικροοργανισμού, ανοίγει το παράθυρο αναφοράς του στη GenBank.

Στο νέο παράθυρο επιλέγουμε «FASTA» και ως «Format» πάλι «FASTA».



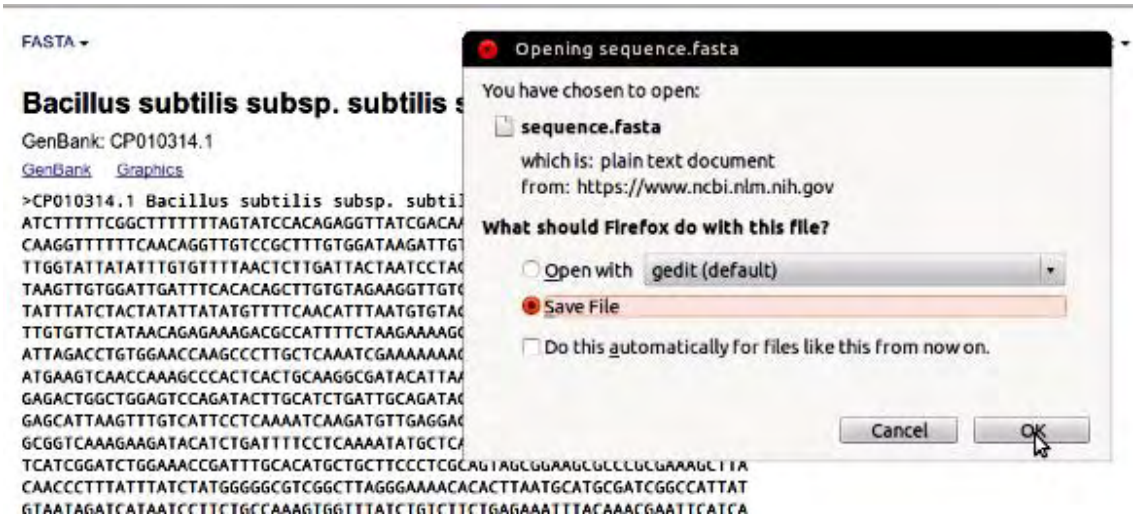
Εικόνα 38 Επάνω και αριστερά στο παράθυρο της Εικόνας 37, κάτω από την ονομασία του μικροοργανισμού και τον κωδικό GenBank, επιλέγουμε FASTA και αναδύεται το εικονιζόμενο παράθυρο. Σε αυτό το παράθυρο επιλέγουμε FASTA και ως Format πάλι FASTA.

Κατόπιν, και όπως φαίνεται στην Εικόνα 39, επιλέγουμε «Send to» και από εκεί «Complete Record», ενώ ως «Choose Destination» επιλέγουμε «File». Ως «Format» ορίζουμε την επιλογή «FASTA» και εκτελούμε «Create File».



Εικόνα 39 Επιλέγουμε Send to και από εκεί Complete Record, ενώ ως Choose Destination επιλέγουμε File. Ως Format ορίζουμε την επιλογή FASTA και εκτελούμε Create File.

Με την εκτέλεση της παραπάνω ενέργειας θα δημιουργηθεί ένα αρχείο τύπου «sequence.fasta» που θα περιέχει τη νουκλεοτιδική αλληλουχία του γονιδιώματος του πιο συγγενικού μικροοργανισμού, το οποίο θα αποθηκεύσουμε στον υπολογιστή μας με τον τρόπο που φαίνεται στο παράθυρο της Εικόνας 40.



Εικόνα 40 Εκτελώντας Create File στο παράθυρο της Εικόνας 39, θα δημιουργηθεί ένα αρχείο τύπου «sequence.fasta» που θα περιέχει τη νουκλεοτιδική αλληλουχία του γονιδιώματος του πιο συγγενικού μικροοργανισμού, το οποίο θα αποθηκεύσουμε στον υπολογιστή μας.

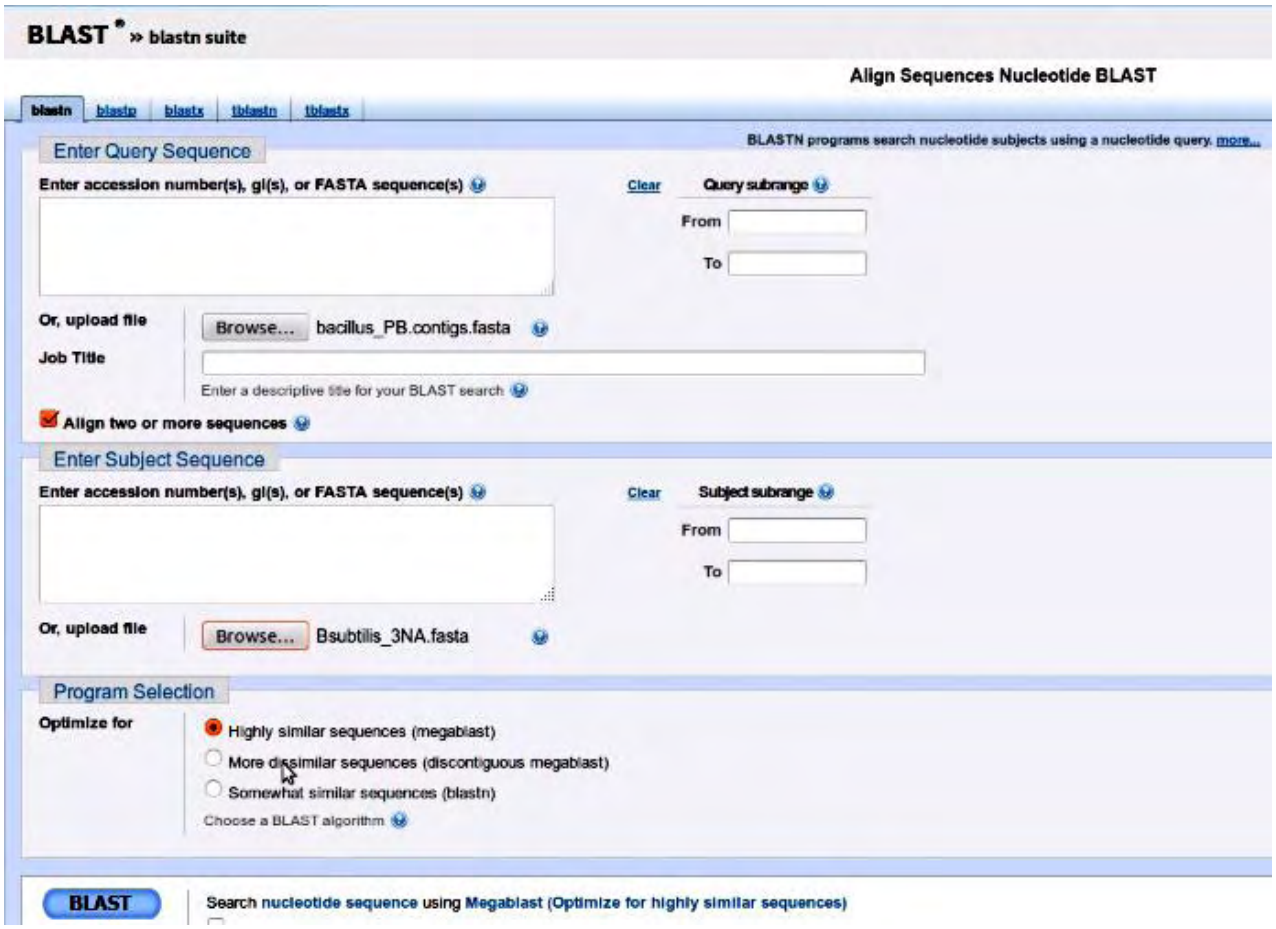
3.9 Σύγκριση του συναρμολογημένου γονιδιώματος με άλλο γονιδίωμα αναφοράς με το πρόγραμμα Blast2seq και με στιγμοπίνακα

Το πρόγραμμα Blast2seq (BLAST 2 sequences), βασίζεται στο BLAST για τη στοίχιση δύο νουκλεοτιδικών ή πρωτεϊνικών αλληλουχιών. Ενώ το πρότυπο πρόγραμμα BLAST χρησιμοποιείται ευρέως για την αναζήτηση ομόλογων αλληλουχιών σε βάσεις δεδομένων νουκλεοτιδίων και πρωτεϊνών, χρειάζεται, συχνά, να συγκρίνονται μόνο δύο αλληλουχίες οι οποίες είναι ήδη γνωστές ότι είναι ομόλογες. Σε τέτοιες περιπτώσεις η αναζήτηση σε ολόκληρη τη βάση δεδομένων θα ήταν άσκοπη και χρονοβόρα. Το Blast2seq χρησιμοποιεί τον αλγόριθμο BLAST για σύγκριση αλληλουχιών DNA-DNA ή πρωτεΐνης-πρωτεΐνης (BLAST 2 SEQUENCES, a new tool for comparing protein and nucleotide sequences).

Το Blast2seq λειτουργεί, επίσης, ως on line εφαρμογή και διατίθεται ελεύθερα από το NCBI στον σύνδεσμο:

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&LINK_LOC=align2seq

Στη μηχανή αναζήτησης πληκτρολογούμε «blast2seq» και επιλέγοντας την τοποθεσία «Nucleotide BLAST: Align two or more sequences using BLAST - NIH» αναδύεται ένα παράθυρο με ονομασία «BLAST® » blastn suite», όπως αυτό της Εικόνας 41.

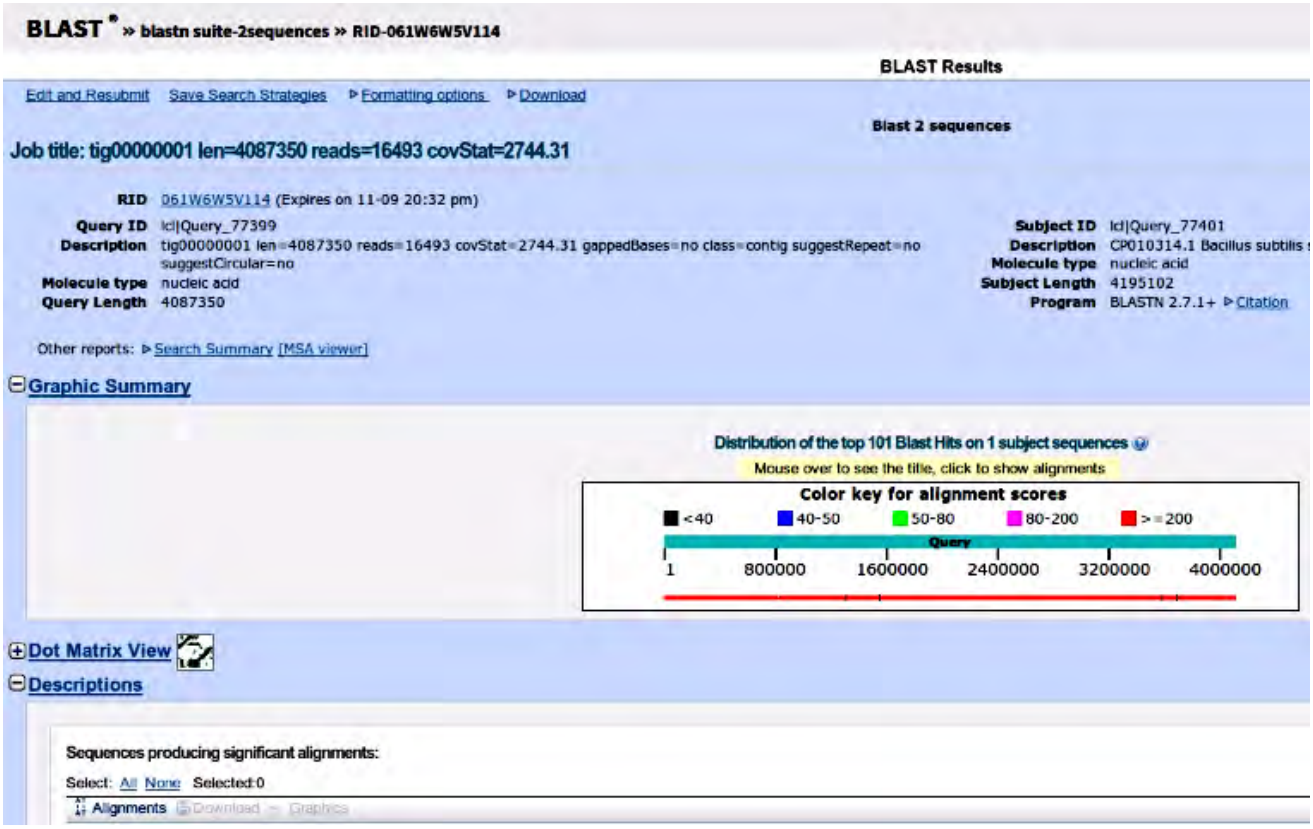


Εικόνα 41 Το παράθυρο BLAST® » blastn suite. Στη θέση Enter Query Sequence μεταφορτώνουμε, από το Browse, το αρχείο «.contigs.fasta» του υπό μελέτη μικροοργανισμού, ενώ στη θέση Enter Subject Sequence το αρχείο «sequence.fasta», με τη νουκλεοτιδική αλληλουχία του γονιδιώματος του πιο συγγενικού μικροοργανισμού. Στην ενότητα Program Selection - Optimize for επιλέγουμε Highly similar sequences (megablast) και εκτελούμε BLAST.

Στη θέση «Enter Query Sequence» θα μεταφορτώσουμε, από το «Browse», το αρχείο «.contigs.fasta» του υπό μελέτη μικροοργανισμού, ενώ στη θέση «Enter Subject Sequence» το αρχείο «sequence.fasta», το οποίο είχαμε αποθηκεύσει στον υπολογιστή μας, με τη νουκλεοτιδική αλληλουχία του γονιδιώματος του πιο συγγενικού μικροοργανισμού. Στην ενότητα

«Program Selection - Optimize for» επιλέγουμε «Highly similar sequences (megablast)» και τέλος εκτελούμε «BLAST».

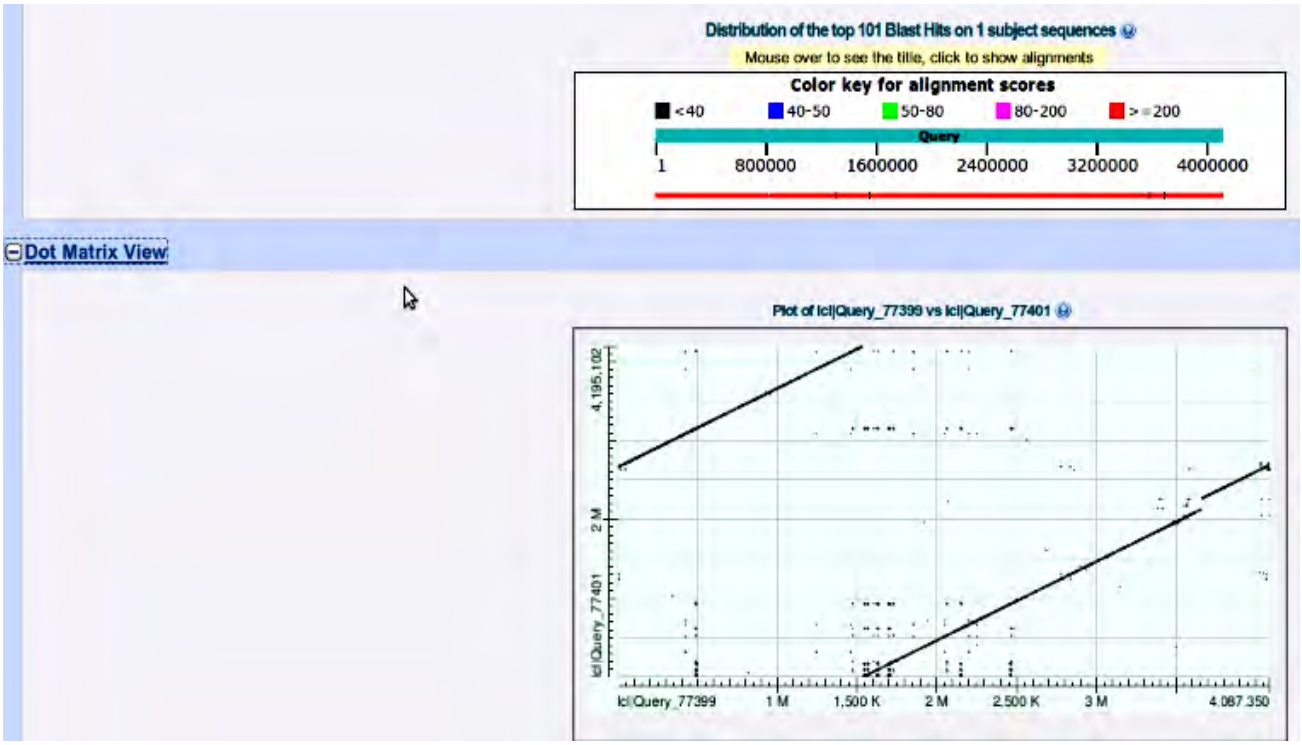
Με την εκτέλεση του «BLAST» αναδύεται το παράθυρο «BLAST Results» που απεικονίζεται στην Εικόνα 42. Στο γράφημα της ενότητας «Graphic Summary» απεικονίζεται με γαλάζιο χρώμα η νουκλεοτιδική αλληλουχία επερώτησης (Query) του υπό μελέτη μικροοργανισμού, ενώ από κάτω της βρίσκεται μια μοναδική κόκκινη γραμμή, η οποία αντιστοιχεί στην αλληλουχία του πιο στενά συγγενικού μικροοργανισμού που ετέθη ως αλληλουχία «Subject».



Εικόνα 42 Με την εκτέλεση του BLAST αναδύεται το παράθυρο BLAST Results. Στο γράφημα της ενότητας Graphic Summary απεικονίζεται με γαλάζιο χρώμα η νουκλεοτιδική αλληλουχία επερώτησης (Query) του υπό μελέτη μικροοργανισμού, ενώ από κάτω της βρίσκεται μια μοναδική κόκκινη γραμμή, η οποία αντιστοιχεί στην αλληλουχία του πιο στενά συγγενικού μικροοργανισμού που ετέθη ως αλληλουχία Subject.

Στο παράθυρο «BLAST Results», επιλέγοντας την ενότητα «Dot Matrix View», αναδύεται ένας στιγμοπίνακας (Dot Plot), όπως αυτός της Εικόνας 43. Σε γενικές γραμμές, η αποτύπωση σε στιγμοπίνακα (Dot Matrix View) της

νουκλεοτιδικής αλληλουχίας των γονιδιωμάτων δύο μικροοργανισμών του ιδίου είδους θα εμφανίζεται ως μία ενιαία γραμμή.



Εικόνα 43 Στο παράθυρο BLAST Results, επιλέγοντας την ενότητα Dot Matrix View, αναδύεται ένας στιγμοπίνακας (Dot Plot).

3.10 Βάσεις δεδομένων γονιδίων παθογονικότητας

Ο εντοπισμός των γονιδίων παθογονικότητας παρουσιάζει εξαιρετικά μεγάλο ενδιαφέρον, από κλινικής, τοξικολογικής, αλλά και εγκληματολογικής άποψης. Η κλινική σημασία των γονιδίων αυτών σε σχέση με τις λοιμώδεις νόσους με τις οποίες σχετίζονται, η τοξικολογική δράση των προϊόντων που κωδικοποιούν και οι επιπτώσεις που μπορεί να επέλθουν για τη δημόσια υγεία, αλλά και για την οικονομία, από την έξαρση επιδημιών ή επιθέσεων με εγκληματολογικά κίνητρα, καθιστούν πολύ σημαντική την ανάλυση των μικροβιακών γονιδιωμάτων για τον εντοπισμό και την ταυτοποίηση των γονιδίων αυτών και των προϊόντων τους.

Εφόσον προσδιοριστεί η πλήρης νουκλεοτιδική αλληλουχία ενός μικροβιακού γονιδιώματος, τότε για τον αντίστοιχο μικροοργανισμό δύναται να διερευνηθεί η παρουσία γονιδίων παθογονικότητας στο γονιδίωμά του. Αυτή η εξέταση διενεργείται σε διάφορες βάσεις δεδομένων, οι οποίες είναι διαθέσιμες διαδικτυακά και είναι οι εξής:

- VFDB: Virulence Factors Database

http://www.mgc.ac.cn/VFs/

Η βάση δεδομένων VFDB είναι μια ηλεκτρονική πηγή για τη διαχείριση των πληροφοριών σχετικά με τους παράγοντες παθογένειας των παθογόνων βακτηρίων. Από την ίδρυσή της, το 2004, η VFDB έχει αφιερωθεί στην παροχή επικαιροποιημένης γνώσης σχετικά με τους λοιμογόνους παράγοντες διαφόρων σημαντικών βακτηριακών μικροοργανισμών. Το κίνητρο για την κατασκευή της VFDB ήταν διπλό. Πρώτον, να παράσχει λεπτομερή κάλυψη των κυριοτέρων λοιμογόνων παραγόντων των βακτηριακών παθογόνων, με τα χαρακτηριστικά της δομής τους και τους μηχανισμούς που χρησιμοποιούνται από αυτά τα παθογόνα για να παρακάμψουν τους αμυντικούς μηχανισμούς του ξενιστή και να προκαλέσουν ασθένειες. Δεύτερον, να παράσχουν στους ερευνητές επικαιροποιημένες γνώσεις σχετικά με την ευρεία ποικιλία των μηχανισμών που χρησιμοποιούνται από τους βακτηριακούς παθογόνους παράγοντες, με σκοπό να διασαφηνιστούν οι μηχανισμοί που προκαλούν τις βακτηριακές ασθένειες, ώστε να αναπτυχθούν νέες ορθολογικές προσεγγίσεις για τη θεραπεία και την πρόληψη των λοιμωδών νοσημάτων (Chen et al., 2016 / Chen et al., 2012 / Yang et al., 2008 / Chen et al., 2005).

- ResFinder - DTU

https://cge.cbs.dtu.dk/services/ResFinder/

Με τη χρήση της βάσης δεδομένων ResFinder δίνεται η δυνατότητα προσδιορισμού των γονιδίων αντιμικροβιακής ανθεκτικότητας και, επίσης, δύναται να εντοπιστούν χρωμοσωμικές μεταλλάξεις σε ολικά ή σε μερικά δεδομένα αλληλούχισης βακτηριακών νουκλεοτιδικών αλληλουχιών (Zankari et. Al., 2012).

- BTXpred

http://crdd.osdd.net/raghava/btxpred/

Στόχος της BTXpred είναι να ταυτοποιηθεί εάν οι αλληλουχίες των αμινοξέων που εισάγονται σε αυτή, ως δεδομένα, είναι βακτηριακές τοξίνες και συγκεκριμένα ενδοτοξίνες ή εξωτοξίνες, ενώ μελετάται και η λειτουργία τους. Ο διακομιστής της βάσης λειτουργεί χρησιμοποιώντας Support Vector Machines, (SVMs), κρυφά Μαρκοβιανά μοντέλα (Hidden Markov Model, HMM) και το PSI-Blast.

Ο διακομιστής της βάσης επιτρέπει στους χρήστες να προβλέπουν τις βακτηριακές τοξίνες με ακρίβεια 96,07%, να ταξινομούν τις βακτηριακές τοξίνες σε εξωτοξίνες και ενδοτοξίνες με ακρίβεια 95,71%, να ταξινομήσουν, με συνολική ακρίβεια 100%, τις εξωτοξίνες σε επτά διαφορετικές λειτουργίες ανάλογα με τους μοριακούς στόχους τους, οι οποίοι είναι οι εξής: i) να ενεργοποιούν την αδενυλική κυκλάση, ii) να ενεργοποιήσουν την γουανυλική κυκλάση, iii) να προκαλούν τροφική δηλητηρίαση, να είναι: iv) νευροτοξίνες, v) κυτταροτοξίνες μακροφάγων, vi) κενοτοπιοτοξίνες και vii) κυτταροτοξίνες που ενεργοποιούνται από θειόλη (Saha και Raghava, 2007).

- CARD (The Comprehensive Antibiotic Resistance Database)

https://card.mcmaster.ca/

Η CARD είναι μια βιοπληροφορική βάση δεδομένων γονιδίων ανθεκτικότητας, των προϊόντων τους και των συναφών τους φαινοτύπων. Συγκεκριμένα, παρέχει δεδομένα, μοντέλα και αλγόριθμους που σχετίζονται με τη μοριακή βάση της αντιμικροβιακής ανθεκτικότητας. Επίσης, παρέχει νουκλεοτιδικές αλληλουχίες αναφοράς και μονονουκλεοτιδικούς πολυμορφισμούς (Single Nucleotide Polymorphisms, SNPs) (Jia et al., 2017 / McArthur και Wright, 2015 / McArthur et al., 2013).

- T3DB

http://www.t3db.ca/

Η βάση δεδομένων Toxin and Toxin Target Database (T3DB), η οποία πρόκειται να μετονομαστεί σε «Toxic Exposome Database», συνδυάζει λεπτομερή δεδομένα για τις τοξίνες με ολοκληρωμένες πληροφορίες του στόχου της τοξίνης. Η βάση δεδομένων περιλαμβάνει 3.673 τοξίνες που περιγράφονται από 41.733 συνώνυμα, συμπεριλαμβανομένων ρύπων, παρασιτοκτόνων, φαρμάκων και τοξινών των τροφίμων, τα οποία συνδέονται με 2.087 αντίστοιχα αρχεία των στόχων των τοξινών. Συνολικά, υπάρχουν

42.471 τοξίνες και αντίστοιχες σχέσεις τοξινών και στόχων. Κάθε καταγραφή τοξινών (ToxCard) περιέχει πάνω από 90 πεδία δεδομένων και περιλαμβάνει πληροφορίες, όπως χημικές ιδιότητες, τιμές τοξικότητας, μοριακές και κυτταρικές αλληλεπιδράσεις, καθώς και ιατρικές πληροφορίες. Αυτές οι πληροφορίες έχουν εξαχθεί από περισσότερες από 18.143 πηγές, οι οποίες περιλαμβάνουν άλλες βάσεις δεδομένων, κυβερνητικά έγγραφα, βιβλία και επιστημονική βιβλιογραφία.

Το επίκεντρο της βάσης T3DB είναι η παροχή πληροφοριών για τους μηχανισμούς της τοξικότητας και των πρωτεϊνών στόχων για κάθε τοξίνη. Αυτή η διπλή φύση της T3DB, στην οποία τα αρχεία των τοξινών και των στόχων τους συνδέονται διαδραστικά και προς τις δύο κατευθύνσεις, την καθιστά μοναδική από τις υπάρχουσες βάσεις δεδομένων. Οι εφαρμογές της T3DB περιλαμβάνουν πρόβλεψη του μεταβολισμού των τοξινών, πρόβλεψη των αλληλεπιδράσεων τοξίνης-φαρμάκου και γενική ευαισθητοποίηση του κοινού σχετικά με την κάθε εξεταζόμενη τοξίνη, καθιστώντας την εφαρμόσιμη σε διάφορους τομείς. Συνολικά, η ποικιλία και η προσβασιμότητα της T3DB την καθιστούν πολύτιμο πόρο τόσο για τον περιστασιακό χρήστη όσο και για τον προηγμένο ερευνητή (Wishart et al., 2015 / Lim et al., 2010).

- DBETH (Database of Bacterial ExoToxins for Human)

http://www.hpppi.iicb.res.in/btox/

Με τη βάση δεδομένων DBETH διεξάγεται ανάλυση αλληλουχιών των αμινοξέων που εισάγονται σε αυτή ως δεδομένα. Πρόκειται για μια βάση δεδομένων με αλληλουχίες, δομές, δίκτυα αλληλεπίδρασης και αναλυτικά αποτελέσματα για 229 εξωτοξίνες, από 26 διαφορετικά παθογόνα στελέχη βακτηρίων για τον άνθρωπο. Όλες οι τοξίνες ταξινομούνται σε 24 διαφορετικές κατηγορίες. Στόχος της DBETH είναι να παρέχει μια ολοκληρωμένη βάση δεδομένων για παθογόνες βακτηριακές εξωτοξίνες του ανθρώπου.

Η DBETH παρέχει, επίσης, μια ειδική πλατφόρμα στους χρήστες της, για να αναγνωρίζουν πιθανές αλληλουχίες που ομοιάζουν με εξωτοξίνη, μέσω μεθόδων που βασίζονται στην ομολογία (Homology based), καθώς και σε μεθόδους που δε βασίζονται στην ομολογία (Non-homology based). Στην προσέγγιση που βασίζεται στην ομολογία, οι χρήστες μπορούν να

αναγνωρίσουν πιθανές αλληλουχίες που ομοιάζουν με εξωτοξίνη, είτε εκτελώντας BLASTp για τις αλληλουχίες της τοξίνης, ή εκτελώντας ανάλυση HMMER χρησιμοποιώντας κρυφά Μαρκοβιανά μοντέλα για τις πρωτεϊνικές επικράτειες (domains) της τοξίνης που ταυτοποιούνται, μέσω της DBETH, ως παθογόνες βακτηριακές εξωτοξίνες που προσβάλλουν τον άνθρωπο. Σε ό,τι αφορά στον μη βασισμένο στην ομολογία τομέα της DBETH, χρησιμοποιείται μια μέθοδος πρόβλεψης της τοξίνης που βασίζεται σε SVMs για τον εντοπισμό δυνητικών εξωτοξινών (Chakraborty et al., 2012).

- VICMpred (Prediction of Virulence factors, Information molecule, Cellular process and Metabolism molecule in the Bacterial proteins)

http://crdd.osdd.net/raghava/vicmpred/

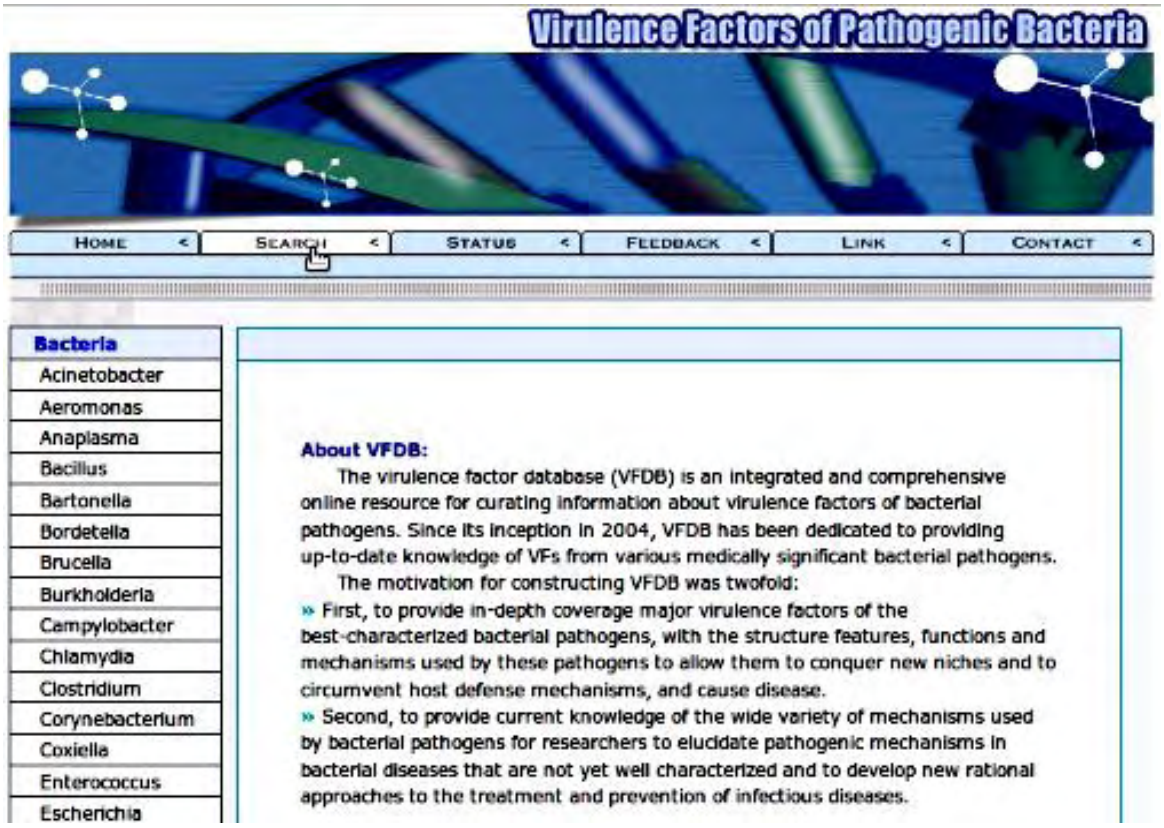
Στη βάση VICMpred διεξάγεται ευρεία λειτουργική ταξινόμηση των βακτηριακών πρωτεϊνών σε λοιμογόνους παράγοντες, μόρια πληροφοριών, μόρια κυτταρικών διαδικασιών και σε μόρια μεταβολισμού. Ο διακομιστής VICMpred, βάσει της αλληλουχίας των αμινοξέων, χρησιμοποιεί μέθοδο που βασίζεται σε SVMs για την πρόβλεψη τοξινών και άλλων λειτουργικών πρωτεϊνών. Η συνολική ακρίβεια αυτού του διακομιστή ανέρχεται στο 70,75% (Saha και Raghava, 2006).

3.11 Εντοπισμός γονιδίων παθογονικότητας με τη βάση δεδομένων VFDB

Στην παράγραφο 3.6 αναφέρθηκε ότι το γονιδίωμα ενός, υπό μελέτη, μικροοργανισμού συναρμολογήθηκε από τα δεδομένα των contigs, τα οποία βρίσκονται αποθηκευμένα εντός του αρχείου «.contigs.fasta». Ένα αρχείο «.contigs.fasta» μπορεί να εισαχθεί για ανάλυση στη βάση δεδομένων VFDB για τον εντοπισμό γονιδίων παθογονικότητας.

Σε διαδικτυακή μηχανή αναζήτησης πληκτρολογούμε VFDB και επιλέγουμε «VFDB: Virulence Factors Database». Ανοίγει το παράθυρο της Εικόνας 44 από το οποίο, επιλέγοντας με τον κέρσορα σε οποιοδήποτε σημείο επάνω στο εικονιζόμενο λογότυπο, αναδύεται το παράθυρο της Εικόνας 45 που μας εισάγει στον περιβάλλοντα χώρο της βάσης.

Εικόνα 44 Λογότυπος της βάσης δεδομένων VFDB**.**



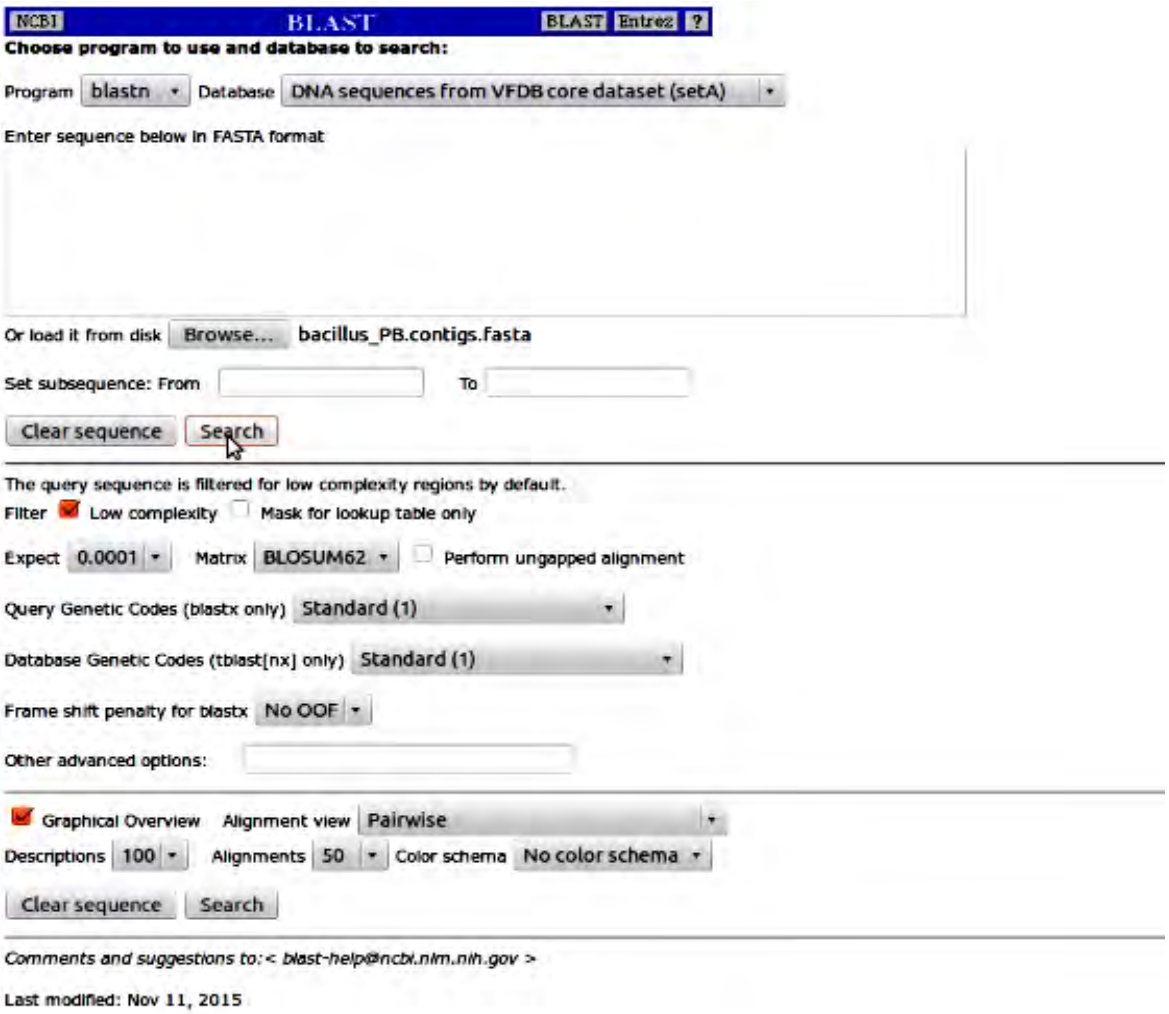
Εικόνα 45 Περιβάλλοντας χώρος της βάσης δεδομένων VFDB.

Επιλέγοντας «SEARCH» μεταφερόμαστε στο παράθυρο «Query page» της Εικόνας 46.

Στην ενότητα «*Blast Search*» επιλέγουμε «Regular BLAST».

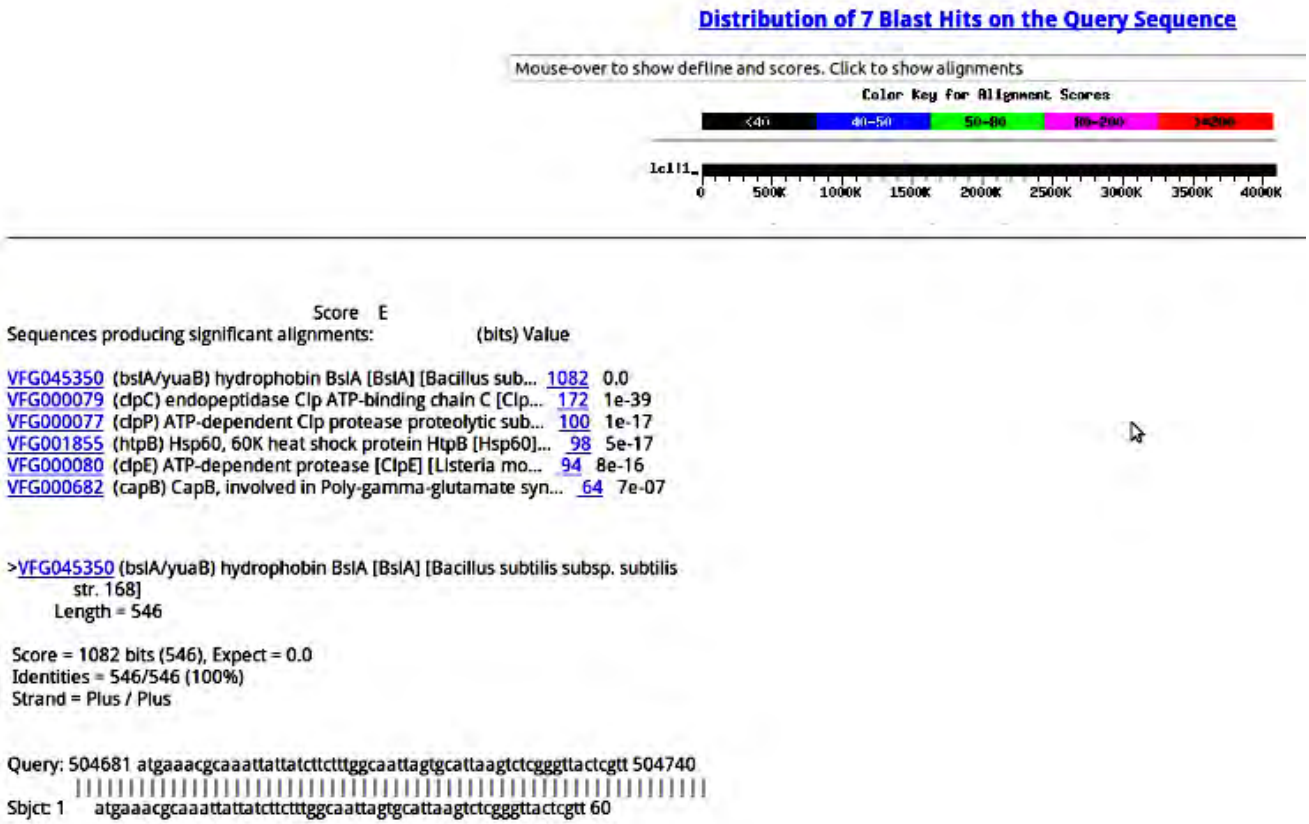
Εικόνα 46 Επιλέγοντας SEARCH στο παράθυρο του περιβάλλοντα χώρου της βάσης, μεταφερόμαστε στο παράθυρο Query page. Στην ενότητα *Blast Search* επιλέγουμε Regular BLAST.

Ανοίγει το παράθυρο της Εικόνας 47 και από την επιλογή «Browse» ανεβάζουμε το επιθυμητό αρχείο, τύπου «.contigs.fasta», το οποίο αφορά στον υπό μελέτη μικροοργανισμό. Στην υποενότητα «Expect» θέτουμε την τιμή «0,0001» και, εν συνεχεία, εκτελούμε «Search» που βρίσκεται στο επάνω ήμισυ του παραθύρου.



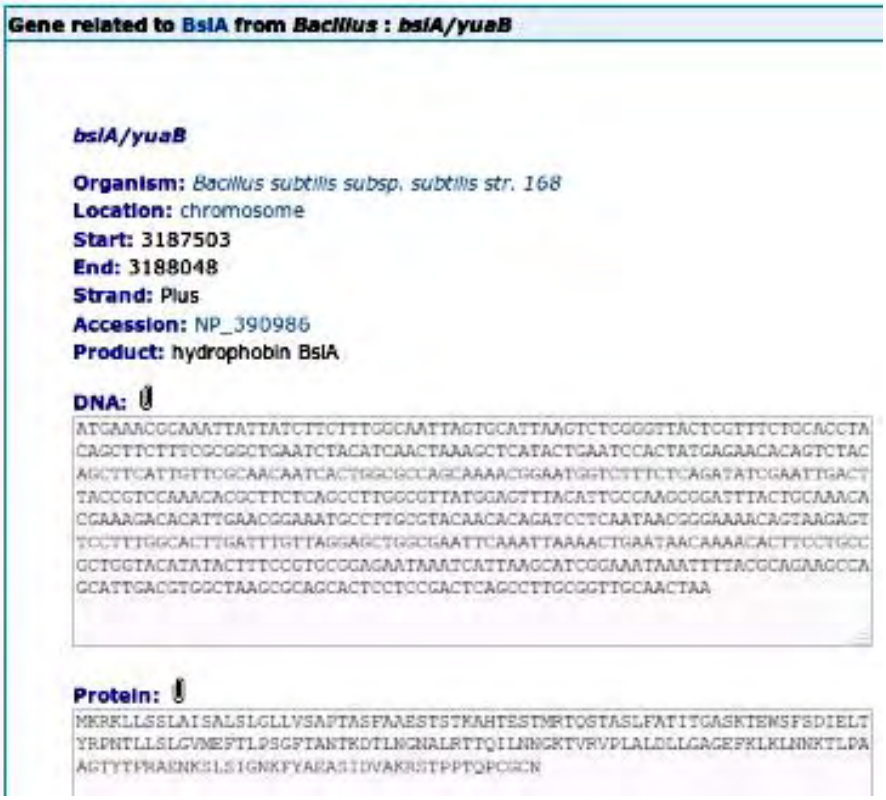
Εικόνα 47 Από την επιλογή Browse ανεβάζουμε το επιθυμητό αρχείο, τύπου «.contigs.fasta», το οποίο αφορά στον υπό μελέτη μικροοργανισμό. Στην υποενότητα Expect θέτουμε την τιμή «0,0001» και, εν συνεχεία, εκτελούμε Search που βρίσκεται στο επάνω ήμισυ του παραθύρου.

Αναδύεται ένα παράθυρο με ονομασία «BLAST Search Results», το κάτω μέρος του οποίου απεικονίζεται στην Εικόνα 48. Στο παράθυρο αυτό δίδονται οι πληροφορίες που προέκυψαν για το αρχείο «.contigs.fasta». Συγκεκριμένα, ο χρήστης πληροφορείται σχετικά με τον συνολικό αριθμό τον ευρεθέντων γονιδίων, της ονομασίας τους, της νουκλεοτιδικής τους αλληλουχίας, του συνόλου των παραγόντων τοξικολογικού ενδιαφέροντος που κωδικοποιούνται από αυτά τα γονίδια, καθώς και οι αντίστοιχοι μικροοργανισμοί στο γονιδίωμα των οποίων έχουν ανευρεθεί να κωδικοποιούνται οι εν λόγω μολυσματικοί παράγοντες.



Εικόνα 48 Το κάτω τμήμα του παραθύρου BLAST Search Results. Στο παράθυρο αυτό δίδονται οι πληροφορίες που προέκυψαν για το αρχείο «.contigs.fasta». Επιλέγοντας κάποιον από τους κωδικούς καταχώρησης των γονιδίων που απεικονίζονται με μπλε χρώμα, αριστερά, δίνεται η δυνατότητα να αναζητηθούν περισσότερες πληροφορίες.

Στο παράδειγμα της Εικόνας 48, επιλέγοντας κάποιον από τους κωδικούς καταχώρησης των γονιδίων που απεικονίζονται με μπλε χρώμα, δίνεται η δυνατότητα να αναζητηθούν περισσότερες πληροφορίες σχετικά με τα εν λόγω γονίδια, τους οργανισμούς στους οποίους έχουν ταυτοποιηθεί, τη νουκλεοτιδική και την πρωτεϊνική τους αλληλουχία και άλλα. Για παράδειγμα, επιλέγοντας των κωδικό «VFG045350» ανοίγει το παράθυρο της Εικόνας 49, το οποίο αναφέρεται στο γονίδιο *bslA/yuaB* και στον μολυσματικό παράγοντα BslA που κωδικοποιείται από το εν λόγω γονίδιο. Επιπλέον, αναφέρεται το γένος, το είδος και το στέλεχος του μικροοργανισμού στον οποίο έχει ταυτοποιηθεί.



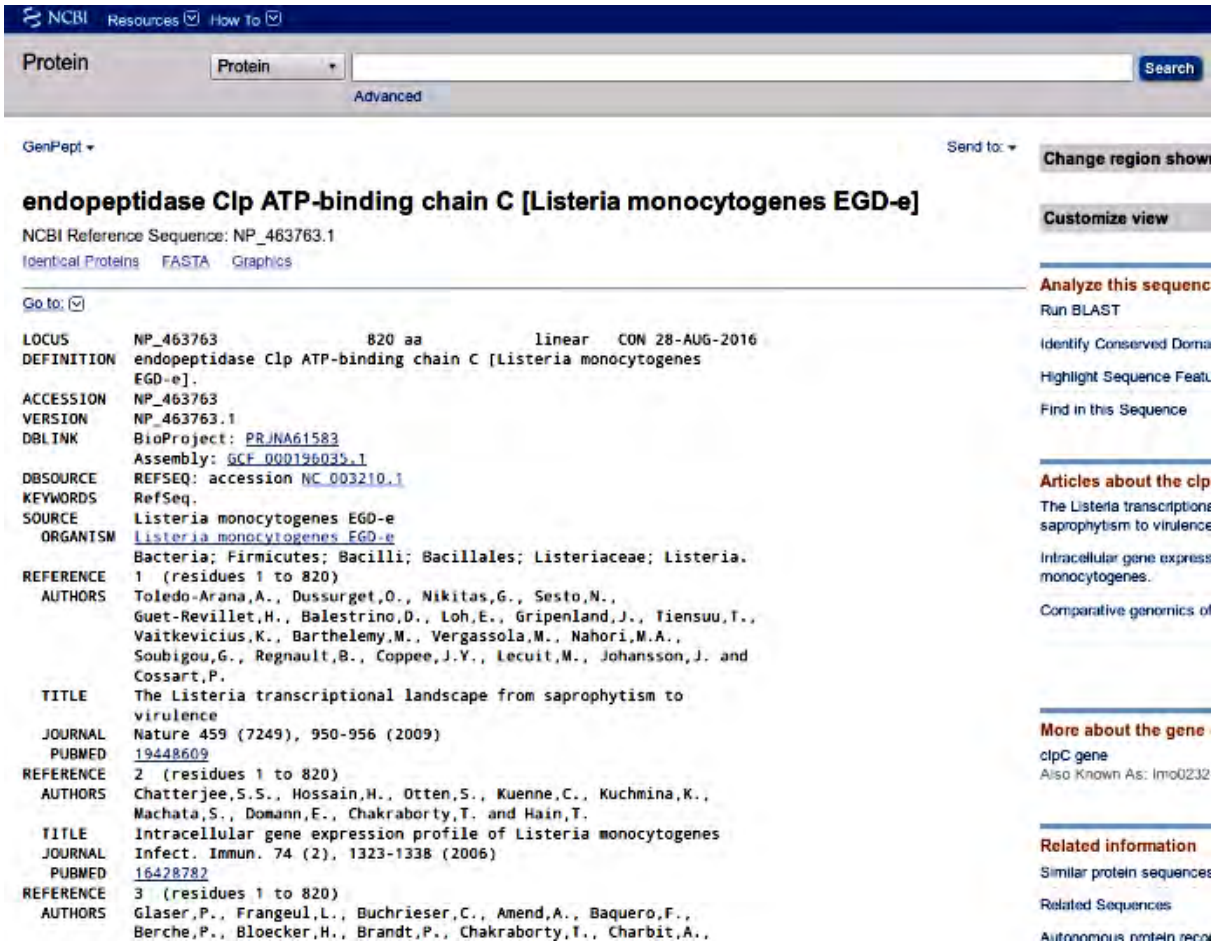
Εικόνα 49 Επιλέγοντας έναν κωδικό καταχώρησης από το παράθυρο της Εικόνας 48, αναδύεται το εικονιζόμενο παράθυρο, το οποίο αναφέρεται σε ένα συγκεκριμένο γονίδιο παθογονικότητας και στον μολυσματικό παράγοντα που κωδικοποιεί. Επιπλέον, αναφέρεται το γένος, το είδος και το στέλεχος του μικροοργανισμού στον οποίο έχει ταυτοποιηθεί.

Με τον ίδιο τρόπο, επιλέγοντας «VFG000079» αναδύεται το παράθυρο της Εικόνας 50, το οποίο αναφέρεται στο γονίδιο *clpC* και στον αντίστοιχο (ClpC) μολυσματικό παράγοντα που κωδικοποιείται από αυτό το γονίδιο.

Εικόνα 50 Επιλέγοντας τον κωδικό VFG000079 της Εικόνας 48, αναδύεται το εικονιζόμενο παράθυρο, το οποίο αναφέρεται στο γονίδιο *clpC* και στον αντίστοιχο (ClpC) μολυσματικό παράγοντα που κωδικοποιεί.

Στις Εικόνες 49 και 50 ο κωδικός πρόσβασης της ενότητας «Accession:» παραπέμπει στη σχετική τοποθεσία του NCBI, όπου βρίσκονται καταχωρημένα τα πλήρη στοιχεία για έναν μολυσματικό παράγοντα.

Για παράδειγμα, στην Εικόνα 50 επιλέγοντας τον κωδικό πρόσβασης «NP_463763» ανοίγει το σχετικό παράθυρο του NCBI, όπως απεικονίζεται στην Εικόνα 51, το οποίο αναφέρεται στα πλήρη στοιχεία του μολυσματικού παράγοντα ClpC που παράγεται από το βακτηριακό στέλεχος *Listeria monocytogenes EGD-e*.



Εικόνα 51 Στις Εικόνες 49 και 50 ο κωδικός πρόσβασης της ενότητας «Accession:» παραπέμπει στη σχετική τοποθεσία του NCBI, όπου βρίσκονται καταχωρημένα τα πλήρη στοιχεία για έναν μολυσματικό παράγοντα. Για παράδειγμα, στην Εικόνα 50 επιλέγοντας τον κωδικό πρόσβασης «NP_463763» ανοίγει το εικονιζόμενο παράθυρο του NCBI.

Μέσω, λοιπόν, του εντοπισμού των γονιδίων παθογονικότητας μπορούμε να βρούμε κάθε σχετική πληροφορία για καθέναν από τους γνωστούς μολυσματικούς παράγοντες που κωδικοποιούν. Για παράδειγμα, αναζητώντας για σχετικές πληροφορίες για τον μολυσματικό παράγοντα ClpC που παράγεται από το βακτήριο *Listeria monocytogenes*, μπορούμε να είμαστε σε θέση να γνωρίζουμε ότι ενισχύει την ενδοκυτταρική επιβίωση του βακτηρίου, καθώς προάγει την πρώιμη διαφυγή του από το φαγοσωμικό διαμέρισμα των μακροφάγων και παίζει σημαντικό ρόλο στη λοιμογόνο δράση του (Rouquette et al., 1998).

ΣΥΖΗΤΗΣΗ

4.1 Πλεονεκτήματα και περιορισμοί των τεχνολογιών αλληλούχισης νέας γενιάς

Τα τέσσερα κύρια πλεονεκτήματα της αλληλούχισης νέας γενιάς σε σχέση με την κλασική αλληλούχιση Sanger είναι τα εξής:

- Ταχύτητα
- Κόστος
- Μέγεθος δείγματος
- Ακρίβεια

Η αλληλούχιση νέας γενιάς είναι σημαντικά φθηνότερη, γρηγορότερη, χρειάζεται πολύ λιγότερο DNA και είναι πιο ακριβής και αξιόπιστη από την αλληλούχιση Sanger.

Ο χρόνος, το εργατικό δυναμικό και τα αντιδραστήρια για τη διεξαγωγή της αλληλούχισης νέας γενιάς είναι πολύ λιγότερα, γεγονός που συνεπάγεται σε πολύ χαμηλότερο οικονομικό κόστος. Η πρώτη αλληλούχιση του ανθρώπινου γονιδιώματος κόστισε περίπου 337.600.000 Ευρώ. Υπολογίζεται ότι χρησιμοποιώντας σύγχρονες μεθόδους αλληλούχισης Sanger και με τη βοήθεια των δεδομένων από αλληλουχίες αναφοράς, ένα πλήρες ανθρώπινο γονιδίωμα θα κόστιζε σήμερα, περίπου, 6.800.000 Ευρώ. Ωστόσο, η αλληλούχιση ενός ανθρώπινου γονιδιώματος με τους τελευταίας τεχνολογίας αναλυτές NovaSeq 5000 και NovaSeq 6000 της Illumina κοστίζει μόλις 820 Ευρώ. Περαιτέρω, ο Francis deSouza, διευθύνων σύμβουλος της Illumina, σε συνέντευξή του που δημοσιεύθηκε στις 09/01/2017 στο Forbes (Forbes, Illumina Promises To Sequence Human Genome For $100), αισιοδοξεί ότι σε λιγότερο από μία δεκαετία το κόστος αλληλούχισης του ανθρώπινου γονιδιώματος θα ανέρχεται σε μόλις 82 Ευρώ! Αυτό το ελπιδοφόρο γεγονός ενισχύεται και από τις ενθαρρυντικές επενδύσεις που προτείνονται από τους ειδικούς του κλάδου στις τεχνολογίες της Illumina, καθώς αντιλαμβάνονται το

ευοίωνο μέλλον της εταιρίας και τα αναλογούντα της κέρδη (Crumly Jim (TMFSpeyside) Better Buy: Illumina, Inc. vs. Pacific Biosciences).

Οι υποστηρικτές των μεθοδολογιών αλληλούχισης νέας γενιάς υποστηρίζουν ότι είναι εφικτή η πλήρης αλληλούχιση ενός γονιδιώματος με μεγάλη ακρίβεια. Οι επικριτές υποστηρίζουν ότι παρόλο που με την περιγραφόμενη τεχνική προσδιορίζεται η αλληλουχία μεγάλων περιοχών του DNA, η ικανότητά της να συνδέει σωστά αυτές τις περιοχές θεωρείται ύποπτη, ιδιαίτερα για τα γονιδιώματα με πολλές επαναλαμβανόμενες περιοχές. Ωστόσο, καθώς τα λογισμικά προγράμματα συναρμολόγησης των νουκλεοτιδικών αλληλουχιών εξελίσσονται συνεχώς και η υπολογιστική ισχύς γίνεται φθηνότερη, είναι δυνατό να ξεπεραστεί αυτός ο περιορισμός.

Παρά το γεγονός ότι οι τεχνολογίες αλληλούχισης νέας γενιάς είναι σε θέση να παρέχουν γονιδιωματικές αλληλουχίες ακόμη και σε πληθυσμιακό επίπεδο, προβληματίζουν, ωστόσο, ορισμένοι περιορισμοί. Συγκεκριμένα, η ακρίβεια και η κάλυψη σε όλο το γονιδίωμα μπορεί να περιορίζονται όταν πρόκειται για πλούσιες σε GC γονιδιωματικές περιοχές και για μακριές σε μήκος ομοπολυμερείς περιοχές (Ross et al., 2013). Προβληματισμός υπήρξε και για τα μικρού μήκους reads, καθώς θεωρήθηκε ότι περιορίζουν σοβαρά την ικανότητα του χαρακτηρισμού με ακρίβεια μεγάλων περιοχών με επαναλήψεις, πολλαπλές προσθήκες και διαγραφές βάσεων (indels), καθώς και με διάφορες δομικές παραλλαγές, αφήνοντας σημαντικά τμήματα του γονιδιώματος αδιαφανή ή ανακριβή (Snyder et al., 2010). Στις μέρες μας, όμως, οι σύγχρονοι αναλυτές της Illumina προσδιορίζουν με πολύ μεγάλη ακρίβεια και κάλυψη την αλληλουχία των εν λόγω γονιδιωματικών περιοχών, περιορίζοντας στο ελάχιστο τους ως άνω περιγραφόμενους προβληματισμούς.

Ένα άλλο, γενικό, ζήτημα που έχει προκύψει σχετικά με την εφαρμογή των μεθοδολογιών αλληλούχισης νέας γενιάς, αφορά στη δημιουργία ενός πρότυπου πρωτοκόλλου για την τυποποιημένη και επικυρωμένη επεξεργασία των δεδομένων αλληλούχισης. Το πρωτόκολλο αυτό θα αναφέρεται στη σύγκριση και στον τρόπο αναφοράς της ακρίβειας των δεδομένων των διαφορετικών τεχνολογιών αλληλούχισης και των διεξαγομένων μελετών. Συγκεκριμένα, το 2011, το «Genome in a Bottle Consortium» (GIAB)

οραματίστηκε το «πρότυπο γονιδιώματος» (gold standard genome) (Zook και Salit, 2011).

Ωστόσο, δεδομένων των περιορισμών και των υπαρκτών μεροληψιών (biases) των διαφόρων τεχνολογιών αλληλούχισης, είναι πιθανό ότι η ακριβής γονιδιωματική αλληλούχιση θα συνεχίσει να χρησιμοποιεί έναν συνδυασμό τεχνολογιών. Στην υπέρβαση αυτών των περιορισμών θα μπορούσαν να βοηθήσουν οι βελτιώσεις που επιχειρούνται σε τεχνολογίες που χρησιμοποιούν αλληλουχίες reads μεγάλου μήκους, όπως οι τεχνολογίες Pacific Biosciences (PacBio) και Oxford Nanopore, καθώς και οι μέθοδοι που χρησιμοποιούν συνθετικά reads μεγάλου μήκους, στις οποίες για τα μεγαλύτερα θραύσματα δύναται να προσδιοριστεί η αλληλουχία τους και να συναρμολογηθούν από μικρά reads (Tilgner et al., 2015). Παρά τους ανακύπτοντες προβληματισμούς, η καινοτομία των νέων τεχνολογιών αλληλούχισης νέας γενιάς βαίνει ολοταχώς σε λαμπρό μέλλον (Schadt et al., 2010).

4.2 Υβριδική συναρμολόγηση με δεδομένα των τεχνολογιών Illumina και Pacific Biosciences και προγράμματα συναρμολόγησης

Επειδή, η τεχνολογία αλληλούχισης PacBio παράγει reads μεγάλου μήκους βρίσκει ευρεία εφαρμογή στη *de novo* συναρμολόγηση (Quail et al., 2012) Αναφορικά με τη μονομοριακή τεχνολογία αλληλούχισης σε πραγματικό χρόνο (Single-molecule real-time, SMRT), που χρησιμοποιούν οι βελτιωμένοι αναλυτές PacBio® RS II της Pacific Biosciences, παρόλο που είναι πιο επιρρεπής σε σφάλματα σε σχέση με άλλου είδους τεχνολογίες, έχει το μεγάλο πλεονέκτημα να παράγει αλληλουχίες reads ιδιαίτερα μεγάλου μήκους, ως και 200 φορές μεγαλύτερες από εκείνες που παράγονται από άλλες τεχνολογίες. Τα μακριά μήκη των reads επιτρέπουν την ταχεία συναρμολόγηση της πλήρους αλληλουχίας ενός γονιδιώματος.

Επειδή, τα μεγάλα reads που προκύπτουν από τις τεχνολογίες αλληλούχισης PacBio και Oxford Nanopore, είναι συχνά επιρρεπή σε

σφάλματα, μπορούν να διορθωθούν, αποτελεσματικά, με μια υβριδική προσέγγιση, ακριβώς όπως αυτή που εξετάστηκε στην παρούσα μελέτη, χρησιμοποιώντας συνδυαστικά τα μικρά, αλλά πολύ ακριβή, reads της Illumina (Goodwin et al., 2015 / Salmela και Rivals, 2014 / Koren et al., 2012).

Η τεχνολογία αλληλούχισης της Illumina αποτελεί τη μεθοδολογία αλληλούχισης νέας γενιάς με τη μεγαλύτερη ακρίβεια σε σχέση με τις υπόλοιπες διαθέσιμες τεχνολογίες. Μάλιστα, για τον λόγο αυτό ο αναλυτής MiSeqDX της Illumina ήταν ο πρώτος που πήρε την έγκριση από τον FDA (US Food and Drug Administration) για να χρησιμοποιηθεί στην κλινική πρακτική (Collins και Hamburg, 2013), ανοίγοντας την πόρτα για την αλληλούχιση νέας γενιάς σε κλινικά εργαστήρια.

Για τη συναρμολόγηση χρησιμοποιούνται αρκετά προγράμματα που εφαρμόζουν τις αλγοριθμικές μεθόδους de Bruijn Grahs (DBG), Overlap Layout Consensus (OLC), και Greedy (Miller et al., 2010). Το πρόγραμμα Canu εφαρμόζει τον αλγόριθμο OLC για τη συναρμολόγηση των reads που προκύπτουν από τις τεχνολογίες PacBio και Oxford Nanopore. Σαν το Canu υπάρχουν τα προγράμματα:

- Flye, το οποίο χρησιμοποιεί τον αλγόριθμο DBG και διατίθεται ελεύθερα από τον σύνδεσμο: **https://github.com/fenderglass/Flye** (Lin et al., 2016).
- Miniasm, το οποίο χρησιμοποιεί τον αλγόριθμο OLC και διατίθεται ελεύθερα από τον σύνδεσμο: **https://github.com/lh3/miniasm**

Αποκλειστικά για τη συναρμολόγηση των δεδομένων που αποκτώνται από την τεχνολογία PacBio χρησιμοποιούνται τα προγράμματα:

- FALCON, το οποίο χρησιμοποιεί τον αλγόριθμο OLC και διατίθεται ελεύθερα από τον σύνδεσμο: **https://github.com/PacificBiosciences/falcon**
- HGAP, το οποίο χρησιμοποιεί τον αλγόριθμο OLC και διατίθεται ελεύθερα από τον σύνδεσμο: **https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP**

Το πρόγραμμα Canu, ωστόσο, είναι εκείνο που χρησιμοποιείται περισσότερο και το οποίο προτείνεται και από άλλους ερευνητές, καθώς

δημιουργεί συνεχείς και ακριβείς συναρμολογήσεις, ενώ θεωρείται ένα από τα καλύτερα προγράμματα για δεδομένα των οποίων η κάλυψη βαίνει μειούμενη. (Giordano et al., 2017). Το Canu μπορεί να συναρμολογήσει με αξιόπιστο τρόπο πλήρη μικροβιακά γονιδιώματα και σχεδόν ολοκληρωμένα ευκαρυωτικά χρωμοσώματα, χρησιμοποιώντας είτε την τεχνολογία PacBio ή την Oxford Nanopore, ενώ πέτυχε τη δημιουργία ενός contig με N50 μεγαλύτερο από 21 Mbp από δεδομένα που προέκυψαν από την τεχνολογία PacBio και προέρχονταν από τον άνθρωπο, καθώς και από το δίπτερο έντομο *Drosophila melanogaster* (Φρουτόμυγα) (Koren et al., 2017). Το πρόγραμμα Canu το προτείνουν, επίσης, οι δημιουργοί του προγράμματος Celera Assembler στην ιστοσελίδα τους: http://wgs-assembler.sourceforge.net/wiki/index.php?title=Main_Page.

Με το πρόγραμμα SPAdes διεξάγεται η υβριδική συναρμολόγηση, όπου συνδυάζονται τα δεδομένα των reads της τεχνολογίας PacBio με εκείνα της Illumina. Υπάρχουν και άλλα προγράμματα που λειτουργούν με παρόμοιο τρόπο για την υβριδική συναρμολόγηση των δεδομένων PacBio με μικρότερα reads, όπως εκείνα που προέρχονται από την Illumina. Ορισμένα προγράμματα που χρησιμοποιούν τα δεδομένα PacBio για να υλοποιήσουν την υβριδική συναρμολόγηση είναι τα εξής:

- pacBioToCA, το οποίο χρησιμοποιεί τον αλγόριθμο OLC και διατίθεται ελεύθερα από τον σύνδεσμο: **https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/Depricated---pacBioToCA**
- ECTools, το οποίο χρησιμοποιεί τον αλγόριθμο OLC και διατίθεται ελεύθερα από τον σύνδεσμο: **https://github.com/jgurtowski/ectools**
- Cerulean, το οποίο χρησιμοποιεί τον αλγόριθμο OLC και διατίθεται ελεύθερα από τον σύνδεσμο: **https://sourceforge.net/projects/ceruleanassembler/** (Deshpande et al., 2013).
- dbg2olc, το οποίο χρησιμοποιεί τον αλγόριθμο OLC και διατίθεται ελεύθερα από τον σύνδεσμο: **https://sites.google.com/site/dbg2olc/** (Ye et al., 2016).

- ALLPATHS-LG το οποίο χρησιμοποιεί τον αλγόριθμο DBG και διατίθεται ελεύθερα από τον σύνδεσμο: **http://software.broadinstitute.org/allpaths-lg/blog/**
- MIRA 4, το οποίο χρησιμοποιεί τον αλγόριθμο OLC και διατίθεται ελεύθερα από τον σύνδεσμο: **https://sourceforge.net/projects/mira-assembler/files/**

Το πρόγραμμα SPAdes θεωρείται ένα από τα καλύτερα προγράμματα για την εφαρμογή της υβριδικής συναρμολόγησης (Antipov et al., 2016). Αναφορικά με το SPAdes, αξίζει να αναφερθούν τα αποτελέσματα της μελέτης των (Gurevich et al., 2013), στην οποία συγκρίθηκαν διάφορα προγράμματα γονιδιωματικής συναρμολόγησης, χρησιμοποιώντας τα δεδομένα της αλληλούχισης του απομονωμένου γονιδιωματικού DNA μεμονωμένων κυττάρων του βακτηρίου *Escherichia coli*. Τα προγράμματα αυτά ήταν τα εξής: EULER-SR (Pevzner et al., 2001), Velvet (Zerbino και Birney, 2008), SOAPdenovo (Li et al., 2010), Velvet SC, EULER + Velvet SC (E + V-SC) (Chitsaz et al., 2011), IDBA-UD (Peng et al., 2012) και το SPAdes. Από τη σύγκριση αυτών των προγραμμάτων, αποδείχθηκε ότι το SPAdes και το IDBA-UD ήταν τα καλύτερα. Ειδικότερα, το πρόγραμμα SPAdes είχε τη μεγαλύτερη τιμή N50, και συγκεκριμένα ίση με 99.913, η οποία αποτελεί μέτρηση αξιολόγησης της ποιότητας της συναρμολόγησης. Επιπλέον, χρησιμοποιώντας το γονιδίωμα αναφοράς του βακτηρίου *Escherichia coli* διαπιστώθηκε ότι με το πρόγραμμα SPAdes συναρμολογήθηκε το υψηλότερο ποσοστό του γονιδιώματος, το οποίο αντιστοιχούσε σε 97% και ο μεγαλύτερος αριθμός σε πλήρη γονίδια. Συγκεκριμένα, με το πρόγραμμα SPAdes συναρμολογήθηκαν 4.071 από τα 4.324 γονίδια που βρίσκονται στο γονιδίωμα του εν λόγω βακτηρίου.

4.3 Ανάγκη τυποποιημένης επικύρωσης των τεχνολογιών αλληλούχισης νέας γενιάς

Η καινοτομία που απορρέει από την εφαρμογή των τεχνολογιών αλληλούχισης νέας γενιάς δε μπορεί να πραγματοποιηθεί χωρίς

αποτελεσματικά εργαλεία βιοπληροφορικής για τη διερεύνηση, ανάλυση και ερμηνεία των μεγάλων ποσοτήτων των παραγόμενων δεδομένων. Τα περισσότερα εργαστήρια που προσανατολίζονται στις εν λόγω εφαρμογές δε διαθέτουν εξειδικευμένο προσωπικό του κλάδου της βιοπληροφορικής, ενώ και για εκείνα τα εργαστήρια που απασχολούν, ήδη, ειδικούς στη βιοπληροφορική, θα πρέπει, για την ανάλυση των δεδομένων, να οριστεί και να επικυρωθεί μια ολοκληρωμένη διαδικασία για τη διοχέτευση της εκτέλεσης των εντολών (pipelining), ώστε να διαπιστωθεί ότι ο αλγόριθμος ενός λογισμικού αναλύει αξιόπιστα τα δεδομένα των αλληλουχιών και παράγει ακριβή αποτελέσματα. Πολλά βιοπληροφορικά εργαλεία διατίθενται σε εμπορικές, ακαδημαϊκές και άλλες ανοικτές πηγές. Ωστόσο, τα ειδικά εργαλεία που χρησιμοποιούνται ή αναπτύσσονται, εξαρτώνται σε μεγάλο βαθμό από την ανάγκη και τη σκοπούμενη χρήση του κάθε εργαστηρίου και ενδέχεται να μην έχουν δοκιμαστεί αυστηρά. Ένας κατάλληλος αγωγός διοχέτευσης εντολών (pipeline) για την ανάλυση των δεδομένων πρέπει να εφαρμοστεί και να επικυρωθεί πλήρως. Η ερμηνεία των δεδομένων πρέπει να χρησιμοποιείται επιτυχώς και αποτελεσματικά.

Τώρα που τα εργαστήρια εφαρμόζουν τις τεχνολογίες αλληλούχισης νέας γενιάς, απαιτούνται κοινές κατευθυντήριες οδηγίες και τυποποιημένα πρότυπα επικύρωσης. Η ανάπτυξη των εν λόγω τεχνολογιών και των συναφών εργαλείων βιοπληροφορικής θα συνεχίσει να σημειώνει ταχεία πρόοδο και, χωρίς αμφιβολία, θα υπάρχουν διαθέσιμα ολοένα και πιο ισχυρά εργαλεία. Η διαδικασία επικύρωσης είναι απαραίτητη για την ανάπτυξη μεθόδων και οι μέθοδοι αυτές πρέπει να είναι αξιόπιστες και κατάλληλες για τον κάθε σκοπό. Η επικύρωση περιγράφεται ως η διαδικασία που:

1. αξιολογεί την ικανότητα των διαδικασιών να παράγουν αξιόπιστα αποτελέσματα υπό καθορισμένες συνθήκες,
2. ορίζει αυστηρά τις προϋποθέσεις που απαιτούνται για την απόκτηση των αποτελεσμάτων,
3. καθορίζει τους περιορισμούς των διαδικασιών,
4. προσδιορίζει τομείς της ανάλυσης που πρέπει να παρακολουθούνται και να ελέγχονται,

5. δημιουργεί τη βάση για την ανάπτυξη κατευθυντήριων οδηγιών που αφορούν στην ερμηνεία των αποτελεσμάτων (Budowle et al., 2008 / Budowle et al., 2014).

Κάθε αναλυτής χρησιμοποιεί μια συγκεκριμένη βιοχημική μέθοδο αλληλούχισης και κάθε μία από αυτές τις μεθόδους θα πρέπει να επικυρώνεται γενικά, και στη συνέχεια ειδικά, σύμφωνα με τα χαρακτηριστικά του αναλυτή στον οποίο λαμβάνει χώρα (Metzker, 2010 / Quail et al., 2008 / Berglund et al., 2011 / Lam et al., 2011). Για παράδειγμα, οι βιοχημικές διεργασίες που χρησιμοποιούνται από τους αναλυτές Ion Torrent της Thermo Fisher Scientific και 454 της Roche τείνουν να είναι λιγότερο ακριβείς από τη βιοχημική μέθοδο που εφαρμόζουν οι αναλυτές της Illumina για την αλληλούχιση των ομοπολυμερών αλληλουχιών. Επίσης, ο τύπος της αλληλούχισης, είτε αφορά στον μεμονωμένου άκρου (single-end), κατά τον οποίο διενεργείται αλληλούχιση ενός από τους δύο κλώνους του δίκλωνου μορίου DNA και η οποία ξεκινά μόνο από το ένα άκρο, ή στον ζευγαρωμένων άκρων (paired-end), όπου η αλληλούχιση διεξάγεται και στους δύο κλώνους ξεκινώντας και στα δύο άκρα, μπορεί να επηρεάσει την κάλυψη και την ακρίβεια με διάφορους τρόπους.

Η αλληλούχιση Sanger, η οποία εξακολουθεί να θεωρείται ως το «πρότυπο» ποιότητας αλληλούχισης, επιτρέπει τη διεξαγωγή συγκριτικής ανάλυσης μίας αλληλουχίας στόχου με διαφορετικές τεχνολογίες αλληλούχισης. Ωστόσο, δεν υπάρχει καμία εγγύηση ότι το πρότυπο» παρέχει πάντα το σωστό αποτέλεσμα. Για παράδειγμα, ο Harismendy και οι συνεργάτες του (Harismendy et al., 2009) συνέκριναν τα αποτελέσματα της αλληλούχισης που διεξήχθη σε συνολικά 266 kb αλληλουχίας, η οποία αντιστοιχούσε σε έξι τμήματα γονιδίων, χρησιμοποιώντας τη μέθοδο Sanger, τις τρεις τεχνολογίες αλληλούχισης νέας γενιάς Roche 454, Illumina GA, και ABI SOLiD, καθώς και μία μέθοδο που βασιζόταν σε μια ειδική πλατφόρμα μικροσυστοιχίας (microarray). Τα ποσοστά των ψευδώς αρνητικών και των ψευδώς θετικών μονονουκλεοτιδικών πολυμορφισμών που αποδόθηκαν στην αλληλούχιση Sanger ανέρχονταν σε 0,9% και 3,1%, αντίστοιχα. Επιπλέον, η χαμηλότερη απόδοση και κάλυψη της αλληλούχισης Sanger την καθιστά ως κατώτερη μεθοδολογία σε σχέση με τα δεδομένα που παράγονται από τις

τεχνολογίες αλληλούχισης νέας γενιάς. Ειδικότερα, τα παραγόμενα ανά εκτέλεση δεδομένα της αλληλούχιση νέας γενιάς είναι τόσο ανώτερα από εκείνα της αλληλούχισης Sanger που μόνο περιορισμένες δειγματοληψίες και πολύ μικρές περιοχές μπορούν να συγκριθούν μεταξύ τους.

Αντ' αυτού, η συγκριτική δοκιμή μπορεί να επιτευχθεί καλύτερα μέσω της ανεξάρτητης δοκιμής των αναλυτών νέας γενιάς με πρότυπα αλληλουχιών αναφοράς. Τα πιθανά σφάλματα και οι μεροληψίες που αποτελούν εγγενείς ιδιότητες κάθε αναλυτή μπορούν να προσδιοριστούν και να τεκμηριωθούν καλύτερα με αυτόν τον τρόπο. Για κάθε τύπο δείγματος, καθώς και για κάθε τύπο και μοντέλο αναλυτή, ο ρυθμός και το προφίλ σφάλματος της αλληλουχίας μπορεί να προσδιοριστεί μόνο με εμπειρικές δοκιμές. Τα δεδομένα μπορούν να χρησιμοποιηθούν για να ορίσουν τους περιορισμούς του κάθε αναλυτή, τα οποία θα πρέπει να αποτελούν μέρος της ερμηνείας των τυποποιημένων διαδικασιών λειτουργίας (standard operating procedures, SOPs). Επιπλέον, οι ανεξάρτητες δοκιμές επιτρέπουν τον εντοπισμό των αδυναμιών και τη βελτίωση των δοκιμασιών πριν από την εφαρμογή. Συμπερασματικά, όπου είναι δυνατόν, θα πρέπει να χρησιμοποιούνται ανεξάρτητες αναλύσεις για την τυποποιημένη επικύρωση των μεθόδων αλληλούχισης νέας γενιάς (Budowle et al., 2014).

Η επιστήμη της βιοπληροφορικής είναι βασική και κρίσιμη, λόγω της ανάγκης για διαχείριση του τεράστιου όγκου των παραγόμενων δεδομένων που σχετίζονται με την απαίτηση για απαντήσεις επί τοξικολογικής, εγκληματολογικής, κλινικής και ερευνητικής φύσεως ερωτημάτων, τα οποία αφορούν είτε σε ίχνη ή σε σύνθετα δείγματα. Από τοξικολογικής και εγκληματολογικής άποψης τα δείγματα αυτά αφορούν και σε πιθανά ευρήματα γενετικής μηχανικής, καθώς και σε ενδημικά μικροβιολογικά δείγματα, στα οποία θα εμπεριέχεται το σύνολο των μικροοργανισμών που βρίσκονται σε μια συγκεκριμένη τοποθεσία. Είναι σημαντικό να διαπιστωθεί ότι οι αλγόριθμοι των χρησιμοποιούμενων λογισμικών αναλύουν αξιόπιστα τα δεδομένα της αλληλούχισης για να παράγουν ακριβή τελικά αποτελέσματα.

Οι μετρήσεις ποιότητας που δημιουργούνται κατά τη διάρκεια της αναλυτικής διαδικασίας περιλαμβάνουν: βαθμολογίες ποιότητας (Q-scores), έλεγχος ποιότητας των reads μέσω αποκοπής και αφαίρεσης τμημάτων ή

ολόκληρων των reads (trimming), στοίχιση (alignment), περιεχόμενο σε βάσεις GC, βάθος κάλυψης (coverage), μεροληψία, εναλλακτική ταυτοποίηση βάσεων. Η ταυτοποίηση των βάσεων (base calling), δηλαδή, η αναγνώριση ενός συγκεκριμένου νουκλεοτιδίου που υπάρχει σε κάθε θέση σε ένα read, θα πρέπει να αποτελεί μέρος του λογισμικού ενός οργάνου. Ένα κατώτατο όριο ποιότητας ορίζεται με μια βαθμολογία Q. Ένα κατώτατο όριο της τάξης του Q20 θέτει την ελάχιστη ακρίβεια ταυτοποίησης στο 99% επιτρέποντας μία λανθασμένη ταυτοποίηση στις εκατό βάσεις ενός read, ενώ μια βαθμολογία Q30 θέτει την ακρίβεια στο 99,9% και άρα μία λανθασμένη ταυτοποίηση βάσης στις 1.000 (Ewing και Green, 1998).

Γενικά, θα πρέπει να καθοριστεί ένα όριο βαθμολογίας Q τόσο για τη διαδικασία της επικύρωσης όσο και για την επακόλουθη εφαρμογή. Ωστόσο, δεν υπάρχουν κατευθυντήριες γραμμές που να δείχνουν ότι, για παράδειγμα, μια βαθμολογία Q20 αποτελεί μια απαίτηση. Μια βαθμολογία μικρότερη από Q20 θα μπορούσε να μην επηρεάσει την ακρίβεια, καθώς η μετέπειτα κάλυψη μπορεί να είναι επαρκής. Υπό καθορισμένες συνθήκες και για λόγους διερεύνησης, ή για δύσκολες περιστάσεις, η βαθμολογία ποιότητας μπορεί να τεθεί κάπως πιο χαλαρά. Ωστόσο, θα πρέπει να τεκμηριώνεται η αιτιολόγηση, ή η αξιοπιστία στην εφαρμογή μιας χαμηλότερης βαθμολογίας. Για τη γονιδιωματική ανάλυση προτείνεται να οριστεί η βαθμολογία Q30 και άνω. Συγκεκριμένα, η βιοχημική μεθοδολογία αλληλούχισης της Illumina παρέχει υψηλή ακρίβεια, με τη συντριπτική πλειοψηφία των βάσεων να έχει βαθμολογία ποιότητας Q30 και άνω. Αυτό το επίπεδο της ακρίβειας κρίνεται από την Illumina ως το ιδανικό για μια ευρεία σειρά εφαρμογών αλληλούχισης, συμπεριλαμβανομένης της κλινικής έρευνας (Illumina Sequencing Quality Scores).

Κάθε αναλυτής θα έχει συγκεκριμένους περιορισμούς και σφάλματα αλληλούχισης: φθορά έντασης σήματος, εσφαλμένες προσθήκες και διαγραφές βάσεων, μεροληψία κ.ο.κ. Οι περιορισμοί αυτοί θα πρέπει να περιγραφούν και να καθοριστούν. Η ακρίβεια της ταυτοποίησης των παραλλαγών αλληλουχίας, όπως οι μονονουκλεοτιδικοί πολυμορφισμοί, προσθήκες και διαγραφές, χρωμοσωμικές αναδιατάξεις, εξαρτάται από έναν αριθμό παραγόντων που περιλαμβάνουν την ταυτοποίηση των βάσεων, την

στοίχιση, την επιλογή του γονιδιώματος αναφοράς, της κάλυψης, της βιοχημικής μεθόδου και των εγγενών χαρακτηριστικών του αναλυτή. Επειδή η στοίχιση περιλαμβάνει την οργάνωση των reads με οδηγό μία αλληλουχία αναφοράς, διαφορετικές στρατηγικές στοίχισης μπορούν να παράγουν διαφορετικά αποτελέσματα. Έτσι, οι διαφορές στην στοίχιση θα διαφέρουν ανάλογα με το λογισμικό, οπότε πρέπει να οριστούν κανόνες στοίχισης για λόγους συνέπειας και ανιχνευσιμότητας.

Αναφορικά με τη διαχείριση του λογισμικού βιοπληροφορικής, δεν έχουν ακόμη καθοριστεί ενιαίες κατευθυντήριες γραμμές ή πρωτόκολλα για τη συγκριτική αξιολόγηση των λογισμικών. Έτσι, οι χρήστες πρέπει να επικυρώσουν πλήρως και να τεκμηριώσουν τους χρησιμοποιούμενους αγωγούς διοχέτευσης των εντολών που εφαρμόζονται για τη βιοπληροφορική ανάλυση των δεδομένων τους. Το λογισμικό μπορεί να διατίθεται ελεύθερα, να αγοράζεται από εμπορικούς φορείς, να αναπτύσσεται εσωτερικά ή να προέρχεται από έναν συνδυασμό πηγών. Τα λογισμικά προγράμματα πρέπει να εκτελούν γενική αξιολόγηση των μετρήσεων της ποιότητας, αλλά το λογισμικό ενδέχεται να διαφέρει στην απόδοση και να αποφέρει διαφορετικά αποτελέσματα. Επομένως, θα πρέπει να διεξάγεται ένα προσεκτικός έλεγχος και δοκιμές με σκοπό να επιλεχθεί το κατάλληλο λογισμικό που θα είναι απαραίτητο (Ellard et al., 2012), όχι μόνο για την τυποποιημένη επικύρωση αλλά και για τις αναλύσεις των δεδομένων.

Τέλος, με δεδομένες τις δυνατότητες των τεχνολογιών αλληλούχισης νέας γενιάς, η διαθεσιμότητα βάσεων δεδομένων αναφοράς υψηλής ποιότητας που να περιέχουν αλληλουχίες ολόκληρων μικροβιακών γονιδιωμάτων αποτελεί έναν κύριο στόχο. Για να είναι υψηλής ποιότητας, οι βάσεις δεδομένων πρέπει να είναι επιμελημένες και ακριβείς όσον αφορά στις αλληλουχίες, στα μεταδεδομένα (πληροφορίες σχετικές με τα δεδομένα) και στην κάλυψη της γενετικής ποικιλομορφίας. Η ανάπτυξη βάσεων δεδομένων που θα περιέχουν τις πλήρεις αλληλουχίες ολόκληρων μικροβιακών γονιδιωμάτων θα συμβάλει στην επιτυχή ανίχνευση και παρακολούθηση των παθογόνων, καθώς και των μικροοργανισμών που απασχολούν τον κλάδο της εγκληματολογικής μικροβιολογίας (Sjödin et al., 2013).

4.4 Κλινικές εφαρμογές των τεχνολογιών αλληλούχισης νέας γενιάς

Η κλινική εφαρμογή των τεχνολογιών αλληλούχισης νέας γενιάς για τον *de novo* προσδιορισμό γονιδιωματικών αλληλουχιών, διεξήχθη για πρώτη φορά με την αλληλούχιση του γονιδιώματος του βακτηρίου *Acinetobacter baumannii*, το οποίο ευθύνεται για σοβαρές ενδονοσοκομειακές λοιμώξεις (Smith et al., 2007). Συνεπώς, η γνώση της παθογένεσης, των μηχανισμών ανθεκτικότητας στα αντιβιοτικά, και οι θεραπευτικές προοπτικές έναντι του εν λόγω βακτηρίου αποτελούν σημαντικούς επιστημονικούς στόχους (Lee et al., 2017).

Στις μέρες μας, η κλινική χρήση της αλληλούχισης νέας γενιάς εστιάζεται, κυρίως, στο ανθρώπινο γονιδίωμα, για σκοπούς όπως ο χαρακτηρισμός της μοριακής βάσης του καρκίνου, καθώς και η διάγνωση και η κατανόηση της βάσης των σπάνιων γενετικών διαταραχών. Η τεχνολογία αλληλούχισης νέας γενιάς χρησιμοποιείται και στη μικροβιακή διαγνωστική για την ανακάλυψη νέων παθογόνων.

Από κλινικής άποψης η αλληλούχιση νέας γενιάς διαθέτει πολλά πλεονεκτήματα έναντι των παραδοσιακών μικροβιακών διαγνωστικών μεθόδων, όπως αμερόληπτα πρωτόκολλα, ικανότητα ανίχνευσης δύσκολα ή μη καλλιεργήσιμων μικροοργανισμών και ικανότητα ανίχνευσης μικτών μολύνσεων. Ένα από τα πιο εντυπωσιακά πλεονεκτήματα της αλληλούχισης νέας γενιάς είναι ότι απαιτεί ελάχιστη ή καθόλου προηγούμενη γνώση του παθογόνου, σε αντίθεση με πολλές άλλες διαγνωστικές δοκιμές (Chapter 15 - Next-Generation Sequencing for Pathogen Detection and Identification / High-throughput sequencing for the study of bacterial pathogen biology).

Η υψηλή ανάλυση που προσφέρεται από τις τεχνολογίες αλληλούχισης νέας γενιάς επιτρέπει την εκτίμηση των οδών μετάδοσης των πανδημιών, καθώς και των τοπικών εστιών μόλυνσης, την ταυτοποίηση των μοριακών μηχανισμών των παθογόνων και την εξελικτική ανάλυση των βακτηριακών πληθυσμών κατά τη μόλυνση μεμονωμένων ασθενών. Η πρόοδος και η έκβαση των λοιμωδών νοσημάτων προσδιορίζεται από τη δυναμική των αλληλεπιδράσεων ξενιστή-παθογόνου και πρόσφατες μελέτες που

χρησιμοποιούν μεθοδολογίες αλληλούχισης νέας γενιάς έχουν προσφέρει νέες γνώσεις σχετικά με την εξέλιξη των βακτηριακών παθογόνων κατά τη διάρκεια του εποικισμού και της μόλυνσης (Bryant et al., 2013 / Golubchik et al., 2013 / Young et al., 2012 / Lieberman et al., 2011).

Οι παραδοσιακές μέθοδοι της βακτηριακής γονοτύπησης έχουν, συχνά, περιορισμένη χρήση για επιδημιολογικές έρευνες σχετικά με τις επιδημικές εξάρσεις μολυσματικών ασθενειών, λόγω του χαμηλού επιπέδου της διακριτικής τους ισχύος (Köser et al., 2012). Το επίπεδο της γενετικής παραλλακτικότητας που παρατηρείται μεταξύ στελεχών επιδημιολογικά συνδεδεμένων με την εκδήλωση μιας νόσου, μπορεί να συμβάλει στον προσδιορισμό της απόδοσης της ευθύνης σε ένα μεμονωμένο ή σε πολλαπλά στελέχη (Reuter et al., 2013 / Eyre et al., 2012 / Köser et al., 2012, Gardy et al., 2011). Επιπλέον, η αλληλούχιση νέας γενιάς έχει εφαρμοστεί αναδρομικά για την ανάλυση επιδημιών που προκλήθηκαν από στελέχη του βακτηρίου *Staphylococcus aureus* σε νοσοκομεία, αποκαλύπτοντας τη δυναμική των επιδημιών που περιορίζονται σε ενδονοσοκομειακούς χώρους (Eyre et al., 2012 / Köser et al., 2012), αλλά και μεταξύ του νοσοκομείου και του εξωνοσοκομειακού χώρου (Harris et al., 2013). Η κατευθυντικότητα της μετάδοσης, αν και συχνά διφορούμενη, μπορεί μερικές φορές να συναχθεί από τον συνδυασμό των επιδημιολογικών δεδομένων και της ανάλυσης των αλληλουχιών του γονιδιώματος ενός βακτηριακού πληθυσμού (Bryant etal., 2013 / Lieberman et al., 2011). Επίσης, η ανάπτυξη καινοτόμων βιοπληροφορικών αλγορίθμων διευκόλυνε την ταυτοποίηση των γεγονότων της μετάδοσης, ενώ ταυτόχρονα κατέδειξε την ετερογένεια που υπάρχει στη μολυσματικότητα των βακτηριακών πληθυσμών (Eyre et al., 2013).

Η ικανότητα αλληλούχισης μεγάλου αριθμού στενά συγγενικών απομονωθέντων προϊόντων επιτρέπει την υψηλής ανάλυσης κατασκευή φυλογενέσεων, οι οποίες μπορεί να παρέχουν χρήσιμες πληροφορίες σχετικά με τις διαδικασίες στις οποίες βασίζεται η εμφάνιση και η εξάπλωση των παθογόνων στελεχών. Οι βακτηριακοί πληθυσμοί συσσωρεύουν τυχαίες μεταλλάξεις με την πάροδο του χρόνου μέσω εγγενών ρυθμών μετάλλαξης που είναι συγκεκριμένοι για κάθε μικροοργανισμό και τη συναφή του οικολογία (McAdam et al., 2014). Οι εκτιμήσεις του ρυθμού μετάλλαξης για ένα

δεδομένο βακτηριακό πληθυσμό επιτρέπουν την κατασκευή χρονικά διαβαθμισμένων φυλογενετικών δέντρων (Gray et al., 2011). Ειδικότερα, η εφαρμογή των φυλογενετικών μεθόδων κατά Bayes σε αλληλουχίες στενά συγγενικών βακτηρίων, μπορεί να επιτρέψει μια χρονικά κλιμακούμενη ανακατασκευή του εξελικτικού ιστορικού και της γεωγραφικής τους διάδοσης και μπορεί, επίσης, να αποκαλύψει γενετικά γεγονότα που σχετίζονται με την εμφάνιση συγκεκριμένων βακτηριακών στελεχών.

Η φυλογενετική ανάλυση που βασίζεται στη χρήση των τεχνολογιών αλληλούχισης νέας γενιάς έχει βελτιώσει στην κατανόηση της ικανότητας των βακτηριακών παθογόνων να αλλάζουν είδη ξενιστών και να προσαρμόζονται, ώστε να μπορούν να επιβιώσουν και να εξαπλωθούν σε νέους πληθυσμούς ξενιστών (Spoor et al., 2013 / Price et al., 2012 / Lowder et al., 2009). Συγκεκριμένα, αρκετές μελέτες έχουν εντοπίσει τα ζώα ως δεξαμενές για αναδυόμενα βακτηριακά στελέχη ικανά να προκαλέσουν ασθένειες στον άνθρωπο (Lebreton et al., 2013 / Spoor et al., 2013 / Price et al., 2012). Για παράδειγμα, οι Price και οι συνεργάτες του εξέτασαν τη συσχέτιση πολλαπλών ξενιστών του στελέχους ST398 του βακτηρίου *Staphylococcus aureus* εφαρμόζοντας μεθοδολογία αλληλούχισης νέας γενιάς, παρέχοντας στοιχεία για την εμφάνιση ανθεκτικών σε αντιβιοτικά στελεχών που προκύπτουν από τη χρήση αντιβιοτικών στην κτηνοτροφική βιομηχανία (Price et al., 2012). Ομοίως, ο Spoor και οι συνεργάτες του κατέδειξαν ότι οι αγελάδες είναι μια πιθανή δεξαμενή νέων στελεχών του βακτηρίου *Staphylococcus aureus* με την ικανότητα για εξάπλωση πανδημίας στον άνθρωπο (Spoor et al., 2013).

Εκτός από την κατανόηση της δυναμικής της μετάδοσης, η αλληλούχιση νέας γενιάς επιτρέπει και την υλοποίηση υψηλής ανάλυσης των γενετικών συσχετισμών της εξειδίκευσης του ξενιστή. Για παράδειγμα, σε μία από τις λίγες μελέτες συσχέτισης σε επίπεδο γονιδιώματος (genome-wide association studies, GWAS) των βακτηρίων, ο Shepphard και οι συνεργάτες του αναγνώρισαν μια γονιδιωματική περιοχή που κωδικοποιεί τα συστατικά της βιοσύνθεσης της βιταμίνης B5, η οποία σχετίζεται με την προσαρμογή του βακτηρίου *Campylobacter jejuni* που χρησιμοποιεί ως ξενιστή τα βοοειδή (Sheppard et al., 2013). Επίσης, μέσω της εφαρμογής της αλληλούχισης νέας

γενιάς σε γονιδιώματα των απομονωθέντων στελεχών DT2 του βακτηρίου *Salmonella typhimurium*, διαπιστώθηκε ότι η προσαρμογή του στελέχους DT2 στο περιστέρι δεν προκαλείται από την απόκτηση νέου γενετικού υλικού, αλλά από πολυμορφισμούς στο προϋπάρχον γενετικό υλικό (Kingsley et al., 2013).

4.5 Τοξικολογικές και εγκληματολογικές εφαρμογές των τεχνολογιών αλληλούχισης νέας γενιάς

Εκτός από τις εφαρμογές σε απομονωμένα δείγματα DNA βιολογικών υλικών ανθρώπων και ζώων, μελέτες που αξιοποιούν τις τεχνολογίες αλληλούχισης νέας γενιάς έχουν διεξαχθεί με σκοπό την ανεύρεση τοξικολογικών παραγόντων. Ως ένα σημαντικό παράδειγμα μπορεί να αναφερθεί η εξέταση της ποιότητας του πόσιμου νερού για την ανεύρεση επικίνδυνων μολυσματικών παραγόντων που αφορούν σε μικροοργανισμούς και ιούς, σε τοξίνες, καθώς και σε προϊόντα βιοαποικοδόμησης των μολυσματικών παραγόντων.

Ενώ οι περισσότεροι μικροοργανισμοί στο επεξεργασμένο πόσιμο νερό είναι αβλαβείς, οι εστίες ασθενειών που σχετίζονται με τους παθογόνους παράγοντες στο πόσιμο νερό μπορεί να οφείλονται σε ακατάλληλη επεξεργασία, σε ζημίες του δικτύου υδροδότησης και σε μολυσμένα νερά των πηγών. Αρκετές πρόσφατες μελέτες έχουν χρησιμοποιήσει τις τεχνολογίες αλληλούχισης νέας γενιάς για να εξετάσουν τους μικροβιακούς πληθυσμούς, συμπεριλαμβανομένων των παθογόνων, στις πηγές ύδατος, τις βρύσες και στα διάφορα στάδια της επεξεργασίας του πόσιμου νερού (Tan et al., 2015).

Ορισμένες από τις πιο πρόσφατες μελέτες στις οποίες εφαρμόστηκαν οι τεχνολογίες αλληλούχισης νέας γενιάς για τη μικροβιακή και τοξικολογική έρευνα στο πόσιμο νερό περιλαμβάνουν τις εξής:

- Inkinen et al. (2017). Bacterial community changes in copper and PEX drinking water pipeline biofilms under extra disinfection and magnetic water treatment.

- Pereira et al. (2017). Development of a genus-specific next generation sequencing approach for sensitive and quantitative determination of the Legionella microbiome in freshwater systems.
- Lee et al. (2017). Novel Primer Sets for Next Generation Sequencing-Based Analyses of Water Quality.
- Liu et al. (2017). Bacterial community radial-spatial distribution in biofilms along pipe wall in chlorinated drinking water distribution system of East China.

Στο πεδίο της εγκληματολογικής μικροβιολογίας (Budowle et al., 2003 / Budowle et al. (Eds): Microbial Forensics 2, 2011 / Breeze et al. (Eds): Microbial Forensics, 2005). η αλληλούχιση νέας γενιάς, σε συνδυασμό με τις ισχυρές δυνατότητες της βιοπληροφορικής, προσφέρουν ένα ισχυρό εργαλείο για τον χαρακτηρισμό εγκληματολογικών πειστηρίων που περιλαμβάνουν μικροοργανισμούς που βρίσκονται σε μια σκηνή εγκλήματος, άγνωστους μικροοργανισμούς, γενετικά τροποποιημένους μικροοργανισμούς και στοιχεία γενετικής μηχανικής, μικροοργανισμούς που βρίσκονται σε εξαιρετικά χαμηλή αφθονία. Επιπλέον, δύναται να ταυτοποιηθεί η προέλευση των μικροοργανισμών, η ευαισθησία τους σε αντιβιοτικά και το προφίλ της μολυσματικότητάς τους (Budowle et al., 2013, The microbial forensics pathway for use of massively-parallel sequencing technologies).

Για την επιστήμη της εγκληματολογικής μικροβιολογίας, η αξία της αλληλούχισης νέας γενιάς έγκειται στο γεγονός ότι δημιουργεί μεγάλες ποσότητες δεδομένων υψηλής ποιότητας και στην ταχύτητα με την οποία εκτελούνται οι αναλύσεις, γεγονός που συνεπάγεται στην ταχύτητα συλλογής των αποδεικτικών στοιχείων και στην ταχύτητα του χαρακτηρισμού και της αποτίμησης των πειστηρίων προς την επίλυση των σχετικών εγκληματολογικών υποθέσεων. Ένα άλλο προτέρημα των τεχνολογιών αλληλούχισης νέας γενιάς είναι ότι η υψηλή τους απόδοση συνδυάζεται με μειωμένο κόστος αλληλούχισης ανά νουκλεοτίδιο, ή ανά γονιδίωμα, σε σύγκριση με την αλληλούχιση Sanger (Wetterstrand KS. DNA sequencing costs) και η ευκολία στη χρήση τους που παρέχεται από την μεγάλης κλίμακας δυνατότητα αυτοματισμού των σύγχρονων αναλυτών. Έτσι λοιπόν, με την εφαρμογή των μεθοδολογιών αλληλούχισης νέας γενιάς, εκατομμύρια

αντιδράσεων αλληλούχισης μπορούν να εκτελεσθούν ταυτόχρονα και μαζικά σε μια μεμονωμένη εκτελεστική διαδικασία που λαμβάνει χώρα σε μόνον έναν αναλυτή (Brenner et al., 2000) / Shendure και Ji, 2008). Επίσης, επειδή η απόδοση και η ακρίβεια είναι ιδιαίτερα αυξημένες, περισσότερα δείγματα μπορούν να αναλύονται σε μία μόνο εκτελεστική διαδικασία χωρίς να θυσιάζεται το βάθος της κάλυψης, ή μπορούν να αναλύονται πιο σύνθετα δείγματα σε μεγαλύτερο βάθος κάλυψης.

Με τις τεχνολογίες αλληλούχισης νέας γενιάς είναι δυνατόν να εντοπιστούν, γρήγορα και με αμερόληπτο τρόπο, οι μη ανιχνεύσιμες με άλλες μεθόδους γονιδιωματικές διαφορές μεταξύ στενά συγγενικών απομονωθέντων βακτηριακών στελεχών. Αυτή η εξέλιξη είναι ιδιαίτερα σημαντική για τις πιο θανατηφόρες βακτηριακές ασθένειες που προκαλούνται από βακτηριακά στελέχη που έχουν μεταξύ τους εξαιρετικά χαμηλά επίπεδα γενετικής ποικιλομορφίας και, άρα, είναι δύσκολο να διακριθούν το ένα από το άλλο. Η ανάλυση ολόκληρων των γονιδιωμάτων των παθογόνων προβλέπεται να είναι όλο και πιο σημαντική για τον σκοπό αυτό. Στο πλαίσιο της εγκληματολογικής μικροβιολογίας, η πλήρης ανάλυση της αλληλουχίας ολόκληρου του γονιδιώματος θεωρείται η απόλυτη μέθοδος για συγκρίσεις μεταξύ στελεχών, καθώς δίνει την απαραίτητη και έγκυρη πληροφόρηση για την ταυτοποίηση, τον χαρακτηρισμό και την τελική αποτίμηση - τα τρία κύρια στάδια της εγκληματολογικής έρευνας (Sjödin et al., 2013).

ΒΙΒΛΙΟΓΡΑΦΙΑ

Aarestrup FM, Brown EW, Detter C, Gerner-Smidt P, Gilmour MW, et al. (2012). Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. Emerg Infect Dis. 18(11): e1. doi: 10.3201/eid/1811.120453

Adalja AA, Toner E, Inglesby TV. (2015). Clinical Management of Potential Bioterrorism-Related Conditions. New England Journal of Medicine. 372(10): 954–962. doi: 10.1056/NEJMra1409755

Advantages of Biologics as Weapons Bioterrorism: A Threat to National Security or Public Health Defining Issue? MM&I 554 University of Wisconsin–Madison and Wisconsin State Laboratory of Hygiene, September 30, 2008

Advantages of paired-end and single-read sequencing, https://www.illumina.com/science/technology/next-generation-sequencing/paired-end-vs-single-read-sequencing.html

Al Jazeera English (2010) UN investigates Haiti outbreak. Al Jazeera Americas. Doha: Al Jazeera Satellite Network. http://www.aljazeera.com/news/africa/2010/10/2010102841412141967.html

Altschul Stephen, Gish Warren, Miller Webb, Myers Eugene, Lipman David (1990). Basic local alignment search tool. Journal of Molecular Biology. 215 (3): 403-410. doi: 10.1016/S0022-2836(05)80360-2

Anderson S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Research. 9(13): 3015-27. doi: 10.1093/nar/9.13.3015

Antipov D, Korobeynikov A, McLean JS, Pevzner PA. (2016). hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics. 32 (7): 1009-15. doi: 10.1093/bioinformatics/btv688

ASCII, From Wikipedia, the free encyclopedia, https://en.wikipedia.org/wiki/ASCII

Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J. (2014). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat. Biotechnol. 33(3):296-300. doi: 10.1038/nbt.3103

Bacillus anthracis Incident, Kameido, Tokyo, 1993, CDC

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology. 19: 455–477. doi: 10.1089/cmb.2012.0021

BBC News (2010) Haiti cholera outbreak: Nepal troops not tested. BBC News South Asia. London: British Broadcasting Corporation, http://www.bbc.co.uk/news/world-south-asia-11949181

Benčić Z, Sinha R. (1972). Cholera carriers and circulation of cholera vibrios in the community. Int J Epidemiol. 1(1):13-14

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 456 (7218): 53-9. doi: 10.1038/nature07517.

Berglund EC, Kiialainen A, Syvänen AC. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. Investig Genet. 2:23. doi: 10.1186/2041-2223-2-23

Biostars, «How to assemble contigs?», https://www.biostars.org/p/253222/

Bioterrorism | Anthrax | CDC, www.cdc.gov

Bioterrorism Agents/Diseases, https://emergency.cdc.gov/agent/agentlist-category.asp, CDC

Bioterrorism Overview. Centers for Disease Control and Prevention, www.cdc.gov

BLAST Basic Local Alignment Search Tool, Blast Program Selection Guide https://www.ncbi.nlm.nih.gov/blast/BLAST_guide.pdf

BLAST 2 SEQUENCES, a new tool for comparing protein and nucleotide sequences. https://www.researchgate.net/publication/240157167_BLAST_2_SEQUENCES_a_new_tool_for_comparing_protein_and_nucleotide_sequences

Botulism - Emergency Preparedness & Response, CDC

Breeze RG, Budowle B, Schutzer SE. (Eds): Microbial Forensics. Amsterdam: Academic Press; 2005

Brenner S, Williams SR, Vermaas EH, Storck T, Moon K, McCollum C, Mao JI, Luo S, Kirchner JJ, Eletr S, DuBridge RB, Burcham T, Albrecht G. (2000). In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. Proc Natl Acad Sci U S A. 15;97(4): 1665-70

Brucellosis - Emergency Preparedness & Response CDC

Bryant JM., Grogono DM., Greaves D, Foweraker J, Roddick I, Inns T, Reacher M, Haworth CS, Curran MD, Harris SR. (2013). Whole-genome sequencing to identify transmission of Mycobacterium abscessus between patients with cystic fibrosis: a retrospective cohort study. Lancet. 381(9877):1551-60. doi: 10.1016/S0140-6736(13)60632-7

Budowle B, Connell ND, Bielecka-Oder A, Colwell RR, Corbett CR, Fletcher J, Forsman M, Kadavy DR, Markotic A, Morse SA, Murch RS, Sajantila A, Schmedes SE, Ternus KL, Turner SD, Mino S. (2014). Validation of high throughput sequencing and microbial forensics applications. Investigative Genetics. 5: 9. doi: 10.1186/2041-2223-5-9

Budowle B, Schutzer SE, Breeze RG, Keim PS, Morse SA. (Eds): Microbial Forensics. 2. Amsterdam: Academic Press; 2011

Budowle B, Schutzer SE, Einseln A, Kelley LC, Walsh AC, Smith JA, Marrone BL, Robertson J, Campos J. (2003). Public health. Building microbial forensics as a response to bioterrorism. Science. 301(5641):1852-3

Budowle B, Schutzer SE, Morse SA, Martinez KF, Chakraborty R, Marrone BL, Messenger SL, Murch RS, Jackson PJ, Williamson P, Harmon R, Velsko SP. (2008). Criteria for validation of methods in microbial forensics. Appl Environ Microbiol. 74(18): 5599-607. doi: 10.1128/AEM.00966-08

Budowle B, Schmedes SE, Murch RS. The Science and Applications of Microbial Genomics. Washington DC: The National Academies Press; 2013. The microbial forensics pathway for use of massively-parallel sequencing technologies; pp. 117–133

Burkholderia mallei, From Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Burkholderia_mallei

Canu Quick Start – canu 1.6 documentation,
http://canu.readthedocs.io/en/latest/quick-start.html

Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics. 13:375. doi: 10.1186/1471-2164-13-375

Casey RM. (2005). BLAST Sequences Aid in Genomics and Proteomics. Business Intelligence Network

Centers for Disease Control and Prevention (2010) Update: cholera outbreak - Haiti, 2010. MMWR Morb Mortal Wkly Rep 59: 1473–1479

Chaisson MJ, Tesler G. (2012). Mapping single molecule sequencing reads using Basic Local Alignment with Successive Refinement (BLASR): Application and Theory. BMC Bioinformatics. 13:238. https://doi.org/10.1186/1471-2105-13-238

Chakraborty A, Ghosh S, Chowdhary G, Maulik U, Chakrabarti S. (2012). DBETH: a Database of Bacterial Exotoxins for Human. Nucleic Acids Researh. 40 (Database issue): D615-20. doi: 10.1093/nar/gkr942

Chapter 15 - Next-Generation Sequencing for Pathogen Detection and Identification
https://www.sciencedirect.com/science/article/pii/S0580951715000136

Chen LH, Xiong ZH, Sun LL, Yang J, Jin Q. (2012). VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res. 40 (Database issue): D641-D645

Chen LH, Yang J, Yu J, Yao ZJ, Sun LL, Shen Y, Jin Q. (2005). VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 33 (Database issue): D325-D328

Chen LH, Zheng DD, Liu B, Yang J, Jin Q (2016). VFDB 2016: hierarchical and refined dataset for big data analysis-10 years on. Nucleic Acids Res. 44 (Database issue): D694-D697

Chitsaz H, Yee-Greenbaum JL, Tesler G, Lombardo MJ, Dupont CL, Badger JH, Novotny M, Rusch DB, Fraser LJ, Gormley NA, Schulz-Trieglaff O, Smith GP, Evers DJ, Pevzner PA, Lasken RS. (2011). Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. Nat Biotechnol. 29: 915–921. doi: 10.1038/nbt.1966

CNN, http://edition.cnn.com/2008/CRIME/08/06/anthrax.case/index.html

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research. 38 (6): 1767–1771. doi: 10.1093/nar/gkp1137

Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honoré N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby T, Jagels K, Lacroix C, MacLean J, Moule S, Murphy L, Oliver K, Quail MA. (2001). Massive gene decay in the leprosy bacillus. Nature. 409 (6823): 1007-1011. doi: 10.1038/35059006.

Collins FS, Hamburg MA (2013). First FDA authorization for next-generation sequencer. N Engl J Med. 369(25): 2369-71. doi: 10.1056/NEJMp1314561

Coordinating Committee for International Staff Unions and Associations, LEAKED UN REPORT ON HAITI'S CHOLERA OUTBREAK SLAMMED SANITATION AT ITS BASES, http://www.ccisua.org/2016/04/07/leaked-un-report-on-haitis-cholera-outbreak-slammed-sanitation-at-its-bases/

Coverage (genetics) From Wikipedia, the free encyclopedia, https://en.wikipedia.org/wiki/Coverage_(genetics)

Crumly Jim (TMFSpeyside) Dec 11, 2017, Better Buy: Illumina, Inc. vs. Pacific Biosciences, https://www.fool.com/investing/2017/12/11/better-buy-illumina-inc-vs-pacific-

Dan Spiegelman, Université de Montréal, https://www.researchgate.net/post/What_is_N50_Scaffold_in_Genome_sequencing_Technique

de la Cruz F, Davies J. (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. Trends Microbiol. 8(3): 128-33

Delva JG (2010) Quake-hit Haiti battles cholera epidemic, 150 dead. Reuters US. New York City: Thomson Reuters, http://www.reuters.com/article/2010/10/22/us-haiti-cholera-idUSTRE69L21520101022

Deshpande V, Fung EDK, Pham S, Bafna V. (2013). Cerulean: A hybrid assembly using high throughput short and long reads. arXiv preprint arXiv:1307.7933.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 36(16): e105 doi: 10.1093/nar/gkn425

Dunham I. (2005). Genome Sequencing. Encyclopedia of Life Sciences. doi: 10.1038/npg.els.0005378.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science. 323 (5910): 133-8. doi: 10.1126/science.1162986

Ellard S, Charlton R, Lindsay H, Camm N, Watson C, Abb S. Mattocks C, Taylor GR, Charlton R. (2012), Practice Guidelines for Targeted Next Generation Sequencing Analysis and Interpretation. Clinical Molecular Genetics Society and Association of Clinical Cytogenetics. http://www.acgs.uk.com/media/774807/bpg_for_targeted_next_generation_sequencing_may_2014_final.pdf

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA. (2012). Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS One. 7(11):e47768. doi: 10.1371/journal.pone.0047768

epsilon toxin, https://www.vocabulary.com/dictionary/epsilon%20toxin

Experts Q & A, Public Broadcasting Service, 2006-12-15

Ewing B, Green P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8(3):186-94

Eyre DW, Cule ML, Griffiths D, Crook DW, Peto TEA, Walker AS, Wilson DJ. (2013). Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in Clostridium difficile transmission. PLoS Comput Biol. 9(5): e1003059. doi: 10.1371/journal.pcbi.1003059

Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty E.M., Ip CL, Wilson DJ, Didelot X, O'Connor L, Lay R, Buck D, Kearns AM, Shaw A, Paul J, Wilcox MH, Donnelly PJ, Peto TE, Walker AS, Crook DW. (2012). A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance. BMJ Open. 2(3). pii: e001124. doi: 10.1136/bmjopen-2012-001124

Facts About Botulism, CDC

Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B, Mahajan MC, Jabado OJ, Deikus G, Clark TA, Luong K, Murray IA, Davis BM, Keren-Paz A, Chess A, Roberts RJ, Korlach J, Turner SW, Kumar V, Waldor MK, Schadt EE. (2012). Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nat. Biotechnol. 30(12):1232-9. doi: 10.1038/nbt.2432

Farmer P. (2013). Key statistics - facts and figures about the earthquake, cholera, and development challenges in Haiti. Manhattan: United Nations Office of the Secretary-General's Special Adviser for Community-Based Medicine and Lessons from Haiti, http://www.lessonsfromhaiti.org/relief-and-recovery/key-statistics

FASTQ format From Wikipedia, the free encyclopedia, https://en.wikipedia.org/wiki/FASTQ_format

FastQC A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FastQC Report https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html#M0

Feng Y, Zhang Y, Ying C, Wang D, Du C. (2015). Nanopore-based Fourth-generation DNA Sequencing Technology Genomics, Proteomics & Bioinformatics 13(1): 4-16. doi: 10.1016/j.gpb.2015.01.009.

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat. Methods. 7(6): 461-5. doi: 10.1038/nmeth.1459

Forbes, Illumina Promises To Sequence Human Genome For $100 -- But Not Quite Yet,
https://www.forbes.com/sites/matthewherper/2017/01/09/illumina-promises-to-sequence-human-genome-for-100-but-not-quite-yet/#4526f680386d

Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJ, Brinkman FS, Brunham RC, Tang P. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 364(8):730-9. doi: 10.1056/NEJMoa1003176

Genome Atlantic, Here's the @NIH Cost per Genome graph updated for 2017. Wow.,
https://www.genome.gov/images/content/costpergenome_2017.jpg

Gibson G. and Muse SV. A Primer of Genome Science. 3rd ed. P.84

Giordano F, Aigrain L, Quail MA, Coupland P, Bonfield JK, Davies RM, Tischler G, Jackson DK, Keane TM, Li J, Yue JX, Liti G, Durbin R, Ning Z. (2017). De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. Scientific Reports 7(1): 3935. doi: 10.1038/s41598-017-03996-z

GitHub linneas/condetri, https://github.com/linneas/condetri

Gladstone R (2013) Rights advocates suing U.N. over the spread of cholera in Haiti. The New York Times Americas. New York City: The New York Times Company,
http://www.nytimes.com/2013/10/09/world/americas/rights-advocates-suing-un-over-the-spread-of-cholera-in-haiti.html

Glanders, CDC

Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Larner-Svensson H, Fung R, Godwin H, Knox K, Votintseva A. (2013). Within-host evolution of Staphylococcus aureus during asymptomatic carriage. PLoS ONE. 8(5): e61319. doi: 10.1371/journal.pone.0061319

Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR. (2015). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Research. 25(11): 1750-6. doi:10.1101/gr.191395.115

Gray RR, Tatem A., Johnson JA, Alekseyenko AV, Pybus OG, Suchard MA, Salemi M. (2011). Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant Staphylococcus aureus ST239 genome-wide data within a Bayesian framework. Mol Biol Evol. 28(5): 1593-603. doi: 10.1093/molbev/msq319

Gregory B, Waag D. (1997), Military Medicine: Medical aspects of biological warfare (PDF), Office of the Surgeon General, Department of the Army, Library of Congress 97-22242

Gregory TR. (2005). Synergy between sequence and size in Large-scale genomics". Nature Reviews Genetics. 6 (9): 699–708. doi: 10.1038/nrg1674

Guo J, Xu N, Li Z, Zhang S, Wu J, Kim DH, Sano Marma M, Meng Q, Cao H, Li X, Shi S, Yu L, Kalachikov S, Russo JJ, Turro NJ, Ju J. (2008). Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proc. Natl. Acad. Sci. U.S.A. 105 (27) :9145-50. doi: 10.1073/pnas.0804023105

Gurevich A, Saveliev, Vyahhi N, Tesler G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29 (8): 1072–1075. doi: 10.1093/bioinformatics/btt086

Haiti: The Importance Of Communicating About Cholera, http://www.doctorswithoutborders.org/news-stories/field-news/haiti-importance-communicating-about-cholera

Han K, Li ZF, Peng R, Zhu LP, Zhou T, Wang LG, Li SG. Zhang XB, Hu W, Wu ZH, Qin N, Li YZ. (2013). Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu. Scientific Reports. 3: 2101. doi: 10.1038/srep02101

Harismendy O, Ng P, Strausberg R, Wang X, Stockwell T, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 10(3): R32. doi: 10.1186/gb-2009-10-3-r32

Harris SR, Cartwright EJ, Török ME, Holden MT, Brown NM, Ogilvy-Stuart AL, Ellington MJ, Quail MA, Bentley SD, Parkhill J, Peacock SJ. (2013). Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. Lancet Infect Dis. 13(2):130-6. doi: 10.1016/S1473-3099(12)70268-2

Heger M (2013) McGill University team develops rapid genome sequencing technique for outbreak monitoring. Clinical Sequencing News. New York City: GenomeWeb LLC, http://www.genomeweb.com/sequencing/mcgill-university-team-develops-rapid-genome-sequencing-technique-outbreak-monit

Hendriksen RS1, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM, Bortolaia V, Pearson T, Waters AE, Upadhyay BP, Shrestha SD, Adhikari S, Shakya G, Keim PS, Aarestrup FM. (2011). Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. MBio. 2(4):e00157-11. doi: 10.1128/mBio.00157-11

Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J. (2008). De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. Genome Research 18(5):802-9. doi: 10.1101/gr.072033.107

High-throughput sequencing for the study of bacterial pathogen biology (2014) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4150483/

High-Throughput Sequencing of DNA https://study.com/academy/lesson/high-throughput-sequencing-of-dna.html

Hou Y, Lin S. (2009). Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes: Gene Content Estimation for Dinoflagellate Genomes. PLoS One. 2009; 4(9): e6978. doi: 10.1371/journal.pone.0006978

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 8(7): R143

IHB - The DoD Immunization Information and Training Portal

Illumina, ALL SYSTEMS, https://www.illumina.com/systems.html

Illumina, Introduction to Sequencing Quality Scores https://www.illumina.com/science/education/sequencing-quality-scores.html

Illumina, Sequencing Quality Scores, https://www.illumina.com/science/education/sequencing-quality-scores.html

Illumina, Technical Note: Informatics, Understanding Illumina Quality Scores https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_understanding_quality_scores.pdf

Illumina, Technical Note: Informatics, Quality Scores for Next-Generation Sequencing https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

Inkinen J, Jayaprakash B, Ahonen M, Pitkänen T, Mäkinen R, Pursiainen A, Santo Domingo JW, Salonen H, Elk M, Keinänen-Toivola MM. (2017). Bacterial community changes in copper and PEX drinking water pipeline biofilms under extra disinfection and magnetic water treatment. J Appl Microbiol. 2017 Dec 9. doi: 10.1111/jam.13662

InsideDNA, 22 January 2016, Analysing raw sequencing reads with FASTQC for quality control and filtering https://insidedna.me/tutorials/view/fastqc-quality-control-and-filtering-of

Ishoey T, Woyke T, Stepanauskas R, Novotny M, Lasken RS. (2008). Genomic sequencing of single microbial cells from environmental samples. Current Opinion in Microbiology. 11: 198-204. doi: 10.1016/j.mib.2008.05.006

Ivers LC, Walton DA. (2012). The “first” case of cholera in Haiti: lessons for global health. Am J Trop Med Hyg. 86(1):36-8. doi: 10.4269/ajtmh.2012.11-0435

Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. (2015). Improved data analysis for the MinION nanopore sequencer. Nat. Methods. 12(4): 351-6. doi: 10.1038/nmeth.3290

Jenson D, Szabo V, Duke FHI. (2011). Haiti Humanities Laboratory Student Research Team (2011) Cholera in Haiti and other Caribbean regions, 19th century. Emerg Infect Dis. 17(11): 2130–2135. doi: 10.3201/eid1711.110958

Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN1, Doshi S, Courtot M, Lo R, Williams LE, Frye JG, Elsayegh T, Sardar D, Westman EL, Pawlowski AC, Johnson TA, Brinkman FS2, Wright GD, McArthur AG. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Research. 45(D1): D566-D573. doi: 10.1093/nar/gkw1004

Jorgensen JH, Ferraro MJ. (2009). Antimicrobial susceptibility testing: a review of general principles and contemporary practices. Clin Infect Dis.; 49(11): 1749-55

Joy, Bill (2007-03-31), Why the Future Doesn't Need Us: How 21st Century Technologies Threaten to Make Humans an Endangered Species, Random House, ISBN 978-0-553-52835-0

Katz JM (2010) UN probes base as source of Haiti cholera outbreak. The Washington Times. Washington, DC: The Washington Times, LLC, http://www.washingtontimes.com/news/2010/oct/27/un-probes-base-as-source-of-haiti-cholera-outbreak

k-mer, From Wikipedia, the free encyclopedia, https://en.wikipedia.org/wiki/K-mer

Kingsford C, Schatz MC, Pop M. (2010). Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics 11:21. doi: 10.1186/1471-2105-11-21

Kingsley RA, Kay S, Connor T, Barquist L, Sait L, Holt KE, Sivaraman K, Wileman T, Goulding D, Clare S. (2013). Genome and transcriptome adaptation accompanying emergence of the definitive type 2 host-restricted Salmonella enterica serovar Typhimurium pathovar. MBio. 4 e00565–13. doi: 10.1128/mBio.00565-13

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Adam M Phillippy. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat. Biotechnol. 30 (7): 693-700. doi: 10.1038/nbt.2280

Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research. 27(5): 722-736. doi: 10.1101/gr.215087.116

Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, Farrington M, Holden MTG, Dougan G, Bentley SD, Parkhill J (2012). Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog. 8(8):e1002824. doi: 10.1371/journal.ppat.1002824

Köser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR, Sanders M, Enright MC, Dougan G, Bentley SD, Parkhill J, Fraser LJ, Betley JR, Schulz-Trieglaff OB, Smith GP, Peacock SJ. (2012). Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med. 366: 2267-2275

Lam H, Clark M, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, Butte AJ, Ji HP, Snyder M. (2011). Performance comparison of whole-genome sequencing platforms. Nat Biotechnol. 30(1):78-82. doi: 10.1038/nbt.2065

Land M, Hauser L, Jun Se-Ran, Nookaew I, Leuze MR, Ahn Tae-Hyuk, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW (2015). Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 15(2): 141-161, doi: 10.1007/s10142-015-0433-4

Lantagne D, Nair GB, Lanata CF, Cravioto A (2014) The cholera outbreak in Haiti: where and how did it begin? Curr Top Microbiol Immunol. 379:145-64. doi:10.1007/82_2013_331
Lebreton F, van Schaik W, McGuire AM, Godfrey P, Griggs A, Mazumdar V, Corander J, Cheng L, Saif S, Young S. (2013). Emergence of epidemic multidrug-resistant Enterococcus faecium from animal and commensal strains. MBio.;4 e00534–13. doi: 10.1128/mBio.00534-13

Lee CR, Lee JH, Park M, Park KS, Bae IK, Kim YB, Cha CJ, Jeong BC, Lee SH. (2017). Biology of Acinetobacter baumannii: Pathogenesis, Antibiotic Resistance Mechanisms, and Prospective Treatment Options. Front. Cell. Infect. Microbiol. 7: 55. doi: 10.3389/fcimb.2017.00055

Lee E, Khurana MS, Whiteley AS, Monis PT, Bath A, Gordon C, Ryan UM, Paparini A. (2017). Novel Primer Sets for Next Generation Sequencing-Based Analyses of Water Quality. PLoS One. 2017 Jan 24;12(1):e0170008. doi: 10.1371/journal.pone.0170008

Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. Science. 299(5607):682-686

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25(16):2078-9. doi: 10.1093/bioinformatics/btp352

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Research. 20(2): 265-72. doi: 10.1101/gr.097261.109

Lieberman TD, Michel JB, Aingaran M, Potter-Bynoe G, Roux D, Davis MR, Skurnik D, Leiby N, LiPuma JJ, Goldberg JB. (2011). Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet. 43(12): 1275–1280. doi: 10.1038/ng.997

Lim E, Pon A, Djoumbou Y, Knox C, Shrivastava S, Guo AC, Neveu V, Wishart DS. (2010). T3DB: a comprehensively annotated database of common toxins and their targets. Nucleic Acids Research. 38(Database issue): D781-6. 19897546

Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, Pevzner P. (2016). Assembly of Long Error-Prone Reads Using de Bruijn Graphs. PNAS. E8396–E8405, doi: 10.1073/pnas.1604560113

Linux, Από τη Βικιπαίδεια, την ελεύθερη εγκυκλοπαίδεια, https://el.wikipedia.org/wiki/Linux

Lipman DJ, Pearson WR. (1985). Rapid and sensitive protein similarity searches. Science. 227 (4693): 1435-41. doi: 10.1126/science.2983426

Liu J, Ren H, Ye X, Wang W, Liu Y, Lou L, Cheng D, He X, Zhou X, Qiu S, Fu L, Hu B. (2017). Bacterial community radial-spatial distribution in biofilms along pipe wall in chlorinated drinking water distribution system of East China. Appl Microbiol Biotechnol. 2017 Jan;101(2):749-759. doi: 10.1007/s00253-016-7887-8

Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, Mccalmon S, Hagerman RJ, Tassone F, Hagerman PJ. (2013). Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene Sequencing the unsequenceable : Expanded CGG-repeat alleles of the fragile X gene. 23(1):121-8. doi: 10.1101/gr.141705.112

Lowder BV, Guinane CM, Ben Zakour NL, Weinert LA, Conway-Morris A, Cartwright RA, Simpson AJ, Rambaut A, Nübel U, Fitzgerald JR. (2009). Recent human-to-poultry host jump, adaptation, and pandemic spread of Staphylococcus aureus. Proc Natl Acad Sci U S A. 106(46): 19545-50. doi: 10.1073/pnas.0909285106

Maharjan L. (2010). Cholera outbreak looms over capital. The Himalayan Times. Kathmandu: International Media Network Nepal Pvt. Ltd., http://www.thehimalayantimes.com/fullNews.php?headline=Cholera+outbreak+looms+over+capital&NewsID=258974

Mardis ER. (2008). Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics. 9 (1): 387-402. doi:10.1146/annurev.genom.9.081307.164359

McAdam PR, Richardson EJ, Fitzgerald JR. (2014). High-throughput sequencing for the study of bacterial pathogen biology Curr Opin Microbiol. 19(100): 106-113. doi: 10.1016/j.mib.2014.06.002

McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJ, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD. (2013). The Comprehensive Antibiotic Resistance Database. Antimicrobial Agents and Chemotherapy. 57(7): 3348-57. doi:10.1128/AAC.00419-13

McArthur AG, Wright GD. (2015). Bioinformatics of antimicrobial resistance in the age of molecular epidemiology. Current Opinion in Microbiology. 27: 45-50. doi: 10.1016/j.mib.2015.07.004

McCutcheon JP, von Dohlen CD. (2011). An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs. Current Biology. 21 (16): 1366–1372. doi:10.1016/j.cub.2011.06.051

Medvedev P, Scott E, Kakaradov B, Pevzner P. (2011). Error correction of high-throughput sequencing datasets with non-uniform coverage. Bioinformatics. 27 (13): i137–141. doi:10.1093/bioinformatics/btr208

Melioidosis, CDC

Miller JR, Koren S, Sutton G. (2010). Assembly algorithms for next-generation sequencing data. Genomics. 95 (6): 315-27. doi:10.1016/j.ygeno.2010.03.001

Ministère de la Santé Publique et de la Population (2014) Rapport journalier MSPP du 07 Janvier 2014. Port-au-Prince: Ministère de la Santé Publique et de la Population, http://mspp.gouv.ht/site/downloads/Rapport%20Web_07.01_Avec_Courbes_Departementales.pdf

Mira A, Ochman H, Moran NA. (2001). Deletional bias and the evolution of bacterial genomes. Trends in Genetics. 17(10): 589-96.

Metzker ML. (2010). Sequencing technologies - the next generation (PDF). Nat Rev Genet. 11 (1): 31–46. doi:10.1038/nrg2626.

Multiple displacement amplification, From Wikipedia, the free encyclopedia https://en.wikipedia.org/wiki/Multiple_displacement_amplification

Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, Niyogi SK, Kim EJ, Ramamurthy T, Chun J, Wood JL, Clemens JD, Czerkinsky C, Nair GB, Holmgren J, Parkhill J, Dougan G. (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477(7365): 462-5. doi:10.1038/nature10392

Myers EW. (1995). Toward simplifying and accurately formulating fragment assembly.. J Comput Biol.; 2(2): 275-90

Next-Generation Sequencing and Assembly of Bacterial Genomes https://www.researchgate.net/publication/236610137_Next-Generation_Sequencing_and_Assembly_of_Bacterial_Genomes

Nigam PK, Nigam A. (2010). BOTULINUM TOXIN. Indian Journal of Dermatology. 55: 8–14. doi:10.4103/0019-5154.60343

Novak Matt (2016-11-03). "The Largest Bioterrorism Attack In US History Was An Attempt To Swing An Election". Gizmodo

N50, L50, and related statistics, From Wikipedia, the free encyclopedia https://en.wikipedia.org/wiki/N50,_L50,_and_related_statistics

Ochman H. (2005). Genomes on the shrink. Proceedings of the National Academy of Sciences. 102 (34): 11959-11960. doi:10.1073/pnas.0505863102.

Oxford Nanopore Technologies, A nanopore is a very small hole, https://nanoporetech.com/how-it-works

Oxford NANOPORE Technologies, Products, https://nanoporetech.com/index.php/products

Parker AA. (2010). Cholera in Haiti - the climate connection. Circle of Blue. Traverse City: Circle of Blue, http://www.circleofblue.org/waternews/2010/world/hold-cholera-in-haiti-the-climate-connection

Past U.S. Incidents of Food Bioterrorism. In Bioterrorism: A Threat to National Security or Public Health Defining Issue, University of Wisconsin–Madison and the Wisconsin State Laboratory of Hygiene, MM&I 554, September 30, 2008

Peng Y, Leung HCM, Yiu SM, Chin FYL. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 28: 1–8. doi: 10.1093/bioinformatics/bts174

Pereira RP, Peplies J, Brettar I, Höfle MG. (2017). Development of a genus-specific next generation sequencing approach for sensitive and quantitative determination of the Legionella microbiome in freshwater systems. BMC Microbiol. 2017 Mar 31;17(1):79. doi: 10.1186/s12866-017-0987-5

Pevzner PA, Tang H. (2001). Fragment assembly with double-barreled data. Bioinformatics 17 Suppl 1:S225-33

Pevzner PA, Tang H, Waterman MS. (2001). An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences of the United States of America. 98(17): 9748-53

Piarroux R, Barrais R, Faucher B, Haus R, Piarroux M, Gaudart J, Magloire R, Raoult D. (2011) Understanding the cholera epidemic, Haiti. Emerg Infect Dis 17(7):1161-8. doi: 10.3201/eid1707.110059

Plague - Frequently Asked Questions (FAQ) About Plague, CDC

Plague Home Page - Division of Vector-Borne Infectious Diseases (DVBID), CDC

Plague Information - Emergency Preparedness & Response, CDC

Poirel L, Jayol A, Bontron S, Villegas MV, Ozdamar M, Türkoglu S, Nordmann P. (2015). The mgrB gene as a key target for acquired resistance to colistin in Klebsiella pneumoniae. J Antimicrob Chemother. 70 (1): 75-80.

Preston Richard (2002). The Demon in the Freezer, Ballantine Books, New York. ISBN 9780345466631

Price LB, Stegger M, Hasman H, Aziz M, Larsen J, Andersen PS, Pearson T, Waters AE, Foster JT, Schupp J, Gillece J, Driebe E, Liu CM, Springer B, Zdovc I, Battisti A, Franco A, Zmudzki J, Schwarz S, Butaye P, Jouy E, Pomba C, Porrero MC, Ruimy R, Smith TC, Robinson DA, Weese JS, Arriola CS, Yu F, Laurent F, Keim P, Skov R, Aarestrup FM. (2012). Staphylococcus aureus CC398: host adaptation and emergence of methicillin resistance in livestock. MBio. 2012; 3(1). pii: e00305-11. doi: 10.1128/mBio.00305-11

PRODUCTS + SERVICES PACBIO SYSTEMS,
http://www.pacb.com/products-and-services/pacbio-systems/

Q fever, Centers for Disease Control and Prevention, https://www.cdc.gov/qfever/

Q Fever - Emergency Preparedness and Response, CDC

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. (2008). A large genome center's improvements to the Illumina sequencing system. Nat Methods. 5(12):1005-10. doi: 10.1038/nmeth.1270

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 13: 341. doi: 10.1186/1471-2164-13-341

Quick J, Quinlan AR, Loman NJ. (2014). Open Access A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. Gigascience.; 3: 22. doi: 10.1186/2047-217X-3-22

Ray DE. (2002). Bioterrorism – the high economic costs of an attack, Policy Pennings, http://www.agpolicy.org/weekpdf/116.pdf, October 2002

Reimer AR, Van Domselaar G, Stroika S, Walker M, Kent H, et al. (2011). Comparative genomics of Vibrio cholerae from Haiti, Asia, and Africa. Emerg Infect Dis. 17(11): 2113–2121

Reuter JA, Spacek D, Snyder MP. (2015). High-Throughput Sequencing Technologies. Molecular Cell 21; 58(4): 586-597. doi: 10.1016/j.molcel.2015.05.004

Reuter S, Harrison TG, Köser CU, Ellington MJ, Smith GP, Parkhill J, Peacock SJ, Bentley SD, Török ME. (2013). A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak. BMJ Open. 3(1). pii: e002175. doi: 10.1136/bmjopen-2012-002175

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. (2011). Integrative Genomics Viewer. Nature Biotechnology 29(1): 24–26. doi: 10.1038/nbt.1754

Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, Chisholm SW. (2009). Whole genome amplification and de novo assembly of single bacterial cells. PLoS ONE. 4: e6864. doi:10.1371/journal.pone.0006864

Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, Supply P, Kalinowski J, Niemann S: Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. PLoS Med 2013, 10:e1001387

Rouquette C, de Chastellier C, Nair S, Berche P. (1998). The ClpC ATPase of Listeria monocytogenes is a general stress protein required for virulence and promoting early bacterial escape from the phagosome of macrophages. Molecular Microbiology. 27(6): 1235-45.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. (2013). Characterizing and measuring bias in sequence data. Genome Biol. 14(5): R51. doi: 10.1186/gb-2013-14-5-r51

Saha S, Raghava GPS. (2007). BTXpred: prediction of bacterial toxins. In Silico Biology. 7(4-5): 405-12

Saha S, Raghava GPS. (2006). VICMpred: SVM-based method for the prediction of functional proteins of gram-negative bacteria using amino acid patterns and composition. Genomics Proteomics & Bioinformatics. 4(1): 42–47. doi: 10.1016/S1672-0229(06)60015-6

Salmela L, Rivals E. (2014). LoRDEC: accurate and efficient long read error correction. Bioinformatics. 30(24):3506-14. doi: 10.1093/bioinformatics/btu538

Schadt EE, Turner S, Kasarskis A. (2010). A window into third-generation sequencing. Hum. Mol. Genet. 19(R2): R227-40. doi: 10.1093/hmg/ddq416

Schloss JA. (2008). How to get genomes at one ten-thousandth the cost. Nature Biotechnology, 26 (10): 1113-5. doi: 10.1038/nbt1008-1113

Sequence alignment, From Wikipedia, the free encyclopedia https://en.wikipedia.org/wiki/Sequence_alignment

Sequence Read Archive, From Wikipedia, the free encyclopedia https://en.wikipedia.org/wiki/Sequence_Read_Archive

Sequencing Coverage, https://www.illumina.com/science/education/sequencing-coverage.html

Shendure J, Ji H. (2008). Next-generation DNA sequencing. Nat Biotechnol. 26(10): 1135-45. doi: 10.1038/nbt1486

Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MCJ, Parkhill J, Falush D. (2013). Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci U S A. 110(29): 11923-7. doi: 10.1073/pnas.1305559110

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics. 15 (2): 121–132. doi: 10.1038/nrg3642

Sjödin A, Broman T, Melefors Ö, Andersson G, Rasmusson B, Knutsson R, Forsman M. (2013). The need for high-quality whole-genome sequence databases in microbial forensics. Biosecurity and Bioterrorism 11 Suppl 1:S78-86. doi: 10.1089/bsp.2013.0007

Smallpox - What CDC Is Doing to Protect the Public From Smallpox, CDC

Smallpox Home, CDC

Smith MG, Gianoulis TA, Pukatzki S, Mekalanos JJ, Ornston LN, Gerstein M, Snyder M. (2007). New insights into Acinetobacter baumannii pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. Genes Dev. 21(5): 601-14

Snyder M, Du J, Gerstein M. (2010). Personal genome sequencing: current approaches and challenges. Genes Dev. 24(5): 423–431. doi:10.1101/gad.1864110

SPAdes 3.11.1 Manual,
http://spades.bioinf.spbau.ru/release3.11.1/manual.html

Spoor LE, McAdam PR, Weinert LA, Rambaut A, Hasman H, Aarestrup FM, Kearns AM, Larsen AR, Skov RL, Fitzgerald JR. (2013). Livestock origin for a human pandemic clone of community-associated methicillin-resistant Staphylococcus aureus. MBio. 4(4). pii: e00356-13. doi: 10.1128/mBio.00356-13

Staden, R (1979). A strategy of DNA sequencing employing computer programs. Nucleic Acids Research. 6 (7): 2601–10. doi: 10.1093/nar/6.7.2601

Tan BoonFei, Ng Charmaine, Nshimyimana Jean Pierre, Loh Lay Leng, Gin Karina Y-H, and Thompson Janelle R. (2015). Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. Frontiers in Microbiology. 6: 1027. doi: 10.3389/fmicb.2015.01027

Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, Harel I, Bustamante C, Rasmussen M, Snyder M. (2015). Comprehensive transcriptome analysis using synthetic long read sequencing reveals molecular co-association of distant splicing events. Nat. Biotechnol. 33(7): 736-42. doi: 10.1038/nbt.3242

Thorvaldsdóttir H, Robinson JT, Mesirov JP. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14(2): 178–192. doi: 10.1093/bib/bbs017

Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. 38 (15): e159. doi: 10.1093/nar/gkq543.

Tularemia - Emergency Preparedness & Response, CDC

Tularemia - Key Facts About Tularemia, CDC

Uemura S, Aitken CE, Korlach J, Flusberg Ba, Turner SW, Puglisi JD. (2010). Real-time tRNA transit on single translating ribosomes at codon resolution. Nature. 464(7291):1012-7. doi: 10.1038/nature08925

United States Geological Survey (2010) Magnitude 7.0 - Haiti region. USGS Earthquake Hazards Program. Reston: United States Geological Survey, http://earthquake.usgs.gov/earthquakes/eqinthenews/2010/us2010rja6/us2010rja6.php

Using Next-Generation Sequencing in Infectious Disease Diagnosis, http://macrogenlab.com/2016/09/02/using-next-generation-sequencing-in-infectious-disease-diagnosis/

Van Leuven JT, Meister RC, Simon C, McCutcheon JP. (2014). Sympatric speciation in a bacterial endosymbiont results in two genomes with the functionality of one. Cell. 158 (6): 1270–80. doi:10.1016/j.cell.2014.07.047

Venter JC. (2006). Shotgunning the Human Genome: A Personal View. Encyclopedia of Life Sciences

Vietri NJ, Purcell BK, Tobery SA, Rasmussen SL, Leffel EK, Twenhafel NA, Ivins BE, Kellogg MD, Webster WM, Wright ME, Friedlander AM. (2009). A Short Course of Antibiotic Treatment Is Effective in Preventing Death from Experimental Inhalational Anthrax after Discontinuing Antibiotics. The Journal of Infectious Diseases. Oxford University Press. 199 (3): 336–41. doi:10.1086/596063

Viral Hemorrhagic Fevers - Emergency Preparedness & Response, CDC

Voelkerding KV, Dames SA, Durtschi JD. (2009). Next Generation Sequencing: From Basic Research to Diagnostics. Clinical Chemistry. 55 (4): 641-58. doi: 10.1373/clinchem.2008.112789

Waldor MK, Mekalanos JJ (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin. Science 272(5270): 1910-4

Wang Y, Yang Q, Wang Z. The evolution of nanopore sequencing. Front. Genet. 2015; 5:449. doi: 10.3389/fgene.2014.00449

WebMD, Cholera, https://www.webmd.com/a-to-z-guides/cholera-11156

Wetterstrand KS. DNA sequencing costs: data from the NHGRI Large-Scale Genome Sequencing Program. 2013. https://www.genome.gov/sequencingcosts/

What is FastQC
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/1%20Introduction/1.1%20What%20is%20FastQC.html

Wheeler DL et al. (2008). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 36(Database issue): D13–D21. doi: 10.1093/nar/gkm1000

Why has melioidosis become a current issue?, CDC

Wishart D, Arndt D, Pon A, Sajed T, Guo AC, Djoumbou Y, Knox C, Wilson M, Liang Y, Grant J, Liu Y, Goldansaz SA, Rappaport SM. (2015). T3DB: the toxic exposome database. Nucleic Acids Research. 43 (Database issue): D928-34. 25378312

Wuthiekanun V, Peacock SJ (2006). Management of melioidosis. Expert review of anti-infective therapy. 4 (3): 445–55. doi: 10.1586/14787210.4.3.445.

Yang J, Chen LH, Sun LL, Yu J, Jin Q. (2008). VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic Acids Res. 36 (Database issue): D539-D542

Ye C, Hill CM, Wu S, Ruan J, Ma ZS. (2016). DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. Sci. Rep. 6, 31900; doi: 10.1038/srep31900

Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, Miller RR, Godwin H, Knox K, Everitt RG (2012). Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. Proc Natl Acad Sci U S A. 109(12):4550-5. doi: 10.1073/pnas.1113219109

Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. (2012). Identification of acquired antimicrobial resistance genes. Journal of Antimicrobial Chemotherapy. 67(11):2640-4. doi: 10.1093/jac/dks261

Zerbino D, Birney E (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research. 18(5): 821-9. doi: 10.1101/gr.074492.107

Zook JM, Salit M. (2011). Genomes in a bottle: creating standard reference materials for genomic variation - why, what and how? Genome Biol. 12: P31. https://doi.org/10.1186/gb-2011-12-s1-p31

Βρουκέλλωση, Από τη Βικιπαίδεια, την ελεύθερη εγκυκλοπαίδεια
https://el.wikipedia.org/wiki/%CE%92%CF%81%CE%BF%CF%85%CE%BA%CE%AD%CE%BB%CE%BB%CF%89%CF%83%CE%B7

Μαύρη πανώλη, Από τη Βικιπαίδεια, την ελεύθερη εγκυκλοπαίδεια https://el.wikipedia.org/wiki/%CE%9C%CE%B1%CF%8D%CF%81%CE%B7_%CF%80%CE%B1%CE%BD%CF%8E%CE%BB%CE%B7

Τουλαραιμία, Από τη Βικιπαίδεια, την ελεύθερη εγκυκλοπαίδεια https://el.wikipedia.org/wiki/%CE%A4%CE%BF%CF%85%CE%BB%CE%B1%CF%81%CE%B1%CE%B9%CE%BC%CE%AF%CE%B1

Τύφος, Από τη Βικιπαίδεια, την ελεύθερη εγκυκλοπαίδεια, https://el.wikipedia.org/wiki/%CE%A4%CF%8D%CF%86%CE%BF%CF%82

Ψιττάκωση, https://blog.doctoranytime.gr/glossary/psittakwsi/