ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Αξιολόγηση αλγορίθμων επεξεργασίας πειραματικών δεδομένων αλληλούχισης νουκλεικών οξέων μεγάλου πλαισίου ανάγνωσης και συγκριτική ανάλυση μεταβλητών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τσαμπανάκης Στέφανος

ΕΠΙΒΛΕΠΩΝ: Δρ. Αλέξανδρος Γεωργακίλας, Καθ. Ε.Μ.Π.

Αθήνα, 2021

Περίληψη

Η ανάλυση της αλληλουχιας ενός γονιδιώματος αποτελεί ένα πολύ σημαντικό κομμάτι στον κλάδο της βιολογίας αφού προσφέρει μια πληθώρα γενετικών πληροφοριών όπου επιστήμονες και επαγγελματίες στον χώρο της Υγείας έχουν την ικανότητα να την χρησιμοποιήσουν σαν διαγνωστικό μέσο νοσημάτων , μέσο πρόληψης ασθενειών καθώς και κατανόησης διάφορων μεταλλάξεων. Για την ακριβή αναπαράσταση γονιδιωμάτων πολλά ερευνητικά ενδιαφέροντα και επιστημονικοί κλάδοι έχουν δημιουργηθεί. Πιο συγκερκιμένα έχει δημιουργηθεί ο κλάδος της βιοπληροφορικής καθώς και των τεχνολογιών αλληλούχισης που τα τελευταία χρόνια βελτιώνονται συνεχώς προσφέροντας όλα και καλύτερες αναπαραστάσεις με μικρότερα σφάλματα σε μικρότερο χρόνο.

Στην παρούσα εργασία έγινε μια βιοπληροφορική ανάλυση του DNA. Πραγματοποιήθηκε δηλαδή ένας ποιοτικός έλεγχος στον οποίο αναλύθηκαν διάφοροι παράγοντες που μπορούν να συμβάλουν στην ποιότητα των παραγόμενων μεγάλων πλαισίων ανάγνωσης (long reads). Πιο συγκεκριμένα οι παράγοντες αυτοί είναι οι διαφορές ανάμεσα στο αν το DNA είναι ανθρώπινο ή όχι, ανάμεσα στον τρόπο που έγινε η προετοιμασία βιβλιοθήκης και τέλος οι διαφορές ανάμεσα σε δύο μοντέλα του προγράμματος GUPPY στο οποίο έγινε η διαδικασία του basecall. Για την εκτίμηση της ποιότητας των reads χρησιμοποιήθηκαν δυο διαφορετικά προγράμματα το FastQC και το NanoPack μέσω των οποίων υπολογίστηκαν οι τιμές Qscore, N50 και τα μήκοι των reads Αρχικά χρησιμοποιήθηκε DNA απο ανθρώπινο επιθηλιακό ιστό (RPE) και DNA από τις ωοθήκες ενός κινέζικου χάμστερ (CHO Chinese Hamster Ovarian). Το CHO DNA προετοιμάστηκε με δύο διαφορετικούς τρόπους προετοιμασίας βιβλιοθήκης την ‘rapid’ μεθοδο και ‘ligation’ και με βάση τα αποτελέσματα τους έγινε εκτίμηση της βέλτιστης μεθόδου προετοιμασίας βιβλιοθήκης . Το επιθηλιακό DNA προετοιμάστηκε μέσω της ligation μεθόδου, όμως η διαδικασία basecall έγινε δυο φορές. Η πρώτη έγινε με το μοντέλο FAST του GUPPY και η δεύτερη μέσω του High Accuracy (HAC) μοντέλου και μέσω των αποτελεσμάτων τους έγινε μια εκτίμηση . Τέλος ανάμεσα στα αποτελέσματα από το επιθηλιακό και το CHO DNA πραγματοποιήθηκε η τελική εκτίμηση ανάμεσα σε ποιο πρότυπο DNA παρουσιάστηκε η μεγαλύτερη ποιότητα στα μεγάλα πλαίσια ανάγνωσης .

Abstract

The anlysis of the DNA sequence of a genome plays an important role in the field of biology since it offers a plethora of genetic information where scientists and medical professionals have the ability to use these informations as disease diagnostic tool, disease prevention tool as well as the understanding of various mutations. For the creation of an accurate representation of a genome various scientific interests and fields have been established. More specifically the filed of bioinformatics has been established as well as various DNA sequencing technologies where the past years they are constantly optimized offering better representations with fewer errors at faster speeds

In this paper we will conduct a bioinformatic analysis of DNA. We will perform a quality check where we will analyze various factors where they can contribute to the quality of the the DNA reads. These factors consists the difference in the quality results whether the DNA was harvested from a human source or from a different organism, the difference in the method where the library preparation was achieved and the difference in the two different basecalling models offered by the program GUPPY. For the assessment of the read quality we will use two different programs FastQC and NanoPack where through these we will calculate the Qscore/N50 values as well as the read length. We will use DNA from Human Retinal Epithelial cell (RPE) and DNA from a ovarian cell originated in a Chinese hamster (CHO). The CHO DNA was prepared using two different library preparations the ligation method and the rapid method and based on the different results which they produced the optimal method will be evaluated. The RPE was prepared using the ligation method but the basecalling was done twice using the two models of GUPPY the high accuracy model(HAC) and FAST model and their results where then evaluated . In the end the results from from CHO and RPE DNA where compared to see which DNA produced the highest quality.

Περιεχόμενα

Σκοπός

Η παρούσα διπλωματική εργασία πραγματεύεται θέματα από το επιστημονικό πεδίο της Βιοπληροφορικής. Πιο συγκεκριμένα αντικείμενο της εργασίας αυτής είναι η βιοπληροφορική ανάλυση γονιδιωμάτων εκτελώντας μια διαδικασία ποιοτικού ελέγχου (quality check) σε έναν αριθμό από κατασκευασμένες αλληλουχίες DNA (reads) . Θα μελετήσουμε διάφορους παράγοντες που μπορουν να επηρεάσουν την ποιότητα αυτών των reads. Αυτοί οι παράγοντες αποτελούν διαφορές στην ποιότητα των αλληλουχιών ,ανάλογα με τον οργανισμό που έγινε η εξαγωγή του πρότυπου DNA , ανάμεσα στην διαδικασία όπου έγινε η προετοιμασία βιβλιοθήκης και μεταξύ FAST και HAC μοντέλου του προγράμματος GUPPY όπου έγινε η διαδικασία basecall.

Εισαγωγή

1.1 Βιοπληροφορική

Η ραγδαία ανάπτυξη της επιστήμης της πληροφορικής έκανε φανερό στην επιστημονική κοινότητα ότι οι δυνατότητες που προσφέρει θα μπορούσαν να χρησιμοποιηθούν αποτελεσματικά από άλλες επιστήμες όπως η Βιολογία, η Ιατρική, η Βιοτεχνολογία. Πιο συγκεκριμένα, τα τελευταία επιτεύγματα της μοριακής βιολογίας, με την ανάλυση και λεπτομερή χαρτογράφηση όλο και περισσοτέρων γονιδιωμάτων, οδήγησαν στην συσσώρευση πλήθους βιολογικών δεδομένων. Κρίθηκε, λοιπόν, απαραίτητη η εύρεση τρόπου αποτελεσματικής και αποδοτικής διαχείρισης όλων αυτών των δεδομένων. Για τον σκοπό αυτό αναπτύχθηκε η επιστημονική περιοχή της Βιοπληροφορικής (Bioinformatics) που αποτελεί ένα κράμα της επιστήμης της βιολογίας και των υπολογιστών.

Μια από τις μεγαλύτερες προκλήσεις που αντιμετωπίζει σήμερα ο κλάδος της βιολογίας είναι η αποδοτικότερη εκμετάλλευση της γνώσης που υπάρχει μέσα στην πληθώρα των βιολογικών δεδομένων που προκύπτουν από την ανάλυση των γονιδιωμάτων διαφόρων οργανισμών. Η ανάγκη,λοιπόν, για πλήρη κατανόηση της γνώσης αυτής καθορίζει εν μέρει και τους στόχους της βιοπληροφορικής:

- Οργάνωση των δεδομένων με κατάλληλο τρόπο που θα κάνει τη διαχείρισή τους ευκολότερη για τους ερευνητές.
- Ανάπτυξη εργαλείων για ανάλυση των δεδομένων καθώς και ερμηνεία των αποτελεσμάτων.
- Ανάπτυξη αποδοτικών αλγορίθμων για μέτρηση ομοιότητας μεταξύ ακολουθιών και αποτίμηση σχέσεων μεταξύ τεραστίων συνόλων δεδομένων.
- Επέκταση των πειραματικών δεδομένων μέσω προβλέψεων(π.χ. πρόβλεψη δομής πρωτεϊνών).
- Εργαλεία για συνδυασμό δεδομένων που θα χρησιμοποιηθούν για κατανόηση φυσικών φαινομένων, εξελικτικών σχέσεων μεταξύ οργανισμών και γενετική προέλευση ασθενειών.

Όπως γίνεται εύκολα κατανοητό οι στόχοι της βιοπληροφορικής, όντας ένας κλάδος που γεννήθηκε από την απαίτηση χρήσης μεθόδων πληροφορικής στην επιστήμη της βιολογίας, είναι άμεσα συνυφασμένοι με τον σημαντικότερο στόχο της βιολογίας που είναι η κατά το δυνατόν πληρέστερη κατανόηση της δομής και λειτουργίας των οργανισμών.

1.2 DNA

1.2.1 Νουκλεοτίδια και Πολυνουκλεοτίδια

Έως και τις αρχές του 20ουαιώνα οι επιστήμονες έκαναν υποθέσεις σχετικά με ένα συγκεκριμένο μόριο, το οποίο δρούσε ως η χημική βάση της κληρονομικότητας, χωρίς όμως να γνωρίζει κανείς ποιο ήταν αυτό. Περίπου το 1950, οι ειδικοί αναγνώρισαν το DNA(δεσοξυρυβονουκλεϊκό οξύ) ως το κληρονομικό μόριο και προσδιόρισαν την κατασκευή του. Το DNAείναι ένα οξύ που αποτελείται από νουκλεοτίδια (θυμίνη, αδενίνη, γουανίνη, κυτοσίνη).Η δομή του DNAείναι η εξής: 2 κλώνοι που φέρουν πάνω τους νουκλεοτίδια, τα οποία τηρούν κανόνες συμπληρωματικότητας(γουανίνη –κυτοσίνη, θυμίνη -αδενίνη περιστρέφονται ενωμένοι με δεσμούς υδρογόνου σε ελικοειδή μορφή(Εικόνα 2.1).Κατά την διάρκεια του σταδίου S–Phasetου κυτταρικού κύκλου το DNAείναι εύκολα ορατό διότι μέσω της διαδικασίας της αντιγραφής συμπυκνώνεται σε μορφή χρωμοσωμάτων, ενώ παρατηρώντας κανείς το κύτταρο σε ουδέτερη φάση, διαπιστώνει ότι το γενετικό υλικό βρίσκεται σε μορφή χαλαρών ινών ακανόνιστου σχήματος. Τα χρωμοσώματα είναι ‘συσκευασμένα’τμήματα DNA, το οποίο περιστρέφεται γύρω απόμία ομάδα πρωτεϊνών τις ιστώνες σαν ένα σχοινί γύρω από χάντρες

Κάθε μόριο DNAσυντίθεται από μικρές δομικές μονάδες, τα νουκλεοτίδια. Ο βασικός άξωνας ενός νουκλεοτιδίου αρθρώνεται από μία πεντόζη (δεοξυριβόζη), ένα σάκχαρο με 5 άτομα άνθρακα, ένα μόριο φωσφορικού οξέος και μια οργανική αζωτούχα βάση (πουρίνη: αδενίνη, γουανίνη ή πυριμιδίνη: θυμίνη, κυτοσίνη)(Εικόνα 2.2). Η συμπληρωματικότητα που ακολουθούν οι βάσεις για την ένωση των δύο κλώνων είναι η εξής: η αδενίνη ενώνεται με θυμίνη ενώ η γουανίνη με κυτοσίνη και τα ζεύγη βάσεων που δημιουργούνται συγκρατούνται με δεσμούς υδρογόνου. Τα νουκλεοτίδια αποτελούν τις γενετικές πληροφορίες όταν αυτά εκφράζονται. Κάθε κλώνος φέρει πάνω του μεμονωμένα νουκλεοτίδια, τα οποία παραταγμένα στην σειρά συνιστούν ένα πολυνουκλεοτίδιο

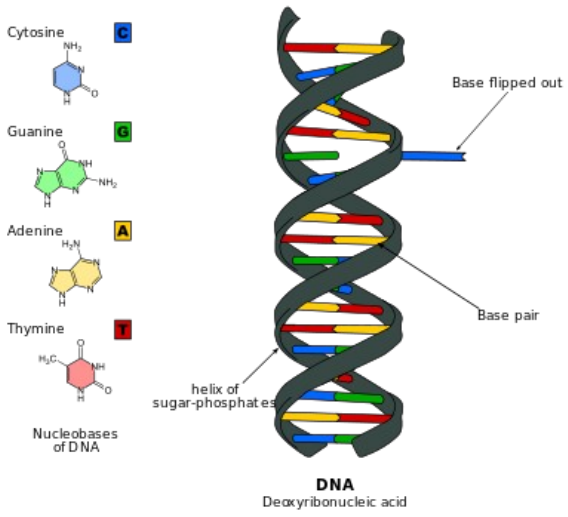


Figure 1: πηγή file:///home/course_user/Desktop/Dna-base-flipping.svg.png Δομή του DNA

1.2.2 Αντιγραφή

Για να επιτευχθεί η ανάπτυξη του οργανισμού και κατ’επέκταση η συνέχιση του είδους και η κληρονομικότητα βασικό στοιχείο είναι η μεταβίβαση του DNAαπό κύτταρο σε κύτταρο και από γενιά σε γενιά. Προϋπόθεση αυτής της διαδικασίας είναι η αντιγραφή του DNA.

Η κατασκευή του DNA θέτει σε εφαρμογή αυτή τη λειτουργία επειδή κάθε κλώνος δύναται να αποτελέσει εκμαγείο ώστε να προσδιορίσει την δημιουργία ενός άλλου κλώνου. Στα πρώτα στάδια αντιγραφής του DNA,η πολυνουκλεοτιδική αλυσίδα διχάζεται, οι κλώνοι του αρχικού μορίου DNA απομακρύνονται ο ένας από τον άλλον ώστε κάθε κλώνος μεμονωμένα να αποτελέσει πρότυπο για την δημιουργία νέας αλυσίδας χάριν στην συμπληρωματικότητα των βάσεων που φέρει. Έτσι παράγονται δύο νέα μόρια DNAπου το κάθε ένα περιέχει ένα νεοσυντιθέμενο κλώνο και έναν κλώνο του αρχικού μορίου, γι’αυτό και ονομάζεται ημισυντηρητικήη διαδικασία της αντιγραφής του.

Η διαδικασία της Αντιγραφής του DNAπεριλαμβάνει 3 στάδια:

Αρχικά ένα ένζυμο, η ελικάση, διαχωρίζει την έλικα του DNA, απομακρύνοντας τους κλώνους τον έναν από τον άλλον. Κάθε ένας από τους δύο διαχωρισμένους κλώνους δέχεται ένα ένζυμο, την DNAπολυμεράση, η οποία δημιουργεί ένα νέο μόριο DNAπου είναι συμπληρωματικό προς κάθε κλώνο. Προσθέτει δηλαδή συμπληρωματικά νουκλεοτίδια σε κάθε κλώνο με σκοπό την δημιουργία δύο νέων μορίων DNA. Την διαδικασία ολοκληρώνει το ένζυμο λιγάση, η οποία ενώνει τα προϊόντα αντιγραφής και τα επιμέρους θραύσματα

1.2.3 Μεταγραφή

Είναι το πρώτο στάδιο της έκφρασης των γονιδίων και αναλύει τον τρόπο με τον οποίο το DNAμετατρέπεται σε RNA. Αυτό επιτυγχάνεται με την συμβολή μίας αλυσίδας του DNA ως πρότυπου, υπό την βοήθεια της RNAπολυμεράσης, με τη διαφοροποίηση ότι η βάση ουρακίλη αντικαθιστά την θυμίνη για την δημιουργία ενός νέου μορίου RNA

Όλη αυτή η διαδικασία επιτελείται σε πυρηνικό επίπεδο και ο σκοπός είναι η γενετική πληροφορία να φτάσει στα ριβοσώματα για την έναρξη της πρωτεϊνοσύνθεσης. Η μεταγραφή τελειώνει όταν η RNAπολυμεράση συναντήσειμία ειδική αλληλουχία λήξης. Στη συνέχεια το RNAθα υποβληθεί σε ακόμη μία διαδικασία, προτού αφήσει τον πυρήνα, κατά την οποία οι μη κωδικοποιούσες περιοχές του (εσόνια) θα πρέπει να απομακρυνθούν. Οι περιοχές που θα κωδικοποιήσουν αμινοξέα (εξόνια) θα συρραφούν μεταξύ τους αφού απομακρυνθούν τα εσόνια

1.2.4 Μετάφραση

Ύστερα από την επεξεργασία, το τελειοποιημένο μόριο RNAονομάζεται αγγελιοφόρο RNAκαι αυτό διότι εγκαταλείπει τον πυρήνα και φέρει μαζί του τις γενετικές πληροφορίες για την πρωτεϊνοσύνθεση. Η μετάφραση πραγματοποιείται στο κυτταρόπλασμα, σε κάποιες δομικές μονάδες που ονομάζονται ριβοσώματα. Η διαδικασία αρχίζει όταν ένα μόριο mRNAσυνδέεται σε μία μικρή ριβοσωμική υπομονάδα. Στη συνέχεια ένα μόριο μεταφορικού (tRNA) συνδέεται με το κωδικόνιο έναρξης του mRNAκαι ονομάζεται μεταφορικό διότι φέρει ένα αμινοξύ. Έπειτα μια μεγάλη ριβοσωμική υπομονάδα συνδέεται με την μικρή δημιουργώντας ένα ολοκληρωμένο ριβόσωμα. Το πρώτο tRNAπου θα φτάσει στο ριβόσωμα φέρει το αμινοξύ μεθειονίνη (Met) αλληλουχία του οποίου είναι το UACκαι θα προσδεθεί στο κωδικόνιο έναρξης AUG. Για να ολοκληρωθεί η έναρξη το αντικωδικόνιο του μορίου tRNAπροσδένεται με το κωδικόνιο του mRNAαφήνοντας το αμινοξύ του στην ριβοσωμική μονάδα. Η διαδικασία συνεχίζεται με τα υπόλοιπα μόρια tRNA, τα οποία θα προσθέτουν τα αμινοξέα τους με συνέπεια να δημιουργηθεί μια πολυπεπτιδική αλυσίδα. Ο τερματισμός της μετάφρασης ολοκληρώνεται όταν το ριβόσωμα

αναγνωρίσει ένα κωδικόνιο λήξης (UAA, UAG,UGA) τα οποία θα σημάνουν την λήξη της μετάφρασης .

1.2.5 Γονίδια και Γονιδίωμα

Κάθε ανθρώπινος οργανισμός είναι μοναδικός. Έτσι και τα είδη των κυττάρων από τα οποία αποτελείται διαφέρουν μεταξύ τους ως προς την μορφή και τη λειτουργία και αυτό γίνεται, διότι κάθε είδος κυττάρου είναι υπεύθυνο για την παράγωγη εξειδικευμένων πρωτεϊνών, γεγονός που οφείλεται στην γονιδιακή έκφραση. Τα γονίδια περιλαμβάνουν στοιχεία που είναι απαραίτητα για την σύνθεση πρωτεϊνών, επομένως, όταν τα γονίδια είναι ανενεργά, σημαίνει ότι η μετάβαση της γενετικής πληροφορίας από το DNAστο RNAκαι στη συνέχεια στην πρωτεΐνη, δηλαδή η γονιδιακή έκφραση, δεν έχει επιτευχθεί. Η απόσταση από το γονίδιο στην πρωτεΐνη είναι μεγάλη και περίπλοκη, ωστόσο υπάρχουν σημεία επιτάχυνσης ή επιβράδυνσης και ενεργοποίησης ή απενεργοποίησης της διαδικασίας. Επιπρόσθετα, ανάλογα τον τύπο των κυττάρων που φέρουν οι ιστοί, η έκφραση των γονιδίων γίνεται επιλεκτικά. Δηλαδή, σε ορισμένα κύτταρα λαμβάνει μέρος η έκφραση συγκεκριμένων γονιδίων, τα οποία όμως δεν εκφράζονται σε άλλα κύτταρα.

Τα γονίδια παίζουν σημαντικό ρόλο στην διακυτταρική επικοινωνία αλλά και στην εσωκυτταρική. Παραδείγματος χάριν, η γενετική πληροφορία εξωθείται από τον πυρήνα στο κυτταρόπλασμα με την συμβολή του RNAκαι ενεργοποιείται η λειτουργία ορισμένων οργανύλιων με σκοπό την πρωτεινοσύνθεση. Όσων αφορά τη διακυτταρική επικοινωνία οι πρωτεΐνες, που αποτελούν παράγωγα γονιδίων, συντελούν στην σηματοδότηση από κύτταρο σε κύτταρο. Το αποτέλεσμα αυτού του σήματος μπορεί να πυροδοτήσει την δημιουργία νέων πρωτεϊνών. Ένα τέτοιο παράδειγμα είναι η κυτταρική σηματοδότηση της ανάπτυξης ενός οργανισμού από έμβρυο σε ενήλικο. Συμπερασματικά, τα γονίδια αποτελούν συγκεκριμένες αλληλουχίες των νουκλεοτιδίων του DNA, στις οποίες συμβαίνουν μεταλλάξεις, όταν αυτές δεν πληρούν τους κανόνες συμπληρωματικότητας ή όταν αυτές διαφέρουν από τον τυπικό πρότυπο μηχανισμό αντιγραφής DNA. Οι μεταλλάξεις ενδέχεται να είναι θετικές ή αρνητικές, όσων αφορά το αναπτυξιακό επίπεδο. Υπάρχει όμως και η περίπτωση να είναι επιβλαβείς οι μεταλλάξεις, γεγονός που γεννά θέματα βιοηθικής. Το είδος των μεταλλάξεων ποικίλλει ανάλογα με το εύρος των νουκλεοτιδίων, παραδείγματος χάριν όταν κάνουμε λόγο για σημειακή μετάλλαξη εννοείται η αντικατάσταση ενός νουκλεοτιδίου του DNAμε ένα άλλο και οι παράγοντες μεταλλάξεων μπορεί να είναι περιβαλλοντικοί, φυσικοί ή χημικοί. Κατ’ επέκταση ως γονιδίωμα ορίζεται το σύνολο των γονιδίων που υπάρχουν στο ανθρώπινο κύτταρο και η εργαστηριακή μελέτη του αναδεικνύεται ολοένα και πιο σημαντική

1.2 Τεχνολογίες Αλληλούχισης

1.2.1 Πρώτης γενιάς τεχνολογίες αλληλούχισης

Ο κλάδος της γενετικής τα τελευταία 30 χρόνια εχει κάνει αλματα με την ανάπτυξη της πρώτης και δεύτερης γενιάς τεχνολογίες αλληλούχισης(DNA sequencing technology) . Η πρώτη γενιά βασίζεται σε δυο μεθόδους που αναπτυχθηκαν παράλληλα την μεθοδο τερματισμού (chain termination method) το 1975 από τον Sanger και Coulson και την χημική μέθοδο από τον Maxam και Gilbert το 1977.

Η μέθοδο Sanger ,που έγινε η πιο συχνά χρησιμοποιούμενη από τις δύο μέθοδος λόγω της μικρότερης πολυπλοκότητας του, αποτελείται από μία ενζυμική καταλυτική αντίδραση που πολυμερίζει τα τμήματα DNA συμπληρωματικά στη μήτρα DNA. Ένας P 32 σημασμένος εκκινητής (λίγα ολιγονουκλεοτίδια με αλληλουχία συμπληρωματική της μήτρας DNA)

υβριδοποιείται σε μια συγκεκριμένη περιοχή της μήτρας DNA παρέχοντας το εναρκτήριο σημείο της σύνθεσης του DNA. Παρουσία της DNA πολυμεράσης συμβαίνει καταλυτικός πολύ μερισμός των τριφωσφορικών δεόξυνουκλεοσιδίων στο DNA. Ο πολυμερισμός συνεχιζόταν μέχρι το ένζυμο να συναντήσει ένα τροποποιημένο νουκλεοσίδιο το οποίο καλείται νουκλεοσίδιο τερματισμού ή τριφωσφορικό διδεόξυ νουκλεοσίδιο στην αναπτυσσόμενη αλυσίδα. Αυτή η μέθοδος πραγματοποιείται σε τέσσερις διαφορετικούς σωλήνες, καθένας από τους οποίους περιέχει την κατάλληλη ποσότητα ενός από τα τέσσερα ddNTPs. Όλα τα δημιουργηθέντα τμήματα έχουν το ίδιο 5’ άκρο, ενώ το 3'άκρο καθορίζεται από το διδεόξυ νουκλεοτίδιο που χρησιμοποιήθηκε στην αντίδραση. Μετά την ολοκλήρωση και των τεσσάρων αντιδράσεων το μίγμα των διαφορετικού μεγέθους DNA τμημάτων διαχωρίζεται με τη διαδικασία της ηλεκτροφόρησης, σε ένα αποδιατακτικό gel ακρυλαμίδης, σε τέσσερα διαφορετικά πηγάδια. Η απεικόνιση των ζωνών γίνεται με αυτοραδιογραφία.

Η ενζυμική μέθοδος ήταν αρκετά χρονοβόρα και ιδιαίτερα επισφαλής. Για το λόγο αυτό, αναπτύχθηκε μια εναλλακτική μέθοδος σήμανσης η οποία αντικατέστησε την ραδιενέργεια. Για την αλληλούχιση ενός τμήματος DNA γίνεται μία σύνθετη αντίδραση τερματισμού. Η αντίδραση πραγματοποιείται παρουσία των τεσσάρων κανονικών τριφωσφορικών δεόξυριβονουκλεοτιδίων σε μεγάλη σχετικά συγκέντρωση και τεσσάρων διδεόξυριβονουκλεοτιδίων σε μικρότερη συγκέντρωση τα οποία είναι σημασμένα το καθένα με διαφορετική φθορίζουσα χημική ομάδα. Με αυτό τον τρόπο έχουμε σχηματισμό από μίγμα προϊόντων τερματισμού που μπορεί να έχουν οποιαδήποτε από τις τέσσερις βάσεις στο 3' άκρο τους. Τα προϊόντα αυτά προκύπτουν από την ενσωμάτωση ενός διδεόξυριβονουκλεοτιδίου σε μία τυχαία θέση κατά τη σύνθεση. Ωστόσο επειδή τέσσερις φθορίζουσες χρωστικές που χρησιμοποιούνται εκπέμπουν φωτεινή ακτινοβολία σε διαφορετικό μήκος κύματος (διαφορετικό χρώμα) η ταυτότητα της βάσης στην οποία τερματίζεται η σύνθεση αντιστοιχεί στο χρώμα του ddNTP που έχει ενσωματωθεί στο 3' άκρο. Τα προϊόντα της αντίδρασης φορτώνονται και αναλύονται στην ίδια διαδρομή του πηκτώματος ή σε ένα τριχοειδές σωληνάκι μιας συσκευής αυτόματης αλληλούχισης. Τα τμήματα διαχωρίζονται ανάλογα με το μέγεθος τους. Τα τμήματα είναι σημασμένα με τα χρώματα που αντιστοιχούν στα τέσσερα διαφορετικά ddNTP ανάλογα με την ταυτότητα του τελευταίου νουκλεοτιδίου τους. Η ανίχνευση των τεσσάρων χρωμάτων φθορισμού των τερματικών προϊόντων τερματισμού γίνεται από το λέιζερ ανιχνευτή της συσκευής αλληλούχισης. Η μέθοδος αλληλούχισης DNA χρησιμοποιείται για γονιδιωματική έρευνα γιατί μπορούν εύκολα να παραχθουν γονιδιωματικές βιβλιοθήκες από εισαγόμενα τμήματα DNA διάφορων μεγεθών αφού η κατασκευή γονιδιώματος επωφελειται απο κατασκευασμένες αλληλουχίες (reads) διάφορων μεγεθών. Ωστόσο λόγω της χαμηλής απόδοσης και του υψηλού κόστους της πρώτης γενιάς νεες μεθοδοι αναπτύχθηκαν οδηγώντας στη δευτερη γενιά αλληλούχισης(second generation sequencing NGS).

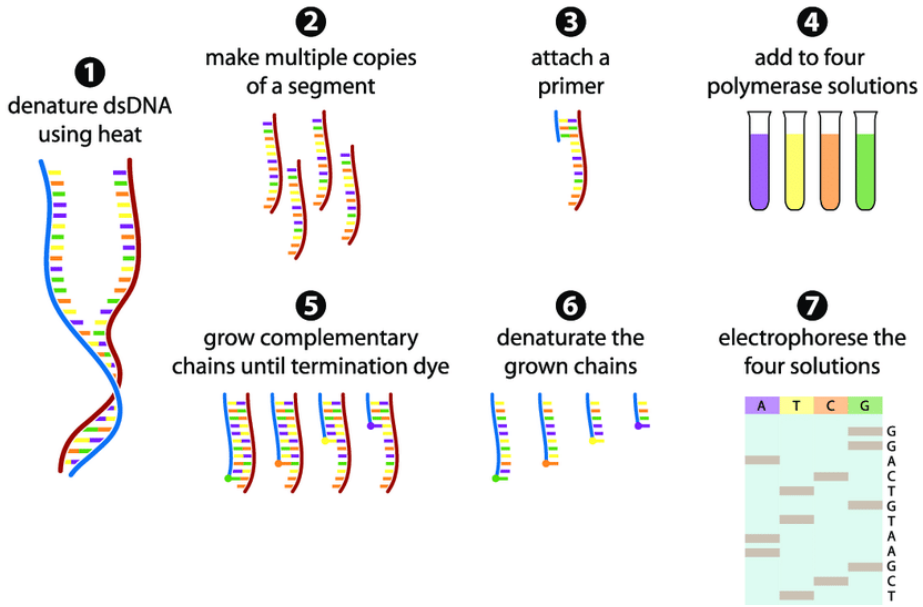


Figure 2: Μέθοδος Sanger πηγή https://www.researchgate.net/figure/The-Sanger-sequencing-method-in-7-steps-1-The-dsDNA-fragment-is-denatured-into-two_fig2_234248746

1.2.2 Δεύτερης γενιάς τεχνολογίας αλληλούχισης

Παρά τις διαρκείς βελτιώσεις στη μέθοδο του Sanger, οι περιορισμοί της μεθόδου συνέχισαν να υπάρχουν και αυτό οδήγησε στην ανάπτυξη νέων τεχνολογιών. Η νέα γενια τεχνολογίας NGS αναπτύχθηκε με τις βασικές διαφορές την ικανότητα ταυτόχρονης αλληλούχισης εκατομμυρίων τμημάτων DNA (massively parallel sequencing technologies) , αλληλούχιση μεγαλύτερων τμημάτων DNA σε μικρότερο χρονικό διάστημα και με μεγαλύτερη ταχύτητα. Στις τεχνολογίες δεύτερης γενιάς ανήκουν κυρίως αυτές της Roche 454 Pyrosequencing, Solid, και Illumina/Solexa.

Το 2005, η εταιρία 454 Life Sciences παρουσίασε την πρώτη νέας γενιάς (NGS) πλατφόρμα αλληλούχισης. Η τεχνολογία ονομάζεται pyrosequencing και πραγματοποιεί αλληλούχιση με σύνθεση, σε πραγματικό χρόνο. Σε μία Picotiter πλάκα κάθε νουκλεοτίδιο δεσμεύεται από την DNA πολυμεράση με αποτέλεσμα την απελευθέρωση πυροφωσφορικού μορίου . Τα ένζυμα ATP σουλφορυλάσης και λουσιφεράσης, μετατρέπουν τα πυροφωσφορικά μόρια με αποτέλεσμα την εκπομπή ορατού φωτός, το οποίο ανιχνεύεται από CCD σύστημα κάμερας. Κάθε τύπος νουκλεοτιδίου (dATP,dCTP,dGTP ΚΑΙ dTTP) πλένεται πάνω στην Picotiter πλάκα και αναλύεται ξεχωριστά για τον κάθε κύκλο αλληλούχισης.

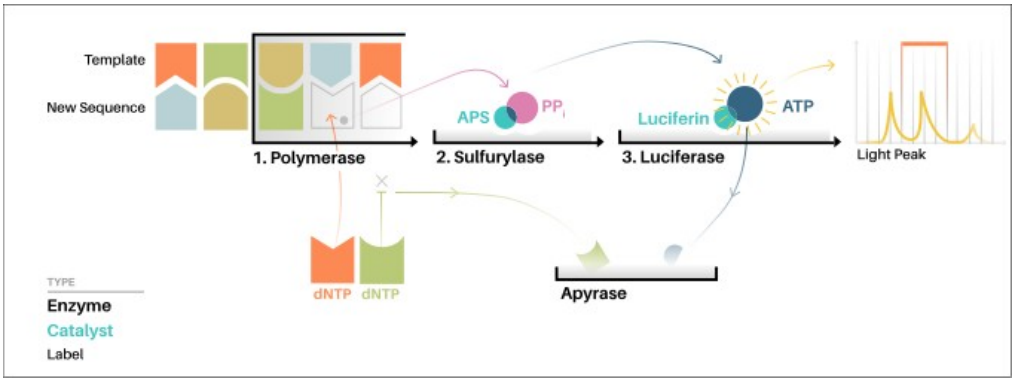


Figure 3: Μέθοδος Sanger πηγή : https://commons.wikimedia.org/wiki/File:How_Pyrosequencing_Works.svg

Το 2007, η εταιρεία Illumina απέκτησε την Solexa, η οποία ανέπτυξε μια πολύ επιτυχημένη τεχνολογία αλληλούχισης των γονιδιωμάτων.Η μέθοδος είναι παρόμοια με αυτή της ηλεκτροφόρησης με τριχοειδή (capillary electrophoresis - CE) δηλαδή ο προσδιορισμός των αζωτούχων βάσεων ενός θραύσματος πραγματοποιείται μέσω των σημάτων που εκπέμπονται. Συγκεκριμένα η διαδικασία sequencing εκτελείται μεσω των εξής βημάτων. Πρώτα γίνεται η προετοιμασία της βιβλιοθήκης (Library Preparation). Τα δίκλωνα μόρια του DNA των δειγμάτων τεμαχίζονται με την βοήθεια διάφορων ενζύμων σε θραύσμτα . Ένα ολιγονουκλεοτίδιο 'T' προσδένεται στα θραύσματα και προεξέχει. Στην συνέχεια συνδέονται και στα δύο άκρα των θραυσμάτων του DNA οι λεγόμενοι adapters. Οι adapters έχουν συγκεκριμένα αλλά διαφορετικά barcodes για το κάθε δείγμα. Τα barcodes είναι μεμονωμένες αλληλουχίες οι οποίες προστίθενται στα δείγματα ώστε να μπορεί να γίνει αναγνώριση του θραύσματος ώστε κατά την ανάλυση των δεδομένων να μπορουμε να το ταυτιστουμε με το δείγμα στο οποίο ανήκει. Μετά την σύνδεση των adapters στα θραύσματα του DNA γίνεται αποδιάταξη των δίκλωνων μορίων σε μονόκλωνα. Μετά το library preparation τα μονόκλωνα μόρια τοποθετούνται πάνω σε μια πλάκα (workflow-glass flow cell).Η κάθε πλάκα αποτελείται εσωτερικά από ολιγονουκλεοτίδια τα οποία είναι συμπληρωματικά ως προς τους adapters, και χωρίζεται σε οχτώ ξεχωριστές λωρίδες. Πραγματοποιείται υβριδισμός (μέσω εναλλαγής υψηλής με χαμηλή θερμοκρασία) μεταξύ των ολιγονουκλεοτιδίων της πλάκας με τους adapters του ενός άκρου των μονόκλωνων θραυσμάτων DNA. Οι ελεύθεροι adapters των μονόκλωνων μορίων υβριδίζονται με τα ολιγονουκλεοτίδια της πλάκας δημιουργώντας γέφυρες (bridge amplification). Μία ισοθερμική πολυμεράση δημιουργεί την συμπλωρωματική αλυσίδα κατασκευάζοντας έτσι δίκλωνες γέφυρες οπου και μετουσιώνονται.Αυτή η διαδικασία γεφύρωσης και μετουσίωσης συμβαινει παράλληλα εκατομύρια φορές και στο τέλος όλοι αυτοί οι συμπληρωματικοι κλώνοι των θραυσματων που έχουν δημιουργηθεί απομακρύνονται. Η κάθε βιβλιοθήκη θραυσμάτων αποτελείται πλέον από εκατοντάδες εκατομμύρια μοναδικά συμπλέγματα (clusters). Η

διαδικασία αυτή ονομάζεται cluster generation. Τέλος πραγματοποιείται η αλληλούχιση(sequencing) όλων των cluster και γίνεται ταυτόχρονα βάση προς βάση με παράλληλο τρόπο χρησιμοποιώντας τέσσερις διαφορετικές φθορίζουσες χρωστικές συνδεδεμένες με τέσσερα διαφορετικά ολιγονουκλεοτίδια (A.T,G και C) . Οι τέσσερις φθορίζουσες με τις βάσεις πλησιάζουν την βάση του cluster αλλά μόνο μία θα ενωθεί μαζί της . Μόλις το λέιζερ - CCD camera ανιχνεύσει ότι όντος η συμπληρωματική βάση είναι σωστή τότε καταγράφεται το χρώμα της φθορίζουσας της βάσης, η φθορίζουσα χρωστική αφαιρείται και μένει η βάση. Το ίδιο γίνεται και για την επόμενη βάση της αλυσίδας του cluster μέχρι να τερματιστεί.

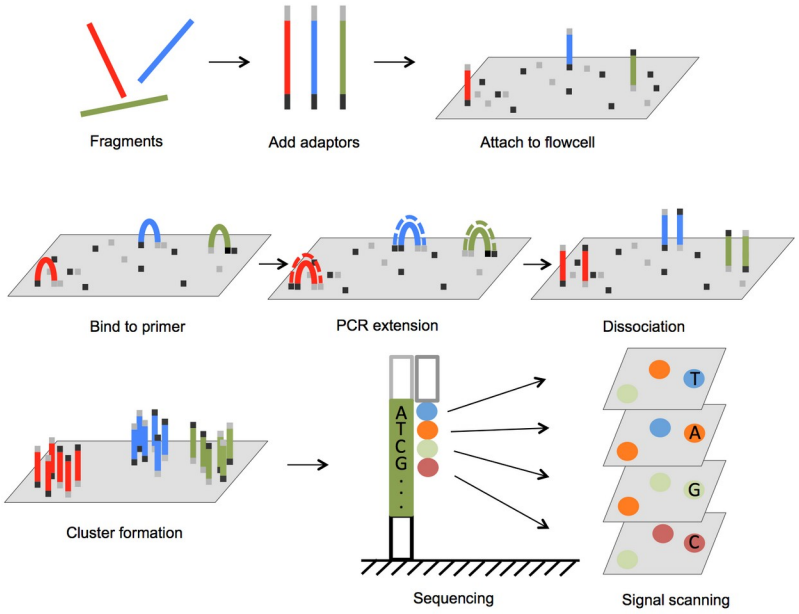


Figure 4: Μέθοδος Illumina/Solexa πηγή: https://www.intechopen.com/books/next-generation-sequencing-advances-applications-and-challenges/next-generation-sequencing-in-aquatic-models

1.2.3 Τρίτης γενιάς τεχνολογίας αλληλούχισης

Η τρίτης γενιάς τεχνολογία αλληλούχησης (third generation sequencing) γνωστό και ως long-read sequencing έχουν την ικανότητα να παράγουν reads της τάξης μεγέθους μεγαλύτερο των 10.000 bp (base pair) σε αντίθεση με της δεύτερης γενιάς που είναι της τάξης των μερικών εκατοντάδων bp. Μέσω των τεχνολογιών της δεύτερης γενιάς έχει πραγματοποιηθεί η μαζική ανάλυση διάφορων μικρών νουκλεοτιδίων αλλά η ανάλυση μεγαλύτερων παραμένει δύσκολη. Συγκεκριμένα επειδή τα γονιδιώματα εμφανίζουν επαναλαμβανόμενες αλληλουχίες, κατά την de novo κατασκευή γονιδιώματος (κατασκευή χωρίς την χρήση ενός γονιδιώματος αναφοράς) με την

χρήση short reads έχει ως αποτέλεσμα να “χαθεί” ένα σημαντικό κομμάτι της γενετικής πληροφορίας καθιστώντας την μελέτη της δομής του γονιδίου αναξιόπιστη. Με την διαθεσιμότητα της τρίτης γενιάς που παράγουν reads κατά μέσω όρο τα 10.000bp, με μερικές τεχνικές να παράγουν μέχρι και 100.000 bp, τα reads διατρέχουν τις επεναλαμβανόμες αλληλουχίες και δίνουν την ικανότητα να κατασκευάσουμε πιο συναφείς γονιδιώματα. Όσο αναφορά την δομική ανάλυση τα long reads έχουν δώσει την ικανότητα για βελτιωμένες ‘split-read’ αναλύσεις ώστε εισαγωγές, διαγραφες, μετατοπίσεις και άλλες διάφορες δομικές μεταβολές να μπορούν εύκολα να αγνωριστούν.

Οι τρίτης γενιάς μπορούν να ταξινομηθούν σε 3 βασικές κατηγορίες .(i) Μέθοδος όπου η DNA πολυμεράση παρατηρείτε καθώς συνθέτει ένα μόριο DNA.(ii) Αλληλούχιση με νανοπορώδες υλικά (Nanopore sequencing) όπου το μόριο DNA περνά μέσα από ένα νανοπορώδες υλικό και κάθε βάση καταγράφεαι.(iii) Άμεση απεικόνιση τού κάθε μορίου DNA με τη χρήση τεχνικές μικροσκοπίας. Οι τρεις εμπορικά διαθέσιμες DNA sequencing τεχνολογίες είναι της Pacific Biosciences (PacBio), το Single Molecule Real Time (SMRT), της Illumina το Synthetic Long-Read και η πλατφόρμα της Oxford Nanopore sequencing.

Το SMRT της Pacific Biosciences είναι η πρώτη μέθοδος που αναπτύχθηκε και δίνει την ικανότητα να παρατηρούμε σε πραγματικό χρόνο την DNA πολυμεράση καθώς συνθέτει ένα κλώνο DNA. Λόγω του μεγέθους της πολυμεράσης που είναι της τάξης των 10 nm σε διάμετρο δυο βασικά προβλήματα έπρεπε να λυθουν. Πρώτον να βρεθεί ένας τρόπος περιορισμού του ενζύμου σε ένα χώρο αρκετά μικρό όπου θα μπορούμε να το παρατηρούμε καθώς συνθετει και ο λόγος σήματος/θορύβου είναι τετοιος ώστε να διακρίουμε τις βάσεις με όσο δυνατόν λιγότερα σφάλματα. Δεύτερον η τιτλοφόρηση των νουκλεοτιδίων να πραγματοποιηθεί με τέτοιο τρόπο ώστε η χρωστική που προσθέτεται να αφαιρείται στην αρχή κάθε κύκλου με σκόπο ώστε ο κλώνος DNA που συνθέτεται πολλές φορές η χρώστικεςπου του προσκολούνται να μην γεμίζουν τον χώρο που γίνεται η σύνθεση.

Το πρώτο εμπόδιο λύθηκε με τη χρήση ενός κυματοδηγού του zero-mode waveguide (ZMW).Η αρχή που χρησιμοποιείται είναι παρόμοια με αυτήν που χρησιμοποιείται προστατευτική οθόνη στην πόρτα των φούρνων μικροκυμάτων.Η οθόνη είναι διάτρητη με τρύπες που είναι πολύ μικρότερες από το μήκος κυμάτων των μικροκυμάτων αλλά πολύ μεγαλύτερες από τα μήκος κύματος του ορατού φωτός που έχει ως αποτέλεσμα οι τρύπες να εμποδίζουν τη διέλευση των πολύ μεγαλύτερων μικροκυμάτων και έτσι εμείς μπορούμε να παρατηρούμε το φαγητό καθώς ζεσταίνεται, ομοίως δουλεύει και ο ZMW. Ο ZMW είναι μία μια τρύπα μερικών δεκάδων νανομέτρων κατεσκευασμένο μέσα σε ένα μεταλικό φιλμ 100 nm το οποίο εναποτίθεται σε γυάλινο υπόστρωμα. Το μικρό μέγεθος του ZMW εμποδίζει το φώς του lazer (που έχει μήκος κύματος 600 nm) να το διαπεράσει και αντί αυτού το φως φθίνει εκθετικά. Αυτό έχει ως αποτέλεσμα όταν εκπέμπουμε τη δέσμη lazer πάνω στο γυαλί μόνο τα πρώτα 30 nm του κυματοδηγού φωτίζονται. Μέσα σε κάθε ZMW μια DNA πολυμεράση προσδένεται με την χρήση της αλληλεπιδρασης βιοτίνης/στρεπταβιδίνη. Διάχυση στο επίπεδο της νανοκλίμακας πραγματοποιείται μέσα σε μερικα μικρά του δευτερολέπτου όπου τα σημασμένα νουκλεοτίδια ταξιδεύουν μέσα στον ZMW περικυκλώνουν την πολυμεράση και μετά πάλι μέσω διάχυσης βγαίνουν έξω από την τρύπα. Καθώς το lazer διαπερνά μόνο τα πρώτα 30 nm για να διεργείρει τα νουκλεοτίδια που έχουν σημανθεί από φθορίζουσες ουσίες εξασφαλίζουμε ότι διεργείρονται μονο όσα είναι μέσα στην τρύπα. Όταν το σωστό νουκλεοτίδιο ανιχνεύεται από την πολυμεράση τότε ενσωματώνεται στο αναπτυσσόμενο σκέλος DNA σε μια διαδικασία που παίρνει milliseconds

δηλαδη 3 ταξείς πιο πολύ απο την ταχύτητα διάχυσης. Αυτή η διαφορά στον χρόνο προκαλεί ένα πιο ενισχυμένο σήμα για ενσωματωμένο έναντι μη ενσωματωμένο νουκλεοτίδιο που έχει ως αποτέλεσμα ένα επιθυμητό υψηλό λόγο σήμα/θορύβου. Καθώς η πολυμεράση αλληλεπιδρα με τον κλώνο κατά την ενσωμάτωση ενός νουκλεοτιδίου η φθορίζουσα ουσία εκπέπει φως χαρακτριστικό για κάθε νουκλεοτίδιο . Το σύστημα καταγράφει αυτην την εκπομπη και την ταυτίζει με το σωστό νουκλεοτίδιο κατασκευάζοντας έτσι την αλληλουχία. Άρα ο ZMW έχει την ικανότητα να καταγράφει το συμβαν της ενσωμάτωσης ενός νουκλεοτιδίου σε πλαίσιο όπου υπάρχει μια σημαντική συγκέντρωση από σημασμένα με φθοριζουσα χρωστκή νουκλεοτίδια. Το δεύτερο εμπόδιο αντιμετωπίζεται με την ενσωμάση της φθορίζουσας χρωστικής στην φωσφορική αλυσίδα αντί στα νουκλεοτίδια αφού κατά την φυσική διαδικασία η αλυσίδα αυτή διασπάται μετά την ενσωμάτωση των νουκλεοτιδίων.

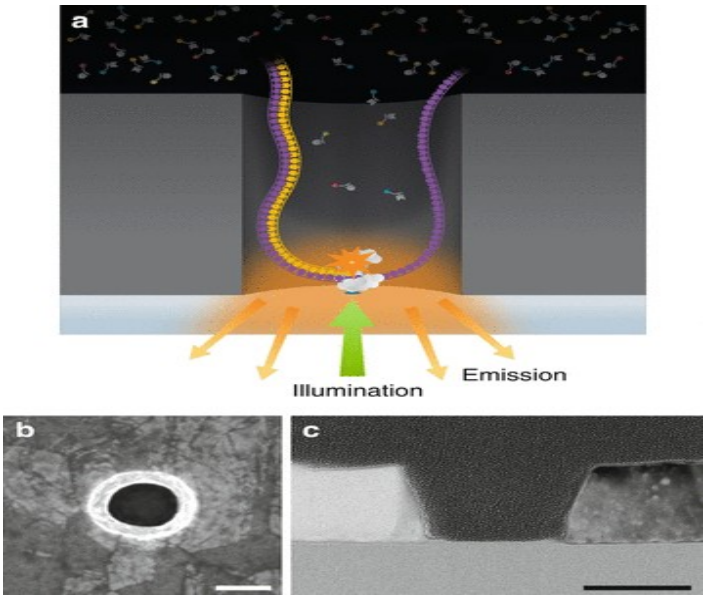


Figure 5: Κυματοδηγός ZMW πηγή: https://link.springer.com/referenceworkentry/10.1007%2F978-3-642-16712-6_499

Οι περισσότερες nanopore sequencing τεχνολογίες βασίζονται στη διεύλεση ενός μορίου DNA ή των αζωτούχων βάσεων από μια τρύπα (πορο) και να ανιχνεύονται οι βάσεις μέσω την επίδραση τους στο ηλεκτρική ροή. Επειδή αυτή η μέθοδος αξιοποιεί μη τροποποιημένο DNA έχει την ικανότητα να πραγματοποιηθεί γρήγορα με εξαιρετικά μικρές ποσότητες υλικού που μελετάμε. Άλλο ένα πλεονέκτημα αυτής της τεχνολογίας είναι το χαμηλό κόστος λόγο της βάσης της σε ηλεκτρικά και όχι οτπικά φαινόμενα. Πιο συσγκεκριμένα η μέθοδος της Oxford Nanopore (από όπου από αυτήν έχουν αλληλουχηθεί τα δεδομένα που θα χρησιμοποιήσουμε σε αυτην την εργασία) βασίζεται σε 3 βιολογικά μόρια που έχουν σχεδιαστεί με τέτοιο τρόπο ώστε να λειτουργούν ως σύστημα. Οι βιολογικός νανοπόρος κατασκευάζεται από έναν τροποποιημένο α-hemolysin πόρο όπου έχει την εξονουκλεάση συνδεδεμένη στο εξωκυτταρικό μέρος του πόρου.

Επίσης ένας συνθετικός σένσορας προσάρταται ομοιπολικά στην εσωτερική επιφάνεια του νανοπόρου. Αυτό το σύστημα περιέχεται σε μια συνθετική διπλής στιβάδας λιπιδίων ώστε όταν το DNA φορτώνεται στο μέρος της εξονουκλεάσης και εφαρμόζοντας μια διαφορά τάσης η εξονουκλεάση μπορεί να αποκόψει ξεχωριστά κάθε νουκλεοτίδιο. Έτσι μόλις αποκοπούν κάθε ένα ανιχνεύεται μέσω της μεταβολή που προκαλούν στο δυναμικό πεδίο και οι πληροφορίες περνούν σε ένα μικροτσίπ το application-specific integrated circuit (ASIC).

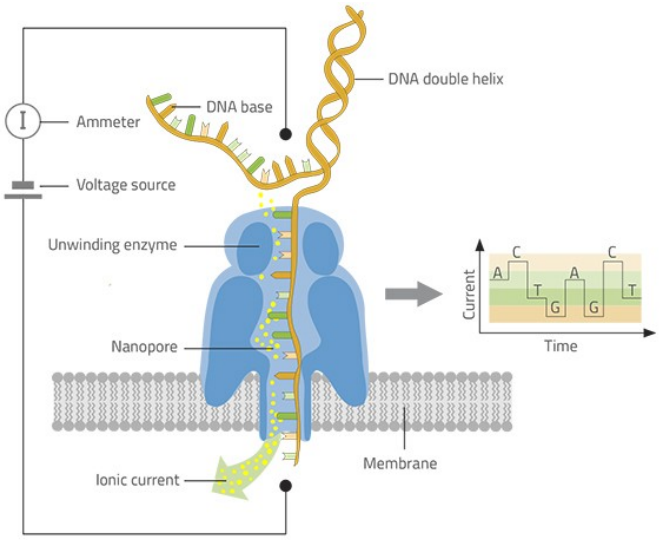


Figure 6: Εικόνα της αρχής λειτουργίας μιας συσκευής προσδιορισμού αλληλουχίας DNA με την χρήση νανοπόρων πηγή:https://www.scienceinschool.org/sites/default/files/articleContentImages/43/dnasequencing/issue43_dnasequencing_fig2.jpg

Η συσκευή MinION της Oxford είναι η μικρότερη εμπορικά διαθέσιμη συσκευή αλληλούχησης με διαστάσεις 10x3x2 cm και ζυγίζει 90 γραμμάρια. Μπορεί να συνδεθεί απευθείας σε μια θύρα USB3 ενός υπολογιστή , ο οποίος πρέπει τουλάχιστον να αποτελείται από ενα solid state drive (SSD), μια κάρτα ram των 8 GB και ένα σκληρό δίσκο με χωρητικότητα μεγαλύτερη των 128 GB , και να επιτελέσει την διαδικασία της αλληλούχισης. Η συσκευή έχει 512 κανάλια και επιτρέπει μέχρι και 512 ανεξάρτητα μορια DNA να αλληλουχηθούν τατόχρονα. Κάθε κανάλι είναι συνδεδεμένο με 4 πηγάδια και μπορεί να παράγει δεδομένα μόνο για ένα την φορά. Η απόδοση στην παραγωγή δεδομένων ανα κάνλι διαφέρει λόγω ότι κάπιοια είναι πιο ενεργά από τα άλλα. Για να μειωθεί ο θόρυβος τα δεδομένα για τις μετρήσεις των ρευμάτων επεξεργάζονται και

μετατρέπονται σε μια αλληλουχία δεδομένων γνωστό και ως squible-plot. Καθώς το DNA περνάει μέσα από τον πόρο η αντίσταση του καθοριζεται απο τις βάσεις που είναι παρόν μέσα στα νουκλεοτίδια και αυτό οδηγεί σε ένα τεράστιο αριθμό καταστάσεων συγκεκριμένα 4^5=1024 διαφορετικές για το κλασσικό συνδυασμό των 4 βασεων. Με την παρουσία τροποποιημένων όπως η 5-methylcytosine ο αριθμός των των πιθανών καταστάσεων αυξάνεται στις 5^5= 3125. Αυτό καθιστά το basecalling ένα μεγάλο πρόβλημα για το machine learning και βασικός παράγοντας που καθορίζει την ποιότητα και χρησιμότητα της αλληλούχισης με νανοπόρους.

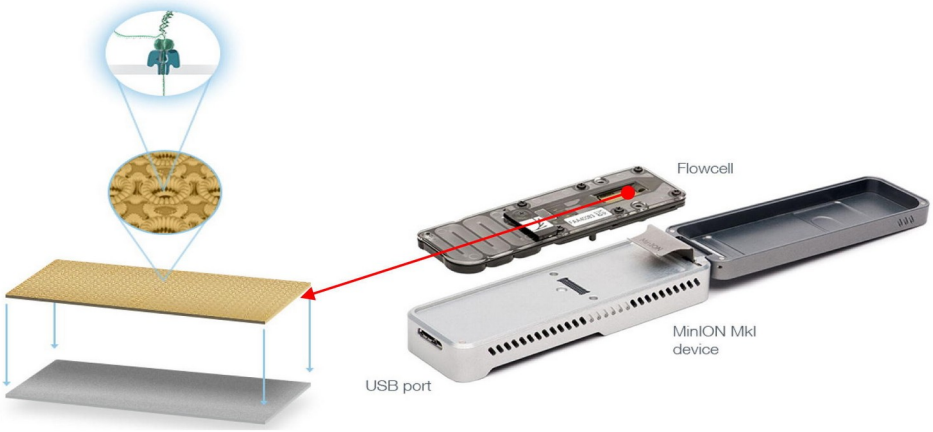


Figure 7: Συσκευή MinIOn πηγή: https://ars.els-cdn.com/content/image/1-s2.0-S1672022916301309-gr1.jpg

1.3 Μεταγονιδιωματική Ανάλυση

1.3.1 Μορφή δεδομένων

Για το data analysis τα περισσότερα εργαλεία της βιοπληροφορικής χρησιμοποιούν αρχεία FASTA ή FASTQ εκτός για κάποιες νέες πλατφόρμες που στα αρχικά τους στάδια χρησιμοποιούν ανεπεξέργαστα δεδεμένα για κάποιες εφαρμογές. Κατά την παρούσα φάση το MinION παράγει ένα

FAST5 αρχείο για κάθε read. Τα FAST5 βασίζονται στον τύπο αρχείου HDF5 Hierarchical Data Format (HDF) τα οποία είναι σχεδιασμένα στο να αποθηκεύουν και να οργανώνουν μεγάλα ποσά δεδομένων.Με αυτήν την ιεαρχική δομή (hierarchical structure) τους τα FAST5 αρχεία δίνουν την δυνατότητα να αποθηκεύουν metadata που συσχετίζονται με κάθε read καθώς και διάφορα γεγονότα που έχουν προεπεξεργαστεί από την συκευή μας. Τα αρχεία FastQ έχουν χαρακτηριστική μορφή. Αποτελούνται από τέσσερις γραμμές για το κάθε read. Η πρώτη γραμμή αρχίζει πάντοτε μετο σύμβολο @και προσδιορίζει το όνομα του read (εικόνα 17). Πολλές φορές (όπως στην περίπτωση αρχείων από Illumina) μπορεί να αναφέρονται πληροφορίες σχετικά με την θέση του read στο flow cell. Στην δεύτερη γραμμή εμφανίζεται η αλληλουχία του read. Δηλαδή A,T,G και C. Η εμφάνιση του γράμματος N δηλώνει ότι η βάση δεν μπόρεσε να διαβαστεί. Η τρίτη γραμμή περιέχει μόνο το σύμβολο ¨+¨ ή άλλοτε μπορεί και να συνοδεύεται από το όνομα του read. Τέλος, η ποιότητα της κάθε βάσης του read εμφανίζεται με κωδικοποιημένη μορφή (ASCII) στην τελευταία γραμμή. Ουσιαστικά στην τελευταία γραμμη, τα σύμβολα αντιστοιχούν σε τιμές Q, για την κάθε μια βάση που αλληλουχίθηκε.Το ASCII (American Standard Code for Information Interchange) είναι μία μορφή κωδικοποίησης κειμένουμε την μορφή χαρακτήρων της αγγλικής αλφαβήτου .

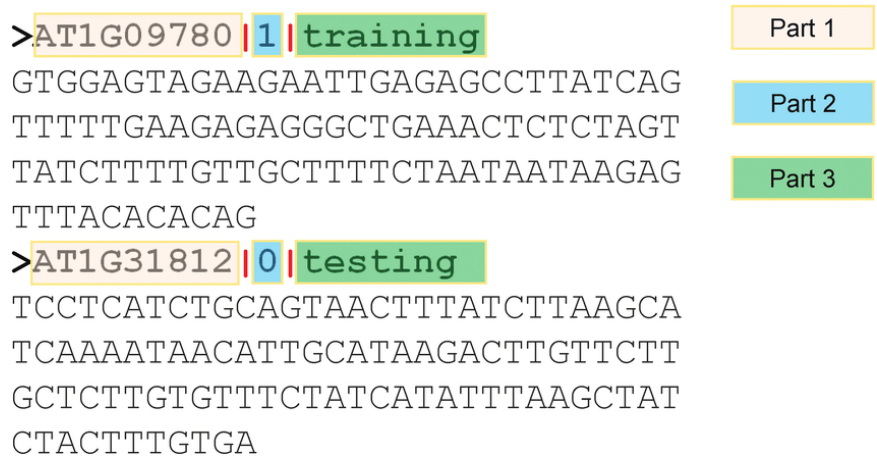


Figure 8: Δομή αρχείων FASTA πηγή : https://www.researchgate.net/profile/Jiangning-Song/publication/331702618/figure/fig2/AS:745823247810566@1554829527399/An-example-of-the-FASTA-format-used-in-iLearn.ppm

Κάθε read παράγεται από ένα από τα 512 κανάλια της MinION συσκευής μας καθώς και κάθε metadata που ακολουθείται από το κάθε read αποθηκεύονται σε ένα χαρακτηριστικό FAST5 αρχείο. Για να διασφαλιστεί η μοναδική ταυτότητα του κάθε αρχείου το κάθε ένα από αυτά ονομάζεται βάσει ενός συνδυασμού του ονόματος του πειράματος ,καναλιού και το πακέτο γονιδιώματος που αλληλουχήθηκε.Μόλις όλα τα δεδομένα που αποθηκεύονται στα FAST5 αρχεία μας συλλεχθούν υποβάλλονται σε μία ανάλυση όπου μεταφράζονται στην επιθυμητή αλληλουχία νουκλεοτιδίων, αυτή η διαδικασία ονομάζεται basecalling. Η ακρίβεια του basecalling υπολογίζεται από το Qscore (Phred quality score) το οποίο αποτελεί κοινή μέτρηση για την αξιολόγηση της ακρίβειας της αλληλούχισης μας.

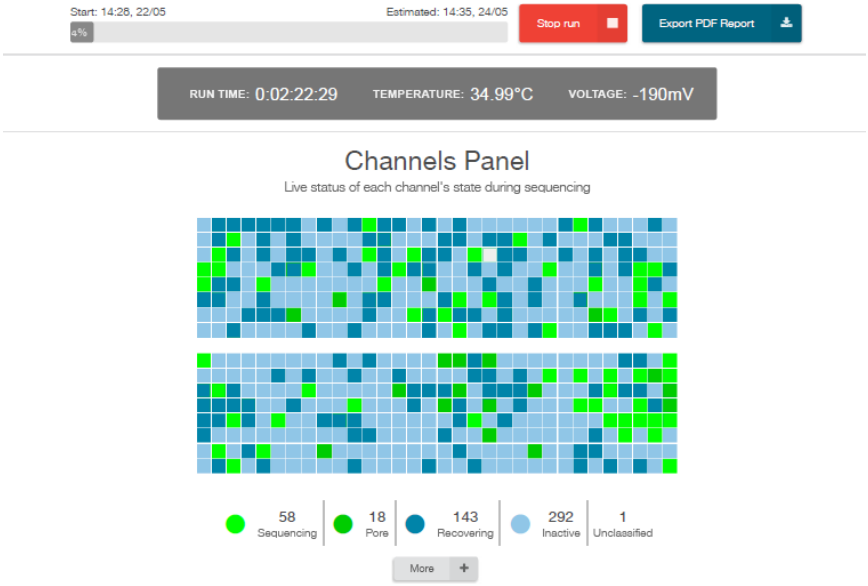


Figure 9: Δραστηριότητα των καναλιών της MinION συσκευής κατά την διαδικασία της αλληλούχισης

1.3.2 Πρότυπο Sanger (Standard Sanger) / βαθμολογία ποιότητας PHRED

Ο βαθμός - τιμή ποιότητας Q (Quality ή Q - score) είναι μια ακέραια τιμή που προκύπτει από την πιθανότητα να έχει γίνει λάθος στην αλληλούχιση μιας συγκεκριμένης βάσης.Αν P = πιθανότητα να έχει γίνει λάθος στην αλληλούχιση τηςσυγκεκριμένης βάσης, τότε:

$$Q = -10 \log(P)$$

Table 1: Phred τιμές

Phred Quality Score (Q)	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%
60	1 in 1000000	99.9999%

Το PHRED ήταν το πρώτο πρόγραμμα το οποίο ανέπτυξε ακριβή και ισχυρή ποιότητα βαθμολόγισης για την κάθε βάση. Έχει τη δυνατότητα υπολογισμού εξαιρετικά υψηλής ακρίβειας αποτελεσμάτων, που συνδέονται λογαριθμικά με τις πιθανότητες λάθους. Η πιο σημαντική χρήση του PHRED score είναι ο αυτόματος προσδιορισμός ακριβείας και ποιότητας των αλληλουχιών. Μπορεί, επίσης να χρησιμοποιηθεί γιανα εκτιμηθεί εάν οι διαφορές μεταξύ των δύο επικαλυπτόμενων ακολουθιών είναι πιο πιθανό να προκύψουν από τυχαία σφάλματα ή από διάφορα αντίγραφα μιας επαναλαμβανόμενης αλληλουχίας. Το Q20 υποδηλώνει ότι η πιθανότητα η βάση να είναι λανθασμένη είναι 0.01 ενώ το Q30 είναι 0.001 . Όπως είναι φανερό, όσο μεγαλύτερο είναι το Quality score (π.χ Q30 πολύ καλής ποιότητας αλληλούχιση) τόσο μεγαλύτερη ακρίβεια έχει η βάση και άρα τόσο μικρότερη είναι η πιθανότητα λάθους. Η αξιολόγιση της ποιότητας του κάθε read είναι πολύ σημαντική διεργασία καθώς υπάρχει το ενδεχόμενο λάθους ανάγνωσης μίας ή περισσότερωνβάσεων εξαιτίας συστηματικού λάθους, που μπορεί να έχει είτε η τεχνολογία της αλληλούχισης που χρησιμοποιήθηκε είτε η ποιότητα της ίδιας της αλληλουχίας.

1.3.3 N50

Το N50 είναι μια στατιστική μονάδα που ορίζει την ποιότητα συναρμολόγησης ως προς τη συνέχεια. Με δεδομένα ένα σύνολο αλληλεπικαλυπτόμενων τμημάτων DNA που μαζί αντιπροσωπεύουν μια συναινετική περιοχή του DNA (contigs) το N50 ορίζεται ως το μήκος αλληλουχίας του μικρότερου contig στο 50% του συνολικού μήκους του γονιδιώματος. Μπορεί να θεωρηθεί ως το σημείο του μισού της μάζας της κατανομής ο αριθμός των βάσεων από όλα τα contigs μεγαλύτερα από το N50 θα είναι κοντά στον αριθμό των βάσεων από όλες τις καμπύλες μικρότερες από το N50. Για παράδειγμα, για 9 contigs με μήκη 2,3,4,5,6,7,8,9 και 10. Το άθροισμά τους είναι 54, το ήμισυ του αθροίσματος είναι 27. Επίσης το μέγεθος του γονιδιώματος είναι επίσης 54. Το 50% αυτής της κατασκευής θα είναι 10 + 9 + 8 = 27 (μισό μήκος της αλληλουχίας). Έτσι το N50 = 8, που είναι το μέγεθος του μικρότερου contig που μαζί με τα μεγαλύτερα contigs περιέχουν τη μισή αλληλουχία ενός συγκεκριμένου γονιδιώματος.

1.3.3 Basecalling

Μόλις όλα τα δεδομένα που αποθηκεύονται στα FAST5 αρχεία μας συλλεχθούν υποβάλλονται σε μία ανάλυση όπου μεταφράζονται στην επιθυμητή αλληλουχία νουκλεοτιδίων, αυτή η διαδικασία ονομάζεται basecalling. Το basecalling είναι ένα πολύ ενεργό επιστημονικό πεδίο όπου η Oxford Nanopore Technologies και ανεξάρτητοι ερευνητές αναπτύσσουν νέες

μεθόδους. Όλοι οι σύγχρονοι basecallers χρησιμοποιούν neural networks (νευρωνικά δίκτυα) όπου εκπαιδεύονται με την χρήση πραγματικών δεδομένων για αυτόν τον λόγο η απόδοση του κάθε basecaller επηρεάζεται από τα εκάστοτε δεδομένα που χρησιμοποιήθηκαν. Αυτό παίζει ιδιαίτερα μεγάλο ρόλο όταν κάνουμε basecall reads στα οποία εμπεριέχονται τροποποιημένες βάσεις. Σε αυτές τις περιπτώσεις η απόδοση μετάβάλεται ανάλογα στο πως αυτές οι τροποποιήσεις και τα μοτίβα τους αναπαριστώνται στο training set τους.

Η ακρίβεια των basecallers μπορεί να εκτιμηθεί στο read level (ακρίβεια των reads μας) και στο επίπεδο consensus sequence (η υπολογισμένη σειρά των πιο συχνών καταλοίπων ,νουκλεοτίδιο είτε αμινοξύ, που βρίσκονται σε κάθε θέση της ευθυγραμισμένης αλληλουχίας μας). Το read accuracy μετράει την σχέση μεταξύ των κάθε basecalled reads μας με την αλληλουχία από μια έμπιστη αλληλουχία αναφοράς. Η ακρίβεια της consensus sequence υπολογίζει την κοινή συναίνεση της αλληλουχίας που έχει κατασκευαστεί από πολλαπλά αλληλοεπικαλυπτόμενα reads που έχουν προέλθει από το ίδιο γονιδιακό σημείο. Η ακρίβεια των consensus sequences βελτιωνεται με μεγαλύτερα reads, κατασκευασμένες αλληλουχίες από 10 reads είναι πιο ακριβείς από αυτές με 100 reads.

1.3.4 Προετοιμασία βιβλιοθήκης (library preparation)

Το φυσικό περιβάλλον στο οποίο ζει και αναπαράγεται ένα είσος, ένας πληθυσμός ή μια βιοκοινότητα από το οποίο παίρνουμε το δείγμα έχει σημαντική επίδραση στην μετέπειτα ανάλυση. Βιότοποι με λίγα μικροβιακά είδη ή με άνισο πληθυσμό από λίγα κυρίαρχα είδη είναι πιο υποσχόμενοι στόχοι σε σχέση με τα ενδιαιτήματα που έχουν πολλά είδηίσης αφθονίας. Ωστόσο, πιο σημαντικό από τον απόλυτο αριθμό των ειδών είναι το επίπεδο της γονιδιακής συνοχής. Ακόμη και φαινομενικά ιδανικοί βιότοποι με σταθερή σύνθεση λίγων κυρίαρχων ειδών μπορεί να είναι δύσκολο να συναρμολογηθούν όταν οι εξελικτικές προσαρμογές έχουν οδηγήσει σε μεγάλα παν-γονιδιώματα και επομένως, σε ένα χαμηλό επίπεδο κλωνικότητας του πληθυσμού. Σε αντίθεση, φαινομενικά ακατάλληλα ενδιαιτήματαπου φιλοξενούν μια πληθώρα ειδών με δυναμικά μεταβαλλόμενες συνθέσεις μπορεί να δώσουν καλή συναρμολόγιση, όταν τα είδη που ευδοκιμούν και κυριαρχούν είναι σε μεγάλο βαθμό κλωνικά

Το DNA που εξάγεται πρέπει να είναι αντιπροσωπευτικό όλων των κυττάρων που υπάρχουν στο δείγμα και πρέπει να ληφθούν επαρκείς ποσότητες υψηλής ποιότητας νουκλεϊκών οξέων για την παραγωγή βιβλιοθήκης και μετέπειτα για την αλληλούχιση. Η επεξεργασία απαιτεί ειδικά πρωτόκολλα για κάθε τύπο δείγματος και είναι διαθέσιμες διάφορες αξιόπιστες μέθοδοι για την εξαγωγή του DNA. Αν η κοινότητα στόχος συνδέεται με έναν ξενιστή (π.χ. ένα ασπόνδυλο ή φυτό), πρέπει να εξασφαλιστεί ότι λαμβάνεται έστω και ελάχιστο DNA του ξενιστή. Ο φυσικός διαχωρισμός και η απομόνωση των κυττάρων από τα δείγματα θα μπορούσαν επίσης να είναι σημαντικά για τη μεγιστοποίηση της απόδοσης του DNA ή την αποφυγή της συν-εξαγωγής των ενζυματικών αναστολέων που θα μπορούσαν να παρεμβαίνουν με μετέπειτα επεξεργασία.

Η παραγωγή βιβλιοθήκης για τις περισσότερες τεχνολογίες αλληλούχισης απαιτεί υψηλές ποσότητες νανογραμμαρίων ή μικρογραμμαρίων DNA, και ως εκ τούτου μπορεί να απαιτηθεί η ενίσχυση του αρχικού υλικού .

Η κατασκευή υψηλής ποιότητας reads εκτός από πιο σύγχρονες συσκευές sequencing και πιο αποδοτικούς αλγόριθμους για basecaling απαιτεί και τεχνικές library preparation όπου θα προσφέρει υψηλής ποιότητας βιβλιοθήκες DNA όπου θα έχουν ένα βελτιωμένο αντίκτυπο στις μετέπειτα γονιδιωματικές αναλύσεις To library preparation είναι η παρασκευή αλυσίδας νουκλεϊκού οξέος, RNA ή DNA σε μια μορφή που είναι συμβατή με το σύστημα αλληλούχισης που θα χρησιμοποιηθεί. Γενικά αποτελείται από τα εξής βήματα : (i)Κατακερματισμός και / ή ταξινόμηση του μεγέθους των στοχευόμενων αλληλουχιών στο επιθυμητό μήκος, (ii)μετατροπή του στόχου σε δίκλωνο DNA,(iii) προσάρτηση ολιγονουκλεοτιδικών adapators (προσαρμογέων) στα άκρα των θραυσμάτων στόχου , και (iv) ποσοτικοποίηση του τελικού προϊόντος βιβλιοθήκης για αλληλούχιση.

Μια από αυτές είναι το Ligation Sequencing Kit. Σε αυτό το κιτ αρχικά το DNA και πιο συγκερκιμένα δικλωνο DNA (ds DNA) εξάγεται από δείγματα Formalin-Fixed Paraffin-Embedded (FFPE). Το FFPE είναι μια μορφή συντήρησης και προετοιμασίας για δείγματα βιοψίας που βοηθούν στην εξέταση, πειραματική έρευνα και διαγνωστική / ανάπτυξη φαρμάκων.Ένα δείγμα ιστού διατηρείται πρώτα σε φορμαλδεΰδη για τη διατήρηση των πρωτεϊνών και των ζωτικών δομών μέσα στον ιστό. Στη συνέχεια, ενσωματώνεται σε ένα μπλοκ παραφίνης όπου αυτό διευκολύνει την κοπή απαιτούμενων μεγεθών για τοποθέτηση σε πλάκες για εξέταση. Αυτή η μέθοδος εξαγωγής έχει αποδειχθεί [1] ότι είναι αναγκαία για προετοιμασία λόγω της αποτελεσματικότητάς τους στη μετατροπή ενός υψηλού ποσοστού των θραυσμάτων DNA εισόδου σε αλληλουχία μορίων βιβλιοθήκης και της ικανότητάς τους να συλλάβουν μικρά θραύσματα DNA . Στην συνέχεια μετά τον κατακερματισμό τα κομμάτια δεν είναι ομοιογενοί δηλαδή στις άκρες περιέχονται “προεξοχές” απο νουκλεοτίδια που δεν είναι συμπληρωμένα για αυτό μια διαδικασία που ονομάζεται end repair απαιτείται όπου διασφαλίζει ότι κάθε μόριο θα απαλαχτεί από τίς προεξοχές και θα περιέχει 5' φωσφορικές και 3' υδροξυλομάδες, δηλαδή μια πολυμεράση συμπληρώνει τα άκρα και καθιστά όλα τα μόρια ομοιόμορφα “αμβλεία”. Επίσης μια ενσωμάτωση που ονομάζεται da tailing πραγματοποιείται όπου μη πρότυπη 5′-μονοφωσφορική δεοξυαδενοσίνη (dAMP) ενσωματώνεται στο 3’ άκρο των θραυσμάτων DNA .To da tailing βοηθά στην αποτροπή του σχηματισμού αλυσομερών κατά τη διάρκεια των σταδίων ligation καθώς και την προώθηση αποτελεσματικής ligation των ειδικών adaptor αλληλουχιών. Αυτό το κιτ προτείνεται σε περιπτώσεις όπου είναι αναγκαία η μέγιστη απόδοση και απαιτείται έλεγχος για το μήκος του read.

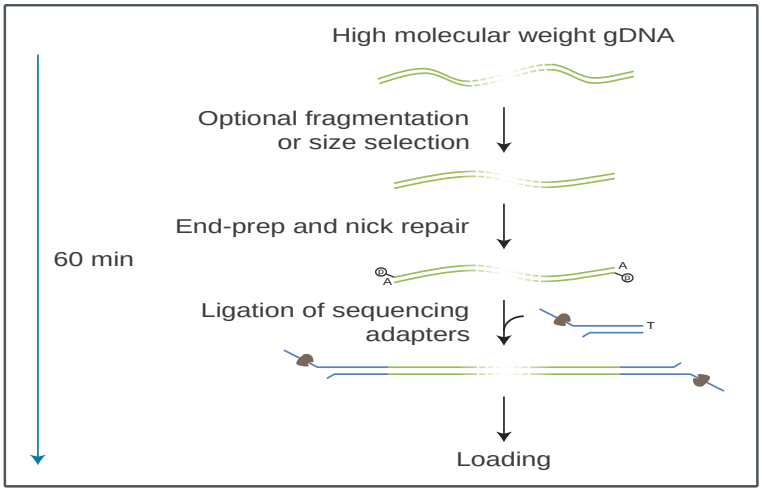


Figure 10: Προετοιμασία βιβλιοθήκης με την ligation μέθοδο πηγή :https://store.nanoporetech.com/us/ligation-sequencing-kit.html

Εκτός από το ligation το rapid Sequencing kit παράγει βιβλιοθήκες αλληλούχισης που εξάγονται απο gDNA μέσα σε 10 λεπτά .Αυτή η μέθοδος library preparation προτείνεται σε περιπτώσεις όπου υπάρχει μικρός χρόνος για προετοιμασία και δεν υπάρχει πρόσβαση σε εξοπλισμός εργαστηρίου δηλαδή τα πλεονεκτήματα του είναι η απλότητα και η ταχύτητα με το μειονέκτημα να μην έχουμε την απόκτηση μέγιστης απόδοσης. Αυτή η μέθοδος βασίζεται ένα υπερδραστικό παράγωγο της Tn5 τρανζοπονάση όπου χρησιμποποιείται για να καταλύσει την in vitro ενσωμάτωση συνθετικών ολιγονουκλεοτιδίων στο επιθυμιτό DNA.Η τρανσποζάση Tn5 είναι μια αλληλουχία DNA που πλαισιώνεται από δύο ανεστραμμένα IS50 στοιχεία όπου το καθένα αποτελείται από μια αλληλουχία των 19 bp. Με την χρήση ενός τεχνικά μεταλαγμένου Tn5 όπου αυξάνεται η δραστικότητα και μεταβάλονται η αλληλουχίες των IS50 σε αλληλουχίες των adaptors μας έχει ως αποτέλεσμα τον κατακερματισμό και την τελική ένωση του συνθετικού adaptor στο 5 'άκρο του DNA.Στο rapid sequencing kit προτείνεται η εισαγωγή από 400 ng gDNA και είναι βελτιστοποιημένο για δείγματα που εμπεριέχουν μεγάλα κομμένα τμήματα DNA (>30 kb) . Σε περιπτώσεις όπου εισάγονται λιγότερα από 400 ng ή υπάρχει παρουσία τμημάτων μικρότερα απο 30 kb μπορεί να οδηγήσει σε μικρότερη απόδοση και μικρότερα reads.

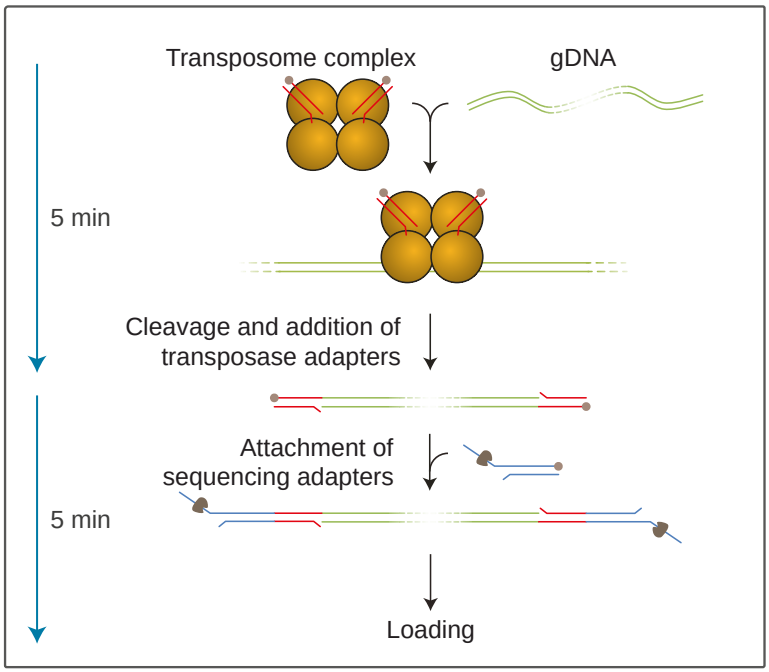


Figure 11: Προετοιμασία βιβλιοθήκης με την rapid μέθοδο πηγή : https://store.nanoporetech.com/us/rapid-sequencing-kit.html

2 Πειραματική διαδικασία

2.1 Μέθοδος

Ένας από τους ποιο σηαντικούς παράγοντες για μια de novo κατασκευή είναι η χρήση μέγαλης ποιότητας και μήκους reads για αυτόν τον λόγω στην εργασία αυτή θα γίνει μια προσπάθεια να μελετήσουμε τις διάφορες μεθόδους και παράγοντες όπου μπορούν να συμβάλουν σημαντικά στην ποιότητα των reads μας . Θα χρησιμοποιηθούν DNA από δυο διαφορετικές πηγές όπου η προετοιμασία βιβλιοθήκης έγινε με δυο διαφορετικούς τρόπους την rapid και ligation μέθοδο και αυτές οι βιβλιοθήκες στη συνέχεια έγιναν basecalled με τα δυο διαφορετικά μοντέλα του GUPPY (table 1). Κατά την διάρκεια αυτής της εργασίας αυτά τα δεδομένα επεξεργάστηκαν μέσω δυο διαφορετικών προγραμμάτων που αναφέρονται στο κεφάλαιο 2.3 και βάσει των αποτελεσμάτων τους

2.2 ΔΕΔΟΜΕΝΑ

Οι κλώνοι DNA που χρησιμοποιήθηκαν για την παραγωγή των δεδομένων αλληλουχίας προήλθαν από δυο βασικές πηγές , μια είναι του ανθρώπου και πιο συγκεκριμένα DNA από επιθηλιακό αμφιβληστροειδή (RPE) και DNA ωοθηκών από Cricetulus griseus γνωστό ως και Κινέζικο χάμστερ (CHO). Στον παρακάτω πίνακα αναπαριστώνται τα διάφορα δεδομένα που χρησιμοποιήθηκαν βάση ποια μέθοδο έγινε η προετοιμασία βιβλιοθήκης (LSK-109 μεθοδος ligation και RAD-004 για rapid μέθοδο) και το μοντέλο FAST/HAC όπου έγινε η διαδικασία του basecall.Επίσης είναι συχνά χρήσιμο να συνδυαστούν όλα τα διαφορετικά fastq αρχεία τα οποία δημιουργήθηκαν κατα τη διαδικασία του basecall σε ένα ενιαίο αρχείο ,όπως και στην δικιά μας περίπτωση και έτσι σε κάθε φάκελο ξεχωριστά μέσω terminal εκτελέστηκε η παρακάτω εντολή

```
cat *.fastq > all_guppy#.fastq
```

Table 2: Γενικές πληροφορίες για τα δεδομένα

Ονομασία των δεδομένων	Πηγή προέλευση των κυττάρων	Μέθοδος προετοιμασίας βιβλιοθήκης	Μοντέλο HAC/FAST
1) all_guppy1.fastq	Ωοθήκες από Cricetulus griseus	Rapid	High accuracy model
2) all_guppy2.fastq	Ωοθήκες από Cricetulus griseus	Ligation	High accuracy model
3) all_guppy3.fastq	DNA επιθηλιακού αμφιβληστροειδή	Ligation	FAST model
4) all_guppy4.fastq	DNA επιθηλιακού αμφιβληστροειδή	Ligation	High accuracy model
5) all_guppy5.fastq	DNA επιθηλιακού αμφιβληστροειδή	Ligation	High accuracy model

2.3 Προγράμματα που χρησιμοποιήθηκαν για quality check

2.3.1 FastQC

Το FastQC της Babraham Bioinformatics είναι ένα πολύ δημοφιλές εργαλείο που έχει γραφτεί σε JAVA και προσφέρει μια επισκόπηση ενός βασικού ποιοτικού ελέγχου (quality control)

των δεδομένων αλληλούχισης . Τα αποτελέσματα αφού αναλύσει τα fastq αρχεία είναι της μορφής html δηλαδή μπορούν να προβληθούν σε ένα πρόγραμμα περιήγησης. Το report αυτό περιέχει ένα αποτέλεσμα για κάθε μονάδα μέτρησης μαζί με την γραφική αναπαράσταση του καθώς και ένα δείκτη όπου στο κάθε αποτέλεσμα το οποίο το πρόγραμμα θεωρεί άμα αυτά είναι αποδεκτά ή όχι. Αυτοί οι δείκτες βασίζοται σε ένα πολύ συγκεκριμένο σύνολο παραδοχών που ισχύουν για έναν πολύ συγκεκριμένο τύπο δεδομένων αλληλούχισης. Άρα τα αποτελέσματα με τον δείκτη “Προσοχή” ή “Αποτυχία” πρέπει να σταματήσουμε και να σκεφτούμε τι σημαίνει αυτό το αποτέλεσμα στο πλαίσιο αυτού του συγκεκριμένου δείγματος και της μεθόδου που χρησιμοποιήθηκε. Τα αποτελέσματα που παράγει είναι τα εξής :

1. Per Base Sequence quality:Γραφική ποιότητας ανάγνωσης ανά βάση όπου στον y αξονα έχουμε το
2. Qscore.Η κόκκινη γραμμή είναι η median value.Το κίτρινο κουτάκι είναι το διατεταρτημοριακό εύρος (25-75%), οι αγκύλες τα 10 % και 90% και η μπλε γραμμή η μέση τιμή.
3. Per sequence quality scores
4. Per base content (γραφημα περιεκτηκότητας σε % ATGC ανά βάση αλληλουχιών)
5. Per sequence GC content (γράφημα περιεκτικότητας σε GC ανά read)
6. Per base N content (γράφημα περιεκτικότητας ποσότητας N)
7. Sequence Length Distribution (γράφημα κατανομής μήκους)
8. Sequence Duplication Levels (γράφημα επιπέδων διπλοτυπιών στα reads)
9. Overrepresented sequences (υπερβολική παρουσίαση αλληλουχιών)
10. Adapter Content

Η χρήση του είναι σχετικά απλή. Μετά την εγκατάσταση του προγράμματος τρέχοντας την εντολή fastqc στο terminal ξεκινά το πρόγραμμα στη συνέχεια επιλέγουμε το αρχείο και η διαδικασία του quality check ξεκινά.

2.3.2 Nanopack

Το NanoPack (https://github.com/wdecoster/NanoPlot) είναι ένα πακέτο αποτελεσματικών scripts γραμμένα βαση την γλώσσα προγραματισμού Python για οπτικοποίηση και επεξεργασία long read αλληλουχιών και είναι διαθέσιμο σε όλα τα βασικά λειτουργικά συστήματα. Η εγκατάσταση του μέσω απο το PYPI και bioconda από το τα δημόσια αποθετήρια είναι απλή και φροντίζει αυτόματα τις εξαρτήσεις στο λειτουργικό σύστημα. Τα εργαλεία για την κατασκευή των γραφικών είναι ευέλικτα και προσαρμόσιμα στις ανάγκες μας .Χρησιμοποιώντας μια μεμονωμένη εντολή NanoPlot ή NanoComp μπορεί να προετοιμαστεί μια πλήρης αναφορά σε html που περιέχει συνοπτικά όλα τα στατιστικά στοιχεία και γραφήματα.

Αρχικά πριν την εγκατάσταση για να λειτουργήσει το πρόγραμμα πρέπει να δημιουργηθεί ένα python περιβάλλον(python enviroment). Τα βήματα που ακολουθήσαμε είναι τα εξής στο terminal:

1)εγκατάσταση των εργαλίων virtualenv μέσω της εντολής

```
sudo apt install virtualenv
```

2) Δημιουργία ενός κατάλογου(directory) python-environments και πλοήγηση σε αυτόν

```
mkdir ~/python-environments && cd ~/python-environments
```

3) Δημιουργία του περιβάλλοντος

```
virtualenv --python=python3 env
```

4) ενεργοποίηση του περιβάλλοντος

```
source env/bin/activate
```

Στην συνέχεια μετά την εγκατάσταση του προγράμματος, για κάθε αρχείο μέσω της εντολής:

NanoPlot -t 2 --fastq [όνομα αρχείου].fastq

εχουμε την παραγωγή μιας στατιστικης σύνοψης σε ένα αρχείο .txt , έναν αριθμό γραφικών και ένα html αρχείο με μια σύνοψη των αποτελεσμάτων.

2.4 Αποτελέσματα και Παρατηρήσεις

2.4.1 Αποτελέσματαν για τις διαφορετικές μεθόδους library preparation (rapid-ligation)

FASTQC:

Per base Sequence quality :

Αρχικά βλέπουμε ότι στην αρχή του διαγράμματος μας έχουμε και στις δυο περιπτώσεις χαμηλό Qscore το οποιο αρχίζει να αυξάνεται. Με την ligation μέθοδο το Qscore φτάνει μέχρι την τιμή 19 δηλαδή 1.3 % πιθανότητα για σφάλμα και στην συνέχεια πέφτει στο 14 δηλαδή 3.9% πιθανότητα για σφάλμα και όπου εκεί σταθεροποιείται . Η rapid αυξάνεται μέχρι την μέγιστη τιμή Qscore =18 (1.6 %) και όπου εκεί σταθεροποιείται.

Per sequence quality scores:

Εδώ αναπαριστόνται οι γραφικές παραστάσεις του συνολικού αριθμού των reads έναντι του μέσου όρου του quality score σε όλο το μήκος των read. Η κατανομή του quality score πρέπει να είναι αρκετά στενη στο ανώτερο εύρος της της γραφικής. Βάση των δυο γραφικών μας

παρατηρούμε ότι η ligation κατανέμεται γύρω από την τιμή 20 του Qscore το οποίο είναι λίγο μεγαλύτερο μεγαλύτερο από την αντίστοιχη της rapid μεθόδου που κατανέμεται γύρω από την τιμή 18. Η πιο σημαντική διαφορά που παρατηρείται είναι ότι με την ligation το μεγαλύτερο μέρος των reads που αλληλουχήθηκαν έχουν καλύτερο Qscore από αυτά της rapid αφού η γκαουσιανή που σχηματίζεται είναι αρκετά στενότερη σε μεγαλύτερο Qscore.

Per base sequence content:

Αυτα τα διαγράμματα αναφέρονται στο ποσοστό των βάσεων που έχουν μετρηθεί για καθένα από τα τέσσερα νουκλεοτίδια σε κάθε θέση για όλα τα reads στα αρχεία.Για ολόκληρο το γονιδίωμα η αναλογία καθε μίας από τις τέσσερις βάσεις θα πρέπει να παραμείνει σχετικά σταθερή κατά το μήκος των reads με %A=%T και %G=%C. Εδώ παρατηρούμε ότι τα αποτελέσματα από την rapid διατηρούν λίγο κσλύτερα την αναμενόμενη αναλογία.

Per Base N Content:

Εάν μια συσκευή αλληλούχησης (sequencer) δεν είναι σε θέση να καλέσει μια βάση με επαρκή εμπιστοσύνη, τότε θα αντικαταστήσει με N παρά μια συμβατική βάση ATGC. Σε αυτά τα γραφήματα απεικονίζεται το ποσοστό των basecalls σε κάθε θέση για την οποία έγινε basecalled ένα N. Στα δικά μας αποτελέσματα παρατηρούμε ότι σε κανένα read είτε της ligation είτε της rapid μεθόδου δεν έγινε basecalled κανένα N.

Per sequence GC content:

Εδώ έχουμε τα διαγράμματα των αριθμό των reads έναντι του GC% ανά κάθε read και την θεωρητική κατανομή που προϋποθέτει μια ομοιόμορφη αναλογία GC σε όλα τα reads. Όπως φάνηκε και στο per base sequence content με την rapid μέθοδο τα reads που πήραμε έχουν την σωστή αναλογία που αναμένουμε να έχουν και για αυτό η καμπύλη σχεδόν ταυτίζεται με την θεωριτική σε αντίθεση με τα reads της ligation μεθόδου που έχουν μια εμφανή απόκλιση από την θεωριτική.

Sequence duplication levels:

Σε αυτές τις γραφικές αναπαριστάται το ποσοστό των reads μιας δεδομένης ακολουθίας , τα οποία υπάρχουν πολλές φορές στο αρχείο.Υπάρχουν γενικά δύο πηγές για αυτά τα αντίφραφα reads, πολλαπλασιασμός μέσω της αλυσιδωτή αντίδραση πολυμεράσης (PCR) στην οποία τα θραύσματα της βιβλιοθήκης έχουν αναπαριστηθεί υπερβολικά λόγω του υπερβολικού εμπλουτισμού ή υπερβολική εκπροσώπηση όμοιων αλληλουχιών. Το πρώτο είναι ανησυχητικό, επειδή τα αντίγραφα πού έχουν δημιουργηθεί μέσω της PCR εσφαλμένα αναπαριστούν την πραγματική αναλογία ακολουθιών στο αρχικό μας υλικό.Το τελευταίο είναι μια αναμενόμενη περίπτωση και δεν προκαλεί ανησυχία γιατί όπως έχουμε αναφερθεί υπάρχουν πολλές αλληλουχίες που επαναλαβάνονται συχνά στο γονιδίωμα. Στα αποτελέσματα για να μειωθούν τις απαιτήσεις για μνήμη στον υπολογιστή αναλύονται μόνο ακολουθίες που εμφανίζονται για πρώτη φορά στις 100.000 ακολουθίες σε κάθε αρχείο.Για την μείωση της ποσότητα των πληροφοριών στο τελικό διάγραμμα, τυχόν αλληλουχίες με περισσότερα από 10 αντίγραφα τοποθετούνται σε ομαδοποιημένους “κάδους” για να δώση μια σαφή εντύπωση του συνολικού επιπέδου αναπαραγωγής του κάθε αντίγραφου χωρίς να χρειάζεται να δείξει κάθε μεμονωμένη τιμή

αναπαραγωγής. Σε διάφορες περιπτώσεις είναι πολυ πιθανό να σχηματίζονται κορυφές χωρίς να καθορίζουν ότι υπάρχουν ανησυχητικά υψηλές αναπαραστάσεις των ίδιων ακολουθιών. Άμα αυτές οι κορυφές επιμείνουν σημαίνει ότι υπάρχει ένας υπερβολικά μεγάλος αριθμός ίδιων αλληλουχιών που μπορεί να υποδηλώνουν είτε ότι έχει γίνει λαθός με το υλικό που αλληλουχήθηκε είτε έχει γίνει λάθος με την PCR. Στα δικά μας αποτελέσματα δεν παρατηρούμε καμία κορυφή.

Nanopack:

Αποτελέσματα από την βιβλιοθήκη που προετοιμαστηκε μέσω του rapid sequencing kit.

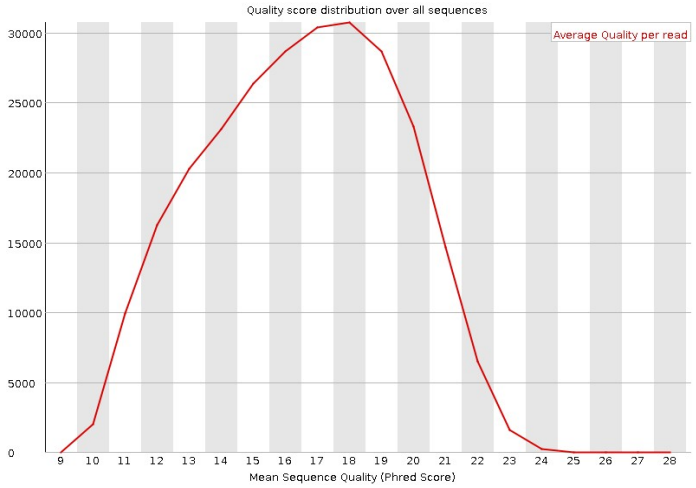
Table 3: Αποτελέσματα για rapid μεθοδο

Μέσο μήκος των reads	3,240.4
Μέσο Qscore	9.7
Median μήκος των reads	1396
Median read Qscore	9.7
Αριθμός των reads	263060
N50	7536
τυπική απόκλιση του μήκους των read	4486.8
Συνολικός αριθμος βασεων	852418457

Table 4: Ποσοστιαία κατανομή των reads με την rapid μεθοδο

	Αριθμός	Ποσοστό
Q>5	263060	100%
Q>7	263060	100%
Q>10	110859	42%
Q>12	15036	6%
Q>15	5	0%

Figure 12: FASTQC per sequence quality



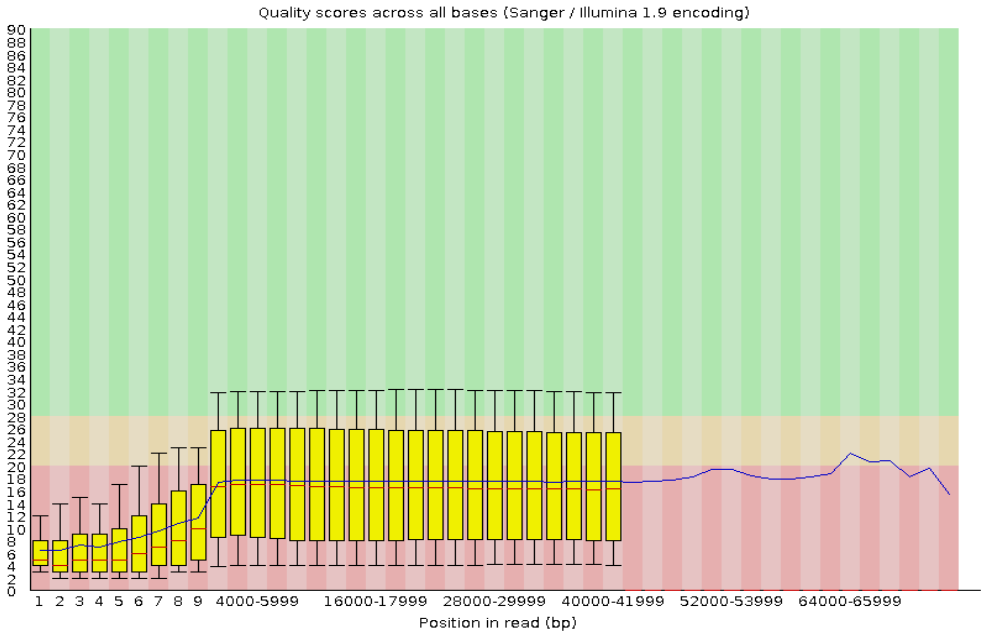


Figure 13: FASTQC per base quality

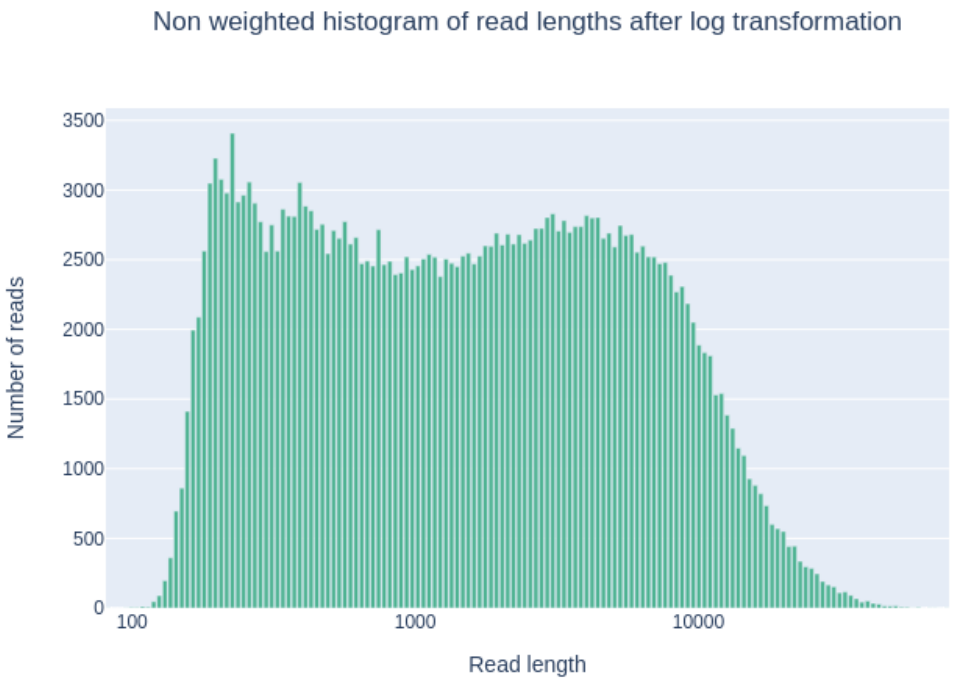


Figure 14: Nanopack κατανομή των reads μετά απο λογαριθμική μετατροπή

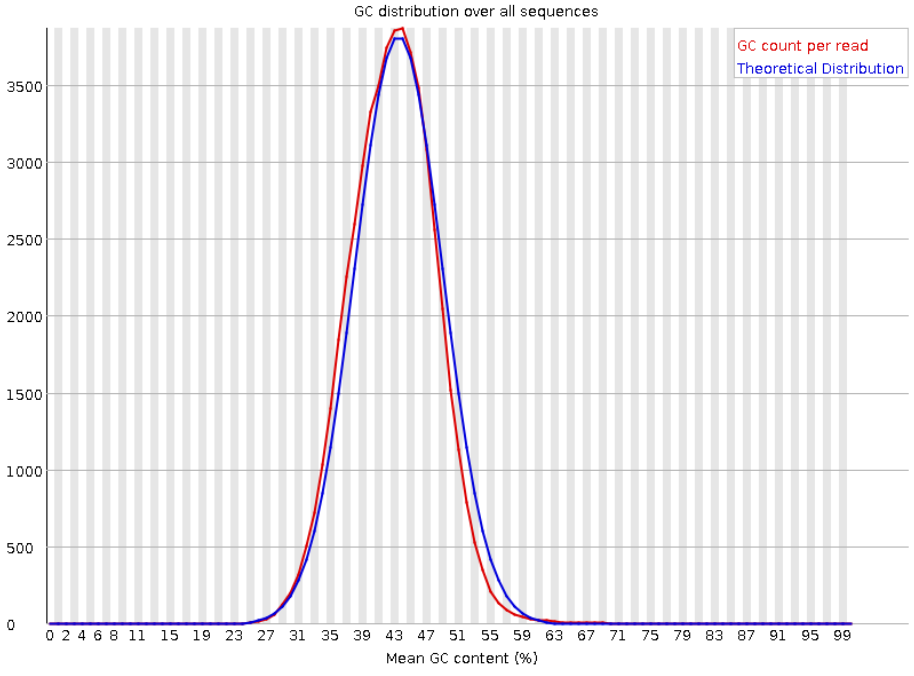


Figure 15: FASTQC per sequence GC content

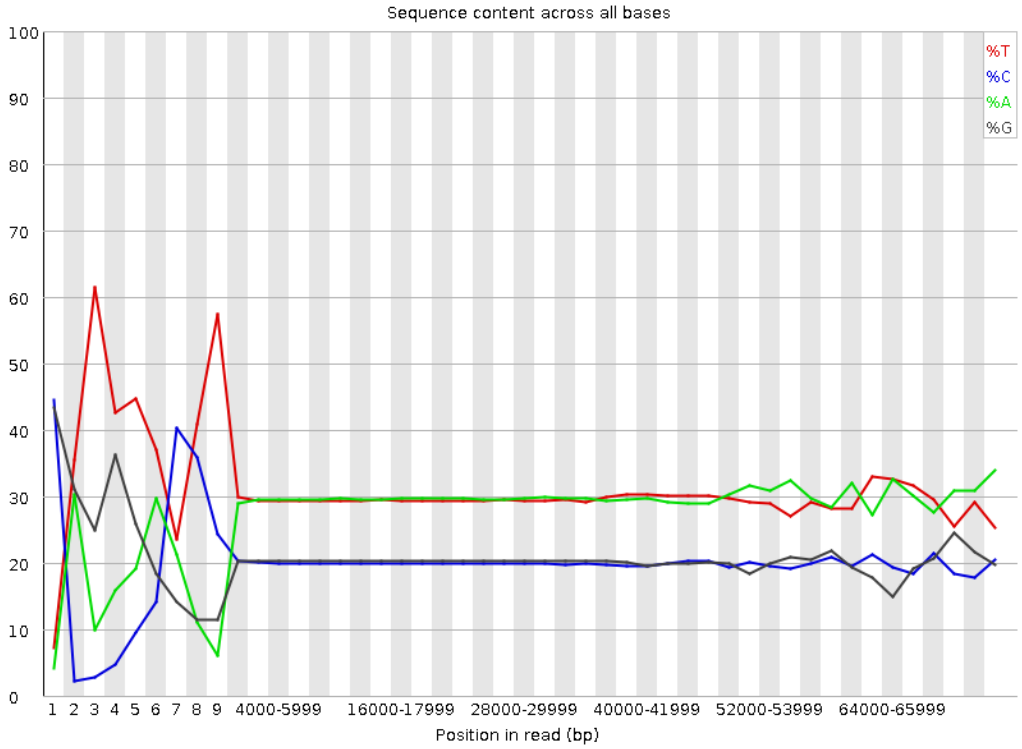


Figure 16: FASTQC ber base sequence content

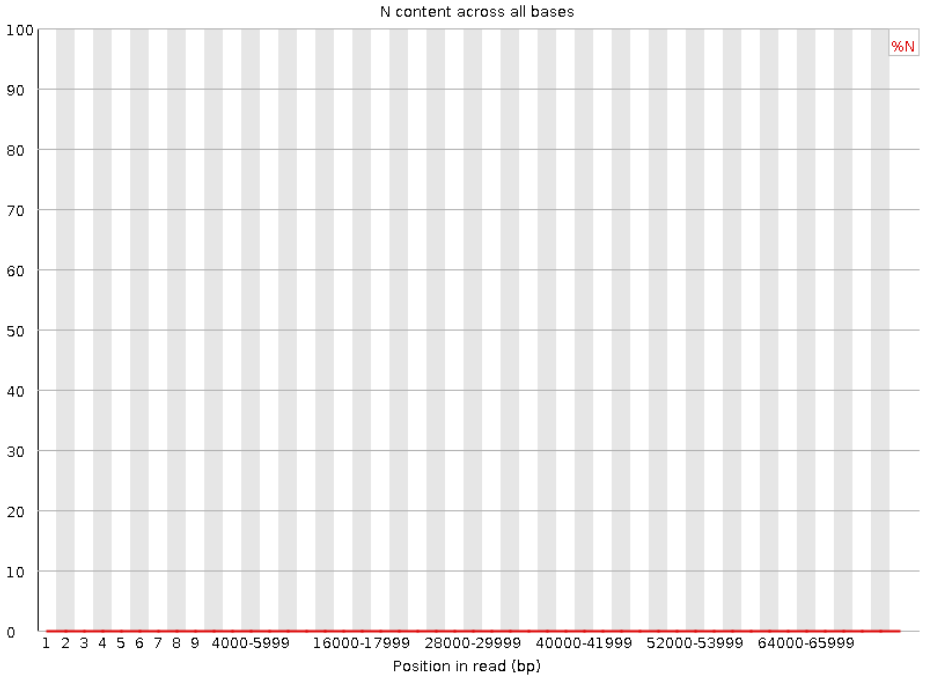


Figure 17: FASTQC N content

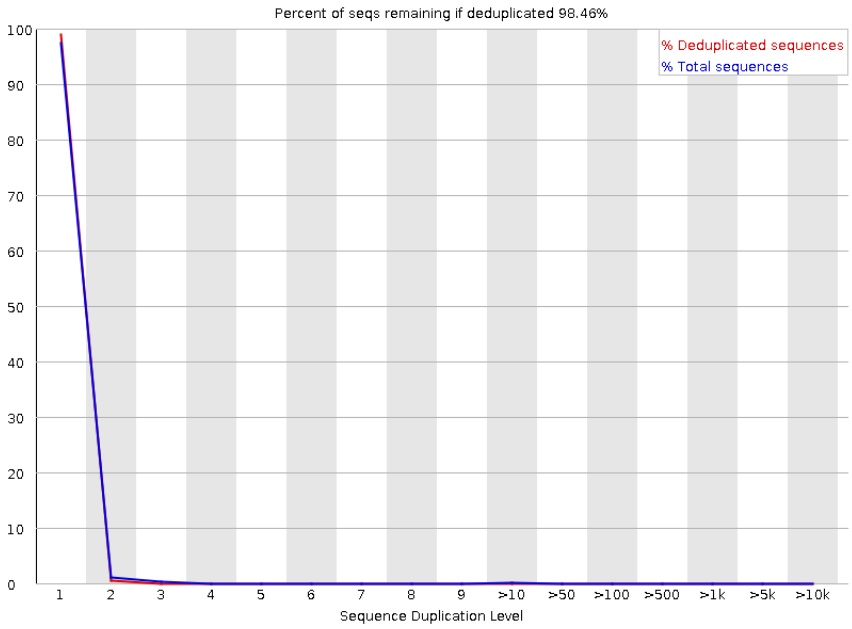


Figure 18: FASTQC duplication levels

Αποτελέσματα από την βιβλιοθήκη που προετοιμαστηκε μέσω του ligation sequencing kit.

Table 5: Αποτελέσματα για ligation μεθοδο

Μέσο μήκος των reads	5,372.2
Μέσο Qscore	10.4
Median μήκος των reads	4578
Median read Qscore	10.7
Αριθμός των reads	764000
N50	8350
τυπική απόκλιση του μήκους των read	4455.8
Συνολικός αριθμος βασεων	4104390261

Table 6: Ποσοστιαία κατανομή των reads με την ligation μεθοδο

	Αριθμός	Ποσοστό
Q>5	764000	100%
Q>7	764000	100%
Q>10	479190	63%
Q>12	141232	19%
Q>15	212	0%

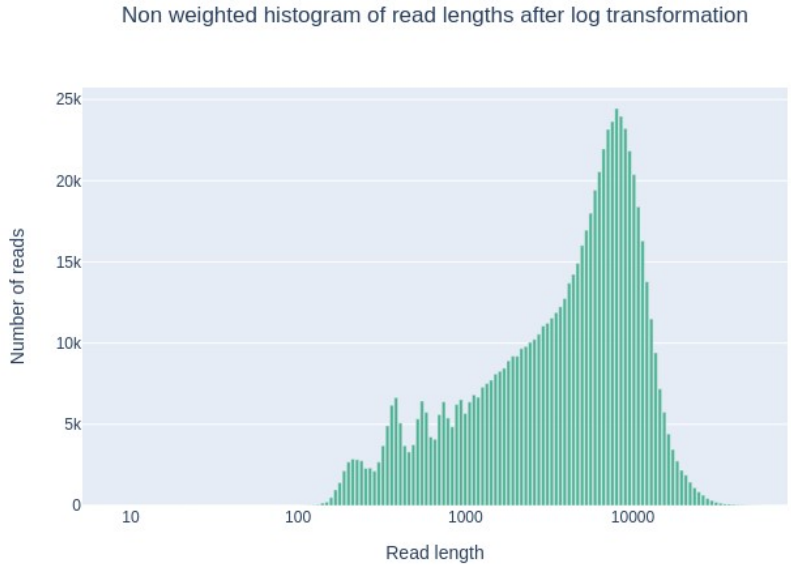


Figure 19: Nanopack κατανομή των reads μετά απο λογαριθμική μετατροπή

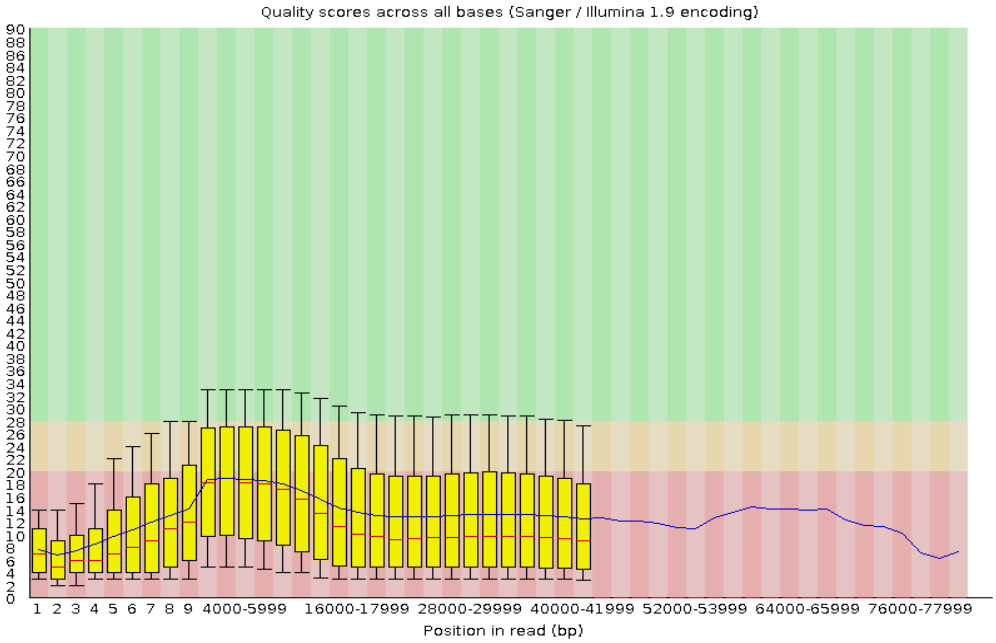


Figure 20: FASTQC per base quality

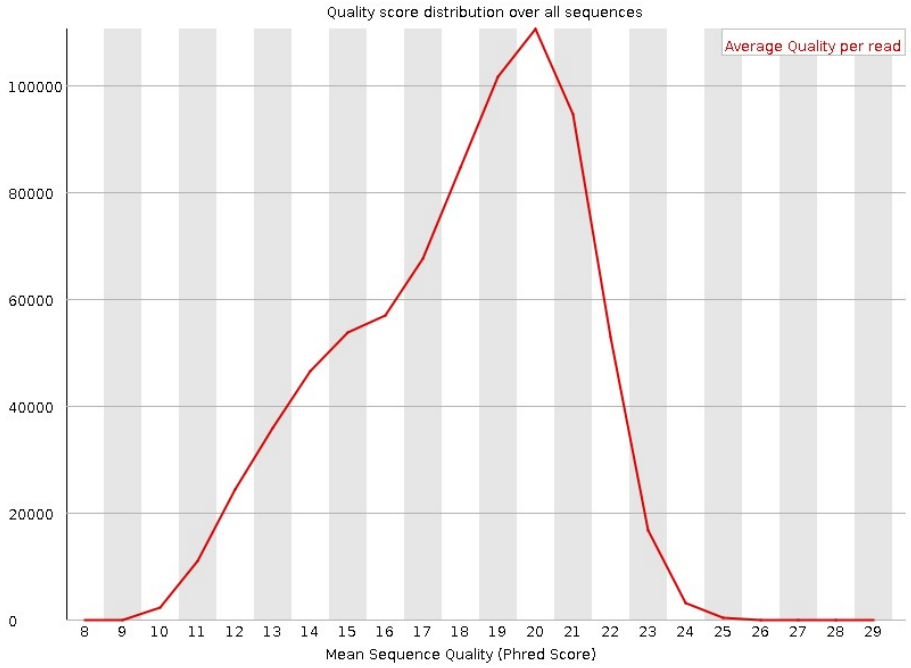


Figure 21: FASTQC per sequence quality

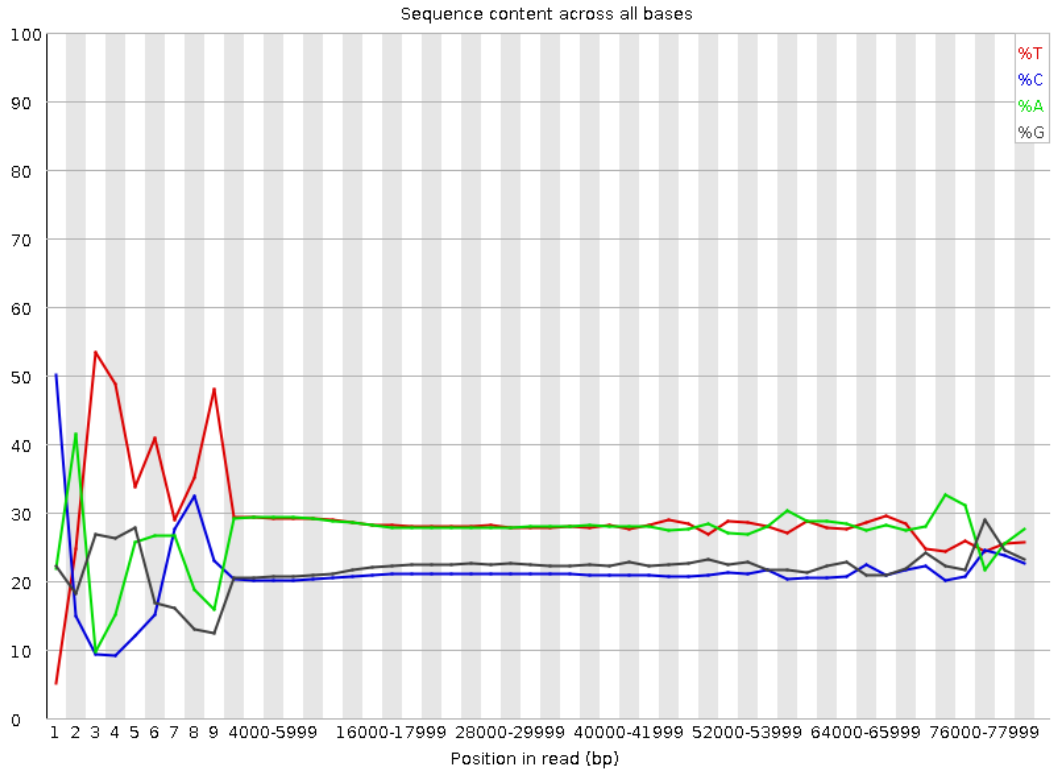


Figure 22: FASTQC ber base sequence content

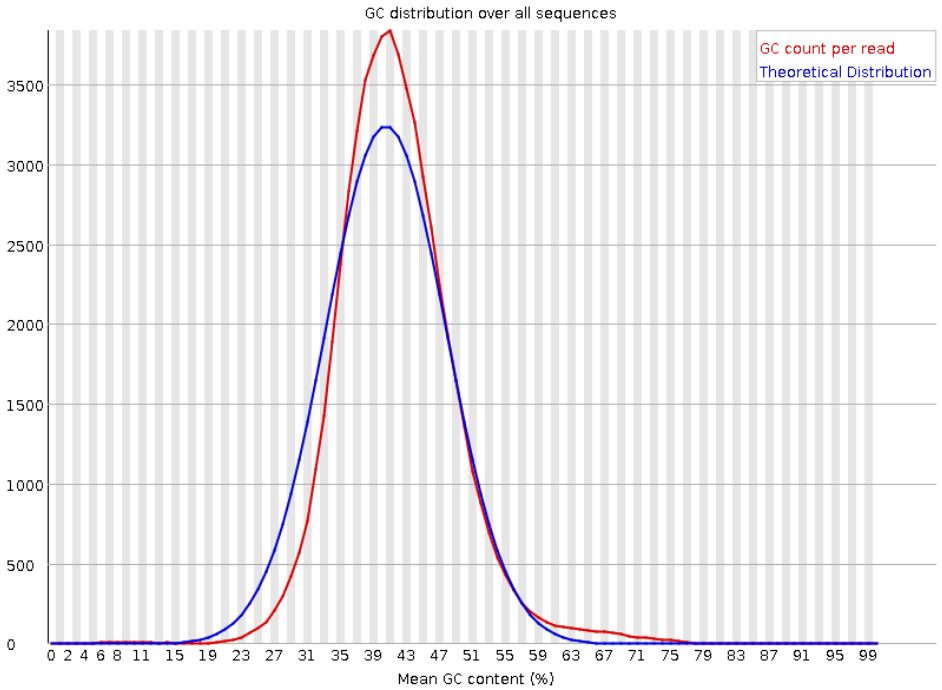


Figure 23: FASTQC per sequence GC content

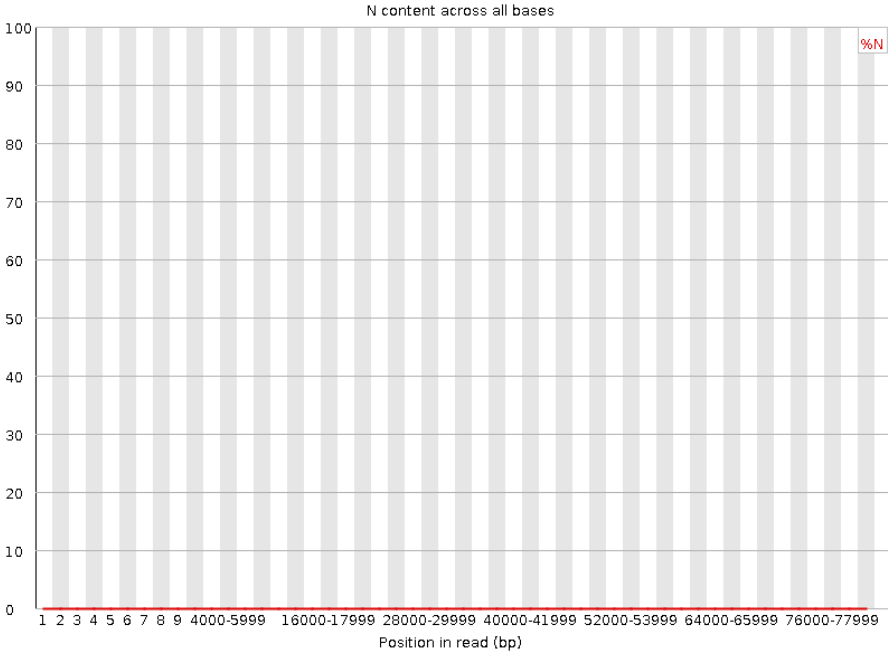


Figure 24: FASTQC N content

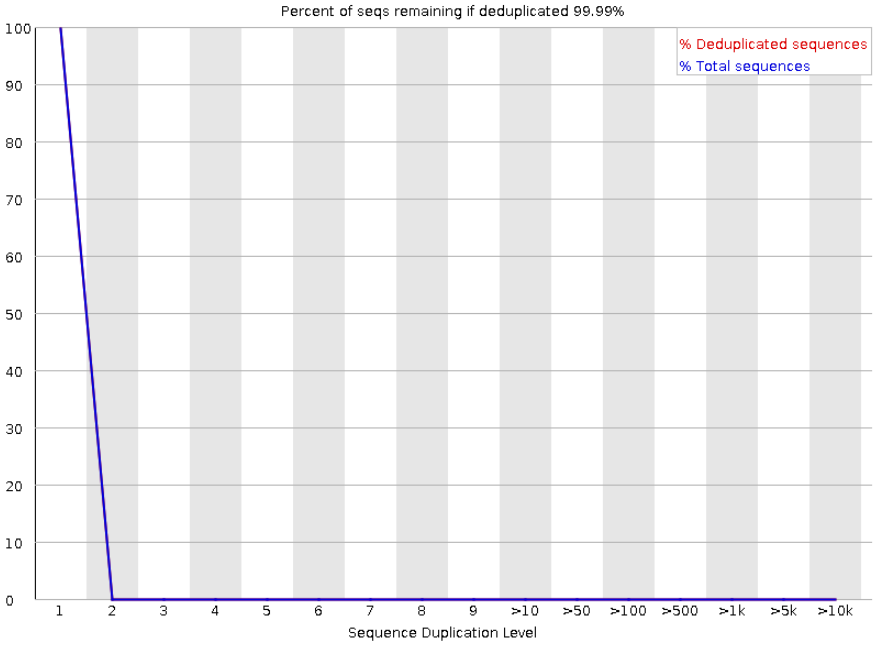


Figure 25: FASTQC duplication levels

2.4.2 Αποτελέσματα για τα δυο βασικά μοντέλα του Guppy (high accuracy model – fast model)

FASTQC:

Per base Sequence quality:

Αρχικά βλέπουμε την αναμενόμενη αύξηση του Qscore και στα δυο διαγράμματα . Με την HAC μέθοδο το Qscore φτάνει και σταθεροποιείται στην τιμή 22 δηλαδή 0.6 % πιθανότητα για σφάλμα δηλαδή να καλεστεί λάθος μια βάση . Η FAST αυξάνεται μέχρι την μέγιστη τιμή Qscore =20 (1%) και όπου εκεί σταθεροποιείται.

Per sequence quality scores:

Βάση των δυο γραφικών μας παρατηρούμε ότι η HAC κατανέμεται γύρω από την τιμή 24 του Qscore το οποίο είναι λίγο μεγαλύτερο από την αντίστοιχη της FAST που κατανέμεται γύρω από την τιμή 22. Η ουσιαστική διαφορά που παρατηρείται μέσω της γκαουσιανή που σχηματίζεται είναι ότι με την HAC τα περισσότερα που αλληλουχήθηκαν reads έχουν καλύτερο Qscore από αυτά της FAST

Per base sequence content:

Τα αποτελέσματα της HAC εδώ είναι καλύτερα από αυτά της FAST αφού η θεωρητικά αναμενόμενη αναλογία παραμένει σταθερή για μεγαλύτερα reads

Per Base N Content:

Σε καμία μέθοδο δεν έγινε basecalled N

Per sequence GC content:

Σε αυτές τις γραφικές δεν παρατηρούμε καμία ουσιαστική διαφορά

Sequence duplication levels:

Στα αποτελέσματα δεν παρατηρούμε καμία κορυφή.

Nanopack:

Αποτελέσματα από τα δεδομένα που το basecall έγινε βάση του high accuracy model

Table 7: Αποτελέσματα για HAC μοντέλο

Μέσο μήκος των reads	8,706.3
Μέσο Qscore	12
Median μήκος των reads	8228
Median read Qscore	12.7
Αριθμός των reads	2171810
N50	11172
τυπική απόκλιση του μήκους των read	5775.3
Συνολικός αριθμος βασεων	18908398659

Table 8: Ποσοστιαία κατανομή των reads με το HAC μοντέλο

	Αριθμός	Ποσοστά
Q>5	2094955	96.5%
Q>7	1949483	89.8%
Q>10	1688163	77.7%
Q>12	1302755	60.0%
Q>15	294004	13.5%

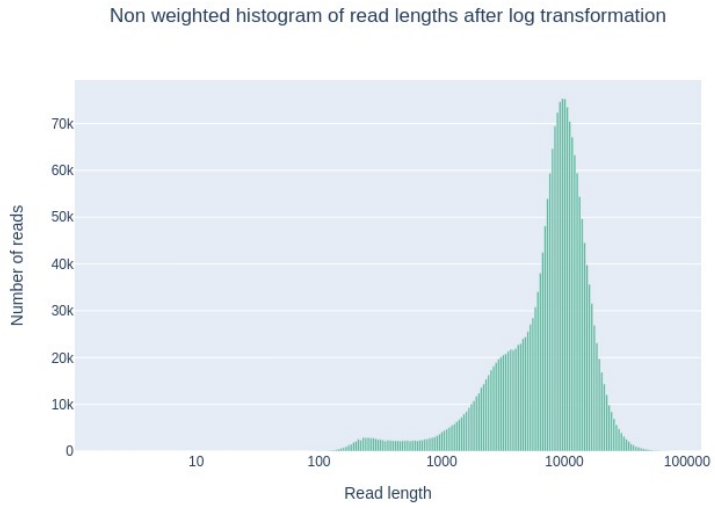


Figure 26: Nanopack κατανομή των reads μετά απο λογαριθμική μετατροπή

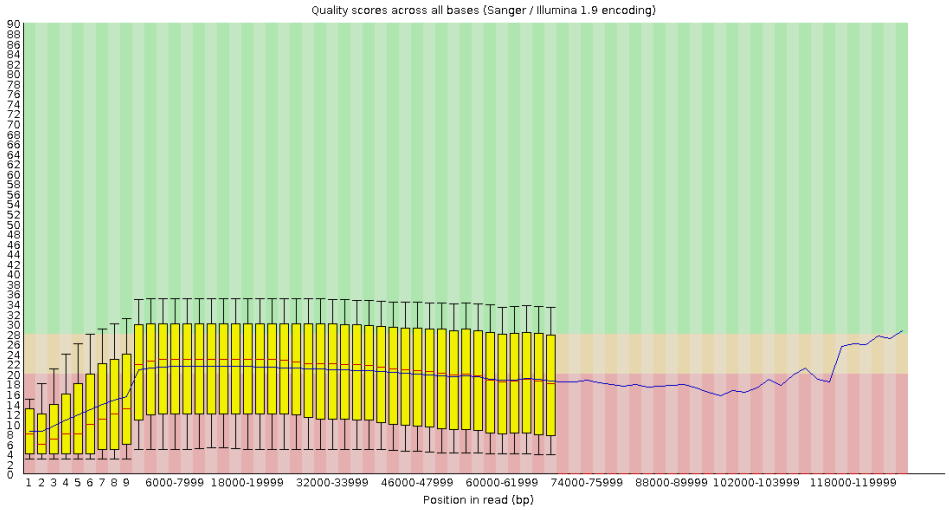


Figure 27: FASTQC per base quality

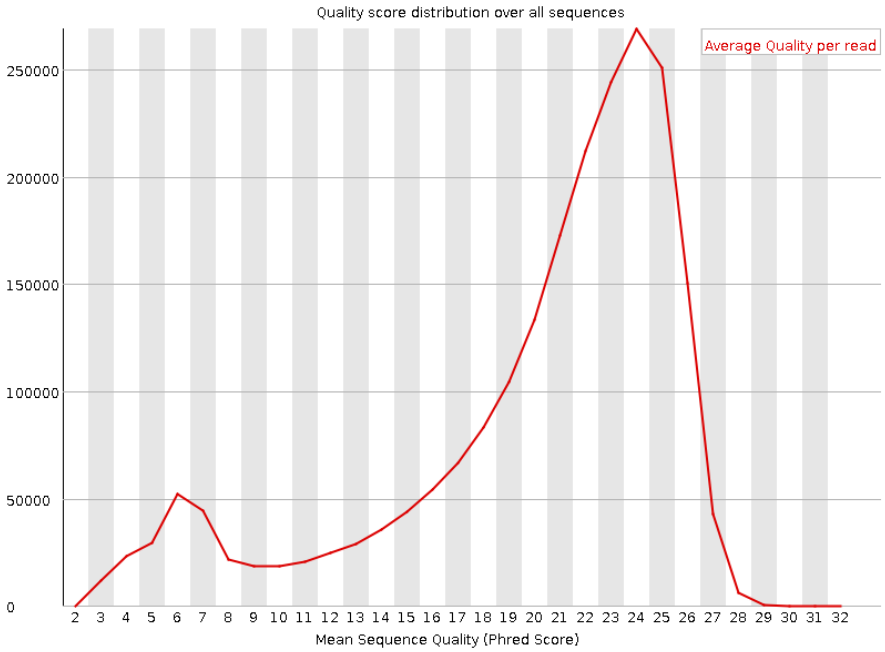


Figure 28: FASTQC per sequence quality

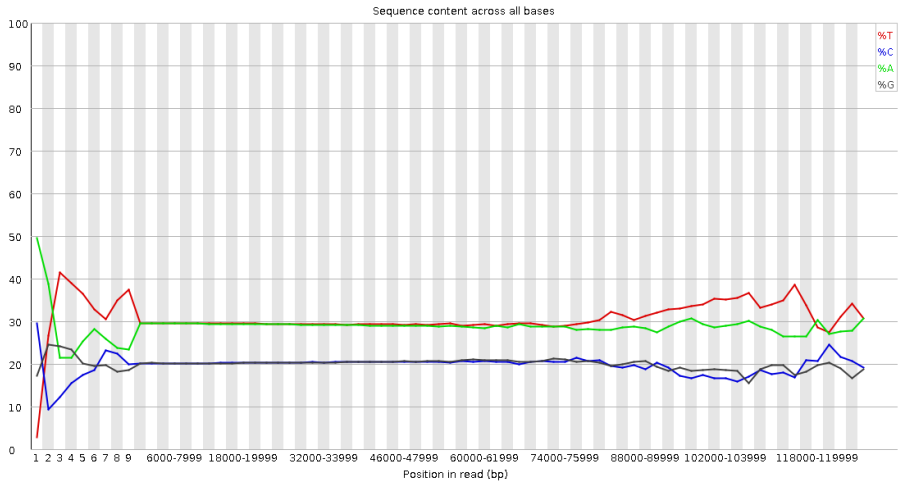


Figure 29: FASTQC ber base sequence content

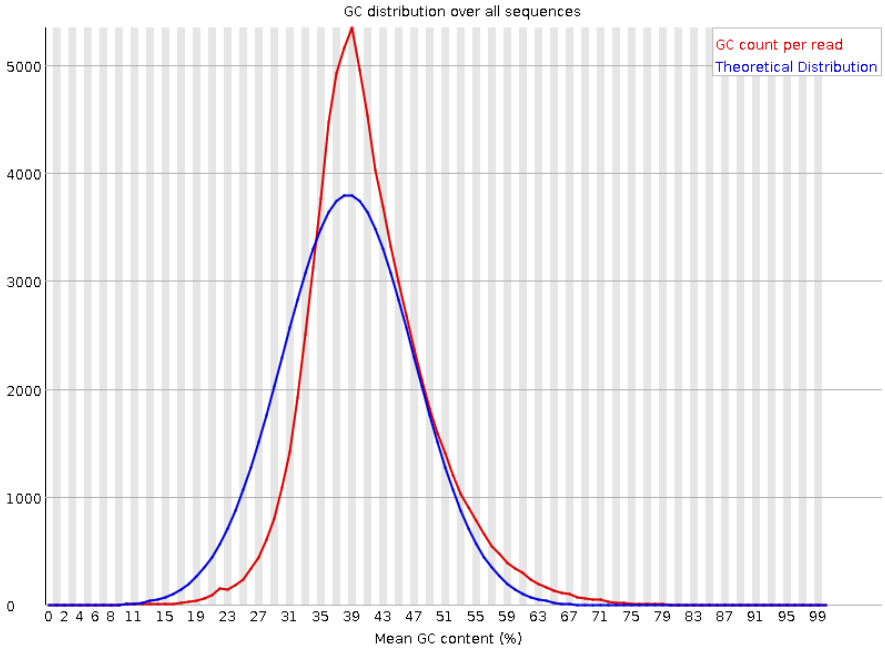


Figure 30: FASTQC per sequence GC content

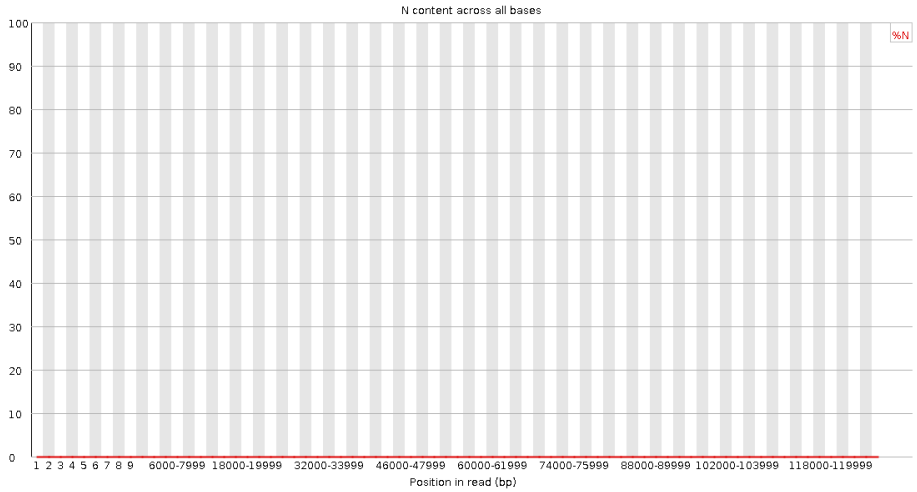


Figure 31: FASTQC N content

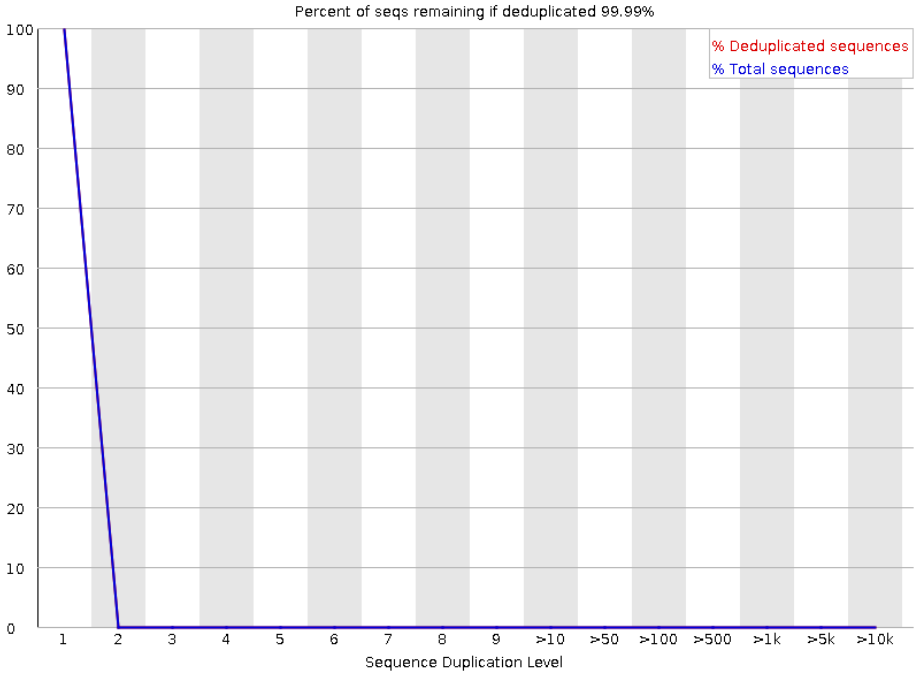


Figure 32: FASTQC duplication levels

Αποτελέσματα από τα δεδομένα που το basecall έγινε βάση του FAST model

Table 9: Αποτελέσματα για FAST μοντέλο

Μέσο μήκος των reads	8,608.5
Μέσο Qscore	10.4
Median μήκος των reads	8152
Median read Qscore	11.1
Αριθμός των reads	2000856
N50	11079
τυπική απόκλιση του μήκους των read	5717.9
Συνολικός αριθμος βασεων	17224439617

Table 10: Ποσοστιαία κατανομή των reads με το FAST μοντέλο

	Αριθμός	Ποσοστά
Q>5	1903295	95.1%
Q>7	1749615	87.4%
Q>10	1334760	66.7%
Q>12	632274	31.6%
Q>15	682	0.0%

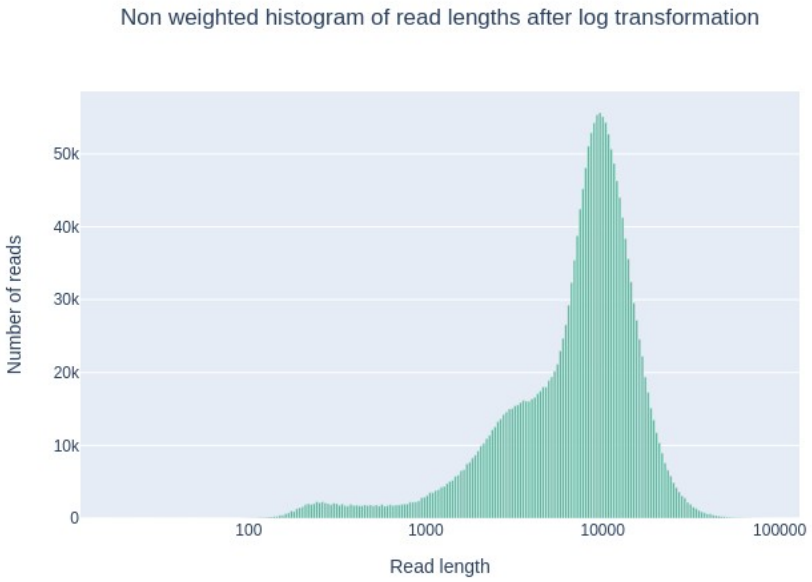


Figure 33: Nanopack κατανομή των reads μετά απο λογαριθμική μετατροπή

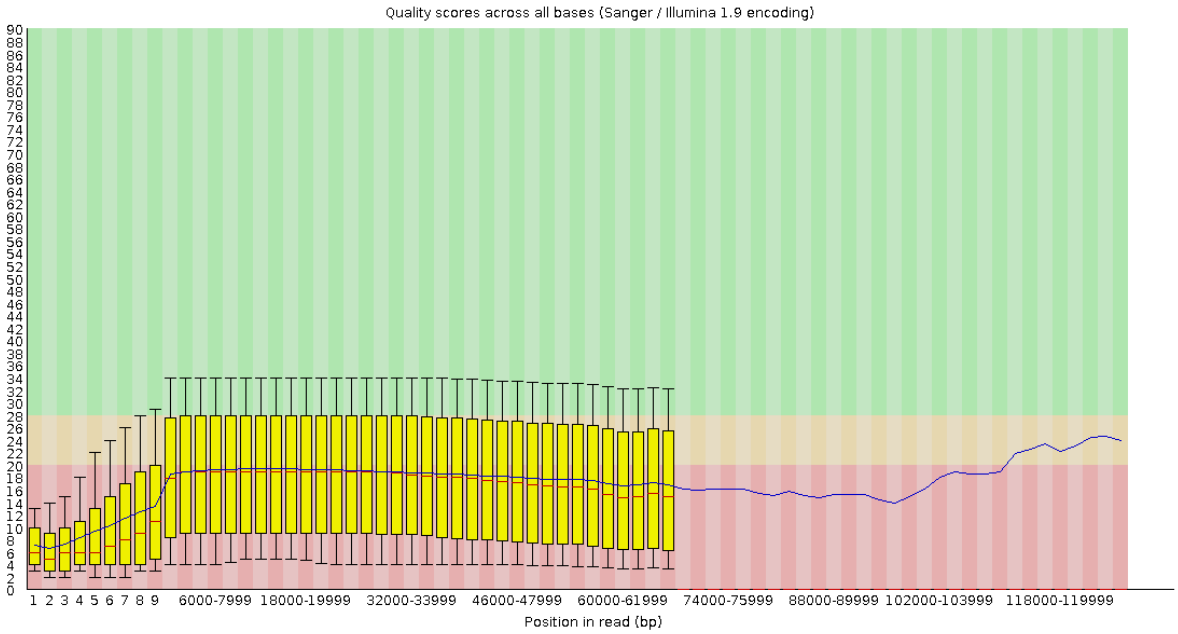


Figure 34: FASTQC per base quality

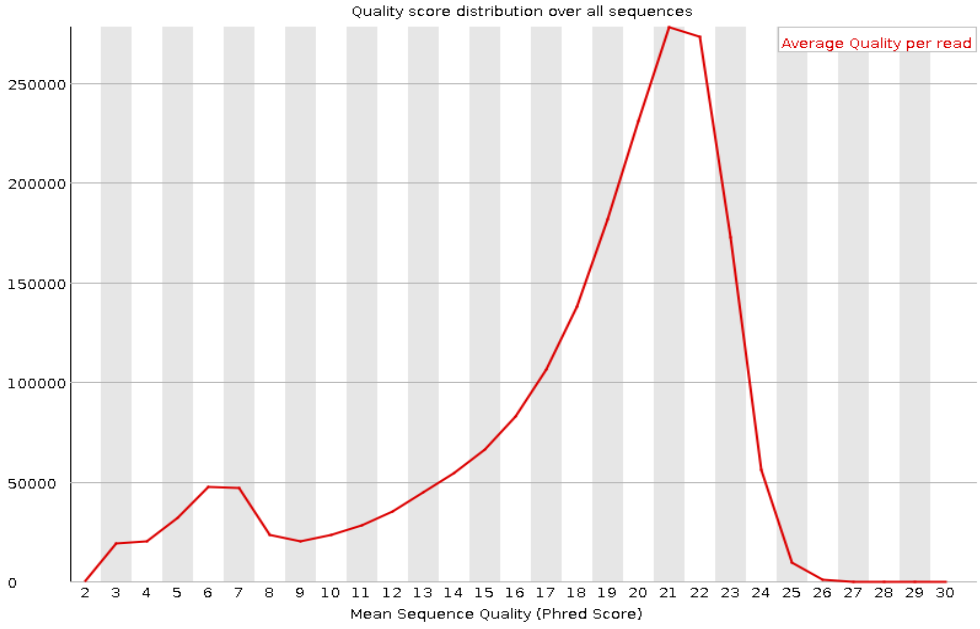


Figure 35: FASTQC per sequence quality

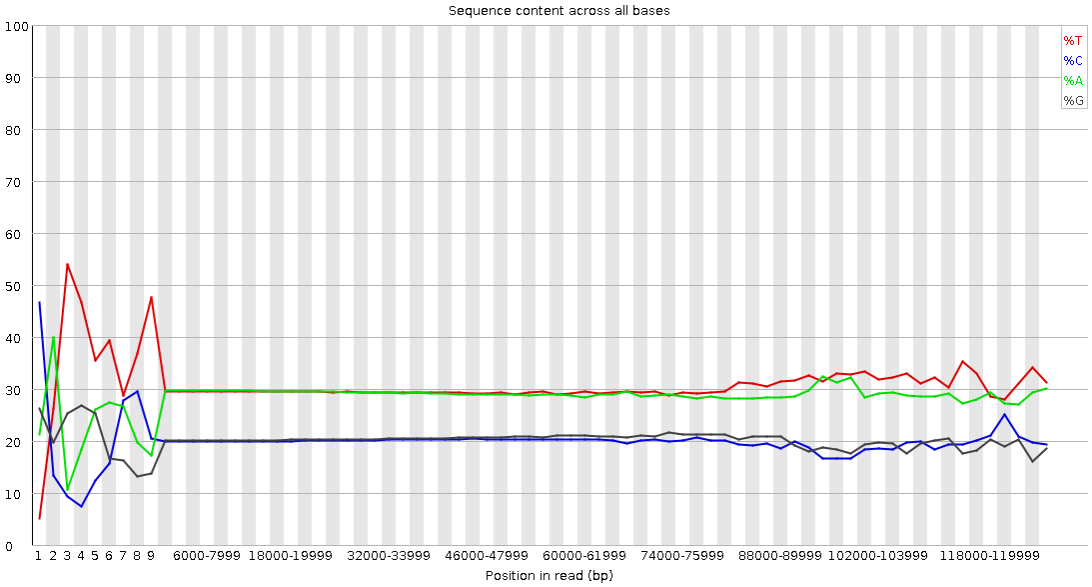


Figure 36: FASTQC ber base sequence content

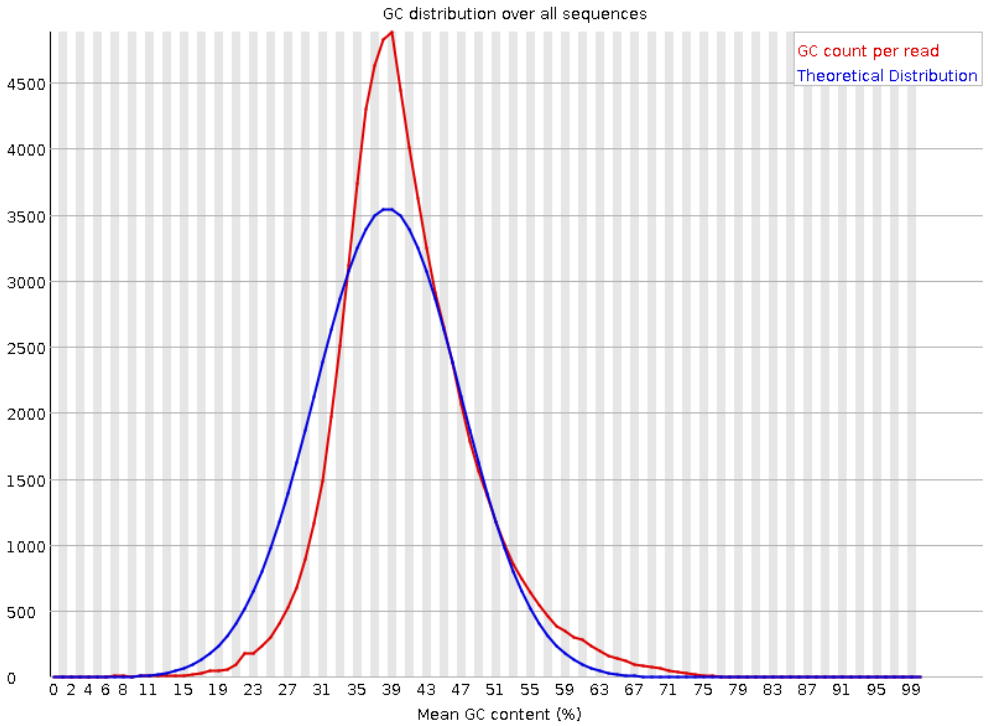


Figure 37: FASTQC per sequence GC content

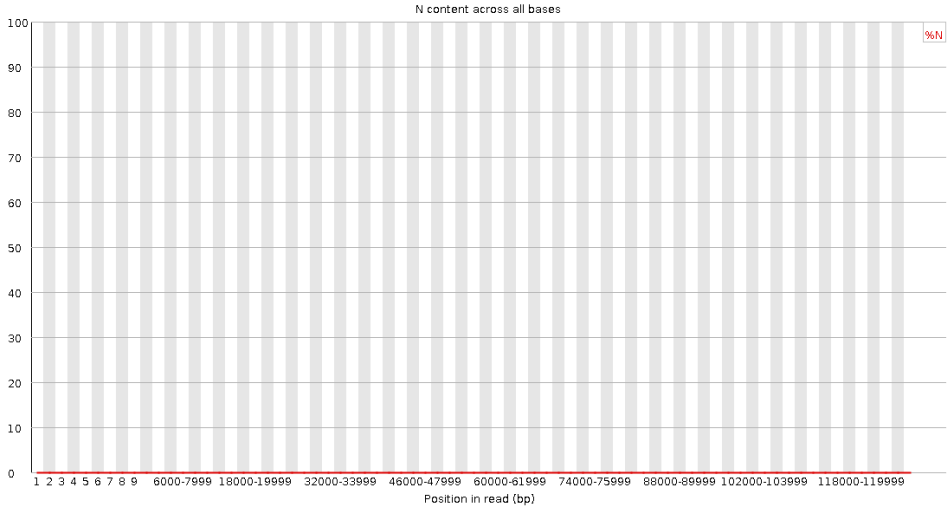


Figure 38: FASTQC N content

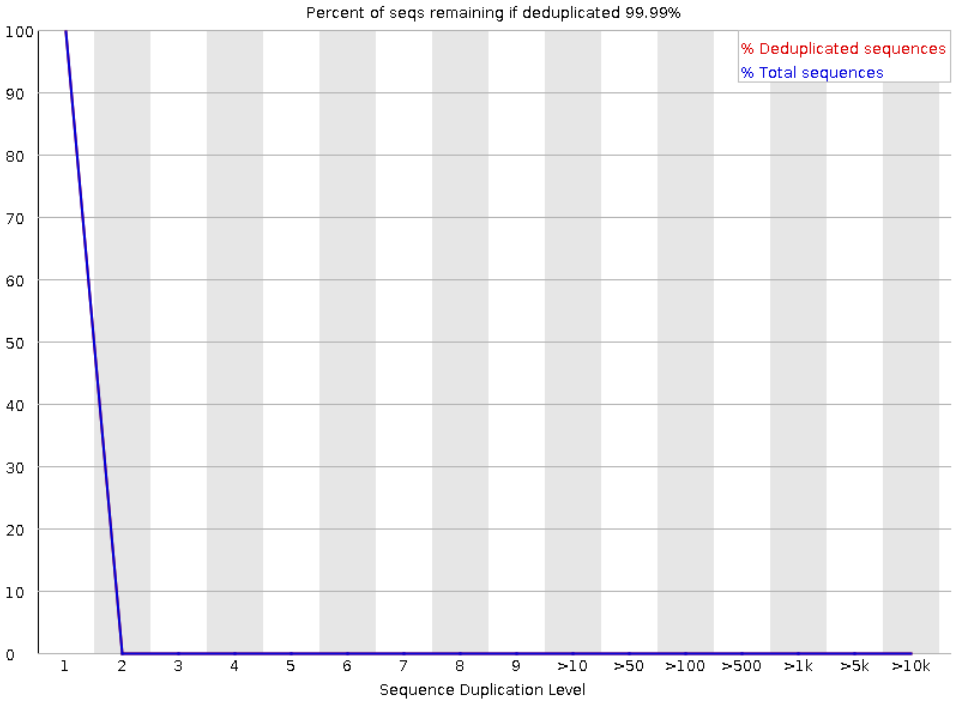


Figure 39: FASTQC duplication levels

2.4.3 Αποτελέσματα για το ανθρώπινο ΔΝΑ και παρατηρήσεις ανάμεσα στις πηγές όπου πήραμε το πρότυπο DNA(RPE-CHO)

FASTQC:

Per base Sequence quality:

Αρχικά βλέπουμε την αναμενόμενη αύξηση του Qscore και στα δυο διαγράμματα .Βάση τα αποτελέσματα του 840 πειράματος δηλαδή του ανθρώπινου DNA το Qscore φτάνει και σταθεροποιείται στην τιμή 20 που αντισοιχεί σε 1 % πιθανότητα για σφάλμα δηλαδή να καλεστεί λάθος μια βάση. Τα αποτελέσματα για DNA χάμστερ την το Qscore φτάνει μέχρι την τιμή 19 δηλαδή 1.3 % πιθανότητα για σφάλμα και στην συνέχεια πέφτει στο 14 δηλαδή 3.9% πιθανότητα για σφάλμα και όπου εκεί σταθεροποιείται

Per sequence quality scores:

Βάση των δυο γραφικών μας παρατηρούμε ότι το DNA ανθρώπου κατανέμεται γύρω από την τιμή 23 του Qscore το οποίο είναι αρκετά μεγαλύτερο από την αντίστοιχη του χάμαστερ που κατανέμεται γύρω από την τιμή 20. Η ουσιαστική διαφορά που παρατηρείται μέσω της γκαουσιανή που σχηματίζεται είναι ότι του ανθρώπου τα περισσότερα reads που αλληλουχήθηκαν έχουν καλύτερο Qscore από αυτά του χάμστερ.

Per base sequence content:

Τα αποτελέσματα του ανθρώπινου DNA είναι καλύτερα από αυτά του χάμστερ αφού η θεωρητικά αναμενόμενη αναλογία παραμένει σταθερή στο ανθρώπινο DNA σε αντίθεση με του χάμστερ όπου αυτή η αναλογία δεν παραμένει σταθερή

Per Base N Content:

Σε καμία μέθοδο δεν έγινε basecalled N

Per sequence GC content:

Σε αυτές τις γραφικές δεν παρατηρούμε καμία ουσιαστική διαφορά

Sequence duplication levels:

Στα αποτελέσματα δεν παρατηρούμε καμία κορυφή.

Nanopack:

Table 11: Αποτελέσματα για RPE DNA

Μέσο μήκος των reads	8,946.8
Μέσο Qscore	11.5
Median μήκος των reads	7681
Median read Qscore	12.1
Αριθμός των reads	2372705
N50	13108
τυπική απόκλιση του μήκους των read	7582.4
Συνολικός αριθμος βασεων	21228128837

Table 12: Ποσοστιαία κατανομή των reads με RPE DNA

	Αριθμός	Ποσοστά
Q>5	2322212	97.9%
Q>7	2195243	92.5%
Q>10	1775926	74.8%
Q>12	1210766	51.0%
Q>15	109596	4.6%

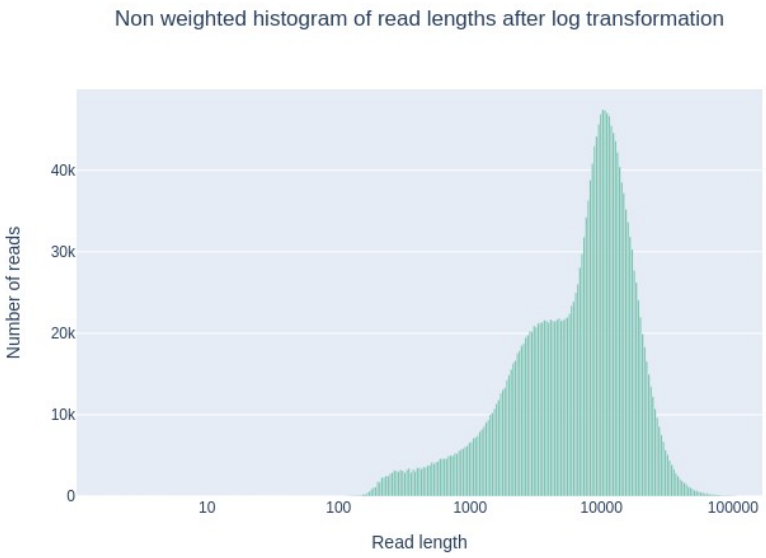


Figure 40: Nanopack κατανομή των reads μετά απο λογαριθμική μετατροπή

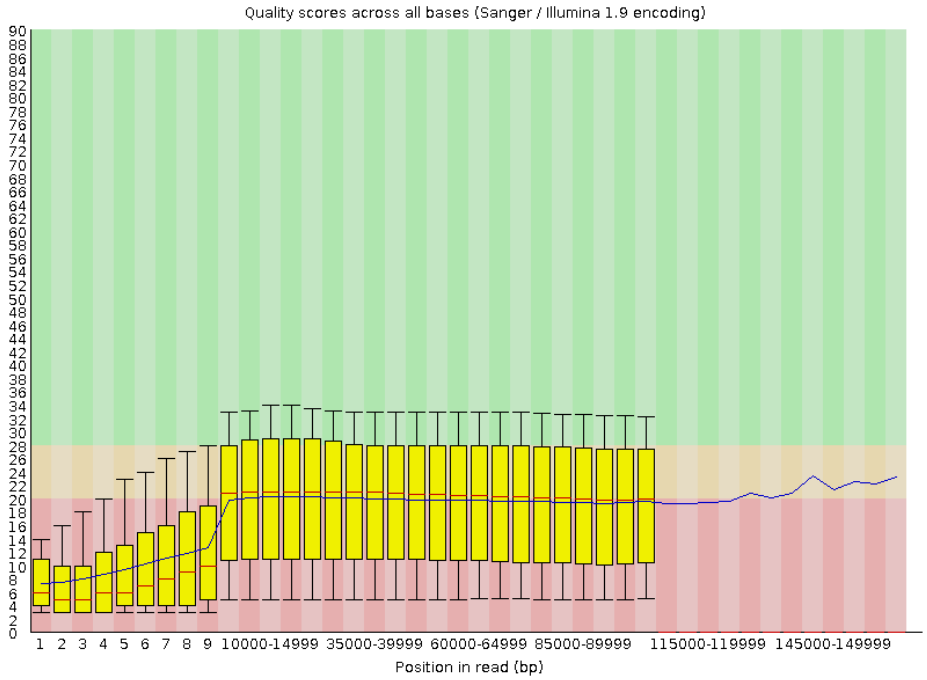


Figure 41: FASTQC per base quality

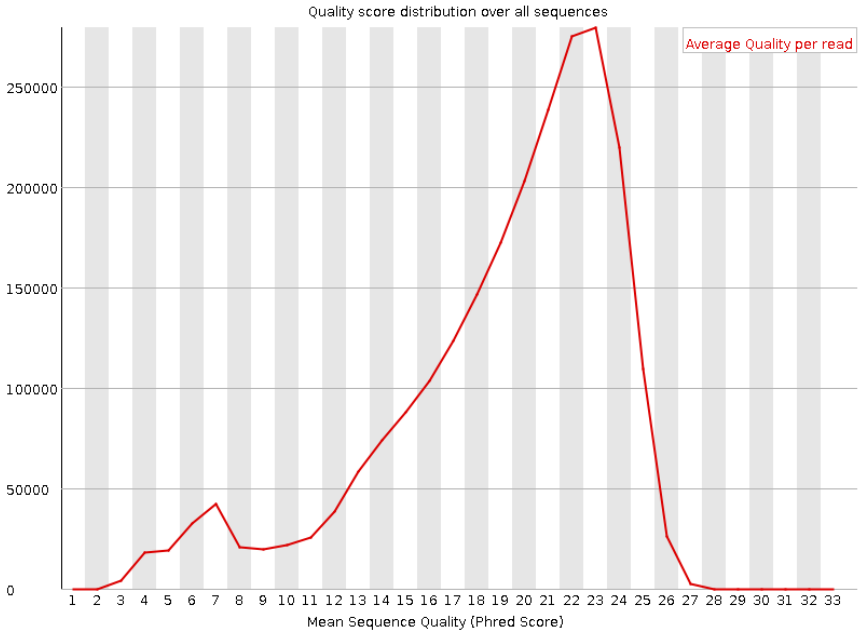


Figure 42: FASTQC per sequence quality

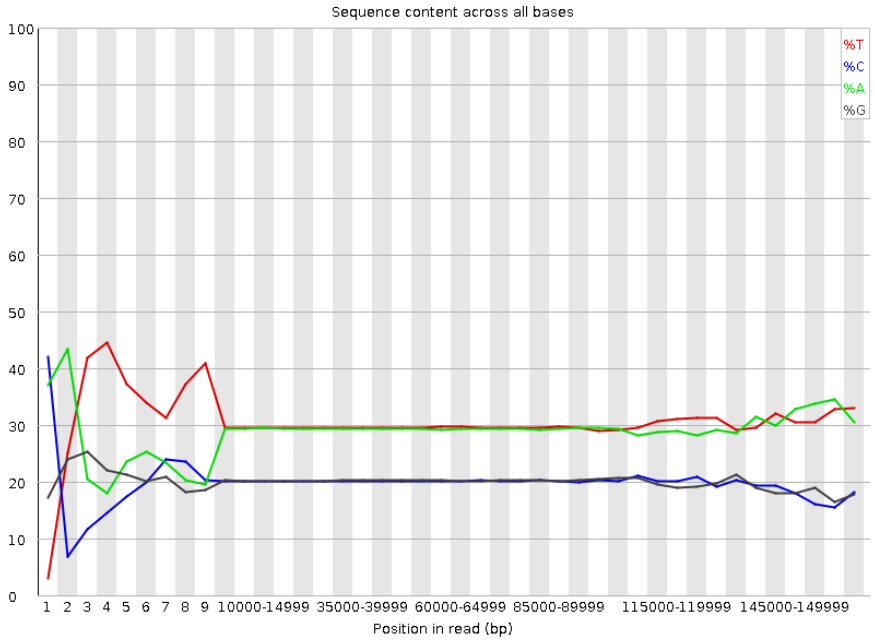


Figure 43: FASTQC ber base sequence content

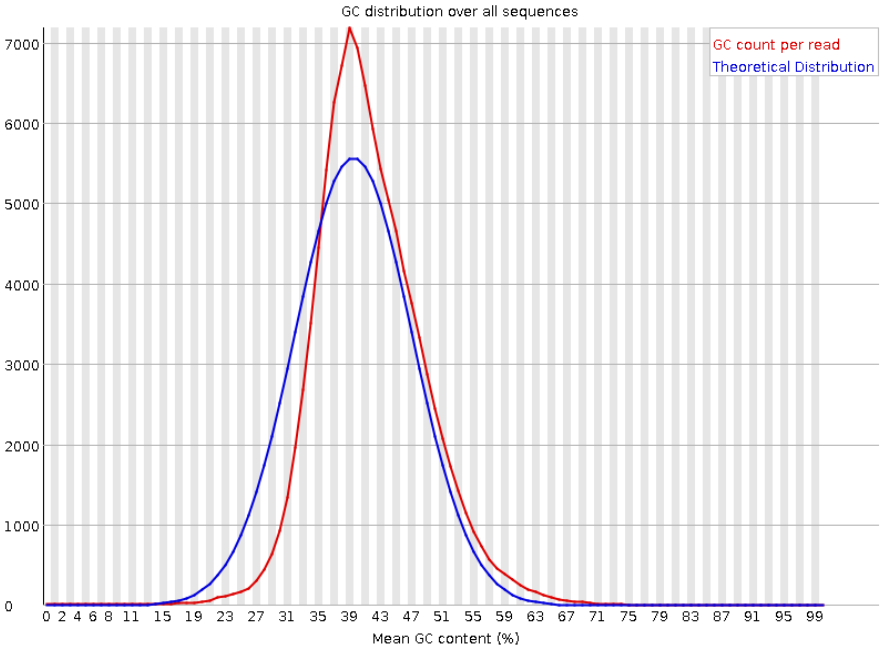


Figure 44: FASTQC per sequence GC content

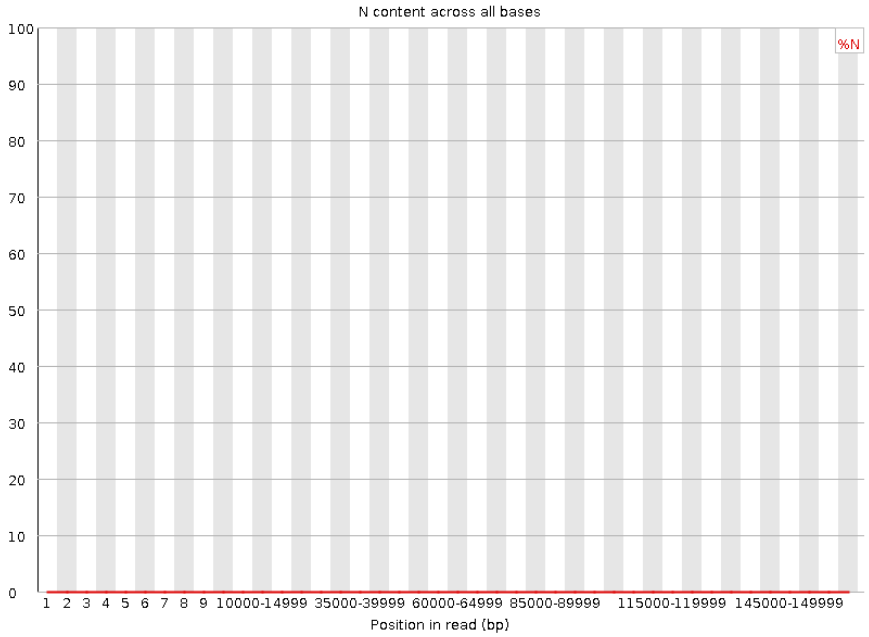


Figure 45: FASTQC N content

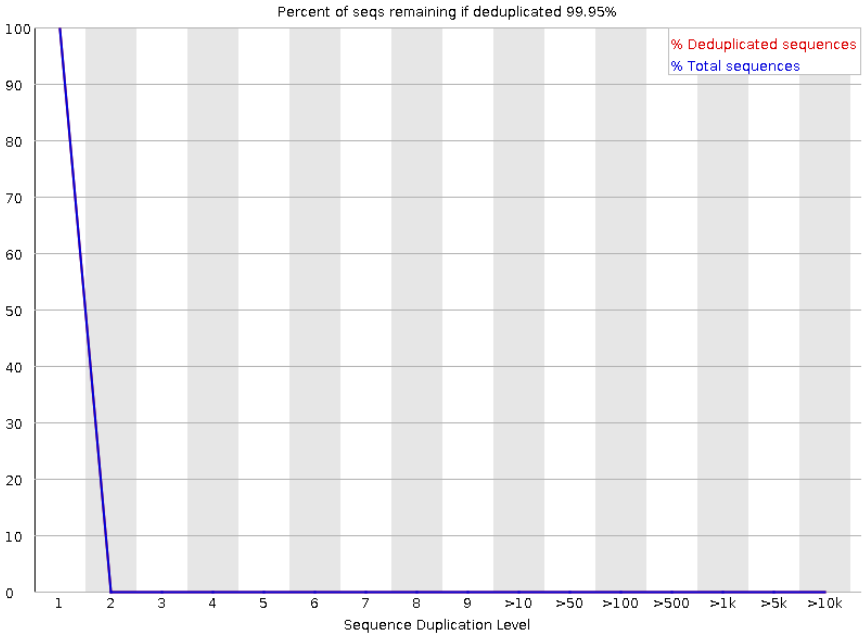


Figure 46: FASTQC duplication levels

2.5 Συζήτηση αποτελεσμάτων

Όπως έχει συζητηθεί ένα πολύ σημαντικό κομμάτι των τεχνολογιών που παράγουν μεγάλα reads είναι η χρησιμότητα τους στη de novo assembly. Για τέτοιου τύπου κατασκευής γονιδιώματος είναι αναγκαία η χρήση contigs που προέρχονται από όσο το δυνατόν μεγαλύτερα reads σε μήκος και ποιότητα. Οι πιο σημαντικές ενδείξεις που οδηγούν σε αποτελέσματα ικανά για ηψυλής ποιότητας de novo κατασκευής είναι η N50 , μέσο μήκος των reads, ποσοστό των reads που εχουν Qscore μεγαλύτερο του 12 και ποσοστό μεγαλύτερου του 15 και μέσο Qscore. Οι συγκρίσεις ανάμεσα σε αυτές τις ενδείξεις και τους παράγοντες που μελετάμε αναπαριστώνται στους παρακάτω διαγράμματα(διαγράμματα 1-5):

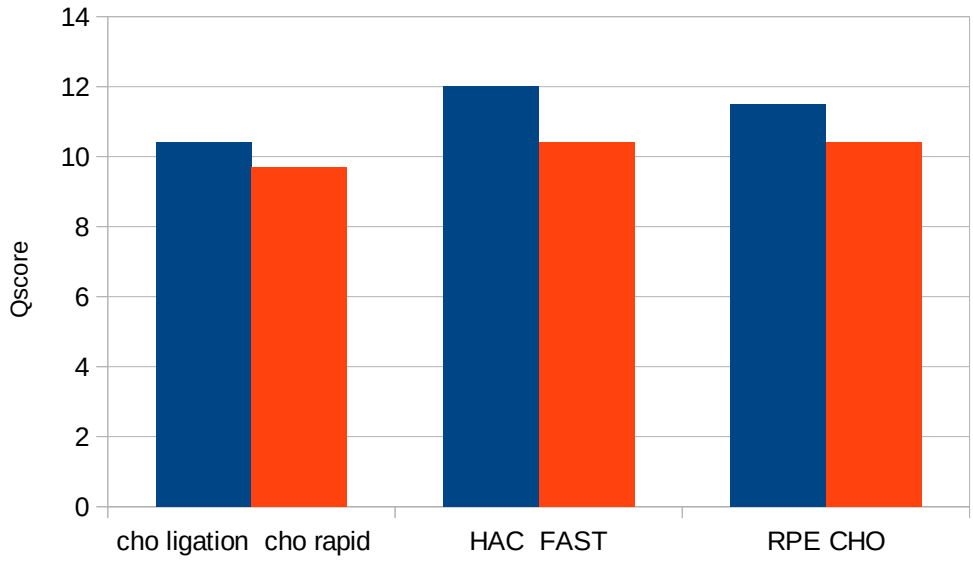


Figure 47: Qscore

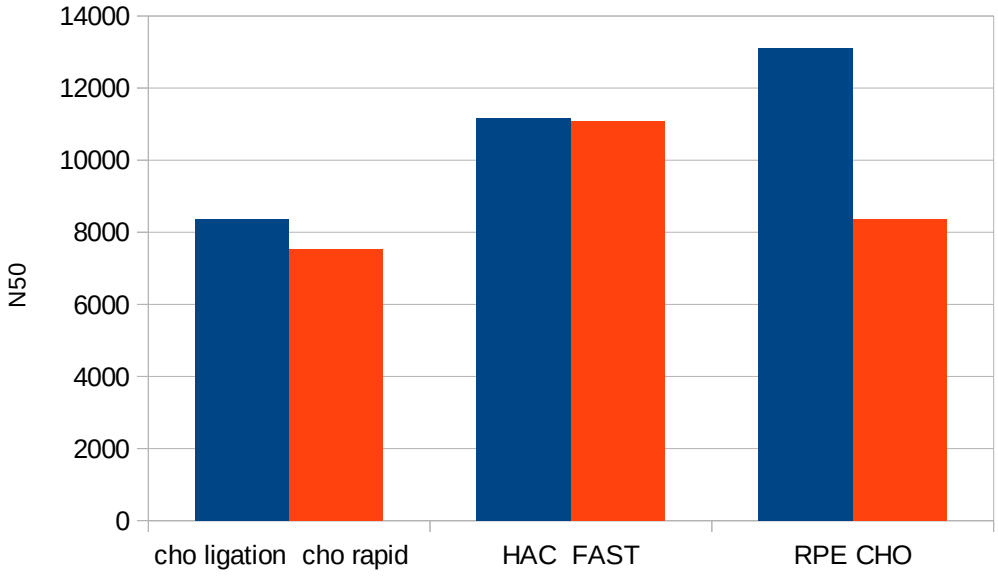


Figure 48: Τιμή N50

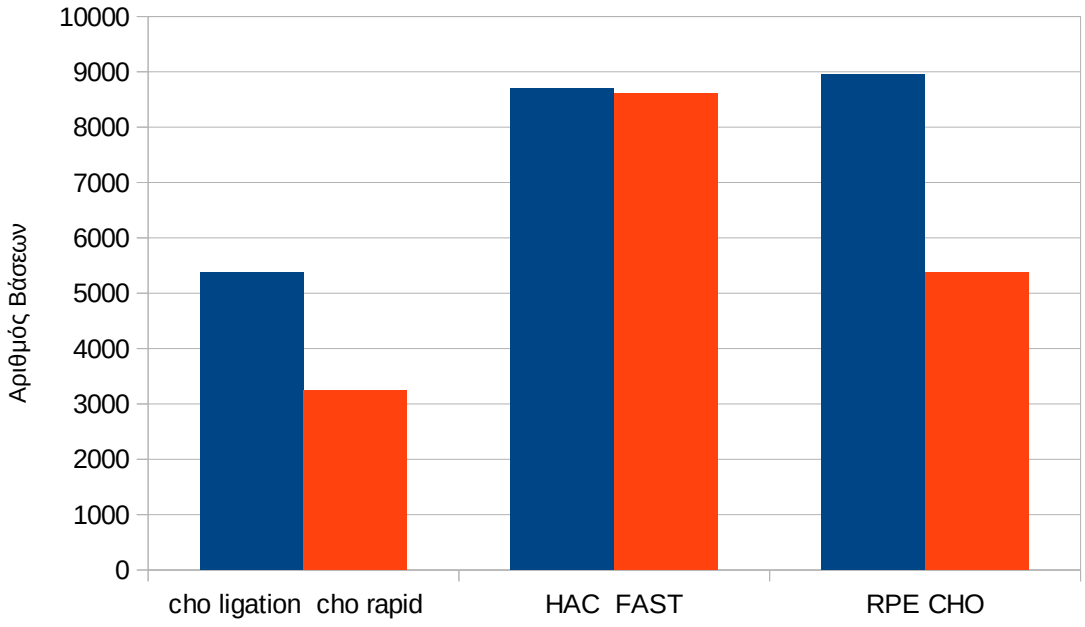


Figure 49: Μέσο μήκος των reads

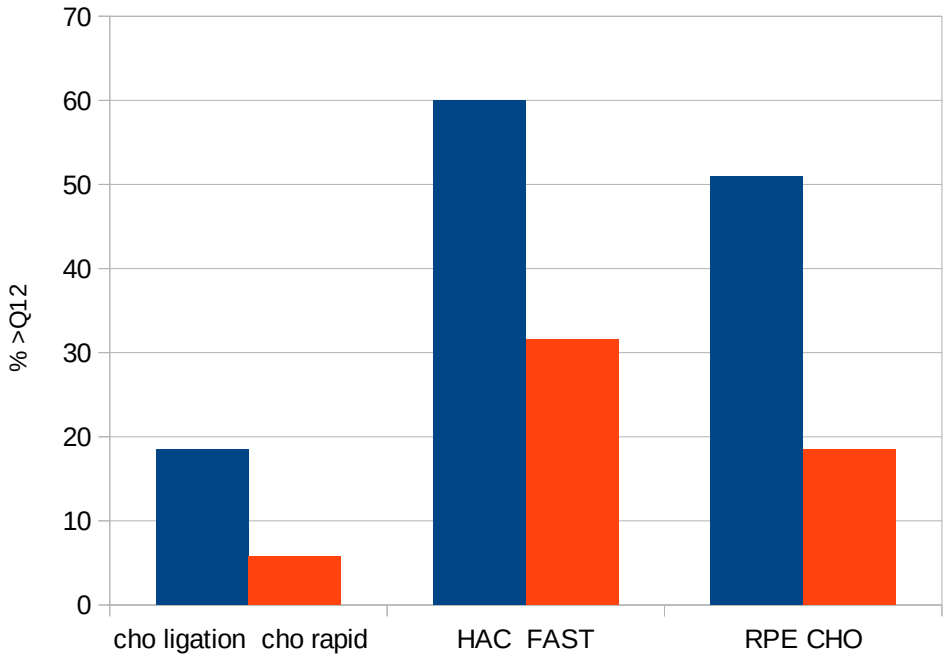


Figure 50: Ποσοστό των reads με Qscore μεγαλύτερο του 12

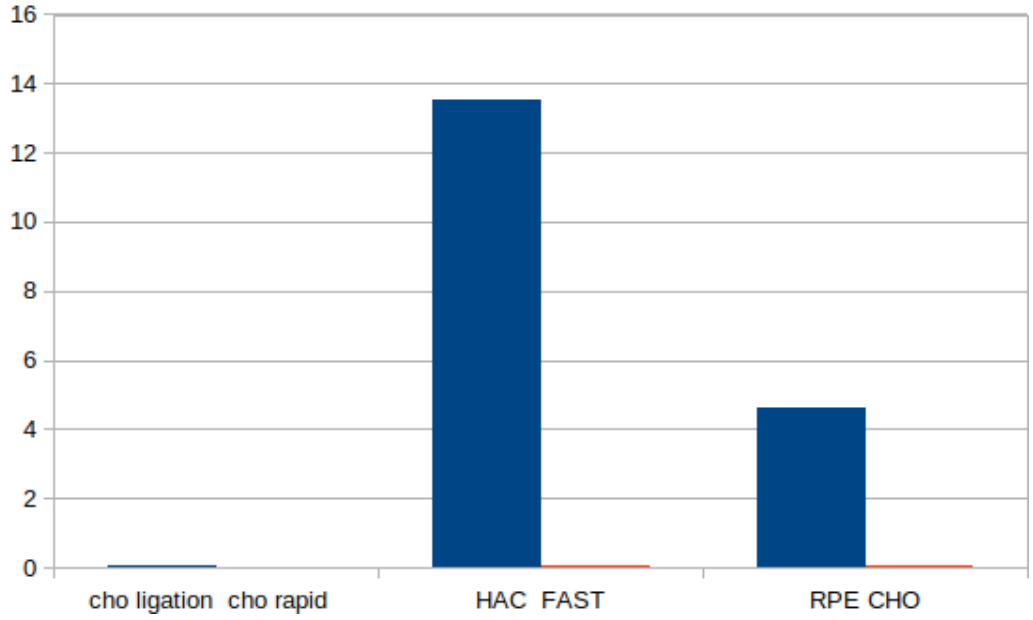


Figure 51: Ποσοστό των reads με Qscore μεγαλύτερο του 15

Βάσει τα παραπάνω δεδομένα η διαφορά του μέσω Qscore μεταξύ της rapid και της ligation είναι σχεδόν αμελητέα με διαφορά με την ligation μέθοδο να παράγονται περισσότερα reads με qscore μεγαλύτερο από 12. Επίσης παρατηρούμε ότι με την ligation μέθοδο παράγονται πολύ περισσότερα reads με μεγαλυτέρο μήκος καθιστώντας αυτήν την μέθοδο προτιμότερη για de novo κατασκευή γονιδιώματος.

Για τις διαφορές μεταξύ των μοντέλων HAC και FAST παρατηρούμε ότι έχουμε σχεδόν ίδιο μήκος στα reads που είναι αναμενόμενο αφού τα δεδομένα προετοιμάστηκαν με την ίδια μέθοδο library preparation. Η σημαντική τους διαφορά είναι είναι οτι με το HAC μοντέλο τα reads έχουν πολύ καλύτερη ποιότητα όπως φαίνεται απο το διάγραμμα 1 και 2 .

Τέλος ανάμεσα στο ανθρώπινο DNA και του χάμστερ. Με την χρήση του RPE έχουμε κατασκευή μεγαλύτερων αλληλουχιών στα reads μας καθώς και υψηλότερο qscore σε αυτά

Βιβλιογραφία

1. Hengyun Lu, Francesca Giordano, Zemin Ning,
Oxford Nanopore MinION Sequencing and Genome Assembly,
Genomics, Proteomics & Bioinformatics,
Volume 14, Issue 5,2016,Pages 265-279, ISSN 1672-0229,

2. Griffiths, Anthony J.F.. "DNA sequencing". Encyclopedia Britannica, Invalid Date, https://www.britannica.com/science/DNA-sequencing. Accessed 17 May 2021.

3. Eric E. Schadt, Steve Turner, Andrew Kasarskis, A window into third-generation sequencing, *Human Molecular Genetics*, Volume 19, Issue R2, 15 October 2010, Pages R227–R240

3.Third-generation sequencing and the future of genomics
Hayan Lee, James Gurtowski, Shinjae Yoo, Maria Nattestad, Shoshana Marcus, Sara Goodwin, W. Richard McCombie, Michael C. Schatz
bioRxiv 048603

4. Wick, R.R., Judd, L.M. & Holt, K.E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* **20,** 129 (2019)

5. Wouter De Coster, Svenn D'Hert, Darrin T Schultz, Marc Cruts, Christine Van Broeckhoven, NanoPack: visualizing and processing long-read sequencing data, *Bioinformatics*, Volume 34, Issue 15, 01 August 2018, Pages 2666–2669

6. Korlach J., Turner S.W. (2013) Zero-Mode Waveguides. In: Roberts G.C.K. (eds) Encyclopedia of Biophysics. Springer, Berlin, Heidelberg.

7. Winston Timp, Jeffrey Comer, Aleksei Aksimentiev,
DNA Base-Calling from a Nanopore Using a Viterbi Algorithm,

8. Jain, M., Fiddes, I., Miga, K. *et al.* Improved data analysis for the MinION nanopore sequencer. *Nat Methods* **12,** 351–356 (2015).

9. T. Laver, J. Harrison, P.A. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz, D.J. Studholme,
Assessing the performance of the Oxford Nanopore Technologies MinION,
Biomolecular Detection and Quantification,Volume 3,2015,Pages 1-8,ISSN 2214-7535,

10. Quail, M.A., Smith, M., Coupland, P. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13,** 341 (2012).

11. Carneiro, M.O., Russ, C., Ross, M.G. *et al.* Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics* **13,** 375 (2012).

12. Jain, M., Olsen, H.E., Paten, B. *et al.* The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17,** 239 (2016).

13. Rougemont, J., Amzallag, A., Iseli, C. *et al.* Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics* **9,** 431 (2008).

14. Article Source: DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads Boža V, Brejová B, Vinař T (2017) DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. PLOS ONE 12(6): e0178751.

15. Silvestre-Ryan, J., Holmes, I. Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing. *Genome Biol* **22,** 38 (2021).

16. Jain, M., Koren, S., Miga, K. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36,** 338–345 (2018).

17. Acharya, K., Khanal, S., Pantha, K. *et al.* A comparative assessment of conventional and molecular methods, including MinION nanopore sequencing, for surveying water quality. *Sci Rep* **9,** 15726 (2019).

18. Haotian Teng, Minh Duc Cao, Michael B Hall, Tania Duarte, Sheng Wang, Lachlan J M Coin, Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning, *GigaScience*, Volume 7, Issue 5, May 2018, giy037

19. Li Y, Wang S, Bi C, Qiu Z, Li M, Gao X. DeepSimulator1.5: a more powerful, quicker and lighter simulator for Nanopore sequencing. Bioinformatics. 2020 Apr 15;36(8):2578-2580. doi: 10.1093/bioinformatics/btz963. PMID: 31913436; PMCID: PMC7178411

20. Ashton, P., Nair, S., Dallman, T. *et al.* MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol* **33,** 296–300 (2015).

21. Leidenfrost, R.M., Pöther, DC., Jäckel, U. *et al.* Benchmarking the MinION: Evaluating long reads for microbial profiling. *Sci Rep* **10,** 5125 (2020).