ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Τμήμα Βιοχημείας και Βιοτεχνολογίας

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Εφαρμογές Μοριακής Βιολογίας - Μοριακή Γενετική - Διαγνωστικοί Δείκτες »

Βιοπληροφορική ανάλυση ανθρώπινου εξονιώματος από δεδομένα νέας γενιάς τεχνολογιών αλληλούχισης

Χριστίνα Σίνη

Επιβλέπων Καθηγητής:
Γρηγόριος Αμούτζιας

Λάρισα 2013

Bioinformatic analysis pipeline for human exome raw data from Next Generation Sequencing (NGS) technologies

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

κ. Γρηγόριος Αμούτζιας (Επιβλέπων).

Λέκτορας 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. Θεολογία Σαραφίδου.

Λέκτορας Μοριακής Γενετικής Ζωϊκών Οργανισμών, Τμήμα Βιοχημείας και Βιοτεχνολογίας , Πανεπιστήμιο Θεσσαλίας.

Abstract

This thesis was carried out during the postgraduate program «Applications in Molecular Biology – Molecular Genetics – Diagnostic Markers», in the Bioinformatics laboratory of Dr. Grigorios Amoutzias, Lecturer of Bionformatics in Genomics, in the Department of Biochemistry and Biotechnology, at the University of Thessaly, Greece. The thesis' title is "Bioinformatic analysis pipeline for human exome raw data from Next Generation Sequencing (NGS) technologies" and deals with the publicly available bioinformatics methods, tools and databases for the analysis of exome data. Particularly, this thesis developed/optimized a bioinformatics protocol/pipeline based on publicly available tools and databases for the analysis of exome sequencing raw paired-end data from Illumina. The developed protocol of programs deals with i) downloading raw data from SRA, ii) quality control and trimming of raw data, iii) reference alignment, iv) further reprocessing of aligned data, v) realignment of problematic regions (indels), vi) estimation of insert size distribution, vii) more accurate recalibration of base quality scores, viii) SNP calling and filtering, ix) SNP annotation and further filtering and finally x) phenotype prediction.

Πρόλογος

Η παρούσα μεταπτυχιακή εργασία διακπεραιώθηκε στα πλαίσια του μεταπτυχιακού προγράμματος «Εφαρμογές Μοριακής Βιολογίας - Μοριακή Γενετική - Διαγνωστικοί Δείκτες», στο εργαστήριο Βιοπληροφορικής του Δρ. Γρηγόριου Αμούτζια, Λέκτορα Βιοπληροφορικής στη Γενωμική, του Τμήματος Βιοχημείας και Βιοτεχνολογίας, του Πανεπιστημίου Θεσσαλίας. Ο τίτλος της εργασίας είναι ¨Ανάλυση του ανθρώπινου εξονιώματος από δεδομένα νέας γενιάς τεχνολογιών αλληλούχισης (Exome sequencing)¨ και πραγματεύεται τις μεθόδους, βιοπληροφορικά εργαλεία και βάσεις δεδομένων που είναι διαθέσιμα στο διαδίκτυο για την ανάλυση δεδομένων ανθρώπινου εξονιώματος. Συγκεκριμένα, η εργασία αυτή ανέπτυξε και βελτιστοποίησε ένα πρωτόκολο βιοπληροφορικής που βασίστηκε σε ελεύθερα διαθέσιμα προγράμματα και βάσεις δεδομένων για την ανάλυση και διαχείριση δεδομένων (paired-end sequence data) ανθρώπινου εξονιώματος που αλληλουχήθηκε με την τεχνολογία της εταιρίας Illumina. Το συγκεκριμένο πρωτόκολο διαχειρίζεται i) την απόκτηση δεδομένων αλληλούχισης από την βάση δεδομένων SRA, ii) τον ποιοτικό έλεγχο των δεδομένων, iii) τη στοίχιση των ακολουθιών στο ανθρώπινο γονιδίωμα αναφοράς, iv) την περαιτέρω επεξεργασία των στοιχισμένων ακολουθιών, v) την επαναστοίχιση προβληματικών περιοχών, vi) τον υπολογισμό της κατανομής του μεγέθους του ενθέματος των paired-end reads, vii) τον βελτιωμένο επανα-υπολογισμό της ποιότητας (Q-score) των αλληλουχημένων βάσεων, viii) τον εντοπισμό SNPs και το φιλτράρισμά τους, ix) τον λειτουργικό σχολιασμό των SNPs και το επιπλέον φιλτράρισμά τους και τέλος, x) την πρόβλεψη φαινοτύπου.

Ευχαριστίες

Τις θερμές μου ευχαριστίες, θα ήθελα να εκφράσω στον επιβλέποντα καθηγητή μου, τον Κ.Γρηγόριο Αμούτζια για την καθοδήγηση, την κατανόηση και την πολύτιμη βοήθειά του, κατά την διάρκεια της διεκπεραίωσης της μεταπτυχιακής αυτής εργασίας. Επίσης, θα ήθελα να ευχαριστήσω την οικογένειά μου καθώς και τους στενούς μου φίλους, για την συμπαράσταση και την υπομονή που έδειξαν όλο αυτό το διάστημα.

Περιεχόμενα

1. Εισαγωγή

1.1 Προς την αλληλούχιση του ανθρώπινου γονιδιώματος

Κατά τη διάρκεια των τελευταίων 20 χρόνων, σημαντικές πρόοδοι στην τεχνολογία οδήγησαν σε μια επανάσταση στον τομέα της βιολογίας (Kevles, 1999) που σημαδεύτηκε από την ακατάπαυστη προσπάθεια για την ¨αποκρυπτογράφηση¨ καταρχήν γονιδίων και αργότερα ολόκληρων γονιδιωμάτων, η οποία ¨έσπειρε¨το πεδίο της γενωμικής (genomics). Καρπός αυτής της προσπάθειας είναι οι αλληλουχίες του γονιδιώματος ιών και ιοειδών (viroids), φυσικών πλασμιδίων, οργανιδίων, ευβακτηρίων, αρχαιοβακτηρίων, μυκήτων, ζώων και φυτών. Ωστόσο, η αποκρυπτογράφηση του ανθρώπινου γονιδιώματος, του γενετικού αποτυπώματος της ζωής, αποτελούσε ουσιαστική πρόκληση. Για το λόγο αυτό καλλιεργήθηκε η ιδέα της αλληλούχισης του ανθρώπινου γονιδιώματος, της εγκυκλοπαίδειας της ζωής και βασικού εργαλείου (International Human Genome Sequencing Consortium, 2001) των μοριακών βιολόγων, που ήταν απαραίτητη για πολλές άλλες εφαρμογές και πειραματικές προσσεγγίσεις, όπως η μεταλλαξιγένεση σημείου, τα πειράματα μεταφοράς γονιδίων και η ανάλυση γονιδιακής ρύθμισης και πρωτεϊνικής έκφρασης (Watson et al., 2007).

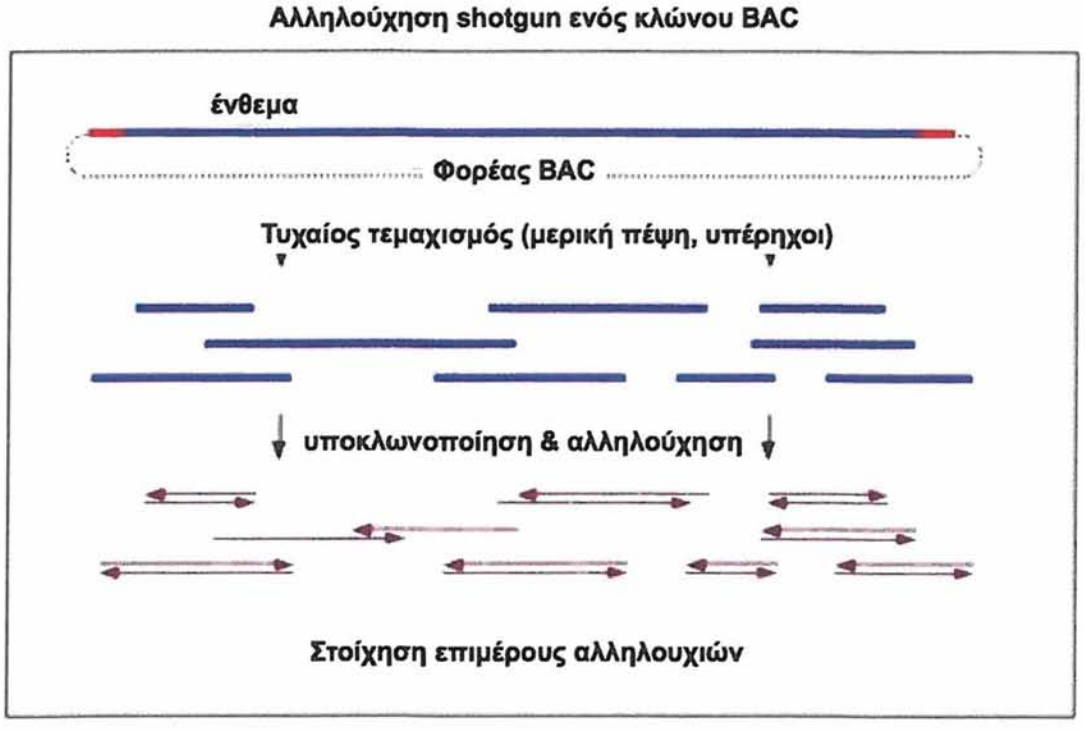
1.1.1 Το Πρόγραμμα του Ανθρώπινου Γονιδιώματος

Το Πρόγραμμα του Ανθρώπινου Γονιδιώματος (Human Genome Project - HGP) ξεκίνησε το 1990 και μέχρι την ολοκλήρωσή του, το 2003, προσέφερε σημαντικές ανακαλύψεις στο πεδίο της αρχιτεκτονικής και λειτουργίας του ανθρώπινου γονιδιώματος (Naidoo et al., 2011). Απώτερος στόχος του ήταν ο προσδιορισμός της αλληλουχίας των 3 δισεκατομμυρίων ζευγών βάσεων από τις οποίες αποτελούνται τα 46 χρωμοσώματα του ανθρώπου (Watson et al., 2007).

Από την αρχή ήταν ξεκάθαρο ότι η αλληλούχιση του ανθρώπινου γονιδιώματος παρουσίαζε τεράστια τεχνικά και λογιστικά προβλήματα, προβλήματα σε έκταση που δεν είχε αντιμετωπίσει η βιολογική έρευνα ποτέ άλλοτε στο παρελθόν. Ένα σημαντικό πρόβλημα αφορούσε τη φύση της ανθρώπινης αλληλουχίας. Τα πειράματα

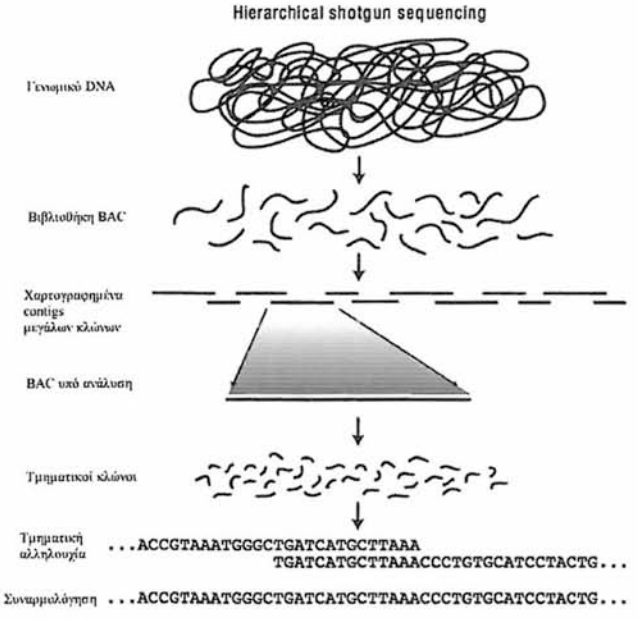
κινητικής υβριδοποίησης που είχαν γίνει κατά τη δεκαετία του 1970, καθώς και οι αλληλουχίες μεμονωμένων γονιδίων είχαν δείξει ότι το ανθρώπινο γονιδίωμα περιέχει πολύ μεγάλη ποσότητα επαναλαμβανόμενου DNA, το οποίο δημιουργεί δυσκολίες στην αλληλούχιση (Watson et al., 2007).

Λίγο μετά την επινόηση μεθόδων για την ανάλυση της αλληλουχίας του DNA αναπτύχθηκε και η στρατηγική της τμηματικής ανάλυσης (shotgun sequencing), η οποία παρέμεινε η κύρια μέθοδος για την ανάλυση σε γενωμικό επίπεδο περίπου 20 χρόνια. Η μέθοδος τελειοποιήθηκε και διευρύνθηκε ώστε να γίνει πιο αποτελεσματική. Για παράδειγμα, η βελτίωση των πρωτοκόλλων για την κατάτμηση και κλωνοποίηση του DNA σε φορείς, επέτρεψε την παρασκευή τμηματικών βιβλιοθηκών με πιο ομοιόμορφη αντιπροσώπευση. Η πρακτική της ανάλυσης και από τα δύο άκρα των δίκλωνων τμημάτων (¨double – barreled¨ shotgun sequencing) που εισήχθει το 1990 από τον Ansorge και άλλους, επέτρεψε τη χρησιμοποίηση «συνδετικών πληροφοριών» (linking information) μεταξύ των κλασμάτων της αλληλουχίας.



Εικ.1: Αλληλούχιση shotgun ενός κλώνου BAC
(http://www.ocf.berkeley.edu/~edy/genome/nhgri.html)

Η ανάλυση της αλληλουχίας ενός μεγάλου γονιδιώματος πλούσιου σε επαναλήψεις όπως το ανθρώπινο, μπορεί να γίνει με δύο τρόπους (International Human Genome Sequencing Consortium, 2001) σύμφωνα με την προσέγγιση της «τυχαίας προσπέλασης ολόκληρου γονιδιώματος», (whole – genome shotgun) που προτάθηκε στις αρχές του 1990 από τον J. Craig Venter. Η μέθοδος αυτή (πρώτος τρόπος), υποστήριζε την αλληλούχιση του γονιδιώματος μέσω της δημιουργίας κλώνων πολλών χιλιάδων μικρών τμημάτων του γονιδιώματος, σε ισάριθμα πλασμίδια, τη μαζική αλληλούχιση τυχαίων κλώνων και τέλος τη συναρμολόγηση αυτών των αναγνώσεων αλληλουχίας χωρίς να είναι γνωστή εκ των προτέρων η σχετική θέση κάθε κλώνου στο γονιδίωμα (Watson et al., 2007). Ο δεύτερος τρόπος είναι η «ιεραρχική τμηματική» ανάλυση (¨hierarchical shotgun sequencing¨ approach), η οποία αναφέρεται και ως «βασισμένη σε χάρτες» (¨map-based¨), «βασισμένη σε BAC (εικ.1)» (¨BAC-based¨) ή «κλώνο – κλώνο» (¨clone - by - clone¨). Η προσέγγιση αυτή συνίσταται στη δημιουργία και οργάνωση ενός συνόλου κλώνων με μεγάλα ένθετα τμήματα (συνήθως 100-200 kb το καθένα) που καλύπτουν το γονιδίωμα και κατόπιν στην τμηματική ανάλυση κατάλληλα επιλεγμένων κλώνων. Επειδή τα δεδομένα για την αλληλουχία αφορούν σε συγκεκριμένα τμήματα συγκεκριμένων κλώνων, το ενδεχόμενο κακής συναρμολόγησης, τόσο σε περιορισμένη όσο και σε ευρύτερη κλίμακα, περιορίζεται. Ωστόσο, μπορεί να προκύψει πρόβλημα από πιθανή αναδιάταξη μερικών κλώνων με μεγάλα ένθετα, αν και ο σχετικός κίνδυνος μπορεί να ελαττωθεί με κατάλληλα μέτρα ποιοτικού ελέγχου (International Human Genome Sequencing Consortium, 2001).



Εικ.2: Βασικές αρχές ιεραρχικής τμηματικής ανάλυσης αλληλουχίας (International Human Genome Sequencing Consortium, 2001).

Τελικά, για διάφορους λόγους, αποφασίστηκε ότι η ανάλυση της αλληλουχίας του ανθρώπινου γονιδιώματος έπρεπε να γίνει με ιεραρχική προσέγγιση (εικ.2). Ήταν φρόνιμο να χρησιμοποιηθεί η συγκεκριμένη προσέγγιση στην πρώτη απόπειρα ανάλυσης ενός γονιδιώματος πλούσιου σε επαναλήψεις (International Human Genome Sequencing Consortium, 2001). Πώς θα μπορούσε άλλωστε να προσδιοριστεί η σωστή θέση των επαναλαμβανόμενων αλληλουχιών του γονιδιώματος, αφού τα διάφορα αντίγραφα προέρχονται από πολλές διαφορετικές θέσεις; Επιπλέον, τα υπολογιστικά προγράμματα συναρμολόγησης αλληλουχιών δεν ήταν τότε σε θέση να διαχειριστούν τον τεράστιο αριθμό αναγνώσεων αλληλουχίας που θα ήταν απαραίτητος για την συναρμολόγηση ακόμη και ενός μικρού γονιδιώματος (Watson et al., 2007). Με την ιεραρχική προσέγγιση, η συχνότητα λαθών συναρμολόγησης στο ολοκληρωμένο προϊόν θα ήταν μάλλον μικρότερη από την αντίστοιχη συχνότητα, για την προσέγγιση σε επίπεδο γονιδιώματος, κατά την οποία θα ήταν δύσκολο να αναγνωρισθούν περιοχές με λάθη συναρμολόγησης (International Human Genome Sequencing Consortium, 2001).

Η πρόχειρη αλληλουχία (draft genome) του Προγράμματος του Ανθρώπινου Γονιδιώματος, η οποία ανακοινώθηκε το 2000 (Watson et al., 2007) ήταν ατελής εξαιτίας των ακάλυπτων περιοχών της ευχρωματίνης το οποίο άγγιζε περίπου το 30% του γονιδιώματος το οποίο δεν καλύφθηκε. Επιπλέον, υπήρχε ένας εκτενής αριθμός κενών (gaps) μεταξύ των συναρμολογημάτων τα οποία καθιστούσαν την γενωνική αλληλουχία δυσανάγνωστη (Naidoo et al., 2011).

Τελικά, η ολοκληρωμένη αλληλουχία περιείχε λιγότερο από ένα σφάλμα ανά 100.000 ζεύγη βάσεων, ξεπερνώντας τον αρχικό ποιοτικό στόχο κατά δέκα φορές. Σε σύγκριση με την πρόχειρη αλληλουχία που είχε δημοσιευτεί, οι βελτιώσεις ήταν εντυπωσιακές. Η συνολική ακρίβεια ως προς τα ζεύγη βάσεων ήταν πολύ υψηλότερη στην ολοκληρωμένη αλληλουχία, όμως το πιο εντυπωσιακό ήταν η μείωση του πλήθους των χασμάτων από 150.000 σχεδόν στην πρόχειρη αλληλουχία σε 281 στην ολοκληρωμένη αλληλουχία (συντελεστής μείωσης >400x). Ενώ ο αριθμός των

σφαλμάτων στην πρόχειρη αλληλουχία ήταν τεράστιος, στην ολοκληρωμένη αλληλουχία πλησίαζε το μηδέν.

Συμπερασματικά, η ολοκληρωμένη αλληλουχία αποτελεί μια πιστή αναπαράσταση μιας αλληλουχίας – αναφοράς (reference – human sequence) του ανθρώπινου γονιδιώματος (Watson et al., 2007).

Συγκεκριμένα, η ολοκληρωμένη αλληλουχία παρέχει ακριβείς πληροφορίες για:

- τον αριθμό και την πυκνότητα των γονιδίων
- τις μη – κωδικοποιούσες σε πρωτεΐνες περιοχές, RNA γονιδίων (ή RNA γονίδια)
- τον αριθμό των αντιγράφων των επαναλαμβανόμενων αλληλουχιών
- την λειτουργική και εξελικτική ταξινόμηση (Naidoo et al., 2011).

1.1.2 Πολυμορφισμοί του DNA

Με την ολοκλήρωση του Προγράμματος του Ανθρώπινου Γονιδιώματος βρέθηκε ότι το DNA αποτελείται από πληθώρα παραλλαγών. Κάποιες από τις κατηγορίες των παραλλαγών, αναφέρονται παρακάτω, διότι συνεισφέρουν στην δημιουργία γενετικών χαρτών και στην έρευνα αυτών καθώς πολλές φορές συντελούν στον προσδιορισμό φαινοτυπικών και γονοτυπικών χαρακτηριστικών. Οι παραλλαγές του DNA μπορεί να είναι κοινές για μια ομάδα ή και μοναδικές για το κάθε άτομο. Κάποιες απο τις κατηγορίες πολυμορφισμών οι οποίες μελετώνται ευρέως είναι :

<u>Μονονουκλεοτιδικός πολυμορφισμός (Single nucleotide polymorphism - SNP)</u>

Μονονουκλεοτιδικός πολυμορφισμός θεωρείται μία αλλαγή σε ένας ζεύγος βάσεων, μία σημειακή μεταλλαγή σε ένα γενετικό τόπο του γονιδιώματος. Πρόκειται για τον πιο συνηθισμένο τύπο πολυμορφισμού DNA καθώς συναντάται με συχνότητα περίπου 1 ανά 350 bp και ευθύνεται περίπου για το 80-85% της ποικιλομορφίας που εμφανίζει η αλληλουχία του ανθρώπινου DNA. Νέα SNPs δημιουργούνται λόγω αυθόρμητων μεταλλαγών οι οποίες συνήθως οφείλονται σε σφάλματα κατά την αντιγραφή (Russell, 2005). Επειδή, λάθη κατά την αντιγραφή συμβαίνουν σπάνια,

εξίσου σπάνιο γεγονός είναι και η εμφάνιση νέου SNP (Kumar et al., 2012). Ανάλογα με την κωδική περιοχή που βρίσκεται το SNP μπορεί να προκαλέσει παρανοηματικές μεταλλαγές στην αντίστοιχη πρωτεΐνη ή σιωπηλές μεταλλαγές.

Εκτιμάται ότι οι μισές παρανοηματικές μεταλλαγές (non-synonymous) που οφείλονται σε SNP προκαλούν γενετικές ασθένειες στον άνθρωπο. Η γονιδιακή λειτουργία μπορεί επίσης να επηρεαστεί από τους μη κωδικούς SNP όταν αυτοί συναντώνται σε υποκινητές ή άλλες ρυθμιστικές περιοχές των γονιδίων (Russell, 2005).

Μονονουκλεοτιδικές παραλλαγές (SNVs)

Οι μονονουκλεοτιδικές παραλλαγές είναι παρόμοιες με τους μονονουκλεοτιδικούς πολυμορφισμούς (SNPs) με την μόνη διαφορά ότι δεν έχουν εντοπιστεί σε μεγάλη συχνότητα πληθυσμού (http://www.populationdiagnostics.com/science.html).

Πολυμορφισμός αριθμού αντιγράφων (CNVs – Copy Number Variation)

Τα πολλαπλά αντίγραφα παραλλαγών (CNVs) αποτελούν το 12% των παραλλαγών του ανθρώπινου DNA. Κάθε τέτοιου είδους παραλλαγή κυμαίνεται από 1Kb νουκλεοτιδικές βάσεις έως αρκετά Mb (Stankiewicz et al., 2010). Τα CNVs μπορούν να προκληθούν από de novo μεταλλάξεις (Carvalho et al., 2007) (δομικές αναδιατάξεις όπως ελλείψεις, διπλασιασμοί, αναστροφές και μετατοπίσεις) ή να κληρονομηθούν (Lupski, 2006). Τα CNVs μπορούν να περιορίζονται σε ένα μόνο γονίδιο ή να περιλαμβάνουν μια συνεχόμενη σειρά από γονίδια. Τα πολλαπλά αντίγραφα παραλλαγών μπορούν να προκαλέσουν φαινοτυπική ποικιλομορφία, πολύπλοκα χαρακτηριστικά συμπεριφοράς και ευαισθησία σε ασθένειες (Freeman et al., 2006)

1.1.3 Μέθοδοι αλληλούχισης

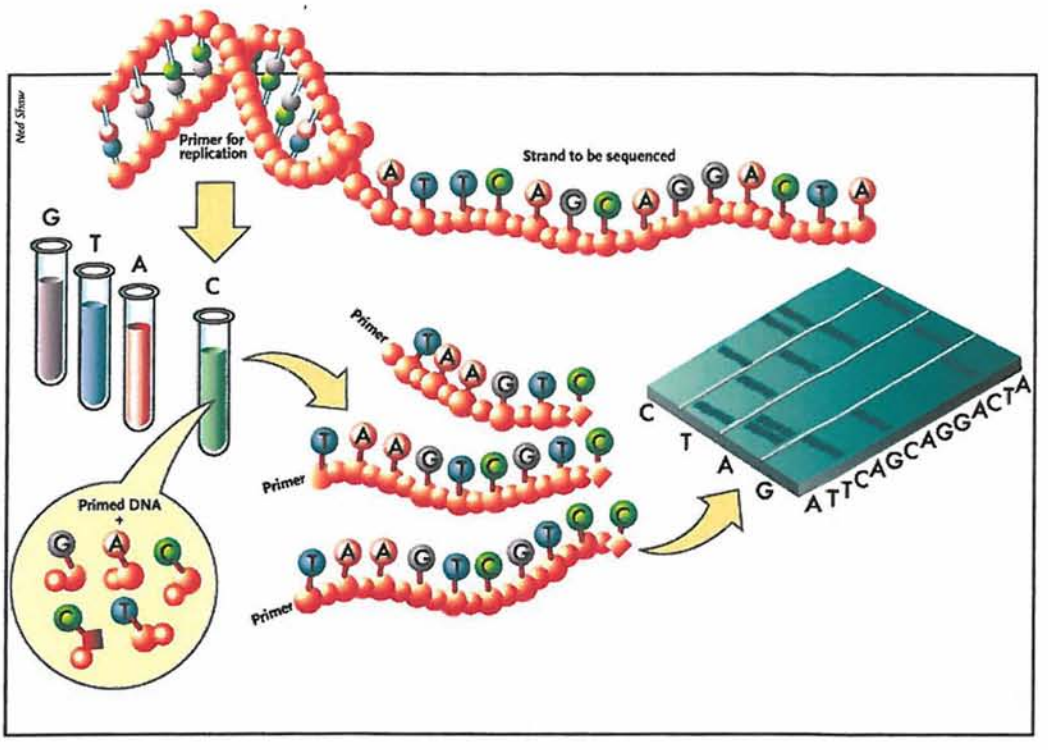
Η επινόηση τεχνικών αλληλούχισης του DNA στα μέσα της δεκαετίας του 1970 σήμανε την έναρξη μιας νέας εποχής για την βιολογική έρευνα, μια εποχή που οδήγησε γρήγορα σε τεράστια διεύρυνση των γνώσεών μας για τα γονίδια, δηλαδή για τη δομή, τη λειτουργία, τη ρύθμιση και την εξέλιξή τους.

Οι τεχνικές αλληλούχισης, αν και δύσχρηστες στην αρχή, γρήγορα υιοθετήθηκαν από πολλά εργαστήρια συντελώντας στην ανάπτυξη της επιστημονικής έρευνας (Watson et al., 2007).

Το 1975 ο Sanger παρουσίασε σε μια διάλεξη του, τη μέθοδο DNA αλληλούχισης, στην οποία γίνεται χρήση ανιχνευτών για τη σύνθεση της αλληλουχίας. Συγκεκριμένα, περιέγραψε μια μέθοδο αλληλούχισης ολιγονουκλεοτιδίων με ενζυμικό πολυμερισμό. Αυτή η μέθοδος ήταν αρχικά γνωστή ώς μέθοδος τερματισμού ή ως μέθοδος διδεοξυνουκλεοτιδίων. 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 πολυμερισμός 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 αυτή μέθοδος αλληλούχισης DNAχρησιμοποιείται για γονιδιωματική έρευνα ως κύριο εργαλείο για την αλληλούχιση τμημάτων DNA.



Εικ.3: Ενζυμική μέθοδος Sanger
http://sulleormedisherlockholmes.protagonista.altervista.org/sulleormedisherlockholmes/Sulle_orme_di_holmes/sequenziamento_dna.html

Το 1977 ο Allan Maxam και ο Walter Gilbert δημοσίευσαν μία μέθοδο η οποία βασίζεται στην αποκοπή με χημικό τρόπο (χημική μέθοδος) των συγκεκριμένων βάσεων τερματισμού. Ο διαχωρισμός των τμημάτων γίνεται με ηλεκτροφόρηση σε τζελ πολυακριλαμίδης (Franca et al., 2002).

Η μέθοδος της ηλεκτροφόρησης σε τζελ πολυακριλαμίδης παρουσίαζε αρκετά προβλήματα όσον αφορά την προετοιμασία του τζελ (χρονοβόρα διαδικασία), του δείγματος και την διατήρηση του τζελ μετά την ηλεκτροφόρηση. Για τους λόγους αυτούς προτάθηκε η τεχνολογία της ηλεκτροφόρησης σε τριχοειδή (Capillary Electrophoresis - CE) όπου η αλληλούχιση του δείγματος πραγματοποιείται σε σωληνάρια σιλικόνης.

Η ενζυμική μέθοδος ήταν αρκετά χρονοβόρα και ιδιαίτερα επισφαλής. Για το λόγο αυτό, αναπτύχθηκε μια εναλλακτική μέθοδος σήμανσης (εικ.3) η οποία αντικατέστησε την ραδιενέργεια. Για την αλληλούχιση ενός τμήματος DNA πραγματοποιείται μία σύνθετη αντίδραση τερματισμού. Η αντίδραση γίνεται παρουσία των τεσσάρων κανονικών τριφωσφορικών δεόξυριβονουκλεοτιδίων σε σχετικά μεγάλη συγκέντρωση και τεσσάρων διδεόξυριβονουκλεοτιδίων σε μικρότερη συγκέντρωση τα οποία είναι σημασμένα το καθένα με διαφορετική φθορίζουσα χημική ομάδα. Έτσι σχηματίζεται μίγμα προιόντων 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΄ακρο. 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 μέγεθός 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 (Watson et al., 2007).

1.2 Νέας γενιάς τεχνολογίες αλληλούχισης (NGS)

Η μέθοδος του Sanger αντικαταστάθηκε το 2004 από τις νέας γενιάς τεχνολογίες αλληλούχισης προσφέροντας υψηλής ευκρίνειας αλληλούχιση περίπου 1000 φορές μεγαλύτερη από την παραδοσιακή.

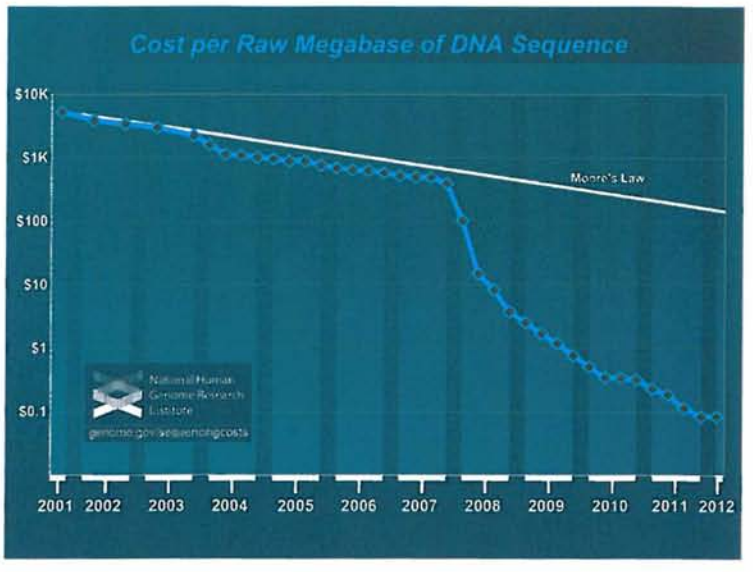
Μία από τις βασικές διαφορές είναι η ικανότητα ταυτόχρονης αλληλούχισης εκατομμυρίων τμημάτων DNA (massively parallel sequencing technologies). Αυτό το χαρακτηριστικό προσφέρει τη δυνατότητα αλληλούχισης μεγάλου αριθμού νουκλεοτιδίων ανά διαδρομή σε σύγκριση με την αλληλούχιση κατά Sanger. Η χημεία της νέας γενιάς τεχνολογιών αλληλούχισης, μαζί με την ικανότητα υψηλής παραγωγικής απόδοσης έχει μειώσει σημαντικά το κόστος της αλληλούχισης. Οι νέες τεχνολογίες αλληλούχισης (πιν.1), που διατίθενται σήμερα, μπορούν να ταξινομηθούν σε δεύτερης και τρίτης γενιάς (Naidoo et al., 2011).

Δεύτερης και τρίτης γενιάς τεχνολογίες αλληλούχισης (Next – Generation sequencing Technology)	
2ης γενιάς NGS	• **Roche /454** • **Illumina/Solexa** • **SOLiD** • **HeliScope**
3ης γενιάς NGS	• **Single-Molecule Real Time (SMRT)** • **Ion Torrent** • **Nanopore**

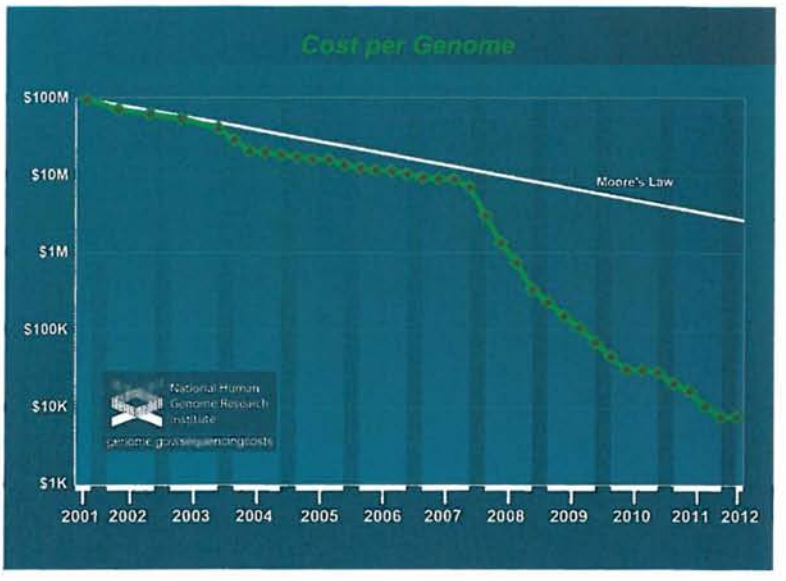
Πίνακας 1:Δεύτερης και τρίτης γενιάς τεχνολογίες αλληλούχισης

Τα τελευταία χρόνια, οι τεχνολογίες νέας γενιάς αλληλούχισης (NGS) παρουσιάζουν μεγάλες βελτιώσεις όσον αφορά στα αποτελέσματα της αλληλούχισης, στο μήκος και

στην ακρίβεια των reads. Η μείωση του κόστους (εικ.4) για την πραγματοποίηση της αλληλούχισης καθώς και η ανάπτυξη του σταδίου ενίσχυσης πριν από την ανάλυση της αλληλουχίας συμβάλλουν ούτως ώστε τα NGS να θεωρούνται οι καλύτερες τεχνικές επιλογής για αλληλούχιση. Σύμφωνα με τον νόμο του Moore (εικ.5) η τεχνολογία αλληλούχισης συνδέεται άμεσα με την υπολογιστική πρόοδο όπως φαίνεται και παρακάτω και μάλιστα αναπτύσσεται πιο γρήγορα (Shokralla et al., 2012).



Εικ.4:κόστος ανά αλληλούχιση http://www.genome.gov/sequencingcost



Εικ.5:κόστος ανά γονιδίωμα (http://www.genome.gov/sequencingcosts/)

Οι τεχνολογίες νέας γενιάς αλληλούχισης μπορούν να ταξινομηθούν σε δύο κατηγορίες (εικ.6) :α) σε αυτές που στηρίζονται σε μεθόδους PCR, όπως είναι η

Roche 454 Genome Sequencer, HiSeq 2000 (illumina), AB SOLiD System, Ion Personal Genome Machine και β) στις ονομαζόμενες 'single-molecule' τεχνολογίες αλληλούχισης (SMS) που δεν χρησιμοποιούν τεχνικές PCR, όπως η HeliScope και η PacBio RS SMRT system (Shokralla et al., 2012). Η τεχνική SMS είναι απλή και το πλεονέκτημά της είναι ότι απαιτείται λιγότερο υλικό δείγματος (<1μg) (Metzker, 2010).

Εικ.6: Σύγκριση των NGS τεχνολογιών (Shokralla et al., 2012)

Category	Platform	Read length (bp)	Max. number of reads/run	Sequencing output/run	Run time
PCR-based NGS technologies	Roche 454 GS FLX	400–500	1×10^6	≤500 Mb	10 h
	Roche 454 GS FLX+	600–800	1×10^6	≤700 Mb	23 h
	Roche 454 GS Junior	400–450	1×10^5	~35 Mb	10 h
	Illumina HiSeq 2000	100–200	6×10^9	≤540–600 Gb	11 d
	Illumina HiSeq 1000	100–200	3×10^9	≤270–300 Gb	8.5 d
	Illumina GAIIx	50–75	6.4×10^8	≤95 Gb	7.5–14.5 d
	Illumina MiSeq	100–150	7×10^6	≤1–2 Gb	19–27 h
	AB SOLiD 5500 system	35–75	2.4×10^9	~100 Gb	4 d
	AB SOLiD 5500 xl system	35–75	6×10^9	~250 Gb	7–8 d
	Ion Torrent -314 chip	100–200	1×10^6	≥10 Mb	3.5 h
	Ion Torrent -316 chip	100–200	6×10^6	≥100 Mb	4.7 h
	Ion Torrent -318 chip	100–200	11×10^6	≥1 Gb	5.5 h
SMS technologies	Helicos HeliScope	30–35	1×10^9	~20–28 Gb	≤1 d
	Pacific Biosciences system	≥1500	50×10^3	~60–75 Mb	0.5 h

Τα NGS μπορούν να αναλύσουν μεγάλα γονιδιώματα σε μικρό σχετικά χρονικό διάστημα , από έναν ή περισσότερους οργανισμούς με αποτέλεσμα την δυνατότητα σύγκρισης αυτών ανάλογα με την εξελικτική τους πορεία. Ένα μείζον πλεονέκτημα των NGS το οποίο θα έχει σημαντικές επιπτώσεις στην ανθρώπινη κοινωνία, είναι η εύρεση γενετικών διαφορών (με ανθρώπινη αλληλούχιση γονιδιώματος) και η ανάλυση αυτών στο πώς επιδρούν σε υγιή αλλά και σε ασθενή άτομα. Επομένως,θα απαντηθούν πολλά ερωτήματα για το πώς τα γονίδια και οι μεταλλάξεις αυτών συμβάλλουν στον ανθρώπινο φαινότυπο (Metzker, 2010).

Με την χρήση των NGS μπορεί να γίνουν τα εξής είδη αλληλούχισης:

- Whole genome sequencing
- Exome sequencing (targeted regions)
- RNA sequencing
- Chip sequencing

(Shendure et al., 2008)

Οι πλατφόρμες νέας γενιάς τεχνολογιών αλληλούχισης (NGS), οι οποίες υπάρχουν στην αγορά από τις διάφορες εταιρίες, είναι πολλές, με ξεχωριστά πλεονεκτήματα και

μειονεκτήματα η κάθε μία. Παρ' όλα αυτά, όλες έχουν ένα κοινό χαρακτηριστικό μοτίβο.

- Template preparation → α) κλωνικά ενισχυμένα πρότυπα που προέρχονται από απλά μόρια DNA ή
  β) μονό πρότυπο DNA (SMS)

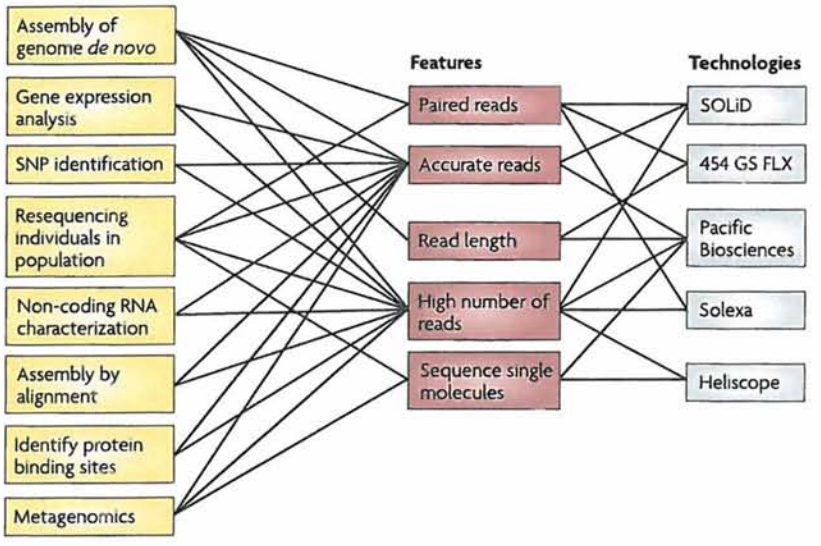
- Sequencing και imaging (αλληλούχιση και απεικόνιση)
  Η απεικόνιση στηρίζεται στα τέσσερα διαφορετικά χρώματα που σηματοδοτούν οι αζωτούχες βάσεις του μορίου που μελετάται.

- Data analysis (ανάλυση δεδομένων)
  Τα δεδομένα που προέρχονται από πλατφόρμες νέας γενιάς τεχνολογιών αλληλούχισης, χρειάζονται ειδική διαχείριση με την βοήθεια της πληροφορικής όπως είναι η αποθήκευση αυτών των δεδομένων, η ανάλυσή τους καθώς και ο ποιοτικός έλεγχος αυτών (Metzker, 2010).

Κάποιες από τις πιο διαδεδομένες πλατφόρμες νέας γενιάς τεχνολογιών αλληλούχισης της εποχής μας είναι:

- Roche 454 - Pyrosequencing (Roche Diagnostics Corp., Branford, CT, USA)
- Illumina/Solexa (Inc., San Diego, CA, USA)
- Nanopore Oxford Technologies
- Ion Torrent (Life Technologies, South San Francisco, CA, USA)
- Pacific Biosciences

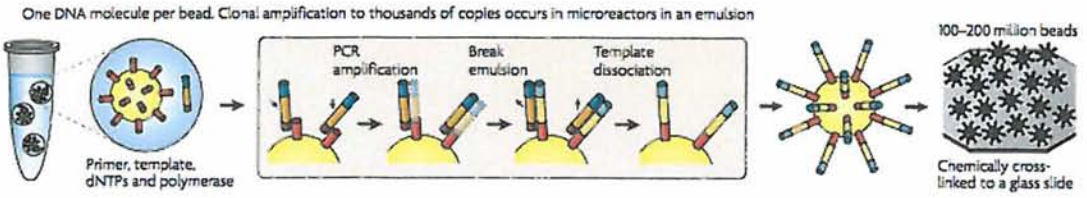
Η επιλογή της πλατφόρμας γίνεται σύμφωνα με το είδος της ανάλυσης που θέλουμε να γίνει. Το παρακάτω διάγραμμα (εικ.7) δείχνει μία πειραματική προσέγγιση στη επιλογή πλατφόρμας.



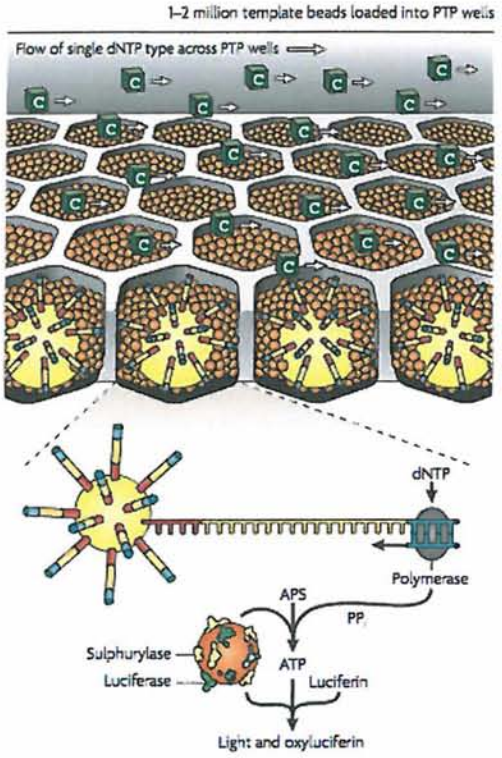
Εικ.7: Προτεινόμενος τρόπος επιλογής πλατφόρμας ανάλογα με την πειραματική προσέγγιση (MacLean et al., 2009)

1.2.1 Roche 454 – Pyrosequencing

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 (εικ.8).



Εικ.8: Μίγμα που αποτελείται από υδατικό έλαιο-γαλάκτωμα δημιουργεί εγκλεισμό σφαιριδίων DNA συμπλοκών σε ενιαία υδατικά σταγονίδια. Ενίσχυση PCR εκτελείται εντός αυτών, για τη δημιουργία σταγονιδίων που περιέχουν σφαιρίδια με χιλιάδες αντίγραφα της ίδιας αλληλουχίας προτύπου. Τα σφαιρίδια μπορούν να προσαρτηθούν χημικά σε μία γυάλινη αντικειμενοφόρο πλάκα ή σε PicoTiter πλάκα. Πολλαπλασιασμός στερεής-φάσης. (Metzker, 2010).



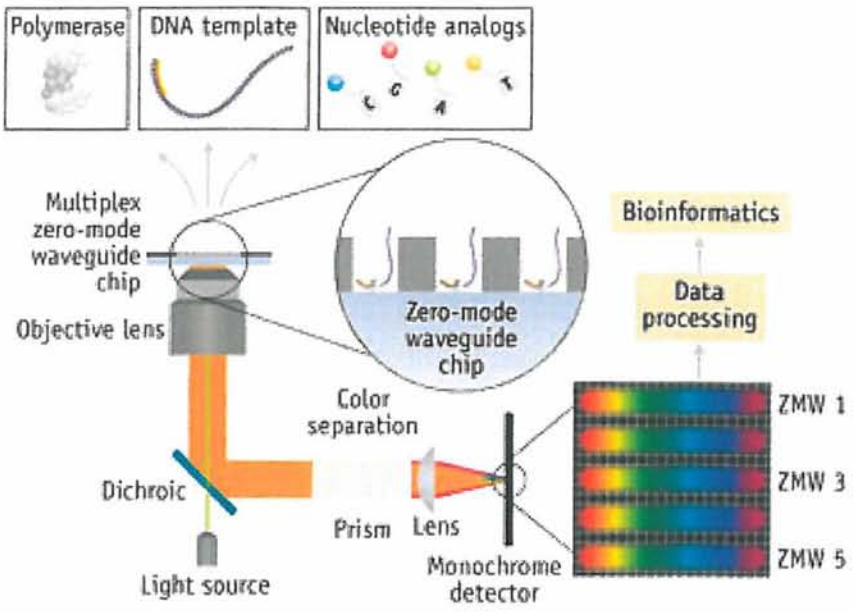
Εικ.9: Roche 454 – Pyrosequencing (Metzker, 2010)

Συγκεκριμένα, σε μία Picotiter πλάκα (εικ.9), κάθε νουκλεοτίδιο δεσμεύεται από την DNA πολυμεράση με αποτέλεσμα την απελευθέρωση πυροφωσφορικού μορίου (Shokralla et al., 2012). Ένζυμα ATP σουλφορυλάσης και λουσιφεράσης, μετατρέπουν τα πυροφωσφορικά μόρια ώστε να γίνει 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. Το μειονέκτημα της τεχνικής είναι η ασύγχρονη χημεία επέκτασης. Τα τελευταία χρόνια, έχουν γίνει ορισμένες βελτιώσεις στην απόδοση της αλληλουχίας των ομοπολυμερών (Rodriguez – Ezpeleta et al., 2012).

1.2.2 Pacific Biosciences

Το 2010, η εταιρία Pacific Biosciences παρουσίασε μια πλατφόρμα DNA – αλληλούχισης μονόκλωνου μορίου σε πραγματικό χρόνο (SMRT – Single Molecule Real Time). Η συγκεκριμένη τεχνολογία αλληλούχισης στηρίζεται σε πραγματικό χρόνο με βάση τον φθορισμό. Για την προετοιμασία του δείγματος δεν απαιτείται στάδιο ενίσχυσης καθώς πραγματοποιείται προσέγγιση αλληλούχισης με σύνθεση μονόκλωνου μορίου (Shokralla et al., 2012). Η πρωτοπορία της πλατφόρμας στηρίζεται σε δύο τεχνολογίες: α) στην σύνδεση της φθορίζουσας χρωστικής με την 5' φωσφορική ομάδα του κάθε δεοξυριβονουκλεοτιδίου και όχι με την βάση (παλιά μέθοδος) και β) στην χρησιμοποίηση μιας νανο-δομής, ονομαζόμενη ως Zero Mode Waveguide (ZMW) στην οποία παίρνει μέρος ο πολυμερισμός του DNA, σε πραγματικό χρόνο (Niedringhaus et al., 2011).

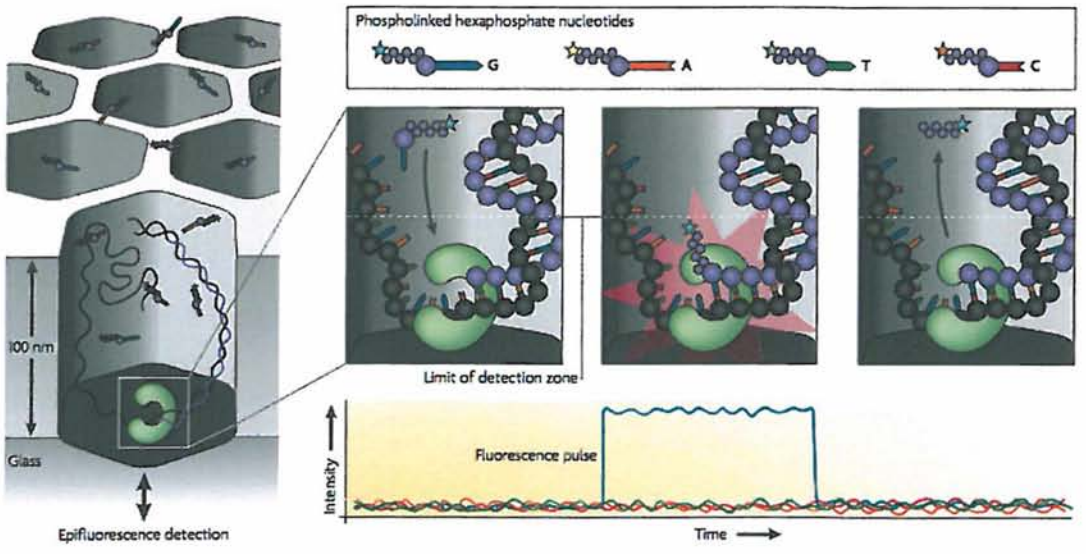
Η νανο-συσκευή Zero Mode Waveguide (ZMW) αποτελείται από δεκάδες χιλιάδες οπές, με διάμετρο νανομέτρων. Κατασκευάζεται με διάτρηση μιας λεπτής μεταλλικής μεμβράνης και υποστηρίζεται από ένα διαφανές υπόστρωμα (Shokralla et al., 2012).



Εικ.10: Τεχνολογία αλληλούχισης της Pacific Biosciences. Η DNA πολυμεράση , ο εκκινητής και τα μόρια της βιβλιοθήκης (δημιουργία συμπλόκου) εισάγονται στην επιφάνεια της νανο-συσκευής (ZMWs). Το σύμπλοκο ακινητοποιείται στον πυθμένα των ZMWs, και προστίθενται φθορίζοντα νουκλεοτίδια. Η οπτική κάμερα και το σύστημα λέιζερ καταγράφει σε πραγματικό χρόνο την δραστικότητα της DNA πολυμεράσης με τα φθορίζοντα μόρια. Κατά τη διαδικασία της ενσωμάτωσης αυτών, η φθορίζουσα ομάδα διαχέεται μακριά, και ακολουθεί μετατόπιση του κλώνου, κάνοντας έτσι χώρο για την ανάγνωση του επόμενου νουκλεοτιδίου της αλυσίδας.
(http://www.bioopticsworld.com/articles/print/volume-5/issue-06/features/dna-sequencing-technologies-the-next-generation-and-beyond.html)

Η διαδικασία της αλληλούχισης, με την τεχνολογία της Pacific Biosciences, υλοποιείται σε δύο στάδια (εικ.10). Στο πρώτο στάδιο, ένα set διαφορετικών φθοριζουσών χρωστικών για την κάθε βάση ενώνεται στην 5' φωσφορική ομάδα του κάθε δεοξυριβονουκλεοτιδίου, ούτως ώστε η κάθε βάση να είναι φασματικά διαφορετική, χωρίς όμως να επηρεάζεται η λειτουργική ικανότητα της DNA πολυμεράσης. Στο δεύτερο στάδιο, το μονόκλωνο μόριο DNA με την DNA πολυμεράση ακινητοποιείται στον πυθμένα του πόρου ZMW της νανο-συσκευής. Τα σύμπλοκα των διαφόρων φθοριζουσών χρωστικών με τις βάσεις κινούνται ελεύθερα

στο χώρο. Κάθε χρονική στιγμή η DNA πολυμεράση (εικ.11) διαλέγει το κατάλληλο σύμπλοκο προς τη βάση του μονόκλωνου DNA. Η βάση του συμπλόκου προσδένεται με την συμπληρωματική της στην μονόκλωνη αλυσίδα και η φθορίζουσα χρωστική απομακρύνεται (Niedringhaus et al., 20011). Στάδιο πλύσης μεταξύ της ροής κάθε νουκλεοτιδίου δεν χρειάζεται με αποτέλεσμα την επιτάχυνση της νουκλεοτιδικής ενσωμάτωσης, καθώς και την βελτίωση της ποιότητας αλληλουχίας (Shokralla et al., 2012). Κατά την απομάκρυνση, το σύστημα λέιζερ καταγράφει το σήμα της κάθε φθορίζουσας χρωστικής. Η ίδια διαδικασία επαναλαμβάνεται μέχρι την ολοκλήρωση της αντιγραφής της μονόκλωνης αλυσίδας και έτσι η αλληλούχιση του δείγματος είναι έτοιμη (Niedringhaus et al., 2011).



Εικ.11: Τεχνολογία αλληλούχισης της Pacific Bioscience (Metzker et al., 2010)

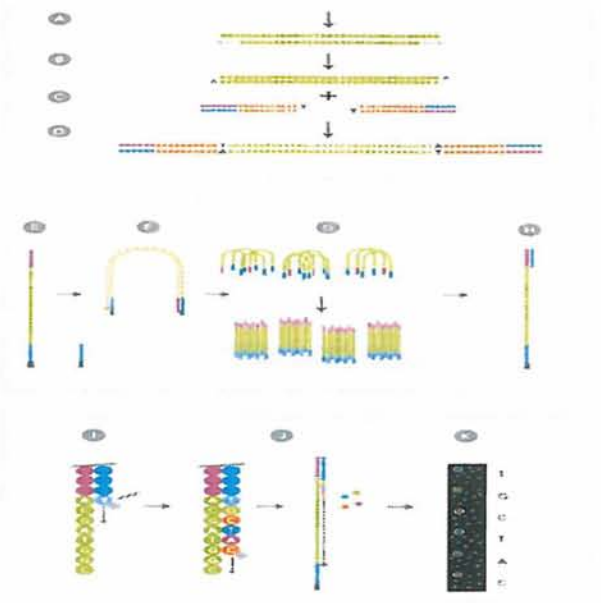
Συμπερασματικά, το είδος της τεχνολογίας που προώθησε η Pacific Biosciences, χρησιμοποιεί τη φυσική ικανότητα της DNA πολυμεράσης να ενσωματώνει δέκα ή περισσότερα νουκλεοτίδια ανά δευτερόλεπτο σε αρκετές χιλιάδες παράλληλες οπές ZMWs (Shokralla et al., 2012). Η αλληλούχιση γίνεται με γρήγορο ρυθμό και δημιουργεί μεγάλου μήκους reads (της τάξεως των 1000 βάσεων) κάτι που αυξάνει την ποιότητα της συναρμολόγησης και τον εντοπισμό SNP (Single- Nucleotide Polymorphisms) (Niedringhaus et al., 2011).

1.2.3 Illumina/Solexa

Το 2007, η εταιρεία Illumina απέκτησε την Solexa, η οποία ανέπτυξε μια πολύ επιτυχημένη τεχνολογία αλληλούχισης των γονιδιωμάτων. Η συγχώνευση αποτέλεσε κλειδί στην ανάπτυξη εργαλείων και ανεπτυγμένων μηχανημάτων αλληλούχισης. Κάποια από τα μηχανήματα της illumina είναι το HiSeq System, HiScan SQ, Genome Analyzer MiSeq. Η επιλογή χρήσης του κάθε μηχανήματος εξαρτάται από το μέγεθος της αλληλουχίας που είναι προς αλληλούχιση. (http://www.illumina.com)

Η τεχνοτροπία πίσω από τα NGS είναι παρόμοια με αυτή της ηλεκτροφόρησης με τριχοειδή (capillary electrophoresis - CE) δηλαδή ο προσδιορισμός των αζωτούχων βάσεων ενός θραύσματος γίνεται από τα σήματα που εκπέμπονται. Κάθε θραύσμα δημιουργείται εκ νέου από ένα κλώνο εκμαγείο. Η τεχνολογία της illumina προσφέρει τη δυνατότητα δημιουργίας εκατομμυρίων αντιδράσεων διαφορετικών δειγμάτων με μαζικό παράλληλο τρόπο χωρίς τον περιορισμό λίγων θραυσμάτων (παλιές τεχνολογίες). Επομένως, μπορούν να αναλυθούν πολλά δείγματα μαζί σε ένα μόνο τρέξιμο. Αυτό επίσης συνεπάγεται και μείωση του χρόνου της αλληλούχισης. (http://www.illumina.com)

Ο τρόπος κατά τον οποίο γίνεται αλληλούχιση γονιδιωμάτων με illumina είναι ο εξής:

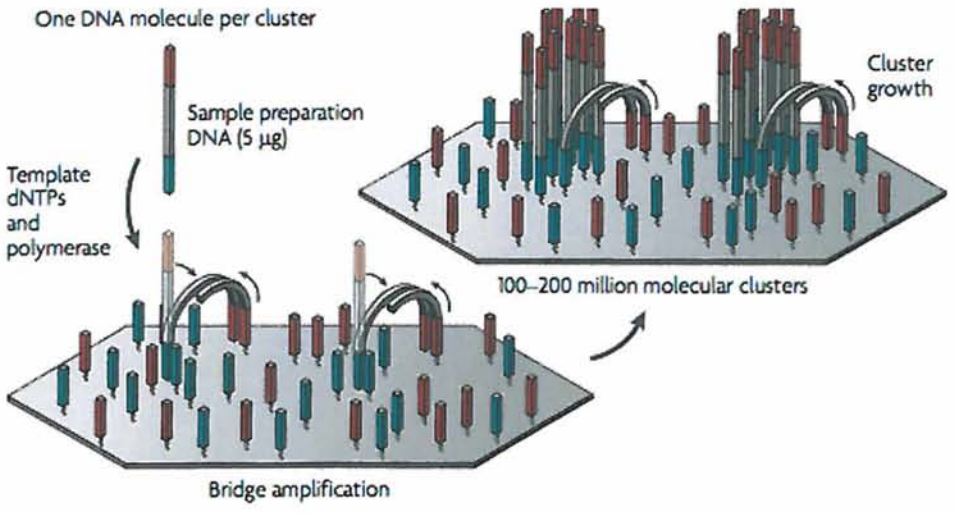


Εικ.12: Διαδικασία τεχνικής αλληλούχισης Illumina (Ansorge, 2009)

I. Προετοιμασία βιβλιοθήκης (Library Preparation)

Τα δίκλωνα μόρια του DNA των δειγμάτων τεμαχίζονται με την βοήθεια ενζύμων σε τυχαία κομμάτια (θραύσματα) (εικ.12A). Ένα ολιγονουκλεοτίδιο 'T' προσδένεται στα θραύσματα και προεξέχει (εικ.12B) (Ansorge, 2009). 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 , κατά την ανάλυση των δεδομένων , με το δείγμα στο οποίο ανήκει. Μετά την σύνδεση των adapters με τα θραύσματα του DNA γίνεται αποδιάταξη των δίκλωνων μορίων σε μονόκλωνα, (εικ.12C) (www.illumina.com/NGS).

II. Δημιουργία συμπλέγματος-cluster (Cluster Generation)

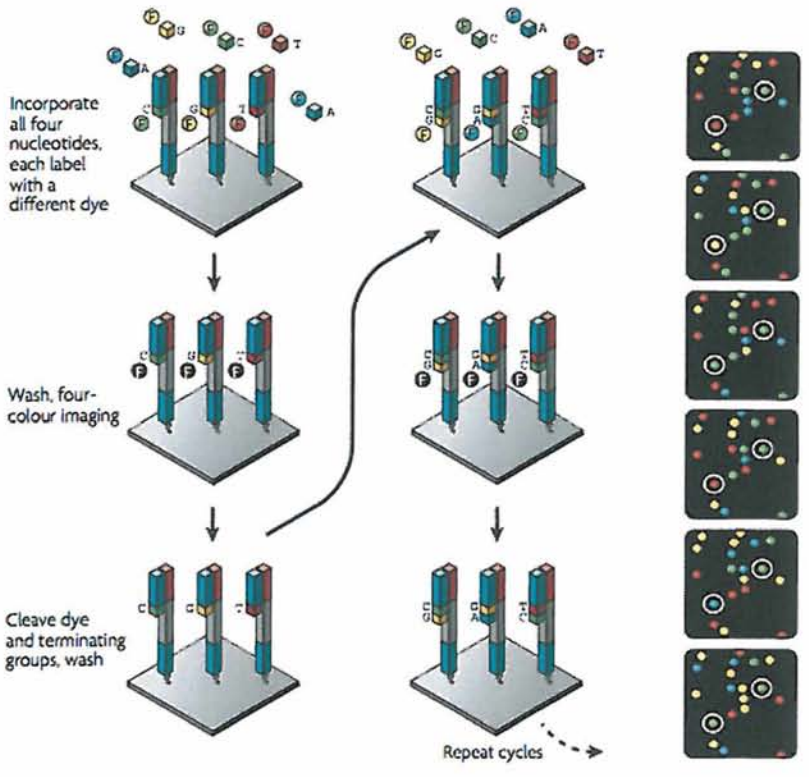


Εικ.13:Ενίσχυση στερεάς φάσης (Metzker et al., 2010)

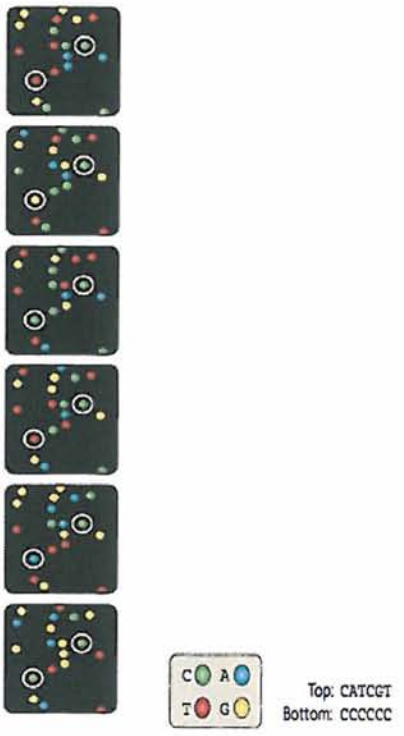
Τα μονόκλωνα μόρια τοποθετούνται πάνω σε μία επιφάνεια-πλάκα εργασίας, (εικ.12 E), (workflow-glass flow cell). Η κάθε πλάκα αποτελείται εσωτερικά από ολιγονουκλεοτίδια τα οποία είναι συμπληρωματικά ως προς τους adapters, και χωρίζεται σε οχτώ ξεχωριστές λωρίδες. Πραγματοποιείται υβριδισμός (μέσω εναλλαγής 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 (εικ.13) – bridge amplification (εικ.12F) (Mardis, 2008). Μία ισοθερμική πολυμεράση ενισχύει για την δημιουργία κλώνων (εικ.12G). Επίσης, και οι adapters της πλάκας δρουν ως εκκινητές για την ενίσχυση (Zhou et al., 2010). Η κάθε βιβλιοθήκη θραυσμάτων αποτελείται πλέον από εκατοντάδες εκατομμύρια μοναδικά συμπλέγματα (clusters). Τα συμπληρωματικά συμπλέγματα αποκόπτονται και απομακρύνονται ξεπλένοντας. Μετά την δημιουργία των clusters οι βιβλιοθήκες είναι έτοιμες προς αλληλούχιση. (www.illumina.com/NGS).

III. Αλληλούχιση (Sequencing)



Εικ.14:Μέθοδος 4 φθοριζουσών χρωστικών αλληλούχισης (Metzker et al., 2010)



Εικ.15:Απεικόνιση βάσεων με CCD camera (Metzker et al., 2010)

Η αλληλούχιση όλων των cluster (εικ.14) γίνεται ταυτόχρονα βάση προς βάση με παράλληλο τρόπο χρησιμοποιώντας τέσσερις διαφορετικές φθορίζουσες χρωστικές συνδεδεμένες με τέσσερα διαφορετικά ολιγονουκλεοτίδια (A.T,G και C) (Zhou et al., 2010). Οι τέσσερις φθορίζουσες με τις βάσεις πλησιάζουν την βάση του cluster αλλά μόνο μία θα ενωθεί μαζί της (εικ.12I). 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 (εικ.12K). Το ίδιο γίνεται και για την επόμενη βάση της αλυσίδας του cluster μέχρι να τερματιστεί (εικ.15). Έτσι δημιουργούνται συμπληρωματικές αλυσίδες των clusters (εικ.12J).

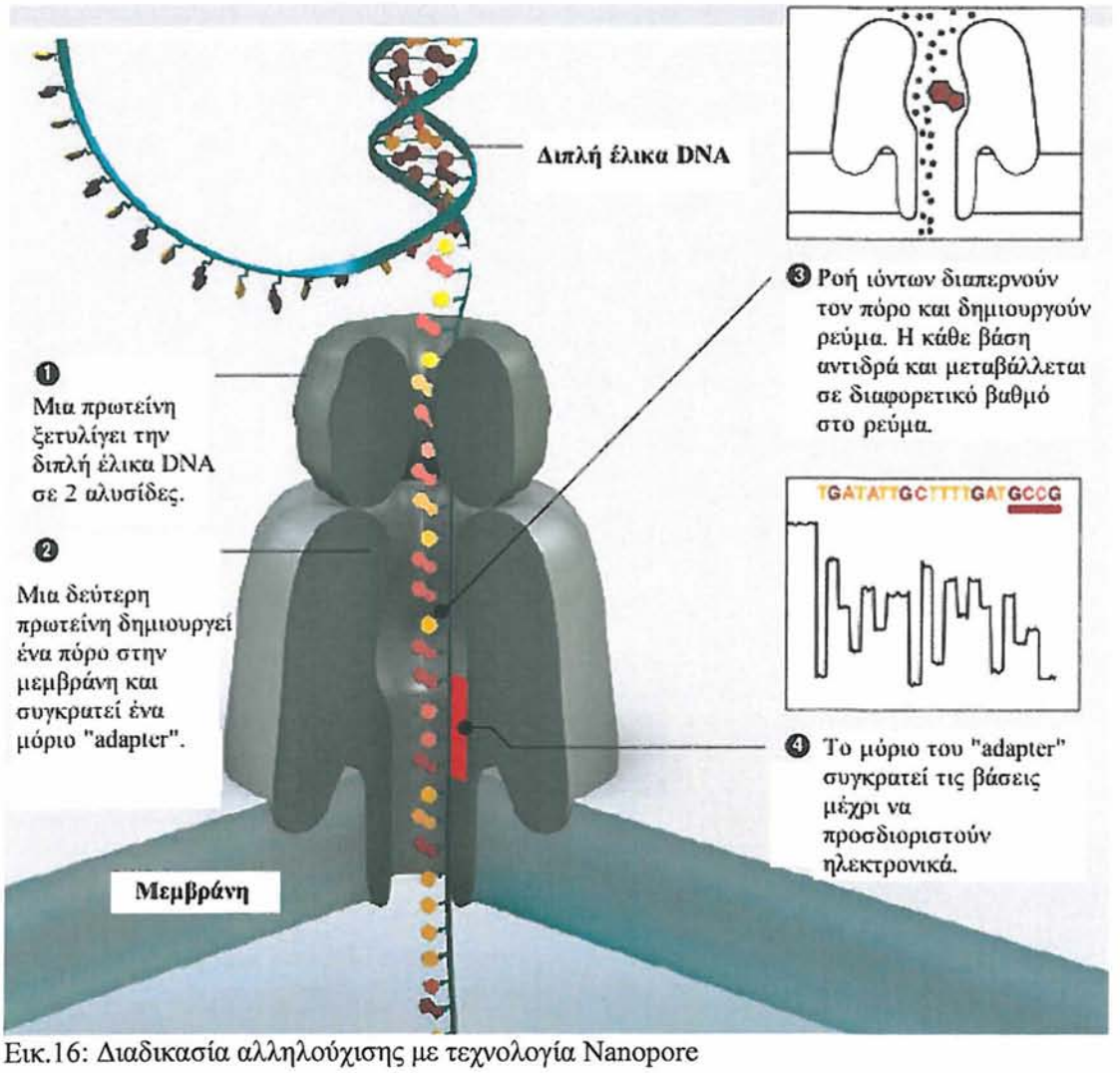
Το ποσοστό σφάλματος της αλληλούχισης μπορεί να μειωθεί εάν η επικάλυψη των reads είναι μεγάλη.

1.2.4 Nanopore Oxford Technologies

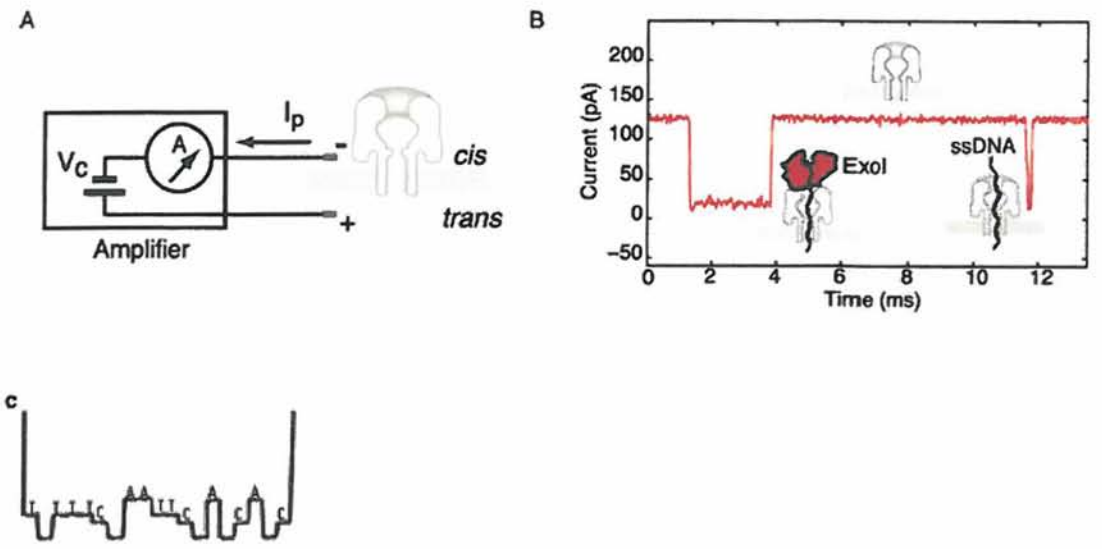
Στο μέλλον, πιστεύεται ότι η τρίτη γενιά αλληλούχισης θα πραγματοποιείται από array που θα αποτελούνται από μικρούς σε σχήμα πρωτεϊνικούς πόρους (nanopores). Η ιδέα της αλληλούχισης, βασισμένη σε nanopores, προτάθηκε πριν από περίπου 20 χρόνια. Τα πρώτα πειράματα διεξήχθησαν το 1996, χρησιμοποιώντας ως πόρους την βιολογική πρωτεΐνη α-αιμολυσίνη. Τα αποτελέσματα των πειραμάτων δεν ήταν τα επιθυμητά καθώς η μέθοδος ήταν ατελής και η τεχνολογική γνώση δεν επαρκούσε (Schneider et al., 2012). Σήμερα, η μέθοδος αυτή είναι πολλά υποσχόμενη, εξαιτίας της ανάπτυξης της τεχνολογίας, καθώς υπολογίζεται ότι θα είναι πιο φτηνή σε κόστος και πιο γρήγορη με ελάχιστη ποσότητα δείγματος. (Akan et al., 2010). Ένα ακόμη πλεονέκτημα αυτής της μεθόδου, είναι η μη σήμανση του μορίου, χωρίς ενίσχυση του δείγματος, για την ανίχνευση των αζωτούχων βάσεων (Anselmetti, 2012). Επίσης, αναμένεται να προσφέρει λύσεις στον περιορισμό των τεχνολογιών αλληλούχισης με short reads και να καταστήσει δυνατή την αλληλούχιση μεγάλων μορίων σε μερικά λεπτά, χωρίς την τροποποίηση ή την προετοιμασία δειγμάτων (Rodriguez-Ezpeleta et al., 2012).

Η τεχνολογία των nanopores για την ανάλυση των νουκλεϊκών οξέων αποτελείται από δύο προσεγγίσεις: i) τα βιολογικά nanopores (Biological nanopores) όπως είναι η αιμολυσίνη και ii) τα συνθετικά nanopores στερεής κατάστασης (Solid-state-Graphene) nanopore. Η κεντρική ιδέα κατά τη διαδικασία της αλληλούχισης με nanapore και στις δύο προσεγγίσεις είναι ίδια, με τη διαφορά ότι τα βιολογικά nanopores στηρίζονται σε μια λιπιδική διπλοστιβάδα ενώ τα συνθετικά nanopores

στηρίζονται σε μία συνθετική μεμβράνη (Rhee et al., 2007). Συγκεκριμένα, το μόριο του δίκλωνου DNA εισέρχεται σε ένα πρωτεϊνικό πόρο (εικ.16). Στον πρωτεϊνικό πόρο στηρίζεται μία άλλη πρωτεΐνη (πολυμεράση) η οποία κατα την διέλευση του δίκλωνου DNA, το ξετυλίγει και έτσι δημιουργούνται δύο αλυσίδες (Venkatesan et al., 2011). Εσωτερικά του πόρου, καθώς ξετυλίγεται η μονόκλωνη αλυσίδα του DNA, περνούν ιόντα και εφαρμόζεται τάση ρεύματος (εικ.17) στα άκρα της μεμβράνης (Rodriguez-Ezpeleta et al., 2012). Το δυναμικό που εφαρμόζεται διαμέσου της μεμβράνης, δημιουργεί ιονικό ρεύμα (Maitra et al., 2012). Στο δυναμικό που δημιουργείται , σε σχέση με την πολικότητα του μορίου, στηρίζεται ο εντοπισμός των βάσεων (A,T,G ή C) την κάθε χρονική στιγμή και ύστερα γίνεται η καταγραφή αυτών (Venkatesan et al., 2011).



Εικ.16: Διαδικασία αλληλούχισης με τεχνολογία Nanopore
(http://www2.technologyreview.com/article/427677/nanopore-sequencing/)

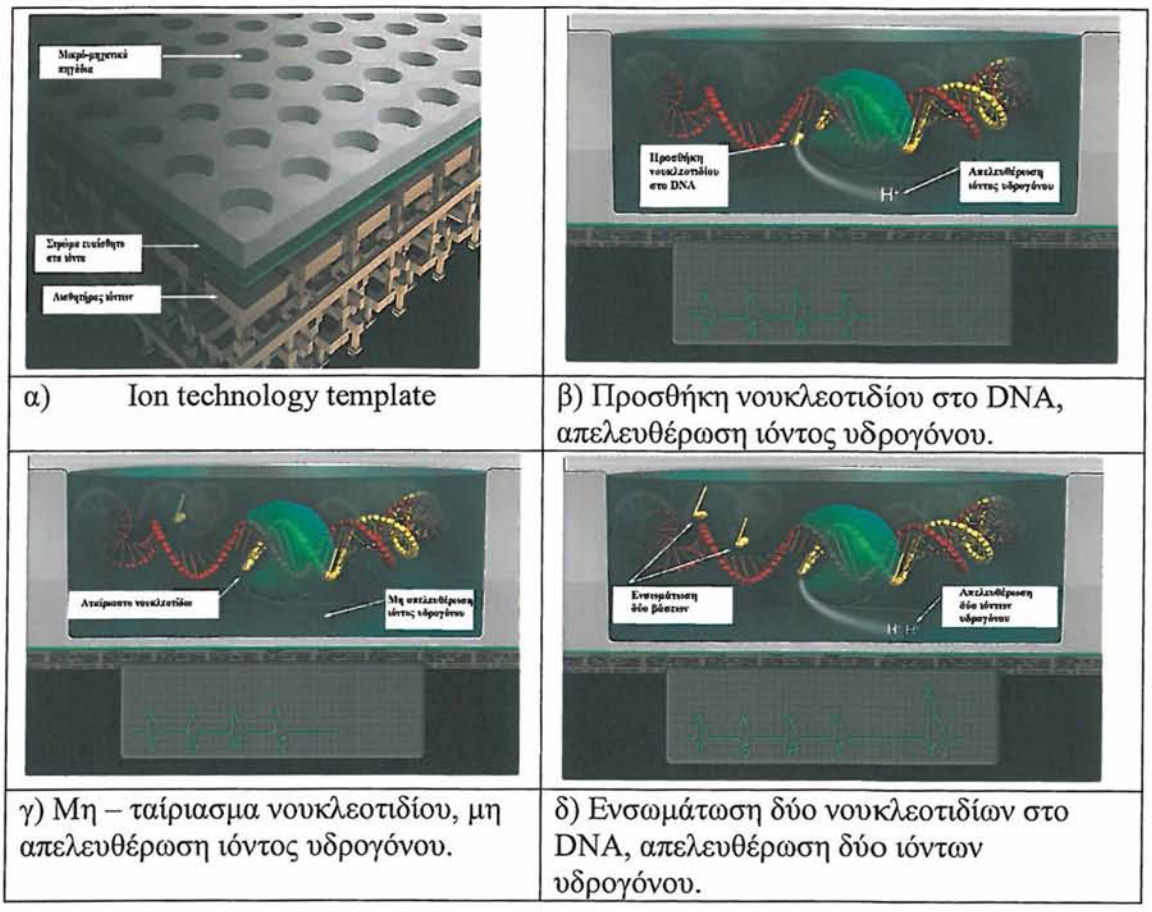


Εικ.17:Α) Δημιουργία διηλεκτρικής έντασης στην μεμβράνη για τον εντοπισμό των βάσεων. (Maitra et al., 2012) Β) Γράφημα βάσεων αλληλουχίας από αλληλούχιση με Nanopore. C) Γράφημα υποθετικής ανάγνωσης αλληλουχίας από αλληλούχιση με Nanopore. (http://www.nature.com/scientificamerican/journal/v294/n1/box/scientificamerican0106-46_BX4.html)

Τέλος, υπάρχουν πολλά αναπάντητα ερωτηματικά για την τεχνική αλληλούχισης Graphene nanopore, όσον αφορά τη χημεία που χρησιμοποιεί και για το πόσο ακριβή (πιστή) είναι η μέθοδος αυτή.

1.2.5 Ion Torrent

Το 2010, η Life Technologies παρουσίασε μία καινούργια τεχνολογία αλληλούχισης την Ion Torrent (Shokralla et al., 2012). Η συγκεκριμένη τεχνολογία (εικ.18) δημιουργεί μια άμεση σύνδεση μεταξύ χημικών και ψηφιακών πληροφοριών, επιτρέποντας έτσι τη γρήγορη, απλή και μαζικά κλιμακούμενη αλληλούχιση. Χρησιμοποιεί την απλή χημεία του νουκλεϊκού οξέος, σύμφωνα με τον Watson, σε μια απίστευτα ισχυρή, τεχνολογία ημιαγωγών. Η αρχή της τεχνολογίας Ion Torrent βασίζεται σε μια αναπτυγμένη βιοχημική διαδικασία, στην οποία ένα νουκλεοτίδιο ενσωματώνεται σε μια αλυσίδα του DNA από μια πολυμεράση, με αποτέλεσμα την απελευθέρωση ιόντων υδρογόνου, ως παραπροϊόν (Pareek et al., 2011).



α) Ion technology template

β) Προσθήκη νουκλεοτιδίου στο DNA, απελευθέρωση ιόντος υδρογόνου.

γ) Μη – ταίριασμα νουκλεοτιδίου, μη απελευθέρωση ιόντος υδρογόνου.

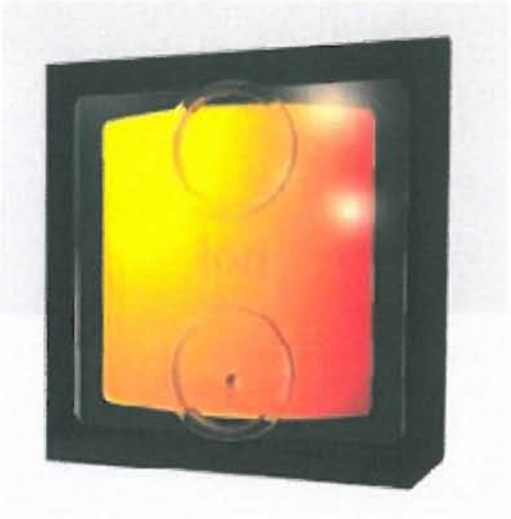
δ) Ενσωμάτωση δύο νουκλεοτιδίων στο DNA, απελευθέρωση δύο ιόντων υδρογόνου.

Εικ.18: Μέθοδος αλληλούχισης Ion Torrent (http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Sequencing/Semiconductor-Sequencing/Semiconductor-Sequencing-Technology/Ion-Torrent-Technology-How-Does-It-Work.html)

Η εκτέλεση της βιοχημικής διαδικασίας γίνεται με μαζικό και παράλληλο τρόπο, σε chip υψηλής πυκνότητας μικρο-μηχανικών πηγαδιών. Κάθε πηγάδι δέχεται ένα μόνο πρότυπο DNA από βιβλιοθήκη. Κάτω από το πηγάδι είναι ένα ιοντικά-ευαίσθητο στρώμα και ένας αισθητήρας ιόντων για την ανίχνευση της αλλαγής της συγκέντρωσης των ιόντων υδρογόνου λόγω της ενσωμάτωσης των νουκλεοτιδίων. (Shokralla et al., 2012).

Το 2012, η Life Technologies παρουσίασε μια νέα γενιά αλληλούχισης, την Ion Proton. Η Ion Proton είναι ένα chip (εικ.19) το οποίο έχει τη δυνατότητα αλληλούχισης του ανθρώπινου γονιδίωματος και του ανθρώπινου εξονιώματος μέσα σε λίγες ώρες (Pareek et al., 2011). Το chip χρησιμοποιεί την τεχνολογία ημιαγωγών CMOS, παρόμοια με εκείνη των ψηφιακών φωτογραφικών μηχανών με την διαφορά

ότι αντί για τον εντοπισμό φωτός, το chip αναγνωρίζει την χημεία των μορίων και το μεταφράζει- αποκωδικοποιεί κατευθείαν σε ψηφιακά δεδομένα. (http://www.vincentabry.com/en/ion-proton-sequencer-decodes-human-genome-in-1-day-for-1000-dollars-1543)



Εικ.19: Ion Proton chip
(http://www.vincentabry.com/en/ion-proton-sequencer- decodes-human-genome-in-1-day-for-1000-dollars-1543)

1.3 Αλληλούχιση εξονιώματος

Οι πρόσφατες εξελίξεις στην ανάπτυξη νέων τεχνολογιών αλληλούχισης NGS) έχουν κάνει την αλληλούχιση εξονιώματος μια τεχνικά εφικτή αλλά και εξαιρετικά αποδοτική μέθοδο επιλογής (Ku et al., 2012).

Τα εξόνια είναι οι περιοχές των γονιδίων τα οποία κωδικοποιούν τις πρωτεΐνες. Το εξονίωμα, είναι όλα τα εξόνια ενός γονιδιώματος. Το εξονίωμα αποτελεί περίπου το 1% του συνολικού γονιδιώματος με 180,000 εξώνια, στον άνθρωπο (Ng et al., 2009). Η αλληλούχιση του εξονιώματος δίνει το πλεονέκτημα στους ερευνητές να μπορούν να εστιάσουν μόνο σε περιοχές με υψηλή βιολογική πληροφορία. (www.illumina.com). Στόχος είναι η κατανόηση των αλληλομορφικών παραλλαγών και η σχέση που έχουν με τον φαινότυπο. Υψηλός εστιασμός γίνεται στα Μενδελικά νοσήματα καθώς οι περισσότερες μεταλλάξεις που ευθύνονται για αυτά συμβαίνουν στα εξόνια (Biesecker et al., 2011).

Πλεονεκτήματα του Exome sequencing

Είναι μία μέθοδος πιο φτηνή και πιο γρήγορη, περίπου 6-10 φορές πιο φτηνή σε σχέση με την αλληλούχιση ολόκληρου του γονιδιώματος (Singleton et al., 2011). Είναι μία τεχνική εύκολη στη χρήση και επιτέπει την ανίχνευση των αιτιών που προκαλούν Μενδελικά νοσήματα (Biesecker et al., 2011), την κατανόηση των αλληλομορφικών παραλλαγών και την σχέση που έχουν με τον φαινότυπο (Singleton et al., 2011).

Μειονεκτήματα του Exome sequencing

Αρκετοί παράγοντες, καθιστούν δύσκολη την στόχευση και τον εμπλουτισμό μόνο του εξονιώματος. Η αλληλούχιση του ανθρώπινου εξονιώματος καλύπτει μόνο το 1% του γονιδιώματος με αποτέλεσμα την μη κάλυψη των ιντρονίων καθώς και άλλων περιοχών που μπορεί να παίζουν σημαντικό λειτουργικό ρόλο και πολύ πιθανόν να επηρεάζουν και τα συμπτώματα της ασθένειας (Mamanova et al., 2010).

Εφαρμογές του Exome sequencing

Η επιτυχής ανάπτυξη της αλληλούχισης του εξονιώματος έχει σημαντικά αποτελέσματα τόσο σε ερευνητικό επίπεδο όσο και στον κλινικό τομέα. Κάποιες από τις εφαρμογές είναι :

1. Εύρεση μενδελικών ασθενειών

   Μία από τις πιο συχνές εφαρμογές της αλληλούχισης ανθρώπινου εξονιώματος είναι η ανίχνευση μενδελικών ασθενειών. Μέχρι στιγμής υπολογίζεται ότι πάνω από 3.000 γονίδια ευθύνονται για μενδελικά νοσήματα. Η ανάλυση του ανθρώπινου εξωνιόματος φέρει καινούριες στρατηγικές εύρεσης μενδελικών νοσημάτων σε οικογένειες με κληρονομούμενο μενδελικό γενετικό νόσημα, ανιχνεύοντας την γενωμική περιοχή του γονιδίου που προκαλεί την μετάλλλαξη (Magewski et al., 2011). Από τον Νοέμβριο του 2009 η αλληλούχιση του εξονιώματος έχει οδηγήσει στην ταυτοποίηση για πάνω απο 30 νέα γονίδια στις μενδελικές ασθένειες. Βέβαια, δεν μπορεί να γίνει πάντα ανίχνευση μιας καινούριας γενετικής ασθένειας με τη χρήση της στρατηγικής του Exome sequencing. Μια ασθένεια μπορεί να προκληθεί από μεταλλάξεις σε διαφορετικά γονίδια και να προκαλέσει τον ίδιο φαινότυπο (ασθένειας). Ένα άλλο μεγάλο πλεονέκτημα είναι η στην ανίχνευση σπάνιων σποραδικών μενδελικών διαταραχών οι οποίες προκαλούνται από de novo μεταλλάξεις όπως για παράδειγμα τα σύνδρομα schinzil-giedion και kabuki. Εκτός από την ανίχνευση σπάνιων μενδελικών de novo νοσημάτων οι μελέτες επικεντρώθηκαν στο ρόλο των κοινών νευρολογικών διαταραχών όπως είναι η διανοητική καθυστέρηση και ο αυτισμός. Η ανίχνευση τους μπορεί να βελτιώσει την διάγνωση εκατομμυρίων ατόμων με μενδελικά νοσήματα (εικ.20),να βελτιώσει τη συμβουλευτική στις οικογένειες και να απελευθερώσει καινούριους θεραπευτικούς στόχους. (Gilissen et al., 2011).

2. Καρκίνος

   Η εφαρμογή των NGS τεχνολογιών έχει προσφέρει αρκετά πλεονεκτήματα στη γονιδιωματική του καρκίνου. Η μαζική παράλληλη αλληλούχιση έκανε εφικτή την κατηγοριοποίηση των σωματικών μεταλλάξεων στον καρκίνο. Μια μεγάλη πρόκληση στη γονιδιωματική ανάλυση του καρκίνου είναι η ανίχνευση της καθοδηγητικής μετάλλαξης (driver mutation). Πρόσφατες

μελέτες για τη λευχαιμία, το μυέλωμα,τους συμπαγείς όγκους συμπεριλαμβανομένων του μαστού, του πνεύμονα και του παγκρέατος έχουν εστιάσει στην ανάλυση στις κωδικές περιοχές (Exome) για να αυξήσουν την πιθανότητα της ανίχνευσης της μετάλλαξης. Αυτό έχει ως αποτέλεσμα τη δημιουργία νέων βιο-δεικτών και τη παρασκευή στοχευμένων φαρμάκων. Παρόλα αυτά η επίτευξη αυτών των στόχων δεν είναι πάντα δυνατή καθώς η ανάλυση των καρκινικών δειγμάτων και καρκινικών γονιδιωμάτων έχουν χαρακτηριστικά τα οποία μπορεί να διαφέρουν από άλλα δείγματα ιστών και γενωμικών αλληλουχιών (Meyersoe et al., 2010).

3. Ανθρώπινη εξέλιξη

   Μεγάλη πρόοδος της στοχευμένης αλληλούχισης έχει παρατηρηθεί στην αλληλούχιση αρχαίου DNA. Οι ερευνητές χρησιμοποίησαν αυτή τη μέθοδο για να στοχεύσουν συγκεκριμένες περιοχές μιτοχονδριακού DNA από 5 δείγματα Neanderthal. Τα αποτελέσματα ήταν η ανίχνευση 88 γνωρισμάτων τα οποία εξελίχθηκαν στους ανθρώπους από την εποχή του διαχωρισμού με τους Neanderthal δίνοντας σημαντικές πληροφορίες για τις διαφορές μεταξύ τους σε γενετικό επίπεδο. Αυτό σημαίνει ότι η τεχνική του εξονιόματος μπορεί να φανεί εξαιρετικά χρήσιμη για να έχουμε μια πιο ολοκληρωμένη άποψη της εξελικτικής μας ιστορίας (Teer et al., 2010).

Συμπερασματικά, η αλληλούχιση ανθρώπινου εξονιώματος μπορεί να προσφέρει διπλό ρόλο στη διάγνωση και στην ανακάλυψη γενετικών διαταραχών. Σε συνδυασμό με τη προβλεπόμενη μείωση του κόστους θα κερδίσει περισσότερο έδαφος σε καθημερινή χρήση από όλο και περισσότερους φορείς.

Disorder	Inheritance	Gene identified	Scope
Congenital chloride diarrhea	Recessive	*SLC26A3*	Exome
Miller syndrome	Recessive	*DHODH*	Exome
Charcot-Marie-Tooth neuropathy	Recessive	*SH3TC2*	Genome
Metachondromatosis	Dominant	*PTPN11*	Genome
Schinzel-Giedion syndrome	Dominant	*SETBP1*	Exome
Nonsyndromic hearing loss	Recessive	*GPSM2*	Exome
Perrault syndrome	Recessive	*HSD17B4*	Exome
Hyperphosphatasia mental retardation syndrome	Recessive	*PIGV*	Exome
Sensenbrenner syndrome	Recessive	*WDR35*	Exome
Cerebral cortical malformations	Recessive	*WDR62*	Exome
Kaposi sarcoma	Recessive	*STIM1*	Exome
Spinocerebellar ataxia	Dominant	*TGM6*	Exome
Combined hypolipidemia	Recessive	*ANGPTL3*	Exome
Complex I deficiency	Recessive	*ACAD9*	Exome
Autoimmune lymphoproliferative syndrome	Recessive	*FADD*	Exome
Amyotrophic lateral sclerosis	Dominant	*VCP*	Exome
Nonsyndromic mental retardation	Dominant	Various	Exome
Kabuki syndrome	Dominant	*MLL2*	Exome
Inflammatory bowel disease	Dominant	*XIAP*	Exome
Nonsyndromic mental retardation	Recessive	*TECR*	Exome
Retinitis pigmentosa	Recessive	*DHDDS*	Exome
Osteogenesis imperfecta	Recessive	*SERPINF1*	Exome
Dilated cardiomyopathy	Dominant	*BAG3*	Exome
Hajdu-Cheney syndrome	Dominant	*NOTCH2*	Exome
Hajdu-Cheney syndrome	Dominant	*NOTCH2*	Exome
Skeletal dysplasia	Recessive	*POP1*	Exome
Amelogenesis	Recessive	*FAM20A*	Exome
Chondrodysplasia and abnormal joint development	Recessive	*IMPAD1*	Exome
Progeroid syndrome	Recessive	*BANF1*	Exome
Infantile mitochondrial cardiomyopathy	Recessive	*AARS2*	Exome
Sensory neuropathy with dementia and hearing loss	Dominant	*DNMT1*	Exome
Autism	Dominant	Various	Exome

Εικ.20: Λίστα ανιχνεύσιμων Μενδελικών ασθενειών με whole genome και exome sequencing (Parla et al., 2011).

1.3.1 Στρατηγικές αλληλούχισης εξονιώματος (capture methods)

Τα τελευταία χρόνια η ανάπτυξη των τεχνολογιών μαζικής παράλληλης αλληλούχισης οδήγησε στην ανάπτυξη νέων μεθόδων σύλληψης (capture) των στοχευμένων – επιθυμητών περιοχών των γονιδιωμάτων. Οι νέες μέθοδοι προσφέρουν περισσότερα πλεονεκτήματα καθώς είναι πιο ευέλικτες αλλά και πιο φτηνές σε σχέση με τις παλαιότερες μεθόδους. Οι παλαιότερες μέθοδοι για μαζική παράλληλη αλληλούχιση χρησιμοποιούσαν την στόχευση συγκεκριμένων περιοχών με την τεχνική της PCR ακολουθούμενη από αλληλούχιση σε τριχοειδή. Αυτή η μέθοδος ήταν πιο χρονοβόρα και κόστιζε περισσότερο. Οι νέες μέθοδοι εξελίσσονται και βελτιώνονται ραγδαία ούτως ώστε η κάλυψη του 1% του γονιδιώματος που αποτελεί το εξονίωμα να είναι όσο το δυνατόν γίνεται πιο αποτελεσματική. Μπορεί να γίνει διάκριση των νέων μεθόδων, όπως αναφέρονται αναλυτικότερα παρακάτω, ανάμεσα σε αυτές που πραγματοποιούνται σε στερεή πλάκα (solid phase) και σε αυτές που δεν είναι (liquid phase) σε στερεή πλάκα (Teer et al., 2010).

Υβριδισμός στερεής φάσης (Solid-phase hybridization)

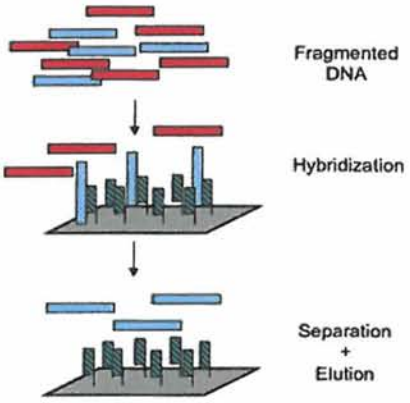
Στον υβριδισμό στερεής-φάσης (εικ.21) χρησιμοποιούνται ανιχνευτές συμπληρωματικοί με τις αλληλουχίες προς ανάλυση και η διαδικασία πραγματοποιείται πάνω σε μια στερεή πλάκα όπως είναι τα microarrays ή τα filters. Το δείγμα του DNA εφαρμόζεται στους ανιχνευτές όπου θα υβριδιστούν μόνο οι στοχευμένες περιοχές. Τα τμήματα που δεν υβριδίθηκαν απομακρύνονται με έκπλυση (Teer et al., 2010). Η πρώτη εταιρία που εφάρμοσε αυτή την τεχνολογία ήταν η Roche/Nimblegen και υιοθετήθηκε στις τεχνολογίες νέας γενιάς (Mamanova et al., 2010).

Υβριδισμός υγρής φάσης (Liquid-phase hybridization)

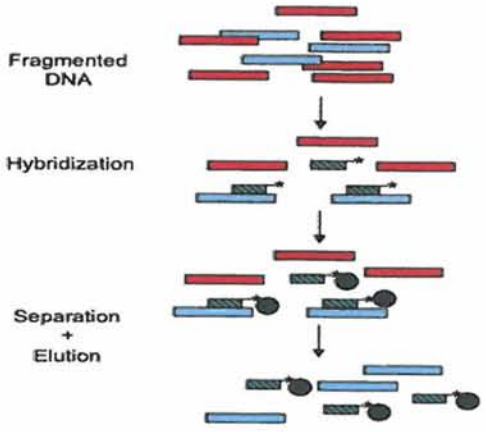
Η τεχνική υβριδισμού υγρής-φάσης (εικ.22) είναι παρόμοια με της στερεής φάσης με τη διαφορά ότι οι ανιχνευτές δεν είναι προσκολλημένοι σε στερεή πλάκα αλλά αντί για αυτό είναι βιοτινυλιωμένοι (Teer et al. , 2010). Οι βιοτινυλιωμένοι ανιχνευτές που είναι δεσμευμένοι με τις συμπληρωματικές αλληλουχίες του δείγματος

προσελκύονται μέσω μαγνητισμού από σφαιρίδια στρεπταβιδίνης. Με την έκλουση το αδέσμευτο γενετικό υλικό απομακρύνεται και το δεσμευμένο θα αλληλουχηθεί. (Mamanova et al., 2010). Την τεχνική αυτή χρησιμοποιεί η εταιρία Agilent με RNA βιοτινυλιωμένους ανιχνευτές (Teer et al., 2010).

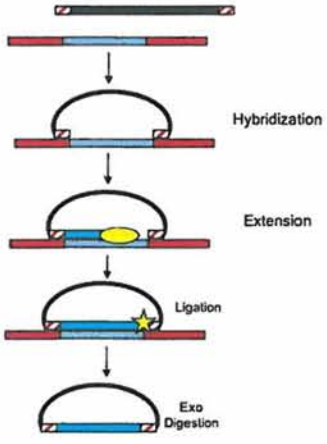
Άλλες δύο μέθοδοι εμπλουτισμού είναι επίσης η MIP-Molecular Inversion Probes (Ενζυμική μέθοδος)(εικ.23) και η PEC – Primer extension capture (εικ.24).



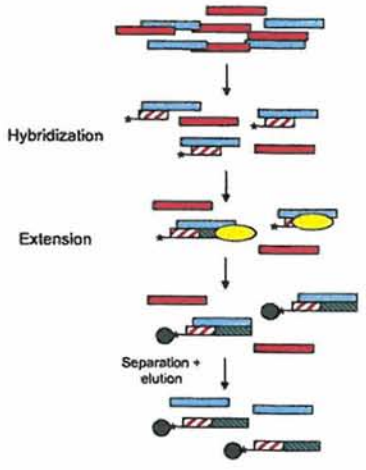
Εικ.21:Διαδικασία υβριδισμού στερεής – φάσης (Teer et al., 2010)



Εικ.22:Διαδικασία υβριδισμού υγρής – φάσης (Teer et al., 2010)

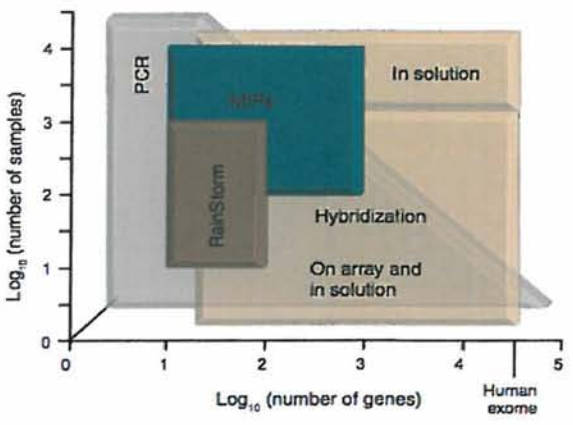


Εικ.23: MIP (Teer et al., 2010)



Εικ.24: PEC (Teer et al., 2010)

Οι παραπάνω μέθοδοι χρησιμοποιούνται για την δέσμευση των επιθυμητών περιοχών του γονιδιώματος το οποίο στη συνέχεια θα αλληλουχηθεί. Το παρακάτω διάγραμμα (εικ.25) απεικονίζει όλες αυτές τις μεθόδους σε σχέση με τον αριθμό των γονιδίων που μπορούν αυτές να δεσμεύσουν. Μπορεί εύκολα να διακριθεί από το διάγραμμα ότι οι νέες τεχνολογίες προσφέρουν και έχουν τη δυνατότητα μεγαλύτερης στόχευσης γονιδίων με αποτέλεσμα να θεωρηθούν καλύτερες και πιο αποδοτικές τεχνικές (Mamanova et al., 2010).



Εικ.25: Διάγραμμα σύγκρισης τεχνικών για την στόχευση περιοχών (Mamanova et al., 2010)

1.3.2 Μέθοδος αλληλούχισης εξονιώματος Illumina

Υπάρχουν δύο βασικές μέθοδοι για αλληλούχιση με την τεχνολογία της illumina, για την δημιουργία των βιβλιοθηκών του δείγματος. Η πρώτη είναι η Amplicon sequencing και η δεύτερη, η Target Enrchment. Η μέθοδος αλληλούχισης είναι και στις δύο παρόμοια, με τη διαφορά ότι η Amplicon sequencing επιτρέπει την αλληλούχιση, μικρών σε μέγεθος επιλεγμένων περιοχών, του γονιδιώματος, σε αντίθεση με την Target Enrchment όπου οι στοχευμένες περιοχές ή γονίδια μπορούν να είναι μεγαλύτερα σε μέγεθος ή ακόμη και να αποτελούν ολόκληρο το εξονίωμα του γονιδιώματος.

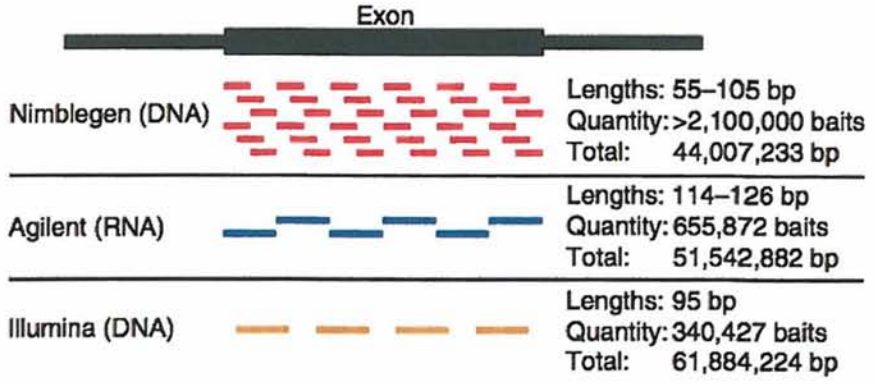
1.3.3 Σύγκριση πλατφόρμων αλληλούχισης εξονιώματος

Για την υλοποίηση της αλληλούχισης εξονιώματος υπάρχουν τρεις δημοφιλείς επιλογές εμπλουτισμού:

- Niblegen / Roche
- Agilent's SureSelect Human All Exon
- Illumina's TruSeq Exome Enrichment
  (Mamanova et al., 2010)

Οι τρεις πλατφόρμες εμπλουτισμού επιδεικνύουν ένα πολύ υψηλό επίπεδο αποτελεσματικότητας των στοχευμένων περιοχών και καλύπτουν ένα πολύ μεγάλο τμήμα του συνολικού εξονιώματος (Clark et al., 2011). Οι προαναφερθήσες τεχνολογίες διαφέρουν μεταξύ τους (εικ.26) ως προς :

i) τις στοχευμένες περιοχές
ii) το μήκος των ανιχνευτών
iii) την πυκνότητα των ανιχνευτών
iv) το είδος του προς ανάλυση μορίου (DNA για την Niblegen και Illumina, RNA για την Agilent)



Εικ.26: Σύγκριση των πλατφόρμων Niblegen, Agilent και Illumina (Clark et al., 2011)

Η Niblegen είναι η μόνη η οποία χρησιμοποιεί υψηλής πυκνότητας ανιχνευτές. Καλύπτει λιγότερες γενωμικές περιοχές από ότι άλλες πλατφόρμες αλλά απαιτεί λιγότερη ποσότητα δείγματος DNA. Η Illumina καλύπτει και μή – μεταφραζόμενες περιοχές, οι οποίες δεν μπορούν να στοχευθούν ούτε από την Niblegen ούτε από την Aginent (Mamanova et al., 2010).

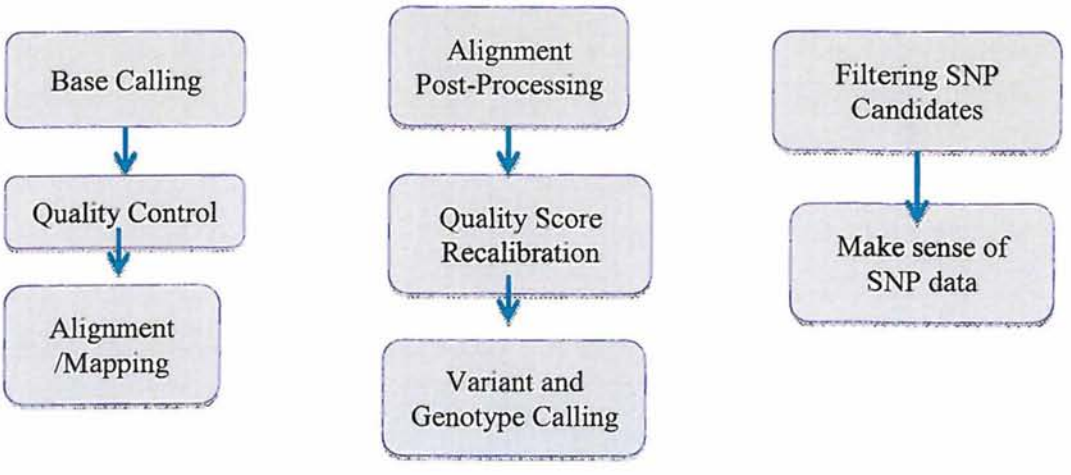
Οι παράγοντες που μπορούν να επηρεάσουν την επίδοση του exome sequencing είναι η ποιότητα του δείγματος, το μήκος των reads καθώς και η φύση του γονιδιώματος αναφοράς (Parla et al., 2011).

1.4 Ανάλυση δεδομένων αλληλούχισης εξονιώματος

Οι τεχνολογίες NGS είναι πλέον πολύ δημοφιλείς στην επιστημονική κοινότητα, λόγω της υψηλής ποιότητας και όγκου δεδομένων που προσφέρουν (Ruffalo et al., 2012). Ο τομέας της Βιοπληροφορικής έχει αναπτύξει και συνεχίζει να αναπτύσσει με ραγδαίο ρυθμό, τα λογισμικά εργαλεία και τις βάσεις δεδομένων για την καλύτερη διαχείριση και ανάλυση των δεδομένων από πλατφόρμες NGS (Lee et al., 2011).

Πολλοί υπολογιστικοί μέθοδοι είναι ήδη διαθέσιμοι για την ανάλυση των γενετικών παραλλαγών χρησιμοποιώντας δεδομένα από NGS. Αυτές οι παραλλαγές περιλαμβάνουν νουκλεοτιδικούς πολυμορφισμούς (SNPs) και δομικές παραλλαγές όπως αριθμούς αντιγράφων, ενθέσεις, διαγραφές, παράλληλοι διπλασιασμοί, αναστροφές και μετατοπίσεις. Η εύρεση και ο χαρακτηρισμός τέτοιων παραλλαγών είναι χρήσιμος σε πολλές εφαρμογές, συμπεριλαμβανομένων στις μελέτες συσχέτισης γονιδιωμάτων, στον προσδιορισμό μεταλλάξεων στον καρκίνο και στη συγκριτική γονιδιωματική (Ruffalo et al., 2012).

Γενικά, η διαδικασία ανάλυσης δεδομένων ανθρώπινου εξονιώματος από δεδομένα NGS , της πλατφόρμας Illumina, απεικονίζεται στο παρακάτω διάγραμμα ροής (εικ.27).



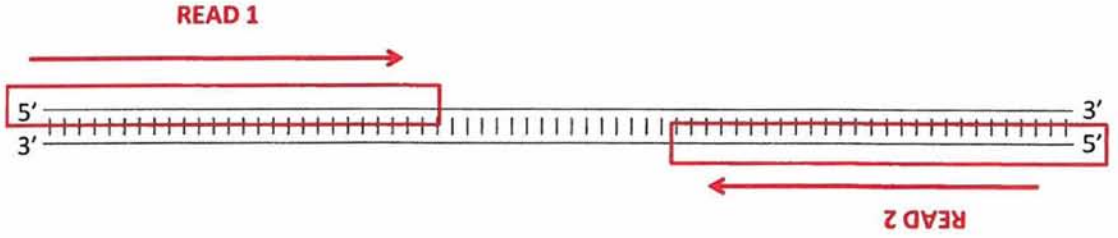
Εικ.27: Γενικό διάγραμμα ροής, ανάλυσης δεδομένων εξονιώματος

Επεξεργασία συμπιεσμένων αρχείων

Ένα από τα μεγαλύτερα προβλήματα που σχετίζονται με τα NGS, είναι η αποθήκευση και ο χειρισμός των δεδομένων. Τα δεδομένα αλληλούχισης καταλαμβάνουν τεράστιο όγκο. Για την επίλυση αυτού του προβλήματος, η εξαγωγή των αρχείων από NGS τεχνολογίες ,είναι σε συμπιεσμένη μορφή (gzip). Αυτό έχει ως αποτέλεσμα την μειωμένη απαίτηση χώρου αποθήκευσης και χρόνου για τη μεταφορά των δεδομένων (Patel et al., 2012). Η αποσυμπίεση των αρχείων γίνεται με τη χρήση εργαλείων ανάλυσης, με εύκολο και γρήγορο τρόπο.

Paired – end data

Ο όρος ¨ paired-end¨, αναφέρεται στην αλληλούχιση των άκρων του ίδιου μορίου DNA – read (εικ.28). Αρχικά, γίνεται αλληλούχιση του ενός άκρου (αλληλούχιση προς τα εμπρός) και στη συνέχεια γίνεται αλληλούχιση του άλλου άκρου (αλληλούχιση προς την αντίστροφη πλευρά). Το αποτέλεσμα είναι η δημιουργία δύο αρχείων,όπου ο ένας έχει τα αλληλουχημένα reads προς τα εμπρός και ο δεύτερος τα αλληλουχημένα reads προς την αντίστροφη πλευρά.



Εικ.28: Τρόπος αλληλούχισης των reads για την δημιουργία δεδομένων paired-end (http://www.cureffi.org/2012/12/19/forward-and-reverse-reads-in-paired-end-sequencing/)

Single – end data

Ο όρος αναφέρεται στην αλληλούχιση του read μόνο από την μία άκρη στην άλλη.

Το πρώτο βήμα της επεξεργασίας των δεδομένων, είναι η διαχείριση των αρχείων της αλληλούχισης. Η εξαγωγή των αρχείων αλληλούχισης από την εταιρεία illumina είναι σε μορφή FastQ.

Μορφή FastQ

Τα αρχεία Fastq έχουν χαρακτηριστική μορφή (εικ.29). Αποτελούνται από τέσσερις γραμμές για το κάθε read.

#	FASTQ Data
1	@SRR032209.2000 length=36
2	GTTGTGGCTGAGATGGGATGTAAACTTGANGANANN
3	+SRR032209.2000 length=36
4	B=A@?@BBB<285:<?8%3;#########!##!#!!

Εικ.29: Μορφή Fastq (Wan et al., 2011)

Η πρώτη γραμμή αρχίζει πάντοτε με το σύμβολο @ και προσδιορίζει το όνομα του read. 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. 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.

ASCII (American Standard Code for Information Interchange)

Το ASCII (εικ.30) είναι μία μορφή κρυπτογράφησης κειμένου με την μορφή χαρακτήρων της αγγλικής αλφαβήτου.

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Εικ.30: Σύστημα κρυπτογράφησης ASCII (http://en.wikipedia.org/wiki/ASCII)

Βαθμός ποιότητας Q (Quality)

Ο βαθμός - τιμή ποιότητας Q μιας βάσης, είναι η πιθανότητα η βάση αυτή να διαβάσθηκε λανθασμένα (error probability P). Για την εύρεση της πιθανότητας δίνονται δύο μαθηματικές εξισώσεις.

Α) Πρότυπο Sanger (Standard Sanger) ή αλλιώς βαθμολογία ποιότητας Phred (quality score).

$$Q_{sanger} = -10x\log_{10} P$$

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 για να εκτιμηθεί εάν οι διαφορές μεταξύ των δύο επικαλυπτόμενων ακολουθιών είναι πιο πιθανό να προκύψουν από τυχαία σφάλματα ή από διάφορα αντίγραφα μιας επαναλαμβανόμενης αλληλουχίας. Σύμφωνα με την εικ.31 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 μικρότερη πιθανότητα λάθους.

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	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%

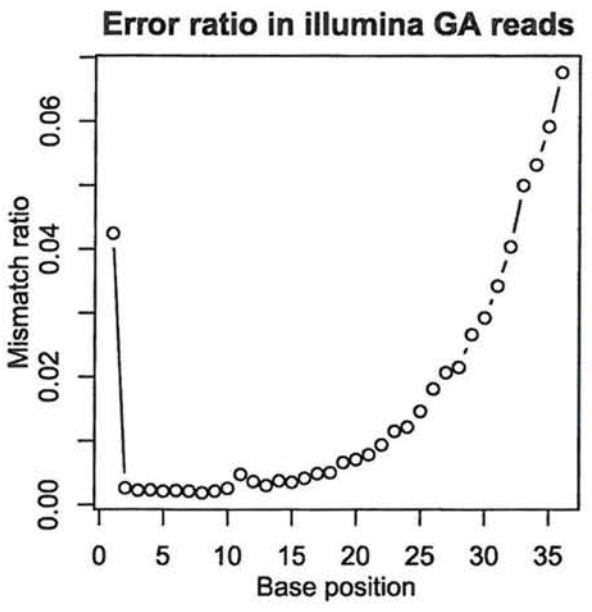
Εικ.31: Phred quality scores

Β) Solexa quality score:

$$Q_{solexa} = -10x\log_{10}\left(\frac{P}{1-P}\right)$$

Το 2004, η εταιρία Solexa εισήγαγε τη δική της έκδοση υπολογισμού ποιότητας της μορφής FastQ. Παρά το γεγονός ότι η μορφή FastQ καταγράφει μόνο ένα αποτέλεσμα ποιότητας, με την εξίσωση της Solexa παράγονται επίσης και άλλα αρχεία όσον αφορά την ποιότητα και των τεσσάρων βάσεων, προκειμένου να μειώσουνε τα χαμηλής ποιότητας αποτελέσματα.

Η αξιολόγιση της ποιότητας του κάθε reads είναι πολύ σημαντική διεργασία καθώς υπάρχει ενδεχόμενο ανακρίβειας της ή των βάσεων εξαιτίας συστηματικού λάθους (εικ.32) που μπορεί να έχει η τεχνολογία της αλληλούχισης που χρησιμοποιήθηκε ή η ποιότητα της ίδιας της αλληλουχίας.

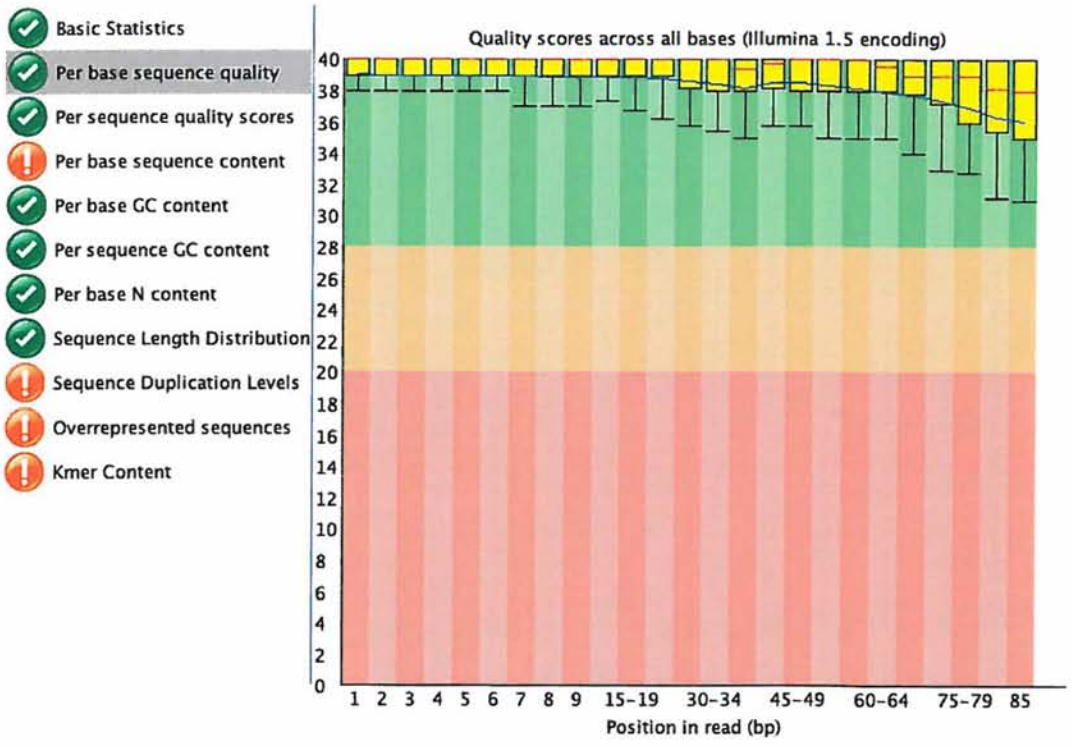


Εικ.32: Συσσώρευση λαθών κατά την ενσωμάτωση φθοριζόντων dNTPs.
(http://openi.nlm.nih.gov/detailedresult.php?img=3096631_pone.0019534.g001&req=4)

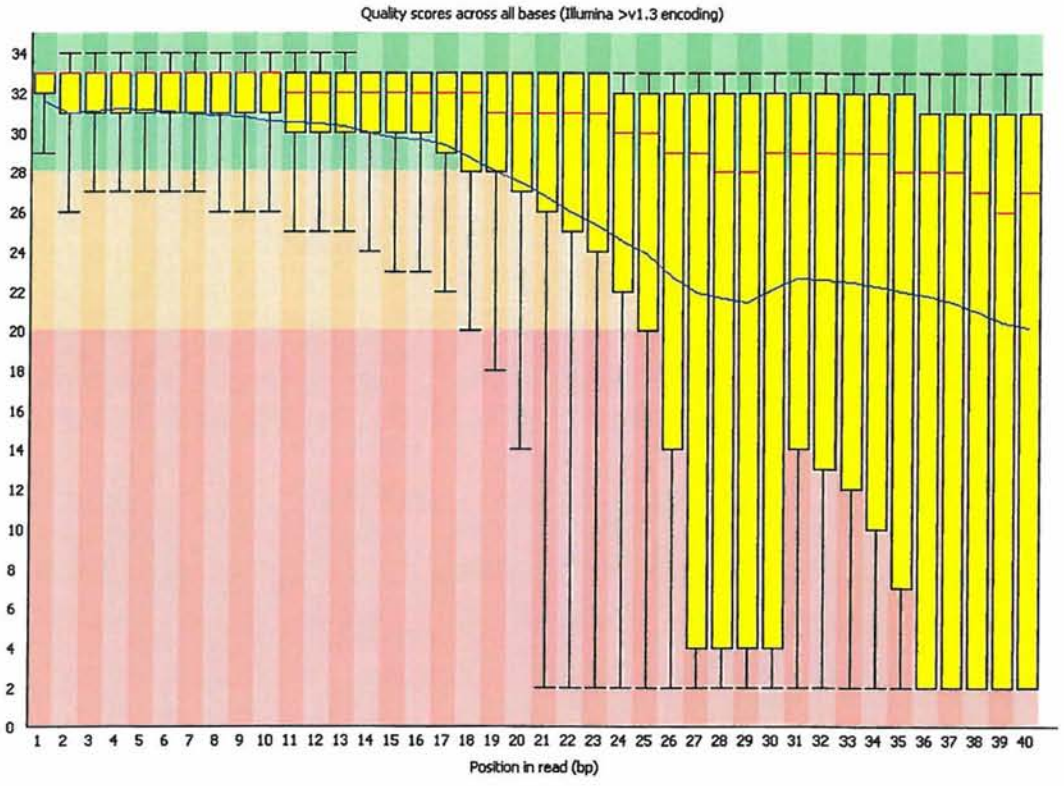
Κάποια από τα προγράμματα ελέγχου ποιότητας των αλληλουχημένων reads είναι: Reaper, FastQC, Bcbio-nextgen., Chipster, GeneProf και Biopieces. Θα πρέπει να σημειωθεί ότι πολλά προγράμματα για την λειτουργία τους στηρίζονται στην γνώση και χρήση δυναμικού προγραμματισμού όπως της Perl, Biopython, Java ή C++. Πριν την επιλογή του προγράμματος θα πρέπει να εξακριβωθούν δύο παράμετροι. Ο πρώτος είναι, να υποστηρίζει το πρόγραμμα τα δεδομένα της συγκεκριμένης τεχνολογίας (π.χ illumina) και ο δεύτερος είναι να δέχεται αρχεία FastQ. Σε περίπτωση που το πρόγραμμα δεν δεχεται τη μορφή FastQ αλλά κάποια άλλη,

υπάρχει η επιλογή μετατροπής του στην επιθυμητή μορφή με την χρήση όμως κάποιων προγραμμάτων που έχουν αυτό το σκοπό.

Παράδειγμα υψηλής (εικ.33) και χαμηλής ποιότητας δεδομένων (εικ.34) με την χρήση του προγράμματος FastQC. Το πρόγραμμα FastQ παρέχει πληροφορίες για την ποότητα του κάθε read,της κάθε βάσης του read, το ποσοστό GC του read, το ποσοστό των αδιάβαστων βάσεων (N), το ποσοστό των διπλασιασμένων reads και την κατανομή μήκους της αλληλουχίας.



Εικ.33: Διάγραμμα απεικόνισης υψηλού βαθμού ποιότητας των βάσεων των reads



Εικ.34: Διάγραμμα απεικόνισης χαμηλού βαθμού ποιότητας των βάσεων των reads

Το δεύτερο βήμα περιλαμβάνει το φιλτράρισμα (trimming) των reads. Όπως αναφέρθηκε και παραπάνω, ο ποιοτικός έλεγχος της κάθε βάσης του read αλλά και επί του συνόλου του κάθε read έχει πραγματοποιηθεί και με βάση αυτό μπορεί να γίνει το φιλτράρισμα. Κατά προτίμηση κόβουμε τα reads που έχουν ποιότητα κάτω από Q20 διότι η πιθανότητα να είναι λάθος θα είναι περισσότερο από 0.01.

Με το φιλτράρισμα μπορούν να επιτευχθούν τρεις στόχοι αποκοπής ανεπιθύμητων περιοχών όπως : i) των adapters των sequence reads ii) όλων των sequence reads από την θέση που εμφανίζεται χαμηλής ποιότητας αλληλούχισης η οποία όμως συνεχίζεται (π.χ εικ.33) και iii) συγκεκριμένων προβληματικών θέσεων για το κάθε sequence read χωριστά (τα κομμένα κομμάτια των sequence reads που έχουν πολύ μικρό μήκος, θα απορριφθούν). Προτεινόμενα προγράμματα για trimming των sequence reads είναι τα ακόλουθα: Condetri, NGS Toolkit, Chipster, Reaper, SeqTrim και SolexaQA.

Σε αυτό το σημείο, προαιρετικά, μπορεί να γίνει αφαίρεση των διπλασιασμένων reads για την καλύτερη στοίχιση των reads στην αλληλουχία αναφοράς. Το Condetri εκτός από το trimming παρέχει και αυτή την δυνατότητα.

Το τρίτο βήμα είναι η στοίχιση (alignment) των reads στην αλληλουχία αναφοράς.Έχουν αναπτυχθεί αρκετά προγράμματα στηριζόμενα στους αλγόριθμους για τη πραγματοποίηση της σωστής ευθυγράμμισης των reads στις χρωμοσωμικές περιοχές. Κάποια από αυτά είναι τα εξής: Burrows Wheeler Aligner (BWA), Bowtie 2, SOAP2 και MAQ.

Θα πρέπει να σημειωθεί ότι η αλληλουχία αναφοράς αντλείται από τις βάσεις δεδομένων UCSC ή Ensembl. Η καινούργια έκδοση της UCSC στην ανθρώπινη αλληλουχία αναφοράς είναι η hg19. Για την καλύτερη εφαρμογή των προγραμμάτων για την ευθυγράμμιση το αρχείο της αλληλουχίας αναφοράς του ανθρώπινου γονιδιώματος θα πρέπει να είναι σε μορφή fasta.

Η μορφή fasta αρχίζει με το σύμβολο " >" στην πρώτη γραμμή και στις επόμενες γραμμές ακολουθεί η αλληλουχία. Πολλές αλληλουχίες σε μορφή fasta μπορούν να συγχωνευτούν σε ενα αρχείο, διαχωρίζοντας η μία από την άλλη από το σύμβολο " >". Η ένωση των αρχείων μπορεί να γίνει στο "τερματικό" με την εντολή cat.

Ένα προτεινόμενο και εύχρηστο πρόγραμμα ευθυγράμμισης στοίχισης με την ανθρώπινη αλληλουχία αναφοράς είναι το BWA (Burrows Wheeler Aligner). Για την λειτουργία του απαιτείται μικρή μνήμη και διαχειρίζεται επιτυχώς δεδομένα από Ilumina.

Για να ξεκινήσει η διαδικασία της ευθυγράμμισης, δημιουργούμε ένα ευρετήριο (index) για την αλληλουχία αναφοράς, υπό μορφή fasta στο BWA. Στην συνέχεια δίνονται εντολές για την στοίχιση των δύο αρχείων, καθώς είναι paired-end, με την αλληλουχία αναφοράς. Το BWA εξάγει τα δεδομένα της ευθυγράμμισης των reads σε μορφή sai. Καθώς όμως πολλά προγράμματα χρησιμοποιούν την μορφή sam και για την συνέχεια της ανάλυσης των δεδομένων (για λόγους συμβατότητας), γίνεται μετατροπή των αρχείων sai σε sam με την βοήθεια εργαλείων – εντολών που περιλαμβάνονται στον aligner BWA.

Μορφή αρχείων SAM

Το αρχείο SAM διακρίνεται: α) στην ενότητα της επικεφαλίδας (header section) και β) στην ενότητα της ευθυγράμμισης (alignment section). Δίνει πληροφορίες σχετικά με:

- την στοίχιση του κάθε read
- τη θέση του read στο contig αναφοράς
- τον προσανατολισμό του read
- την ποιότητα στοίχισης
- δυνατότητα επανευθυγράμμισης των reads

α) Ενότητα Επικεφαλίδας (header section)

Αρχίζει πάντοτε με το σύμβολο " @" και ακολουθούν δύο γράμματα με κωδικό χαρακτήρα. Δίνονται πληροφορίες για την περιοχή του χρωμοσώματος του read στην αλληλουχία αναφοράς. Στην επικεφαλίδα (εικ.35), υπάρχουν γραμμές που αποτελούνται και από πεδία δεδομένων που αναφέρονται ως «TAG: VALUE» (ετικέτες). Τα TAGs είναι προαιρετικά πεδία στο αρχείο SAM. Μια μορφή TAG είναι μια συμβολοσειρά και αποτελείται από δύο χαρακτήρες: το είδος και την τιμή, τα οποία καθορίζουν το περιεχόμενο και τη μορφή της αξίας για την αποθήκευση των πληροφοριών σχετικά με το read και το alignment. Το σύμβολο " * " σημαίνει ότι ο τύπος εγγραφής είναι παρόν. Επίσης οι ετικέτες που περιέχουν πεζά γράμματα προορίζονται για τους τελικούς χρήστες. Κάποιες από τις ετικέτες είναι οι εξής :
@HD: Είναι η ετικέτα επικεφαλίδας ($1^{η}$ γραμμή).
@SQ: Καθορίζει τη σειρά ταξινόμησης της στοίχισης στην αλληλουχία αναφοράς.
@RG: Η ομάδα των reads. Η κάθε ομάδα θα πρέπει να αποτελείται από μοναδικό όνομα- κωδικό (ID) καθώς αργότερα θα τροποποιηθούν για την συγχώνευση των αρχείων SAM και θα πρέπει να μπορούν να διακριθούν μεταξύ τους. Η τιμή του ID χρησιμοποιείται για την καταγραφή της ευθυγράμμισης.
@PG: Καταγραφή του προγράμματος που χρησιμοποιήθηκε.
@CO:Γραμμή σχολιασμού του αρχείου.

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Εικ.35: Παράδειγμα ενότητας επικεφαλίδας του αρχείου SAM (http://samtools.sourceforge.net/SAM1.pdf)

β) Ενότητα ευθυγράμμισης (alignment section).

Η ενότητα ευθυγράμμισης περιέχει πληροφορίες για το κάθε read σχετικά με το πού και το πώς έγινε η στοίχιση στην αλληλουχία αναφοράς. Η κάθε γραμμή στοίχισης, του κάθε read, αποτελείται από 11 υποχρεωτικά πεδία παρέχοντας βασικές πληροφορίες, όπως για παράδειγμα η θέσης της ευθυγράμμισης στην χαρτογράφηση.

Αναλυτικότερα, τα πεδία αυτά (εικ.36) είναι τα παρακάτω και αναφέρουν:

QNAME (query name): Αναφέρει την ομαδοποίηση των alignments (π.χ ζευγαρωμένα alignment ή ένα read το οποίο εμφανίζεται σε πολλαπλά alignment).
FLAG: Περιέχει πληροφορίες περιγράφοντας το alignment (π.χ. τον αριθμό των θραυσμάτων, τον αριθμό των σωστά ευθυγραμμισμένων θραυσμάτων, το πρώτο θραύσμα, το τελευταίο θραύσμα καθώς και πιο read έχει μικρό quality control).
RNAME: Το όνομα της αλληλουχίας αναφοράς.Συνήθως περιέχει το όνομα του χρωμοσώματος.
POS: Αναφορά της θέσης όπου ξεκίνησε η στοίχιση.
MAPQ: Αναφορά της ποιότητας της χαρτογράφισης με βάση την κλίμακα – βαθμονόμηση Phred.
CIGAR: Δίνει πληροφορίες για το ποιες βάσεις ευθυγραμμίστικαν.ποιες βάσεις δεν ταίριαξαν με την αλληλουχία αναφοράς, ποιες διαγράφησαν, τις επιπλέον βάσεις που μπορεί να υπάρχουν καθώς και τις βάσεις που δεν υπάρχουν για ταίριασμα.
SEQ: Η αλληλουχία του εξεταζόμενου read για αυτό το δείγμα.
QUAL: Είναι ένας δείκτης ποιότητας της εξεταζόμενης αλληλουχίας με βάση το σύστημα ASCII.

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{29}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next segment
8	PNEXT	Int	[0,2^{29}-1]	Position of the mate/next segment
9	TLEN	Int	[-2^{29}+1,2^{29}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Εικ.36: Πληροφορίες της ενότητας ευθυγράμμισης του αρχείου SAM.
(http://samtools.sourceforge.net/SAM1.pdf)

Μετατροπή SAM/BAM αρχείων

Τις περισσότερες φορές επειδή τα αρχεία σε μορφή SAM καταλαμβάνουν πολύ χώρο (ένα αρχείο SAM μπορεί να χρειάζεται και 20-30GB), για το λόγο αυτό αλλά και για την καλύτερη ανάγνωση των αρχείων γίνεται μετατροπή αυτών από την μορφή SAM σε BAM. Η μορφή BAM περιέχει τις ίδιες ακριβώς πληροφορίες αλλα σε ένα συμπιεσμένο, αναπροσαρμοσμένο σε δυαδική μορφή, αρχείο. Υπάρχουν διάφορα προγράμματα μετατροπής αρχείων από SAM σε BAM. Το πιο εύχρηστο είναι το Picard, το οποίο έχει την ικανότητα διαχείρισης και χειραγώγισης αρχείων SAM. Η χρήση του Picard στηρίζεται σε προγραμματισμό Java.

Το επόμενο βήμα είναι η αφαίρεση των μη – μοναδικών reads στην στοίχιση. Τα reads που μπορούν να στοιχιθούν σε περισσότερες από μία θέσεις και εξαιτίας του ότι δεν μπορεί να αποφασιστεί σε ποια ακριβώς θέση ουσιαστικά θα πρέπει να βρίσκονται, αφαιρούνται. Για την αφαίρεση χρησιμοποιούνται ειδικά προγράμματα όπως είναι το GATK. Με την ολοκλήρωση αυτών των βημάτων μπορεί να ξεκινήσει η διαδικασία αναγνώρισης των SNPs (SNP-calling).

Το αρχικό βήμα πριν το SNP-calling, είναι ο επανα-υπολογισμός του βαθμού ποιότητας των reads. Η ακριβής ποιότητα του βαθμού των reads είναι απαραίτητη για την χρήση των σύγχρονων αλγορίθμων που ως στόχο έχουν την αναγνώριση των SNPs, καθώς ενσωματώνουν το βαθμό ποιότητας των βάσεων (σύμφωνα με την κλίμακα Phed).

Προγράμματα τα οποία εκτελούν επανέλεγχο της ποιότητας των reads είναι το GATK και το SOAPsnp. Το πιο ευρέως στη χρήση του επανέλεγχου των reads και προτεινόμενο, είναι το GATK.

Στη συνέχεια, αρχίζει η διαδικασία εύρεσης των πολυμορφισμών και του γονοτύπου με την βοήθεια των προγραμμάτων GATK και των εργαλείων SAM. Η ταυτοποίηση των πολυμορφισμών γίνεται στηριζόμενη σε βάσεις δεδομένων που περιέχουν γνωστούς πολυμορφισμούς όπως η dbSNP.Κατόπιν, γίνεται φιλτράρισμα των πολυμορφισμών που ανιχνεύτηκαν με στόχο την μείωση των ψευδώς-θετικών πολυμορφισμών. Εργαλεία που μπορούν να χρησιμοποιηθούν για το φιλτράρισμα των SNPs είναι το GATK, εργαλεία SAM και εργαλεία VCF.

Σε αυτό το σημείο, θα πρέπει να αναφερθεί ότι αρκετά εργαλεία για την εύρεση των πολυμορφισμών των SNPs δημιουργούν αρχεία σε μορφή vcf. Τα αρχεία σε μορφή vcf περιέχουν πληροφορίες για το είδος του πολυμορφισμού, σε ποιο χρωμόσωμα βρίσκεται και τη θέση που κατέχει στο χρωμόσωμα.

Το τελευταίο στάδιο της ανάλυσης, των δεδομένων του ανθρώπινου εξονιώματος, είναι ο σχολιασμός των SNPs. Έχουν αναπτυχθεί αρκετά προγράμματα που ως στόχο έχουν τον αυτοματοποιημένο σχολιασμό γενετικών παραλλαγών. Ένα από αυτά είναι το Annovar. Το Annovar αντλεί πληροφορίες από την βάση δεδομένων UCSC. Εισάγοντας τη λίστα με τις παραλλαγές που ανιχνεύθηκαν, τις σχολιάζει με βάση την περιοχή στην οποία βρίσκονται, εξάγοντας ένα αρχείο σε μορφή excel.

Ένα αρκετά υποσχόμενο εργαλείο σχολιασμών των πολυμορφισμών, είναι το Promethease. Το Promethease, για το σχολιασμό των πολυμορφισμών στηρίζεται στην SNPedia και έχει τη δυνατότητα διαχείρισης δεδομένων από την 23andMe.Για την λειτουργία του χρειάζεται η εισαγωγή του αρχείου των γονοτύπων και η επιλογή του γονιδιώματος αναφοράς. Τα αποτελέσματα εξάγονται σε περίπου τέσσερις ώρες.

Οι πληροφορίες που δίνονται αφορούν:

- μοναδικούς πολυμορφισμούς οι οποίοι δεν έχουν επιβεβαιωθεί ακόμη
- πολυμορφισμοί που εμπλέκονται σε παθολογικές καταστάσεις

- το είδος του μεταβολισμού σε ουσίες και η δοσολογία η οποία θα πρέπει να χορηγείται στο συγκεκριμένο άτομο (π.χ ενδιάμεσοι μεταβολίτες στο αντιπηκτικό βαρφαρίνη, όπου θα πρέπει να χορηγείται συγκεκριμένη δόση).

Για την επεξεργασία των πολυμορφισμών, έχουν αναπτυχθεί προγράμματα οπτικής ανάλυσης που βοηθάνε στην ανίχνευση παραλλαγών σε σύγκριση με την αλληλουχία αναφοράς, καθώς και στη σύγκριση μεταξύ διαφορετικών δειγμάτων. Τέτοια προγράμματα είναι το IGV, το GenomeView και το Tablet.

Το πρόγραμμα οπτικής ανάλυσης Tablet έχει δημιουργηθεί για την επεξεργασία δεδομένων NGS. Γίνεται απεικονίση των contigs και προσφέρει α) την δυνατότητα αναφοράς περιοχών που δεν ταιριάζουν με την αλληλουχία αναφοράς καθώς και β) μία περίληψη των σχολιασμών των πολυμορφισμών σε σχέση με τις πληροφορίες του contig.

2. Υλικά και μέθοδοι

Η ανάλυση και η διαχείριση των δεδομένων, για την πραγματοποίηση της παρούσας εργασίας, έγινε με την χρήση προγραμματισμού, βάσεων δεδομένων και ειδικών προγραμμάτων βιοπληροφορικής. Οι απαιτήσεις του υπολογιστικού συστήματος για την διακπεραίωση της εργασίας ήταν:

- Λογισμικό Linux – Ubuntu 12.04
- Intel® Xeon, CPU E5620, Quad Core 2.40 GHz (16 threads)
- Μνήμη 96GB
- Σκληρός δίσκος 3TB

2.1 Χρήση προγραμματισμού

Η επεξεργασία των δεδομένων και η εκτέλεση των προγραμμάτων έγινε με την Perl. Η Perl είναι μία γλώσσα προγραμματισμού η οποία είναι διαθέσιμη για όλα τα λειτουργικά συστήματα. Χαρακτηριστικά της είναι η μετατροπή αρχείων, η διαχείριση αριθμών, πινάκων (απλοί ή συσχετιστικοί), συμβολοσειρών και σύνδεσης δεδομένων (προσαρμοστικότητα). Το λογισμικό που χρησιμοποιήσαμε στο εργαστήριο ήταν Linux (Ubuntu-12.04) το οποίο είχε ενσωματωμένη την Perl.

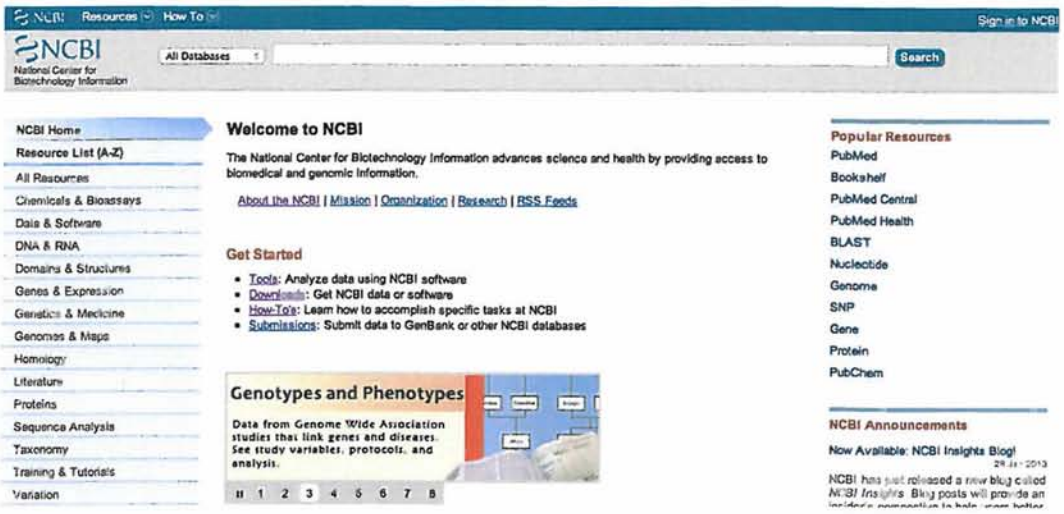
2.2 Βάσεις Δεδομένων

Για την ανάλυση δεδομένων από exome sequencing χρειάστηκε η συμβολή από αρκετές βάσεις δεδομένων, που αναφέρονται παρακάτω:

The National Center for Biotechnology Information (NCBI)

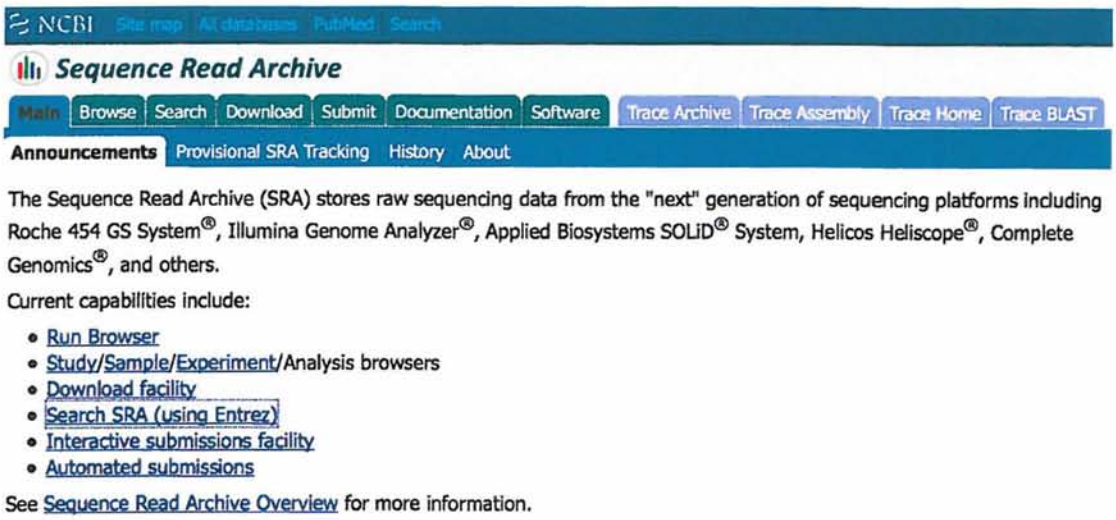
Η NCBI (εικ.37) στεγάζει μια σειρά από βάσεις δεδομένων που σχετίζονται με τη βιοτεχνολογία και τη βιοϊατρική. Σημαντικές βάσεις δεδομένων είναι η GenBank που περιλαμβάνει πληροφορίες για αλληλουχίες DNA και η PubMed, μια βιβλιογραφική βάση δεδομένων για τη βιοϊατρική βιβλιογραφία. Όλες αυτές οι βάσεις δεδομένων είναι διαθέσιμες στο διαδίκτυο μέσω της μηχανής αναζήτησης Entrez (http://en.wikipedia.org/wiki/National_Center_for_Biotechnology_Information).

Από το NCBI κατεβάσαμε τον φάκελο των SNPs σε μορφή VCF για την επαναβαθμολόγιση της ποιότητας των βάσεων μετά την στοίχιση. Το αρχείο σε μορφή VCF περιείχε τα ονόματα των χρωμοσωμάτων με αριθμούς.



Εικ.37: Ιστοσελίδα εθνικού κέντρου πληροφοριών βιοτεχνολογίας-NCNI (http://www.ncbi.nlm.nih.gov)

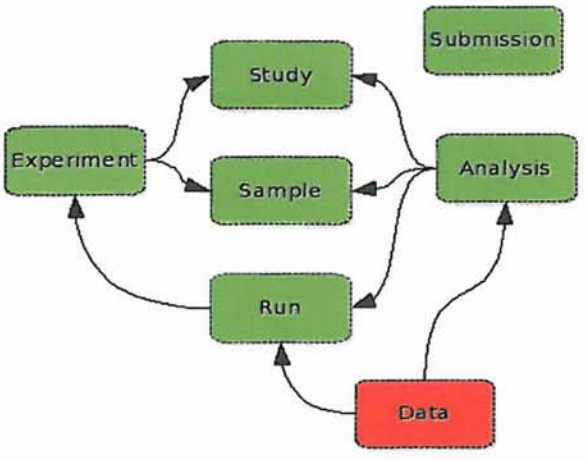
Sequence Read Archive (SRA)



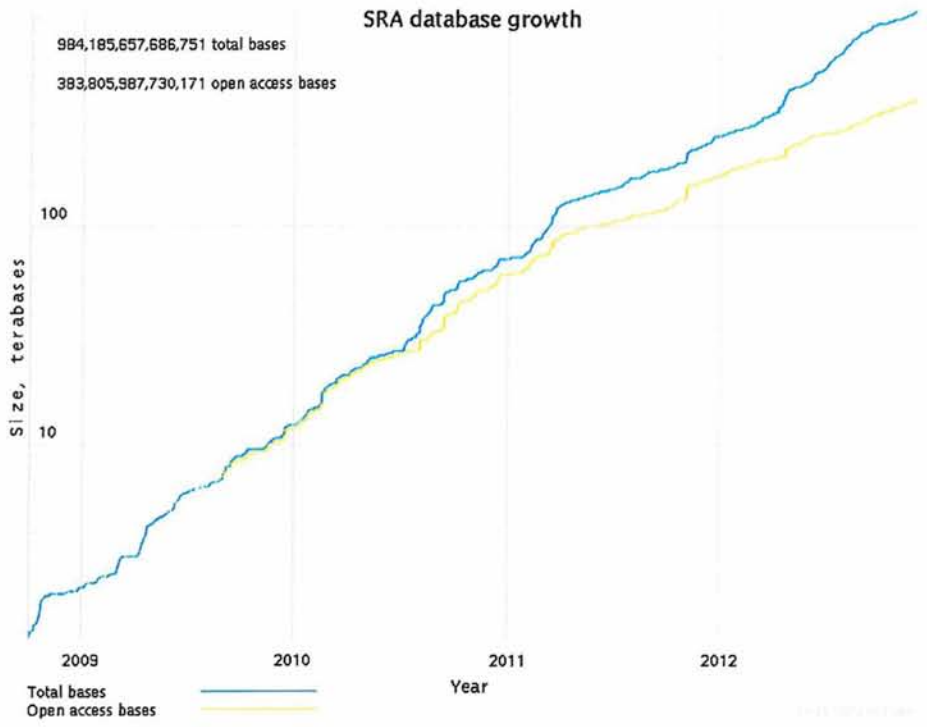
Εικ.38:Βάση δεδομένων SRA (http://www.galter.northwestern.edu/news/index.cfm/2009/9/15/NCBI-Short-Read-Archive-SRA-of-NextGeneration-Sequencing-Data)

Από την SRA αρχικά κατεβάσαμε τα εργαλεία της (SRA-TOOLKIT) και ύστερα αντλήσαμε από εκεί τα δεδομένα της αλληλούχισης του αθρώπινου εξονιώματος σε μορφή fastq από Illumina (αρχείο SRR330441).

Η SRA (εικ.38) περιέχει δεδομένα τα οποία προέρχονται από αλληλούχιση με τεχνολογίες νέας γενιάς (sort reads), όπως: Illumina (Immumina Inc.), Roche/454 (Roche Diagnostics Corp.), SOLiD (Life Technologies Corp.), HeliScope Single Molecule Sequencer (Helicos Biosciences Corp.), Complete Genomics Inc., SMRT (Pasific Biosciences Inc.), Ion Torrent Sysrems Inc (Leinonen et al., 2010). Σκοπός της είναι να παρέχει στους ερευνητές δωρεάν πρόσβαση στα δεδομένα (Kodama et al., 2012). Η αποθήκευση των δεδομένων στην SRA (εικ.39) γίνεται με τέτοιο τρόπο ουτως ώστε να είναι διακριτή η προέλευση των δεδομένων αναφέροντας πληροφορίες σχετικά με τη μελέτη,την προέλευση του εξεταζόμενου δείγματος, το πείραμα, την πλατφόρμα η οποία χρησιμοποιήθηκε, την ανάλυση που έγινε καθώς και τη χρονική περίοδο που υποβλήθηκαν τα δεδομένα (Kodama et al., 2012). Τα τελευταία χρόνια παρατηρείται ανάπτυξη της SRA (εικ.39) καθώς όλο και περισσότερα δεδομένα αποθηκεύονται.



Εικ.39: Δομή δεδομένων στην SRA
(http://www.ebi.ac.uk/ena/about/sra_submissions)



Εικ.40:Ανάπτυξη της βάσης δεδομένων SRA.
(http://www.ncbi.nlm.nih.gov/Traces/sra/?view=announcement)

UCSC Genome Browser

Η μηχανή αναζήτησης γονιδιωμάτων UCSC αναπτύσσεται και συντηρείται από μια διατμηματική ομάδα γονιδιωματικής βιοπληροφορικής του Πανεπιστήμιου της Καλιφόρνιας Σάντα Κρουζ (UCSC). Η ιστοσελίδα περιέχει την ανθρώπινη αλληλουχία αναφοράς, καθώς και προσχέδιες αλληλουχίες άλλων οργανισμών. Επίσης, συνδέεται με την βάση δεδομένων ENCODE (http://genome.ucsc.edu).

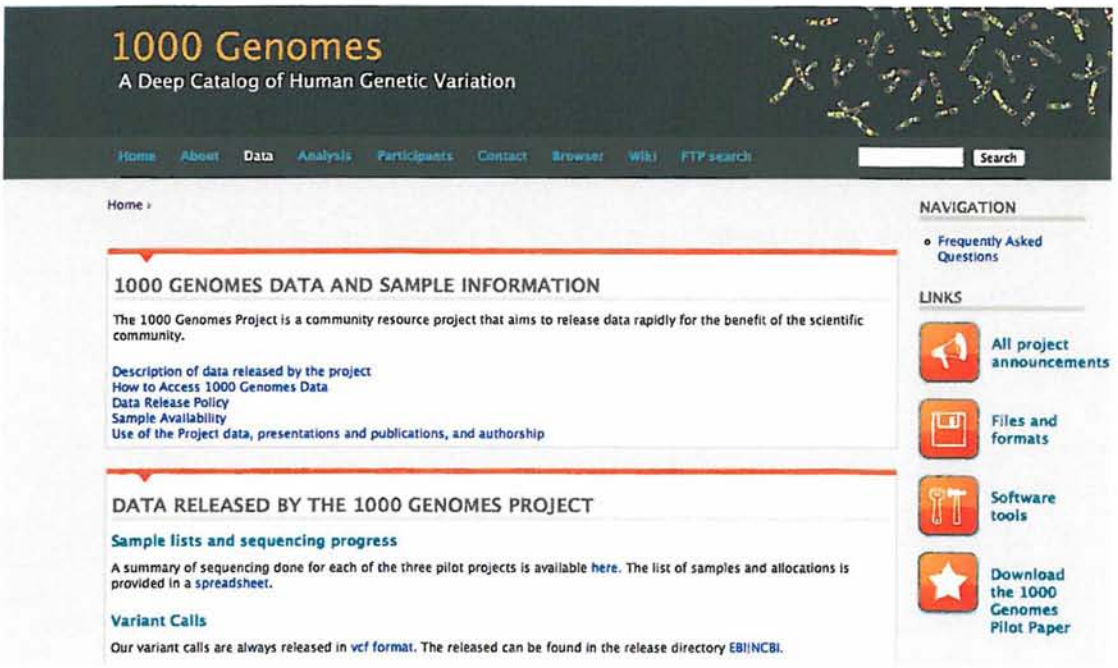
Από την UCSC κατεβάσαμε την ανθρώπινη αλληλουχία αναφοράς hg19 (GRCh37) για να γίνει η στοίχιση με τα δεδομένα μας (alignment). Το αρχείο ήταν σε συμπιεσμένη μορφή και η αλληλουχία του κάθε χρωμοσώματος βρισκόταν σε ξεχωριστό αρχείο. Με την χρήση της Perl συγχωνεύσαμε τα αρχεία των χρωμοσωμάτων σε ένα αρχείο.

1000 Genomes

Το πρόγραμμα των 1000 Ανθρώπινων Γονιδίων αποσκοπεί στην δημιουργία συλλογής και παροχής πληροφοριών για την κατανόηση της γενετικής συμβολής στις ασθένειες σε σχέση με την γεωγραφική προέλευση των ατόμων καθώς και τις λειτουργικές γενετικές παραλλαγές που αυτές φέρουν (εικ.41).

Ο σκοπός του προγράμματος ήταν η ανίχνευση πολυμορφισμών (SNPs και δομικοί πολυμορφισμοί). Τα δείγματα για την πραγματοποίηση του προγράμματος αντλήθηκαν από 14 πληθυσμούς από την Ευρώπη, την Αμερική,την Ανατολική Ασία και την Αφρική. Η ανάλυση των δειγμάτων έγινε τόσο με χαμηλής κάλυψης αλληλούχιση ολόκληρων των γονιδιωμάτων όσο και αλληλούχιση των εξονιωμάτων. Για την διαχείριση των αναλύσεων χρησιμοποιήθηκαν αρκετοί αλγόριθμοι και πολλές πηγές δεδομένων.

Τα αποτελέσματα του προγράμματος ήταν η ανίχνευση: 38 εκατομμυρίων SNPs και 1.4 εκατομμύρια μικρών προσθηκών και απαλοιφών.

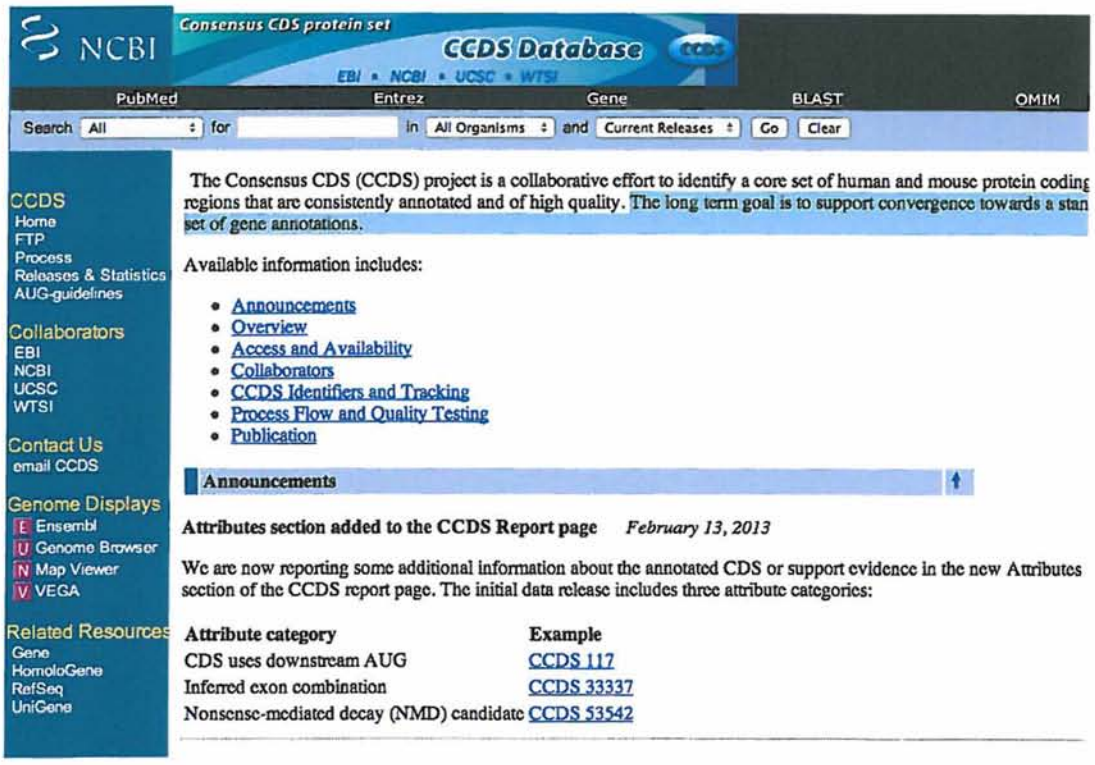


Εικ.41:Ιστοσελίδα παροχής-άντλησης δεδομένων από το 1000 Genomes (http://www.1000genomes.org/data)

Από τη βάση δεδομένων 1000 Genomes αντλήθηκαν τα SNPs του ανθρώπινου γονιδιώματος για την εύρεση των πολυμορφισμών των δεδομένων μας, μέσω του προγράμματος annovar όπως θα δούμε αναλυτικότερα παρακάτω.

CCDS (Consensus coding sequence database)

Η CCDS είναι μια βάση δεδομένων που περιέχει πληροφορίες για τις κωδικές περιοχές του ανθρώπινου γονιδιώματος και του γονιδιώματος του ποντικιού.Η βάση δεδομένων CCDS (εικ.42) συνεργάζεται με τις βάσεις δεδομένων EBI, NCBI, UCSC και WTSI. Ο μακροπρόθεσμος στόχος του CCDS είναι η κατηγοριοποίηση των σχολιασμών του συνόλου των γενετικών παραλλαγών που έχουν βρεθεί ως τώρα (http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi). Δεδομένα από την CCDC αντλήθηκαν μέσω του Annovar.



Εικ.42: Βάση δεδομένων CCDS (http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi)

SEQanswers

Η ιστοσελίδα SEQanswers είναι μια πηγή πληροφοριών και forum συζήτησης για θέματα που αφορούν το NGS. Η κοινότητα αποτελείται από πολλά μέλη παρέχοντας βοήθεια και απαντώντας στις ερωτήσεις των χρηστών της κοινότητας. Σκοπός του ιδρυτή, είναι η κοινότητα να κατέχει μια κεντρική θέση στην εκπαίδευση της νέας γενιάς τεχνολογιών αλληλούχισης γονιδιωμάτων (http://seqanswers.com). Από το SEQanswers όπως και από την Biostar αντλήσαμε πληροφορίες για τον τρόπο ανάλυσης των NGS.

Biostar

Είναι ένας δικτυακός τόπος που επικεντρώνεται στη βιοπληροφορική, στην υπολογιστική γονιδιωματική και στην ανάλυση βιολογικών δεδομένων. Η ιστοσελίδα απαρτίζεται από διάφορες θεματικές ενότητες συζητήσεων πάνω σε αυτά τα επιστημονικά αντικείμενα, βίντεο και προτεινόμενα εργαλεία βιοπληροφορικής. Επίσης, τα μέλη έχουν την δυνατότητα λήψης απαντήσεων και επίλυσης προβλημάτων, με την βοήθεια άλλων χρηστών (http://www.biostars.org).

SNPedia

Είναι μία ιστοσελίδα η οποία περιέχει κατάλογο των πολυμορφισμών και δίνει πληροφορίες σχετικά με την επίδραση αυτών στο φαινότυπο. Περιέχει 36.379 SNPs που μπορεί να εμπλέκονται στον μεταβολισμό φαρμάκων, στις χημικές ουσίες, γενικές πληροφορίες των πολυμορφισμών καθώς και άλλα (http://snpedia.com/index.php/SNPedia). Επίσης, χρησιμοποιείται από το λογισμικό Promethease, το οποίο χρησιμοποιήσαμε στο εργαστήριο για την επεξήγηση των SNPs.

dbSNP

Η βάση δεδομένων dbSNP ανήκει στην NCBI και περιέχει πληροφορίες για τους πολυμορφισμούς SNPs, τοποθετώντας το όνομα (ID) του πολυμορφισμού στην αναζήτηση (http://www.ncbi.nlm.nih.gov/SNP)

2.3 Προγράμματα

Τα προγράμματα τα οποία χρησιμοποιήθηκαν στο εργαστήριο για την ανάλυση και την επεξεργασία των δεδομένων ήταν το fastQC, το Condetri, το BWA, τα εργαλεία Picard, τα εργαλεία SAM, το GATK, το Tablet, το Annovar και το Promethease. Αναλυτικότερα:

FastQC

Το πρόγραμμα FastQC χρησιμοποιήθηκε για τον ποιοτικό έλεγχο τόσο των βάσεων όσο και του συνόλου των reads. Επίσης, με το FastQC εξήγαμε πληροφορίες για το μέγεθος των reads, το ποσοστό των GC βάσεων καθώς και τον αριθμό των διπλασιασμένων reads.

Condetri

Με το Condetri έγινε "κόψιμο" των reads αλλα και των βάσεων που είχαν βαθμό ποιότητας (quality score) χαμηλότερο του 25. Επίσης, πραγματοποιήθηκε αφαίρεση των διπλασιασμένων reads.

BWA

Το BWA προκαλεί στοίχιση (alignment) της αλληλουχίας αναφοράς hg19 με τα δεδομένα των reads μας. Αρχικά, έγινε εισαγωγή (index) της αλληλουχίας αναφοράς στο BWA και στην συνέχεια έγινε η στοίχιση.

Εργαλεία Picard

Τα εργαλεία Picard βοήθησαν στην χειραγώγηση των αρχείων SAM που εξήχθησαν από το BWA, δηλαδή στον καθαρισμό των reads, στο φιλτράρισμα καθώς και στη μετατροπή των αρχείων από SAM σε BAM.

GATK

Το πρόγραμμα GATK χρησιμοποιεί μια σειρά εργαλείων που ως στόχο έχουν την εύρεση των προβληματικών reads (RealignerTargetCreator) και την επαναστοίχιση αυτών (IndelRealigner). Επίσης, με το GATK έγινε επαναβαθμολόγιση της ποιότητας των reads, ανίχνευση και φιλτράρισμα των SNPs.

Tablet

Το πρόγραμμα Tablet προσφέρει απεικόνιση της στοίχισης των reads ως προς το γονιδίωμα αναφοράς.

Annovar

Ο σχολιασμός των SNPs έγινε με το πρόγραμμα Annovar .Για την εκτέλεσή του κατεβάσαμε από τις βάσεις δεδομένων (π.χ 1000 Genome, dbsnp,esp κ.α) τους πολυμορφισμούς. Στην συνέχεια έγινε ο σχολιασμός, με αποτέλεσμα την αναφορά του ονόματος (ID) του πολυμορφισμού, την θέση του στο χρωμόσωμα και την ομοζυγωτία ή ετεροζυγωτία του αλληλομόρφου.

Promethease

Η ολοκλήρωση της ανάλυσής έγινε με το πρόγραμμα Promethease. Το Promethease έδωσε πληροφορίες για τον κάθε πολυμορφισμό σε σχέση με το πως επηρεάζει αυτός τον φαινότυπο, π.χ πιθανότητα εμφάνισης κάποιας νόσου, μοναδικότητα χαρακτηριστικών καθώς και την απόκριση του μεταβολισμού σε χημικές ουσίες .

3. Αποτελέσματα

Η ανάλυση και διαχείριση των δεδομένων που πραγματοποιήθηκε στο εργαστήριο, αφορούσε δεδομένα αλληλούχισης ανθρώπινου εξονιώματος από τεχνολογία Illumina. Τα βήματα και τα αποτελέσματα της ανάλυσης περιγράφονται παρακάτω.

Βήμα 1^{0} : Άντληση αρχείων σε μορφή fasta και fastq από τη βάση δεδομένων SRA (NCBI) και UCSC

Αρχικά, για να μπορέσουμε να εξάγουμε δεδομένα με τη μορφή fastq από την βάση δεδομένων SRA, θα πρέπει να εγκαταστηθεί στον υπολογιστή μας το πακέτο εργαλείων της SRA (SRA-TOOLKIT), από την NCBI. Για να κατεβάσουμε ένα συγκεκριμένο αρχείο για ένα συγκεκριμένο δείγμα, γίνεται χρήση του προγράμματος περιήγησης.

Πριν την εκτέλεση των εργαλείων SRA (SRA-TOOLKIT) πρέπει να τρέξει για πρώτη φορά το διαμορφωμένο -perl script assistant.perl.

Μόλις ανακτηθεί το αρχείο SRA, χρησιμοποιούμε την παρακάτω εντολή:
configuration-assistant.perl SRR330441.sra

Για την εξαγωγή των αρχείων fastq (αν είναι paired-end), τρέχουμε την παρακάτω εντολή, από το SRA-TOOLKIT.
fastq-dump-split-3 SRR330441.sra

Ανάκτηση της ανθρώπινης αλληλουχίας αναφοράς, hg19 (GRCh37), από τη βάση δεδομένων UCSC

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/
Το αρχείο της ανθρώπινης αλληλουχίας αναφοράς είναι σε συμπιεσμένη μορφή. Για την χρήση των δεδομένων θα πρέπει πρώτα να αποσυμπιεστεί. Τα δεδομένα για το κάθε χρωμόσωμα ξεχωριστά βρίσκεται σε διαφορετικό αρχείο. Για την χρήση των δεδομένων θα πρέπει να γίνει συγχώνευση των αρχείων, που περιέχουν τις

πληροφορίες των χρωμοσωμάτων, σε ένα αρχείο. Αυτό πραγματοποιείται με την εντολή cat, δηλαδή:

cat chr1.fa chr2.fa chr3.fa chr4.fa chr5.fa chr6.fa chr7.fa chr8.fa chr9.fa chr10.fa chr11.fa chr12.fa chr13.fa chr14.fa chr15.fa chr16.fa chr17.fa chr18.fa chr19.fa chr20.fa chr21.fa chr22.fa chrX.fa chrY.fa chrM.fa > hg19.fa

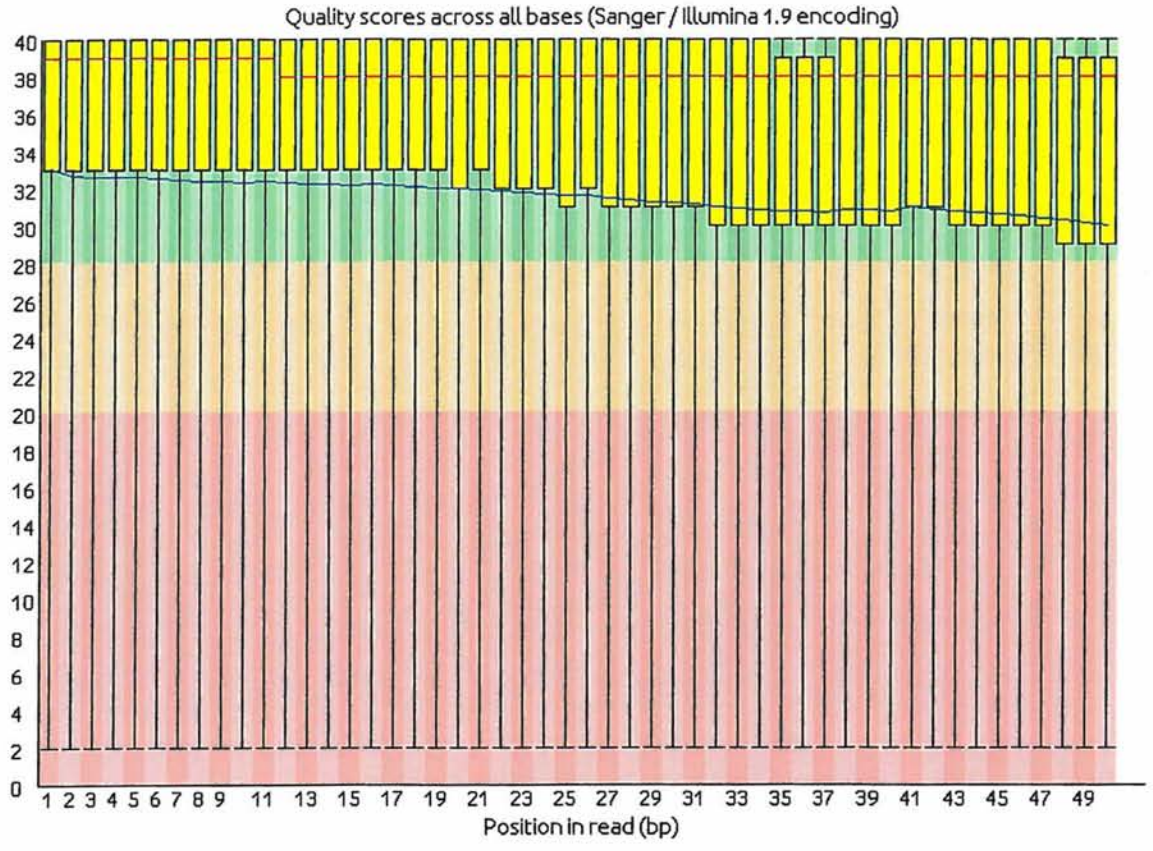
Βήμα 2^{0} : Ποιοτικός έλεγχος και φιλτράρισμα (trimming, PCR duplicate removal) των reads, με τα προγράμματα fastQC & Condetri

Πραγματοποιήθηκε ποιοτικός έλεγχος των reads, με το προγραμμα fastQC, των αρχειων SRR330441_1.fastq και SRR330441_2.fastq (υπάρχουν δύο αρχεία για το ίδιο δείγμα, καθώς τα δεδομένα είναι paired-end). Τα αποτελέσματα που πήραμε ήταν τα εξής:

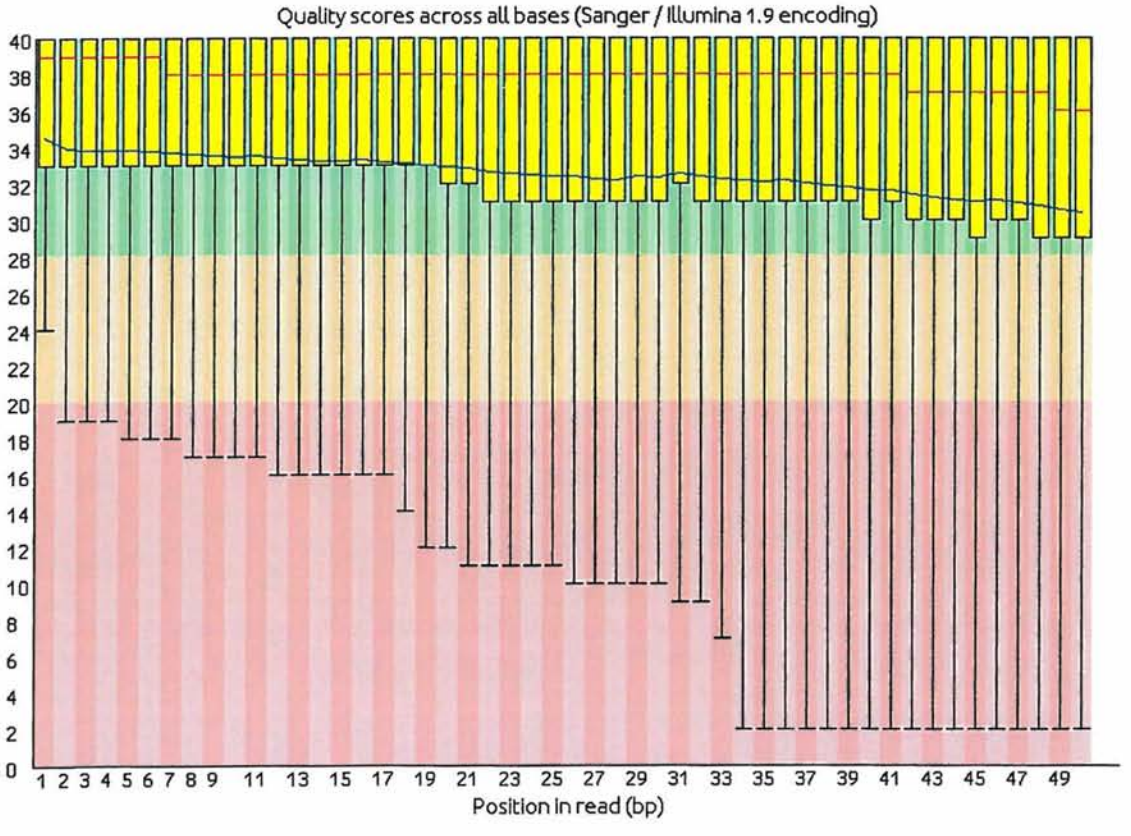
Για το αρχείο SRR330441_1.fastq:
Ο συνολικός αριθμός των reads ήταν 57.090.906.Το μήκος του κάθε reads ήταν 50bp (εικ.45). Το % GC βάσεων στο σύνολο τών reads ήταν 51. Η κωδικοποίηση που χρησιμοποιήθηκε κατά την διαδικασία της αλληλούχισης ήταν Sanger/Illumina 1.9.

Για το αρχείο SRR330441_2.fastq:
Ο συνολικός αριθμός των reads ήταν 57.090.906. Το μήκος του κάθε read ήταν 50bp. Το συνολικό ποσοστό των GC βάσεων των reads ήταν 52 και η κωδικοποίηση ήταν η ίδια με το αρχείο SRR330441_1.fastq.

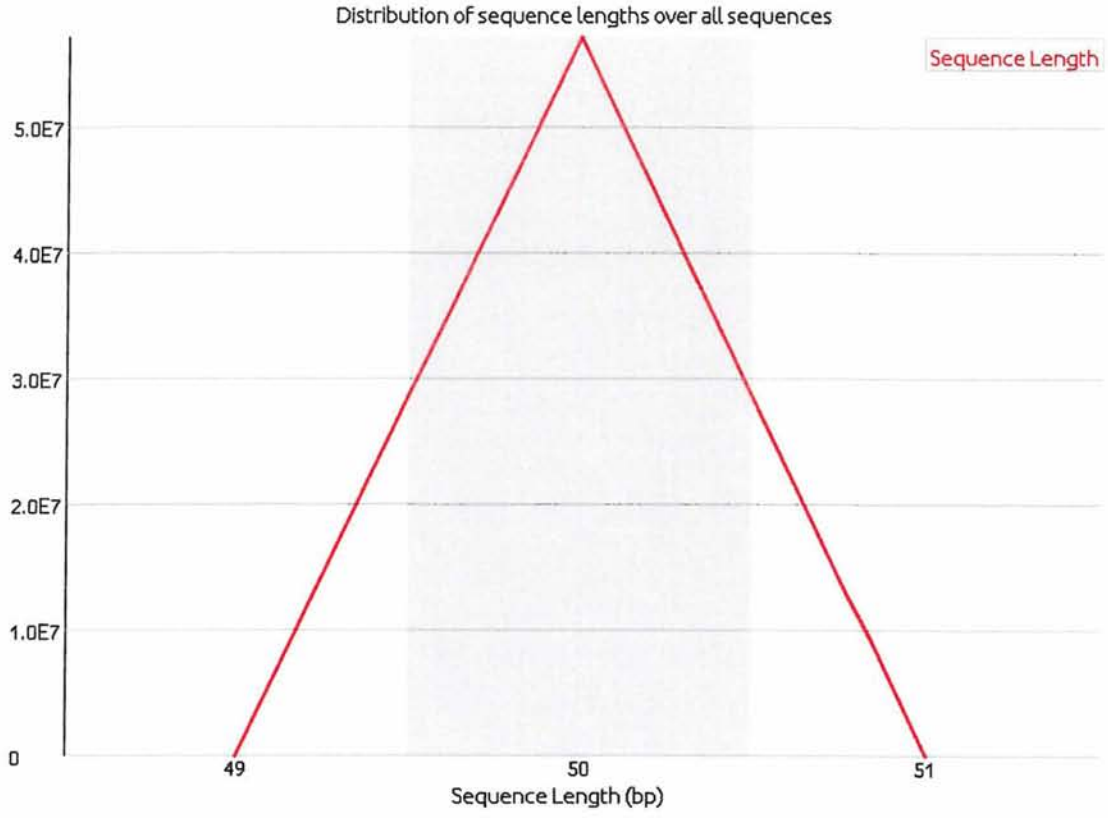
Παρατηρούμε ότι το quality score των βάσεων των reads και για τα δύο αρχεία (SRR330441_1.fastq και SRR330441_2.fastq) είναι αρκετά ικανοποιητικό (εικ.43 και εικ.44) δεν πέφτει κάτω από Q=20 και η πλειοψηφία αυτών έχουν μικρό ποσοστό λάθους, που σημαίνει ότι η διαδικασία της αλληλούχισης έγινε επιτυχώς ή ότι φιλτραρίστηκε πριν κατατεθεί στο SRA.



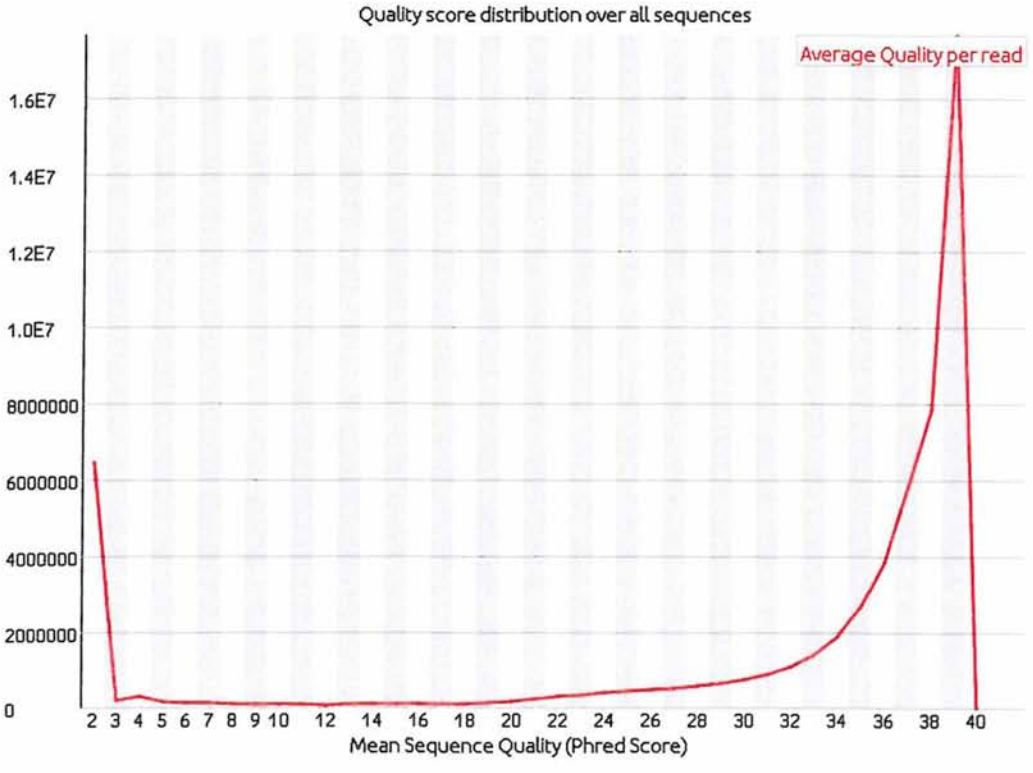
Εκ.43:Διάγραμμα απεικόνισης της ποιότητας όλων των βάσεων των reads για το αρχείο SRR330441_1.fastq



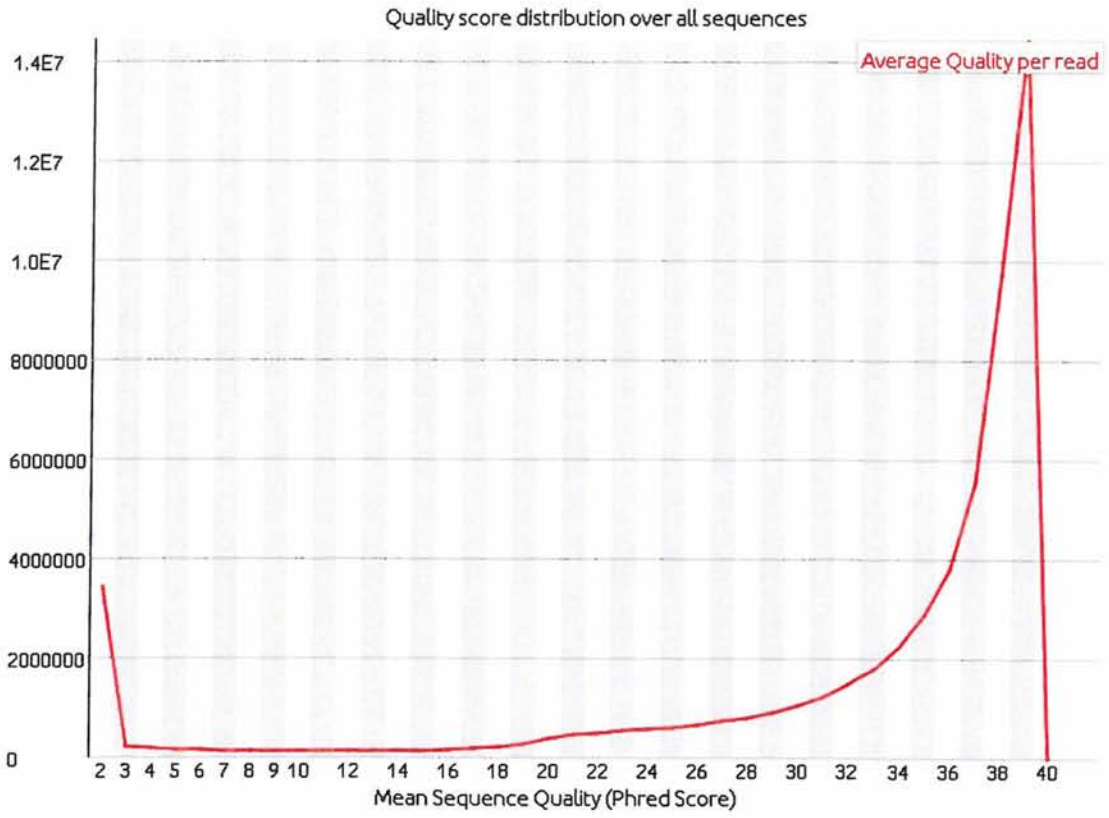
Εκ.44: Διάγραμμα απεικόνισης της ποιότητας όλων των βάσεων των reads για το αρχείο SRR330441_2.fastq



Εικ.45: Διάγραμμα απεικόνισης του μήκους των reads (ίδιο και για τα δύο αρχεία)



Εικ.46: Απεικόνιση μέσου όρου της ποιότητας του κάθε read (αρχείο SRR330441_1.fastq).



Εικ.47: Απεικόνιση μέσου όρου της ποιότητας του κάθε read (αρχείο SRR330441_2.fastq).

Παρατηρούμε ότι η ποιότητα των βάσεων προς το τέλος τως reads μειώνεται. Για το λόγο αυτό, όπως θα δούμε και παρακάτω, έγινε αφαίρεση αυτών των βάσεων ούτως ώστε η στοίχιση με την αλληλουχία αναφοράς (alignment) σε επόμενο βήμα, να είναι πιο ακριβής και να μειωθεί το ποσοστό λάθους.

Στην συνέχεια, έγινε αφαίρεση των περιοχών (trimming) των reads, που δεν εκπληρώνουν τον βαθμό της επιθυμητής ποιότητας με το πρόγραμμα Condetri. Στο "τερματικό" πληκτρολογούμε την εντολή:

./condetri_v2.2.pl -fastq1=SRR330441_1.fastq -fastq2=SRR330441_2.fastq -prefix=SRR330441 -hq=25 -lq=13 -frac=0.8 -minlen=35 -mh=5 -ml=1 -sc=33 –rmN

Τα αρχεία εξάγονται με όνομα SRR330441_trim1.fastq και SRR330441_trim2.fastq. Για την αφαίρεση των διπλασιασμένων reads (με το πρόγραμμα FilterPCRdupl, που είναι μαζί με το Condetri), πληκτρολογούμε την εντολή:

./filterPCRdupl.pl -fastq1=SRR330441_trim1.fastq -fastq2=SRR330441_trim2.fastq -prefix=SRR330441 –cmp=31

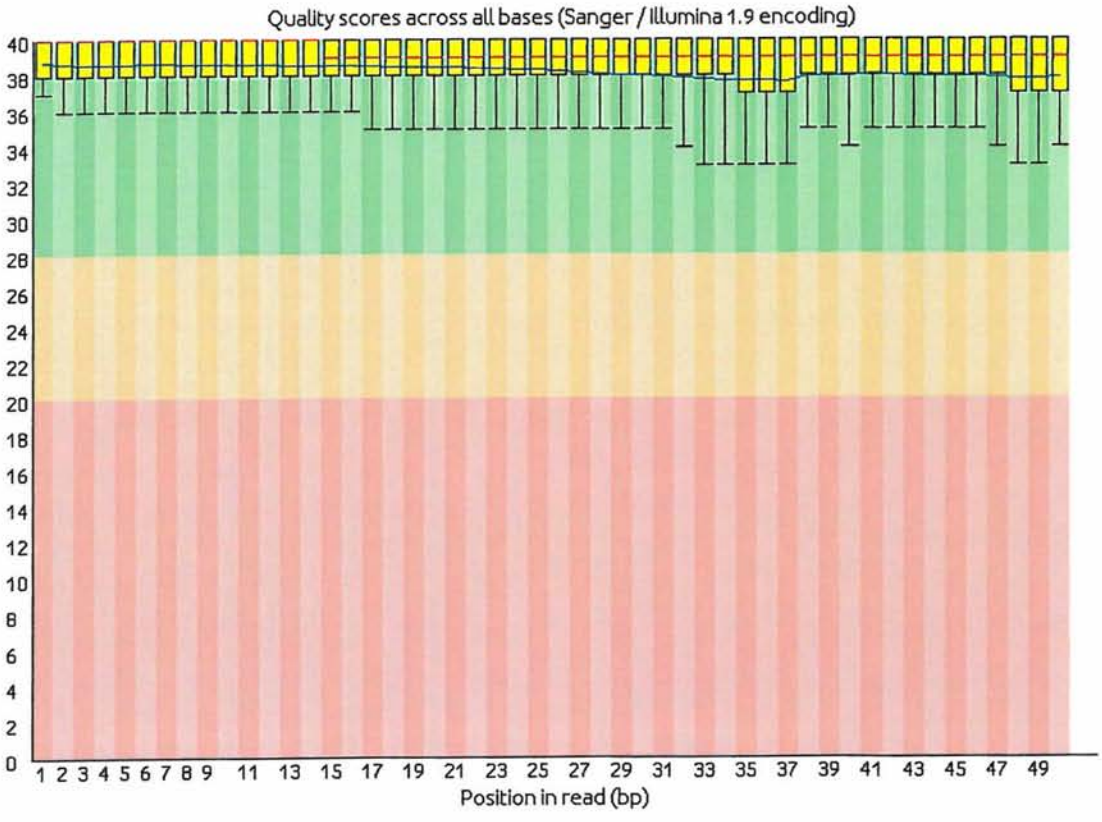
Το όνομα των αρχείων πλέον είναι: SRR330441_uniq1.fastq και SRR330441_uniq2.fastq.

Κατόπιν, έγινε πάλι ποιοτικός έλεγχος των δεδομένων των αρχείων SRR330441_uniq1.fastq (εικ.48) και SRR330441_uniq2.fastq (εικ.49) με την χρήση του προγράμματος fastQC.

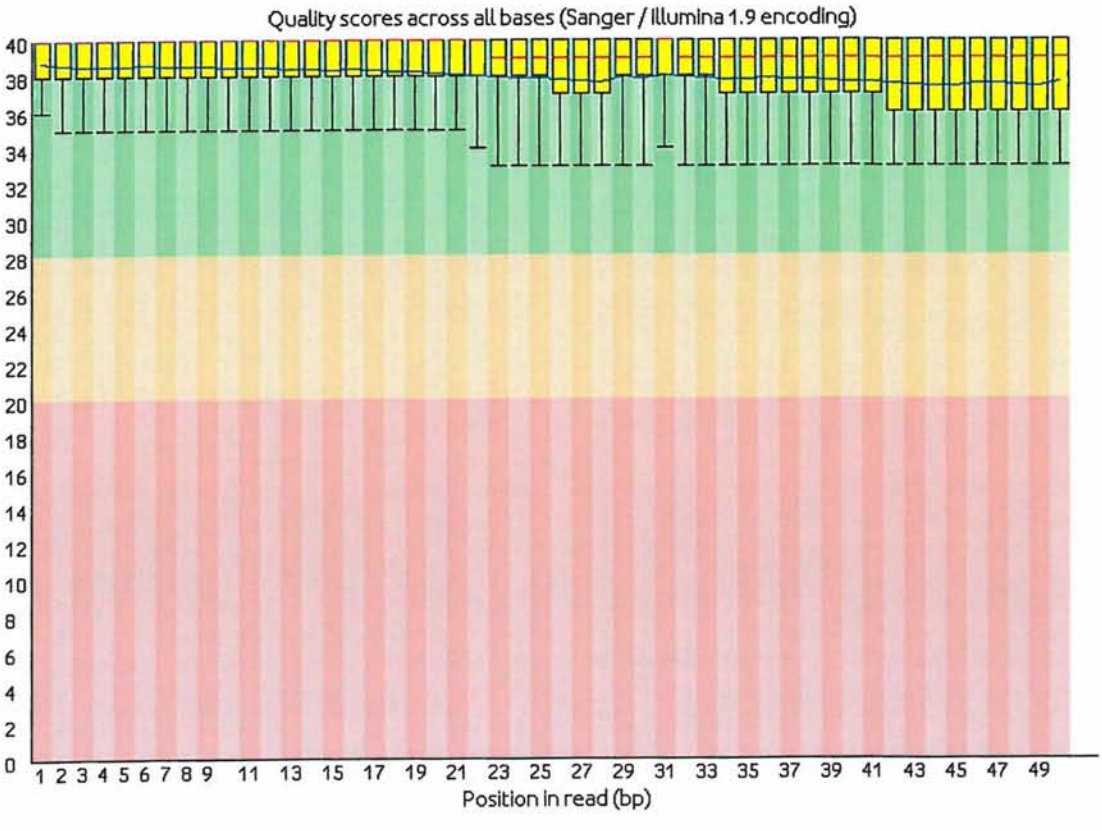
Τα αποτελέσματα που πήραμε (πιν.2) μετά την αφαίρεση των διπλασιασμένων reads ήταν τα εξής: α) για τα δεδομένα του αρχείου SRR330441_uniq1.fastq, οι συνολικές αλληλουχίες ήταν 34.870.773, το μήκος των reads ήταν μεταξύ 35-50 και το % GC βάσεων ήταν 49 για το σύνολο των reads και β) για τα δεδομένα του αρχείου SRR330441_uniq.fastq οι συνολικές αλληλουχίες ήταν 34.870.773, το μήκος των reads ήταν μεταξύ 35-50 και το %GC βάσεων του συνόλου των reads ήταν 49.

Αναλυτικός πίνακας αποτελεσμάτων του βαθμού ποιότητας των reads (πιν.2)

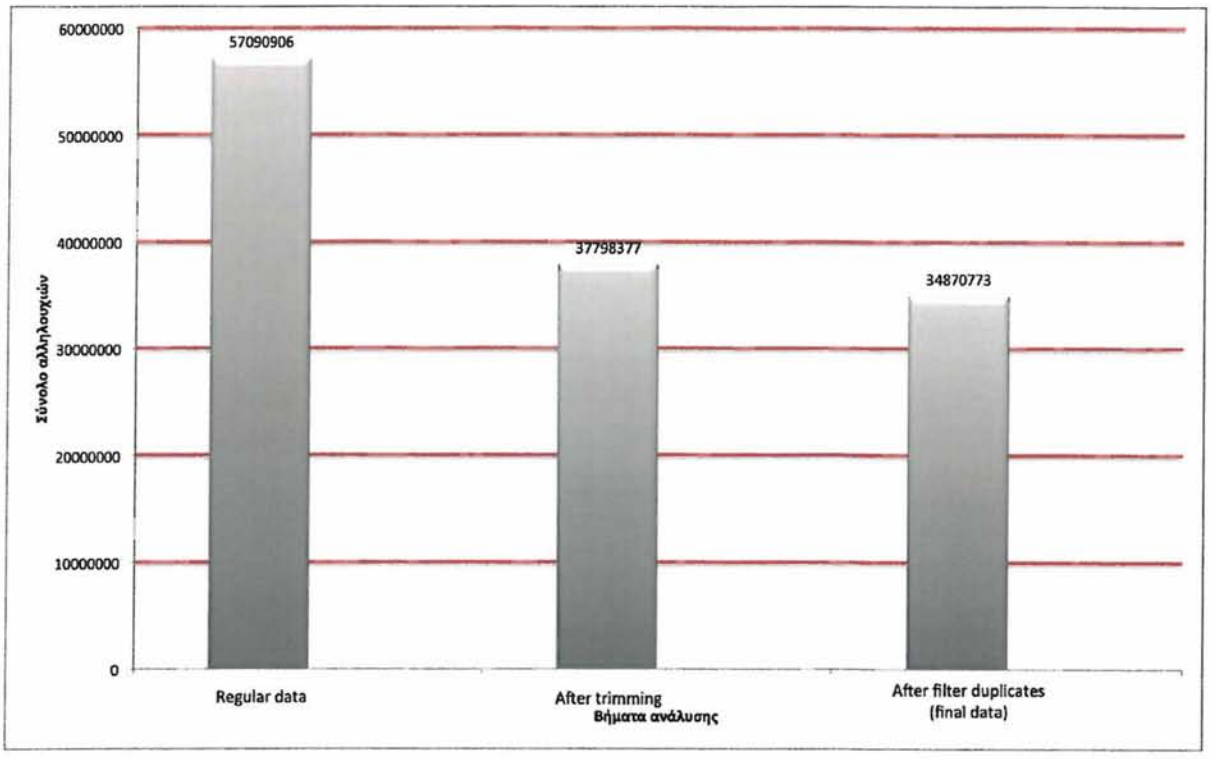
Δεδομένα- αρχεία	Σύνολο αλληλουχιών	Μήκος των reads	%GC
Αρχικά δεδομένα-αρχεία fastq			
SRR330441_1.fastq	57.090.906	50	51
SRR330441_2.fastq	57.090.906	50	52
Δεδομένα μετά το trimming			
SRR330441_trim1.fastq	37.798.377	35-50	49
SRR330441_trim2.fastq	37.798.377	35-50	49
Δεδομένα μετά την αφαίρεση των διπλασιασμένων reads			
SRR330441_uniq1.fastq	34.870.773	35-50	49
SRR330441_uniq2.fastq	34.870.773	35-50	49



Εκ.48:Διάγραμμα απεικόνισης της ποιότητας όλων των βάσεων των reads, μετά το φιλτράρισμα-αφαίρεση των διπλασιασμένων reads, για το αρχείο SRR330441_uniq1.fastq



Εκ.49:Διάγραμμα απεικόνισης της ποιότητας όλων των βάσεων των reads, μετά το φιλτράρισμα-αφαίρεση των διπλασιασμένων reads, για το αρχείο SRR330441_uniq2.fastq



Εικ.50: Ραβδόγραμμα απεικόνισης της μεταβολής των συνολικών αλληλουχιών

Βήμα 3^{0} : Στοίχιση (alignment) των reads στην αλληλουχία αναφοράς με το BWA

Το πρόγραμμα BWA μπορεί να αποκτηθεί ελεύθερα από την ιστοσελίδα:
http://bio-bwa.sourceforge.net/

Αρχικά, δημιουργούμε ένα ευρετήριο (index) για το αρχείο hg19.fa που περιέχει την αλληλουχία αναφοράς, με την εντολη:

./bwa index -a bwtsw -p hg19_index hg19.fa (Διάρκεια: 5155 sec)

Κατόπιν, δημιουργούμε αρχεία σε μορφή .sai, και για τα δύο αρχεία fastq, ξεχωριστά, με τις παρακάτω εντολές :

./bwa aln -t 12 hg19_index SRR330441_uniq1.fastq > SRR330441_uniq1_hg19.sai
(Διάρκεια: 694 sec)

./bwa aln -t 12 hg19_index SRR330441_uniq2.fastq > SRR330441_uniq2_hg19.sai
(Διάρκεια: 689 sec)

Τα αρχεία σε μορφή . sai ,είναι ενδιάμεσα αρχεία και στη συνέχεια θα μετατραπούν σε αρχεία SAM.

Ύστερα, δημιουργήσαμε το αρχείο SAM με το bwa sampe.
./bwa sampe hg19_index SRR330441_uniq1_hg19.sai SRR330441_uniq2_hg19.sai
SRR330441_uniq1.fastq SRR330441_uniq2.fastq >
SRR330441_uniq1_uniq2_hg19.sam
(Διάρκεια: 2837 sec)

Βήμα 4°: Ανάλυση των δεδομένων με το πρόγραμμα Picard

Στη συνέχεια μεταφέραμε το αρχείο SAM (SRR330441_uniq1_uniq2_hg19.sam) στον φάκελο του πρόγραμματος Picard και διαχειριστήκαμε, όπως θα δούμε και παρακάτω, τα εργαλεία του. Τα εγαλεία του Picard στηρίζονται σε προγραμματισμό της Java.

4.1 Καθαρισμός του αρχείου Sam

Η εντολή που δώσαμε ήταν η εξής:

java -Xmx4g -Djava.io.tmpdir=/tmp -jar CleanSam.jar
INPUT=SRR330441_uniq1_uniq2_hg19.sam
OUTPUT=SRR330441_uniq1_uniq2_hg19.Pic1.sam

(Διάρκεια: 10min)

4.2 SortSam

java -Xmx4g -Djava.io.tmpdir=/tmp -jar SortSam.jar
INPUT=SRR330441_uniq1_uniq2_hg19.Pic1.sam
OUTPUT=SRR330441_uniq1_uniq2_hg19.Pic2.sam SORT_ORDER=queryname

Το όνομα του αρχείου που δημιουργήθηκε ήταν:
SRR330441_uniq1_uniq2_hg19.Pic2.sam
(Διάρκεια: 24min)

4.3 Φιλτράρισμα των Reads του αρχείου SAM

java -Xmx4g -Djava.io.tmpdir=/tmp -jar FilterSamReads.jar
INPUT=SRR330441_uniq1_uniq2_hg19.Pic2.sam
OUTPUT=SRR330441_uniq1_uniq2_hg19.Pic3.sam FILTER=includeAligned
SORT_ORDER=coordinate

(Διάρκεια: 38min)

4.4 Σήμανση και απομάκρυνση των διπλασιασμένων reads που προκλήθηκαν στο στάδιο της PCR

java -Xmx4g -Djava.io.tmpdir=/tmp -jar MarkDuplicates.jar
INPUT=SRR330441_uniq1_uniq2_hg19.Pic3.sam
OUTPUT=SRR330441_uniq1_uniq2_hg19.Pic4.sam METRICS_FILE=metrics
CREATE_INDEX=true VALIDATION_STRINGENCY=STRICT
REMOVE_DUPLICATES=true ASSUME_SORTED=true

(Διάρκεια: 20min)

4.5 Μετατροπή του αρχείου SAM σε BAM

Όπως αναφέρθηκε παραπάνω, το αρχείο SAM μετατρέπεται σε μια συμπιεσμένη μορφή, τη μορφή BAM, για λόγους χωρητικότητας και ευκολίας στη διαχείρισή του. Οι πληροφορίες του αρχείου παραμένουν ίδιες.

java -Xmx4g -Djava.io.tmpdir=/tmp -jar SamFormatConverter.jar
INPUT=SRR330441_uniq1_uniq2_hg19.Pic4.sam
OUTPUT=SRR330441_uniq1_uniq2_hg19.Pic4.bam

(Διάρκεια: 17min)

4.6 Προσθήκη ή αντικατάσταση των ομάδων των reads

java -jar AddOrReplaceReadGroups.jar
INPUT=SRR330441_uniq1_uniq2_hg19.Pic4.bam
OUTPUT=SRR330441_uniq1_uniq2_hg19.Pic5.bam SORT_ORDER=coordinate
RGID=1 RGLB=Hum_ex1 RGPL=Illumina RGPU=Hum_ex1 RGSM=Hum_ex1
RGCN=bi RGDS=Hum_ex1 RGDT=2013-1-22 CREATE_INDEX=true
(Διάρκεια: 15min)

4.7 Επικύρωση του αρχείου SAM

java -Xmx4g -Djava.io.tmpdir=/tmp -jar ValidateSamFile.jar
INPUT=SRR330441_uniq1_uniq2_hg19.Pic5.bam OUTPUT=validation.report
VALIDATE_INDEX=true MAX_OUTPUT=100000

Το αποτέλεσμα της εντολής είναι η εξαγωγή του αρχείου
SRR330441_uniq1_uniq2_hg19.Pic5.bam.

(Διάρκεια: 7min)

Βήμα 5ο:Εντοπισμός και επαναστοίχιση προβληματικών θέσεων με το GATK

Σε αυτό το βήμα, γίνεται μεταφορά των αρχείων
SRR330441_uniq1_uniq2_hg19.Pic5.bam και hg19.fa (η αλληλουχία αναφοράς) στον φάκελο GATK. Στη συνέχεια, δίνονται οι παρακάτω εντολές:

5.1 Δημιουργία φακέλου που περιέχει θέσεις που χρειάζονται επαναστοίχιση

java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R hg19.fa -I
SRR330441_uniq1_uniq2_hg19.Pic5.bam –o SRR330441.realigner.intervals
(Διάρκεια: 1h)

Η εντολή που βάλαμε για το αρχείο που θα εξαχθεί (SRR330441.realigner.intervals) θα πρέπει να έχει κατάληξη .realigner.intervals για να συνεχίσουμε στο επόμενο βήμα.

5.2 Επαναστοίχιση προβληματικών θέσεων

java -jar GenomeAnalysisTK.jar -T IndelRealigner -R hg19.fa -I

SRR330441_uniq1_uniq2_hg19.Pic5.bam -targetIntervals
SRR330441_1.realigner.intervals -o SRR330441_uniq1_uniq2_hg19.Pic6.bam
(Διάρκεια: 18min)

Βήμα 6°: Υπολογισμός της κατανομής του μεγέθους ενθέματος μεταξύ paired – end reads

Για το στάδιο αυτό χρησιμοποιούνται προγράμματα του Picard.

Έγινε μεταφορά του αρχείου SRR330441_uniq1_uniq2_hg19.Pic6.bam στον φάκελο Picard-tools και πληκτρολογήσαμε την παρακάτω εντολή:

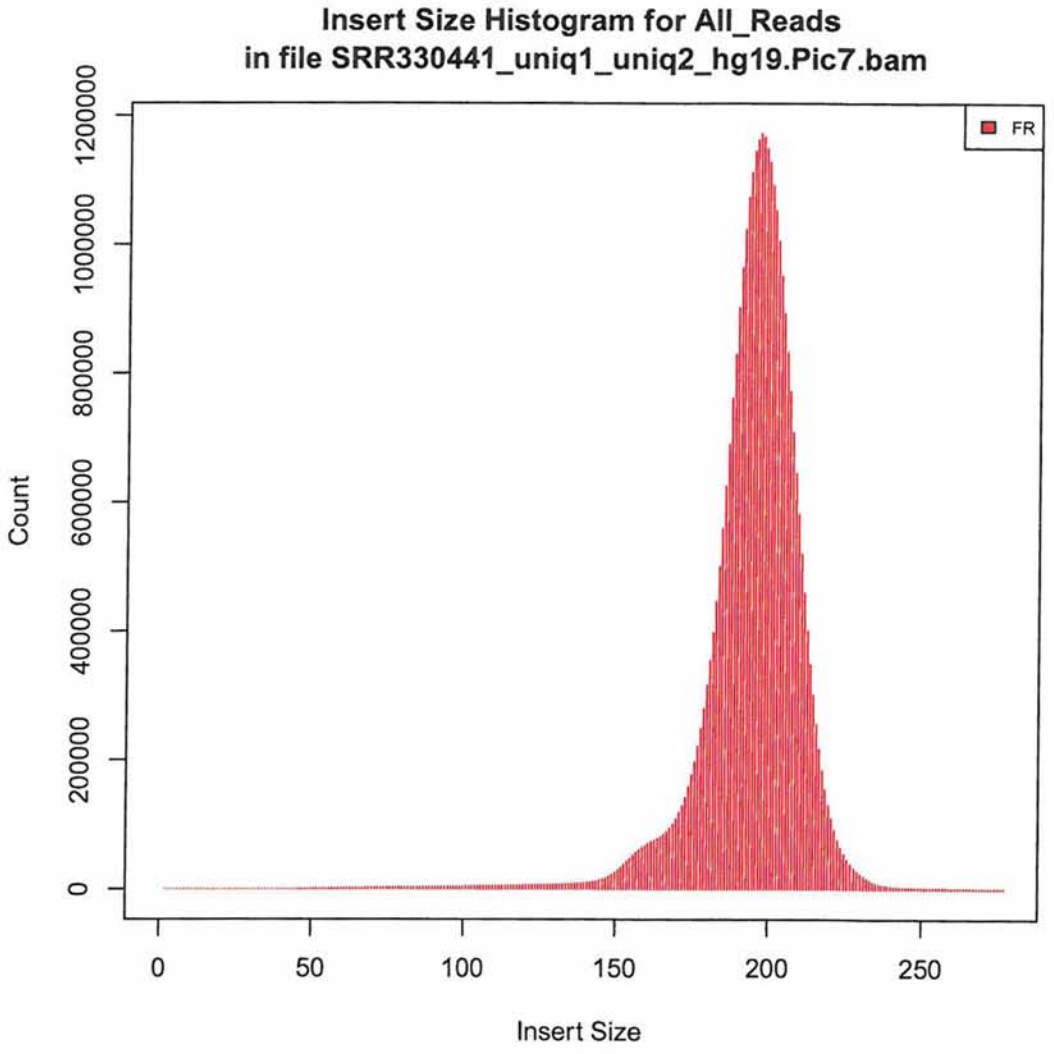
java -Djava.io.tmpdir=/tmp -jar FixMateInformation.jar
INPUT=SRR330441_uniq1_uniq2_hg19.Pic6.bam
OUTPUT=SRR330441_uniq1_uniq2_hg19.Pic7.bam SO=coordinate
VALIDATION_STRINGENCY=LENIENT CREATE_INDEX=true

(Διάρκεια: 26min)
Στη συνέχεια εκτελέστηκε η παρακάτω εντολή:

java -Djava.io.tmpdir=/tmp -jar CollectInsertSizeMetrics.jar
HISTOGRAM_FILE=insert_size_histogram_file_SRR330441
INPUT=SRR330441_uniq1_uniq2_hg19.Pic7.bam
OUTPUT=output_collect_insert_sizes REFERENCE_SEQUENCE=hg19.fa
ASSUME_SORTED=true

(Διάρκεια: 3 min)

Τα αποτελέσματα που πήραμε από αυτή την εντολή ήταν το παρακάτω ιστόγραμμα (εικ.51), που δείχνει την κατανομή του μεγέθους του ενθέματος μεταξύ των paired-end reads για το φάκελο SRR330441_uniq1_uniq2_hg19.Pic7.bam.



Εικ.51: Ιστόγραμμα απεικόνισης του μεγέθους εισαγωγής των reads

Βήμα 7ο: Επαναβαθμολόγιση της ποιότητας (Q-score) των βάσεων

Κατεβάσαμε από τη βάση δεδομένων NCBI το αρχείο με τους ανθρώπινους πολυμορφισμούς SNPs σε μορφή VCF:

ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/VCF/

Το αρχείο με όνομα common_all.vcf, έχει το όνομα των χρωμοσωμάτων μόνο με αριθμούς, ενώ το αρχείο hg19.fa ως, chr1, chrM.

Χρησιμοποιήσαμε ένα perl script για την μετονομασία του VCF αρχείου.

Επαναβαθμολόγιση ποιότητας

Οι εντολές που εκτελέστηκαν ήταν:

java -Xmx4g -jar GenomeAnalysisTK.jar -T BaseRecalibrator -I SRR330441_uniq1_uniq2_hg19.Pic7.bam -R hg19.fa -knownSites corrected_names_common_all_dbSNP137.vcf -o recal_data_SRR330441.grp
(Διάρκεια: 4h)

```
java -jar GenomeAnalysisTK.jar \
  -T PrintReads \
  -R hg19.fa \
  -I SRR330441_uniq1_uniq2_hg19.Pic7.bam \
  -BQSR recal_data_SRR330441.grp \
  -o  SRR330441_uniq1_uniq2_hg19.Pic8.bam
```

(Διάρκεια: 55 min)

Βήμα 8^{ο}: Ανίχνευση των SNPs & φιλτράρισμα με το πρόγραμμα GATK

Η εντολή που δόθηκε ήταν η εξής :

```
java -Xmx4g -jar GenomeAnalysisTK.jar
-glm BOTH
-R hg19.fa
-T UnifiedGenotyper
-I SRR330441_uniq1_uniq2_hg19.Pic8.bam
-o snps.SRR330441_uniq1_uniq2_hg19.Pic8.vcf
-metrics snps.metrics
-stand_call_conf 50.0
-stand_emit_conf 10.0
-dcov 1000
-A DepthOfCoverage
-A AlleleBalance
```

(Διάρκεια: 6h)

Στη συνέχεια έγινε φιλτράρισμα των SNPs:

java -Xmx4g -jar GenomeAnalysisTK.jar -R hg19.fa -T VariantFiltration --variant snps.SRR330441_uniq1_uniq2_hg19.Pic8.vcf -o snps.filtered.SRR330441_uniq1_uniq2_hg19.Pic8.vcf --clusterWindowSize 10 --filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)" --filterName "HARD_TO_VALIDATE" --filterExpression "DP < 5 " --filterName "LowCoverage" --filterExpression "QUAL < 30.0 " --filterName "VeryLowQual" --filterExpression "QUAL > 30.0 && QUAL < 50.0 " --filterName "LowQual" --filterExpression "QD < 1.5 " --filterName "LowQD"

(Διάρκεια: 30sec)

Βήμα 9ο: Λειτουργικός σχολιασμός των SNPs με τη χρήση του προγράμματος Annovar

Μετατροπή του αρχείου SRR330441_uniq1_uniq2_hg19.Pic8.vcf σε μοφή του annovar.

./convert2annovar.pl -format vcf4 snps.filtered.SRR330441_uniq1_uniq2_hg19.Pic8.vcf > snps.filtered.SRR330441_uniq1_uniq2_hg19.Pic8.annovar

(Διάρκεια: λίγα δευτερόλεπτα)

Στην συνέχεια κατεβάσαμε βάσεις δεδομένων για το annovar.

./annotate_variation.pl -buildver hg19 -downdb refgene humandb/

./annotate_variation.pl -buildver hg19 -downdb phastConsElements46way humandb/

./annotate_variation.pl -buildver hg19 -downdb genomicSuperDups humandb/

./annotate_variation.pl -buildver hg19 -downdb snp137 -webfrom annovar humandb/

./annotate_variation.pl -buildver hg19 -downdb avsift -webfrom annovar humandb/

./annotate_variation.pl -buildver hg19 -downdb ljb_all -webfrom annovar humandb/

./annotate_variation.pl -buildver hg19 -downdb 1000g2012apr -webfrom annovar humandb/

./annotate_variation.pl -buildver hg19 -downdb esp6500_all -webfrom annovar humandb/

./summarize_annovar.pl -out snps.filtered.SRR330441_uniq1_uniq2_hg19.Pic8.annovar.annotated -buildver hg19 -verdbsnp 137 -ver1000g 1000g2012apr -veresp 6500 -remove -alltranscript snps.filtered.SRR330441_uniq1_uniq2_hg19.Pic8.annovar humand

Τα αρχεία που εξήχθησαν ήταν τα ακόλουθα:

snps.filtered.SRR330441_uniq1_uniq2_hg19.Pic8.annovar.annotated.hg19_ALL.sites.2012_04_filtered

snps.filtered.SRR330441_uniq1_uniq2_hg19.Pic8.annovar.annotated.hg19_avsift_filtered

snps.filtered.SRR330441_uniq1_uniq2_hg19.Pic8.annovar.annotated.hg19_esp6500_all_ filtered

snps.filtered.SRR330441_uniq1_uniq2_hg19.Pic8.annovar.annotated.hg19_ljb_all_filtered

snps.filtered.SRR330441_uniq1_uniq2_hg19.Pic8.annovar.annotated.hg19_snp137_filtered
snps.filtered.SRR330441_uniq1_uniq2_hg19.Pic8.annovar.annotated.exome_summary
snps.filtered.SRR330441_uniq1_uniq2_hg19.Pic8.annovar.annotated.genome_summary

Τα αρχεία αυτά περιέχουν πληροφορίες των πολυμορφισμών που ανιχνεύθηκαν κατά την ανάλυση όπως το όνομα των SNPs, τη θέση, την ομοζυγωτία ή ετεροζυγωτία των αλληλομόρφων κ.τ.λ.

Βήμα 10ο: Πρόβλεψη φαινοτύπου με τη χρήση του προγράμματος Promethease

Η επεξήγηση των SNPs έγινε με τη χρήση του προγράμματος Promethease. Αρχικά τροποποιήσαμε το παρακάτω αρχείο, snps.filtered.SRR330441_uniq1_uniq2_hg19.Pic8.annovar.annotated.exome_summary (μορφή .csv), στην απαιτούμενη μορφή του Promethease (.txt μορφή).

Κατόπιν, εισάγαμε το αρχείο στο Promethease και πήραμε πληροφορίες (πιν.3) για τους πολυμορφισμούς των δεδομένων μας, σε σχέση με το πώς από αυτά μπορεί να επιδράσουν στον φαινότυπο.

Αποτελέσματα σχολιασμού των πολυμορφισμών (πιν.3)

Πολυμορφισμός (SNP)	Σημασία	Κατάσταση	Συχνότητα Πληθυσμού	Περιγραφή
rs307377(C;T)	4.0	**Good**	None	Επιπλέον δυνατότητα στη γεύση, σπάνιο T αλληλόμορφο προσδίδει καλύτερη ανίχνευση γεύσης umami
gs191	3.1	**Bad**		Μειωμένος μεταβολισμός σε μη στεροειδή αντι-φλεγμονώδη φάρμακα, όπου προκαλείται παράγοντας κινδύνου για γαστρεντερική αιμορραγία με λήψη των παρακάτω φαρμάκων: ακεκλοφενάκη, σελεκοξίμπη, δικλοφενάκη, ιβουπροφαίνη, indomethazine, λορνοξικάμη, μελοξικάμη, ναπροξένη, η πιροξικάμη, τενοξικάμη και valdecoxib.
gs227	3.0			Ετερόζυγωτία σε 3 SNPs που είναι γνωστό ότι επηρεάζουν την αντίληψη ικανότητας στη γεύση πικρότητας.Τα 3 SNPs είναι rs10246939, rs1726866, rs713598 στο γονίδιο TAS2R38
gs161	2.5			CYP2C9, ενδιάμεσοι μεταβολιστές αποτελούν

				το 30% του πληθυσμού. Μπορούν να απαιτούν ελαφρά διαφορετικές δοσολογίες για φάρμακα όπως ταμοξιφένη, βαρφαρίνη, fluvastin, και πολλά μη στεροειδή αντι-φλεγμονώδη όπως είναι η ασπιρίνη, η ιβουπροφαίνη και η ναπροξένη.
rs16969968(A;G)	2.5	**Bad**	45.1%	Ελαφρώς υψηλότερο κίνδυνο για την εξάρτηση από τη νικοτίνη, χαμηλότερο κίνδυνο για την εξάρτηση από την κοκαΐνη
rs5888(C;T)	2.5	**Bad**	53.6%	3x υψηλότερο κίνδυνο για την ηλικιοεξαρτώμενη εκφύλιση της ωχράς κηλίδας
rs12252(C;T)	2.5	**Bad**	None	Μειωμένη ανθεκτικότητα στη γρίπη
rs3743930(C;G)	2.5	**Bad**	None	Κανένας φορέας οικογενούς μεσογειακού πυρετού
rs6265(A;G)	2.4		33.6%	Διαταραχή κινητικών δεξιοτήτων μάθησης. Κίνδυνος νόσου του Alzheimer για μη ApoE4 μεταφορείς, επηρεάζεται από την ετερόζυγη μορφή rs6265.
rs6025(A;G)	2.3	**Bad**	2.7%	Επιρρεπής σε θρόμβωση
rs4149056(C;T)	2.1	**Bad**	28.3%	Μειωμένη απόκριση ορισμένων φαρμάκων 5x αυξημένος κίνδυνος μυοπάθειας για χρήστες στατινών
rs1052133(C;G)	2.1	**Bad**	28.6%	Φυσιολογικό για τον κίνδυνο καρκίνου της ουροδόχου κύστης 1.9x αυξημένο κίνδυνο για καρκίνο της χοληδόχου κύστης
rs1815739(C;T)	2.1		58.4%	Μειωμένοι μυς
gs239	2.0	**Bad**		Γυναίκες που φέρουν τουλάχιστον ένα T σε SNPs rs7501331 και rs12934922 παρουσιάζουν 69% χαμηλότερη ικανότητα να μετατροπής της βήτα-καροτίνης σε ρετινόλη
rs283413(G;T)	2.0		0.9%	3x υψηλότερο κίνδυνο εκδήλωσης νόσου

				Πάρκινσον
rs2476601(A;G)	2.0	**Bad**	21.6%	2,5x κίνδυνο για διαβήτη τύπου 1, RA, νόσος Addison (είναι αρκετά σπάνια διότι φαίνεται να παρέχει ευαισθησία για τέσσερις διαφορετικές αυτοάνοσες διαταραχές) 1.5x κίνδυνο για τη ρευματοειδή αρθρίτιδα 1.5x κίνδυνο για ΣΕΛ 1.8x κίνδυνο για θυρεοειδίτιδα Hashimoto.
rs1136287(C;T)	2.0	**Bad**	36.3%	1.5x αυξημένο κίνδυνο εκφύλιση της ωχράς κηλίδας
rs1676486(A;G)	2.0		37.3%	1.4x κίνδυνο για την LDH
rs2298566(A;C)	2.0		42.5%	Αυξημένο κίνδυνο στεφανιαίας νόσου
rs1051730(C;T)	2.0	**Bad**	45.1%	1.3 φορές αυξημένο κίνδυνο καρκίνου του πνεύμονα
rs2274223(A;G)	2.0	**Bad**	49.6%	0,5x αυξημένο κίνδυνο για καρκίνο του στομάχου και του οισοφάγου (εμφανίζονται σε Κινέζους Han)
rs17576(A;G)	2.0		51.3%	Υψηλότερο κίνδυνο για έμφραγμα του μυοκαρδίου, καρκίνο του πνεύμονα, και χρονια αποφρακτική πνευμονοπάθεια στους καπνιστές
rs1050152(C;T)	2.0	**Bad**	57.5%	2.1x αυξημένο κίνδυνο της νόσου του Crohn
rs1061170(C;T)	2.0	**Bad**	58.3%	2.5x κίνδυνο για την AMD υψηλότερη θνησιμότητα μεταξύ των nonagenarians
rs6152(A;G)	2.0	None		Για άνδρα,αποτροπή εμφάνιση φαλάκρα διότι η μητέρα μεταφέρει ένα αντίγραφο του χρωμοσώματος X SNP
rs5400(C;T)	1.7		19.1%	Σημαντικά υψηλότερη κατανάλωση γλυκόζης
rs11523871(A;C)	1.6		49.2%	1.6x αυξημένο κίνδυνο καρκίνου του μαστού για τις γυναίκες άνω των 60 ετών
rs3764880(A;G)	1.5		13.8%	Πιθανόν 1,2 - 1,8x αυξημένη ευαισθησία για φυματίωση στις γυναίκες
rs6746030(A;G)	1.5		22.1%	Αυτό το SNP φαίνεται να

				επηρεάζει την αντίληψη του πόνου
rs2464196(C;T)	1.5		45.5%	~ 1.5x αυξημένο κίνδυνο καρκίνου του πνεύμονα
rs2241880(C;T)	1.5	**Bad**	56.2%	1.4x αυξημένο κίνδυνο για νόσο του Crohn σε Καυκάσιους
rs5219(C;T)	1.5	None		1,3 φορές αυξημένο κίνδυνο για διαβήτη τύπου 2
rs1799782(C;T)	1.3		11.7%	1,3 φορές αυξημένο κίνδυνο για καρκίνο της στοματικής κοιλότητας μεταξύ των Ασιατών
rs1042713(A;G)	1.3		46.9%	1,3 φορές αυξημένο κίνδυνο επιδείνωσης του άσθματος με χρήση συσκευών για εισπνοές κατα την παιδική ηλικία
rs2549782(G;T)	1.3		48.2%	1,3 x αυξημένο κίνδυνο για προεκλαμψία στους περισσότερους πληθυσμούς. Σημείωση ότι αυτό είναι σχετικό με το εμβρυϊκό γονότυπο και όχι με το μητρικό
gs184	1.2	**Good**		Φυσιολογική ικανότητα γεύση σε πικρό
rs9306160(C;T)	1.2		47.7%	0,75 x (μειωμένος) κίνδυνος για μετάσταση σε LN-/ER + ασθενείς με καρκίνο του μαστού
rs1229984(A;G)		None	0.0%	0.56x μειωμένο κίνδυνο καρκίνου του στόματος / λάρυγγα
rs2230201(A;G)		None	0.0%	1.4x κίνδυνο του λύκου
rs2232165(C;T)		None	4.6%	Αυξημένο κίνδυνο για βαριά κατανάλωση αλκοόλ
rs7951(C;T)		None	4.7%	1.4x κίνδυνο του λύκου
rs2250889(C;G)		None	9.2%	1.46x υψηλότερο κίνδυνο για καρκίνο του πνεύμονα
rs867186(A;G)		None	17.7%	EPCR H3 απλότυπος, Μειώνει ή αυξάνει τον κίνδυνο φλεβικής θρομβοεμβολής
rs1799853(C;T)		None	19.0%	CYP2C9 * 2 μεταφορέας, κατά μέσο όρο 20% μείωση στο μεταβολισμό της βαρφαρίνης
rs11631797(A;G)		None	23.1%	Συνήθως καφέ χρώμα ματιών
rs140504(A;G)		None	24.8%	1.4x αυξημένο κίνδυνο για διπολική διαταραχή
rs2515641(C;T)		None	25.7%	

rs2273535(A;T)	None	26.2%	Υψηλό κίνδυνο καρκίνου
rs2653349(A;G)	None	27.7%	~ 1.5x αυξημένο κίνδυνο για πονοκεφάλους
rs1042714(C;G)	None	30.8%	
rs523349(C;G)	None	32.3%	Αυξημένο κίνδυνο καρκίνου των ωοθηκών
rs324420(A;C)	None	32.7%	Φυσιολογικό
rs6897932(C;T)	None	32.7%	1,3x αυξημένο κίνδυνο για σκλήρυνση κατά πλάκας
rs4961(G;T)	None	33.6%	1.8x αυξημένο κίνδυνο για υψηλή πίεση του αίματος
rs6971091(A;G)	None	35.4%	2x αυξημένο κίνδυνο για οικογενειακή παχυσαρκία
rs1801253(C;G)	None	36.9%	Εξαρτάται από το rs1801252
rs16890979(C;T)	None	38.2%	1.7x κίνδυνο ουρικής αρθρίτιδας
rs27044(C;G)	None	38.5%	1.4x μεγαλύτερο κίνδυνο για σπονδυλίτιδα
rs2074190(C;T)	None	38.9%	2.2X κίνδυνο για AIA
rs1012729(A;G)	None	39.8%	Μεταβολή αρτηριακής πίεσης σε παιδιά
rs8192678(A;G)	None	39.8%	Υψηλότερη αρτηριακή πίεση
rs2304256(A;C)	None	41.1%	1.6x αυξημένο κίνδυνο για SLE
rs2287622(C;T)	None	42.0%	1.7x μεγαλύτερο κίνδυνο για ενδοηπατική χολόσταση της κύησης
rs2470890(C;T)	None	42.9%	
rs30187(C;T)	None	44.2%	1.4x μεγαλύτερο κίνδυνο για σπονδυλίτιδα
rs669(A;G)	None	44.6%	Ενδεχόμενη άυξηση κινδύνου για Αλτσχάιμερ
rs1148259(G;T)	None	45.1%	Ενδιάμεση αναλογία σφιγγομυελίνης
rs1801274(C;T)	None	45.1%	Περίπλοκο, γενικά μεγαλύτερος κίνδυνος για την πρόοδο του καρκίνου
rs2270968(A;C)	None	45.5%	
rs3184504(C;T)	None	45.5%	Αυξημένος κίνδυνος για celiac ασθένεια
rs4633(C;T)	None	45.9%	Υψηλός κίνδυνος για καρκίνο του ενδομητρίου
rs4680(A;G)	None	46.0%	
rs1131532(C;T)	None	46.8%	Φυσιολογικό
rs1049550(C;T)	None	46.9%	0.62x μειωμένο κίνδυνο για σαρκοείδωση
rs3900940(C;T)	None	46.9%	Αυξημένος κίνδυνος στεφανιαίας νόσου. Καλύτερη ανταπόκριση στις στατίνες
rs438034(C;T)	None	46.9%	Σε περιπτώσεις καρκίνου του μαστού μικρό

				ποσοστό επιβίωσης των ασθενών
rs602662(A;G)		None	46.9%	
rs662(A;G)		None	46.9%	0.65x μικρότερος κίνδυνος του καρκίνου των ωοθηκών. Υψηλότερος κίνδυνος στεφανιαίας νόσου, σε ορισμένες μελέτες
rs855791(C;T)		None	46.9%	0,1 g / dL χαμηλότερη αιμοσφαιρίνη κατά μέσο όρο
rs1800860(A;G)		None	48.2%	10% μικρότερα νεφρά ως νεογνό
rs12934922(A;T)	1.0		49.2%	Μειωμένη μετατροπή της βήτα-καροτίνης σε ρετινόλη
rs7501331(C;T)	1.0		49.5%	Μειωμένη μετατροπή της βήτα-καροτίνης σε ρετινόλη
rs25487(A;G)		None	50.0%	2x υψηλότερος κίνδυνος για καρκίνο του δέρματος
rs3740066(A;G)		None	50.8%	1.6x κίνδυνος για ICP
rs2305480(C;T)		None	51.3%	Φυσιολογικό
rs13181(G;T)		None	52.2%	1.12x αυξημένος κίνδυνος για καρκίνο του δέρματος
rs6313(C;T)		None	52.2%	Υψηλός κίνδυνος για RA
rs693(C;T)		None	52.2%	Αυξημένα λιπίδια
rs1799983(G;T)		None	53.8%	Αυξημένος κίνδυνος προεκλαμψίας
rs2227928(C;T)		None	57.5%	Μειωμένη απόκριση σε καρκίνο του παγκρέατος
rs6277(C;T)		None	58.7%	1.4x υψηλότερος κίνδυνος σχιζοφρένειας
rs12150220(A;T)		None	59.5%	Ελαφρώς αυξημένος κίνδυνος για διάφορες αυτοάνοσες νόσους
rs1143674(A;G)		None	None	1,3 φορές αυξημένος κίνδυνος για αυτισμό
rs396991(G;T)		None	None	Περίπλοκο,γενικά μεγαλύτερο κίνδυνο για καρκίνο
rs4894(A;C)		None	None	1.78x αυξημένος κίνδυνος για σχιζοφρένεια στους άνδρες
rs5020278(A;G)		None	None	Γεύση σε γλυκό% 22 % 22 όσφρηση
rs5092(A;G)		None	None	Μικρή πιθανότητα αύξησης βάρους με πρόσληψη ολανζαπίνης
rs854560(A;T)		None	None	Υψηλότερος κίνδυνος για καρδιακή νόσο και διαβητική αμφιβληστροειδοπάθεια
rs1048661(G;T)	0.5	None	None	Ενδεχομένως υψηλότερος

				κίνδυνος γλαυκώματος
rs601338(A;G)	0.1	**Bad**	45.5%	Ευπαθής σε λοιμώξεις Norovirus
rs1726866(C;T)	0.1		46.9%	Ανίχνευση πικρής γεύσης
rs10246939(C;T)	0.1		47.8%	Ανίχνευση πικρής γεύσης
rs713598(C;G)	0.1		51.6%	Ανίχνευση πικρής γεύσης
rs12021720(C;T)	0.05		17.7%	Ασθένεια ούρων Maple Syrup, αλλά φαίνεται ακίνδυνη
rs1126809(A;G)	0.0	**Good**	33.8%	Μικρή αύξηση του κινδύνου καρκίνου του δέρματος
rs688(C;T)	0.0		44.2%	Φυσιολογικό κίνδυνο για Αλτσχάιμερ.Στις γυναίκες υψηλότερα επίπεδα ολικής και LDL χοληστερόλης
rs1802710(C;T)	0.0	**Good**	46.0%	Κοινό στο ολοκλ.γονιδίωμα
rs1800858(A;G)	0.0		48.7%	Πιθανώς φυσιολογικό
rs272879(C;G)	0.0	**Good**	50.8%	Κοινό στο ολοκλ.γονιδίωμα
rs272893(A;G)	0.0	**Good**	52.2%	Κοινό στο ολοκλ.γονιδίωμα
rs509749(A;G)	0.0		56.6%	Κοινό, ελαφρά αύξηση του κινδύνου στους ΣΕΛ
rs1800974(A;G)	0.0	**Good**	57.8%	Κοινό στο ολοκλ.γονιδίωμα

4. Συζήτηση

Η ανάλυση των SNPs είχε ως αποτέλεσμα την ανίχνευση και τον σχολιασμό 1.305 γονοτύπων. Οι πληροφορίες ενός πολυμορφισμού αφορούν: την πιθανότητα δημιουργίας παθολογικών καταστάσεων στον συμμετέχοντα, την συχνότητα που εμφανίζεται στον πληθυσμό της Ευρώπης και τον αριθμό της βαρύτητα που έχει ο κάθε πολυμορφισμός σύμφωνα με την βαθμολόγηση του Promethease. Επίσης, αναφέρονται πολυμορφισμοί οι οποίοι προσδίδουν μοναδικά χαρακτηριστικά στο άτομο, καθώς και τον βαθμό απόκρισης του μεταβολισμού του ατόμου σε διάφορες χημικές ουσίες.

Το Promethease αξιολογεί τους πολυμορφισμούς, ανάλογα με τις έρευνες που έχουν γίνει και με την συχνότητα που αυτοί βρίσκονται στον πληθυσμό, με βάση την κλίμακα από 0-10, με 0 όταν ο πολυμορφισμός είναι κοινός και χωρίς ιδιαίτερη σημασία και 10 όταν αποτελεί σημαντική πληροφορία.

Συγκεκριμένα, τα δεδομένα του ατόμου που αναλύθηκαν, έδειξαν την ύπαρξη του πολυμορφισμού rs307377(C;T), ο οποίος προσδίδει στο άτομο αυξημένη αίσθηση της γεύσης και ειδικότερα στην πικρή και αυτό οφείλεται στο σπάνιο T αλληλόμορφο.

Επίσης, το άτομο πιθανόν να εμφανίζει μειωμένο μεταβολισμό σε μη στεροειδή αντιφλεγμονώδη φάρμακα, όπου προκαλείται παράγοντας κινδύνου για γαστρεντερική αιμορραγία στη λήψη των φαρμάκων ακεκλοφενάκη, σελεκοξίμπη, δικλοφενάκη, ιβουπροφαίνη, indomethazine, λορνοξικάμη, μελοξικάμη, ναπροξένη, η πιροξικάμη, τενοξικάμη και valdecoxib. Η βαρύτητα της αναφοράς του συγεκριμμένου SNP στον συμμετέχοντα, έχει συντελεστή 4.0. Ο rs4149056(C;T) δηλώνει την μειωμένη απόκριση ορισμένων φαρμάκων και υπάρχει 5 φορές μεγαλύτερος κίνδυνος μυοπάθειας για χρήστες στατινών.

Οι πολυμορφισμοί rs16969968(A;G), rs16969968(A;G) και rs1136287(C;T) δηλώνουν πιθανότητα κινδύνου για εκφύλιση της ωχράς κηλίδας ενώ υπάρχει 2.1 φορές αυξημένος κίνδυνος της νόσου Chron λόγω της παρουσίας των rs1050152(C;T) και rs2241880(C;T).

Μία αξιοσημείωτη ανίχνευση είναι εκείνη του πολυμορφισμού rs2476601(A;G) ο οποίος προσδίδει 2,5 φορές κίνδυνο για διαβήτη τύπου 1, RA, νόσος Addison (είναι αρκετά σπάνια διότι φαίνεται να παρέχει ευαισθησία για τέσσερις διαφορετικές αυτοάνοσες διαταραχές), 1.5φορές κίνδυνο για τη ρευματοειδή αρθρίτιδα και 1.5 φορές κίνδυνο για Συστηματικό Ερυθηματώδη Λύκο (ΣΕΛ).

Εξαιτίας της εύρεσης του rs12252(C;T) , ο συμμετέχοντας μπορεί να εμφανίσει μειωμένη ανθεκτικότητα στη γρίπη και με τον rs601338(A;G) να είναι και ευπαθής σε λοιμώξεις Norovirus.

Ο συμμετέχοντας υπάρχει πιθανότητα να είναι επιρρεπής σε θρόμβωση (rs6025(A;G)), να παρουσιάσει διαταραχή κινητικών δεξιοτήτων μάθησης και κίνδυνο νόσου του Alzheimer για μη ApoE4 μεταφορείς (επηρεάζεται από την ετερόζυγη μορφή rs6265(A;G)).

Αξιόλογα αποτελέσματα είναι και η ανίχνευση των: rs1051730(C;T) που μπορεί να προκαλέσει (1,3 φορές αυξημένο κίνδυνο) καρκίνο του πνεύμονα και του rs1052133(C;G), με 1.9φορές μεγαλύτερο κίνδυνο για καρκίνο της χοληδόχου κύστης.

Τέλος, ο rs16969968 (A;G)), που βρίσκεται και στον συμμετέχοντα, εμφανίζει στο 45.1% του πληθυσμού των Καυκάσιων ελαφρώς υψηλότερο κίνδυνο για την εξάρτηση από τη νικοτίνη και χαμηλότερο κίνδυνο για την εξάρτηση από την κοκαΐνη.

Συμπερασματικά, η ανάλυση του ανθρώπινου εξονιώματος του συμμετέχοντα, συνετέλεσε στην ανίχνευση σημαντικών πολυμορφισμών για την εμφάνιση παθολογικών καταστάσεων, το οποίο μπορεί να βοηθήσει στην έγκαιρη πρόληψη αλλά και διάγνωση των νοσημάτων.

Σε αυτό το σημείο τίθενται ζητήματα βιοηθικής διότι θα πρέπει να αναλογιστούμε την βαρύτητα και την αξιοπιστία των αποτελεσμάτων, καθώς αυτά στηρίζονται στις

πιθανότητες. Ποιός καθορίζει την βαρύτητα και πότε θα πρέπει να ανακοινώνονται τα αποτελέσματα στον συμμετέχοντα;
Παρά το γεγονός ότι η αλληλούχιση του εξονιώματος γίνεται πιο προσιτή στο ευρύ κοινό παρ' όλα αυτά τίθεται ένα μεγάλο και σημαντικό δίλημμα στην επιστημονική κοινότητα (Cassa et al., 2012) για ηθικά, νομικά και κοινωνικά θέματα σε σχέση με τον έλεγχο και την χρήση της αλληλουχίας του ατομικού ανθρώπινου γονιδιώματος (Robertson et al., 2003).

Η κοινωνία θα πρέπει να ενημερωθεί επαρκώς τόσο για τα θετικά όσο και για τα αρνητικά της ανάλυσης του ατομικού γονιδιώματος. Η αλληλούχιση του ανθρώπινου γονιδιώματος μπορεί να παρέχει πληροφορίες σχετικά με την ανταπόκριση σε φάρμακα, στην πρόληψη και στην διάγνωση καθώς και στην ανακάλυψη της λειτουργίας γονιδιών (Robertson et al., 2003).

Όμως, πολλές φορές τα αποτελέσματα της αλληλούχισης είναι διφορούμενα διότι δεν είναι γνωστή η λειτουργία πολλών γονιδίων. Επίσης, η υπεύθυνη διαχείριση της μεγάλης ποσότητας της γενετικής πληροφορίας είναι μια ακόμη διαδικασία που πρέπει να ξεπεραστεί. Έτσι μια πρόωρη ένταξη της ES/WGS στην κλινική φροντίδα ίσως οδηγήσει σε συνέπειες στον τομέα της υγείας (Trakadis, 2012). Δεν θα πρέπει να αγνοείται ότι η διαδικασία της αλληλούχισης όπως και όλες οι τεχνολογίες φέρουν κάποιο ποσοστό αποτυχίας (Kaye et al., 2010).

Οι ηθικές ανησυχίες που προκύπτουν από την αποκάλυψη των αποτελεσμάτων περιστρέφονται γύρω από τα στερεότυπα - στιγματισμό και το δικαίωμα της ιδιοκτησίας (Hull et al., 2011).

Στερεότυπα και στιγματισμός

Τα μέλη ενός πληθυσμού είναι πιο πιθανόν να προσβληθούν από κάποιες ασθένειες σε σχέση με μέλη άλλων πληθυσμών καθώς πολλοί πολυμορφισμοί ποικίλουν σε συχνότητα μεταξύ πληθυσμών. Αυτό είναι πιθανόν να αποτελέσει στερεότυπα για πληθυσμούς (Foster et al, 2006).

Το δικαίωμα της ιδιοκτησίας

Το δικαίωμα της ιδιοκτησίας πάνω στο γονιδίωμα θεωρείται ένα βασικό ζήτημα. Μόνο το άτομο που παρέχει εθελοντικά το γενετικό του υλικό μπορεί να μεταβιβάσει μέσω γραπτής συναίνεσης το δικαίωμα να χρησιμοποιήσουν το DNA του ή την πληροφορία που αυτό περιέχει. Μια τέτοια πολιτική είναι ουσιώδης για την προστασία του ατόμου.

Υπέρ της αποκάλυψης των πληροφοριών

Τα επιχειρήματα στηρίζονται στην υποστήριξη βασικών αξιών όπως αυτονομίας, σεβασμού, φιλανθρωπίας, αμοιβαιότητας. Γενικά, οι σύμμαχοι υπέρ της αποκάλυψης των αποτελεσμάτων φαίνεται πως στηρίζουν την άποψή τους στην γενικότερη υποχρέωση όλων μας για την σωτηρία της ανθρώπινης ζωής ή την πρόληψη σοβαρών ασθενειών μέσω της πληροφόρησης (Miller et al., 2008).

Κατά της αποκάλυψης των πληροφοριών

Στηρίζονται σε δεοντολογικά θέματα όπως για παράδειγμα ότι: α) η έρευνα δεν έχει ως σκοπό να ωφελήσει τον συμμετέχοντα (Cho et al., 2008) β) Οι πληροφορίες ενδέχεται να εμπεριέχουν σφάλματα στην αλληλούχιση και επομένως τα αποτελέσματα είναι εσφαλμένα και γ) τα αποτελέσματα που προκύπτουν θα πρέπει να είναι απόρρητα (Steinsbekk et al., 2012).

Προκειμένου να βρεθεί οριοθέτηση για το πότε πρέπει να αποκαλύπτονται τα αποτελέσματα έχουν δημοσιοποιηθεί κάποιοι περιορισμοί και προϋποθέσεις.

Το 2010, η NHLBI (National Heart, Lung and Blood Institute) παρουσίασε τις προϋποθέσεις που πρέπει να πληρούνται για την αποκάλυψη των αποτελεσμάτων στους συμμετέχοντες και οι οποίες είναι:

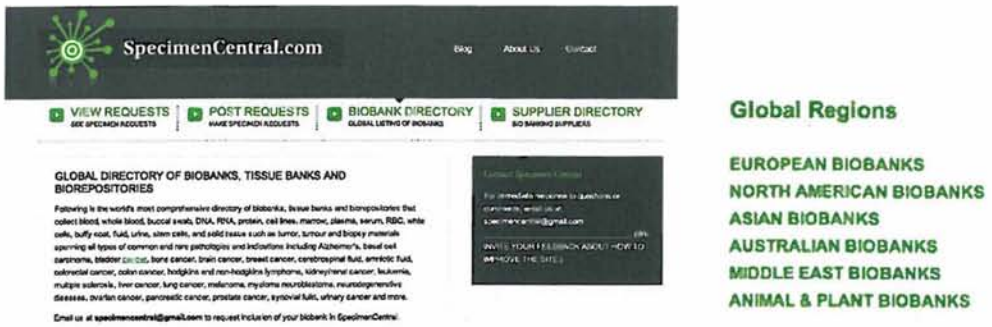
1. Η γενετική ανακάλυψη θα πρέπει έχει σημαντικές επιπτώσεις για την υγεία του συμμετέχοντα.
2. Το γενετικό εύρημα θα πρέπει να είναι αναστρέψιμο. Δηλαδή να υπάρχουν θεραπευτικές και προληπτικές παρεμβάσεις που θα έχουν την δυνατότητα να αλλάξουν την κλινική κατάσταση της ασθένειας.
3. Τα αποτελέσματα θα πρέπει να είναι σύμφωνα με τους ισχύοντες νόμους.

4. Ο συμμετέχοντας θα πρέπει με γραπτή συγκατάθεση για την έρευνα να έχει επιλέξει να ενημερωθεί για τα ατομικά γονιδιωματικά του αποτελέσματα. (Cassa et al., 2012)

Ο Steven Pinker, ο πρώτος άνθρωπος του οποίου αλληλουχήθηκε ολόκληρο το γονιδίωμα, έγραψε : " Άνθρωποι που έχουν μεγαλώσει με τον εκδημοκρατισμό των πληροφοριών δεν θα ανεχτούν κανονισμούς – πρωτόκολλα, που θα τους κρατήσουν μακριά από τα δικά τους γονιδιώματα".

Συμπερασματικά, θα πρέπει λοιπόν ταυτόχρονα με την εξέλιξη της επιστήμης και των τεχνολογιών να διευκρινιστούν διεξοδικώς οι επιπτώσεις των καινοτομιών που δημιουργούν ηθικά ζητήματα (Soden et al., 2012). Επιπρόσθετα, απαιτείται περαιτέρω ανάπτυξη: εναλλακτικών διαδικασιών που να διασφαλίζουν τον συμμετέχοντα, παγκόσμιων μηχανισμών διαχείρισης των ερευνών και των δεδομένων αυτών και τέλος βελτίωση των συνθηκών αποκάλυψης των αποτελεσμάτων (Kaye, 2010).

Η δημιουργία όλο και μεγαλύτερης συλλογής ανθρώπινων βιολογικών δειγμάτων και πληροφοριών, οδήγησε στην ανάγκη αποθήκευσης: των κλινικών δειγμάτων, των πληροφοριών (data) και των αποτελεσμάτων. Η αποθήκευση αυτών γίνεται από εξουσιοδοτημένες, νόμιμες βιο-τράπεζες (Biobanks) (εικ.53), σε ολόκληρο τον κόσμο, που ως στόχο έχουν την ανάλυση, την ασφαλή αποθήκευση των δειγμάτων γενετικού υλικού, τη διαχείριση αυτού και των αποτελεσμάτων, σύμφωνα με την συγκατάθεση του παροχέα του δείγματος (Wolf et al., 2012). Ευχόμαστε να γίνει σωστή χρήση και διαχείριση των δειγμάτων και πληροφοριών.



Εικ.53: Ιστοσελίδα κατοχής λίστας όλων των βιο-τραπεζών παγκοσμίως (http://www.specimencentral.com/biobank-directory.aspx#European%20Biobanks)

Βιβλιογραφία

Akan, Pelin, Henrik Stranneheim, Preben Lexow, and Joakim Lundeberg. "Design and Assessment of Binary DNA for Nanopore Sequencing." *Genome Biology* 11, no. Suppl 1 (2010): P4.

Altmann, André, Peter Weber, Daniel Bader, Michael Preuß, Elisabeth B. Binder, and Bertram Müller-Myhsok. "A Beginners Guide to SNP Calling from High-throughput DNA-sequencing Data." *Human Genetics* 131, no. 10 (August 11, 2012): 1541–1554.

Anselmetti, Dario. "Nanopores: Tiny Holes with Great Promise." *Nature Nanotechnology* 7, no. 2 (2012): 81–82.

Ansorge, W., A. Rosenthal, B. Sproat, C. Schwager, J. Stegemann, and H. Voss. "Non-radioactive Automated Sequencing of Oligonucleotides by Chemical Degradation." *Nucleic Acids Research* 16, no. 5 (1988): 2203–2206.

Ansorge, Wilhelm, Brian Sproat, Josef Stegemann, Christian Schwager, and Martin Zenke. "Automated DNA Sequencing: Ultrasensitive Detection of Fluorescent Bands During Electrophoresis." *Nucleic Acids Research* 15, no. 11 (1987): 4593–4602.

Biesecker, Leslie G, Kevin V Shianna, and Jim C Mullikin. "Exome Sequencing: The Expert View." *Genome Biology* 12, no. 9 (2011): 128.

Cassa, C. A., S. K. Savage, P. L. Taylor, R. C. Green, A. L. McGuire, and K. D. Mandl. "Disclosing Pathogenic Genetic Variants to Research Participants: Quantifying an Emerging Ethical Responsibility." *Genome Research* 22, no. 3 (December 6, 2011): 421–428.

Cho, Mildred K. "Understanding Incidental Findings in the Context of Genetics and Genomics." *The Journal of Law, Medicine & Ethics* 36, no. 2 (June 2008): 280–285.

Clark, Michael J, Rui Chen, Hugo Y K Lam, Konrad J Karczewski, Rong Chen, Ghia Euskirchen, Atul J Butte, and Michael Snyder. "Performance Comparison of Exome DNA Sequencing Technologies." *Nature Biotechnology* 29, no. 10 (September 25, 2011): 908–914.

Cochrane. "Video-Supervised Classification of Sonar Data for Mapping Seafloor

Habitat." 185–194. Alaska Sea Grant, 2008.

Foster, M. W. "Ethical Issues in Medical-sequencing Research: Implications of Genotype-phenotype Studies for Individuals and Populations." *Human Molecular Genetics* 15, no. 90001 (April 15, 2006): R45–R49.

França, Lilian T. C., Emanuel Carrilho, and Tarso B. L. Kist. "A Review of DNA Sequencing Techniques." *Quarterly Reviews of Biophysics* 35, no. 02 (August 20, 2002). http://www.journals.cambridge.org/abstract_S0033583502003797.

Freeman, J. L. "Copy Number Variation: New Insights in Genome Diversity." *Genome Research* 16, no. 8 (June 29, 2006): 949–961. "Geneticresearch.pdf".

Hull S.C., Goldenberg, A.J., .B.S. Wilfond, and R.R. Sharp. "Patient Perspectives on Group Benefits and Harms in Genetic Research." *Public Health Genomics* 14, no. 3 (2011): 135–142.

International Human Genome Sequencing Consortium, "Initial sequencingand analysis of the human genome*", Nature,* (2001).

Kaye, Jane, Paula Boddington, Jantina De Vries, Naomi Hawkins, and Karen Melham. "Ethical Implications of the Use of Whole Genome Methods in Medical Research." *European Journal of Human Genetics* 18, no. 4 (2009): 398–403.

Kevles J.Daniel,Hood Leroy, "*The Code of Codes: Scientific and social Issues in the Human Genome Project*", 1992, Harvard University Press.

Kodama, Y., M. Shumway, R. Leinonen, and on behalf of the International Nucleotide Sequence Database Collaboration. "The Sequence Read Archive: Explosive Growth of Sequencing Data." *Nucleic Acids Research* 40, no. D1 (October 18, 2011): D54–D56.

Ku, Chee-Seng, David N. Cooper, Constantin Polychronakos, Nasheen Naidoo, Mengchu Wu, and Richie Soong. "Exome Sequencing: Dual Role as a Discovery and Diagnostic Tool." *Annals of Neurology* 71, no. 1 (January 2012): 5–14.

Lander, Eric S., and Michael S. Waterman. "Genomic Mapping by Fingerprinting Random Clones: a Mathematical Analysis." *Genomics* 2, no. 3 (1988): 231–239.

Lee, H. C., K. Lai, M. T. Lorenc, M. Imelfort, C. Duran, and D. Edwards.

"Bioinformatics Tools and Databases for Analysis of Next-generation Sequence Data." *Briefings in Functional Genomics* 11, no. 1 (December 19, 2011): 12–24.

Leinonen, R., R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tarraga, Y. Cheng, I. Cleland, et al. "The European Nucleotide Archive." *Nucleic Acids Research* 39, no. Database (October 23, 2010): D28–D31.

Leinonen, R., H. Sugawara, M. Shumway, and on behalf of the International Nucleotide Sequence Database Collaboration. "The Sequence Read Archive." *Nucleic Acids Research* 39, no. Database (November 9, 2010): D19–D21.

Lupski, James R., and Pawel Stankiewicz. "Genomic Disorders: Molecular Mechanisms for Rearrangements and Conveyed Phenotypes." *PLoS Genetics* 1, no. 6 (2005): e49.

Maitra, Raj D., Jungsuk Kim, and William B. Dunbar. "Recent Advances in Nanopore Sequencing." *ELECTROPHORESIS* 33, no. 23 (December 2012): 3418–3428.

Majewski, J., J. Schwartzentruber, E. Lalonde, A. Montpetit, and N. Jabado. "What Can Exome Sequencing Do for You?" *Journal of Medical Genetics* 48, no. 9 (July 5, 2011): 580–589.

Mamanova, Lira, Alison J Coffey, Carol E Scott, Iwanka Kozarewa, Emily H Turner, Akash Kumar, Eleanor Howard, Jay Shendure, and Daniel J Turner. "Target-enrichment Strategies for Next-generation Sequencing." *Nature Methods* 7, no. 2 (February 2010): 111–118.

Mardis, Elaine R. "A Decade's Perspective on DNA Sequencing Technology." *Nature* 470, no. 7333 (February 10, 2011): 198–203.

McVean, Gil A., David M. Altshuler (Co-Chair), Richard M. Durbin (Co-Chair), Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, et al. "An Integrated Map of Genetic Variation from 1,092 Human Genomes." *Nature* 491, no. 7422 (October 31, 2012): 56–65.

Metzker, Michael L. "Sequencing Technologies — the Next Generation." *Nature Reviews Genetics* 11, no. 1 (December 8, 2009): 31–46.

Miller, Franklin G., Michelle M. Mello, and Steven Joffe. "Incidental Findings in Human Subjects Research: What Do Investigators Owe Research Participants?" *The Journal of Law, Medicine & Ethics* 36, no. 2 (June 2008): 271–279.

Naidoo, Nasheen, Yudi Pawitan, Richie Soong, David N. Cooper, and Chee-Seng Ku.

"Human Genetics and Genomics a Decade After the Release of the Draft Sequence of the Human Genome." *Human Genomics* 5, no. 6 (2011): 577–622.

Ng, Sarah B., Emily H. Turner, Peggy D. Robertson, Steven D. Flygare, Abigail W. Bigham, Choli Lee, Tristan Shaffer, et al. "Targeted Capture and Massively Parallel Sequencing of 12 Human Exomes." *Nature* 461, no. 7261 (August 16, 2009): 272–276.

"Niedringhaus Et Al Landscape of Next Generation Sequencing Technologies Pdf Free Ebook Download from Www.eeb.uconn.edu.htm", n.d.

Niedringhaus, Thomas P., Denitsa Milanova, Matthew B. Kerby, Michael P. Snyder, and Annelise E. Barron. "Landscape of Next-Generation Sequencing Technologies." *Analytical Chemistry* 83, no. 12 (June 15, 2011): 4327–4341.

Pareek, Chandra Shekhar, Rafal Smoczynski, and Andrzej Tretyn. "Sequencing Technologies and Genome Sequencing." *Journal of Applied Genetics* 52, no. 4 (June 23, 2011): 413–435.

Patel, Ravi K., and Mukesh Jain. "NGS QC Toolkit: a Toolkit for Quality Control of Next Generation Sequencing Data." *PloS One* 7, no. 2 (2012): e30619.

Parla, Jennifer S., Ivan Iossifov, Ian Grabill, Mona S. Spector, Melissa Kramer, and W. Richard McCombie. "A Comparative Analysis of Exome Capture." *Genome Biol* 12, no. 9 (2011): R97.

"Pawel Stankiewicz, Christine J. Shaw, Marjorie Withers, Et Al- Serial Segmental Duplications During Primate Evolution Result in Complex Human Genome Architecture.htm", n.d.

Pushpendra Kumar, Gupta. "Comparative DNA Sequence Analysis Involving Wheat, Brachypodium and Rice Genomes Using Mapped Wheat ESTs." *Triticeae Genomics and Genetics* (2012). http://bio.sophiapublisher.com/html-337-16-tgg.

Russel J. Peter, "iGenetics", 2005, Πανεπιστημιακές Εκδόσεις Μπάσδρα.

Robertson, John A. "The $1000 Genome: Ethical and Legal Issues in Whole Genome Sequencing of Individuals." *American Journal of Bioethics* 3, no. 3 (2003): 35–42.

Rodriguez – Ezpeleta Naiara , Michael Hackenberg,Ana M.Aransay, *"Bioinformatics for High Throughput Sequencing*",2012,Springer.

Ruffalo, M., M. Koyuturk, S. Ray, and T. LaFramboise. "Accurate Estimation of Short Read Mapping Quality for Next-generation Genome Sequencing." *Bioinformatics* 28, no. 18 (September 7, 2012): i349–i355.

Schneider, Grégory F., and Cees Dekker. "DNA Sequencing with Nanopores." *Nature Biotechnology* 30, no. 4 (2012): 326–328.

Shendure, Jay, and Hanlee Ji. "Next-generation DNA Sequencing." *Nature Biotechnology* 26, no. 10 (October 2008): 1135–1145.

Shokralla, Shadi, Jennifer L. Spall, Joel F. Gibson, and Mehrdad Hajibabaei. "Next-generation Sequencing Technologies for Environmental DNA Research." *Molecular Ecology* 21, no. 8 (April 2012): 1794–1805.

Singleton, Andrew B. "Exome Sequencing: a Transformative Technology." *The Lancet Neurology* 10, no. 10 (October 2011): 942–946.

Sobreira, Nara LM, Elizabeth T. Cirulli, Dimitrios Avramopoulos, Elizabeth Wohler, Gretchen L. Oswald, Eric L. Stevens, Dongliang Ge, Kevin V. Shianna, Jason P. Smith, and Jessica M. Maia. "Whole-genome Sequencing of a Single Proband Together with Linkage Analysis Identifies a Mendelian Disease Gene." *PLoS Genetics* 6, no. 6 (2010): e1000991."Whole-genome Sequencing of a Single Proband Together with Linkage Analysis Identifies a Mendelian Disease Gene." *PLoS Genetics* 6, no. 6 (2010): e1000991.

Soden, Sarah E, Emily G Farrow, Carol J Saunders, and John D Lantos. "Genomic Medicine: Evolving Science, Evolving Ethics." *Personalized Medicine* 9, no. 5 (July 2012): 523–528.

Teer, J. K., and J. C. Mullikin. "Exome Sequencing: The Sweet Spot Before Whole Genomes." *Human Molecular Genetics* 19, no. R2 (August 12, 2010): R145–R151.

Trakadis, Yannis J. "Patient-controlled Encrypted Genomic Data: An Approach to Advance Clinical Genomics." *BMC Medical Genomics* 5, no. 1 (2012): 31.

Venkatesan, Bala Murali, and Rashid Bashir. "Nanopore Sensors for Nucleic Acid Analysis." *Nature Nanotechnology* 6, no. 10 (September 18, 2011): 615–624.

Wan, W., Y. Chong, L. Ge, H. Noh, A. D. Stone, and H. Cao. "Time-Reversed Lasing and Interferometric Control of Absorption." *Science* 331, no. 6019 (February 17, 2011): 889–892.

Watson D. James, Richardr M.Myers, Caudy A.Amy, Witkowski A. Jan, "Ανασυνδιασμένο DNA", 2007, Πανεπιστημιακές Εκδόσεις Μπάσδρα.

Wolf, Susan M., Brittney N. Crock, Brian Van Ness, Frances Lawrenz, Jeffrey P. Kahn, Laura M. Beskow, Mildred K. Cho, et al. "Managing Incidental Findings and Research Results in Genomic Research Involving Biobanks and Archived Data Sets." *Genetics in Medicine* 14, no. 4 (April 2012): 361–384.

Zhou, Xiaoguang, Lufeng Ren, Qingshu Meng, Yuntao Li, Yude Yu, and Jun Yu. "The Next-generation Sequencing Technology and Application." *Protein & Cell* 1, no. 6 (July 7, 2010): 520–536.

Πηγές από το διαδίκτυο

http://www.ocf.berkeley.edu/~edy/genome/nhgri.html

http://www.populationdiagnostics.com/science.html

http://sulleormedisherlockholmes.protagonista.altervista.org/sulleormedisherlockholmes/Sulle_orme_di_holmes/sequenziamento_dna.html

http://www.genome.gov/sequencingcost

http://www.bioopticsworld.com/articles/print/volume-5/issue-06/features/dna-sequencing-technologies-the-next-generation-and-beyond.html

http://investor.pacificbiosciences.com/releasedetail.cfm?ReleaseID=732039

http://www.illumina.com

www.illumina.com/NGS

http://www.cbcb.umd.edu/research/assembly_primer.shtml

http://www2.technologyreview.com/article/427677/nanopore-sequencing/

http://www.nature.com/scientificamerican/journal/v294/n1/box/scientificamerican0106-46_BX4.html

http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Sequencing/Semiconductor-Sequencing/Semiconductor-Sequencing-Technology/Ion-Torrent-Technology-How-Does-It-Work.html

http://www.vincentabry.com/en/ion-proton-sequencer-decodes-human-genome-in-1-day-for-1000-dollars-1543

http://www.cureffi.org/2012/12/19/forward-and-reverse-reads-in-paired-end-sequencing

http://en.wikipedia.org/wiki/ASCII

http://samtools.sourceforge.net/SAM1.pdf

http://www.ncbi.nlm.nih.gov

http://genome.ucsc.edu

http://www.galter.northwestern.edu/news/index.cfm/2009/9/15/NCBI-Short-Read-Archive-SRA-of-NextGeneration-Sequencing-Data

http://www.ebi.ac.uk/ena/about/sra_submissions

http://www.ncbi.nlm.nih.gov/Traces/sra/?view=announcement

http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?view=list_arrivals

http://www.ebi.ac.uk

http://www.1000genomes.org/data

http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi

http://seqanswers.com

http://www.biostars.org

http://snpedia.com/index.php/SNPedia

http://www.ncbi.nlm.nih.gov/SNP

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips

http://www.wikipedia.org

ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/VCF

http://openi.nlm.nih.gov/detailedresult.php?img=3096631_pone.0019534.g001&req=4