ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών

του Τμήματος Βιοχημείας και Βιοτεχνολογίας

ΕΦΑΡΜΟΓΕΣ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ - ΜΟΡΙΑΚΗ ΓΕΝΕΤΙΚΗ-ΔΙΑΓΝΩΣΤΙΚΟΙ ΔΕΙΚΤΕΣ

ΤΣΟΥΧΛΟΥ ΠΑΡΑΣΚΕΥΗ

«Βιοπληροφορική και εξελικτική ανάλυση δεδομένων RNA-Sequencing από Δίθυρα»

ΛΑΡΙΣΑ 2015

Bioinformatics and evolutionary analysis of RNA-Sequencing data from the Bivalvia

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

κ. Αμούτζιας Γρηγόριος (Επιβλέπων)

Επίκουρος Καθηγητής Βιοπληροφορικής στη Γενωμική, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας

κ. Μαμούρης Ζήσης

Καθηγητής Γενετικής Ζωϊκών Πληθυσμών, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας

κα. Ιμσιρίδου Αναστασία

Αναπληρώτρια Καθηγήτρια Γενετικής Μηχανικής και Βιοτεχνολογίας, Τμήμα Τεχνολογίας

Τροφίμων, Αλεξάνδρειο Τεχνολογικό Εκπαιδευτικό Ίδρυμα Θεσσαλονίκης

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του Μεταπτυχιακού Προγράμματος Σπουδών «*ΕΦΑΡΜΟΓΕΣ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ - ΜΟΡΙΑΚΗ ΓΕΝΕΤΙΚΗ- ΔΙΑΓΝΩΣΤΙΚΟΙ ΔΕΙΚΤΕΣ*», του Τμήματος Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας, υπό την επίβλεψη του Επίκουρου Καθηγητή κ. Γρηγόριου Αμούτζια. Θα ήθελα λοιπόν, να τον ευχαριστήσω θερμά για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον αντικείμενο καθώς και για την αμέριστη βοήθεια και καθοδήγηση του, καθ' όλη την διάρκεια της εκπόνησης της παρούσας εργασίας. Ευχαριστώ θερμά τον Καθηγητή Ζήση Μαμούρη για τις πολύτιμες γνώσεις που μου προσέφερε κατά τη διάρκεια του μεταπτυχιακού προγράμματος και για την πολύτιμη βοήθειά του κατά την εκπόνηση της εργασίας αυτής. Ευχαριστώ θερμά και την Αναπληρώτρια Καθηγήτρια Αναστασία Ιμσιρίδου για την πολύτιμη βοήθεια της στην εκπόνησης της εργασίας, για τον καταλυτικό της ρόλο και για την βοήθειά της στην συλλογή των δειγμάτων. Ακόμα, θα ήθελα να ευχαριστήσω θερμά την κα. Σοφία Γαληνού-Μητσούδη,

Καθηγήτρια ΑΤΕΙΘ, ΣΤΕΦ, Τμήμα Πολιτικών Μηχανικών ΑΤΕΙ Θεσσαλονίκης για τον καταλυτικό της ρόλο στην πραγματοποίηση αυτής της ερευνητικής προσπάθειας και για την πολύτιμη βοήθειά της στην συλλογή των δειγμάτων. Ευχαριστώ θερμά και τον Δρ. Αθανάσιο Μανούση για την συλλογή δειγμάτων. Τέλος, θα ήθελα να ευχαριστήσω και τους Θέμη Γιαννούλη και Στυλιανή Γεωργίου για την απομόνωση του ολικού RNA από τα δείγματα Πίννας.

1 ΠΕΡΙΛΗΨΗ

Η εξέλιξη των Διθύρων είναι ένα θέμα που ακόμα δεν έχει διαλευκανθεί. Η παρούσα διπλωματική εργασία μελετά τις εξελικτικές σχέσεις μεταξύ των βασικών ομάδων των Διθύρων μέσω της βιοπληροφορικής ανάλυσης μοριακών ακολουθιών των αντιπροσώπων κάθε ομάδας. Συγκεκριμένα, η φυλογενετική ανάλυση στηρίχθηκε σε δημοσιευμένα γενωμικά δεδομένα καθώς και σε δεδομένα RNA-SEQ, από τεχνολογίες νέας γενιάς (next generation sequencing). Για την Pinna nobilis παράχθηκαν νέα δεδομένα RNA-SEQ. Με προγράμματα βιοπληροφορικής μπόρεσαν να ανακατασκευαστούν *de novo* τα μεταγραφώματα κάθε οργανισμού, αφού πρώτα ελέγχθηκε η ποιότητά τους. Τα μεταγραφώματα χρησιμοποιήθηκαν για την πρόβλεψη των πρωτεϊνών και στη συνέχεια έγινε η εύρεση των ορθόλογων γονιδίων με τη μέθοδο του ανταποδωτικού blast (best reciprocal blastp) και με γονιδίωμα αναφοράς το δημοσιευμένο Crassostrea gigas. Η ανάλυση συνεχίστηκε με τον περαιτέρω εντοπισμό μιας ομάδας 143 γονιδίων που είχαν ορθόλογα σε 14 οργανισμούς και μιας άλλης ομάδας 785 γονιδίων που είχαν ορθόλογα γονιδία σε 7 οργανισμούς. Από αυτές τις ομάδες γονιδίων κατασκευάστηκαν φυλογενετικά δέντρα και εκτιμήθηκαν οι εξελικτικές σχέσεις των ομάδων Archiheterodonta, Palaeoheterodonta, Imparidentia, Anomalodesmata, Pteriomorphia & Protobranchia.

ABSTRACT

The evolution of Bivalves is still an unresolved issue for the resolution of which, the morphological and molecular data are being investigated. This thesis studies the evolutionary relationships established between the major taxonomic groups in the class of Bivalves, through the bioinformatics analysis of Genomic and RNA-SEQ data from the representatives of each group. Initially, data were downloaded from public databases and in the case of *Pinna nobilis*, new RNA-SEQ data were generated. Then, the transcripts of each organism were reconstructed *de novo* with bioinformatics programs, after quality testing. The transcripts were used for the

prediction of proteins that were later used for the detection of orthologous genes with the method of best reciprocal blastp. The Genome of *Crassostrea gigas* was used as a reference point. The analysis continued with the identification of a set of 143 genes with orthologs in 14 organisms and another set of 785 genes with orthologs in 7 organisms. These two sets of genes were used for the reconstruction of phylogenetic trees. The results were compared with other recently published phylogenetic trees.

2 ΕΙΣΑΓΩΓΗ

2.1 Σκοπός της παρούσας εργασίας

Η παρούσα εργασία είχε ως σκοπό να μπορέσει να εξιχνιάσει τις φυλογενετικές σχέσεις μεταξύ των οργανισμών της ομάδας των Δίθυρων, του φύλου Μαλάκια καθώς και να καθορίσει τη φυλογενετική θέση του οργανισμού Pinna nobilis, που είναι το μεγαλύτερο δίθυρο της Μεσογείου. Για τον σκοπό αυτό χρησιμοποιήθηκαν δεδομένα αλληλούχισης νέας γενιάς (RNA-SEQ), δημοσιευμένα γονιδωματικά δεδομένα και σύγρονες βιοπληροφορικές/γονιδιωματικές αναλύσεις. Αυτή η άνευ προηγουμένου χρήση ενός τεράστιου όγκου δεδομένων αναλύθηκε με τις πιο σύγχρονες βιοπληροφορικές μεθόδους και οδήγησε σε πολύ πιο αξιόπιστα αποτελέσματα.

Στην ομάδα των διθύρων ανήκουν πολλά οικονομικά σημαντικά είδη. Ο απώτερος σκοπός ήταν να χρησιμοποιηθεί η νέα αυτή και πολύ πιο αξιόπιστη φυλογένεση για να μπορέσουμε να μελετήσουμε περαιτέρω γονίδια που εμπλέκονται στην ανάπτυξη των διθύρων και που θα μπορούσαν να αποτελέσουν στόχους για βιοτεχνολογικές εφαρμογές. Με βάση μια αξιόπιστη φυλογένεση θα μπορέσουμε στο μέλλον να εκτιμήσουμε κατά πόσον τα αποτελέσματα μια μοριακής μελέτη που πραγματοποιείται σε ένα δίθυρο μπορούν ή όχι να επεκταθούν και σε άλλα είδη αυτής της ομάδας.

2.2 Μαλάκια (Molluscs)

2.2.1 Γενικά χαρακτηριστικά του Φύλου Μαλάκια (Molluscs)

Οι οργανισμοί του φύλου Μαλάκια αγγίζουν περίπου τους 50.000 σε αριθμό και αποτελούν μια από τις μεγαλύτερες και ποικιλόμορφες ομάδες στο ζωϊκό βασίλειο. Το κύριο χαρακτηριστικό των Μαλακίων είναι το **μαλακό** τους σώμα, από το οποίο

προέρχεται και η ονομασία τους. Ο τρόπος διαβίωσής τους ποικίλει και μπορούν να βρεθούν σε μεγάλο εύρος ενδιαιτημάτων, όπως για παράδειγμα, σε τροπικές περιοχές έως και πολικές, ή σε μεγάλα υψόμετρα έως και χαμηλές μικρές λίμνες. Βέβαια, το μεγαλύτερο ποσοστό των ειδών εντοπίζεται στις θάλασσες.

Τα Μαλάκια ανήκουν στους κοιλωματικούς οργανισμούς και υπάγονται στον κλάδο των Πρωτοστομίων, παρουσιάζοντας συγκεκριμένο τρόπο ανάπτυξης. Μορφολογικά, τα Μαλάκια έχουν σώμα με αμφίπλευρη συμμετρία και απουσία μεταμέρειας, και κοίλωμα που περιορίζεται σε συγκεκριμένα όργανα, όπως η καρδιά και οι γονάδες. Κοιλιακά, υπάρχει το 'πόδι' το οποίο βοηθάει στην κίνηση του οργανισμού, ενώ ραχιαία, υπάρχει ο μανδύας (ο οποίος είναι αποτέλεσμα αναδιπλώσεων του σώματος) που οδηγεί στην έκκριση στου οστράκου (απουσία σε μερικά είδη), ενώ ο ίδιος μπορεί να διαφοροποιηθεί σχηματίζοντας βράγχια ή πνεύμονες. Τα εσωτερικά συστήματα λειτουργίας είναι πλήρως ανεπτυγμένα με χαρακτηριστικό το περίπλοκο πεπτικό τους σύστημα που φέρει ένα καινοτόμο όργανο απόξεσης, το λεγόμενο 'ξύστρο΄.

Τα Μαλάκια με βάση τον τύπο του οστράκου και του ποδιού χωρίζονται σε εφτά ομοταξίες (Figure **1**): τα Γαστρόποδα (Gastropoda), τα Δίθυρα (Bivalvia), τα Σκαφόποδα (Scaphopoda), τα Κεφαλόποδα (Cephalopoda), τους Σωληνόγαστρους (Neomeniomorpha), τα Ουροβοθριωτά (Chaetodermomorpha) και τα Πολυπλακοφόρα (Polyplacophora).

Figure 1. Βασικές φυλογενετικές ομάδες του Φύλου των Μαλακίων (Kocotetal *et al.*, 2011).

2.2.2 Οικογένειες της Ομοταξίας των Δίθυρων (Bivalvia)

Τα Δίθυρα είναι μια από τις μεγαλύτερες ομοταξίες των Μαλακίων και αποτελούνται από οργανισμούς που ζουν σε υφάλμυρα ή γλυκά νερά. Το όστρακο τους αποτελείται από δύο θυρίδες ενωμένες με ένα ραχιαίο σύνδεσμο, οι οποίες διατηρούνται κλειστές με έναν προσαγωγό μυ.

Τα Δίθυρα χωρίζονται σε πέντε βασικές ομάδες, με βάση τα μορφολογικά χαρακτηριστικά τους (NCBI taxonomy, 2014) (Table 1):

Heteroconchia	Ισομερείς προσαγωγοί μυς οδοντοφόρου, μερικώς καμπυλωτό ξύστρο
Paloeoheterodonta	Όχι πλήρως οριοθετημένα όρια του μανδύα, εξωτερικός

	σύνδεσμος πίσω από το περιόστρακο, ξύστρο σαν ανάποδο 'V'
Anomalodesmata	Συνήθως έλλειψη ή ατελώς ανεπτυγμένο ξύστρο
Pteriomorphia	Έλλειψη πραγματικού ξύστρου, ύπαρξη όμως οδοντοφόρου μυ
Protobranchia	Ειδικά διαμορφομένα κτενίδια (βράγχια) για διήθηση και πρόσληψη τροφής

Table 1. Βασικά μορφολογικά χαρακτηριστικά των ομάδων των Δίθυρων (Carter *et al.*, 2006; Newell, 1965)

Οστόσο, νεώτερες μελέτες αναγνωρίζουν 6 κύριες ομάδες (Bieler *et al.*, 2014; Gonzalez *et al.*, 2015).

2.2.3 Βασικά χαρακτηριστικά της Ομοταξίας των Γαστρόποδων (Gastropoda)

Τα γαστρόποδα είναι η μεγαλύτερη ομάδα του φύλου «Μαλάκια», στην οποία ανήκουν τα σαλιγκάρια, τα κοχύλια, οι πεταλίδες και κάποια πολύ εξελιγμένα σαλιγκάρια που ζουν και αναπνέουν στην ξηρά. Χαρακτηριστικό εσωτερικό τους γνώρισμα είναι η συστροφή της μανδυακής κοιλότητας. Έχουν όστρακο με μια θύρα και σώμα ασύμμετρο. Στην συγκεκριμένη εργασία χρησιμοποιήσαμε ως εξωομάδα για τις φυλογενετικές αναλύσεις μας το γαστρόποδο Lottia gigantea (Figure **2**).

Figure 2. Εικόνα του Γαστρόποδου Lottia gigantea. Πηγή: Wikipedia (http://en.wikipedia.org/wiki/Lottia_gigantea).

2.2.4 Εξέλιξη του Φύλου των Μαλακίων (Molluscs)

Τα Μαλάκια προέρχονται από την εξελικτική γραμμή των Μεταζώων, τα οποία, οδήγησαν στη δημιουργία αυτού του φύλου, μέσω της γραμμής των απογόνων των Πρωτοστομίων (Balavoine *et al.*, 1998). Έχουν βρεθεί απολιθώματα σε γεωλογικά στρώματα της εποχής του Καμβρίου. Έχει γίνει η υπόθεση πως ο ΄κοινός πρόγονος΄ των Μαλακίων ήταν πιθανότατα ένας σκωληκόμορφος οργανισμός με γλιστερή κοιλιακή επιφάνεια και ραχιαία επιφάνεια αποτελούμενη από έναν μανδύα με ασβεστολιθικά έλυτρα, ο οποίος οδήγησε εξελικτικά στη δημιουργία διαφόρων ομοταξιών των Μαλακίων, με συγκεκριμένα ξεχωριστά χαρακτηριστικά. Η ομοταξία των Ουροβοθριωτών είναι κοντινότερη στην προγονική μορφή. Οι Σωληνόγαστροι στερούνται στερεού οστράκου, γεγονός που δηλώνει την πρώιμη διαφοροποίησή τους από τις άλλες ομοταξίες. όπως και τα Πολυπλακοφόρα. Για την ομοταξία των

Γαστερόποδων υποστηριζόταν μέχρι πρόσφατα η πολυφυλετικότητα, όμως πρόσφατες μελέτες (Zapata *et al.*, 2014) συνηγορούν στη μονοφυλετικότητά τους. Τα Δίθυρα εμφανίζονται να είναι αδελφά τάξα με τα Σκαφόποδα, σύμφωνα με τον Hickman. Η εξέλιξη, με βάση τους γεωλογικούς χρόνους, των διαφόρων ομοταξιών των Μαλακίων φαίνεται στην παρακάτω εικόνα (Figure **3**).

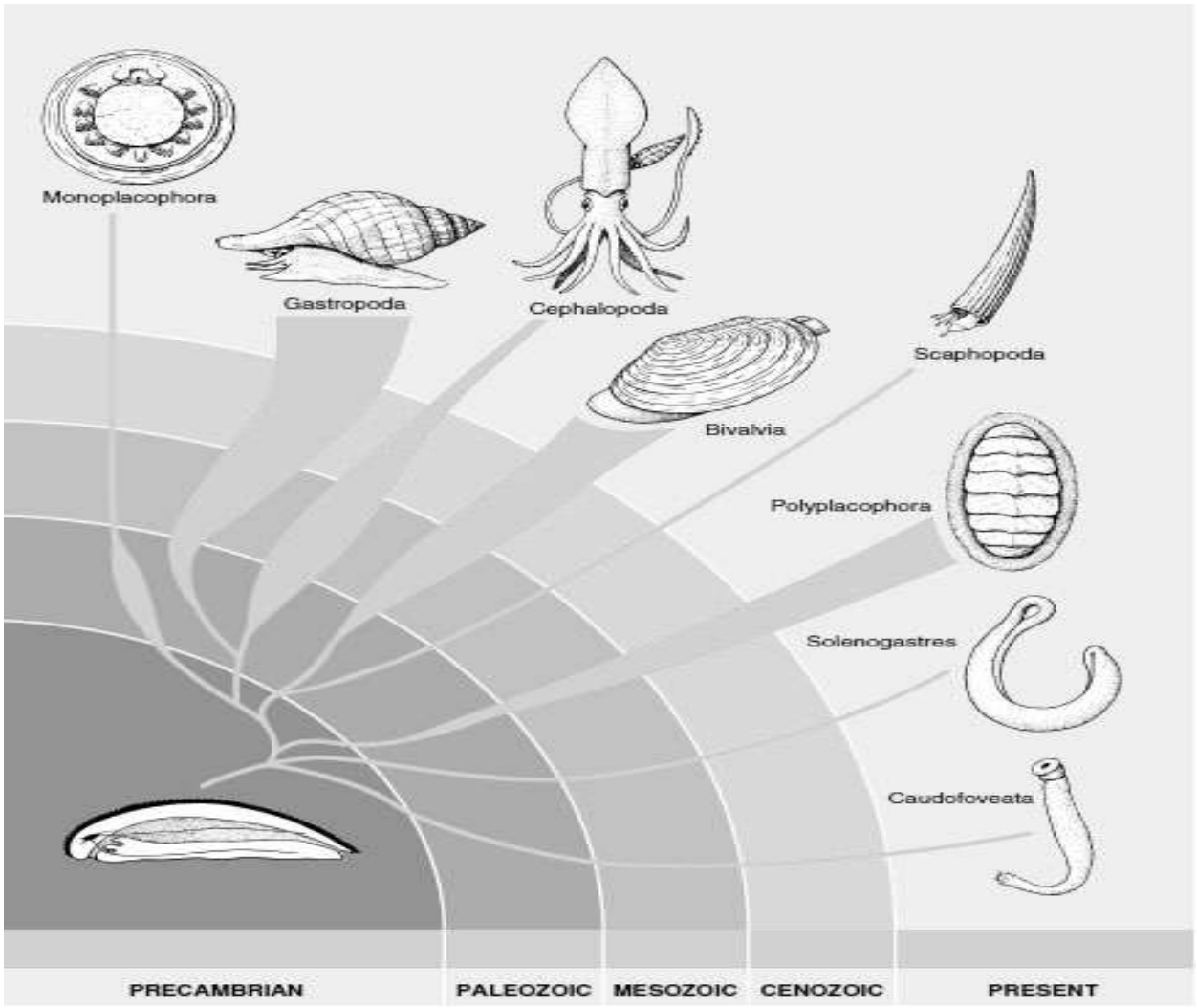


Figure 3. Προέλευση των Ομοταξιών των Μαλακίων (Hickman *et al.*, 11nd Edition, 2005).

2.2.5 Εξέλιξη της Ομοταξίας των Δίθυρων (Bivalvia)

Υποστηρίζεται πως η ομάδα Rostroconchia, η οποία είναι μια εξαφανισμένη ομάδα των Μαλακίων, είναι ο κοντινότερος συγγενής των σημερινών Δίθυρων (Waller, 1998). Στην πρώιμη μορφή τους, αποτελούνταν από ένα όστρακο που έμοιαζε με μονή βαλβίδα, το οποίο μετετράπη σε ένα όστρακο με δύο θυρίδες ανάμεσα στις οποίες υπήρχε κενό, ενώ δεν είχαν καθόλου οδοντοφόρους μυς. Επίσης, οι ομάδες των Δίθυρων μελετώνται και όσον αφορά τη μονοφυλετικότητα, την παραφυλετικότητα και την ομαδοποίησή τους. Μια κατάταξη των ομάδων είναι όπως (Figure **4**), καθώς και στα παρακάτω δέντρα που ακολουθούν στην (Figure **5**) και παρουσιάζουν δεδομένα από διαφορετικές εργασίες.

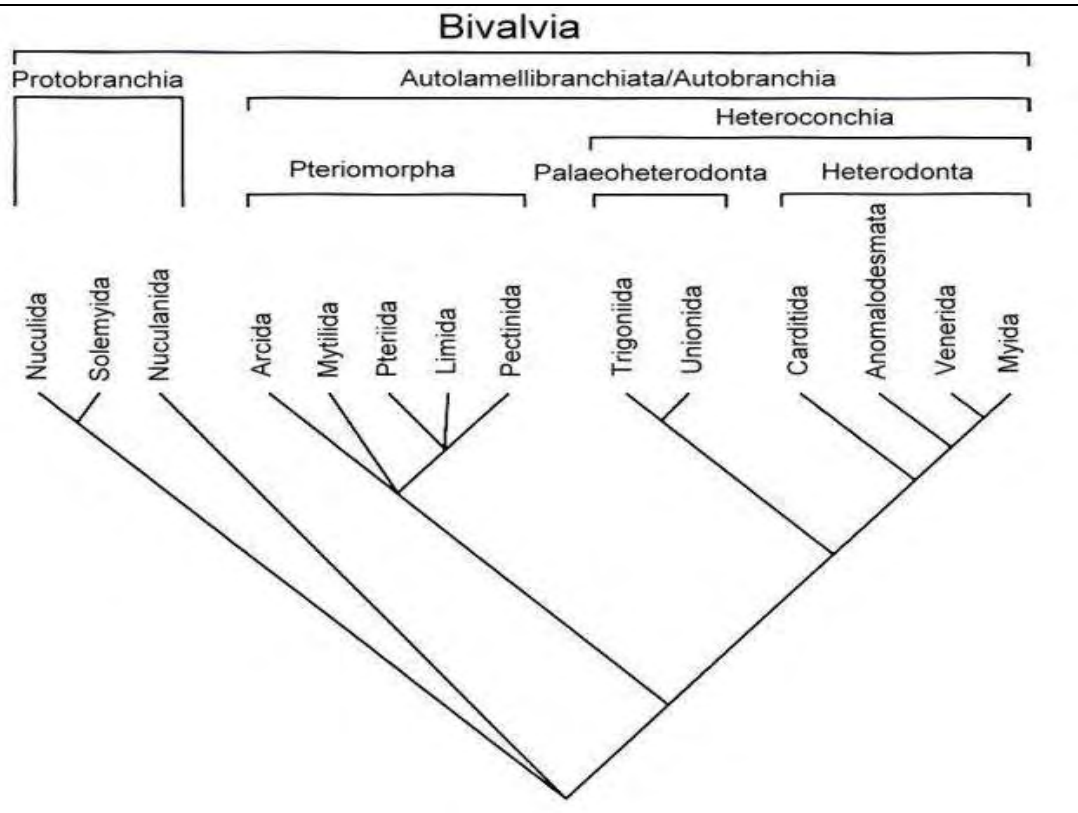


Figure 4. Κλαδόγραμμα της εξέλιξης των ομάδων των Δίθυρων (Bieler, 2007)

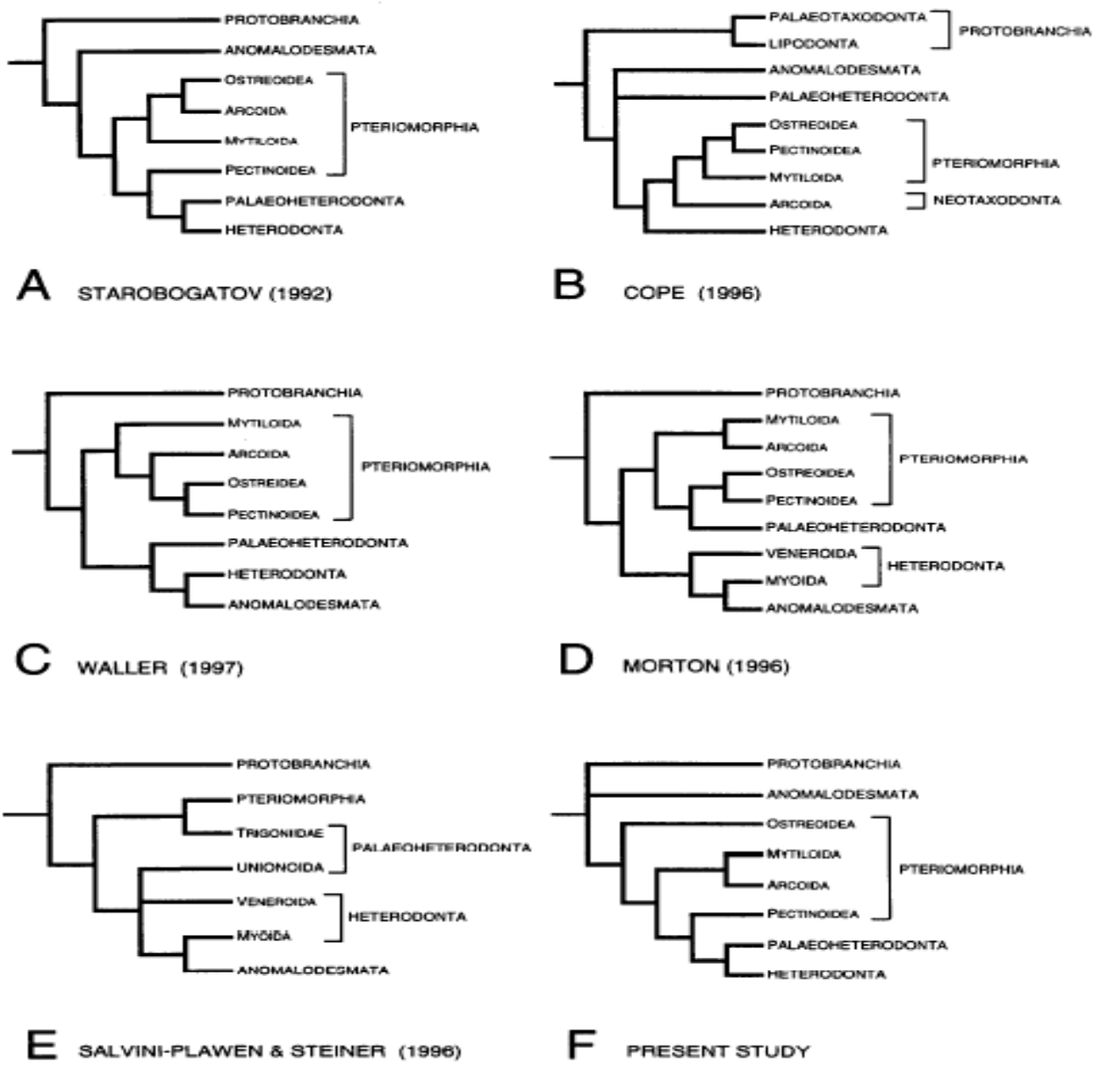


Figure 5. Υποθετικές φυλογενήσεις των ομάδων των Δίθυρων (A-F) με βάση μορφολογικά χαρακτηριστικά τους (Adamkewicz *et al.*, 1997)

2.3 Φυλογενετική και Φυλογενωμική

2.3.1 Ο κλάδος της Φυλογένεσης και η ανάπτυξη του

Από τα αρχαία χρόνια είχαν γίνει πολλές προσπάθειες για την ομαδοποίηση των οργανισμών με βάση κοινά γνωρίσματα. Το πρώτο ολοκληρωμένο ταξινομικό σύστημα το ανέπτυξε ο Λινναίος (18ος αιώνας) ενώ στη συνέχεια, με βάση την

Θεωρία της Εξέλιξης των ειδών του Δαρβίνου η ταξινόμιση άρχισε να γίνεται με βάση την κοινή προέλευση των οργανισμών και όχι τα μορφολογικά τους γνωρίσματα, τα οποία μπορεί να είναι αποτέλεσμα συγκλίνουσας εξέλιξης. Με τον όρο συγκλίνουσα εξέλιξη αναφερόμαστε στο φαινόμενο κατά το οποίο δύο είδη μοιάζουν μεταξύ τους όχι λόγω κοινής καταγωγής, αλλά λόγω προσαρμογής σε όμοιες οικολογικές συνθήκες. Η συγκλίνουσα εξέλιξη μπορεί να οδηγήσει και στο φαινόμενο της ομοπλασίας όπου ένας χαρακτήρας είναι κοινός σε δύο ή περισσότερα είδη, αλλά δεν υπάρχει στον κοινό τους πρόγονο. Ένα χαρακτηριστικό παράδειγμα είναι τα φτερά στο περιστέρι (πτηνό) και την νυχτερίδα (θηλαστικό) τα οποία δεν υπήρχαν στον κοινό τους πρόγονο, αλλά εμφανίστηκαν και στα δύο ανεξάρτητα και σε διαφορετικούς χρόνους (Figure 6). Στην φυλογένεση των οργανισμών, όταν χρησιμοποιούμε φαινοτυπικούς χαρακτήρες υπάρχει πάντα ο κίνδυνος αυτοί να είναι προϊόν ομοπλασίας και όχι ομολογίας και επομένως να οδηγήσει σε λάθη. Αντιθέτως, είναι πολύ δύσκολο έως αδύνατο να συναντήσουμε ομοιότητες σε μοριακές ακολουθίες που να οφείλονται σε ομοπλασία. Για αυτό το λόγο πλέον χρησιμοποιούνται οι μοριακές ακολουθίες όταν θέλουμε να κάνουμε φυλογένεση οργανισμών.

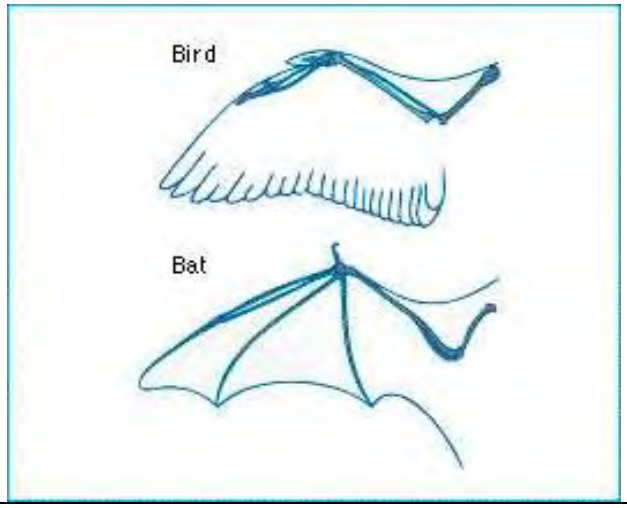


Figure 6. Φαινόμενο ομοπλασίας στα πτηνά και τις νυχτερίδες.

Με βάση της φυλογενετικές αναλύσεις έχει δημιουργηθεί το δέντρο της ζωής, που

απεικονίζει τις εξελικτικές σχέσεις μεταξύ των οργανισμών (Figure **7**), με ένα κοινό πρόγονο που εμφανίστηκε πριν από 3,7 δισεκατομμύρια χρόνια πριν.

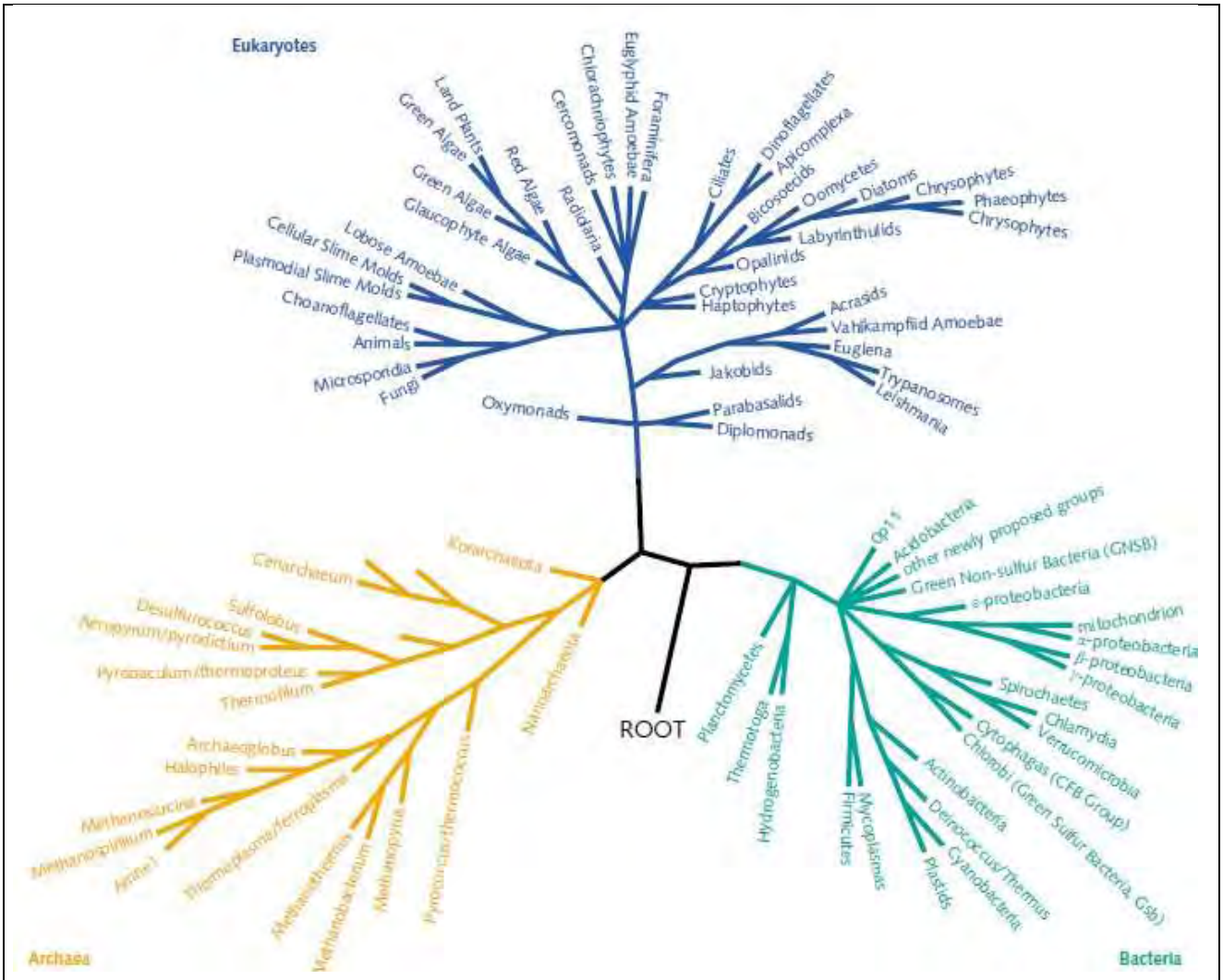


Figure 7. Το δέντρο της ζωής και οι μεγάλες ομάδες των οργανισμών, όπως είχε προκύψει από τις μελέτες και αναλύσεις που είχαν γίνει μέχρι το 2008 (*©Leonard Eisenberg, 2008).*

2.3.2 Φυλογενετικά δέντρα

Τα φυλογενετικά δέντρα είναι μια γραφική αναπαράσταση των φυλογενετικών σχέσεων μιας ομάδας από οργανισμούς ή γονίδια και μπορεί να είναι με ρίζα (rooted) όπου εμφανίζεται ο κοινός πρόγονος ή χωρίς ρίζα (unrooted) όπου δεν εμφανίζεται ο κοινός πρόγονος (Figure **8**).

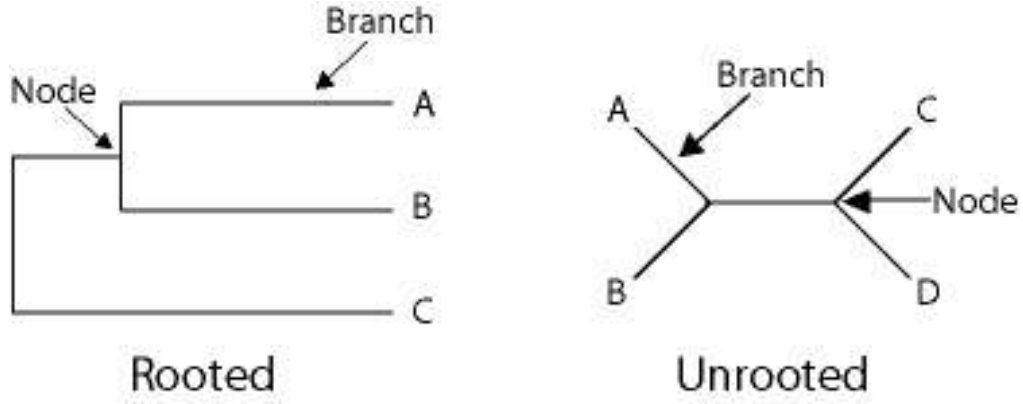


Figure 8. Παρουσίαση ενός δέντρου με ρίζα (rooted) και ενός άρριζου δέντρου (unrooted). Σημειώνονται οι κόμβοι (nodes) και οι κλάδοι (branches).

Η φυλογένεση μπορεί να γίνει με βάσει 4 διαφορετικά είδη δεδομένων:

- Μορφολογικών χαρακτηριστικών
- Χρωμοσωμικών χαρακτηριστικών
- Μοριακών ή Βιοχημικών χαρακτήρων
- Ηθολογικών ή Οικολογικών χαρακτηριστικών

Αρχικά, για τη συστηματική κατάταξη των οργανισμών οι ερευνητικές ομάδες χρησιμοποιούσαν ως κριτήριο κυρίως, τα μορφολογικά χαρακτηριστικά κάθε οργανισμού που μελετούσαν και σε ανατομική σύγκριση με άλλους παρόμοιους οργανισμούς μπορούσαν να εξάγουν συμπεράσματα για την ταξινόμηση του κάθε είδους. Δηλαδή, τα συμπεράσματα προέκυπταν από παρατηρήσεις στον φαινότυπο, γεγονός το οποίο μπορούσε να οδηγήσει σε λανθασμένα συμπεράσματα λόγω του φαινομένου της ομοπλασίας. Γι αυτό το λόγο η επιστημονική κοινότητα στράφηκε προς τα μοριακά δεδομένα, δηλαδή την κατασκευή φυλογενετικών δέντρων βασιζόμενοι σε μοριακές πληροφορίες που είχαν συλλέξει και επεξεργαστεί. Αρχικά, όταν ήταν ακόμα δύσκολο να ακριβό να συλλεχθούν μοριακές ακολουθίες, οι περισσότερες μελέτες στηρίζονταν στη μελέτη ενός κοινού γονιδίου (ορθόλογου) για τους υπό μελέτη οργανισμούς.

Βέβαια, όπως και σε κάθε μεθοδολογία έτσι και στην μοριακή φλογένεση υπάρχουν κάποια μειονεκτήματα που μπορεί να οδηγήσουν σε εσφαλμένα συμπεράσματα.Μερικά από αυτά είναι:

1. Η στοίχιση των νουκλεοτιδίων είναι δύσκολη, λόγω μεγάλης απόκκλισης σε κάποιες περιοχές.
2. Τα είδη που μελετώνται μπορεί να εξελίσσονται με διαφορετικούς ρυθμούς και να προκύψουν προβλήματα στον υπολογισμό της γενετικής απόστασης.
3. Η σύγχιση μεταξύ ορθόλογων (ειδογένεση) και παράλογων (διπλασιασμός) γονιδίων εκ των οποίων τα τελευταία θα πρέπει να αναγνωρίζονται και να αποκλείονται από τη μελέτη.

Τα προβλήματα αυτά, με την πρόοδο της τεχνολογίας και την συνεχή ανάπτυξη νέων αλγόριθμων με μικρό ποσοστό σφαλμάτων, μειώνονται. Με την ανάπτυξη της τεχνολογίας αλληλούχισης ολόκληρων των γονιδιωμάτων όμως, οι μελέτες έχουν αρχίσει να γίνονται σε επίπεδο ολόκληρου γονιδιώματος και όχι απλώς ενός ή μερικών γονιδίων. Επομένως, ο όρος Φυλογενετική (Phylogenetics) έχει αρχίσει να εξελίσσεται προς τον όρο Φυλογενωμική (Phylogenomics), (Comas *et al.*, 2007).

Η Φυλογενωμική είναι το κομμάτι της φυλογενετικής ανάλυσης το οποίο κάνει φυλογένεση με βάσει όχι ένα, αλλά πάρα πολλά γονίδια ενός οργανισμού.Η ανάπτυξη αυτού του τομέα έγινε με σκοπό:

- τον εντοπισμό ορθόλογων γονιδίων σε αλληλουχούμενα γονιδιώματα
- την διαλεύκανση και τον καθορισμό είτε κοντινών είτε βαθιών εξελικτικών σχέσεων μεταξύ ειδών
- τον εντοπισμό της οριζόντιας μεταφοράς γονιδίων, ένα μειονέκτημα που η φυλογενετική δυσκολεύεται να το ξεπεράσει.

Ένα παράδειγμα όπου με τη βοήθεια της γονιδιωματικής και της φυλογενωμικής βρέθηκε η αλήθεια είναι αυτό όπου πρόσφατα εντοπίστηκε η αιτία μιας επιδημίας χολέρας στην Αϊτή.

2.4 Τεχνολογίες Αλληλούχισης Νέας Γενιάς (Next Generation Sequencing, NGS)

Ο βασικός λόγος, όπου πλέον η φυλογενετική εξελίσσεται σε φυλογενωμική, με πιο περίπλοκες βιοπληροφορικές αναλύσεις αλλά και πιο αξιόπιστα αποτελέσματα είναι η ραγδαία ανάπτυξη των τεχνολογιών αλληλούχισης νέας γενιάς που συνέβη την τελευταία δεκαετία.

Με την προσπάθεια για την αλληλούχιση του Ανθρώπινου Γονιδιώματος το οποίο ξεκίνησε το 1990, κατέστη αναγκαία η εξέλιξη των τεχνολογιών αλληλούχισης, όπου μέχρι τότε η χρονοβόρα/κοστοβόρα μέθοδος κατά Sanger (197) ήταν αυτή που χρησιμοποιούσαν ευρέως. Με την πρόοδο της τεχνολογίας άρχισαν να αναπτύσσονται τεχνολογίες δεύτερης γενιάς ή αλλιώς νέες τεχνολογίες αλληλούχισης, αντικαθιστώντας σε μεγάλο ποσοστό τις παλαιότερες μεθόδους λόγω, εξοικονόμησης χρόνου (ταυτόχρονη αλληλούχιση πολλών τμημάτων DNA ταυτόχρονα). Ιδιαίτερη έμφαση δόθηκε στη μείωση του κόστους, το οποίο φάνηκε αρχικά να ακολουθεί το νόμο του Moore, αλλά στη συνέχεια να ακολουθεί μια ακόμα πιο ραγδαία μείωση (Figure **9**). Κάποιοι πλέον μιλάνε για ένα νέο νόμο του Moore στις τεχνολογίες αλληλούχισης (Figure **10**).

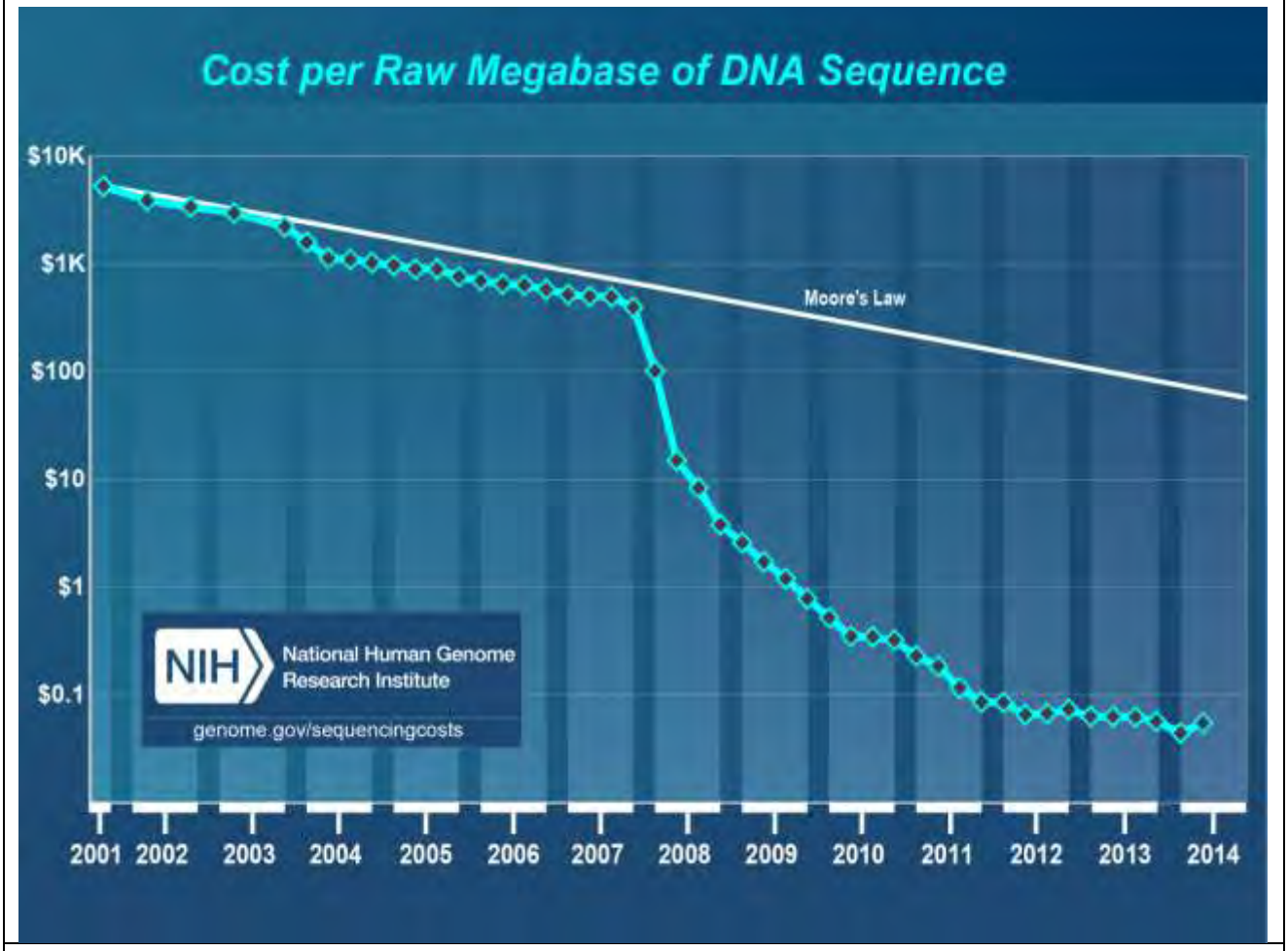


Figure 9. Η εξέλιξη του κόστους αλληλούχισης ανά Μεγαβάση (Megabase) το διάστημα 2001 με 2014 (http://www.genome.gov/sequencingcosts/).

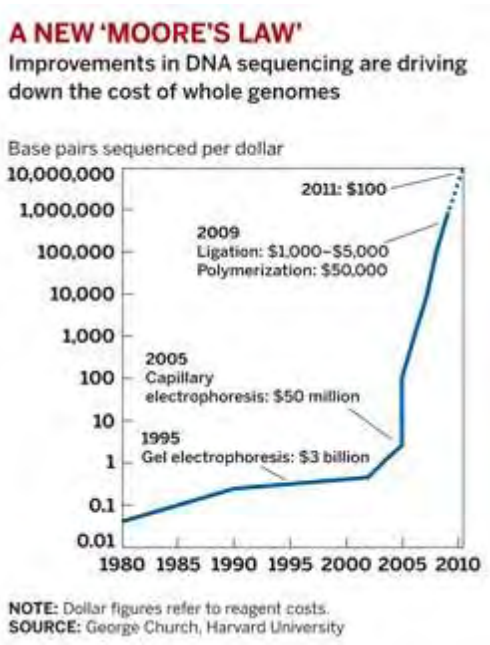
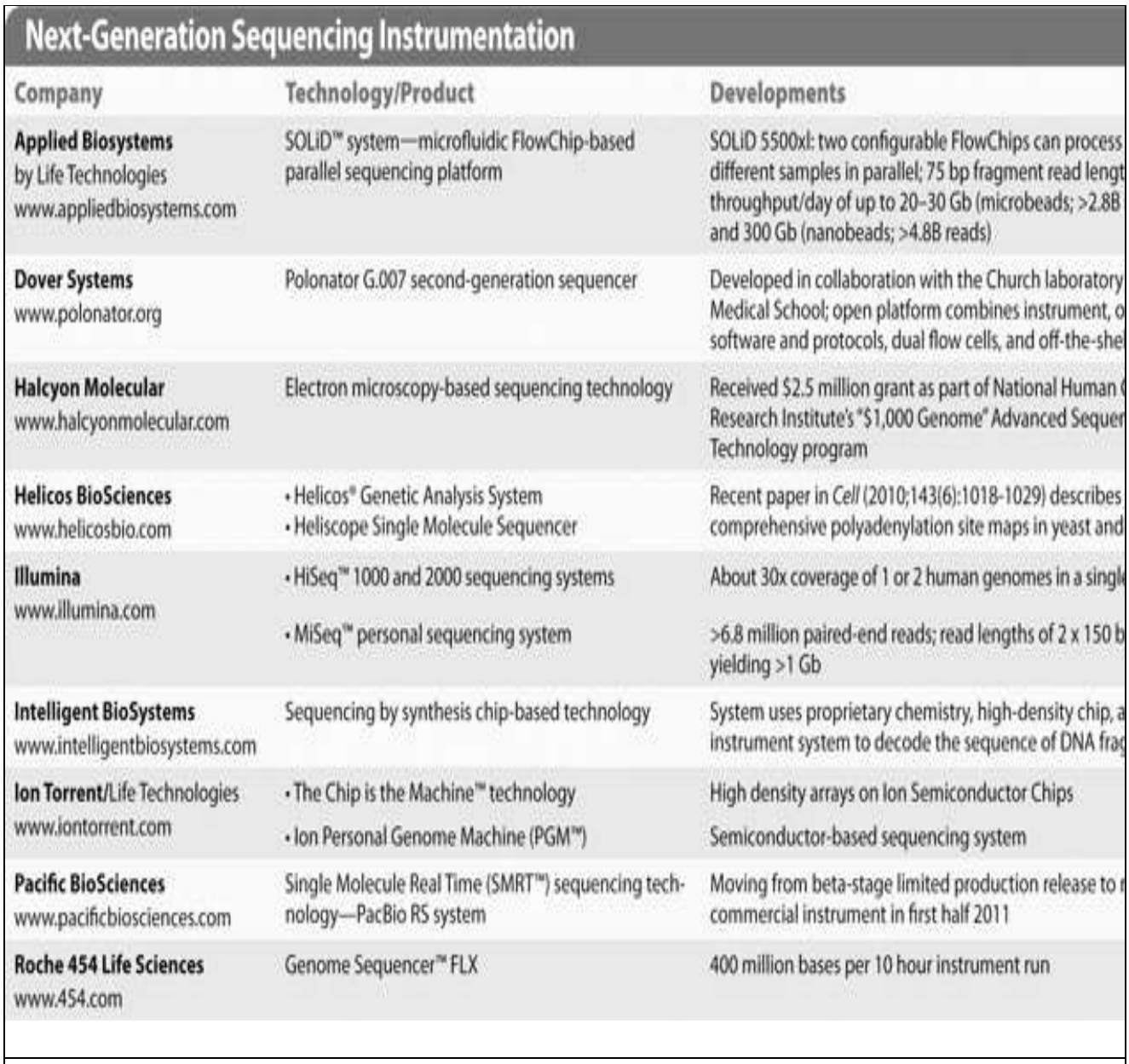


Figure 10. Ο 'νέος' νόμος του Moore, 2011 (http://cen.acs.org/articles/87/i50/DNA-Sequencing-Forges-Ahead.html).

Έτσι, δόθηκε λοιπόν, ιδιαίτερη έμφαση στην ανάπτυξη των νέων τεχνολογιών αλληλούχισης με αποτέλεσμα να υπάρχουν επιτυχημένες τωρινές τεχνολογίες δεύτερης γενιάς και ήδη οι εταιρείες αυτών των τεχνολογιών να έχουν φτάσει στης τρίτης γενιάς τεχνολογίες και να υπόσχονται ένα ακόμη πιο αποδοτικό μέλλον στην τεχνολογία της αλληλούχισης.

Στις τεχνολογίες πρώτης γενιάς ανήκει και η μέθοδος κατά Sanger, η οποία σε αντίθεση με τις τεχνολογίες νέας γενιάς απαιτεί μεγάλες ποσότητες DNA και δεν μπορεί να επεξεργαστεί πολλά δεδομένα ταυτόχρονα. Στις τεχνολογίες δεύτερης γενιάς ανήκουν κυρίως αυτές της Roche 454 Pyrosequencing, Solid, Illumina και Pacific Biosciences. Οι τεχνολογίες τρίτης γενιάς περικλύουν την τεχνολογία Ion torrent/Ion proton και Oxford nanopore. Στην Figure **11** εμφανίζονται οι σημαντικότερες εταιρείες και οι τεχνολογίες που έχουν αναπτύξει καθώς και τα πλεονεκτήματα κάθε μιας.



Next-Generation Sequencing Instrumentation

Company	Technology/Product	Developments
Applied Biosystems by Life Technologies www.appliedbiosystems.com	SOLiD™ system—microfluidic FlowChip-based parallel sequencing platform	SOLiD 5500xl: two configurable FlowChips can process different samples in parallel; 75 bp fragment read lengt throughput/day of up to 20–30 Gb (microbeads; >2.8B and 300 Gb (nanobeads; >4.8B reads)
Dover Systems www.polonator.org	Polonator G.007 second-generation sequencer	Developed in collaboration with the Church laboratory Medical School; open platform combines instrument, o software and protocols, dual flow cells, and off-the-she
Halcyon Molecular www.halcyonmolecular.com	Electron microscopy-based sequencing technology	Received $2.5 million grant as part of National Human Research Institute's "$1,000 Genome" Advanced Sequer Technology program
Helicos BioSciences www.helicosbio.com	• Helicos® Genetic Analysis System • Heliscope Single Molecule Sequencer	Recent paper in *Cell* (2010;143(6):1018-1029) describes comprehensive polyadenylation site maps in yeast and
Illumina www.illumina.com	• HiSeq™ 1000 and 2000 sequencing systems	About 30x coverage of 1 or 2 human genomes in a singl
	• MiSeq™ personal sequencing system	>6.8 million paired-end reads; read lengths of 2 x 150 b yielding >1 Gb
Intelligent BioSystems www.intelligentbiosystems.com	Sequencing by synthesis chip-based technology	System uses proprietary chemistry, high-density chip, a instrument system to decode the sequence of DNA fra
Ion Torrent/Life Technologies www.iontorrent.com	• The Chip is the Machine™ technology	High density arrays on Ion Semiconductor Chips
	• Ion Personal Genome Machine (PGM™)	Semiconductor-based sequencing system
Pacific BioSciences www.pacificbiosciences.com	Single Molecule Real Time (SMRT™) sequencing technology—PacBio RS system	Moving from beta-stage limited production release to commercial instrument in first half 2011
Roche 454 Life Sciences www.454.com	Genome Sequencer™ FLX	400 million bases per 10 hour instrument run

Figure 11. Οι εταιρείες και οι τεχνολογίες που έχουν παρουσιάσει καθώς και τα πλεονεκτήματα κάθε μιας (http://www.genengnews.com/gen-articles/range-of-ngs-applications-rises-quickly/3575/).

2.4.1 Η τεχνολογία της Illumina

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 (Figure **12**).

Η τεχνολογία της illumina έχει ομοιότητες 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. Επομένως, μπορούν να αναλυθούν πολλά δείγματα μαζί σε ένα μόνο τρέξιμο. Αυτό επίσης συνεπάγεται και μείωση του χρόνου/κόστους της αλληλούχισης (http://www.illumina.com).

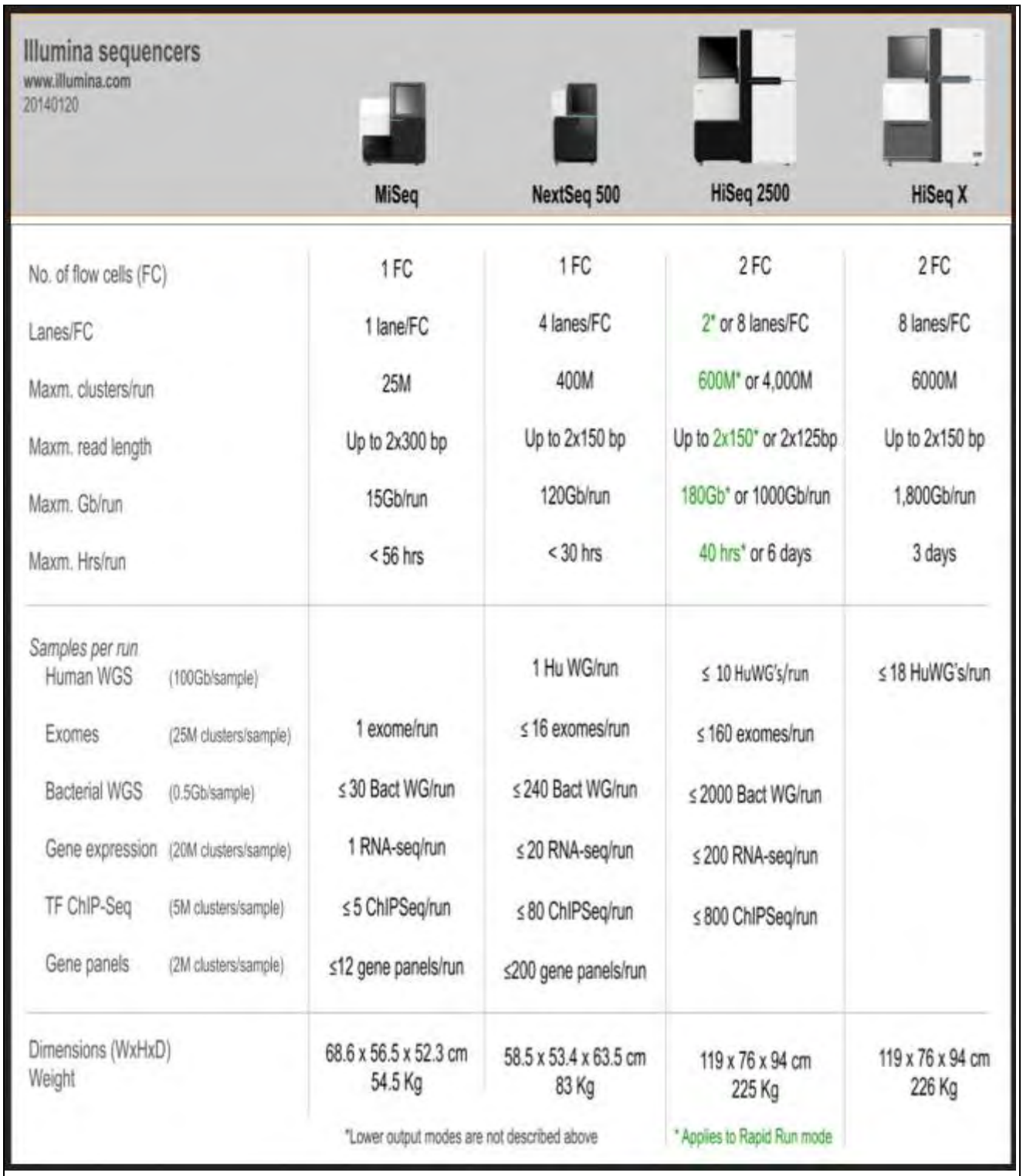


Figure 12. Μηχανήματα αλληλούχισης της Illumina (http://www.illumina.com/).

Το πρωτόκολλο που ακολουθείται από την Illumina για την αλληλούχιση των γονιδιωμάτων είναι το εξής:

- **<u>Προετοιμασία βιβλιοθήκης (Library Preparation)</u>**

Αρχικά, με τη χρήση μη ειδικών περιοριστικών ενζύμων γίνεται τεμαχισμός του δίκλωνου DNA των δειγμάτων, ώστε να δημιουργηθούν τυχαία κομμάτια γενετικού υλικού. Έπειτα, συμβαίνει πρόσδεση ενός ολιγονουκλεοτιδίου θυμιδίνης, όπου προσδένεται στα θραύσματα με το ίδιο να προεξέχει (Ansorge, 2009). Στην συνέχεια συνδέονται και στα δύο άκρα των θραυσμάτων του DNA οι λεγόμενοι αντάπτορες (adapters), οι οποίοι έχουν συγκεκριμένα αλλά διαφορετικά barcodes για το κάθε δείγμα (Figure **13**).Τα barcodes είναι μεμονωμένες αλληλουχίες οι οποίες προστίθενται στα δείγματα για να μπορεί να γίνει, κατά την ανάλυση των δεδομένων, ταυτοποίηση του θραύσματος με το δείγμα στο οποίο ανήκει. Μετά την σύνδεση των ανταπτόρων με τα θραύσματα του DNA το γενετικό υλικό ηλεκτροφορείται και ύστερα γίνεται επιλογή κομματιών συγκεκριμένων μεγεθών, περίπου 200-300 bp. Κατόπιν, με την τεχνική της PCR ενσωματώνονται στα επεξεργασμένα κομμάτια τα άκρα P5 και P7, για να μπορεί να συμβεί αποδιάταξη των δίκλωνων μορίων σε μονόκλωνα (Figure **14**).

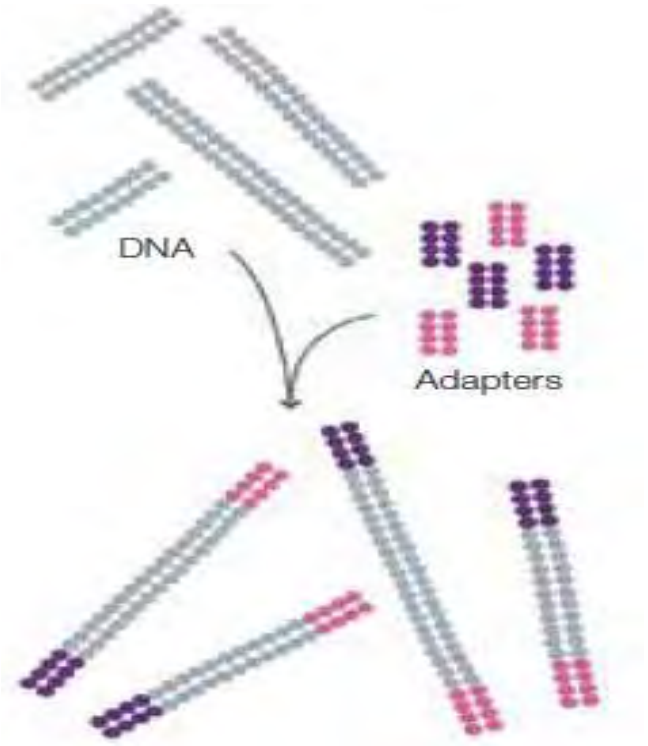


Figure 13. Κατακερματισμένο DNA τυχαία και η ενσωμάτωση των ανταπτόρων στα άκρα (http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf).

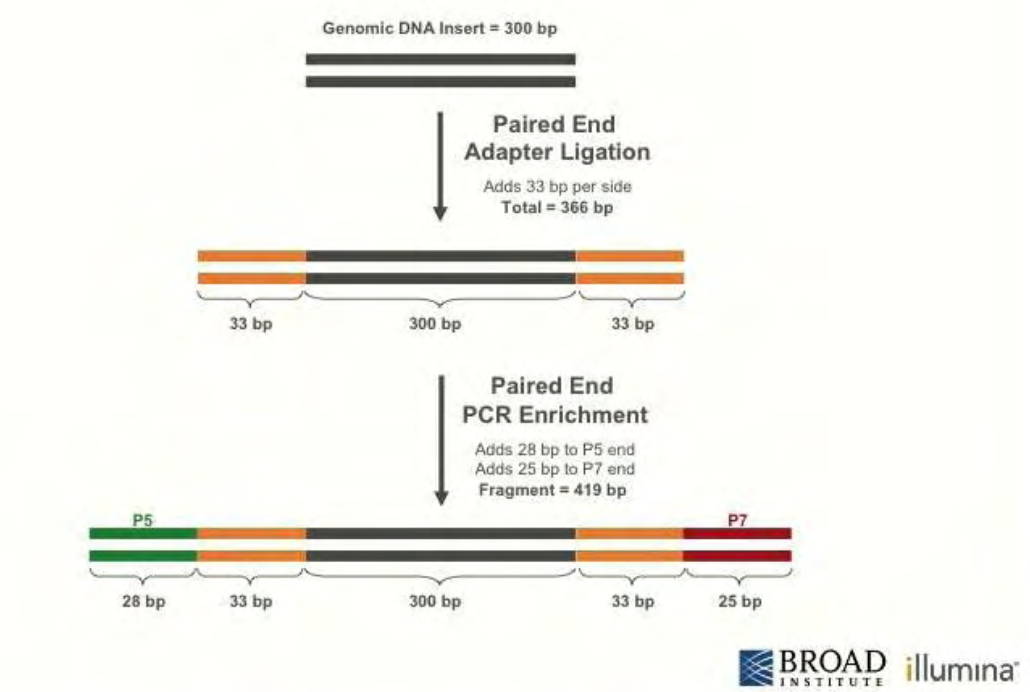


Figure 14. Σύνδεση ανταπτόρων και ενσωμάτωση P5 και P7 άκρων (http://openwetware.org/wiki/Image:Libprep3.jpg).

- **<u>Δημιουργία συμπλέγματος-cluster (Cluster Generation)</u>**

Η διαδικασία συνεχίζεται με την τοποθέτηση των μονόκλωνων μορίων επάνω σε μια επιφάνεια-πλάκα εργασίας (workflow-glass flow cell) (Figure **15**).

Figure 15. Πολλά δείγματα μπορούν να φορτώνονται στις οκτώ λωρίδες για ταυτόχρονη ανάλυση σε ένα σύστημα Illumina Sequencing, (http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.

pdf).

Η κάθε πλάκα αποτελείται εσωτερικά από ολιγονουκλεοτίδια τα οποία είναι συμπληρωματικά ως προς τους αντάπτορες και χωρίζεται σε οχτώ ξεχωριστές λωρίδες. Ακολουθεί τυχαίος υβριδισμός (μέσω εναλλαγής υψηλής με χαμηλή θερμοκρασία) μεταξύ των ολιγονουκλεοτιδίων της πλάκας με τους αντάπτορες του ενός άκρου των μονόκλωνων θραυσμάτων DNA. Όπως φαίνεται στην Figure **16**, οι ελεύθεροι αντάπτορες των μονόκλωνων μορίων υβριδίζονται με τα ολιγονουκλεοτίδια της πλάκας δημιουργώντας γέφυρες (bridge amplification), (Mardis, 2011). Πρέπει να σημειωθεί πως μία ισοθερμική πολυμεράση ενισχύει για την δημιουργία κλώνων καθώς και το ότι οι αντάπτορες της πλάκας δρουν ως εκκινητές για την ενίσχυση (Zhou *et al.*, 2010). Αποτέλεσμα της παραπάνω διαδικασίας είναι ότι κάθε βιβλιοθήκη θραυσμάτων αποτελείται πλέον από εκατοντάδες εκατομμύρια μοναδικά συμπλέγματα (clusters). Τα συμπληρωματικά συμπλέγματα αποκόπτονται και απομακρύνονται ξεπλένοντας. Οι αντίστροφοι κλώνοι διασπώνται και ξεπλένονται, ενώ τα άκρα μπλοκάρονται και ο εκκινητής αλληλούχισης υβριδοποιείται με τα DNA.Έτσι, μετά τη δημιουργία συμπλέγματος, οι βιβλιοθήκες είναι έτοιμες για αλληλούχιση.

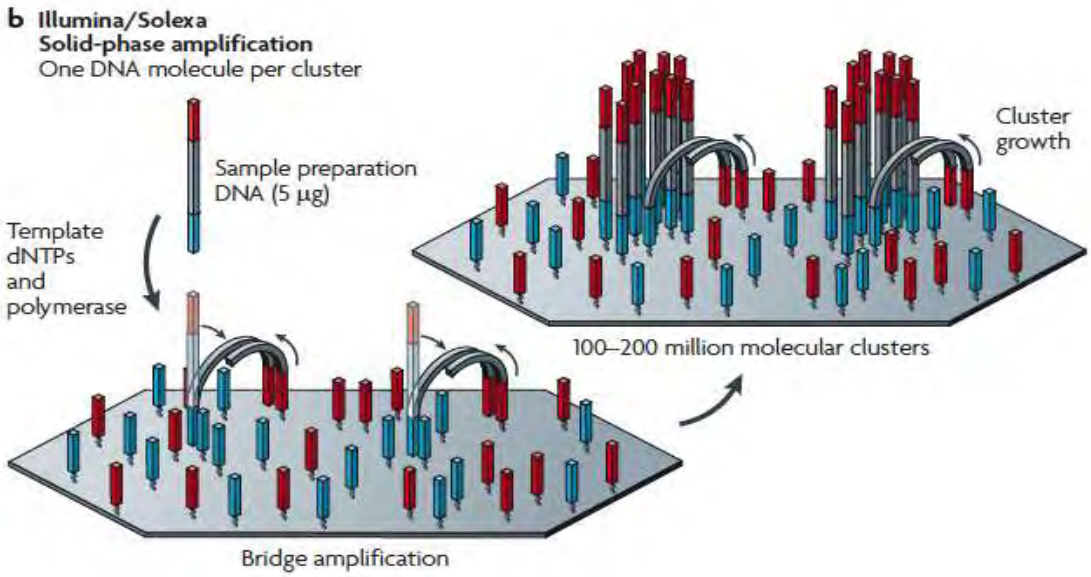


Figure 16. Δημιουργία γεφυρών και ομαδοποίηση (Metzker, 2010).

- **<u>Αλληλούχιση (Sequencing)</u>**

Στο τελευταίο βήμα πραγματοποιείται η αλληλούχιση όλων των clusters που 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) (Figure **17**). Οι τέσσερις φθορίζουσες με τις βάσεις πλησιάζουν την βάση του cluster αλλά μόνο μία θα ενωθεί. Και οι τέσσερις βάσεις ανταγωνίζονται μεταξύ τους για να συνδεθούν με το εκμαγείο. Αυτός ο ανταγωνισμός εξασφαλίζει την υψηλότερη δυνατή ακρίβεια. Μόλις το λέιζερ – CCD camera ανιχνεύσει τη συμπληρωματική βάση (από το χρώμα που εκπέμπει), η φθορίζουσα χρωστική αφαιρείται και μένει η βάση μετά από ξέπλυμα. Το ίδιο γίνεται και για την επόμενη βάση της αλυσίδας του cluster μέχρι να τερματιστεί. Έτσι δημιουργούνται συμπληρωματικές αλυσίδες των clusters.

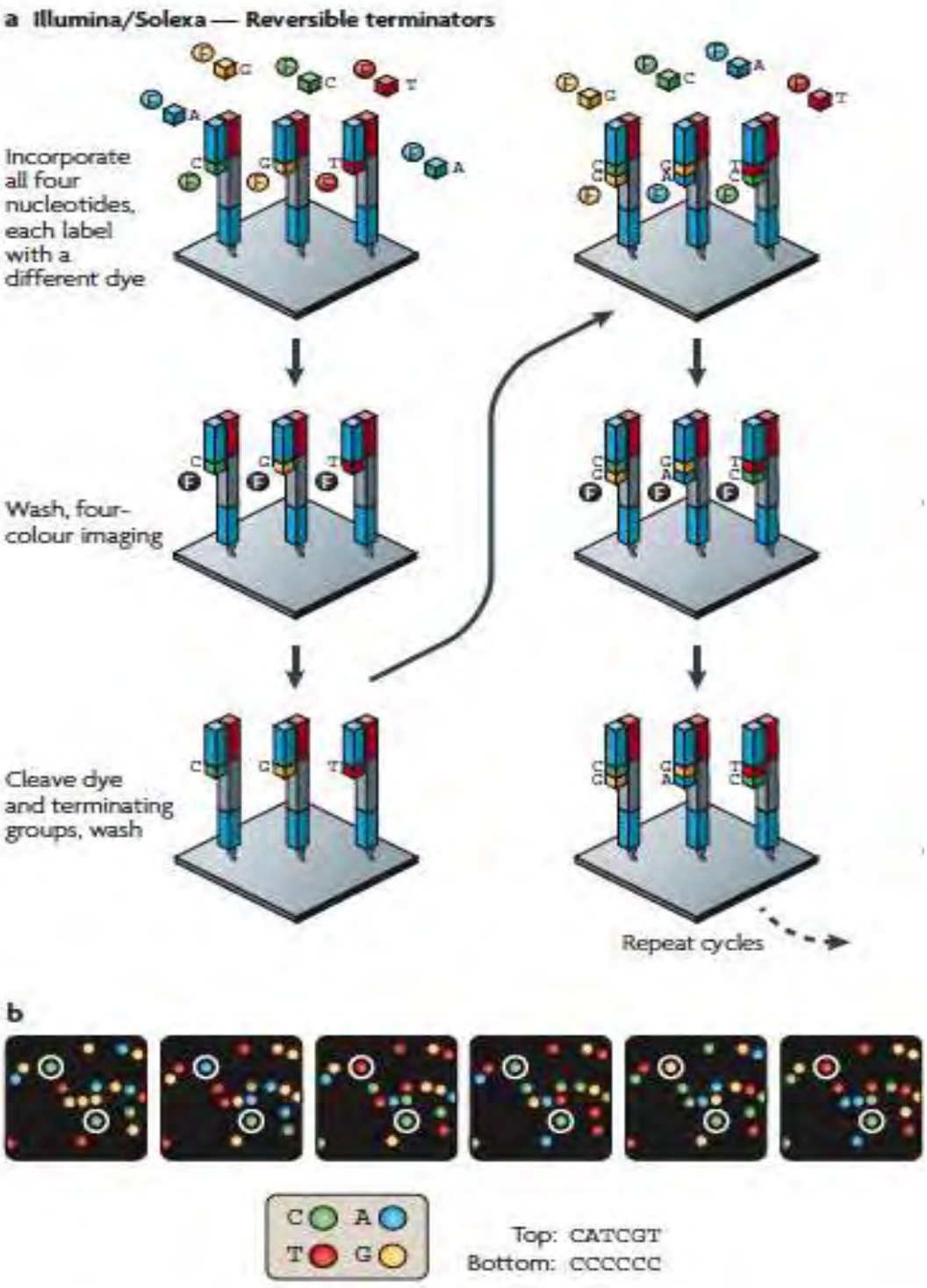


Figure 17. a. Ενσωμάτωση των βασεων, έκπλυση και δημιουργία αλυσίδας. b. Οι εικόνες των τεσσάρων χρωμάτων επισημαίνουν τα δεδομένα αλληλούχισης από δύο κλωνικά ενισχυμένα πρότυπα (Metzker, 2010).

Η τεχνολογία αλληλούχισης νέας γενιάς, όπως αναφέρθηκε παραπάνω κερδίζει

συνεχώς έδαφος λόγω της αξιοπιστίας της και της συνεχόμενης μείωσης του κόστους. Αυτό έχει σας συνέπεια την αλληλούχιση του ολικού RNA ή ολικού mRNA οργανισμών, το οποίο φέρει πολλά πλεονεκτήματα, με σκοπό τον εντοπισμό γονιδίων που εκφράζονται (expression profiling), τη μελέτη δηλαδή σε επίπεδο μεταγραφώματος. Οι κυριότερες πλατφόρμες που χρησιμοποιούνται για αλληλούχιση του RNA είναι:

- 454™ Titanium
- 454 GS-FLX+
- Illumina HiSeq™ 2000 (Official Service Provider)
- Illumina MiSeq™
- SOLiD v4 (Official Service Provider)
- SOLiD 5500xl
- Ion Torrent PGM™

Ένα από τα πλεονεκτήματα της αλληλούχισης του RNA, κυρίως με την τεχνολογία της Illumina, είναι ότι δίνεται η δυνατότητα στους ερευνητές να βρούν πόσα reads αντιστοιχούν σε κάθε mRNA, το οποίο είναι ανάλογο του επιπέδου έκφρασης του συγκεκριμένου μεταγράφου. Έτσι, επιτρέπεται ποσοτικοποίηση της γονιδιακής έκφρασης η οποία είναι συγκρίσιμη μεταξύ διαφορετικών γονιδιών, κάτι που δεν ισχύει για τα Microarreys. Η τεχνολογία RNA-Seq δίνει την ευκαιρία να μην χρειάζονται υποθέσεις για τον πειραματικό σχεδιασμό και μπορεί να επιτρέψει την έρευνα σε είδη δίχως άλλες πληροφορίες. Πέρα από την ανάλυση της έκφρασης των γονιδίων, η αλληλούχιση του RNA μπορεί να χρησιμοποιηθεί για την ανακάλυψη εναλλακτικού ματίσματος, έκφραση ειδικών αλληλομόρφων καθώς και σπάνια ή καινούργια μετάγραφα, (http://res.illumina.com/documents/products/datasheets/datasheet_rnaseq_analysis.pdf) (Table **2**).

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Table 2. Πλεονεκτήματα της μεθόδου RNAseq σε σχέση με άλλες μεθόδους αλληλούχισης μεταγράφων(transcriptomic) (Wang *et al.*,2009).

2.5 Βιοπληροφορική ανάλυση

2.5.1 Βάσεις Δεδομένων με δημοσιευμένα δεδομένα RNA-SEQ

Τα δεδομένα από τις διάφορες αλληλουχίσεις αποθηκεύονται συνήθως στη βάση δεδομένων του NCBI (National Center for Biotechnology Information) και πιο συγκεκριμένα στη βάση δεδομένων SRA (Sequence Read Archive, http://www.ncbi.nlm.nih.gov/sra), αλλά και σε άλλες βάσεις δεδομένων όπως αυτή της Ensembl, που αφορά κυρίως ολόκληρα γονιδιώματα (http://www.ensembl.org/index.html). Τα δεδομένα που αποθηκεύονται στη SRA συνοδεύονται από πληροφορίες που αφορούν τη μελέτη, πληροφορίες του βιολογικού δείγματος, τα πειραματικά δεδομένα, το είδος της πλατφόρμας, την ανάλυση αλλά και το χρόνο υποβολής των δεδομένων (Kodama *et al.*, 2012).

2.5.2 Μορφή, φίλτρα και έλεγχος ποιότητας της αλληλούχισης

Μορφή

- 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 (Figure **18**). 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 το ένα έχει τα αλληλουχημένα reads προς τα εμπρός και το δεύτερο τα αλληλουχημένα reads προς την αντίστροφη πλευρά.

- Single – end data

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

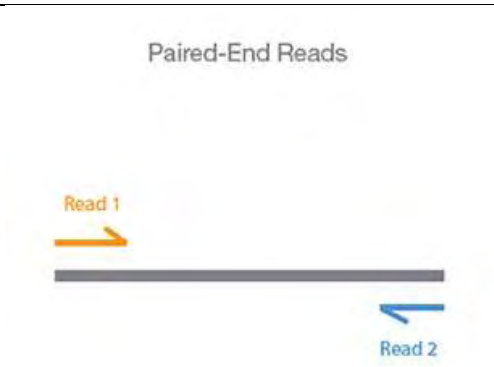


Figure 18. Paired-end reads (http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html)

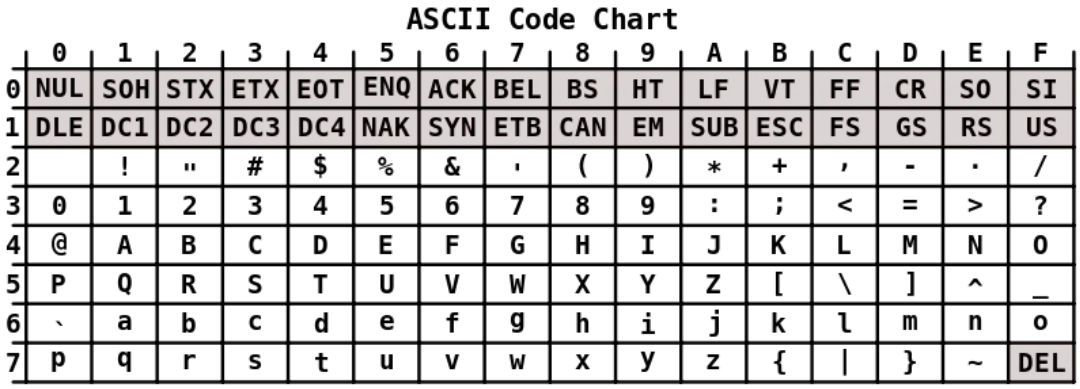
Τα δεδομένα αλληλούχισης αποθηκεύονταιι σε μορφή FastQ και κάθε τέτοιο αρχείο περιλαμβάνει την ακολουθία καθώς επίσης και ένα σκορ για την βεβαιότητα με την οποία εντοπίστηκε σωστά η κάθε συγκεκριμένη βάση (Figure **19**). Ένα αρχείο FASTQ έχει την παρακάτω δομή:

1. Το αναγνωριστικό της αλληλουχίας το οποίο πάντα ξεκινάει με το σύμβολο ΄@΄ και πρέπει να έχει αυτήν την ακολουθία: @<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos> <read>:<is filtered>:<control number>:<index sequence>
2. Η αλληλουχία, π.χ AAAAUGCUACGACACAGACGCAUAGCACGU
3. Το σύμβολο '+΄ το οποίο είναι αναγνωριστικό για το σκορ της ποιότητας(Quality score identifier)
4. Σκόρ ποιότητας, δηλαδή η ποιότητα της κάθε βάσης στη συγκεκριμένη αλληλουχία σε μορφή του ASCII (American Standard Code for Information Interchange)

```
#
1    @EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG
2    AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
3    +
4    BBBBCCCC?<A?BC?7@@???????DBBA@@@@A@@
```

Figure 19. Παράδειγμα ενός fastq αρχείου.

Το ASCII (American Standard Code for Information Interchange) είναι μία μορφή κωδικοποίησης κειμένου με την μορφή χαρακτήρων της αγγλικής αλφαβήτου (Figure **20**).



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

Figure 20. Σύστημα κωδικοποίησης ASCII (http://en.wikipedia.org/wiki/ASCII).

Το σκορ ποιότητας της κάθε βάσης στην τέταρτη γραμμή σε ένα αρχείο FASTQ κωδικοποιείται με γράμματα του ASCII table. Αυτό το σκορ ποιότητας ονομάζεται Q-score και είναι ο αρνητικός λογάριθμος της πιθανότητας να έχει αναγνωστεί λάθος η συγκεκριμένη βάση, πολλαπλασιαζόμενο με το 10.

$$Q = -10 \log_{10} P$$

Η αξιολόγιση της ποιότητας της κάθε βάσης ενός sequence read είναι πολύ σημαντική καθώς υπάρχει συχνά το ενδεχόμενο λάθους ανάγνωσης μίας ή περισσότερων βάσεων εξαιτίας συστηματικού λάθους, που μπορεί να οφείλεται σε

διάφορους λόγους. Συγκεκριμένα, η τεχνολογία της Illumina εμφανίζει συστηματικά λάθη, ιδιαίτερα προς το τέλος του sequence read (Figure **21**).

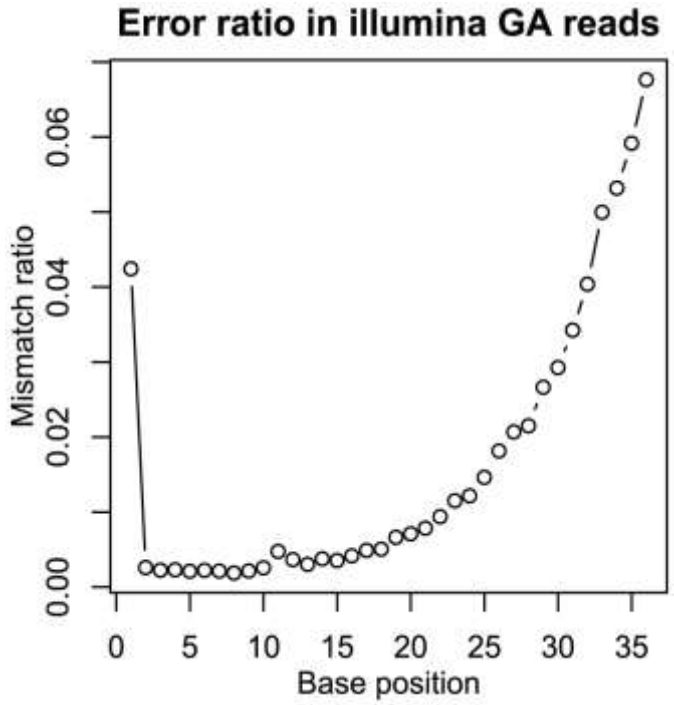


Figure 21. Κατανομή λάθους ανάγνωσης βάσεων σε Illumina reads. Το πρόβλημα εντοπίζεται στη συσσώρευση λαθών κατά την ενσωμάτωση φθοριζόντων dNTPs. Ο βαθμός λάθους στα GA reads εξαρτάται από τη θέση κάθε βάσης στο read. Ο βαθμός του mismatch ανάμεσα στα mapped reads και στην αλληλουχία αναφοράς σε σχέση με τον ολικό αριθμό των mapped reads απεικονίζεται σε διάγραμμα έναντι της θέσης της κάθε βάσης στα reads. Ο βαθμός του mismatch αυξάνεται μαζί με τη θέση της βάσης υποδεικνύοντας τη μείωση της ακρίβειας των base calls (http://openi.nlm.nih.gov/detailedresult.php?img=3096631_pone.0019534.g001&req=4).

Η ποιότητα των αλληλουχήσεων ελέγχεται με διάφορα προγράμματα βιοπληροφορικής, όπως το FastQC (Figure **22**, Figure **23**, Figure **24**) το οποίο στηρίζεται στη γλώσσα προγραμματισμού java και μπορεί να επεξεργαστεί τα δεδομένα και να παρουσιάσει διάφορες παραμέτρους της ποιότητας τους όπως:

- ποιότητα των βάσεων της κάθε αλληλούχισης
- ποσοστό περιεκτικότητας σε GC
- ποσότητα αδιάβαστων βάσεων

- ποσοστό διπλασιασμένων αλληλουχιών
- κατανομή μήκους των αλληλουχιών

FastQC Report

Fri 30 May 2014
good_sequence_short.txt

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content

Produced by FastQC (version 0.11.1)

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Figure 22. Παρουσίαση των βασικών στοιχείων μιας καλής αλληλούχισης με το πρόγραμμα FastQC (http://www.bioinformatics.babraham.ac.uk).

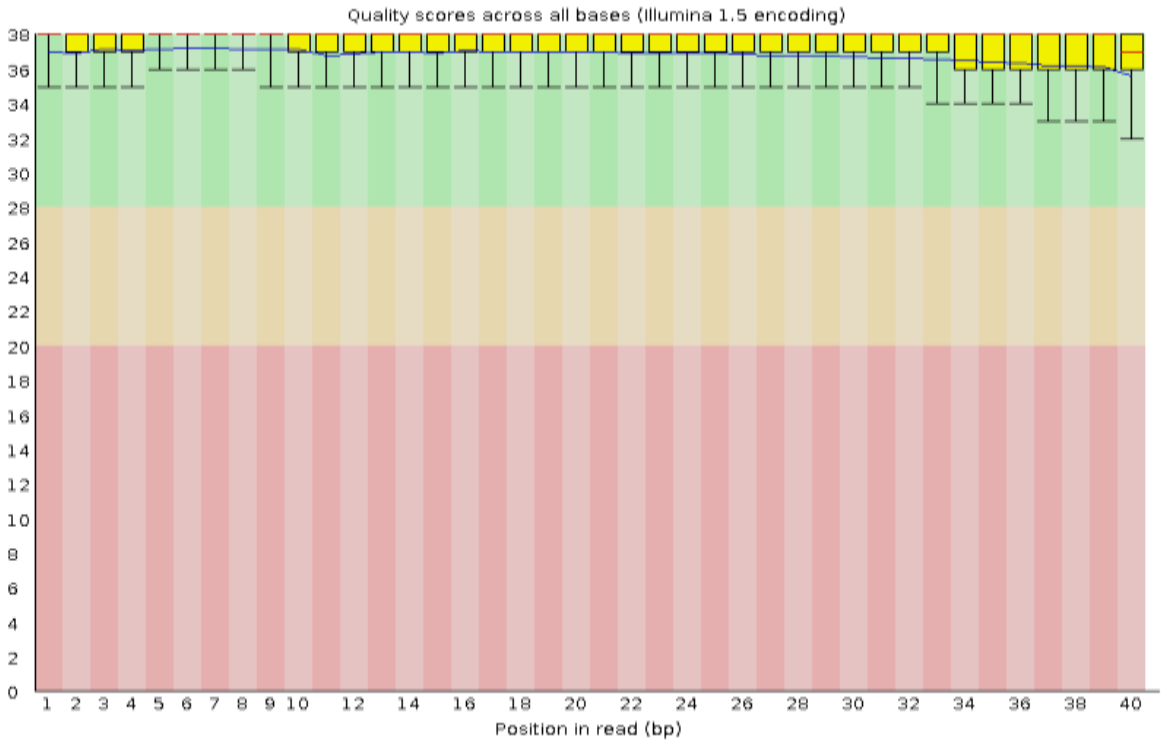


Figure 23. Διαγραμματική απεικόνιση υψηλής ποιότητας αλληλούχισης με το πρόγραμμα FastQC (http://www.bioinformatics.babraham.ac.uk).

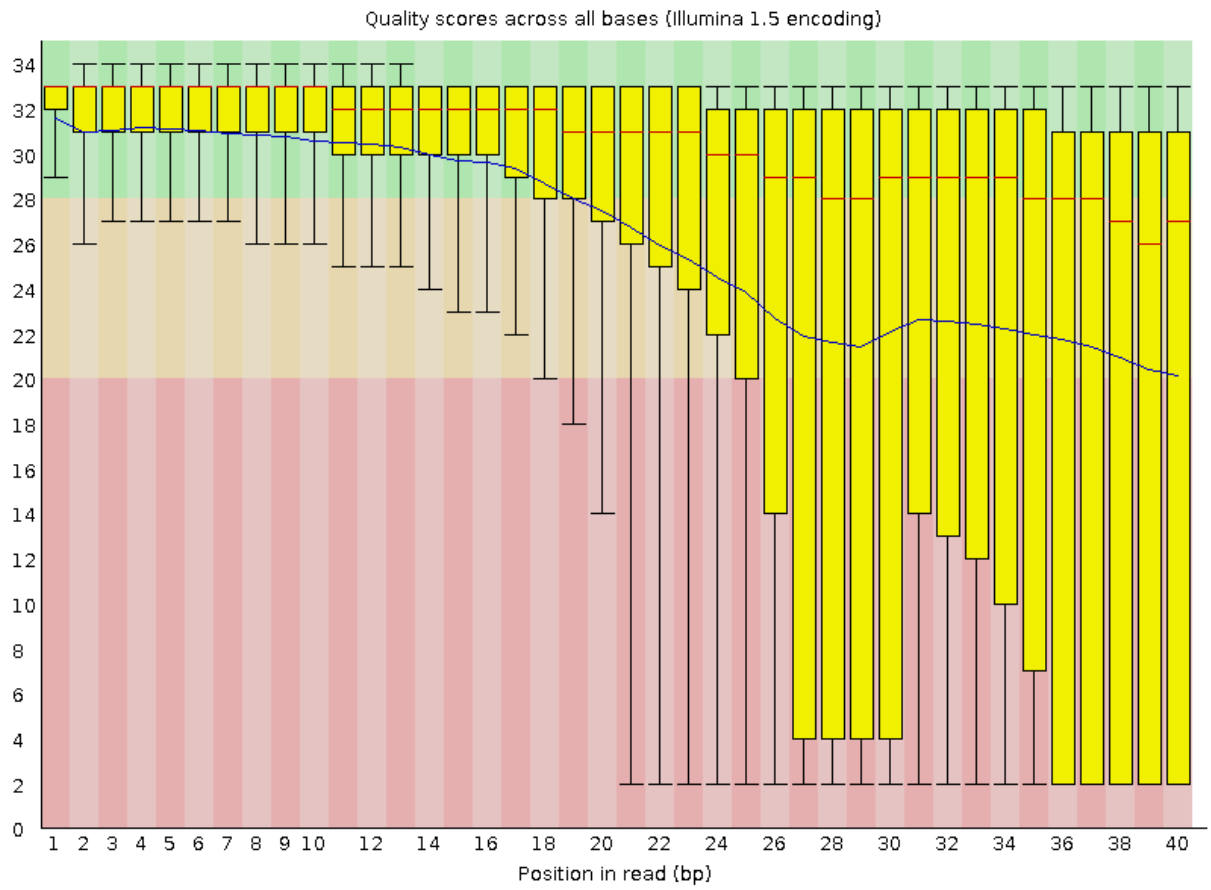


Figure 24. Διαγραμματική απεικόνιση χαμηλής ποιότητας αλληλούχισης με το πρόγραμμα FastQC (http://www.bioinformatics.babraham.ac.uk).

Φίλτράρισμα αλληλουχιών χαμηλής ποιότητας

Σε αλληλουχίσεις με χαμηλή ποιότητα Q score τα δεδομένα πρέπει να φιλτραριστούν για να βελτιωθεί η ποιότητα τους και να μπορέσουν να χρησιμοποιηθούν για περαιτέρω αναλύσεις.Το φιλτράρισμα, γνωστό με τον αγγλικό όρο 'trimming΄ μπορεί να πραγματοποιηθεί με διάφορα προγράμματα όπως το Condetri (© 2011 Smeds, Künstner). Το πρόγραμμα αυτό περιέχει και το πρόγραμμα filterPCRduplicates το οποίο αφαιρεί τις διπλασιασμένες αλληλουχίες που προέκυψαν από την διαδικασία της τεχνικής PCR. Κατά τη διαδικασία του 'trimming' με το Condetri το πρόγραμμα

διαβάζει την αλληλουχία από το 3'άκρο και εξάγει καλής ποιότητας σειρές αλληλούχισης.Σε paired-end αλληλουχίες το φιλτράρισμα γίνεται κατά ζεύγη και αν δεν είναι καλής ποιότητας ένα από τα δύο τότε σώζεται μόνο το ένα(single-end) ή διαγράφεται. Η διαδικασία περιέχει δύο βήματα (Figure **25**, Figure **26**).

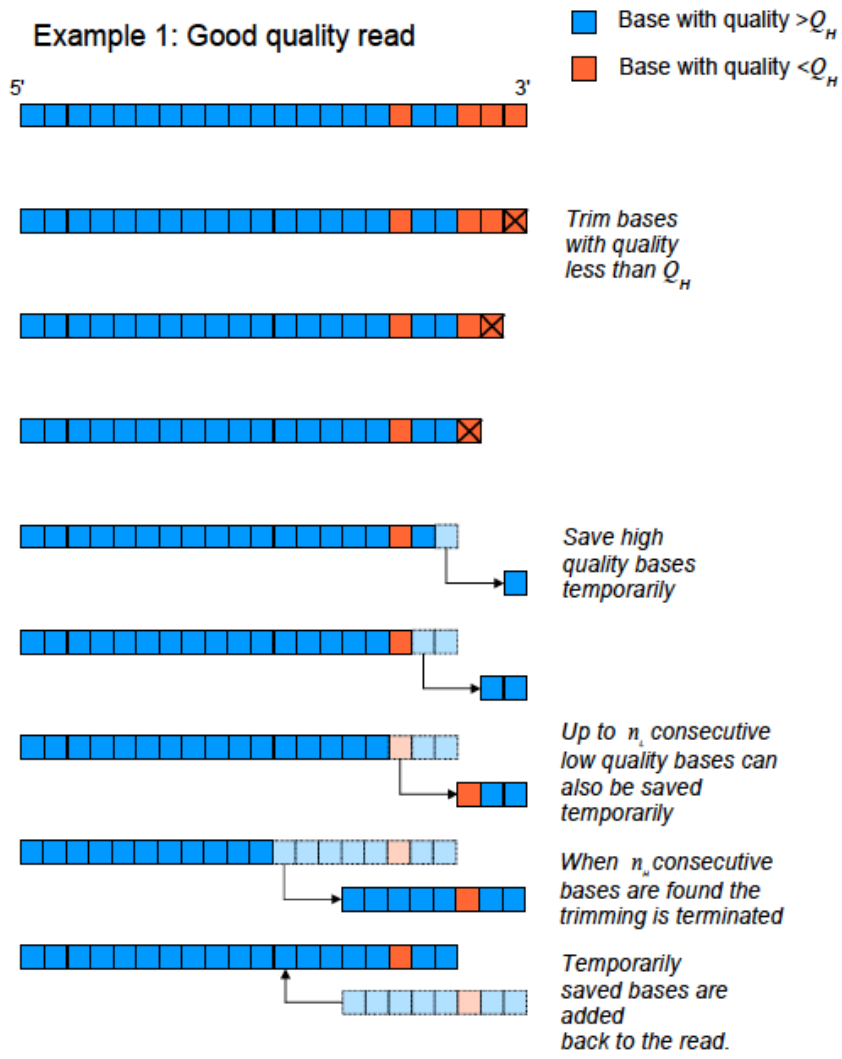


Figure 25. Trimming σε read υψηλής ποιότητας (Smeds and Künstner, 2011).

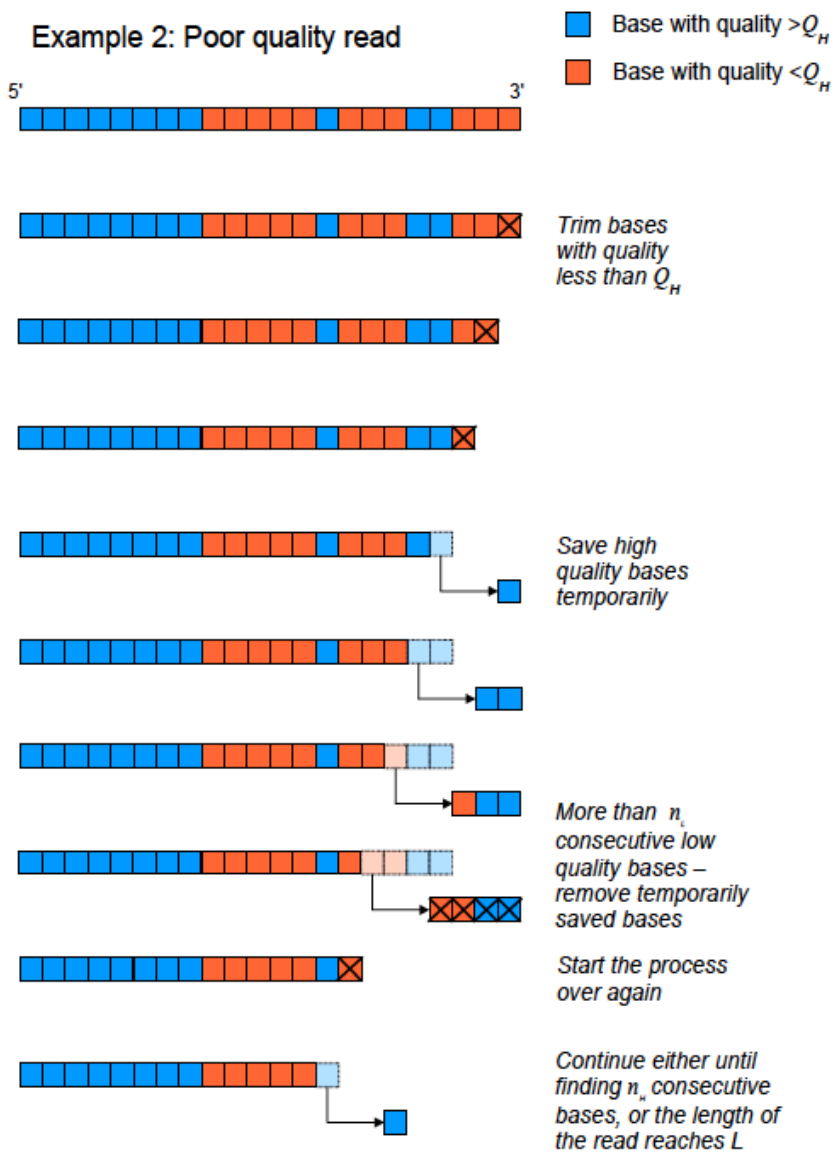


Figure 26. Trimming σε read χαμηλής ποιότητας (Smeds and Künstner, 2011).

Ύστερα, οι αλληλουχίες επεξεργάζονται και με το πρόγραμμα filterPCRduplicates το οποίο παρουσιάζει τις τελικές μοναδικές (files.unique) ακολουθίες.

2.5.3 Συναρμολόγηση (Assembly) των contigs

Με βάση την αλληλοεπικάλυψη μεταξύ των διαφόρων sequence reads γίνεται η συναρμολόγιση των διαφόρων contigs και mRNAs. Ένας από τους αλγόριθμους που

χρησιμοποιείται για την de novo συναρμολόγιση των contigs από δεδομένα RNAsequencing είναι αυτός του προγράμματος Trinity (Broad Institute and the Hebrew University of Jerusalem). Το πρόγραμμα αυτό αποτελείται από τρία διαφορετικά software το Inchworm, Chrysalis και Butterfly τα οποία εκτελούν συγκεκριμένες διαδικασίες για την τελική συναρμολόγηση (Figure **27**) και βασίζεται στα γραφήματα de Bruijn (Figure **28**). Στο τέλος της διαδικασίας τα αποτελέσματα αποθηκεύονται σε fasta μορφή (Figure **29**).

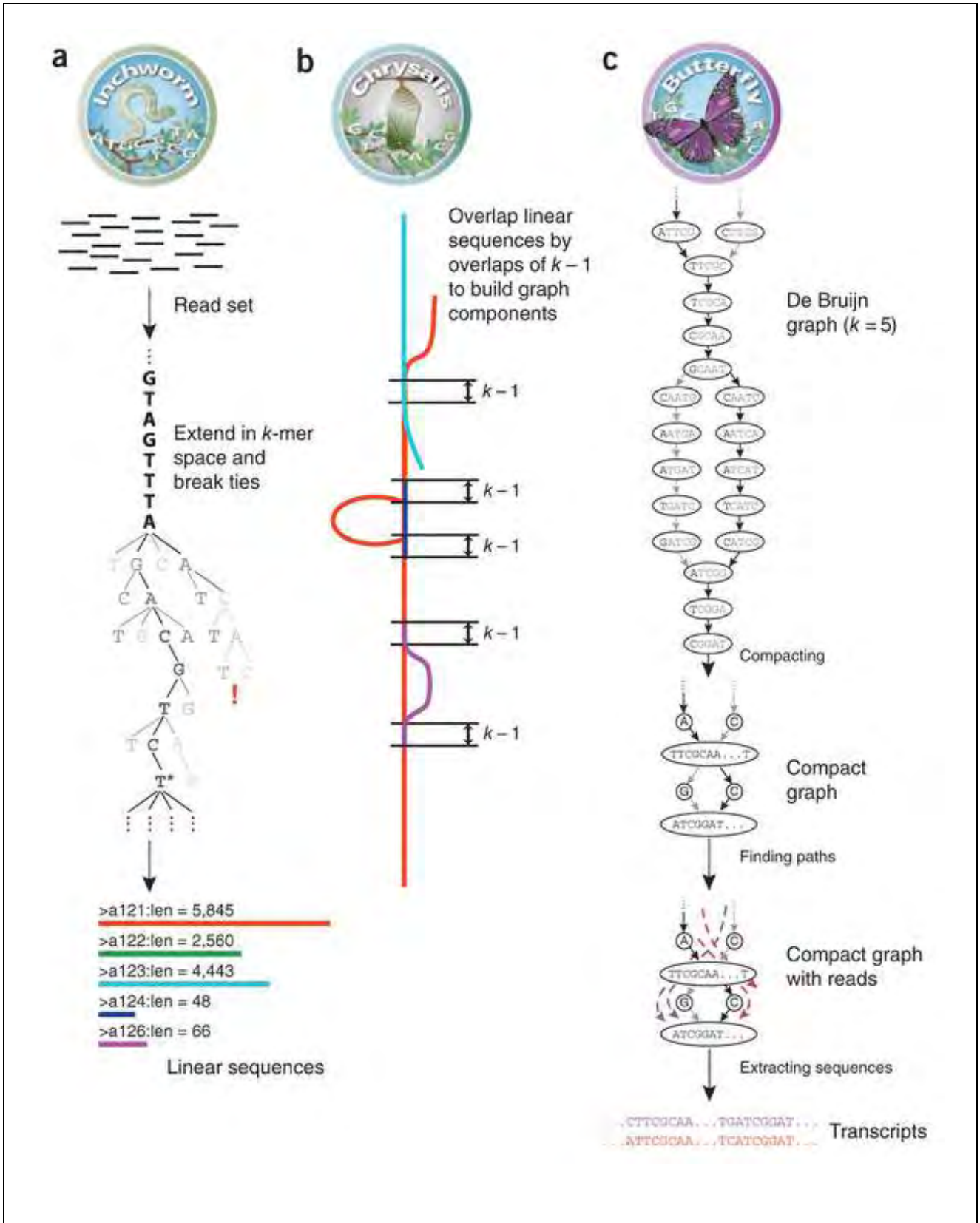


Figure 27. Σχηματική αναπαράσταση των τριών διαφορετικών βημάτων(a,b,c) του προγράμματος Trinity για την κατασκευή και ανασύσταση των RNA μεταγράφων (http://www.nature.com/nbt/journal/v29/n7/full/nbt.1883.html)

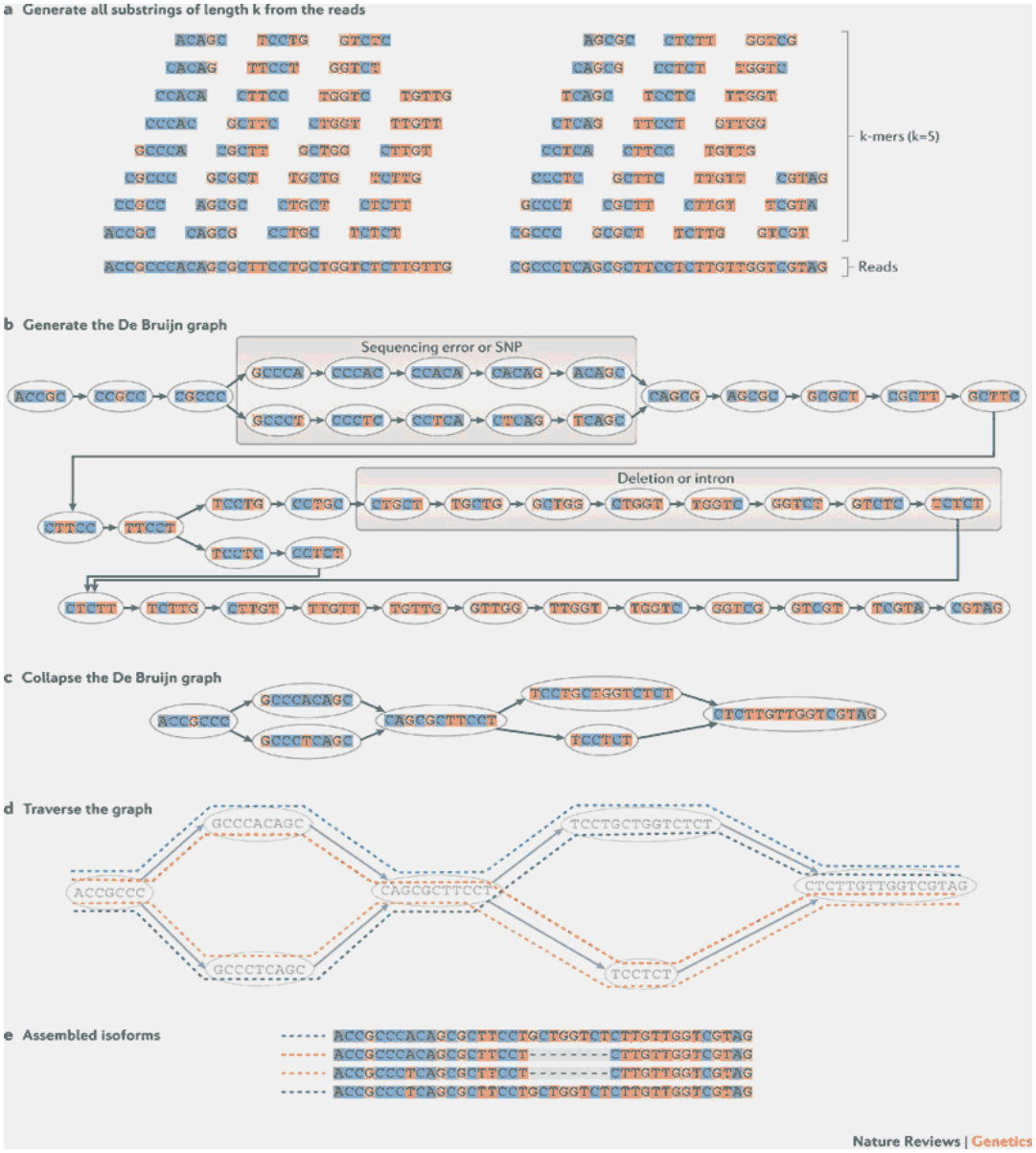


Figure 28. Γράφημα de Bruijn για τη συναρμολόγηση μεταγράφων (Martin *et al.*,2011).

>c2_g0_i1 len=2364 path=[0:0-587 588:588-1076 1146:1077-2363]
GAGCTCTTCAGGAGGGGGAATGTGCTTGTGGTTTTTGGTCTTGTGCATTTT
GTGACAAA
GAATTCCCTTTTGAATCGCGCTGTTCCCTTGAAACCCTGGAGCCTCTGGTT
CAAGCAGC
CAGTCAGTCTGTGCAGTGTCCCTGACGTCATCCGGCGTATGCATAAGCTC
TGCTATTGT
TTACCGCTAGAGCAGGGCTGAGGACTGCAGTCTCTGCTGCTGCTCGCAGA
CCTGCCCTG

Figure 29. Παράδειγμα ακολουθίας contig σε fasta μορφή μετά το Trinity (http://trinityrnaseq.sourceforge.net/advanced_trinity_guide).

2.5.4 Πρόβλεψη πεπτιδίων (peptide prediction) και Ομαδοποίηση πρωτεϊνών (Clustering)

Για την πρόβλεψη πεπτιδίων από συναρμολογημένα mRNAs χρησιμποποιείται το πρόγραμμα TransDecoder (Brian Haas and Alexie Papanicolaou) που είναι μέρος του προγράμματος Trinity και θα μπορούσε να χαρακτηριστεί ως 'ηλεκτρονικό ριβόσωμα'.

Όταν ολοκληρωθεί η 'μετάφραση' των mRNAs γίνεται χρήση του προγράμματος Usearch (®Robert Edgar, drive5.com/usearch) το οποίο χρησιμοποιεί ένα σύνολο διαφορετικών αλγόριθμων (π.χ smallmem/uclust) για την ταξινόμηση (sorting) και την ομαδοποίηση (clustering) των πρωτεϊνών με βάση την ομοιότητά τους (Figure **30**). Για κάθε ομάδα ακολουθιών που είναι πολύ όμοιες μεταξύ τους επιλέγεται μια πρωτεΐνη-αντιπρόσωπος (centroid).

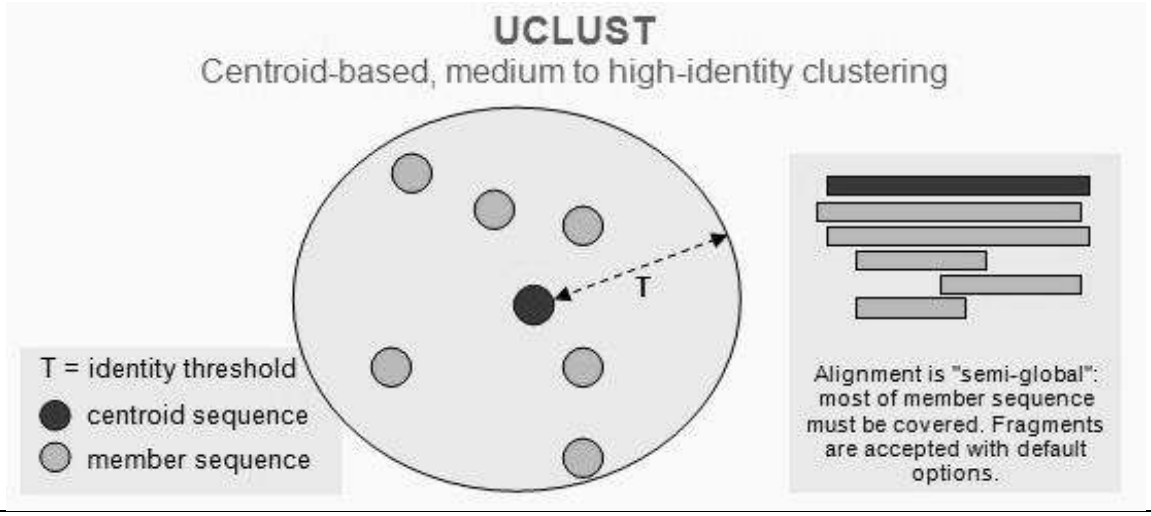


Figure 30. Λειτουργία του αλγόριθμου του uclust (drive5.com/usearch/manual/uclust_algo.html).

2.5.5 Φυλογενετική ανάλυση (Phylogenetics)

Όπως αναλύσαμε και στη αρχή, για να μελετήσουμε τις εξελικτικές σχέσεις ανάμεσα σε διαφορετικούς οργανισμούς ή μεταξύ παρόμοιων οργανισμών χρησιμοποιούμε τις φυλογενετικές αναλύσεις.Φυλογενετική, λοιπόν είναι η εξελικτική σχέση που προκύπτει μετά από μοριακές και μορφολογικές αναλύσεις μεταξύ των οργανισμών.Τα βασικά βήματα για την διαδικασία της φυλογένεσης είναι πέντε:

- εύρεση των ομόλογων ακολουθιών
- πολλαπλή στοίχιση των ομόλογων ακολουθιών
- Επιλογή ενός εξελικτικού μοντέλου
- κατασκευή φυλογενετικού δέντρου
- αξιολόγηση της εμπιστοσύνης των διαφόρων κλάδων του δέντρου

2.5.5.1 Εύρεση ομόλογων ακολουθιών

Υπάρχουν δύο είδη στοίχισης :

1. Η ολική στοίχιση (global alignment)

2. Η τοπική στοίχιση (local alignment)

Στην ολική στοίχιση γίνεται προσπάθεια για να στοιχιθούν όσο το δυνατόν περισσότεροι χαρακτήρες σε όλο το μήκος των δύο αλληλουχιών. Η στοίχιση αυτή χρησιμοποιείται για ακολουθίες οι οποίες δεν έχουν αποκλίνει σε μεγάλο ποσοστό και έχουν παρόμοιο μέγεθος. Μια από τις κλασσικές μεθόδους που βασίζεται στον δυναμικό προγραμματισμό είναι η Needleman-Wunsch. Αντίθετα, στην τοπική στοίχιση η διαδικασία περιλαμβάνει νησίδες στοίχισης και όχι όλο το μήκος και μπορούν να στοιχιθούν ακολουθίες που είναι απομακρυσμένες με συντηρημένες μόνο κάποιες περιοχές τους. Πιο συχνά, η χρήση αυτής της στοίχισης γίνεται για την αντιστοίχιση του mRNA με γενωμικό DNA η οποία βασίζεται στις κλασσικές μεθόδους δυναμικού προγραμματισμού Smith-Waterman και ευρετικών-heuristics μεθόδων, όπως το Blast.

Η ομάδα προγραμμάτων Blast (blast.ncbi.nlm.nih.gov, Altschul et al., 1990) είναι μια σειρά από υπολογιστικά εργαλεία που χρησιμοποιούνται για τον εντοπισμό ομόλογων ακολουθιών. Υπάρχουν διαφορετικά προγράμματα του Blast και η διαφοροποίηση τους βασίζεται στο είδος της ακολουθίας της βάσης δεδομένων και της ακολουθίας επερώτησης (Blastn, Blastp, Blastx, tBlasn, tBlastx) (Table **3**).

Program	Database	Query	Typical uses
BLASTN	Nucleotide	Nucleotide	Mapping oligonucleotides, cDNAs, and PCR products to a genome; screening repetitive elements; cross-species sequence exploration; annotating genomic DNA; clustering sequencing reads; vector clipping
BLASTP	Protein	Protein	Identifying common regions between proteins; collecting related proteins for phylogenetic analyses
BLASTX	Protein	Nucleotide translated into protein	Finding protein-coding genes in genomic DNA; determining if a cDNA corresponds to a known protein
TBLASTN	Nucleotide translated into protein	Protein	Identifying transcripts, potentially from multiple organisms, similar to a given protein; mapping a protein to genomic DNA
TBLASTX	Nucleotide translated into protein	Nucleotide translated into protein	Cross-species gene prediction at the genome or transcript level; searching for genes missed by traditional methods or not yet in protein databases

Table 3. Τα βασικά προγράμματα του BLAST και οι χρήσεις τους.

Μια γρήγορη μέθοδος εντοπισμού ορθόλογων ακολουθιών μεταξύ δύο οργανισμών με βάση το εργαλείο Blast, είναι το ανταποδωτικό Blasta (Best Reciprocal Blast) όπου μία ακολουθία Α από ένα οργανισμό όταν βρίσκει ως καλύτερο blast-hit μια ακολουθία Β από ένα άλλο οργανισμό θα πρέπει να ισχύει και το αντίστροφο, για να θεωρηθούν ως ορθόλογες. Η μέθοδος αυτή είναι απλή, γρήγορη και σχετικά αυστηρή.

2.5.5.2 Πολλαπλή στοίχιση (Multiple Sequence Alignment)

Η πολλαπλή στοίχιση ακολουθιών περιλαμβάνει τη στοίχιση περισσότερων από δύο ομόλογες ακολουθίες, ώστε να εντοπιστούν οι συντηρημένες περιοχές μιας οικογένειας (Figure **31**). Αυτού του είδους οι στοιχίσεις μπορούν να γίνουν με κάποια Ευρετική μέθοδο ή τον Δυναμικό προγραμματισμό, μόνο όμως όταν είναι λίγες οι ακολουθίες. Ένα από τα προγράμματα που χρησιμοποιείται ευρέως είναι το Muscle (Edgar, 2004). Εν συνεχεία θα πρέπει να γίνουν οι απαραίτητες διορθώσεις στη στοίχιση είτε χειροκίνητα είτε μέσω προγραμμάτων όπως το Gblocks (www.phylogeny.fr/version2_cgi/one_task.cgi?task_type=gblocks) το οποίο κόβει τις μη καλά συντηρημένες περιοχές, με βάση κάποια κριτήρια που ορίζει ο χρήστης. Το Seaview είναι ένα πρόγραμμα που επιτρέπει την πολλαπλή στοίχιση ακολουθιών καθώς επίσης και την χειροκίνητη διόρθωση της στοίχισης.

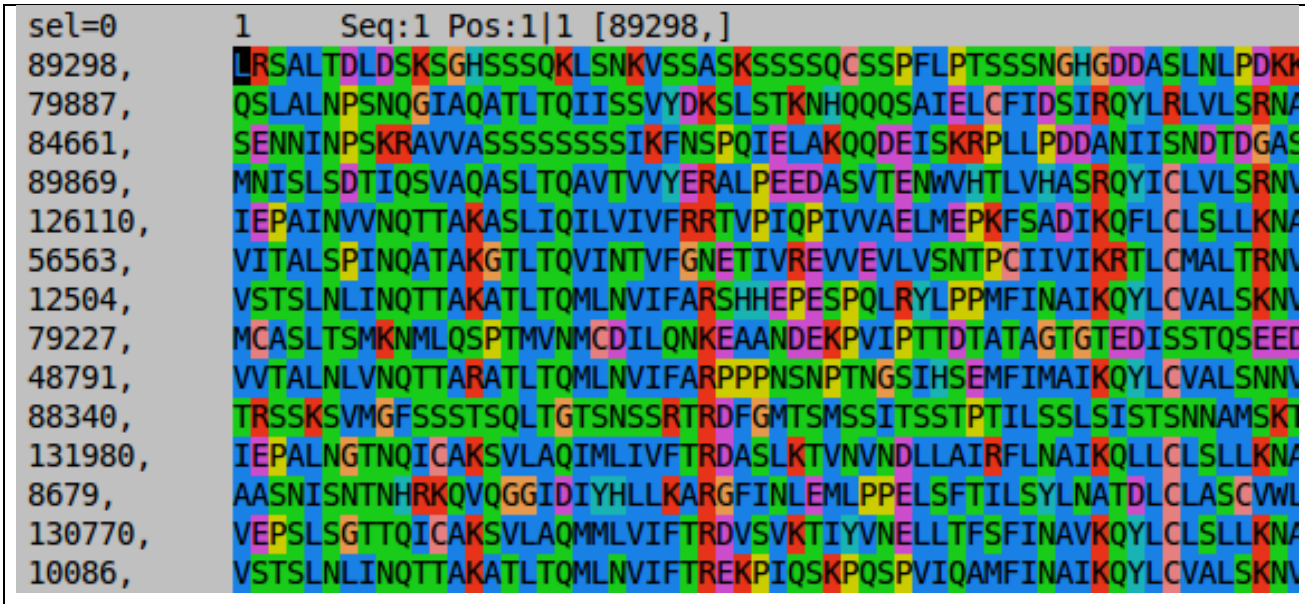


Figure 31. Παράδειγμα πολλαπλής στοίχισης πρωτεϊνών με το πρόγραμμα Seaview.

2.5.5.3 Κατασκευή Φυλογενετικών δέντρων

Υπάρχουν δύο μέθοδοι κατασκευής φυλογενετικών δέντρων:

- *Μέθοδοι βασιζόμενες σε αποστάσεις:*

Οι μέθοδοι αυτές βασίζονται στην ανομοιότητα (απόσταση) που μπορεί να έχουν οι στοιχισμένες ακολουθίες. Οι πιο γνωστές μέθοδοι είναι οι UPGMA και Neighbor Joining που βασίζονται στην ομαδοποίηση και οι Fitch-Margoliash και Minimum Evolution (ελάχιστης εξέλιξης), που βασίζονται στη βελτιστοποίηση. Σύμφωνα με τις μεθόδους αυτές υπολογίζουμε την παρατηρούμενη απόσταση από την στοίχιση, δηλαδή βλέπουμε σε ποιές θέσεις δεν ταιριάζουν οι χαρακτήρες. Όμως, η παρατηρούμενη απόσταση δεν συμπίπτει με την πραγματική (εξελικτική) απόσταση, λόγω πιθανών πολλαπλών αντικαταστάσεων στην ίδια θέση.Όσο μεγαλύτερη είναι η απόσταση, τόσο πιο πιθανό να έχουν συμβεί πολλές αντικαταστάσεις στην ίδια θέση, με αποτέλεσμα να υποεκτιμάται η πραγματική γενετική απόσταση. Επομένως, για την κατασκευή των σωστών φυλογενετικών δέντρων πρέπει πρώτα να υπολογιστεί η παρατηρούμενη απόσταση και στη συνέχεια με την επιλογή κάποιου κατάλληλου εξελικτικού μοντέλου να γίνει η διόρθωση της παρατηρούμενης απόστασης σε πραγματική/γενετική (Figure **32**). Αυτά τα εξελικτικά μοντέλα είναι στατιστικά και κάνουν κάποιες παραδοχές. Αν η απόσταση είναι πολύ μεγάλη, υπάρχει πιθανότητα να έχει επέλθει κορεσμός και επομένως δεν είναι δυνατόν να γίνει σωστή διόρθωση.

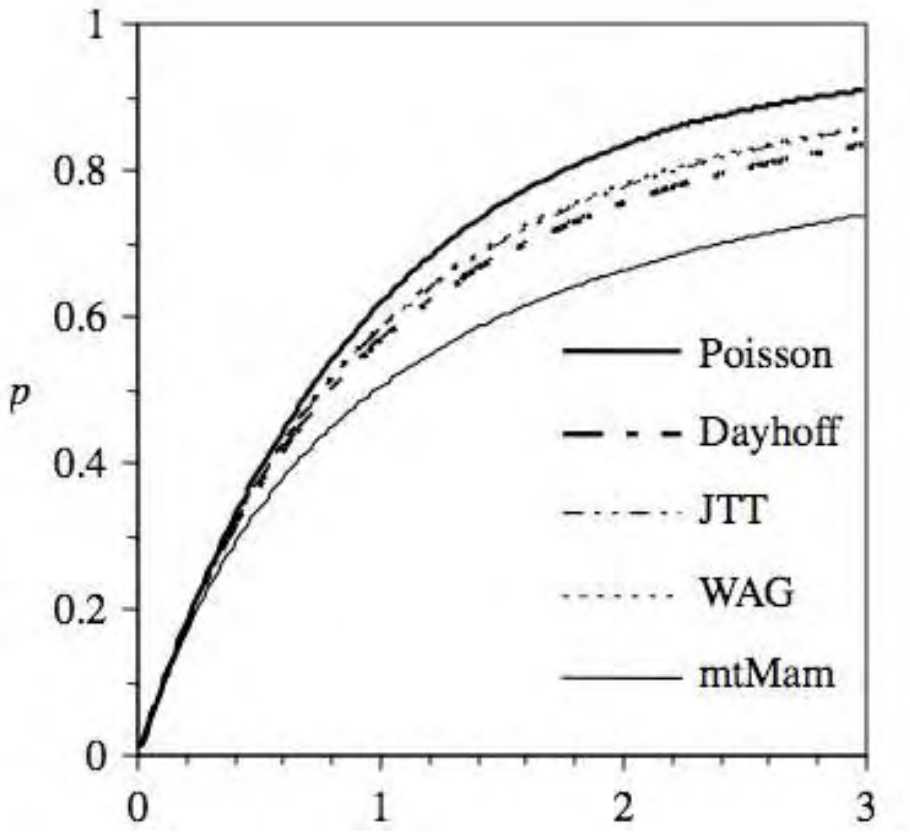


Figure 32. Διόρθωση των παρατηρούμενων αποστάσεων σε πραγματικές για πρωτεΐνες, με διάφορα εξελικτικά μοντέλα.

- *Μέθοδοι βασιζόμενες σε χαρακτήρες:*

Στις μεθόδους αυτές δεν υπολογίζεται η απόσταση μεταξύ των ακολουθιών, αλλά με βάση τους χαρακτήρες που συναντώνται σε μια στοίχιση και ένα συγκεκριμένο εξελικτικό μοντέλο γίνεται προσπάθεια να βρεθεί το καλύτερο δέντρο που να εξηγεί τα παρατηρούμενα δεδομένα. Οι σημαντικότερη μέθοδος αυτής της κατηγορίας είναι η Μέγιστη Πιθανοφάνεια (Maximum Likelihood). Το πλεονέκτημα αυτής της κατηγορίας μεθόδων είναι ότι επιτρέπουν την ανακατασκευή προγονικών ακολουθιών, κάτι που δεν είναι δυνατό με τις μεθόδους αποστάσεων. Τα τελευταία χρόνια έχει φανεί ότι η μέθοδος της Μέγιστης Πιθανοφάνειας είναι η προτιμότερη, οστώσο έχει μεγάλο υπολογιστικό κόστος. Επομένως, η επιλογή της μεθόδου κατασκευής φυλογενετικών δέντρων εξαρτάται από την υπολογιστική ισχύ και τα χρονικά περιθώρια που έχει ο ερευνητής.

Το PhyML είναι ένα πολύ δημοφιλές λογισμικό που χρησιμοποιεί τη μέθοδο της Μέγιστης Πιθανοφάνειας για την κατασκευή φυλογενετικών δέντρων. Η αξιοπιστία των παραγόμενων δέντρων και των επιμέρους κλάδων ελέγχεται με τη μέθοδο bootstrap (επαναδειγματοληπτική αξιολόγηση δέντρων) ή το Approximately-Likelihood Ratio Test (aLRT) (Anisimova *et al.*, 2006).

2.5.5.4 Αξιολόγηση της αξιοπιστίας των επιμέρους κλάδων ενός δέντρου

Η πιο δημοφιλής μέθοδος για την εκτίμηση της αξιοπιστίας των επιμέρους κλάδων ενός δέντρου είναι το bootstrap. Τα βασικά βήματα της μεθόδου αυτής είναι:

- Τυχαία δειγματοληψία θέσεων της πολλαπλής στοίχισης.
- Μια θέση μπορεί να επιλεγεί περισσότερες από μια φορές ή και καμία.
- Δημιουργία μιας νέας αλλαγμένης πολλαπλής στοίχισης.
- Η διαδικασία επαναλαμβάνεται 100-1000 φορές.
- Για κάθε νέα πολλαπλή στοίχιση, υπολογίζεται το αντίστοιχο δένδρο.
- Τα νέα δένδρα συγχωνεύονται σε ένα νέο δένδρο (consensus tree).
- Boostrap → συχνότητα εμφάνισης ενός κόμβου.
- Bootstrap 70% → 95% διάστημα εμπιστοσύνη.

Το Jacknife είναι μια μέθοδος παρόμοια με το Bootstrap, με την διαφορά ότι επιλέγονται τυχαία (δίχως αντικατάσταση) οι μισές στήλες της πολλαπλής στοίχισης. Το πρόβλημα είναι ότι τα νέα δένδρα δημιουργούνται από λιγότερα δεδομένα.

3 ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

3.1 Συλλογή και αλληλούχιση δειγμάτων Pinna nobilis

Δείγματα της Pinna nobilis συλλέχθηκαν από την περιοχή της Νέας Μηχανιώνας του νομού Θεσσαλονίκης Η συλλογή των δειγμάτων έγινε τον μήνα Απρίλιο, όταν τα νέα άτομα βρίσκονται στην πιο έντονη φάση ανάπτυξης. Συγκεκριμένα, συλλέχθηκε μανδύας και γάγγλια από ζωντανά δείγματα. Οι ιστοί αυτοί καταψύχθηκαν σε ξηρό πάγο και αργότερα στους -80^{o}C. Η αλληλούχιση του ολικού mRNA έγινε στο Beijing Genomics Institute με τεχνολογία Illumina HiSeq 1000 και με ακολουθίες paired-end reads που το μήκος τους ήταν 91 βάσεις

3.2 Συγκέντρωση δημοσιευμένων δεδομένων RNA ακολουθιών και Γονιδιωμάτων

Στην ανάλυση χρησιμοποιήθηκαν συνολικά 16 οργανισμοί της οικογένειας των Δίθυρων (Bivalvia) και 1 οργανισμός της οικογένειας των Γαστρόποδων (Gastropoda) ως εξωομάδα, όπως παρουσιάζεται στον παρακάτω συγκεντρωτικό πίνακα (Table 4). Σε πολλούς οργανισμούς, λόγω λίγων δεδομένων που βρέθηκαν στο SRA (κυρίως από αλληλούχιση με την τεχνολογία 454 GS FLX) χρειάστηκε να γίνει συνένωση των αρχείων. Ένας οργανισμός, λοιπόν για να αποκτήσει επαρκή δεδομένα για την περαιτέρω βιοπληροφορική επεξεργασία αποτελείτω από πολλά αρχεία που έφεραν αλληλουχίες RNA τα οποία είχαν συνενωθεί με την εντολή cat των Linux (cat file1 file2 > file3).

Οργανισμός	Φυλογενετική ομάδα	Είδος δεδομένων	Τεχνολογία αλληλούχισης
Arctica_islandica	Bivalvia	Αλληλουχίες RNA	454 GS FLX Titanium
Astarte_sulcata	Bivalvia	Αλληλουχίες RNA	Illumina HiSeq 2000
Corbicula_fluminea	Bivalvia	Αλληλουχίες RNA	Illumina Genome Analyzer IIx
Crassostrea_virginica	Bivalvia	Αλληλουχίες RNA	Illumina HiSeq 2000
Crassostrea_gigas	Bivalvia	Ολόκληρο το γονιδίωμα	Genome
Laternula_elliptica	Bivalvia	Αλληλουχίες RNA	454 GS FLX Titanium
Lottia_gigantia	Gastopoda	Ολόκληρο το γονιδίωμα	Genome
Mercenaria_campechiensis	Bivalvia	Αλληλουχίες RNA	Illumina HiSeq 2000
Myochama_anomioides	Bivalvia	Αλληλουχίες RNA	Illumina HiSeq 2000
Mytilus_edulis	Bivalvia	Αλληλουχίες RNA	454 GS FLX Titanium
Mytilus_galloprovincialis	Bivalvia	Αλληλουχίες RNA	Illumina Genome Analyzer II
Neotrigonia_margaritacea	Bivalvia	Αλληλουχίες RNA	Illumina HiSeq 2000
Pinctada_fucata	Bivalvia	Ολόκληρο το γονιδίωμα	Genome
Pinna_nobilis	Bivalvia	Αλληλουχίες RNA	Illumina HiSeq 2000
Ruditapes_philippinarum	Bivalvia	Αλληλουχίες RNA	Illumina HiSeq 2000
Solemya_velum	Bivalvia	Αλληλουχίες RNA	Illumina Genome Analyzer IIx
Villosa_lienosa	Bivalvia	Αλληλουχίες RNA	Illumina HiSeq 2000

Table 4. Είδος δεδομένων των οργανισμών που συμμετείχαν στην ανάλυση

Για την Crassostrea gigas, Pinctada fucata και Lottia gigantia έγινε χρήση ολόκληρου του γονιδιώματος, όπως αυτό βρέθηκε σε βάσεις δεδομένων του διαδικτύου (Ensembl Genomes http://ensemblgenomes.org). Η συλλογή των υπόλοιπων δεδομένων έγινε μέσω της βάσης δεδομένων του SRA (Sequence Read Archive) του

NCBI (National Center for Biotechnology Information) (http://www.ncbi.nlm.nih.gov/sra) από την οποία τα αρχεία αποθηκεύτηκαν στο υπολογιστικό σύστημα ως SRRfiles.sra.

3.3 Βιοπληροφορικές αναλύσεις

Τα αρχεία που κατεύηκαν από το SRA σε μορφή .sra μετατράπηκαν σε αρχεία της μορφής fastq μέσω του SRAtoolkit. Οι οργανισμοί από τους οποίους χρησιμοποιήθηκε ολόκληρο το γονιδίωμα τους αποθηκεύτηκαν απευθείας σε fasta μορφή. Εν συνεχεία έγινε έλεγχος της ποιότητας των sequence reads των αρχείων fastq με το πρόγραμμα FastQC. Το φιλτράρισμα των αλληλουχιών (trimming στο 3' άκρο της αλληλουχίας) έγινε με το πρόγραμμα Condetri και έπειτα με το πρόγραμμα PCRduplicates. Όλοι οι φάκελοι ελέγχθηκαν ξανά με το πρόγραμμα FastQC.

Η συναρμολόγηση των φιλτραρισμένων sequence reads έγινε με το πρόγραμμα Trinity για κάθε οργανισμό και κατασκευάστηκαν τα contigs. Κατόπιν με το πρόγραμμα Cap3 ενώθηκαν πολύ όμοια contigs με μεγάλη αλληλεπικάλυψη σε ένα contig.

Τα αρχεία των μεταγράφων σε που ήταν σε μορφή fasta χρησιμοποιήθηκαν στο πρόγραμμα TransDecoder για να προβλεφθούν τα πεπτίδια που κωδικοποιούν.

Στη συνέχεια, το πρόγραμμα Usearch χρησιμοποιήθηκε για να ομαδοποιηθούν οι πρωτεϊνικές ακολουθίες με ποσοστό 95% και παραπάνω και για κάθε cluster βρέθηκε το αντιπροσωπευτικό πεπτίδιο (centroid).

Αφού για κάθε οργανισμό εντοπίστηκαν τα πεπτίδια που εκφράζονται στα υπό ανάλυση δεδομένα, με τη χρήση του blastp έγινε ανταποδωτικό blast με σημείο αναφοράς το πλήρες γονιδίωμα της Crassostrea gigas ώστε για κάθε οργανισμό να βρεθούν τα ορθόλογα γονίδιά του.

Με τη βοήθεια κατάλληλων προγραμμάτων της Perl που γράφτηκαν στο εργαστήριο δημιουργήθηκε ένας πίνακας όπου για κάθε γονίδιο της Crassostrea gigas εμφανιζόταν και η ορθόλογη ακολουθία στον αντίστοιχο οργανισμό. Στον πίνακα αυτό η κάθε γραμμή αντιπροσώπευε ένα γονίδιο ενώ η κάθε στήλη έναν οργανισμό. Με βάση αυτόν τον πίνακα και με άλλα προγράμματα της Perl, για κάθε γονίδιο της Crassostrea gigas μαζεύονταν αυτόματα σε ένα αρχείο οι ορθόλογες ακολουθίες από τα άλλα υπό εξέταση είδη. Έτσι, για κάθε γονίδιο και τα ορθόλογά του, υπήρχε ένα αρχείο με τις ακολουθίες τους σε μορφή fasta.

Καθένα από τα παραπάνω αρχεία υπέστησαν πολλαπλή στοίχιση με το πρόγραμμα Muscle (Seaview software). Για την αυτοματοποίηση της παραπάνω διαδικασίας γράφτηκαν προγράμματα στην γλώσσα Perl. Στη συνέχεια, για κάθε ομάδα στοιχισμένων ορθολόγων έγινε αυτόματη επιδιόρθωση της πολλαπλής στοίχισης με το πρόγραμμα Gblocks (Seaview software). Κατόπιν, με ένα άλλο πρόγραμμα της Perl, όλες οι ομάδες στοιχισμένων ορθολόγων ενώθηκαν σε μια υπερ-πολλαπλή στοίχιση (alignment concatenation). Αυτό επέτρεψε την πιο εύκολη διαχείριση των δεδομένων. Προκειμένου να γίνει φυλογενετική ανάλυση από την παραπάνω υπερ-πολλαπλή στοίχιση, με το πρόγραμμα Prottest έγινε έλεγχος για το ποιό είναι το πιο κατάλληλο εξελικτικό μοντέλο. Στη συνέχεια, κατασκευάστηκαν φυλογενετικά δέντρα με τις μεθόδους Neighbor Joining (Poisson model) και μέσω της μεθόδου maximum likelihood στο πρόγραμμα PhyML. Για την προβολή και επεξεργασία των φυλογενωμικών δέντρων έγινε χρήση των προγραμμάτων Seaview και Treedyn.

3.4 Υπολογιστικό Περιβάλλον εργασίας

Η συλλογή και ανάλυση των δεδομένων έλαβε χώρα σε υπολογιστικό σύστημα Intel® Xeon, 2XCPU E5620, Quad Core 2.40 GHz με μνήμη 96GB και σκληρό δίσκο 3TB.Το λογισμικό χρήσης ήταν Linux-Ubuntu 12.04 το οποίο είχε ενσωματωμένη τη γλώσσα προγραμματισμού Perl η οποία χρησιμοποιήθηκε εκτενώς

για την επεξεργασία των δεδομένων της ανάλυσης.

4 ΑΠΟΤΕΛΕΣΜΑΤΑ

Το πρώτο μέρος των αποτελεσμάτων αφορά την πρόβλεψη γονιδίων, ενώ το δεύτερο μέρος αφορά τη φυλογενετική ανάλυση. Οι χρόνοι της ολοκλήρωσης κάθε προγράμματος δεν ήταν σταθεροί ανάμεσα στα διάφορα δεδομένα λόγω διαφορετικότητας όγκου. Οι οργανισμοί που χρησιμοποιήθηκαν τελικά στις φυλογενετικές αναλύσεις ήταν 14 αντί για 17, από τους οποίος συλλέξαμε δεδομένα αρχικά. Οι τρείς οργανισμοί Arctica islandica, Laternula elliptica και Ruditapes philippinarum δεν χρησιμοποιήθηκαν λόγω των λιγοστών δεδομένων που προέκυψαν μετά τον εντοπισμό των ορθόλογων γονιδίων.

4.1 Εντοπισμός Γονιδίων και πρωτεϊνών

Αρχικά, έγινε η λήψη των αρχείων .sra (Run → Download) από το Sequence Read Archive. Για τα αρχεία που είχαν αλληλουχηθεί με τεχνολογία Illumina η μετατροπή του αρχείου από .sra σε .fastq έγινε με την εντολή:

./fastq-dump -M 50 –split-3 file.sra,

-M (minReadLen): η παράμετρος αυτή λειτουργεί σαν φίλτρο για το μήκος των ακολουθιών.
-split-3: στην περίπτωση που τα reads είναι paired-end χωρίζεται το κάθε ζευγάρι σε διαφορετικούς φακέλους → file_1.fastq/file_2.fastq.

Για τα αρχεία που είχαν αλληλουχηθεί με τεχνολογία 454 GS Flex η μετατροπή έγινε με τις εντολές:

./sff-dump file.sra, για την μετατροπή του φακέλου sra σε sff

και στη συνέχεια για την μετατροπή σε fastq μορφή:

./sff_extract -o file.fastq --min_left_clip [number] file.sff,

-o: είναι το output, δηλαδή ο φάκελος που θα δημιουργηθεί
--min_left_clip: φιλτράρει τις αλληλουχίες στο αριστερό άκρο μέχρι εκεί που ορίζει ο αριθμός που βάζουμε [number], ώστε να απομακρυνθούν όλοι οι αντάπτορες.

Τα αρχεία από την αλληλούχιση με την τεχνολογία 454 GS Flex περιείχαν στο αριστερό τους άκρο αντάπτορες, οι οποίοι θα δημιουργούσαν πρόβλημα στην περαιτέρω ανάλυση μας και έπρεπε να αφαιρεθούν με την παραπάνω εντολή.

Πριν και μετά το φιλτράρισμα με το πρόγραμμα Condetri και filterPCRduplicates όλα τα αρχεία που είχαν μετατραπεί σε fastq μορφή αναλύθηκαν με το πρόγραμμα FastQC, ώστε να εκτιμηθεί η ποιότητα των αλληλουχιών καθώς και η αποτελεσματικότητα του trimming. Η επεξεργασία των δεδομένων στο FastQC έγινε με την εντολή :

./fastqc file.fastq

Για το trimming των reads τα αρχεία μεταφέρθηκαν στο directory του προγράμματος Condetri v2.2.pl, όπου δώθηκε για κάθε αρχείο fastq η παρακάτω εντολή.

για single-end αλληλουχίες:
./condetri_v2.2.pl -fastq1=file.fastq -prefix=file -hq=25 -lq=13 -frac=0.8 -minlen=35 -mh=5 -ml=1 -sc=33 -rmN

και για paired-end αλληλουχίες:
./condetri_v2.2.pl -fastq1=file_1.fastq -fastq2=file_2.fastq -prefix=file -hq=25 -lq=13

-frac=0.8 -minlen=35 -mh=5 -ml=1 -sc=33 -rmN

-fastq1/fastq2: είναι το input file
-prefix: το όνομα του αρχείου που θα δημιουργηθεί
-hq: high quality threshold
-lq:low quality threshold
-minlen: το ελάχιστο μήκος που είναι αποδεκτό
-sc: Ανάλογα με το αν τα δείγματα αλληλουχήθηκαν με τεχνολογία της Illumina 1.5 ή αργότερα, χρησιμοποιείται διαφορετικό σύστημα βαθμολόγισης της ποιότητας των reads. (ASCII-sc 64 ή 33)
-rmN: αφαίρεση μη διαβασμένων βάσεων

Στη συνέχεια όσα αρχεία ήταν paired-end υπέστησαν περαιτέρω φιλτράρισμα με το πρόγραμμα filterPCRduplicates με σκοπό να αφαιρεθούν αλληλουχίες. Η εντολή που δόθηκε ήταν η παρακάτω:

./filterPCRdupl.pl -fastq1=file_trim1.fastq -fastq2=file_trim2.fastq -prefix=file -cmp=31

-cmp: ο αριθμός των βάσεων που συγκρίνεται σε κάθε read από κάθε ζεύγος για να εκτιμηθεί αν είναι διπλασιασμένα τα reads λόγω PCR.

Τα αποτελέσματα του PCRduplicates αποθηκεύτηκαν σε δύο φακέλους: file_unique1.fastq, file_unique2.fastq.

Μετά από όλες τις παραπάνω διαδικασίες, ο αριθμός των sequence reads που απέμειναν για κάθε οργανισμό συνοψίζονται στον παρακάτω πίνακα (Table **5**)

Organism	*fastq format SAMPLES*	*after_Condetri/PCRdupl*
Corbicula fluminea	33543565	11052677
Villosa lienosa	21203214	10893052
Solemya velum	5963543	3142358
Mytilus edulis	2203934	226083
Mytilus galloprovincialis	39878184	12691407
Crassostrea virginica	26428921	14156135
Astarte sulcata	37671660	4470207
Mercenaria campechiensis	17015685	2876616
Myochama anomioides	38532955	9317153
Neotrigonia margaritacea	24061675	12460125

Table 5. Ο αριθμός των sequence reads για κάθε οργανισμό πριν και μετά το trimming/PCRdupl.

Κατόπιν, μεταφέρθηκαν όλοι οι φάκελοι στο εγκατεστημένο πρόγραμμα Trinity (trinityrnaseq_r20131110) και με την παρακάτω εντολή για κάθε φάκελο ξεκίνησε η de novo συναρμολόγιση των contigs/mRNAs:

για single-end:

./Trinity.pl –seqType fq –JM 90G --single file_trim.fastq –CPU 12 --out file_run

και για paired-end:

./Trinity.pl –seqType fq –JM 90G --left file_unique1.fastq –right file_unique2.fastq –CPU 12 --out file_run

-seqType: η μορφή στην οποία βρίσκονται τα αρχεία

--JM: Jellyfish Memory, δηλαδή πόση μνήμη του υπολογιστή θα χρησιμοποιήσει το πρόγραμμα για την μέτρηση των k- mers

-CPU: ο αριθμός των CPUs

--out: ο φάκελος που περιέχει τα αποτελέσματα

Τα αποτελέσματα αποθηκεύτηκαν σε ένα φάκελο file_run, για κάθε οργανισμό και περιείχαν το αρχείο Trinity.fasta με τις ακολουθίες του κάθε contig/mRNA σε fasta

format. Εν συνεχεία, τα αρχεία μεταφέρθηκαν στο directory του προγράμματος Cap3 και εκτελέστηκε η εντολή:

./cap3 Trinity.fasta > file_run.cap3_out

και δημιουργήθηκαν δύο αρχεία, τα .contigs, και τα .singlets τα οποία ενώθηκαν σε ένα αρχείο με την εντολή cat του linux και μετονομάστηκαν σε [όνομα_οργανισμού]_transcripts.fasta:

Το επόμενο βήμα στην επεξεργασία των δεδομένων ήταν η πρόβλεψη των πεπτιδίων από τα αρχεία με τα μετάγραφα που είχαμε δημιουργήσει για κάθε οργανισμό. Όλα τα transcripts.fasta files μεταφέρθηκαν στο directory του προγράμματος TransDecoder το οποίο ανήκει στα συμπληρωματικά προγράμματα του Trinity (trinity-plugins). Η εντολή που χρησιμοποιήθηκε για το peptide prediction ήταν:

./TransDecoder -t file_transcripts.fasta

-t: το input file

Όταν ολοκληρώθηκε η επεξεργασία των δεδομένων από το πρόγραμμα, δημιουργήθηκαν τέσσερα αρχεία για κάθε οργανισμό, τα οποία ήταν όλα σε fasta μορφή. Το αρχείο file_transcripts.fasta.transdecoder.pep περιείχε τις πρωτεΐνες που είχαν προβλεφθεί για κάθε οργανισμό από το TransDecoder.

Ύστερα τα αρχεία με τις αλληλουχίες των πρωτεϊνών υπέστησαν επεξεργασία με το πρόγραμμα Usearch. Αρχικά έγινε sorting των ακολουθιών με βάση το μέγεθος με την παρακάτω εντολή:

./usearch -sortbylength file_transcripts.fasta.transdecoder.pep -output

file_transcripts.fasta.transdecoder.pep.sorted -minseqlength 50

-sortbylength: input file
-output: το αρχείο που θα περιέχει το αποτέλεσμα
-minseqlength: ελάχιστο μέγεθος της κάθε αλληλουχίας

Στη συνέχεια, έγινε clustering των ακολουθιών με τον αλγόριθμο smallmem με την εντολή:

./usearch -cluster_smallmem file_transcripts.fasta.transdecoder.pep.sorted -id 0.95 -centroids file_095_centroids.fasta -uc file_clusters.uc

-cluster_smallmem: input file
-id: ποσοστό τάυτισης ακολουθιών που θα ομαδοποιηθούν, στην ανάλυση μας αυτό τέθηκε στο 95%
-centroids: output file (εκπρόσωποι του κάθε cluster)
-uc: αποτέλεσμα του clustering

Τα αποτελέσματα από το TransDecoder και από το Usearch φαίνονται στον παρακάτω πίνακα (Table **6**):

Organism	*after_TransDeco*	*after_Usearch*
Corbicula fluminea	16209	15868
Villosa lienosa	27697	25100
Solemya velum	15815	14769
Mytilus edulis	9640	9115
Mytilus galloprovincialis	37854	35925
Crassostrea virginica	27904	21466
Astarte sulcata	10297	9879
Mercenaria campechiensis	5716	5411
Myochama anomioides	10767	10189
Neotrigonia margaritacea	17048	15791
Pinna nobilis	36346	27810

Table 6. Αριθμός αλληλουχιών για κάθε οργανισμό μετά τα προγράμματα TransDecoder και Usearch.

Στο σημείο αυτό πρέπει να σημειωθεί πως κατέβηκαν από τις βάσεις δεδομένων του διαδικτύου έτοιμα τα πρωτεώματα των οργανισμών Crassostrea gigas, Pinctada fucata και Lottia gigantea.

Κατόπιν, με χρήση της γλώσσας προγραμματισμού Perl δημιουργήθηκε ένα script για την ένωση όλων των γραμμών κάθε πρωτεϊνικής αλληλουχίας σε μια, ώστε να μπορέσει να γίνει σωστά η περαιτέρω ανάλυση. Επιπλέον, για τη σωστή λειτουργία της εύρεσης των ορθόλογων γονιδίων και προς πρακτικής διευκόλυνσης έγινε μετονομασία όλων των ακολουθιών ώστε να αντικατασταθεί το id κάθε πρωτεϊνικής αλληλουχίας, κάθε οργανισμού με το όνομα του οργανισμού και έναν μετρητή που ξεκινάει από το 1.

Με το τέλος της πρόβλεψης των γονιδίων και τις απαραίτητες τροποποιήσεις που αναφέρθηκαν παραπάνω, ακολούθησε η φυλογενετική ανάλυση.

4.2 Φυλογενετική Ανάλυση

4.2.1 Εντοπισμός ορθόλογων ακολουθιών

Για την εύρεση των ορθόλογων γονιδίων χρειάστηκε να γίνει ανταποδωτικό blastp με αλληλουχία αναφοράς αυτή του Crassostrea gigas. Αρχικά, δημιουργήθηκαν βάσεις δεδομένων blastp από όλους τους οργανισμούς με το πρόγραμμα makeblastdb στο directory του Blast/ncbi-blast-2.2.29+/bin δίνοντας την εντολή:

./makeblastdb -in file -out BlastDB_file_peps -dbtype prot

-in: input file
-out: το αρχείο με το αποτέλεσμα
-dbtype: το είδος των ακολουθιών

Έπειτα, έλαβε χώρα το blastp για κάθε κατεύθυνση. Στην πρώτη κατεύθυνση, οι ακολουθίες της Crassostrea gigas ήταν ακολουθίες επερώτησης ενώ στην δεύτερη κατεύθυνη οι ακολουθίες της Crassostrea gigas ήταν η βάση δεδομένων. Οι εντολές που χρησιμοποιήθηκαν για κάθε blastp, ήταν οι παρακάτω για κάθε περίπτωση αντίστοιχα:

./blastp -query file_Crassostrea_gigas_seq -db BlastDB_file_organism.peps -evalue 1e-10 -out Blastp_results_Crassostrea_gigas_vs_DBfile_organism.peps -outfmt "6 qacc sacc evalue qstart qend sstart send bitscore qlen slen length pident ppos" -max_target_seq 1 -num_threads 2

./blastp -query file_organism_seq -db BlastDB_Crassostrea_gigas.peps -evalue 1e-10 -out Blastp_results_organism_vs_DB_Crassostrea_gigas.peps -outfmt "6 qacc sacc evalue qstart qend sstart send bitscore qlen slen length pident ppos" -max_target_seq

1 -num_threads 2

-query: η αλληλουχία επερώτησης
-db: το αρχείο που θα χρησιμοποιηθεί ως βάση δεδομένων
-evalue: παράμετρος στατιστικής σημαντικότητας.
-out: το αρχείο που θα δημιουργηθεί με τα αποτελέσματα
-outfmt: το μοντέλο της φόρμας που θα έχουν τα αποτελέσματα από το blastp
-max_target_seq: την ποσότητα των αποτελεσμάτων που θέλουμε. Στην περίπτωση αυτή ορίστηκε 1 γιατί πήραμε την αλληλουχία με την μεγαλύτερη ομοιότητα και όχι τις υπόλοιπες ομόλογες της
-num_threads: πόσα threads θα χρησιμοποιήσει ο υπολογιστής για να τρέξει το blast

Με τα αποτελέσματα που προέκυψαν από κάθε blastp και με την δημιουργία ενός script στη γλώσσα προγραμματισμού Perl εντοπίστηκαν τα ανταποδωτικά χτυπήματα για κάθε οργανισμό με το Crassostrea gigas. Έτσι, δημιουργήθηκαν αρχεία για όλους τους οργανισμούς, τα οποία περιείχαν το καθένας τα ανταποδωτικά χτυπήματα μεταξύ του Crassostrea gigas και του αντίστοιχου οργανισμού (file_organism_rbh).

Στη συνέχεια, από τα παραπάνω αρχεία και με τη βοήθεια Perl scripts δημιουργήθηκε ένας πίνακας ορθόλογων ακολουθιών για κάθε γονίδιο της Crassostrea gigas. Στον πίνακα αυτό, η κάθε γραμμή είναι ένα γονίδιο της Crassostrea gigas και η κάθε στήλη έχει την ορθόλογη ακολουθία από τον συγκεκριμένο οργανισμό. Προκειμένου να γίνει η φυλογενωμική ανάλυση, χρειαζόμασταν γονίδια που είχαν ορθόλογες ακολουθίες σε όλους τους υπό εξέταση οργανισμούς. Έτσι, όταν στην φυλογενωμική μας ανάλυση θελήσαμε να συμπεριλάβουμε 14 συνολικά οργανισμούς, υπήρχαν 143 γονίδια που είχαν ορθόλογα και στους 14 οργανισμούς (δείτε παρακάτω Table **7**). Όταν στην φυλογενωμική μας ανάλυση θελήσαμε να συμπεριλάβουμε 7 συνολικά οργανισμούς (έναν μόνο αντιπρόσωπο από κάθε σημαντική εξελικτική ομάδα), τότε υπήρχαν 785

γονίδια που είχαν ορθόλογα και στους 7 οργανισμούς.

No	*Ορθολ.Γονίδια(acces.num.)*	*Λειτουργία*
1	EKC18676	UPF0661 TPR repeat-containing protein C16D10.01c
2	EKC27782	AP-1 complex subunit sigma-2
3	EKC19429	Zinc finger protein ZPR1
4	EKC19642	Nucleolar GTP-binding protein 1
5	EKC28616	Splicing factor, arginine/serine-rich 4
6	EKC24761	Have not found
7	EKC25822	Catalase
8	EKC26938	NADH dehydrogenase
9	EKC26790	Transport protein Sec61 subunit alpha isoform 2
10	EKC29307	FAM50-like protein
11	EKC26369	AP-3 complex subunit delta-1
12	EKC26388	Gamma-aminobutyric acid receptor-associated protein
13	EKC30683	Mitochondrial-processing peptidase subunit beta
14	EKC30690	Troponin T, skeletal muscle
15	EKC29991	Proteasome subunit alpha type-4
16	EKC29663	Protein disulfide-isomerase
17	EKC29671	40S ribosomal protein S9
18	EKC28502	Cold shock domain-containing protein E1
19	EKC20919	Putative 39S ribosomal protein L24, mitochondrial
20	EKC19106	40S ribosomal protein S5
21	EKC25786	Leucine-zipper-like transcriptional regulator 1
22	EKC29001	Malate dehydrogenase, mitochondrial
23	EKC28935	Signal transducing adapter molecule 2
24	EKC26210	Putative phosphoglycerate mutase
25	EKC20496	Serologically defined colon cancer antigen 1-like protein
26	EKC19324	Splicing factor 3A subunit 3
27	EKC19328	Eukaryotic translation initiation factor 2 subunit 3, Y-linked
28	EKC17550	ribosomal protein L12
29	EKC19012	Lysosomal aspartic protease
30	EKC27411	ADP,ATP carrier protein

31	EKC21190	T-complex protein 1 subunit zeta
32	EKC24037	26S protease regulatory subunit 6A
33	EKC24065	39S ribosomal protein L38, mitochondrial
34	EKC24079	Have not found
35	EKC29343	Eukaryotic translation initiation factor 2 subunit 2
36	EKC17829	Ribosomal protein L21
37	EKC25158	Malate dehydrogenase
38	EKC20975	Pre-mRNA-splicing factor SLU7
39	EKC30446	HEAT repeat-containing protein 1
40	EKC27484	NADH dehydrogenase
41	EKC21628	Triosephosphate isomerase
42	EKC18261	Integral membrane protein 2A
43	EKC23652	40S ribosomal protein S26
44	EKC25464	60S ribosomal protein L14
45	EKC30295	Brain protein 16
46	EKC30320	40S ribosomal protein S11
47	EKC24756	Ectoine hydroxylase
48	EKC24666	DnaJ-like protein subfamily C member 2
49	EKC23857	Proteasome subunit beta type
50	EKC23867	Non-selenium glutathione peroxidase\x3b Peroxiredoxin-6
51	EKC30426	39S ribosomal protein L44, mitochondrial
52	EKC28543	NADH dehydrogenase
53	EKC21733	Adipophilin
54	EKC29597	SWI/SNF complex subunit SMARCC2
55	EKC29598	Coatomer subunit zeta-1
56	EKC29767	Guanine nucleotide-binding protein-like 3-like protein
57	EKC29780	26S protease regulatory subunit 4
58	EKC18743	Stress-induced-phosphoprotein 1
59	EKC20508	Lupus La-like protein
60	EKC20990	Complement component 1 Q subcomponent-binding protein, mitochondrial
61	EKC18281	Peptidyl-prolyl cis-trans isomerase
62	EKC20231	Calmodulin
63	EKC20238	Tubulin-folding cofactor B
64	EKC21442	Septin-7
65	EKC20167	60S ribosomal protein L4
66	EKC25665	NADH dehydrogenase
67	EKC18078	Kelch domain-containing protein 4
68	EKC27694	Eukaryotic translation initiation factor 5
69	EKC23295	Glyceraldehyde-3-phosphate dehydrogenase
70	EKC19336	Eukaryotic translation initiation factor 3 subunit G-A
71	EKC23492	DNA excision repair protein ERCC-1
72	EKC23503	NADH dehydrogenase
73	EKC22590	Putative rRNA-processing protein EBP2
74	EKC30127	Histidine triad nucleotide-binding protein 1
75	EKC17390	ATP-dependent RNA helicase DDX56
76	EKC19222	Sperm-associated antigen 1
77	EKC28150	Succinyl-CoA ligase
78	EKC18045	Bifunctional protein NCOAT
79	EKC30192	THO complex subunit 1
80	EKC30200	Serine/threonine-protein kinase RIO1

81	EKC30210	Aconitate hydratase, mitochondrial
82	EKC25698	Cytochrome c oxidase subunit 5A, mitochondrial
83	EKC25710	Protein BTG1
84	EKC19760	Serine/threonine-protein kinase Pim-3
85	EKC29329	Nucleoredoxin
86	EKC26112	Cytoplasmic dynein 1 intermediate chain 2
87	EKC24376	Protein AATF
88	EKC22548	Putative ribosomal RNA methyltransferase NOP2
89	EKC23900	Transforming growth factor-beta-induced protein ig-h3
90	EKC23905	Calreticulin
91	EKC20118	Tax1-binding protein 1-like protein B
92	EKC20119	Microtubule-associated serine/threonine-protein kinase-like protein
93	EKC30472	Neutral and basic amino acid transport protein rBAT
94	EKC30477	Nucleolar protein 58
95	EKC19905	RRP12-like protein
96	EKC29910	Sulfurtransferase
97	EKC26844	Aldo-keto reductase family 1 member B10
98	EKC27987	Radixin
99	EKC27989	PIH1 domain-containing protein 1
100	EKC28114	26S proteasome non-ATPase regulatory subunit 7
101	EKC22747	Tribbles-like protein 2
102	EKC18239	Y-box factor-like protein
103	EKC25062	NADH dehydrogenase
104	EKC19371	V-type proton ATPase subunit F
105	EKC27073	Sperm surface protein Sp17
106	EKC30968	RNA-binding protein 28
107	EKC20023	40S ribosomal protein S8
108	EKC28714	Hydroxysteroid dehydrogenase-like protein 2
109	EKC28361	BCCIP-like protein
110	EKC30364	Ankyrin repeat domain-containing protein 45
111	EKC20273	Spliceosome RNA helicase BAT1
112	EKC27637	BRCA1-A complex subunit BRE
113	EKC20435	S-adenosylmethionine synthase
114	EKC17881	ribosomal protein L8
115	EKC31203	p21-activated protein kinase-interacting protein 1-like protein
116	EKC21221	Microfibrillar-associated protein 1
117	EKC18102	mRNA turnover protein 4-like protein
118	EKC28224	ETS-related transcription factor Elf-3
119	EKC28232	RuvB-like 2
120	EKC31151	DNA ligase

121	EKC19714	Hydroxyacyl-coenzyme A dehydrogenase, mitochondrial
122	EKC19603	Nuclear factor erythroid 2-related factor 2
123	EKC24582	Have not found
124	EKC18214	Mitotic spindle assembly checkpoint protein MAD1
125	EKC21617	Severin
126	EKC28337	Proteasome subunit beta type
127	EKC24477	Presenilin-2
128	EKC24481	26S protease regulatory subunit 6B
129	EKC28777	Protein SET
130	EKC28782	Transcription initiation factor TFIID subunit 1
131	EKC19731	Cytochrome c1, heme protein, mitochondrial
132	EKC25184	60S ribosomal protein L3
133	EKC29138	FACT complex subunit SSRP1
134	EKC18892	Pre-mRNA-splicing factor syf2
135	EKC18898	Phospholipase D1
136	EKC26686	Phosphate carrier protein, mitochondrial
137	EKC30771	Eukaryotic translation initiation factor 3 subunit B
138	EKC21427	Polycystin-2
139	EKC26094	Blastula protease 10
140	EKC18189	Protein arginine N-methyltransferase 1
141	EKC17230	Eukaryotic initiation factor 4A-II
142	EKC20311	Myosin regulatory light chain A, smooth adductor muscle
143	EKC23947	40S ribosomal protein S10\x3b Ribosomal protein S10

Table 7. Λίστα με τα 143 γονίδια του Crassostrea gigas, που χρησιμοποιήθηκαν στη φυλογενετική ανάλυση των 14 οργανισμών (τα γονίδια αναφέρονται με το accession number τους).

4.2.2 Πολλαπλή στοίχιση και κατασκευή φυλογενωμικών δέντρων

Στη συνέχεια έγινε πολλαπλή στοίχιση των πρωτεϊνων κάθε οργανισμού με τον αλγόριθμο Muscle μέσα στο πρόγραμμα Seaview και έπειτα, οι επιμέρους πολλαπλές στοιχίσεις από τα 143 και 785 γονίδια ενώθηκαν σε μία υπέρ-πολλαπλή στοίχιση αντίστοιχα. Έπειτα, η κάθε μία υπέρ-πολλαπλή στοίχιση φιλτραρίστηκε για καλά συντηρημένες περιοχές με το πρόγραμμα Gblocks, εκτελώντας την παρακάτω εντολή:

./Gblocks concatenate_fasta_aln -t=p -b4=8 -b5=n

-t: τύπος αλληλουχίας
-b4: το μικρότερο μέγεθος του block
-b5: κενά ανάμεσα στα blocks, στη περίπτωση αυτής της ανάλυσης επιλέχθηκε να μην επιτρέπονται

Η πρώτη υπέρ-πολλαπλή στοίχιση των 143 γονιδίων για 14 οργανισμούς περιείχε 87.451 θέσεις, ενώ μετά από το φιλτράρισμα με το πρόγραμμα Gblocks οι καλά συντηρημένες θέσεις μειώθηκαν στις 11.351. Αντίστοιχα, η δεύτερη υπέρ-πολλαπλή στοίχιση των 785 γονιδίων για 7 οργανισμούς περιείψε 477.514 θέσεις, ενώ μετά τη χρήση του Gblocks οι καλά συντηρημένες θέσεις ήταν 90.825. Πρέπει να σημειωθεί πως οι παράμετροι που χρησιμοποιήθηκαν για το Gblocks ήταν αυστηρές.

Στο τελικό στάδιο δημιουργήθηκαν τα φυλογενετικά δέντρα για τις δύο υπέρ-πολλαπλές στοιχίσεις και στη συνέχεια έγινε η απεικόνιση τους. Αρχικά επιλέχθηκε η δημιουργία φυλογενετικών δέντρων με τη μέθοδο BioNJ (μέθοδος απόστασης), μοντέλο Poisson και 500 Bootstrap. Η Lottia gigantea χρησιμοποιήθηκε ως 'out-group'. Τα δύο δέντρα αποθηκεύτηκαν σε μορφή .pdf αλλά και σαν αρχεία .nwk.

Στη συνέχεια, δημιουργήθηκαν δύο φυλογενετικά δέντρα μέσω του προγράμματος PhyML, όπου επιλέχθηκε το μοντέλο WAG, τέσσερις κατηγορίες εξελικτικών ρυθμών (4 rate categories) και αναζήτηση του καλύτερου φυλογενετικού δέντρου με την μέθοδο SPR. Η αξιολόγιση των επιμέρους κάδων έγινε με το approximate likelihood ratio test (aLRT). Τα δέντρα που δημιουργήθηκαν αποθηκεύτηκαν σε μορφή .pdf και στη συνέχεια σε .nwk.

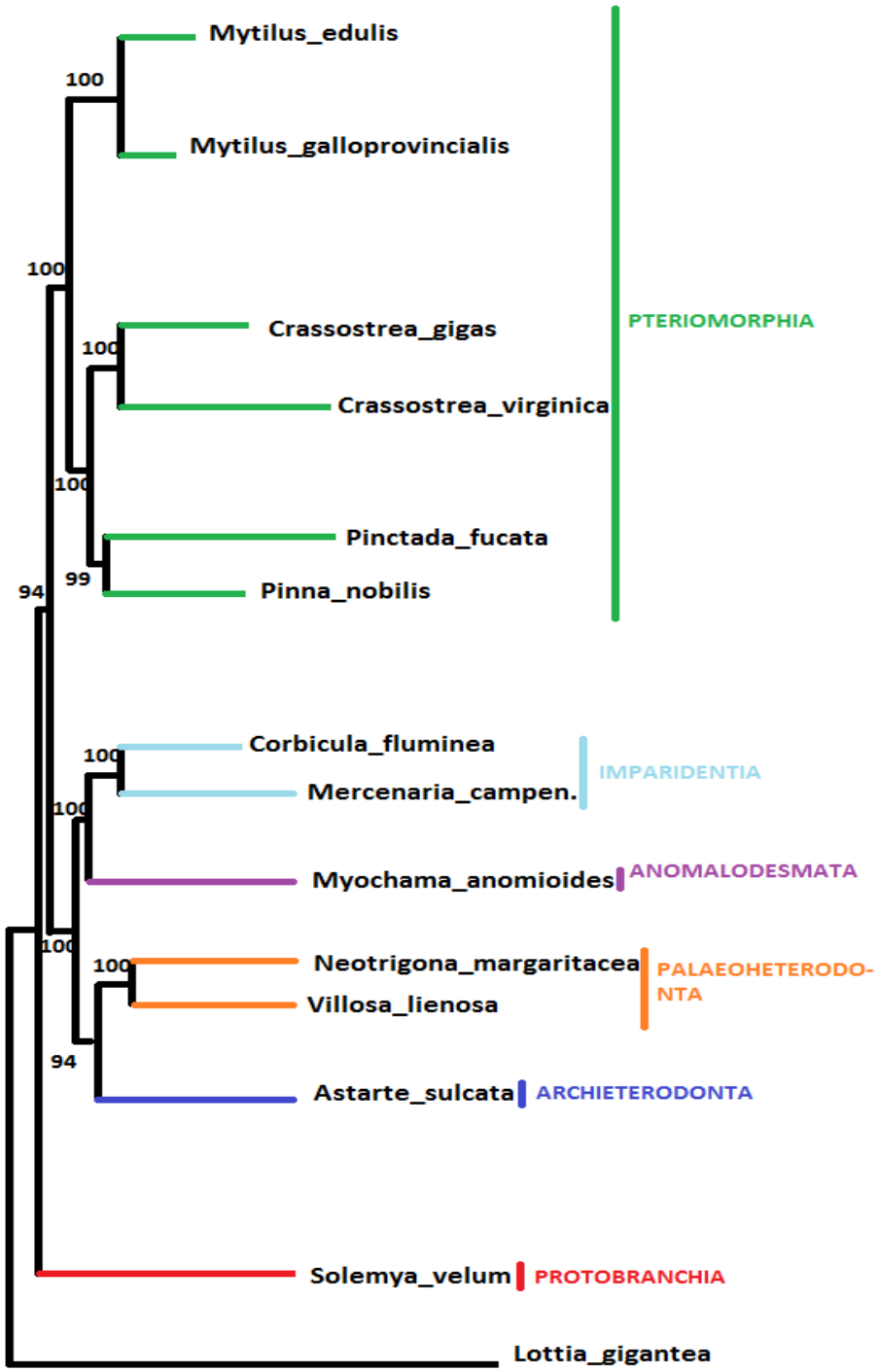


Figure 33. Φυλογενετικό δέντρο με τη μέθοδο απόστασης BioNJ, μοντέλο Poisson και 500 bootstraps για τα 143 γονίδια από 14 οργανισμούς.

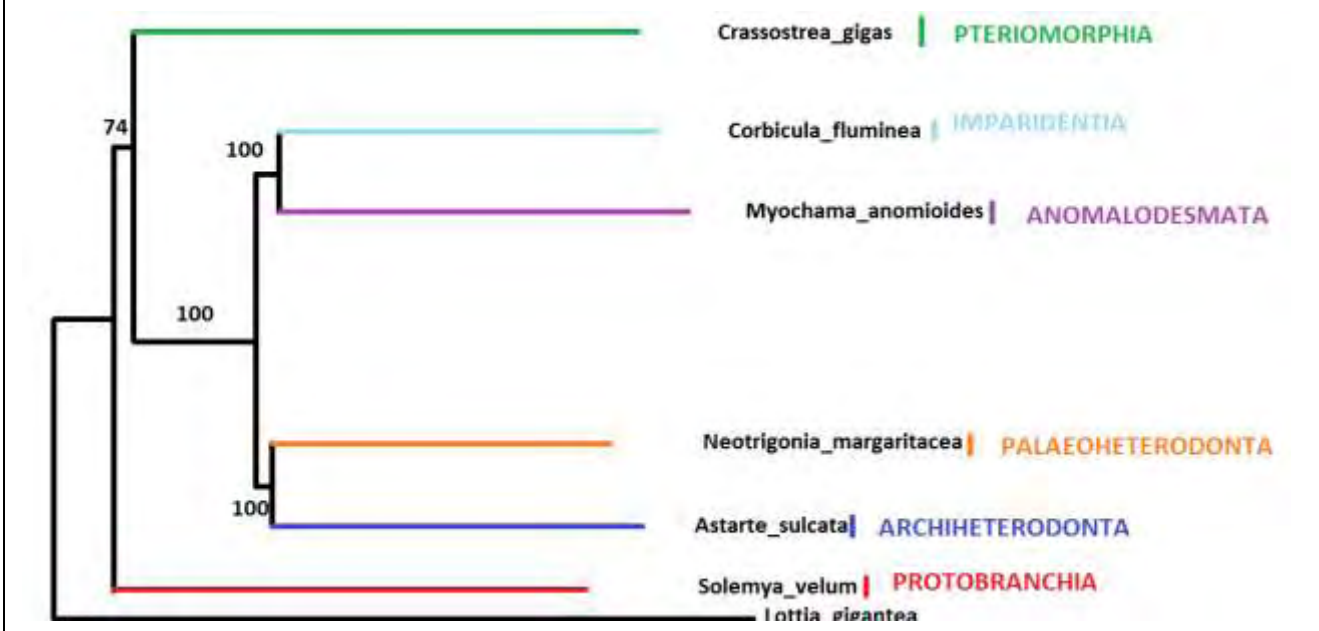


Figure 34. Φυλογενετικό δέντρο με τη μέθοδο απόστασης BioNJ, μοντέλο Poisson και 500 bootstraps για τα 785 γονίδια από 7 οργανισμούς.

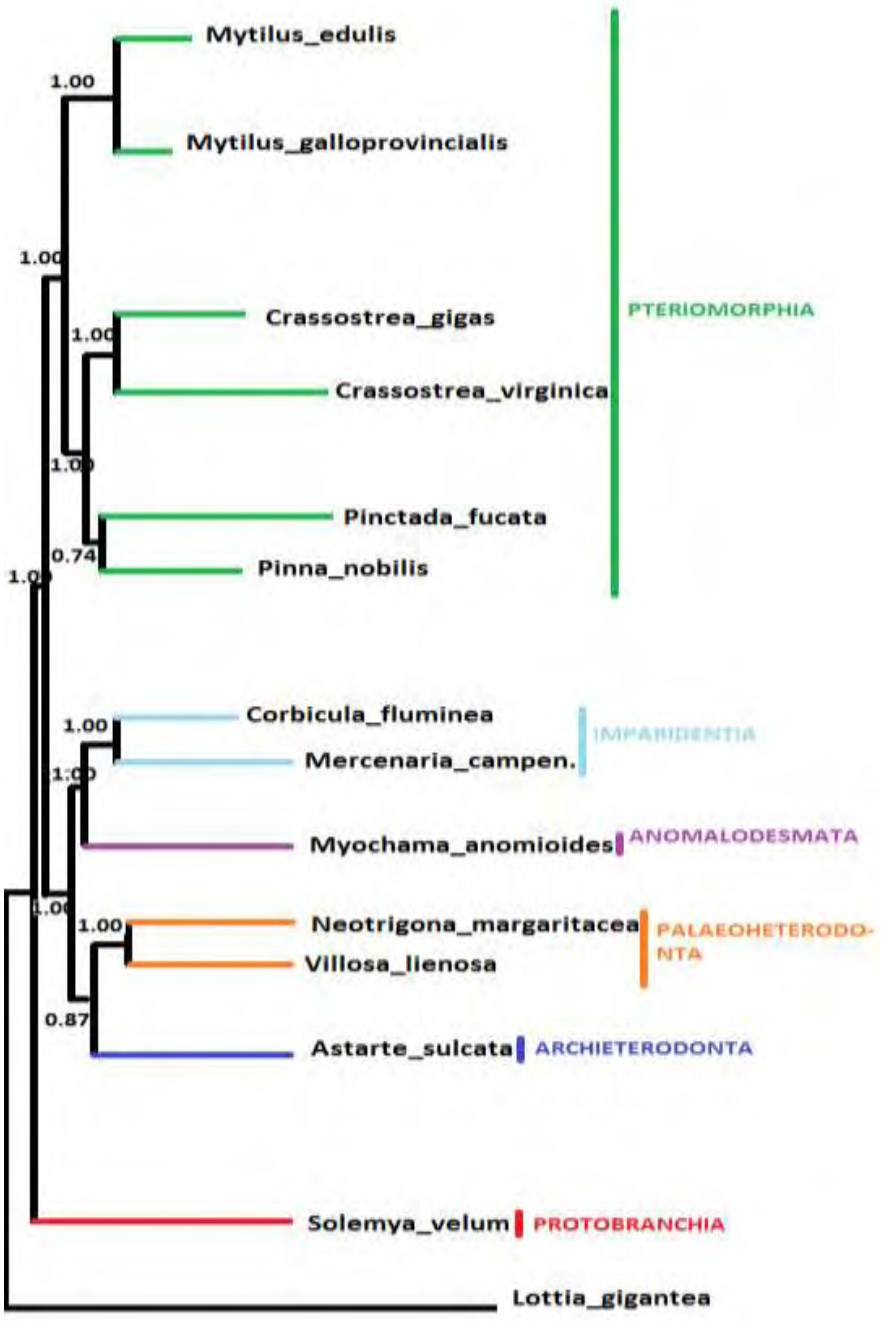


Figure 35. Φυλογενετικό δέντρο με τη μέθοδο χαρακτήρων Maximum Likelihood, μοντέλο WAG, 4 crate categories, SPR, aLRT για τα 143 γονίδια από 14 οργανισμούς.

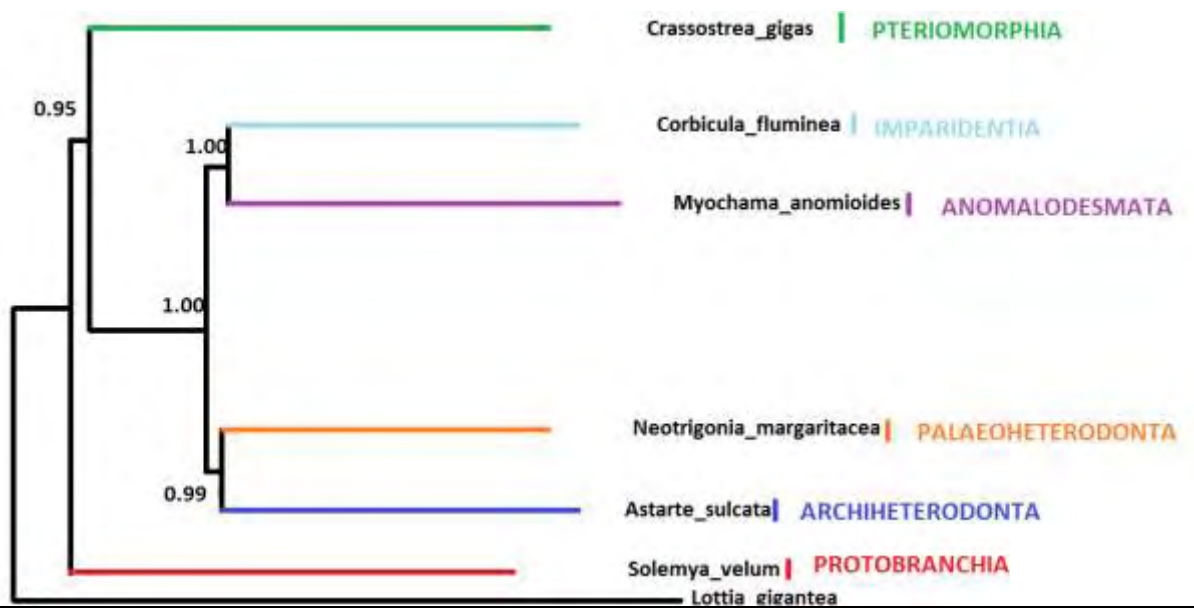


Figure 36. Φυλογενετικό δέντρο με τη μέθοδο χαρακτήρων Maximum Likelihood, μοντέλο WAG, 4 crate categories, SPR, aLRT για τα 785 γονίδια από 7 οργανισμούς.

5 ΣΥΖΗΤΗΣΗ

Η παρούσα εργασία είχε ως σκοπό να μπορέσει να εξιχνιάσει τις φυλογενετικές σχέσεις μεταξύ των οργανισμών της εξελικτικής ομάδας των Δίθυρων, καθώς και να καθορίσει τη φυλογενετική θέση του οργανισμού Pinna nobilis, που είναι το μεγαλύτερο δίθυρο της Μεσογείου. Στην ομάδα των διθύρων ανήκουν πολλά οικονομικά σημαντικά είδη. Ο απώτερος σκοπός ήταν να χρησιμοποιηθεί η νέα αυτή και πολύ πιο αξιόπιστη φυλογένεση για να μπορέσουμε να μελετήσουμε περαιτέρω γονίδια που εμπλέκονται στην ανάπτυξη των διθύρων και που θα μπορούσαν να αποτελέσουν στόχους για βιοτεχνολογικές εφαρμογές. Με βάση μια αξιόπιστη φυλογένεση θα μπορέσουμε στο μέλλον να εκτιμήσουμε κατά πόσον τα αποτελέσματα μια μοριακής μελέτη που πραγματοποιείται σε ένα δίθυρο μπορούν ή όχι να επεκταθούν και σε άλλα είδη αυτής της ομάδας.

Για τον σκοπό αυτό χρησιμοποιήθηκαν δεδομένα αλληλούχισης νέας γενιάς (RNA-SEQ), δημοσιευμένα γονιδωματικά δεδομένα και σύγρονες βιοπληροφορικές/γονιδιωματικές αναλύσεις. Τελικά, χρησιμοποιήθηκαν δύο ομάδες φυλογενωμικών δεδομένων. Στην μια ομάδα αναλύθηκαν 143 γονίδια που είχαν ορθόλογα σε 14 οργανισμούς (13 δίθυρα και 1 γαστρόποδο-εξωομάδα) και στην δεύτερη ομάδα αναλύθηκαν 785 γονίδια που είχαν ορθόλογα σε 7 οργανισμούς (6 δίθυρα, ένα αντιπρόσωπο από κάθε σημαντική εξελικτική ομάδα και 1 γαστρόποδο-εξωομάδα) Αυτή η άνευ προηγουμένου χρήση ενός τόσο μεγάλου όγκου δεδομένων αναλύθηκε με τις πιο σύγχρονες βιοπληροφορικές μεθόδους και οδήγησε σε πολύ πιο αξιόπιστα αποτελέσματα τα οποία συγκρίνονται με πρόσφατες αναλύσεις από άλλες ερευνητικές ομάδες. Παλαιότερα αποτελέσματα άλλων μελετών είναι αντικρουόμενα και αντικαθίστανται από πιο μοντέρνα δεδομένα και αναλύσεις.

Για παράδειγμα, το NCBI taxonomy αναφέρει πέντε εξελικτικές ομάδες των Δίθυρων, τα: 1) Heteroconchia, 2) Paleoheterodonta, 3) Anomalodesmata, 4)

Pteriomorphia, 5) Protobranchia.

Η μελέτη των Bieler *et al.*, 2014 αποτελεί την πιο πρόσφατη μελέτη που ενσωματώνει ένα πλήθος μορφολογικά δεδομένα μαζί με μοριακά. Συγκεκριμένα, μελετήθηκαν 103 δίθυρα είδη ως προς τα μορφολογικά τους χαρακτηριστικά ενώ επίσης έγιναν μοριακές φυλογενέσεις και για 5 γονίδια δείκτες (Figure **37**). Για ένα μέρος από τα 103 δίθυρα ενσωματώθηκαν μοριακά δεδομένα από μια άλλη μελέτη που χρησιμοποίησε 4 άλλα γονίδια δείκτες και έτσι για αυτό το υποσύνολο ειδών οι μοριακές φυλογενέσεις βασίστηκαν σε 9 γονίδια δείκτες (Figure **38**). Οι μελέτες τους κατέληξαν στο να προτείνουν 6 μονοφυλετικές ομάδες στα δίθυρα, τα Protobranchia, Pteriomorphia, Palaeoheterodonta, Archiheterodonta, Anomalodesmata και Imparidentia.

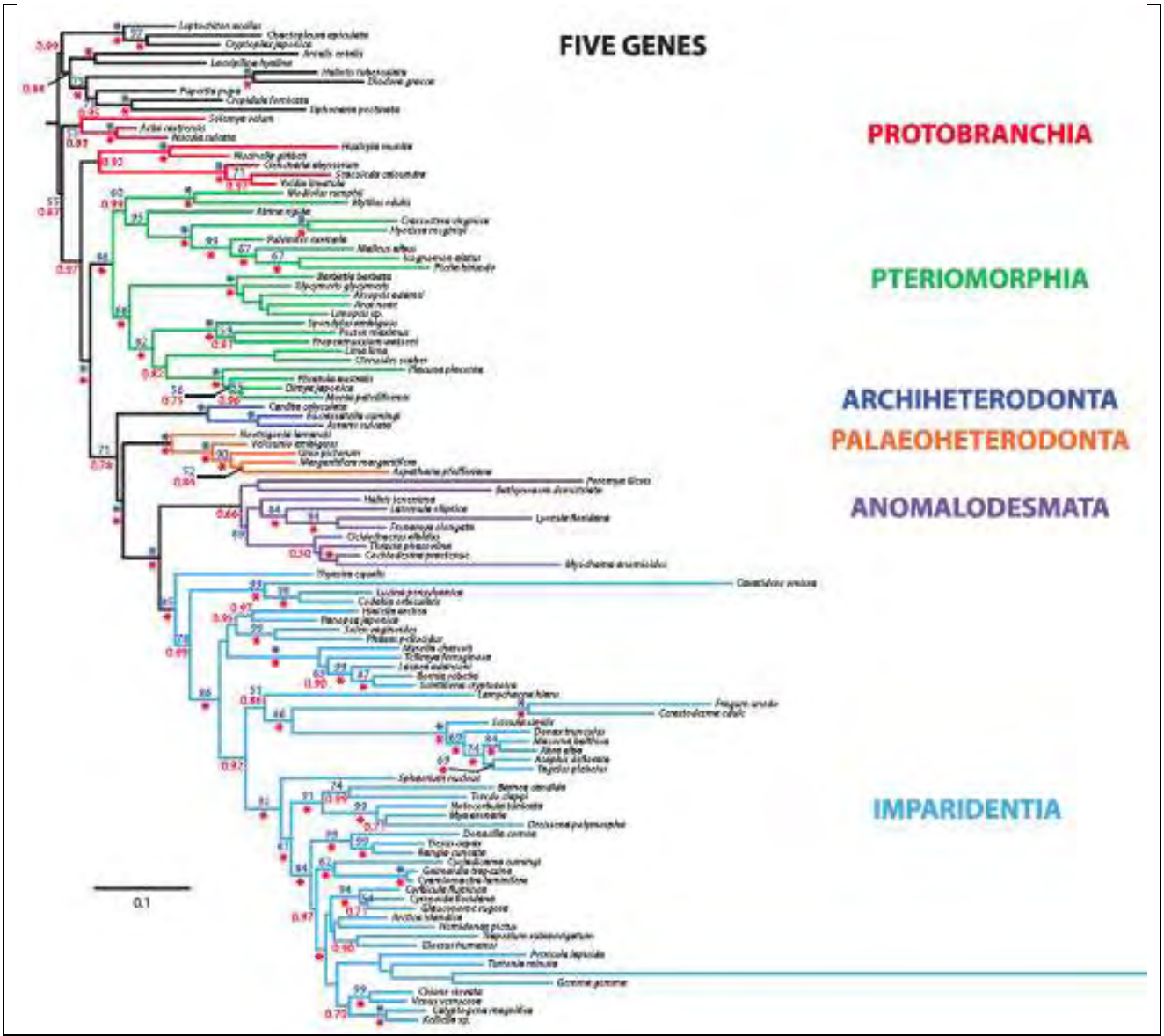


Figure 37. Φυλογενετική ανάλυση Maximum Likelihood των Bieler *et al.*, 2014 βασισμένη σε 5 γονίδια.

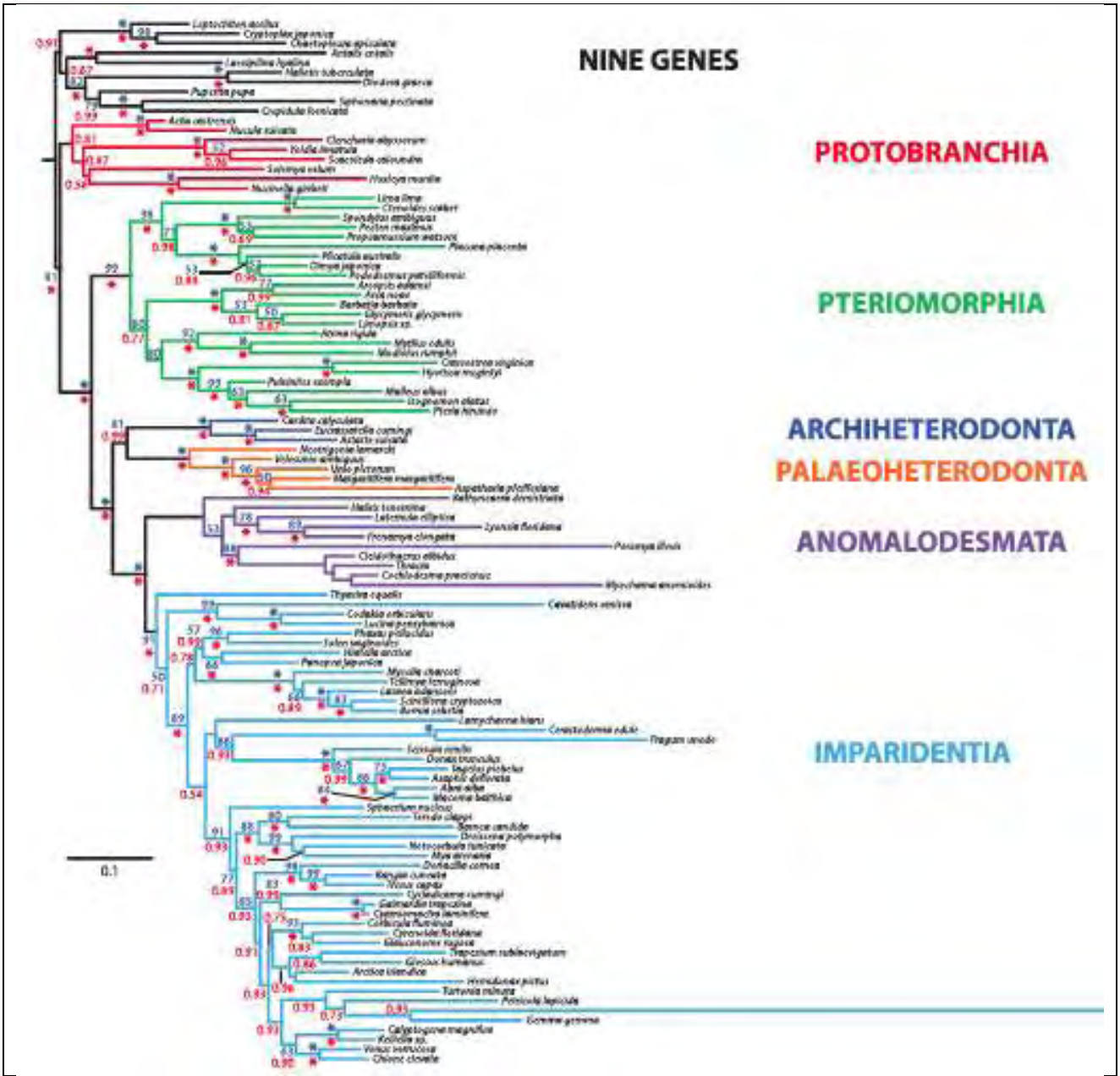


Figure 38. Φυλογενετική ανάλυση Maximum Likelihood των Bieler *et al*., 2014 βασισμένη σε 9 γονίδια.

Η πιο πρόσφατη φυλογενετική μελέτη των Gonzalez *et al.*, 2015 χρησιμοποιεί δεδομένα RNA-SEQ και πρωτεώματα από 31 δίθυρα. Τα δεδομένα τους συναρμολογήθηκαν σε γονίδια με μεθόδους πολύ όμοιες με τις δικές μας. Οστόσο, η ανίχνευση ορθόλογων γονιδίων έγινε με εντελώς διαφορετικό τρόπο, χρησιμοποιώντας το πρόγραμμα OMA που δημιουργεί supermatrices. Το φυλογενωμικό τους δέντρο φαίνεται στην (Figure **39**).

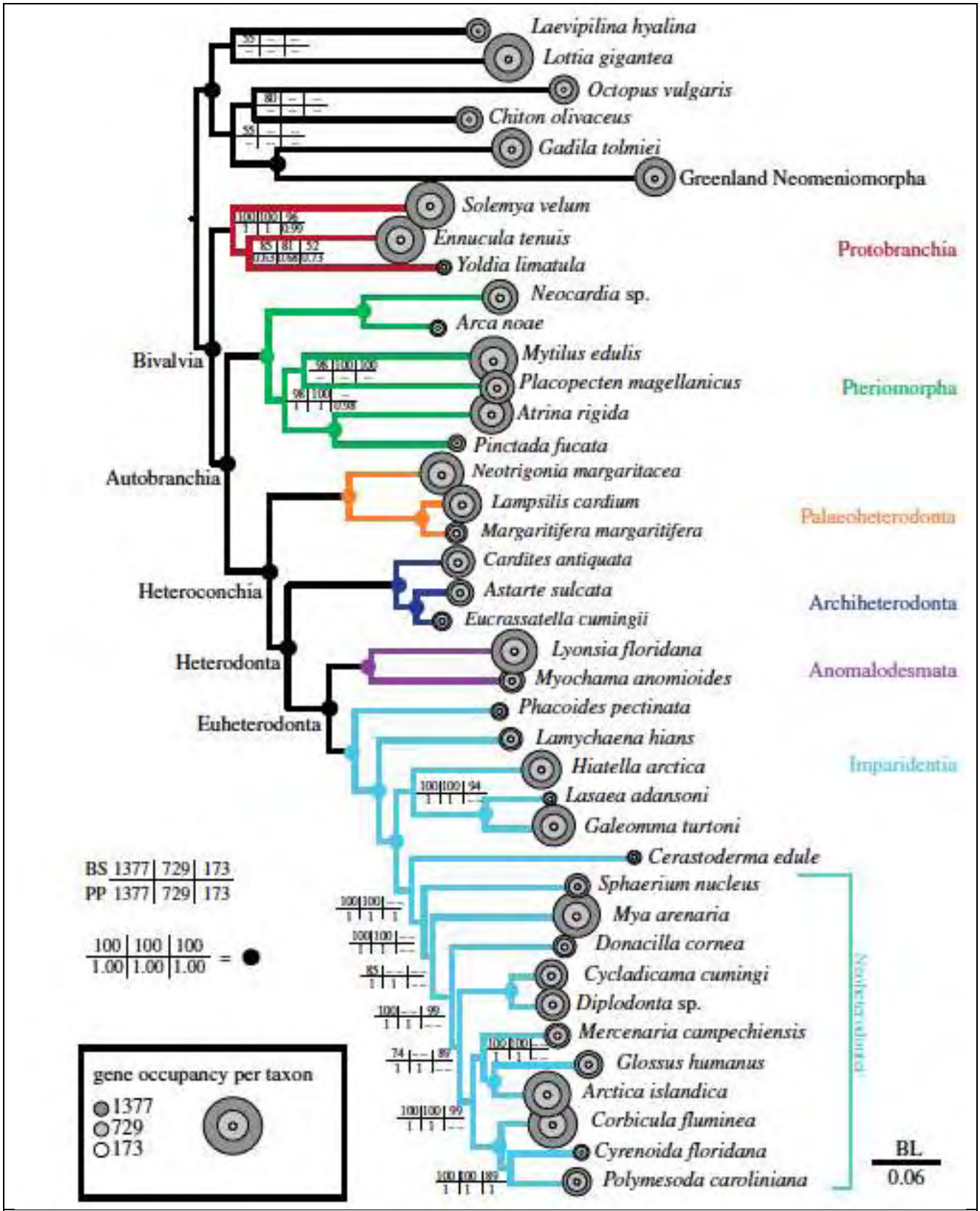


Figure 39. Φυλογενωμική ανάλυση Maximum Likelihood των Gonzalez *et al*., 2015 βασισμένη σε μέθοδο supermatrix από δεδομένα RNA-SEQ.

Τα δικά μας φυλογενωμικά δέντρα, που έγιναν με 2 διαφορετικές μεθόδους και 2 διαφορετικά σετ 143 & 785 γονιδίων για δύο διαφορετικά σετ 14 και 7 ειδών συμφωνούν απόλυτα μεταξύ τους, αλλά και με την φυλογενετική ανάλυση των 9 γονιδίων στη μελέτη των Bieler *et al.,* 2014. Η φυλογενετική ανάλυση των 5 γονιδίων στη μελέτη των Bieler *et al.,* 2014 συμφωνεί απόλυτα με την μελέτη των Gonzalez *et al.,* 2015. Η μόνη διαφορά μεταξύ των δύο αυτών αποτελεσμάτων είναι ότι στην πρώτη περίπτωση (δικές μας αναλύσεις και ανάλυση 9 γονιδίων των Bieler et al., 2014) τα Paleoheterodonta είναι αδελφά τάξα με τα Archiheterodonta ενώ στην δεύτερη περίπτωση (ανάλυση 5 γονιδίων των Bieler *et al.,* 2014 και ανάλυση των Gonzalez *et al.,* 2015) τα Paleoheterodonta αποκλίνουν πριν τα Archiheterodonta. Σε όλες τις άλλες κυριότερες εξελικτικές σχέσεις όλες οι αναλύσεις συμφωνούν μεταξύ τους.

Το χαμηλό κόστος των τεχνολογιών νέας γενιάς μαζί με την ραγδαία ανάπτυξη της Βιοπληροφορικής υπόσχεται να παραδώσει ένα τεράστιο όγκο εξελικτικών μοριακών δεδομένων, τα οποία θα μας αποκαλύψουν με εξαιρετικά μεγάλη ακρίβεια εξελικτικές διαδικασίες που συνέβησαν σε διάφορες χρονικές κλίμακες, από εκατοντάδες εκατομύρια χρόνια μέχρι πολύ πρόσφατα γεγονότα. Αυτό με την σειρά του θα μας επιτρέψει να κατανοήσουμε εις βάθος βασικούς μηχανισμούς εξέλιξης και προσαρμογής και επιπλέον θα μας επιτρέψει να κατανοήσουμε ποιές λειτουργικές μελέτες που συμβαίνουν σε ένα οργανισμό μοντέλο μιας εξελικτικές ομάδας μπορούν να έχουν εφαρμογή σε κοντινούς ή μακρινούς συγγενείς με οικονομικό ενδιαφέρον.

6 ΒΙΒΛΙΟΓΡΑΦΙΑ

Adamkewicz, S.L., Harasewych, M.G., Blake, J., Saudek, D., Bult, C.J. (1997). A Molecular Phylogeny of the Bivalve Mollusks. Mol.Biol.Evol. 14(6):619-629.

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

Anisimova, M., Gascuel, O. (2006). Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Syst Biol 55 (4): 539-552.*

Ansorge, W.J. (2009). Next-generation DNA sequencing techniques.New Biotechnology, vol. 25, Issue 4, pages 195–203.Published by Elsevier B.V.

Bieler, R, Mikkelsen, P.M. (2007) A look at the branches. Engeser 15:41, 21.

Bieler, R, Mikkelsen, P.M., Collins, T.M., Glover, E.A., Gonzáles, V.L., Graf, D.L., Giribet, G., et al. (2014). Investigating the Bivalve Tree of Life – an exemplar-based approach combining molecular and novel morphological characters. CSIRO, Invertebrate Systematics, 28, 32-115.

Balavoine, G., Adoutte, A., (1998). One or three Cambrian Radiations. Science 280, 5362, pp.397-398.

Carter, J.G, Campell, D.C., Campell, M.R., (2006). Morphological Phylogenetics of the early Bivalvia. International Congress on Bivalvia, 22-27 July 2006, Universitat Autònoma de Barcelona, Catalunya, Spain.

Comas, I., Moya, A., Gonzalez-Candelas, F. (2007). From Phylogenetics to Phylogenomics: The Evolutionary Relationships of Insect Endosymbiotic 7-Proteobacteria as a Test Case. Syst. Biol. 56(1):1-16.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.32(5):1792-1797.*

González V.L., Andrade S.C., Bieler R., Collins T.M., Dunn C.W., Mikkelsen P.M., Taylor J.D., Giribet G. (2015).A phylogenetic backbone for Bivalvia: an RNA-seq approach.Proc Biol Sci. 2015 Feb 22;282(1801).

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. (2010).New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst.Biol.59 (3): 307-321.*

Hickman, C.P., Roberts, L.S., Larson, A., (eds) 2005. Integrated Principles of Zoology. McGraw-Hill Higher Education, 11th Edition.

Huang, X., Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. Genome Res.9:868-877.

Kocot, K.M., Cannon, J.T., Todt, C., Citarella, M.R., Kohn, A.B., Meyer, A., et al. (2011). Phylogenomics reveals deep molluscan relationships. Nature 477, 452-456.

Kodama, Y., Shumway, M., Leinonen, R., International Nucleotide Sequence Database Collaboration (2012). The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 40, D54-56.

Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. Nature 470, 198–203.

Martin, J.A., Wang, Z. (2011).Next-generation transcriptome assembly.Nature Reviews Genetics 12, 671-682.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat. Rev. Genet. 11, 31–46.

Newell, N.D. (1965). Classification of the Bivalvia. American Museum of Natural History, NY, no.2206.

Patel, R.K., Mukesh, J. (2012).NGS QC Toolkit: a Toolkit for Quality Control of Next Generation Sequencing Data.PloS One 7, no. 2 : e30619.

Ronaghi, M. (2001).Pyrosequencing Sheds Light on DNA Sequencing.Genome Res. 11: 3-11.Cold Spring Harbor Laboratory Press.

Sanger, F., Nicklen, S., Coulson, A.R. (1977).DNA sequencing with chain-terminating inhibitors.Proc Natl Acad Sci U S A, 74(12): 5463–5467.

Sharma, P.P., Gonzáles, V.L., Kawauchi, G.Y., Andrade, S.C., Guzmán, A., Collins, T.M., Glover, A.E., Harper, E.M., Healy, J.M., Mikkelsen, P.M., Taylor, J.D., Bieler, R., Giribet, G. (2012). Phylogenetics analysis of four nuclear protein-encoding genes largely corroborates the traditional classification of Bivalvia (Mollusca). Mol. Phylogenet. Evol., dx.doi.org/10.1016/j.ympev.2012.05.025.

Smeds, L., and Künstner, A. (2011). ConDeTri--a content dependent read trimmer for Illumina data. PloS One 6, e26314.

Smith, S.A, Wilson, N.G, Goetz, F.E., Feehery, C., Andrade, S.C.S, Rouse, G.W., Giribet, G., Dunn, C.W., (2011). Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature 480, 364-367.

Takeushi, T., Takeshi, K., Koyanagi, R., Gyoja, F., Tanaka, M., Ikuta, T., Shoguchi, E., Satoh., N., et al. (2012). Draft Genome of the Pearl Oyster Pinctada Fucata: A Platform for Understanding Bivalve Biology. DNA Research 19, 117-130.

Venier, P., De Pitta, C., Bernante, F., Varotto, L., Lanfranchi, G., et al. (2009). MytiBase: a knowledgebase of mussel transcribed sequences. BMC Genomics, 10:72, 1471-2164.

Wang, Z., Gerstein, M., Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10, 57-63.
Wiley, E.O., Chakrabarty, P., Craig, M.T., Davis, M.P., Holcroft, N.I., Mayden, R.L., Smith, WM.L. (Eds) (2011). Will the Real Phylogeneticists Please Stand Up?. Zootaxa, 2946, 1–142.

Zapata, F., Wilson, N.G., Howison, M., et al., (2014). Plylogenomic analyses of deep gastropod relationships reject Orthogastropoda. BioRxiv, 10.1101/007039.

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Wang, J., et al. (2012). The oyster genome reveals stress adaptation and complexity of shell formation. Nature 11413 490, 49.

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., and Yu, J. (2010). The next-generation sequencing technology and application. Protein Cell 1, 520–536.