ΜΠΣ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

ΜΠΕΛΙΑΣ ΜΙΧΑΗΛ

Χρονολόγηση γεγονότων διασποράς επιδημίας HIV-1 υποτύπου Β με μεθόδους Μπεϋζιανής συμπερασματολογίας

ΑΘΗΝΑ, ΕΤΟΣ
Οκτώβριος 2016

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη ΒΙΟΣΤΑΤΙΣΤΙΚΗ που απονέμει η Ιατρική Σχολή και το Τμήμα Μαθηματικών του Εθνικού & Καποδιστριακού Πανεπιστημίου Αθηνών

Εγκρίθηκε την……. από την εξεταστική επιτροπή:

Δ. ΠΑΡΑΣΚΕΥΗΣ	**ΕΠ.ΚΑΘΗΓΗΤΗΣ**	**……**
Γ. ΜΑΓΙΟΡΚΙΝΗΣ	**ΛΕΚΤΟΡΑΣ**	**……**
Ν.ΠΑΝΤΑΖΗΣ	**ΛΕΚΤΟΡΑΣ**	**……**

Εισαγωγή

Η γενετική πληροφορία όλων των ζωντανών οργανισμών είναι αποθηκευμένη στα νουκλεϊνικά οξέα (DNA ή RNA για κάποιους ιούς), που θεωρούνται ως οι γενετικές οδηγίες και χρησιμοποιούνται από τον οργανισμό με σκοπό την ανάπτυξη, τη λειτουργία και την αναπαραγωγή του. Τα νουκλεϊνικά οξέα αποτελούνται από νουκλεοτιδικές βάσεις.

Υπάρχουν πέντε νουκλεοτιδικές βάσεις:

- Αδενίνη
- Κυτοσίνη
- Γουανίνη
- Θυμίνη
- Ουρακίλη

Οι πρώτες τρεις είναι κοινές και στα RNA και στα DNA μόρια, η Θυμίνη εμφανίζεται μόνο στο DNA, ενώ η Ουρακίλη σε μόρια RNA.

Επιδημιολογία

Η Επιδημιολογία είναι ο επιστημονικός κλάδος, ο οποίος έχει ως στόχο να διερευνήσει την κατανομή και την αιτιολογία των παραγόντων που σχετίζονται με τα νοσήματα, τη συχνότητα και την κατανομή των νοσημάτων στους πληθυσμούς, την αξιολόγηση των υπηρεσιών υγείας και, επιπλέον, να εφαρμόσει τα ευρήματα ερευνών με σκοπό τον πρόληψη και αντιμετώπιση των νοσημάτων και γενικότερα τη βελτίωση της δημόσιας υγείας. Πολλές μέθοδοι

μπορούν να χρησιμοποιηθούν για τη διεξαγωγή επιδημιολογικών ερευνών: περιγραφικές μελέτες και μελέτες παρατήρησης χρησιμοποιούνται για να διερευνηθεί η κατανομή και αναλυτικές μελέτες για την αποσαφήνιση των καθοριστικών παραγόντων της νόσου (World Health Organization, 2016).

Μια κατηγορία επιδημιολογικής έρευνας, είναι η μοριακή επιδημιολογία, η οποία ορίζεται ως η μελέτη γενετικών και περιβαλλοντικών παραγόντων για τη διερεύνηση της πιθανής συσχέτισης τους με νοσήματα σε ανθρώπινους πληθυσμούς. Η μοριακή επιδημιολογία θεωρείται ένας συνδυασμός της μοριακής βιολογίας και της επιδημιολογίας. Η συμβολή των μοριακών επιδημιολογικών μελετών έχει ζωτική σημασία για τη βελτίωση της δημόσιας υγείας, δεδομένου ότι μπορεί να συμβάλλει στη βελτίωση της μελέτης της παθογένειας της νόσου αλλά και στην επιδημιολογική επιτήρηση της νόσου (Foxman, 2001). Επίσης η μοριακή επιδημιολογία βρίσκει εφαρμογές στη μελέτη επιδημιών από ταχέως εξελισόμενα παθογόνα όπως οι ιοί. Αξίζει να σημειωθεί ότι τα αποτελέσματα μοριακών επιδημιολογικών ερευνών συνέβαλαν στην αναγνώριση της προέλευσης και των διαδρομών μετάδοσης πληθώρας λοιμωδών νοσημάτων, όπως επίσης και στην αξιολόγηση της αποτελεσματικότητας των προληπτικών μέτρων τα οποία έχουν διενεργηθεί (λ.χ. εμβολιασμοί).

Human Immunodeficiency Virus (HIV)

Ο ιός της ανθρώπινης ανοσοανεπάρκειας είναι ένας λεντιϊός (υποκατηγορία των ρετροϊών), σχετίζεται μετο σύνδρομο της επίκτητης ανοσοανεπάρκειας (AIDS) (Weiss, 1993). Ο HIV αποτελείται από δύο μεγάλες κατηγορίες, που κωδικοποιούνται ως HIV-1 και HIV-2. Και οιδύο θεωρούνται αποτέλεσμα μεταδόσεων του ιού SIV που μολύνει πρωτεύοντα θηλαστικά της Αφρικής (Sharp και Hahn, 2011).

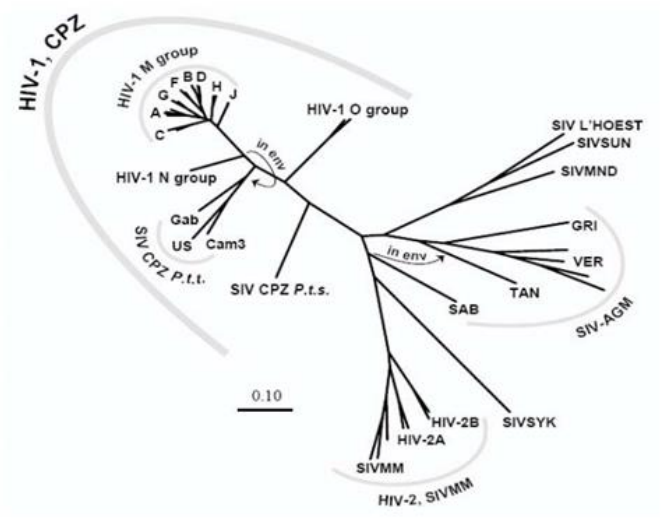
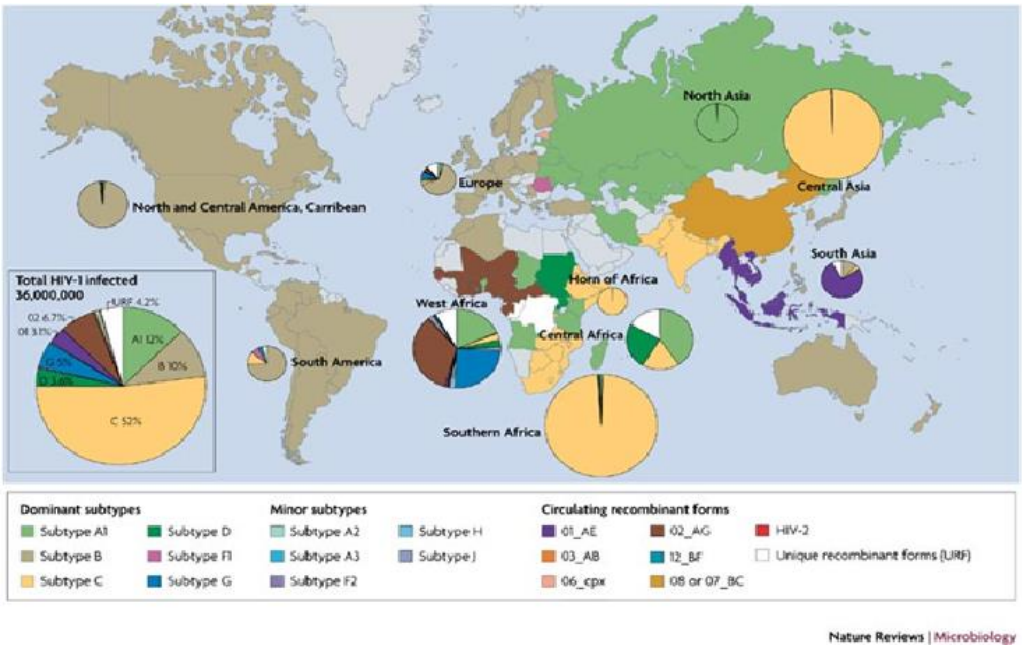


Fig 1.1 Φυλογενετική σχέση λεντιιών πρωτεόντων θηλαστικών. Το pol γονίδιο χρησιμοποιήθηκε για την κατασκευή του δέντρου.(Bette Korber and Watkins, n.d.)

Οι εν λόγω μεταδόσεις φαίνεται να προϋπήρχαν από την αρχή του 20ου αιώνα και είχαν συνήθως μια περιορισμένη έκταση. Ωστόσο, μια μετάδοση στο νότιο-ανατολικό Καμερούν φαίνεται να δημιούργησε την ομάδα Μ του HIV-1 (Greene, 2007), η οποία είναι η κύρια αιτία της μεγάλης επιδημίας HIV-1 που ταυτοποιήθηκε για πρώτη φορά στις αρχές της δεκαετίας του '80. Αντιθέτως, ο HIV-2 και άλλες μικρές ομάδες του HIV-1 (ομάδα Ν, Ο και Ρ) είχαν περιορισμένη εξάπλωση και μπορούν να εντοπιστούν ως επί το πλείστον σε μικρές επιδημίες της Κεντρικής Αφρικής. Όπως φαίνεται στο σχήμα 1.1, όλες αυτές οι ομάδες μπορούν να θεωρηθούν ως ανεξάρτητες μεταδόσεις του SIV σε ανθρώπους.

Περισσότερο από το 90% των κρουσμάτων του HIV οφείλονται σε στελέχη της ομάδας Μ του HIV-1, η οποία, σε πρόσφατη έρευνα, εκτιμάται ότι έχει μολύνει τουλάχιστον 60 εκατομμύρια ανθρώπους και έχει προκαλέσει περισσότερους από 25 εκατομμύρια θανάτους, από την αρχή της επιδημίας. Η ομάδα Μ HIV-1

διαχωρίζεται περαιτέρω σε 10 υποτύπους οι οποίοι έχουν λάβει το όνομά τους από τα πρώτα γράμματα του λατινικού αλφαβήτου (από A έως J), και θεωρείται ότι ισαπέχουν γενετικά η μία από την άλλη. Αρχικά, η ταξινόμηση των HIV-1 υπότυπων βασίστηκε σε περιοχές που είχαν επιλεχθεί από συγκεκριμένα γονίδια. Η μέση εντός-υποομάδας γενετική μεταβλητότητα για τα γονίδια gag, env και pol του HIV είναι περίπου 15%, 25% και 10% αντίστοιχα (Camacho R., 2006). Ωστόσο, μετά την αύξηση του αριθμού των ερευνών σε ικά στελέχη, ως αποτέλεσμα της φθηνότερης και ταχύτερης μεθόδου αλληλούχισης, πλέον η ταξινόμηση της φυλογένειας του HIV-1 βασίζεται σε νουκλεοτιδικές αλληλουχίες από πολλαπλά γονίδια και μερικές φορές ολόκληρο το γονιδίωμα, αποκαλύπτοντας σε πολλές περιπτώσεις ανασυνδυασμούς (Peeters, 2001).



Σχήμα 1.2 Η συχνότητα του κάθε HIV-1 υποτύπου και οι ανασυνδυασμένες μορφές εκτιμήθηκαν σε κάθε χώρα με βάση δημοσιευμένα ευρήματα. Οι χώρες έχουν χρωματική κωδικοποίηση με βάση τον κυρίαρχο υπότυπο HIV-1 της ομάδας M. Οι χώρες με γκρι χρώμα έχουν χαμηλό επίπεδο επιπολασμού του HIV-1 ή δεν έχουν εκπροσωπηθεί στην επιστημονική βιβλιογραφία που σχετίζεται με τον ιό HIV-1. Τα διαγράμματα πίτας απεικονίζουν το ποσοστό κάθε υποτύπου ή ανασυνδυασμένης μορφής σε κάθε γεωγραφική περιοχή. Το μέγεθος των πιτών

είναι ανάλογο με τον αριθμό των HIV-1 μολυσμένων ατόμων στη συγκεκριμένη περιοχή. (Arien et al., 2007).

Σε πρόσφατες μελέτες παγκόσμιας κλίμακας, οι κύριοι υπότυποι του HIV-1 είναι οι A, B και C, με τον C να καλύπτει περισσότερο από το ήμισυ των παγκόσμιων μολύνσεων.

Όπως φαίνεται στο Σχήμα 1.2:

• Ο υπότυπος A είναι κοινός στη Δυτική Αφρική και τη Ρωσία.

• Ο υπότυπος B είναι η κυρίαρχη μορφή της επιδημίας στην Ευρώπη, τη Βόρεια Αφρική, την Αμερικανική ήπειρο, την Ιαπωνία και την Αυστραλία.

• Ο υπότυπος C είναι η κυρίαρχη μορφή στη Νότια Αφρική, την Ανατολική Αφρική, την Ινδία, το Νεπάλ και τα μέρη της Κίνας.

•Ο υπότυπος D βρίσκεται στην Ανατολική και Κεντρική Αφρική.

•Ο υπότυπος CRF01_AE βρίσκεται στη Νοτιοανατολική Ασία και θεωρείται η κυρίαρχη μορφή για ετεροφυλόφιλους.

•Ο υπότυπος F έχει σημαντική παρουσία στην κεντρική Αφρική, τη Νότια Αμερική και την Ανατολική Ευρώπη.

• Ο υπότυπος G βρίσκεται κυρίως στην Αφρική και την κεντρική Ευρώπη.

• Οι υπότυποι H, K και CRF04_cpx παρατηρούνται σε μικρές επιδημίες, κυρίως στην κεντρική Αφρική.

• Τέλος, ο υπότυπος J βρίσκεται κυρίως σε μικρές επιδημίες στη Βόρεια, Κεντρική και Δυτική Αφρική, καθώς και την Καραϊβική.

Προέλευση Υποτύπου επιδημία Β

Όπως προαναφέρθηκε, ο υπότυπος B είναι η κυρίαρχη μορφή του HIV στο «δυτικό κόσμο», συμπεριλαμβανομένης της δυτικής και κεντρικής Ευρώπης, της Βόρειας Αφρικής, της αμερικανικής ηπείρου, στο σύνολό της, αλλά και της Ωκεανίας, καλύπτοντας περισσότερο από το 70% των λοιμώξεων ((Osmanov et al., 2002) και (Hemelaar et al., 2011)), σε αυτές τις ηπείρους. Ο υπότυπος B συνήθως παρατηρείται σε πολλές άλλες χώρες της Νοτιοανατολικής Ασίας, της Μέσης Ανατολής, της Νότιας Αφρικής και της Ρωσίας μεταξύ ομοφυλόφιλων ανδρών (MSM) (Buonaguro et al., 2007). Αξίζει να σημειωθεί ότι ο υπότυπος B είναι παγκοσμίως η κυρίαρχη μορφή μεταξύ των MSM ,δηλαδή άνδρες έρχονται σε σεξουαλική επαφή με άνδρες ,αλλά και ο πιο διαδεδομένος υπότυπος σε όλο τον κόσμο. Πιθανολογείται ότι είναι ο αιτιολογικός παράγοντας για το 11% περίπου όλων των περιπτώσεων του ιού HIV σε όλο τον κόσμο.

Προηγούμενες μελέτες έχουν καταλήξει στο συμπέρασμα ότι τοπικές επιδημίες ήταν ενεργές στην Κινσάσα από τα μέσα του '40 (Faria et al., 2014), ενώ η παγκόσμια εξάπλωση φαίνεται να έχει τις ρίζες της στις αρχές της δεκαετίας του '60. Στα μέσα της δεκαετίας του '60, επαγγελματίες από την Αϊτή, που πήγαν στην Αφρική στις αρχές της δεκαετίας, επέστρεψαν πίσω στις πατρίδες τους λόγω πολιτικών αναταραχών ((Junqueira και Matos Almeida, 2016) και (Gilbert etal., 2007)), εισάγοντας έτσι τον ιό HIV στην Καραϊβική. Αξίζει να σημειωθεί ότι η Αϊτή ήταν ένα νησί με ομοφυλοφιλικό τουρισμό για τους πολίτες των ΗΠΑ στη δεκαετία του '60, γεγονός που ερμηνεύει τον μετέπειτα υψηλότερο επιπολασμό της HIV λοίμωξης μεταξύ των MSM σε σχέση με τους ετεροφυλόφιλους ((Paraskevis et al., 2009) και (Beyrer et al., 2013)) αλλά και τη γρήγορη εισαγωγή στις ΗΠΑ. Μέσα από συστηματικές φυλοδυναμικές και φυλογεωγραφικές έρευνες προτείνεται ότι η αρχική εισαγωγή του ιού HIV στις ΗΠΑ ήταν μέσω του δικτύου των MSM και μπορεί να χρονολογηθεί στα τέλη της δεκαετίας του '60 (1969 ~ (1966-1972)) (Junqueira και Matos Almeida, 2016), (Gilbert et al., 2007). Επίσης, έχουν βρεθεί και μετέπειτα εισαγωγές στις ΗΠΑ, αλλά η πλειοψηφία των λοιμώξεων προτείνεται ότι προέρχεται από την πρώτη εισαγωγή της Αϊτής (Osmanov et al., 2002), (Hemelaar et al., 2011).

Στην Ευρώπη, ο υπότυπος είναι υπεύθυνος σχεδόν για το 70% (Abecasis et al., 2013) των HIV λοιμώξεων, με πολλαπλές εισαγωγές (Paraskevis et al., 2009), μέσω του δικτύου MSM ως επί το πλείστον. Επιπλέον, έχει επισημανθεί ότι αρκετές χώρες, όπως η Ελλάδα, η Πορτογαλία, η Σερβία και η Ισπανία ήταν μεταβατικές προς την Ευρωπαϊκή επιδημία, ενώ η Αυστρία, το Βέλγιο, η Δανία, η Γαλλία, η Γερμανία, το Λουξεμβούργο, η Πορτογαλία, η Σουηδία, η Ελβετία και η Ολλανδία ήταν οι κύριοι μεταναστευτικοί στόχοι.

Στην Ωκεανία, ο πληθυσμός των ατόμων που έχουν μολυνθεί με τον ιό HIV είναι ιδιαίτερα μικρός και η μελέτη της προέλευσης και της διάδοσης του ιού είναι περιορισμένη. Τέλος, οι κυρίαρχοι υπότυποι της Ασίας και της Αφρικής είναι οι A και C, ενώ ο υπότυπος B περιορίζεται σε δίκτυα MSM (Gilbert et al., 2007).

AIDS

Η HIV μόλυνση οδηγεί στο Σύνδρομο Επίκτητης Ανοσολογικής Ανεπάρκειας (AIDS), μια ασθένεια η οποία αναγνωρίστηκε για πρώτη φορά το 1981, όταν αυξημένος αριθμός νεαρών MSM υπέκυπταν από ασυνήθιστες ευκαιριακές λοιμώξεις και σπάνια κακοήθειες (Greene, 2007). Ο ρετροϊός, που είναι γνωστός ως ιός ανθρώπινης ανοσοανεπάρκειας τύπου 1 (HIV-1), ταυτοποιήθηκε ως ο αιτιολογικός παράγων μιας από τις πλέον θανατηφόρες μολυσματικές ασθένειες στην ανθρώπινη ιστορία ((Barre-Sinoussietal, 1983) (Galloetal, 1984) (Popovicetal, 1984)). Οι τρόποι μετάδοσης του HIV είναι κυρίως σεξουαλικοί, διαδερμικοί, περιγεννητικοί και αιματογενείς

Πρόσφατες έρευνες υποδεικνύουν ότι σχεδόν οκτώ στους δέκα ενήλικες μολύνθηκαν από τον ιό μετά από έκθεση του βλεννογόνου, καταλήγοντας στο συμπέρασμα ότι το AIDS μπορεί να θεωρηθεί ως μια σεξουαλικώς μεταδιδόμενη ασθένεια κατά κύριο λόγο ((Hladík και McElrath, 2008) (Cohen et al, 2011)). Παρά το γεγονός ότι η αντιρετροϊκή θεραπεία έχει μειώσει τους θανάτους που

σχετίζονται με το AIDS, η πρόσβαση στη θεραπεία δεν είναι καθολική, και οι προοπτικές εύρεσης θεραπείας ενός αποτελεσματικού εμβολίου είναι αβέβαιες ((Barouch, 2008) (Richman et al, 2009)). Έτσι, το AIDS θα εξακολουθήσει να αποτελεί σημαντική απειλή για τη δημόσια υγεία για τις επόμενες δεκαετίες. Ένα από τα σημαντικότερα χαρακτηριστικά των λεντιϊών (ρετροϊών) είναι η εκτεταμένη γενετική μεταβλητότητα τους, αποτέλεσμα του υψηλού ποσοστού σφάλματος κατά τη διάρκεια του ανασυνδυασμού του RNA μέσω του ενζύμου της αντίστροφης μεταγραφάσης, και ο γρήγορος χρόνος ζωής των ιοσωματίων (Drosopouloset al., 1998).

Φυλογενετική Ανάλυση

Στοίχιση των αλληλουχιών

Το γονιδίωμα υπόκειται σε αλλαγές με την πάροδο του χρόνου, που ονομάζονται μεταλλάξεις και ορίζονται ως οι μεταβολές του γονιδιώματος ενός οργανισμού ή ενός ιού.

Οι μεταλλάξεις ταξινομούνται ως εξής:

• Αντικαταστάσεις, όπου μια βάση αντικαθιστά μια άλλη

• Προσθέσεις, όπου εισάγονται μέσα στην αλληλουχία του DNA επιπλέον ζεύγη βάσεων

• Διαγραφές όπου ένα τμήμα του DNA έχει χαθεί ή διαγραφεί.

Προκειμένου να αναλυθεί η γενετική σχέση μεταξύ των διαφόρων παθογόνων που συλλέγονται (κυρίως του ίδιου παθογόνου σε διαφορετικούς μολυσμένους ασθενείς), θα πρέπει αρχικά να επιλεχθεί ένα κοινό γονίδιο κατάλληλο για τη σύγκριση. Το είδος της μελέτης υποδεικνύει το μέρος του γονιδιώματος που θα

πρέπει να χρησιμοποιήσουμε τις περισσότερες φορές. Στα παθογόνα, ο ρυθμός μεταλλάξεων είναι πολύ υψηλός, με αποτέλεσμα ο προσδιορισμός του κοινού τμήματος του να χρειάζεται μετέπειτα μια διαδικασία που ονομάζεται στοίχιση των DNA (ή RNA) αλληλουχιών. Η Στοίχιση των αλληλουχιών επιτυγχάνεται μέσω υπολογιστών ως επί το πλείστον, με βάση scores, τα οποία αυξάνονται όταν ταιριάζουν και μειώνονται όταν δεν ταιριάζουν οι νουκλεοτιδικές βάσεις.

Σε αυτή την ενότητα, αναφερόμαστε στις κύριες μεθόδους στοίχισης, οι οποίες είναι:

- Ανά ζεύγος στοίχιση

- Μέθοδοι Dot-matrix (Gibbs και McIntyre, 1970)

- Δυναμικός Προγραμματισμός

- Πολλαπλή στοίχιση ακολουθιών

- Δυναμικός Προγραμματισμός

- Προοδευτικοί μέθοδοι

Η Στοίχιση των αλληλουχιών αξιολογεί την ομοιότητα των αλληλουχιών, δίνοντας έτσι μια πρώτη ομαδοποίηση, και θεωρείται ως το πρωταρχικό βήμα για οποιαδήποτε περαιτέρω ανάλυση.

Εξελικτικά μοντέλα

Η βασική ιδέα στα εξελικτικά μοντέλα είναι ότι ακολουθούν μια διαδικασία Markov, δηλαδή ότι υπάρχει μια τυχαία διαδικασία που υφίσταται μεταβάσεις από μια κατάσταση σε μια άλλη. Σε όλες τις αλυσίδες Markov, υπάρχει ένας πίνακας μετάβασης που αντιπροσωπεύει την πιθανότητα να βρεθεί στην κατάσταση i, δεδομένου ότι ήταν στην κατάσταση j.

Στην περίπτωσή μας, ο πίνακας μετάβασης είναι:

$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{GA}(t) & p_{CA}(t) & p_{TA}(t) \\ p_{AG}(t) & p_{GG}(t) & p_{CG}(t) & p_{TG}(t) \\ p_{AC}(t) & p_{GC}(t) & p_{CC}(t) & p_{TC}(t) \\ p_{AT}(t) & p_{GT}(t) & p_{CT}(t) & p_{TT}(t) \end{pmatrix}$$

όπου οι δείκτες είναι τα νουκλεοτίδια (A, T, G, C) και t είναι ο χρόνος και πρόκειται για μια συνεχή γενίκευση του μοντέλου Markov.

Τα πιο κοινά μοντέλα της εξέλιξης DNA είναι:

- JC69 μοντέλο (JUKES and CANTOR, 1969)

Το Jukes-Cantor μοντέλο είναι το πιο απλό εξελικτικά μοντέλο, αφού υποθέτει ότι όλες οι συχνότητες των νουκλεοτιδικών βάσεων είναι ίσες, δηλαδή ότι $\pi_A = \pi_T = \pi_G = \pi_C = \frac{1}{4}$ και ότι οι ρυθμοί αντικατάστασης είναι ίσοι οπότε η μόνη παράμετρος του μοντέλου αυτού είναι το μ, δηλαδή ο συνολικός ρυθμός αντικατάστασης.

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \kappa \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

Και ο πίνακας μετάβασης :

$$P(t) = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu}) \end{pmatrix}$$

Λαμβάνοντας υπόψη το ποσοστό των περιοχών που διαφέρουν μεταξύ δύο αλληλουχιών, η εξελικτική απόσταση μεταξύ τους μπορεί να εκτιμηθεί ως:

$\hat{d}$ = - $\frac{3}{4}$ $ln(1 - \frac{4}{3}p)$

- K80 μοντέλο (Kimura, 1980)

Στο μοντέλο Kimura όλες οι συχνότητες των νουκλεοτιδικών βάσεων είναι ακόμα ίσες, δηλαδή ότι $\pi_A = \pi_T = \pi_G = \pi_C = \frac{1}{4}$, αλλά οι μεταβάσεις και οι μεταστροφές θεωρούνται ότι έχουν διαφορετικές πιθανότητες.

Ο πίνακας των ρυθμών αντικατάστασης είναι:

$$Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$$

Η εξελικτική απόσταση μεταξύ δύο ακολουθιών είναι $\hat{d}$ = - $\frac{1}{2}\ln((1 - 2p - q)\sqrt{1 - 2q})$

- Μοντέλο F81 (Felsenstein, 1981)

Το μοντέλο Felsenstein υποθέτει ότι οι νουκλεοτιδικές βάσεις είναι άνισες ($\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$)

Ο πίνακας των ρυθμών αντικατάστασης είναι:

$$Q = \begin{pmatrix} * & \pi_C & \pi_A & \pi_G \\ \pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \pi_G \\ \pi_T & \pi_C & \pi_A & * \end{pmatrix}$$

- Μοντέλο HKY85 (Hasegawa, Kishino και Yano) (Hasegawa et al., 1985)

Το μοντέλο Hasegawa, Kishino και Yano είναι ένας συνδυασμός των δύο παραπάνω μοντέλων, επειδή επιτρέπει την ανισότητα των συχνοτήτων βάσεων($\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$)αλλά και διαφορετικούς ρυθμούς μεταβάσεων και μεταστροφών .

Ο πίνακας των ρυθμών αντικατάστασης είναι:

$$Q = \begin{pmatrix} * & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & * \end{pmatrix}$$

- TN93 model (K and M, 1993)

Το μοντέλο Tamura και Nei επιτρέπει την ανισότητα των συχνοτήτων βάσεων ($\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$) και οι μεταβάσεις επιτρέπεται να είναι άνισες ((A<->G) μεταξύ τους (C<->T)), αλλά οι μεταστροφές πρέπει να έχουν ίδιο ρυθμό.

Ο πίνακας των ρυθμών αντικατάστασης είναι:

$$Q = \begin{pmatrix} * & \kappa_1 \pi_C & \pi_A & \pi_G \\ \kappa_1 \pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \kappa_2 \pi_G \\ \pi_T & \pi_C & \kappa_2 \pi_A & * \end{pmatrix}$$

- GTR: Generalised time-reversible (Tavare, 1986) (S, 1986)

Τέλος, το μοντέλο με τις περισσότερες άγνωστες παραμέτρους και σχεδόν χωρίς καμία παραδοχή για ισότητα είναι το μοντέλο GTR. Όλες οι συχνότητες βάσης και όλες οι τιμές είναι άνισες ($\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$).

Αν συμβολίσουμε τις παραμέτρους ρυθμού μετάβασης,

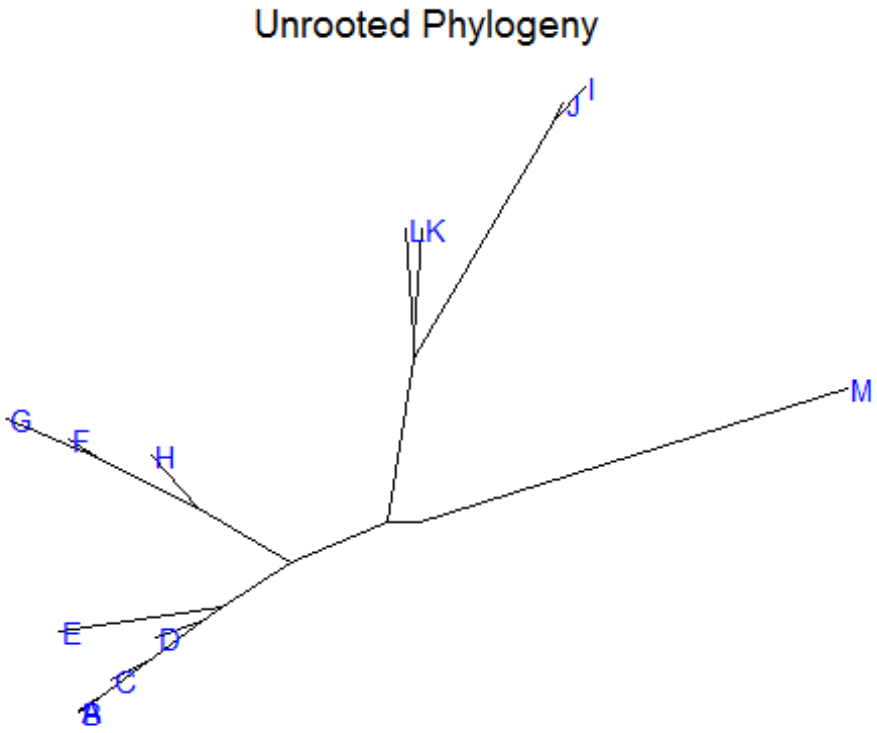
$$\begin{pmatrix} \alpha = r(T \to C) = r(C \to T) \\ \beta = r(T \to A) = r(A \to T) \\ \gamma = r(T \to G) = r(G \to T) \\ \delta = r(C \to A) = r(A \to C) \\ \epsilon = r(C \to G) = r(G \to C) \\ \eta = r(A \to G) = r(G \to A) \end{pmatrix}$$

τότε ο πίνακας των ρυθμών αντικατάστασης είναι:

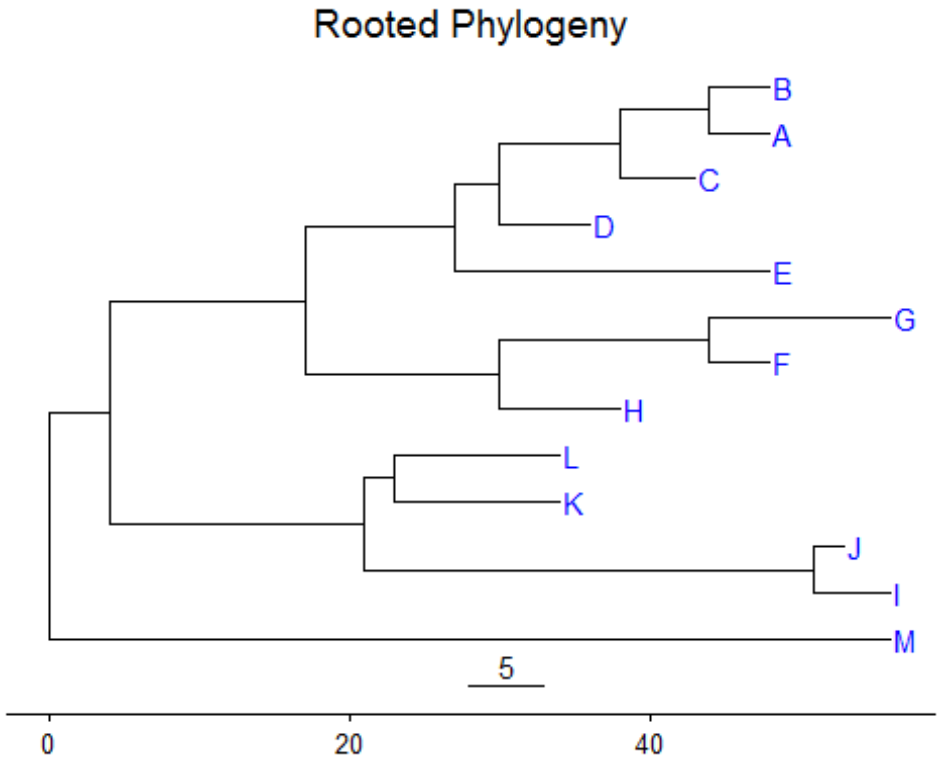
$$Q = \begin{pmatrix} -(\alpha\pi_C + \beta\pi_A + \gamma\pi_G) & \alpha\pi_C & \beta\pi_A & \gamma\pi_G \\ \alpha\pi_T & -(\alpha\pi_T + \delta\pi_A + \epsilon\pi_G) & \delta\pi_A & \epsilon\pi_G \\ \beta\pi_T & \delta\pi_C & -(\beta\pi_T + \delta\pi_C + \eta\pi_G) & \eta\pi_G \\ \gamma\pi_T & \epsilon\pi_C & \eta\pi_A & -(\gamma\pi_T + \epsilon\pi_C + \eta\pi_A) \end{pmatrix}$$

Φυλογένεια (Φυλογενετικά δέντρα)

Αφού στοιχίσουμε τις αλληλουχίες DNA ή RNA ενός παθογόνου, διάφορες στατιστικές αναλύσεις μπορούν να εκτελεστούν ανάλογα με το ερευνητικό ερώτημα που έχουμε θέσει. Προκειμένου να εξεταστεί η εξελικτική ιστορία και οι σχέσεις μεταξύ των αλληλουχιών που έχουν συλλεχθεί, θα πρέπει να διεξαχθεί μια φυλογενετική ανάλυση. Η αξιολόγηση της παρατηρούμενης τοπολογίας ενός παθογόνου με την εφαρμογή των εξελικτικών μοντέλων, οδηγεί στην ανακάλυψη των προαναφερθεισών σχέσεων. Το αποτέλεσμα αυτών των αναλύσεων είναι ένα φυλογενετικό δέντρο, όπου οι άκρες είναι οι παθογόνα που παρατηρήθηκαν στη μελέτη (Jansson, 2015). Τα φυλογενετικά δέντρα μπορεί να έχουν ρίζα ή όχι, ανάλογα με την προγονική πληροφορία. Τα δένδρα τα οποία δεν έχουν ρίζα απεικονίζουν τη σχετικότητα των άκρων (αλληλουχιών) χωρίς κάποια παραδοχή για τον κοινό τους πρόγονο ή το εξελικτικό μοντέλο.

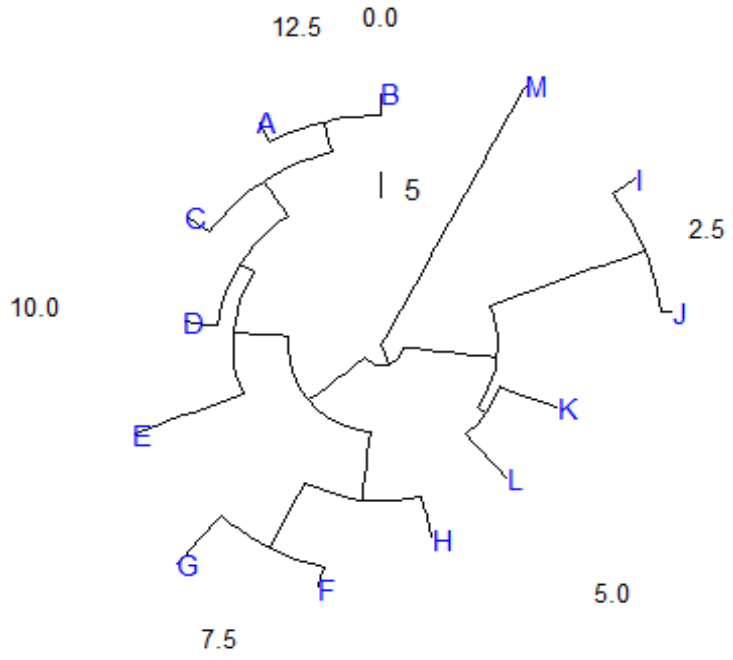


Σε φυλογενετικά δέντρα που έχουν ρίζα υπονοείται η ύπαρξη ενός κοινού προγόνου και, ως εκ τούτου, οι αλληλουχίες συνδέονται με έναν θεωρητικό πρόγονο. Αυτό επαναλαμβάνεται για τους κοινούς προγόνους μέχρι τη ρίζα του δέντρου που είναι ο κοινός πρόγονος για όλες τις αλληλουχίες. Είναι προφανές ότι μπορούμε να παράγουμε πάντα δέντρα χωρίς ρίζα από δέντρα με ρίζα, αλλά το αντίστροφο δεν είναι πάντα εφικτό χωρίς την υπόθεση ενός μοντέλου εξέλιξης ή ρυθμών διασποράς (Romero, 2004).

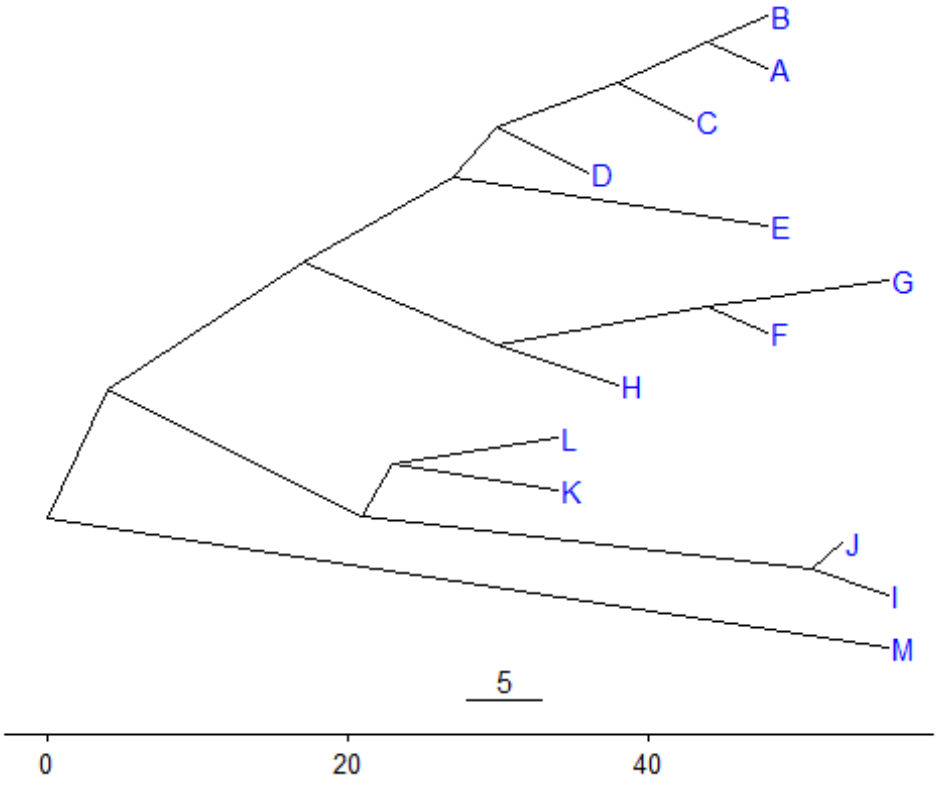


Υπάρχουν επίσης διαφορετικά σχήματα των φυλογενετικών δέντρων, όπως κυκλικά και κεκλιμένα φυλογενετικά δέντρα.

Fan Shaped Phylogeny



Slanted Phylogeny



Ποικίλες μέθοδοι έχουν αναπτυχθεί για την κατασκευή ενός φυλογενετικού δέντρου . Οι πιο διαδεδομένες είναι:

- Μέθοδοι Αποστάσεων
- Μέγιστης Φειδωλότητας
- Μέγιστης πιθανοφάνειας
- Μπεϋζιανές Μέθοδοι

Μέθοδοι Αποστάσεων

Οι Μέθοδοι Αποστάσεων βασίζονται στον υπολογισμό της γενετικής απόστασης μεταξύ των αλληλουχιών. Η γενετική απόσταση συχνά υπολογίζεται ως το κλάσμα του αριθμού των μεταλλάξεων διαιρούμενο με το συνολικό μήκος της αλληλουχίας (τα κενά μπορεί να χρησιμοποιηθούν ως αναντιστοιχίες δηλαδή ως μεταλλάξεις επίσης). Με αυτόν τον τρόπο, ένας πίνακας NxN αποστάσεων δημιουργείται με όλες (Rohlf, 2005) τις γενετικές αποστάσεις των αλληλουχιών. Αξίζει να αναφερθεί ότι με αυτή τη μέθοδο η πραγματική γενετική απόσταση υποεκτιμάται. Με την προσθήκη ενός εξελικτικού μοντέλου διορθώνουμε το υποεκτιμώμενη παρατηρούμενη γενετική απόσταση στην πραγματική .

Οι κύριες μέθοδοι κλαδοποίησης που βασίζονται σε μεθόδους αποστάσεων είναι:

- Nearest Neighbour: Μια μέθοδος ομαδοποίησης από κάτω προς τα πάνω (agglomerative), που δημιουργήθηκε από τους Naruya Saitou και Masatoshi Nei για την ανακατασκευή φυλογενετικών δέντρων από την εξελικτική δεδομένων απόστασης.

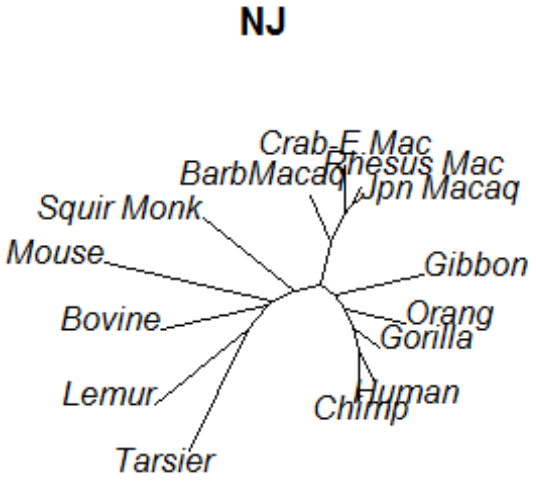
  Ο αλγόριθμος μπορεί να συνοψιστεί σε τέσσερα βήματα:

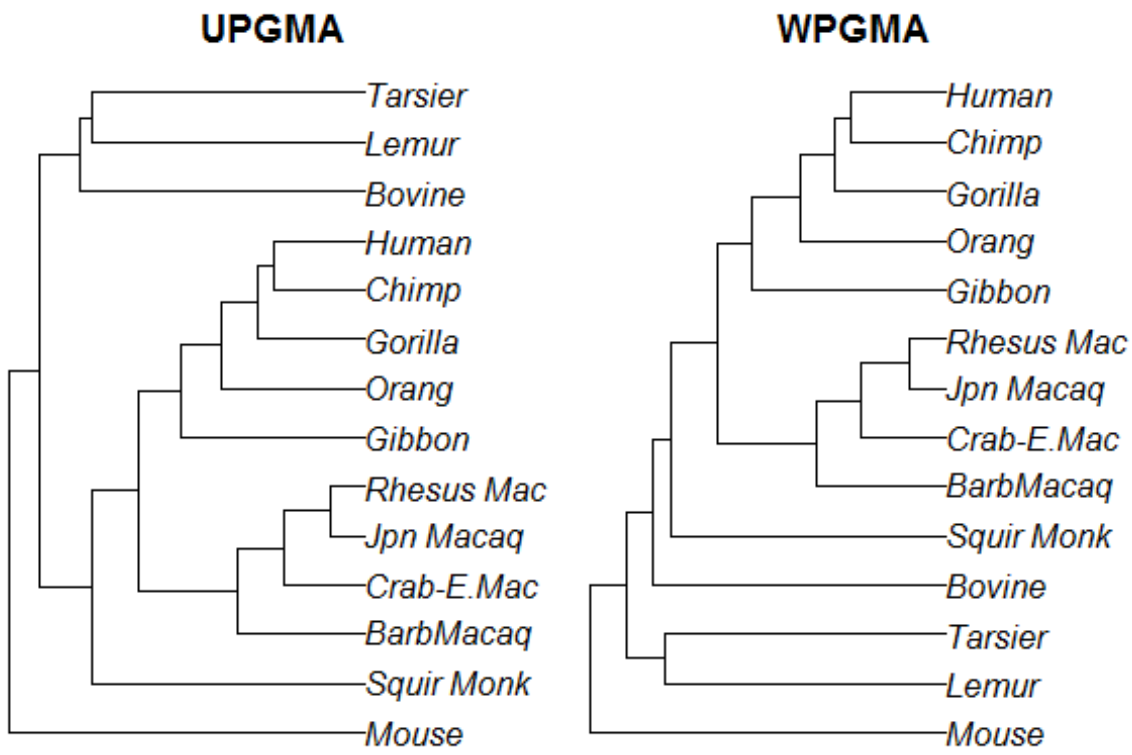
  - ✓ ***Βήμα 1ο:*** Ορίζουμε τις αλληλουχίες ή τις λειτουργικές ταξινομικές μονάδες (OTUs)
  - ✓ ***Βήμα 2ο:*** Υπολογίζουμε τις γενετικές αποστάσεις μεταξύ των OTUs
  - ✓ ***Βήμα 3ο:*** Συνδέουμε τα OTUs με την ελάχιστη απόσταση

- ✓ ***Βήμα 4ο:*** Αν όλες οι OTUs είναι ενωμένες σταματάμε αλλιώς επαναλαμβάνουμε το Βήμα 3ο.

Σε κάθε στάδιο, η μέθοδος αυτή προσπαθεί να ελαχιστοποιήσει τη γενετική απόσταση η οποία είναι ίση με τη μέση τιμή του μήκους διακλάδωσης δύο Otus. Η μέθοδος ξεκινάει με μια αστεροειδή τοπολογία (Nei M., 1987), δεν υποθέτει σταθερό ρυθμό εξέλιξης και έτσι παράγει δένδρα χωρίς ρίζα. Για να ξεπεραστεί αυτή η δυσκολία, έχουν προταθεί πολλές παραλλαγές.

- UPGMA & WPGMA: Οι Unweighted and Weighted Pair-Group Methods είναι και αυτές agglomerative μέθοδοι ομαδοποίησης, στις οποίες η αρχική υπόθεση είναι ότι υπάρχει ένα υπερμετρικό (ultrametric) δέντρο, δηλαδή ότι ο αριθμός των μεταλλάξεων είναι ανάλογος με τη χρονική απόσταση ενός κόμβου προς τον πρόγονο και ότι οι μεταλλάξεις πραγματοποιούνται με τον ίδιο ρυθμό σε όλα τα μονοπάτια. Η διαφορά μεταξύ των δύο μεθόδων είναι ότι στην WPGMA ο υπολογισμός της απόστασης μεταξύ 2 ομάδων (OTUs) υπολογίζεται από ένα απλό μέσο όρο, ενώ στο UPGMA οι μέσοι σταθμίζονται με τον αριθμό των taxa σε κάθε βήμα.





Η μέθοδος NJ είναι λιγότερο απαιτητική σε υπολογιστική ισχύ μέθοδος. Οι μέθοδοι UPGMA & WPGMA είναι αρκετά καλές μέθοδοι σε σύνολα δεδομένων όπου η παραδοχή του μοριακού ρολογιού ισχύει. Τα φυλογενετικά δέντρα τα οποία λαμβάνουν προγράμματα για τη διερεύνηση της Φυλοδυναμικής ενός ιού συχνά στηρίζονται στην υπόθεση ύπαρξης μοριακού ρολογιού. Έτσι, συχνά χρησιμοποιούνται οι UPGMA ή WPGMA ως αρχικές τοπολογίες.

Μέθοδοι μεγίστης Φειδωλότητας

Η μέθοδος μέγιστης Φειδωλότητας είναι μια μέθοδος χαρακτήρων, στην οποία οι αλληλουχίες παράγουν το εξελικτικό δέντρο (ή δέντρα) και ελαχιστοποιούν

τον αριθμό των βημάτων που απαιτούνται για να δημιουργηθούν οι παρατηρούμενες μεταβολές στις αλληλουχίες από τις κοινές προγονικές αλληλουχίες (Mount, 2008).

Οι μέθοδοι μέγιστης Φειδωλότητας δεν βασίζονται σε κάποιο εξελικτικό μοντέλο, αλλά στο κριτήριο της Φειδωλότητας, δηλαδή να γίνονται όσο το δυνατόν λιγότερες μεταλλάξεις. Απαιτούν μεγάλη υπολογιστική ισχύ ειδικά σε δεδομένα με περισσότερες από 20 αλληλουχίες. Ο λόγος είναι ότι ο αριθμός των δέντρων που πρέπει να υπολογισθούν είναι πολύ μεγάλος και για αυτό έχουν αναπτυχθεί διάφορες ευριτικές μέθοδοι προκειμένου να επιταχυνθεί η διαδικασία, όπως είναι ο αλγόριθμος Close-Neighbour-Interchange (M. Nei και Kumar, 2000) που εφαρμόζεται στο πρόγραμμα MEGA. Η προαναφερθείσα μέθοδος υποθέτει ότι ρυθμός μεταβολής κατά μήκος κλάδων είναι παρόμοιος, υπόθεση που σε αρκετές περιπτώσεις δεν είναι αληθής. Οι μέθοδοι μέγιστης Φειδωλότητας έχουν επίσης επικριθεί, καθώς η φύση δεν λειτουργεί ρητά με την προϋπόθεση της Φειδωλότητας.

Μέθοδοι Μεγίστης Πιθανοφάνειας

Οι μέθοδοι μεγίστης Πιθανοφάνειας στηρίζονται στην εκτίμηση των παραμέτρων του μοντέλου, δεδομένων των αλληλουχιών που έχουμε λάβει. Αυτό επιτυγχάνεται με την εύρεση των τιμών των παραμέτρων, που μεγιστοποιούν την πιθανοφάνεια, δηλαδή την πιθανότητα να λάβουμε τις παρατηρήσεις με δεδομένες τις παραμέτρους. Στα φυλογενετικά δέντρα ειδικότερα, οι μέθοδοι μέγιστης πιθανοφάνειας προσπαθούν να βρουν το δέντρο που μεγιστοποιεί την πιθανότητα να παραχθούν τα παρατηρούμενα δεδομένα, βάσει ενός μοντέλο νουκλεοτιδικής (ή πρωτεϊνικής) αντικατάστασης. Για παράδειγμα, βάσει των δεδομένων D και του μοντέλου M, να βρεθεί το δέντρο T έτσι ώστε η πιθανότητα να παρατηρήσουμε τα δεδομένα, δεδομένου του δέντρου και του μοντέλου ($Pr(D|T, M)$) να μεγιστοποιείται.

Τα κυριότερα πλεονεκτήματα αναφορικά με τις μεθόδους μεγίστης Πιθανοφάνειας είναι τα παρακάτω:

- ✓ Κατάλληλη για ακολουθίες DNA, διότι μπορούμε να μοντελοποιήσουμε τη στοχαστική διαδικασία
- ✓ Robust, Εύρωστη ακόμη και αν υπάρχουν πολλές παραβιάσεις του εξελικτικού μοντέλου.

- ✓ Απαιτείται ένα σαφές μοντέλο εξέλιξης, που να μπορεί να εφαρμοστεί στα δεδομένα.

- ✓ Ανώτερη από τις μεθόδους αποστάσεων επειδή χρησιμοποιεί όλη την πληροφορία της αλληλουχίας και αξιολογεί διαφορετικές τοπολογίες δέντρων
- ✓ Παράγει πιο ρεαλιστικά μήκη κλαδιών σε σχέση με αυτά των μεθόδων αποστασεων και της μέγιστης Φειδωλότητας.
- ✓ Χρησιμοποιεί πληροφορία που προέρχεται από περιοχές όπου οι μέθοδοι της μέγιστης Φειδωλότητας δεν προσφέρουν πληροφορία.

Τα κυριότερα μειονεκτήματα για τις μεθόδους μεγίστης Πιθανοφάνειας είναι τα παρακάτω:

- ❖ Υπολογιστικά επίπονες και αργές
- ❖ Το αποτέλεσμα εξαρτάται από το μοντέλο που χρησιμοποιήθηκε και οι πληροφορίες που προέρχονται από τις περιοχές δεν εκτιμώνται στις μεθόδους μέγιστης Φειδωλότητας (Εκλαμβάνεται και ως πλεονέκτημα ανάλογα τα δεδομένα)

- ❖ Σε περίπλοκα δεδομένα, κατά τα οποία δεν μπορούμε να βρούμε σαφές εξελικτικό μοντέλο, η εφαρμογή ενός επιδέχεται μεγάλης κριτικής.

Μπεϋζιανές Μέθοδοι

Μέχρι τώρα έχουμε διερευνήσει τις φυλογένειες από τη σκοπιά της κλασσικής στατιστικής, οι οποίες θεωρούνται συγκεκριμένες αλλά με άγνωστες ποσότητες. Μια πιο πρόσφατη μπεϋζιανή προσέγγιση έχει αναπτυχθεί στην οποία η φυλογένεια θεωρείται και αυτή τυχαία μεταβλητή. Η μπεϋζιανή προσέγγιση βασίζεται στο Θεώρημα Bayes από το οποίο λαμβάνει και το όνομά της. Αν

έχουμε δύο γεγονότα Α και Β, τότε η πιθανότητα να συμβεί το Α δεδομένου ότι έχει συμβεί το Β είναι:

$$\mathrm{P(A \mid B)} = \frac{\mathrm{P(B \mid A)} \cdot \mathrm{P(A)}}{\mathrm{P(B)}}$$

Όπου

- P(A), είναι η πιθανότητα να συμβεί το Α ενδεχόμενο και ονομάζεται **εκ των προτέρων πιθανότητα**
- P(B) είναι η πιθανότητα να συμβεί το Β, και ονομάζεται **περιθώρια πιθανοφάνειας**.

- $\mathrm{P(A \mid B)}$, είναι η πιθανότητα συμβεί το Α ενδεχόμενο δεδομένου ότι ισχύει το Β και ονομάζεται **εκ των υστέρων πιθανότητα**.

- $\mathrm{P(B \mid A)}$ είναι η πιθανότητα να συμβεί το Β δεδομένου ότι ισχύει το Α στην περίπτωση που το Α είναι τα δεδομένα και ονομάζεται **συνάρτηση πιθανοφάνειας**. Η **συνάρτηση πιθανοφάνειας** υποδεικνύει τη συμβατότητα των στοιχείων με τη συγκεκριμένη υπόθεση.

  Το θεώρημα του Bayes γενικεύεται ως εξής:

  Έστω $\{\mathrm{B}_1, \mathrm{B}_2, \ldots\}$ μια διαμέριση ενός δειγματικού χώρου Β.

Τότε για κάθε$\mathrm{B_i}$ που ανήκει στο χώρο Β ισχύει ότι:

$$\Pr(\mathrm{B_i}|\mathrm{A}) = \frac{\Pr(\mathrm{A}|\mathrm{B_i})\Pr(\mathrm{B_i})}{\Pr(\mathrm{A})} = \frac{\Pr(\mathrm{A}|\mathrm{B_i})\Pr(\mathrm{B_i})}{\sum_j \Pr(\mathrm{A}|\mathrm{B_j})\Pr(\mathrm{B_j})}$$

Ενώ για συνεχείς μεταβλητές:

$$f_{\mathrm{X}|\mathrm{Y}}\left(\mathrm{X}|\mathrm{Y}\right) = \frac{f_{\mathrm{y}|\mathrm{x}}(\mathrm{y}|\mathrm{x}) f_X(\mathrm{x})}{f_Y(\mathrm{y})} = \frac{f_{\mathrm{y}|\mathrm{x}}(\mathrm{y}|\mathrm{x}) f_X(\mathrm{x})}{\int f_{\mathrm{y}|\mathrm{x}}(\mathrm{y}|\mathrm{x}) f_X(\mathrm{x})}$$

Η γενική αυτή Μπεϋζιανή προσέγγιση εφαρμόζεται στην κατασκευή του φυλογενετικού δέντρου, όπου εκ των υστέρων πιθανότητα είναι η πιθανότητα ενός δέντρου δεδομένου των αλληλουχιών P(Tree|D) ~ f(Tree|D). Η εκ των προτέρων πιθανότητα ενός δέντρου είναι η P(Tree) ~ f(Tree), ενώ η συνάρτηση πιθανοφάνειας είναι η P(D|Tree) ~ f(D|Tree). Αν συνδυάσουμε τα παραπάνω, έχουμε εκ των υστέρων κατανομή των δέντρων P(Tree|D) ~ $f_{\mathrm{Tree}|\mathrm{D}}(\mathrm{Tree}|\mathrm{D})$.. Μετά την εύρεση της εκ των υστέρων πιθανότητας όλων των δέντρων, επιλέγουμε αυτό με τη μεγαλύτερη πιθανότητα, ώστε να αντιπροσωπεύσει καλύτερα την φυλογένεια, η οποία δίνεται ως τελικό αποτέλεσμα. Όπως εύκολα μπορούμε να συμπεράνουμε, ένα μεγάλο πρόβλημα σε αυτή την προσέγγιση είναι ο υπολογισμός του ολοκληρώματος στον παρονομαστή. Προκειμένου να ληφθεί η τελική λύση MCMC (Monte Carlo Markov Chains), συνήθως χρησιμοποιούνται αλγόριθμοι δειγματοληψίας.

Φυλοδυναμική Ανάλυση

Η έρευνα για το πώς οι επιδημιολογικές και εξελικτικές διαδικασίες αλληλεπιδρούν με τις φυλογένειες θεωρείται σε γενικές γραμμές το κύριο θέμα της φυλοδυναμικής ανάλυσης. Η κυρίαρχη κατηγορία έρευνας στη φυλοδυναμική είναι οι φυλογενετικές μελέτες ιών, λόγω του γρήγορου ρυθμού εξέλιξης και μετάλλαξης αυτών. Ο γενικός ορισμός μπορεί να αναλυθεί περαιτέρω σε τρεις στόχους, που συνήθως διερευνώνται στις έρευνες, προκειμένου να μας βοηθήσει στον εντοπισμό, την πρόβλεψη και την επίλυση ιογενών προβλημάτων.

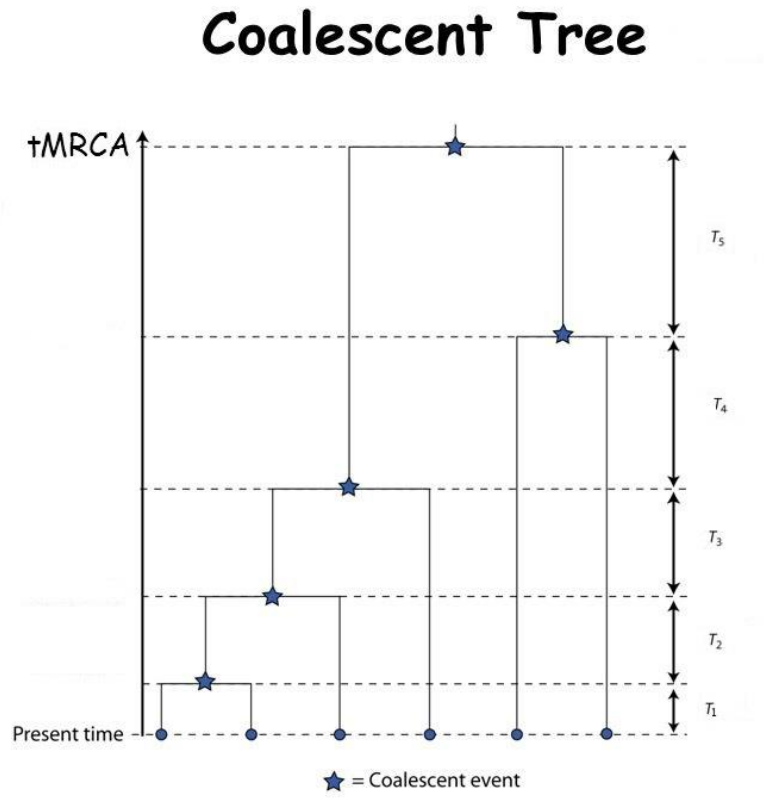
- Προέλευση Ιών: Οι ιοί εξελίσσονται εκατομμύρια φορές πιο γρήγορα από οργανισμούς, όπως ο άνθρωπος. Αυτό το χαρακτηριστικό μπορεί να μας βοηθήσει στην εκτίμηση μοριακών μοντέλων, παρέχοντας έτσι ένα ρυθμό εξέλιξης του ιού ανά έτος. Μετά τον υπολογισμό του ρυθμού εξέλιξης σε πραγματικές μονάδες χρόνου, είναι δυνατό να συμπεράνουμε την

ημερομηνία του πιο πρόσφατου κοινού πρόγονου, εντοπίζοντας έτσι την ημερομηνία προέλευσης και μερικές φορές ακόμη και τον τόπο προέλευσης του ιού.

- Διασπορά και Μετάδοση των Ιών: Μια άλλη πτυχή της φυλοδυναμικής είναι η μελέτη του «πώς η επιδημία εξαπλώνεται σε όλο το χρόνο». Φυλοδυναμικά μοντέλα μπορούν να παρέχουν πληροφορίες σχετικά με επιδημιολογικές παραμέτρους που είναι δύσκολο να εκτιμηθούν μέσω των παραδοσιακών μέσων επιτήρησης, συμβάλλοντας έτσι στη δημόσια υγεία μέσα από την πρόβλεψη και την πρόληψη περαιτέρω εξάπλωσης του ιού.
- Αξιολόγηση των παρεμβάσεων: Τέλος, οι φυλοδυναμικές αναλύσεις μπορούν να μας βοηθήσουν να αξιολογήσουμε την αποτελεσματικότητα των στρατηγικών παρέμβασης για τον περιορισμό ή την θεραπεία μια νόσου, όπως προγράμματα εμβολιασμού ή άλλες παρεμβάσεις.

Θεωρία σύγκλισης

Ένα σημαντικό βήμα στην εξέλιξη των Φυλοδυναμικών αναλύσεων ήταν η Coalescent Theory όπως περιγράφεται από τον Kingman ((J. F. C. Kingman, 1982), (J. Kingman, 1982)) και ορίζεται ως ένα μαθηματικό πρόβλημα της στοχαστικής διαδικασίας. Το n-coalescent ορίζεται ως η Μαρκοβιανή αλυσίδα συνεχούς χρόνου σε ένα πεπερασμένο σύνολο καταστάσεων. Αυτό περιγράφει τις σχέσεις μεταξύ ενός δείγματος μελών (n) που προέρχονται από έναν μεγάλο πληθυσμό. Η θεωρία αυτή μπορεί να χρησιμοποιηθεί για τη συσχέτιση της γενετικής ποικιλομορφίας και της δημογραφικής ιστορίας ενός πληθυσμού αλλά και ως ένα μοντέλο για την επίδραση της γενετικής παρέκκλισης στη γενεαλογία των προγόνων. Ο χρόνος θεωρείται συνεχής, και το t_j, που είναι οι χρόνοι διάρκειας για κάθε κόμβο, θεωρείται ότι ακολουθεί εκθετική κατανομή ((Tajima, 1983) (R., 1991) Nordborg, .D.)) ή γενικευμένα γάμα κατανομή



Ο συνολικός χρόνος για το γενικό κοινό απόγονο συνήθως συμβολίζεται ως tMRCA, είναι το άθροισμα των $T_1 + T_2 + T_3 + \ldots + T_n$, και επίσης κατανέμεται εκθετικά (ή Γάμμα) $(\mathrm{Exp}(\mathrm{sum}_{i=1}^{k} t_i, \lambda))$

Η coalescent theory υποθέτει ότι τα δεδομένα λαμβάνονται από έναν Wright-Fisher πληθυσμό, ο οποίος θεωρεί την ύπαρξη ως:

(i) Σταθερό μέγεθος πληθυσμού N
(ii) Διακριτές γενεές
(iii) Πλήρης μίξη των μεταδόσεων

Διαδικασία Γέννησης θανάτου

Μια άλλη προσέγγιση είναι τα μοντέλα γέννησης - θανάτου,τα οποία είναι μια ειδική περίπτωση της στοχαστικής Μαρκοβιανής διαδικασίας συνεχούς χρόνου, με μόνο δύο μεταβάσεις:

• Γεννήσεις, όπου ο πληθυσμός αυξάνεται κατά ένα και

• Θάνατοι, όπου ο πληθυσμός μειώνεται κατά ένα.

Οι διαδικασίες γέννησης - θανάτου (Kendall, 1948) χρησιμοποιούνται συχνά ως εναλλακτική λύση έναντι των Coalescent, όταν οι υποθέσεις πληθυσμού Wright-Fisher δεν ισχύουν ή αν τα δεδομένα δεν μπορούν να θεωρηθούν ένα μικρό δείγμα των εν λόγω πληθυσμού.

Molecular Clock Assumption

Μετά τον καθορισμό των tMRCA, δηλαδή του πιο πρόσφατου κοινού μας προγόνου ,ένα άλλο στοιχείο που πρέπει να λάβουμε υπόψιν είναι ο ρυθμός σύμφωνα με τον οποίο συμβαίνουν οι μεταλλάξεις. Στις αρχές της δεκαετίας του '60, ο όρος του μοριακού ρολογιού επινοήθηκε από τους Zuckerkandl και Pauling, οι οποίοι παρατήρησαν ότι η μοριακή εξέλιξη λαμβάνει χώρα σε ένα σχεδόν σταθερό ρυθμό με την πάροδο του χρόνου (ZuckerkandlE, 1962). Σχεδόν πέντε δεκαετίες αργότερα, μια σειρά από μελέτες έχουν αποδείξει ότι η υπόθεση αυτή δεν είναι καθολική και δεν μπορεί να εφαρμοστεί σε όλα τα μοριακά δεδομένα. Για παράδειγμα, αυτή η πολύ αυστηρή παραδοχή δεν είναι εφαρμόσιμη στους ιούς, όπως τον ιό HIV και συνιστάται η χρήση πιο χαλαρών συνθηκών (Drummond et al., 2006).

Οι κυριότερες «διορθώσεις» του μοριακού ρολογιού που έχουν αναπτυχθεί, προκειμένου να καμφθεί η υπόθεση του σταθερού ρυθμού αντικατάστασης κατά τη διάρκεια του χρόνου, είναι :

- Τοπικά Μοριακά Ρολόγια (Local molecular clocks) ((Kishino and Hasegawa, 1990), (Rambaut and Bromham, 1998), (Yang and Yoder, 2003); (Drummond

and Suchard, 2010)): Εφαρμόζεται κυρίως σε στενά συνδεόμενες γενεαλογίες που έχουν το ίδιο ρυθμό εξέλιξης και οι ρυθμοί ομαδοποιούνται ανά υπο-κλάδο.
- Μεικτές Διαδικασίες Poisson Compound Poisson process (Huelsenbeck, 2000): Υποθέτουμε ότι εμφανίζονται αλλαγές στο ρυθμό κατά μήκος των γενεαλογιών, σύμφωνα με μια σημειακή διαδικασία και διακριτά γεγονότα αλλαγής του ρυθμού μεταβολής. Ως νέος ρυθμός θεωρείται ο παλιός πολλαπλασιασμένος με έναν γ-κατανεμημένο πολλαπλασιαστή.

- Αυτοσυσχετιζόμενοι ρυθμοί: Ρυθμοί αντικατάστασης εξελίσσονται σταδιακά κατά το μήκος του δέντρου

– Log-normally κατανεμημένοι ρυθμοί: ο ρυθμός σε έναν κόμβο αντλείται από μία κανονική λογαριθμική κατανομή με μέσο ίσο με το συνολικό ρυθμό όλου του κλάδου((Thorne et al., 1998), (Kishino et al., 2001), (Thorne and Kishino, 2002))

– Cox-Ingersoll-Ross Process: ο ρυθμός ενός θυγατρικού κλάδου προσδιορίζεται από μια μη-κεντρική χ ^ 2 κατανομή. Η διαδικασία αυτή περιλαμβάνει μια παράμετρο που καθορίζει την ένταση της δύναμης που οδηγεί τη διαδικασία στη στάσιμη κατανομή της (Lepage et al., 2006).

- Ασυσχέτιστοι ρυθμοί: Ο ρυθμός κάθε κλάδου λαμβάνεται από μια παραμετρική κατανομή όπως η κανονική λογαριθμική ή η εκθετική και θεωρείται ότι είναι μεταξύ τους ασυσχέτιστοι ((Drummond et al., 2006), (Rannala and Yang, 2007), (Lepage et al., 2007)).
- Πεπερασμένο μείγμα μοντέλων στους ρυθμούς των κλάδων: Οι κλάδοι λαμβάνουν διακριτούς ρυθμούς, σύμφωνα με τη διαδικασία Dirichlet (Heath et al., 2011).

Το συμπέρασμα είναι ότι το χαλαρό μοριακό ρολόι είναι ένα στατιστικό μοντέλο της μοριακής εξέλιξης που επιτρέπει στο ρυθμό εξέλιξης να ποικίλει μεταξύ των οργανισμών (στην περίπτωσή μας των δειγμάτων του HIV) (Ho, 2013).

Τα μοντέλα χαλαρού μοριακού ρολογιού λαμβάνουν υπόψιν τη διακύμανση του ρυθμού μεταξύ γενεαλογιών και έχουν προταθεί προκειμένου να επιτευχθεί καλύτερη εκτίμηση στη χρονολόγηση γεγονότων διασποράς (Drummond et al., 2006). Αντιπροσωπεύουν μια ενδιάμεση θέση μεταξύ των «αυστηρών» υποθέσεων του μοριακού ρολογιού και του μοντέλου Felsenstein των πολλαπλών ρυθμών (Felsenstein, 2001). Παλαιότερα, τέτοιες μπεϋζιανές προσεγγίσεις ήταν αδύνατες λόγω του υπολογισμού δύσκολων ολοκληρωμάτων και κατέστησαν δυνατές μέσω τεχνικών MCMC, που εξερευνούν ένα σταθμισμένο φάσμα τοπολογιών ενώ ταυτόχρονα εκτιμούν τις παραμέτρους του επιλεγμένου μοντέλου αντικατάστασης. όμως, πρέπει να υπενθυμίσουμε ότι η χρονολόγηση των γεγονότων διασποράς, που προκύπτει χρησιμοποιώντας ένα μοριακό ρολόι, βασίζεται σε στατιστική συμπερασματολογία και όχι σε άμεση απόδειξη.

Σκοπός της έρευνας

Ο ιός της ανθρώπινης ανοσοανεπάρκειας (HIV) είναι ο αιτιολογικός παράγοντας του συνδρόμου επίκτητης ανοσοανεπάρκειας (AIDS) και ανήκει στην κατηγορία των ρετροϊών. Ο ιός χαρακτηρίζεται από εκτενή γενετική ετερογένεια και ταξινομείται παγκοσμίως σε 4 ομάδες: την ομάδα M (major), την ομάδα O (outlier), την ομάδα N (new) και την ομάδα P που απομονώθηκε πρόσφατα στην Αφρική. Η επιδημία του HIV-1, που είναι συνέπεια μετάδοσης από τον αντίστοιχο ιό, μολύνει χιμπαντζήδες (SIVcpz) και έχει εκτιμηθεί ότι συνέβη στην Αφρική στις αρχές του 20ου αιώνα. Μετά την αρχική μετάδοση του ιού στους ανθρώπους, ο επιπολασμός της λοίμωξης HIV ήταν πολύ χαμηλός και η επιδημία ήταν εστιασμένη στην περιοχή της Κ. Αφρικής. Η διασπορά στο Δυτικό κόσμο συνέβη πρώτα στις ΗΠΑ μέσω της Αϊτής, που αποτέλεσε ενδιάμεσο σταθμό μετάδοσης από την Αφρική στις ΗΠΑ. Αυτό συνέβη γιατί αρκετοί Αϊτινοί επισκέφθηκαν την Κ. Αφρική τη δεκαετία του 1960 μέσω ενός προγράμματος της UNESCO και επέστρεψαν στην Αϊτή μετά από μερικά χρόνια. Τη δεκαετία του 1970, η

επιδημία εξελίχθηκε γρήγορα στις ΗΠΑ και στον υπόλοιπο δυτικό κόσμο λόγω της ασυμπτωματικής φάσης της λοίμωξης που διαρκεί περίπου για μια δεκαετία. Την περίοδο αυτή (δεκαετία 1970) εκτιμάται ότι συνέβησαν οι περισσότερες μεταδόσεις από τις ΗΠΑ στον υπόλοιπο δυτικό κόσμο και κυρίως στην Ευρώπη. Σε προηγούμενη μελέτη, εκτιμήσαμε τα γεγονότα παγκόσμιας διασποράς του υποτύπου Β και βρήκαμε ότι η Β. Αμερική έχει αποτελέσει πηγή μετάδοσης της επιδημίας αυτού του υποτύπου. Μέχρι σήμερα, δεν έχει εκτιμηθεί η χρονολόγηση της διασποράς αυτών των γεγονότων - μεταδόσεων. Λόγω του ότι η επιδημία δεν είχε διαγνωσθεί πριν το 1981, τα μοριακά δεδομένα αποτελούν τη μόνη αξιόπιστη πηγή για εκτιμήσεις που αφορούν στο παρελθόν της επιδημίας. Σκοπός της μελέτης είναι η εκτίμηση της χρονολόγησης των δεδομένων διασποράς της επιδημίας του υποτύπου Β με μεθόδους Μπεϋζιανής συμπερασματολογίας. Τα δεδομένα έχουν συλλεγεί σε συνέχεια συστηματικής ανασκόπησης της βιβλιογραφίας και η προηγούμενη ανάλυση (φυλογεωγραφία) είχε πραγματοποιηθεί σε αλληλουχίες περισσοτέρων από 8,000 οροθετικών. Η εκτίμηση της χρονολόγησης θα βασιστεί σε αυτά τα δεδομένα σε συνδυασμό με μεθόδους μέγιστης πιθανοφάνειας προς εκτίμηση της αρχικής τοπολογίας. Στην εργασία θα γίνει προσπάθεια να χρονολογηθεί για πρώτη φορά ένας τόσο μεγάλος αριθμός μοριακών δεδομένων.

Μέθοδοι

Συλλογή δεδομένων

Η μελέτη αυτή βασίζεται σε προηγούμενη μελέτη (Magiorkinis, 2009). Το σύνολο δεδομένων στην Magiorkinis et al έρευνα του 2009 αντλείται από 8.370 αλληλουχίες HIV-1 υποτύπου B, από 79 χώρες, σε όλο τον κόσμο. Η επιλογή όπως περιγράφεται στη προαναφερθείσα έρευνα έγινε με δύο στρατηγικές:

- Η προσθήκη δύο αναγνωρισμένων και Ευρωπαϊκών ερευνών (CATCH και SPREAD),
- Συστηματική έρευνα μέσα από τη βάση δεδομένων PubMed.

Ένα μεγάλο μέρος των δεδομένων που προέρχεται κυρίως από δύο ευρωπαϊκές μελέτες:

1) Την CATCH (combined Analysis of Resistance Transmission over Time of Chronically and Acute Infected HIV Patients) και

2) τη SPREAD (Strategy to Control SPREAD of HIV Drug Resistance).

Η SPREAD μελέτη συμπεριέλαβε 4.480 νεοδιαγνωσθέντες ασθενείς από τους οποίους ελήφθη δείγμα DNA την περίοδο 9/2002 - 12/2007, από 20 ευρωπαϊκές χώρες και το Ισραήλ. Η εκπροσώπηση της παγκόσμιας διασποράς του υποτύπου B σχεδιάστηκε για να συμπεριλάβει αντιπροσωπευτική δειγματοληψία από όλες τις χώρες ((Vercauteren et al., 2009), (Wensing, 2008)). Στην CATCH μελέτη συγκεντρώθηκαν 2.208 αλληλουχίες από 18 ευρωπαϊκές χώρες και το Ισραήλ την περίοδο 1996-2002 (Wensing et al., 2005). Και στις δύο μελέτες, οι ασθενείς ήταν ενήλικες χωρίς αντιρετροϊκή θεραπεία, ενώ για την τρέχουσα εργασία χρησιμοποιήθηκαν μόνο οι αλληλουχίες του υποτύπου B (τόσο από την CATCH όσο και από τη SPREAD). Το σύνολο των δεδομένων εμπλουτίστηκε με αλληλουχίες ανά τον κόσμο, οι οποίες συγκεντρώθηκαν μέσω συστηματικής

βιβλιογραφικής αναζήτησης στη βάση δεδομένων του PubMed, με τη χρήση συγκεκριμένων λέξεων – κλειδιών, όπως: "HIV-1", "molecular epidemiology", "resistance", "subtype B" και "pol", καθώς και συνδυασμούς αυτών.

Στη συνέχεια, εφαρμόστηκαν για την υπό-δειγματοληψία τα ακόλουθα κριτήρια:

i) Σε περιπτώσεις κατά τις οποίες περισσότερες από μία μελέτη ήταν διαθέσιμες για μία χώρα, συμπεριλήφθηκαν μόνο αλληλουχίες που απομονώνονται από διαφορετικές περιοχές της χώρας.

ii) Σε περιπτώσεις στις οποίες μελέτες από την ίδια χώρα δεν περιγράφονταν οι περιοχές δειγματοληψίας, λάβαμε υπόψη μόνο αλληλουχίες από τη μεγαλύτερη μελέτη για να αποφύγουμε περιττές αλληλουχίες.

iii) Ομοίως, για τις μελέτες που πραγματοποιήθηκαν στα ίδια κέντρα ή πόλεις.

iv)Από διαχρονικές μελέτες που αφορούν κυρίως στην αντοχή σε αντιρετροϊκή θεραπεία, συμπεριλάβαμε μόνο την παλαιότερη διαθέσιμη αλληλουχία ανά ασθενή και, τέλος,

v) Εξαιρέθηκαν μελέτες που αφορούν σε μετάδοση από τη μητέρα στο παιδί (Magiorkinis, 2009).

Σε προηγούμενες μελέτες, διάφορες μονοφυλετικές συστάδες εξήχθησαν χρησιμοποιώντας φυλογενετικά δέντρα Μέγιστης Πιθανοφάνειας. Εκείνα ταυτοποιήθηκαν ως υποδέντρα με έναν κοινό προγονικό κόμβο, των οποίων τα στελέχη είχαν υποβληθεί σε μια συγκεκριμένη περιοχή, και που αντιπροσωπεύαν περισσότερο από το 75% του συνόλου των στελεχών εντός του υποδέντρου. Κάθε μονοφυλετική συστάδα έπρεπε να αποτελείται από περισσότερες από 10 αλληλουχίες, προκειμένου να είναι επιλέξιμη για ένταξη. Αυτή η διαδικασία διεξήχθη με το χέρι, με οπτική επιθεώρηση χρησιμοποιώντας το πρόγραμμα Dendroscope (Huson and Scornavacca, 2012). Αυτή η διαδικασία μας εξήγαγε 25 συστάδες, με 3.510 αλληλουχίες συνολικά, που κατανέμονται, όπως φαίνεται στον παρακάτω πίνακα (Magiorkinis, 2009).

Μέγεθος μονοφυλετικών συστάδων

Cluster ID	*Total*	*Cluster ID*	*Total*

1	Cluster 0	4	Cluster 13	33
2	Cluster 1	207	Cluster 14	33
3	Cluster 2	118	Cluster 15	33
4	Cluster 3	185	Cluster 16	36
5	Cluster 4	143	Cluster 17	37
6	Cluster 5	61	Cluster 18	39
7	Cluster 6	53	Cluster 19	44
8	Cluster 7	51	Cluster 20	46
9	Cluster 8	58	Cluster 21	143
10	Cluster 9	37	Cluster 22	241
11	Cluster 10	46	Cluster 23	1790
12	Cluster 11	32	Cluster 24	5
13	Cluster 12	30	Cluster 25	5
	Overall	3510		

Επεξεργασία δεδομένων

Το πρόγραμμα που χρησιμοποιήθηκε για τη χρονολόγηση του HIV-1 υποτύπου Β ήταν η έκδοση BEAST 1.8.0, στην οποία πρέπει να αναφέρονται τα στοιχεία του έτους κατά το οποίο συλλέχθηκε το δείγμα του DNA. Με αυτόν τον τρόπο, δημιουργήθηκε ένα μοναδικό ID για κάθε αλληλουχία, συνδυάζοντας τις ήδη γνωστές πληροφορίες που φαίνονται παρακάτω:

• ID

• Η Χώρα από την οποία συλλέχθηκε το δείγμα DNA

• Η περιοχή όπου ανήκει η χώρα

• Η μονοφυλετική συστάδα όπου ανήκει, σύμφωνα με τις παλαιότερες έρευνες, και, τέλος,

• Η αλληλουχία του DNA.

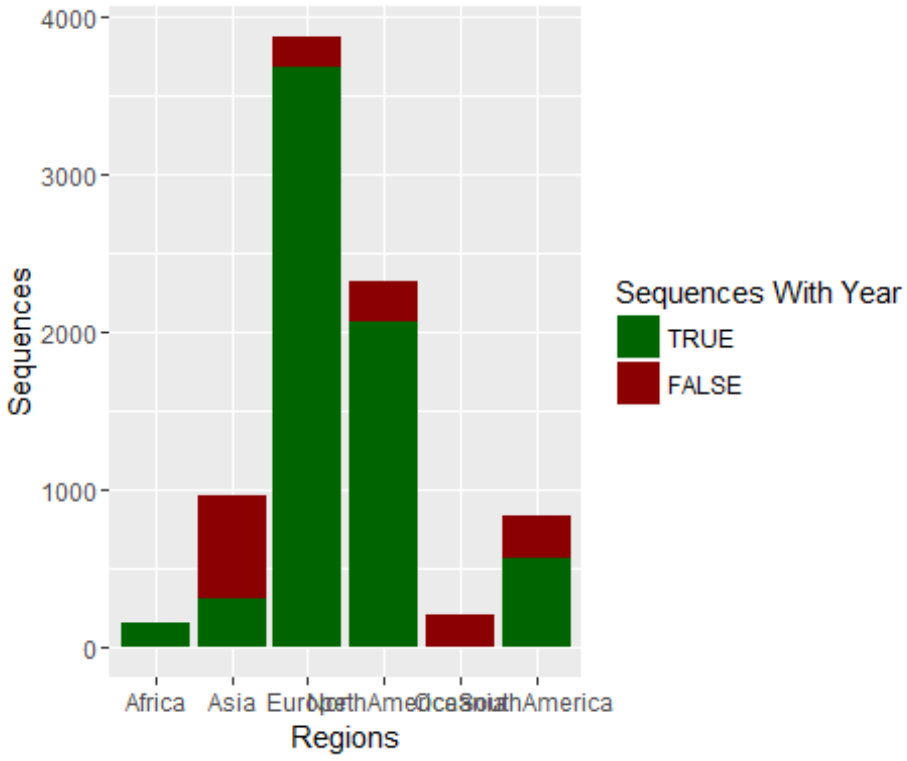
Σε διάφορες περιπτώσεις, η πληροφορία του έτους κατά το οποίο ελήφθη το δείγμα DNA από τους ασθενείς δεν έχει δηλωθεί, συνεπώς 1.603 αλληλουχίες, χωρίς αυτή την πληροφορία, αποκλείστηκαν από τη μελέτη. Το μέγεθος των δεδομένων μας μειώθηκε σε 6.767 αλληλουχίες και αυτό επηρέασε τις μονοφυλίες που είχαν αναφερθεί σε προηγούμενες έρευνες, όπως φαίνεται στον ακόλουθο πίνακα.

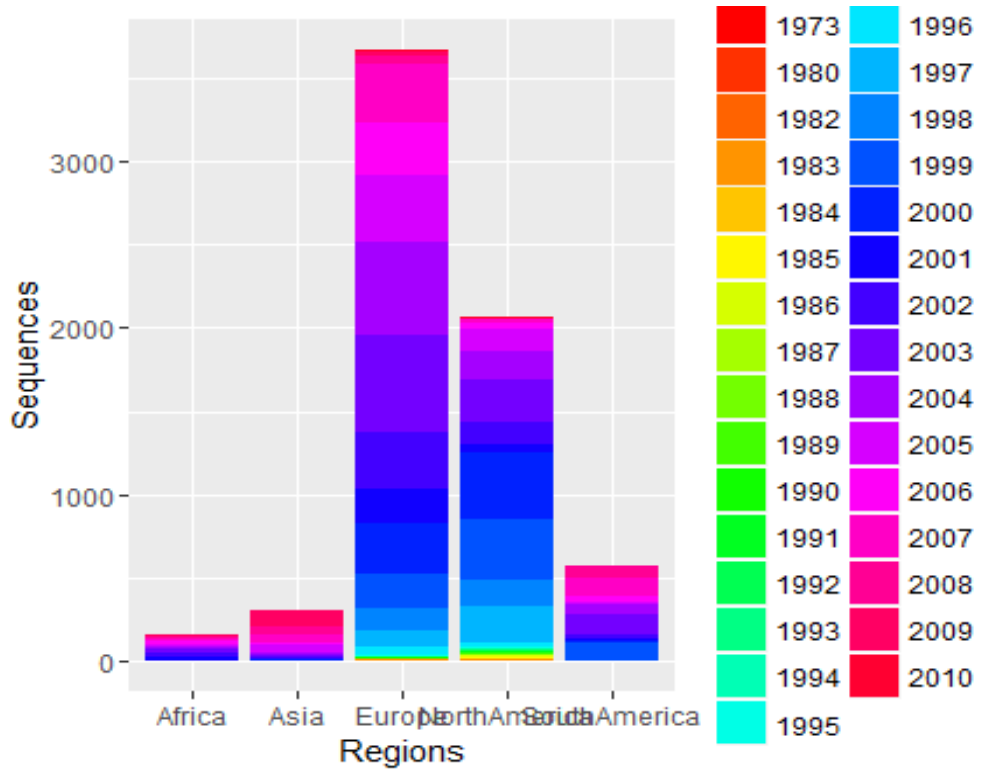
	Cluster ID	*With Sampling Years Frequency*	*Total*	*Percentage*
1	Cluster 0	4	4	100
2	Cluster 1	100	207	48.31
3	Cluster 2	16	118	13.56
4	Cluster 3	132	185	71.35
5	Cluster 4	91	143	63.64
6	Cluster 5	40	61	65.57

7	Cluster 6	15	53	28.3
8	Cluster 7	25	51	49.02
9	Cluster 8	27	58	46.55
10	Cluster 9	17	37	45.95
11	Cluster 10	39	46	84.78
12	Cluster 11	20	32	62.5
13	Cluster 12	19	30	63.33
14	Cluster 13	22	33	66.67
15	Cluster 14	26	33	78.79
16	Cluster 15	17	33	51.52
17	Cluster 16	27	36	75
18	Cluster 17	28	37	75.68
19	Cluster 18	34	39	87.18
20	Cluster 19	33	44	75
21	Cluster 20	40	46	86.96
22	Cluster 21	136	143	95.1
23	Cluster 22	230	241	95.44
24	Cluster 23	1728	1790	96.54
25	Cluster 24	0	5	0
26	Cluster 25	4	5	80
	Overall	2870	3510	81.77

24 Monophyletic Clustered Data-Set sizes as extracted in previews studies

Όπως μπορούμε να παρατηρήσουμε, η συστάδα 24 διεγράφη εντελώς, ενώ παρατηρήθηκε μεγάλη διακύμανση που κυμαίνεται από 13,5% έως 100% του ποσοστού των αλληλουχιών που περιλήφθηκαν. Επιπλέον, οι αλληλουχίες χωρίς πληροφορίες για την ημερομηνία δειγματοληψίας δεν ήταν αναλογικά κατανεμημένες μεταξύ των διαφόρων περιφερειών και έτσι οδηγηθήκαμε στον πλήρη αποκλεισμό της Ωκεανίας και τη δυσανάλογη συρρίκνωση των δεδομένων της Ασίας, σε σύγκριση με άλλες περιοχές από την έρευνά μας.





The regional distribution divided by year for the Dataset (with year information)

Η μείωση της δειγματοληψίας των αλληλουχιών

Οι αλληλουχίες στοιχίστηκαν με τη χρήση του προγράμματος mafft (έκδοση 7), ένα πρόγραμμα στοίχισης που χρησιμοποιεί τη μέθοδο στοίχισης πολλαπλών αλληλουχιών για λειτουργικά συστήματα Unix (Katoh, 2002). Η Speed-oriented μέθοδος προτιμήθηκε καθώς η accuracy-oriented μέθοδος θεωρήθηκε χρονοβόρα για 6.767 αλληλουχίες. Εξήχθησαν φυλογενετικά δέντρα μέγιστης πιθανοφάνειας για την εκ νέου επαλήθευση των μονοφυλετικών συστάδων, με τη χρήση του προγράμματος RaXmL με έκδοση 8.2.4 (Stamatakis, 2014) στο www.phylo.org server (Miller et al., 2010), ενώ χρησιμοποιήθηκε ένα τυχαίο παγκόσμιο δείγμα 5.000 αλληλουχιών ως σημείο αναφοράς που αντλήθηκε από τη LANL βάση δεδομένων (Foleyetal, 2015), που παρουσιάζεται στο Παράρτημα, με τις ακόλουθες επιλογές:

a. raxmlHPC-PTHREADS-SSE3 (multicore).
b. -T 4 (threads)
c. -f a (+ ML search)
d. -N auto MRE IGN - (a stopping Criterion) (Pattengale et al., 2010)
e. -m GTRGAMMA (GTRGAMMA model GTR with Gamma Distribution of rates across sites)
f. -p 123 -x 123 (Bootstrap seeds)

Προτείνεται η auto MRE_IGN επιλογή, προκειμένου να καθοριστεί το τέλος του φυλογενετικού bootstrapping, ιδιαίτερα για datasets με περισσότερες από 200 αλληλουχίες ((Stamatakis, 2014), (Pattengale et al., 2010)).

Το Generalised time-reversible μοντέλο ετερογένειας με γάμμα του ρυθμού αντικατάστασης θεωρήθηκε κατάλληλο, λόγω της ανισότητας των συχνοτήτων νουκλεοτιδικών βάσεων και των ρυθμών μεταλλάξεων (Posada and Crandall, 2001).

Οι 6.767 αλληλουχίες χωρίστηκαν σε δύο data sets ανάλογα με το αν ανήκουν σε μονοφυλετικές συστάδες, έτσι ώστε να καθοριστεί η μέθοδος υποδειγματοληψίας που θα χρησιμοποιηθεί. Οι αλληλουχίες που ανήκουν σε μονοφυλετικές ομάδες αναμένεται εξ ορισμού να έχουνμικρές γενετικές

αποστάσεις και κοινό τόπο προέλευσης, για αυτό μπορούμε να λάβουμε λιγότερες αλληλουχίες ώστε να εκπροσωπηθούν στην τελική φυλοδυναμική ανάλυση.

- 2.870 αλληλουχίες έχουν ομαδοποιηθεί σε 24 μονοφυλετικές ομάδες, ενώ
- 3.897 αλληλουχίες δεν θα μπορούσαν να ταξινομηθούν σε κάθε συστάδα. Τρία τυχαία δείγματα των 300, 500 και 700 αλληλουχιών εξήχθησαν, για να συνδυαστούν με τα δείγματα των μονοφυλετικών ομάδων.

Υποδειγματοληψία των μονοφυλετικών ομάδων

Δύο διαφορετικές στρατηγικές για την υποδειγματοληψία των μονοφυλετικών ομάδων διενεργήθηκαν:

• Ημιτυχαία επιλογή από κάθε μονοφυλετική ομάδα Και

• Επιλογή των πιο απόμακρων γενετικά αλληλουχιών.

Στην ημιτυχαία (semi-random) στρατηγική, 100 αλληλουχίες εξήχθησαν τυχαία, διατηρώντας τους παρακάτω τέσσερις κανόνες:

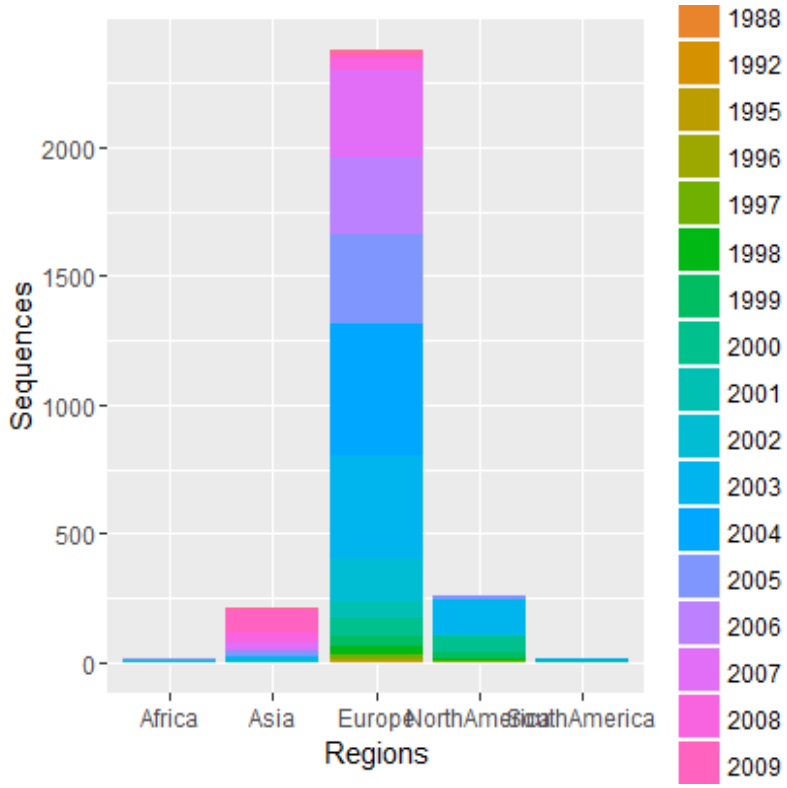
• Και οι 24 μονοφυλετικές ομάδες θα πρέπει να εκπροσωπούνται

• Όλες οι περιφέρειες θα πρέπει να δειγματοληφθούν από δείγμα αναλογικά, έτσι ώστε να διατηρηθεί η χωροδυναμική του ιού.

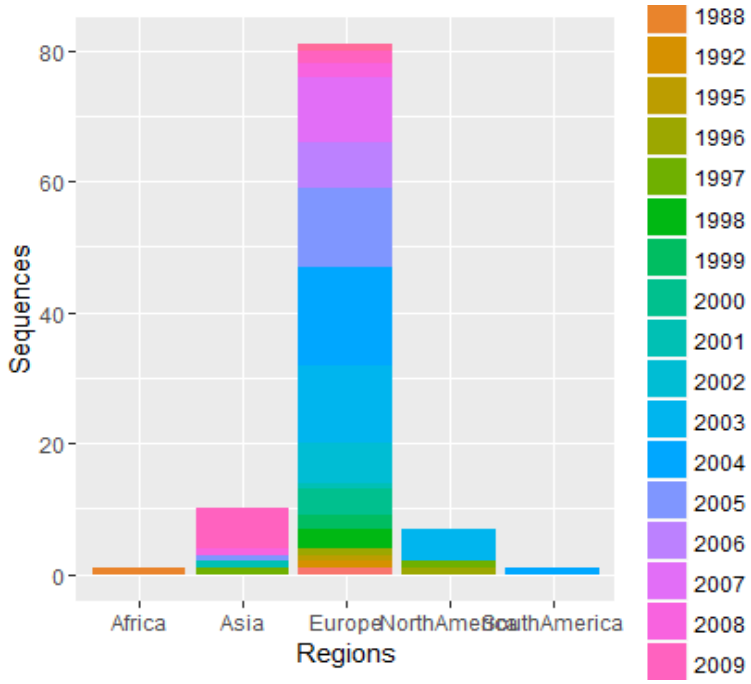
• Όλα τα έτη πρέπει να εκπροσωπούνται, προκειμένου να αυξηθεί το παράθυρο δειγματοληψίας των αλληλουχιών.

• Ένα έτος δεν μπορεί να εκπροσωπείται περισσότερο από 10 φορές, προκειμένου να βοηθήσει την αυτόματη βαθμονόμηση της Φυλοδυναμικής ανάλυσης.

Οι μονοφυλετικές ομάδες ήταν:



Η περιφερειακή και ανά έτος κατανομή των αλληλουχιών πριν την υποδειγματοληψία



Η περιφερειακή και ανά έτος κατανομή των αλληλουχιών πριν την υποδειγματοληψία.

Cluster	*sampled number of Sequences*	*Cluster*	*sampled number of Sequences*
0	3		
1	4	13	2
2	3	14	2
3	3	15	1
4	3	16	1
5	1	17	1
6	2	18	1
7	1	19	2
8	1	20	3
9	1	21	5
10	1	22	7
11	1	23	49
12	1	25	1

Αριθμός αλληλουχιών ανά μονοφυλετική ομάδα

Επιλογή βάσης γενετικής απόστασης

Για τη δεύτερη στρατηγική, επιλέχθηκαν 100 αλληλουχίες σύμφωνα με το μέγεθος της κάθε μονοφυλετικής ομάδας και τη γενετική απόσταση των αλληλουχιών. Για τις μικρές μονοφυλίες μεγέθους κάτω των 10 αλληλουχιών πήραμε 1 τυχαία αλληλουχία, για τις μεσαίου μεγέθους μονοφυλίες λάβαμε δείγμα ανάλογα με τη γενετική απόσταση που είχαν (2-4 επιλογές), ενώ για το Cluster 23 λάβαμε τυχαίο δείγμα 30 αλληλουχιών. Ένα boot strap consensus tree έχει εξαχθεί για κάθε μονοφυλετική ομάδα (Stamatakis, 2014), προκειμένου να προσδιοριστεί η γενετική απόσταση των αλληλουχιών (Appendix).

Clusters	*Samples*	*Clusters*	*Samples*
Cluster 0	1	Cluster 13	2
Cluster 1	4	Cluster 14	4
Cluster 2	3	Cluster 15	2
Cluster 3	3	Cluster 16	2

Cluster 4	4	Cluster 17	4
Cluster 5	4	Cluster 18	4
Cluster 6	2	Cluster 19	2
Cluster 7	2	Cluster 20	2
Cluster 8	4	Cluster 21	4
Cluster 9	2	Cluster 22	4
Cluster 10	3	Cluster 23	30
Cluster 11	3	Cluster 24	0
Cluster 12	4	Cluster 25	1
		Overall	100

Αριθμός αλληλουχιών ανά μονοφυλετική ομάδα

Χρονολόγηση αλληλουχιών

Ο χρόνος για τον πιο κοινό απόγονο (tMRCA) εκτιμήθηκε για τα έξι διαφορετικά παγκόσμια datasets. Χρησιμοποιήθηκε μπεϋζιανή προσέγγιση με το πρόγραμμα BEAST, έκδοσης 1.8.0 (Drummond and Bouckaert, n.d.) με GTR + G ως μοντέλο νουκλεοτιδικής αντικατάστασης. Η διόρθωση του μοριακού ρολογιού ήταν η uncorrelated log normal relaxed clock model (Drummond et al., 2006). Η απαραμετρική coalescent προσέγγιση επιλέχθηκε ως η καταλληλότερη, ενώ εξήχθη και το διάγραμμα Bayesian skyline (Drummond and Bouckaert, n.d.) Το καλύτερο φυλογενετικό με μεθόδους μεγίστης πιθανοφάνειας εξήχθη χρησιμοποιώντας το πρόγραμμα RaXmL (Stamatakis, 2014) και τοποθετήθηκε ως αρχικό δέντρο για να επιταχυνθεί η σύγκλιση της MCMC αλυσίδας. Χρησιμοποιήθηκαν 100 τυχαίοι operators σε μικρές αναλύσεις μοριακού ρολογιού με 30×10^6 (30,000,000) γενιές και burnin 3×10^6 ((3,000,000)), ενώ η δειγματοληψία ορίστηκε ανά χίλιες επαναλήψεις. Η διαδικασία αυτή μάς έδωσε ποιους operators να χρησιμοποιήσουμε, δηλαδή ποιο βήμα, πόσο συχνά και με ποια βαρύτητα θα κινείται ο MCMC αλγόριθμος.

Τελικά, διενεργήθηκε η ανάλυση μοριακού ρολογιού χρησιμοποιώντας Markov chain Monte Carlo (MCMC) αλγόριθμο και τρέξαμε 4 φορές κάθε data set για 300×10^6 (300,000,000) generations με burnin 30×10^5 δειγματοληψία ορίστηκε σε ανά χίλιες επαναλήψεις. Η σύγκλιση αξιολογήθηκε χρησιμοποιώντας το πρόγραμμα Tracer v. 1.6 (Rambaut A, 2014) και το Estimated Sample Size (ESS) έπρεπε να είναι μεγαλύτερο από 200. Το consensus δέντρο για κάθε ανάλυση εκτιμήθηκε οπτικά με το πρόγραμμα Tree Annotator (Drummond and Bouckaert, n.d.).

Λειτουργικά συστήματα και Στατιστικά πακέτα

Για τη διπλωματική χρησιμοποιήθηκαν 2 λειτουργικά συστήματα :

- Windows 8.1 - Windows 10
- Ubuntu Linux 15.04 - 16.04

Το βασικό στατιστικό πακέτο που χρησιμοποιήθηκε ήταν η R (R Core Team, 2015), R version 3.2.3 (2015-12-10), με την προσθήκη του R-Studio (R Studio Team, 2015) `RStudioDesktop 0.99.892`. Επιπλέον, χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python 2.7.11 για τον χειρισμό των αλληλουχιών.

Οι βιβλιοθήκες της R που χρησιμοποιήθηκαν ήταν οι παρακάτω:

- Για φυλογενετική ανάλυση, απεικόνιση δέντρων και χειρισμό δεδομένων οι:

a. phangorn (Schliep, 2011)

b. ape (Paradis et al., 2004)

c. Rphylip (Scott A. Chamberlain, 2013)

d. phytools (Revell, 2012)

e. ips (Heibl, 2008 onwards)

f. XML (Lang and CRAN Team, 2015)

g. ggtree (Yu et al., submitted)

- Άλλα πακέτα:

a. ggplot2 (Wickham, 2009)

b. knitr (Leisch and Peng, submitted)

c. rworldmap (South, 2011)

d. png (Urbanek, 2013)

e. RMySQL (Ooms et al., 2015)

Για τη συγγραφή, χρησιμοποιήθηκαν το πακέτο `Rmarkdown` (Allaire et al., 2016) και η **LaTeX**.

Αποτελέσματα

Στην παρούσα μελέτη, έχουμε χρησιμοποιήσει στοιχεία αλληλουχιών που έχουν συγκεντρωθεί από προηγούμενες μελέτες, ανά τον κόσμο (Magiorkinis, 2009), (Paraskevis et al., 2009). Το αρχικό σύνολο δεδομένων αποτελούνταν από 8.370 αλληλουχίες του HIV-1 υποτύπου Β που συλλέχθηκε από μολυσμένους ενήλικες ασθενείς. Αρχικά, αφού «επεξεργάστηκε» και αφαιρέσαμε αλληλουχίες που δεν ανέφεραν την ημερομηνία συλλογής του DNA, πραγματοποιήθηκε φυλογενετική ανάλυση για την επιβεβαίωση των μονοφυλετικών υποομάδων στις οποίες είχαν ομαδοποιηθεί σε πρότερες έρευνες (Magiorkinis, 2009), (Paraskevis et al., 2009). Μετέπειτα, στις αλληλουχίες οι οποίες ομαδοποιήθηκαν σε μονοφυλετικές υποομάδες διενεργήθηκε υποδειγματοληψία ακολουθώντας δύο στρατηγικές, μία ημιτυχαία δειγματοληψία 100 αλληλουχιών, με σεβασμό της αναλογικότητας των περιοχών αλλά και τη διεύρυνση του παράθυρου δειγματοληψίας και μια με επιλογή των 100 πιο γενετικά διαφορετικών αλληλουχιών εντός μιας μονοφυλίας. Από την άλλη πλευρά, σε όσες αλληλουχίες δεν ομαδοποιήθηκαν σε μονοφυλετικές υποομάδες διενεργήθηκαν τρεις τυχαίες δειγματοληψίες, και ελήφθησαν 3 datasets που περιέχουν 300, 500 και 700 αλληλουχίες, αντίστοιχα. Ο συνδυασμός των δύο παραπάνω συνόλων οδήγησε στη δημιουργία έξι datasets, μεγέθους 400, 600, 800 αλληλουχιών εις διπλούν.

Φυλογενετικά δέντρα μεγίστης πιθανοφάνειας, τα οποία ελήφθησαν με την χρήση του προγράμματος RaXmL, χρησιμοποιήθηκαν ως αρχικά δέντρα για την γρηγορότερη σύγκλιση της αλυσίδας. Εξ αρχής, είχαμε θέσει ως κριτήριο σύγκλισης την τιμή των 200 για το ESS (Effective Sampling Size) για τις παραμέτρους ενδιαφέροντος: likelihood, prior, posterior, treeModel.rootHeight και ucld.mean.

Σύμφωνα με τους Drummond et al. (2006), το Effective Sampling Size (ESS) μιας παραμέτρου του δείγματος μιας αλυσίδας MCMC είναι ο αριθμός των πραγματικά ανεξάρτητων επιλογών από την εκ των υστέρων πιθανότητα, με τις οποίες η Μαρκοβιανή αλυσίδα είναι ισοδύναμη, και υπολογίζεται με τη μέτρηση της συσχέτισης μεταξύ δειγμάτων της αλυσίδας. Οι μέθοδοι που προτείνονται για να βελτιστοποιηθεί η τιμή της ESS είναι οι εξής:

• Η αύξηση του μήκους της αλυσίδας, δηλαδή του αριθμού των γενεών.

• Ο συγχρονισμός παραμέτρων κίνησης MCMC

• Η αύξηση της συχνότητας δειγματοληψίας

• Ο συνδυασμός αποτελεσμάτων πολλαπλών ανεξάρτητων MCMC αναλύσεων.

Ο συγχρονισμός των παραμέτρων κίνησής του MCMC αλγόριθμου, που διενεργήθηκε για τη βελτιστοποίηση της εκ των προτέρων κίνησης της Μόντε Κάρλο Μαρκοβιανής Αλυσίδας (Drummond et al., 2006) έγινε με τη διεξαγωγή μικρών αναλύσεων μοριακού ρολογιού με 30,000,000 γενιές και 3,000,000 burn-in, χρησιμοποιώντας τυχαίες τιμές παραμέτρων (operators).

Για την αξιολόγηση της κίνησης της αλυσίδας χρησιμοποιήσαμε το πρόγραμμα Tracer, έκδοσης 1.6 (Rambaut A, 2014), με κριτήρια την τιμή του ESS (Effective Sampling Size) και με οπτικά κριτήρια σύγκλισης τις τιμές των παραμέτρων με την καλύτερη απόδοση. Για να μπορέσουμε να συγκρίνουμε τα τελικά αποτελέσματα των Monte Carlo αλληλουχιών διατηρήσαμε σταθερό seed σε κάθε ανάλυση

Οι παράμετροι οι οποίοι μας ενδιέφεραν κυρίως ήταν οι :

- Posterior
- prior
- likelihood
- treemodel.rootheight
- ucld.mean

Ενδεικτικά:

	file 1	file 2	file 3	file 4	file 5	file 6	file 7	file 8	file 9	file 10
Posterior	44	193	78	18	109	36	105	55	43	78
prior	44	149	77	43	83	44	79	73	107	78
likelihood	36	93	56	18	68	23	49	64	35	91
treemodel.routeheight	202	267	30	136	156	101	51	114	22	24
ucld.mean	193	331	143	247	181	158	203	154	227	234
	2	4	1	2	3	1	2	2	2	1

Διεξήχθησαν τέσσερις αναλύσεις μοριακού ρολογιού για κάθε σύνολο δεδομένων, με τη χρήση του προγράμματος BEAST, έκδοσης 1.8.0, με 300,000,000 γενιές, 30,000,000 burn-in και δειγματοληψία ανά 1,000 βήματα. Το αποτέλεσμα ήταν να επιτευχθεί σύγκλιση μόνο για τα μικρά σύνολα δεδομένων μεγέθους 400 αλληλουχιών και στις 4 αναλύσεις. Επίσης, αξίζει να αναφερθεί ότι στο συγχρονισμό παραμέτρων και τις τελικές αναλύσεις μοριακού ρολογιού, τα ημιτυχαία δείγματα είχαν ταχύτερη σύγκλιση και πέτυχαν καλύτερες τιμές στο ESS, γεγονός το οποίο μας οδηγεί στο συμπέρασμα ότι ενδεχομένως η δειγματοληψία ως προς την περιφέρεια και την ημερομηνία να είναι καλύτερη από εκείνη που έχει ως βάση τη γενετική ετερογένεια.

Διατηρήσαμε τα αποτελέσματα της «καλύτερης» ανάλυσης, η οποία ήταν η ημιτυχαία μεγέθους 400 αλληλουχιών. Ο χρόνος του πιο πρόσφατου κοινού πρόγονου (tMRCA) εκτιμήθηκε από την παράμετρο treeModel.rootHeight, 50 χρόνια πριν από τις 1-1-2009 που ήταν η τελευταία ημερομηνία δειγματοληψίας και με διάστημα αξιοπιστίας σε ποσοστό 95% για την περίοδο από 40,8 έως 60,5 έτη.

Lower 95%	**Mean**	**Upper 95%**
1948	1959	1969

tMRCA εκτίμηση με 95% CI

Αν r_i είναι ο ρυθμός για τον ί-οστό κλάδο, τότε το ucld.mean είναι απλά το άθροισμα των r_i για όλα τα i διαιρούμενα με τον αριθμό τους (2n-2). Είναι, δηλαδή, ο απλός αριθμητικός μέσος των ρυθμών αντικατάστασης των κλάδων. Δεδομένου ότι ορισμένοι κλάδοι αντιπροσωπεύουν πολύ περισσότερο χρόνο συγκριτικά με άλλους, αυτό ο ρυθμός δεν θα είναι απαραίτητα ο ίδιος με το συνολικό αριθμό των αντικαταστάσεων ανά τοποθεσία διαιρούμενο με το συνολικό ποσό του χρόνου που αντιπροσωπεύει το δέντρο. Έτσι, το αποτέλεσμα του μέσου ρυθμού σημειακής μετάλλαξης του HIV είναι 0.0024 ανά θέση και ανά

έτος, ενώ το αποτέλεσμα του ucld.stdevήταν≃ 0.27 με διάστημα Αξιοπιστίας 95% [0.23, 0.31] , το οποίο μας φανερώνει ότι τα δεδομένα μας έχουν μια συμπεριφορά η οποία μπορεί να προσομοιαστεί με εκείνη ενός ρολογιού. Επιπλέον, στο Bayesian Skyline Plot φαίνεται ότι το μέγεθος του πληθυσμού αυξήθηκε αστραπιαία κατά τη δεκαετία του '60 και του '70, κατά προσέγγιση, ενώ φαίνεται να επιβραδύνεται και να σταθεροποιείται κατά τη δεκαετία του '80.

Συζήτηση

Στη παρούσα διπλωματική, χρονολογήσαμε την πηγή προέλευσης του HIV υποτύπου B με τη χρήση ενός μεγάλου dataset 8.370 αλληλουχιών. Δεδομένου ότι η υπολογιστική ισχύς μιας χρονολόγησης τόσων πολλών αλληλουχιών είναι τεράστια και η πιθανότητα εύρεσης αποτελέσματος μηδαμινή, θεωρήθηκε αναγκαία η υποδειγματοληψία του προαναφερθέντος dataset, κατά τέτοιο τρόπο ώστε το δείγμα να είναι αντιπροσωπευτικό του αρχικού συνόλου δεδομένων.

Η Μπεϋζιανή προσέγγιση έχει αποδειχθεί ως βέλτιστη για τη χρονολόγηση του HIV, αλλά και γενικότερα για ταχέως μεταλλασσόμενους ιούς (Drummond et. al., 2005) (Magiorkinis et.al, 2009), καθώς συνδυάζει πρότερη γνώση και έτσι μπορεί

να δώσει αξιόπιστα αποτελέσματα πιο γρήγορα. Σε αυτή τη λογική, χρησιμοποιήσαμε φυλογενετικά δέντρα μεγίστης πιθανοφάνειας ως αρχική τοπολογία, ενώ διενεργήθηκε συγχρονισμός των παραμέτρων της κίνησης της αλυσίδας Μόντε Κάρλο.

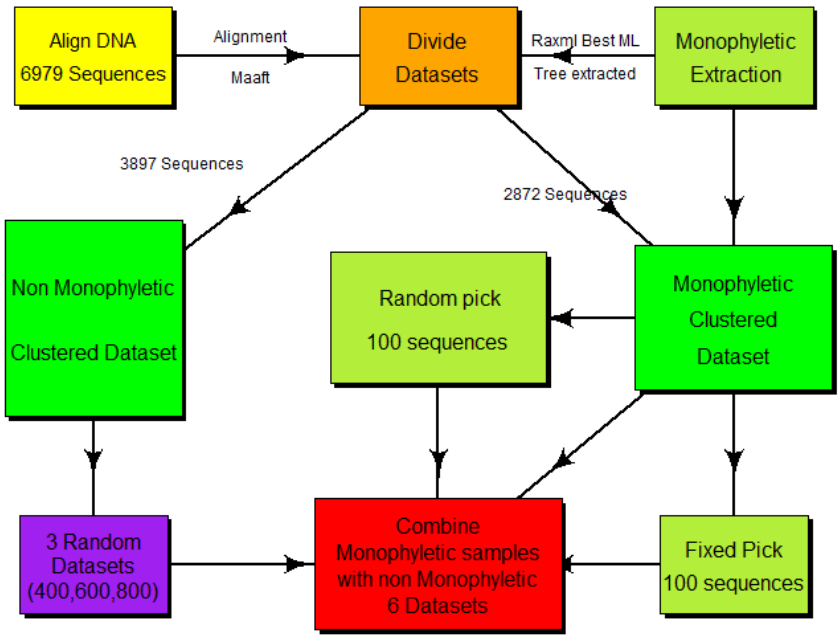
Στην εργασία, μέσω της ανάλυσης, φάνηκε ότι ενδεχομένως η υποδειγματοληψία με κριτήρια την περιοχή και την ημερομηνία συλλογής των δειγμάτων υπερτερεί από εκείνη των γενετικών αποστάσεων. Επιπλέον, ο συγχρονισμός των παραμέτρων της αλυσίδας Monte Carlo φαίνεται να παίζει πολύ σημαντικό ρόλο στην ταχύτητα της σύγκλισης της και ενδεχομένως να χρίζει περισσότερης έρευνας.

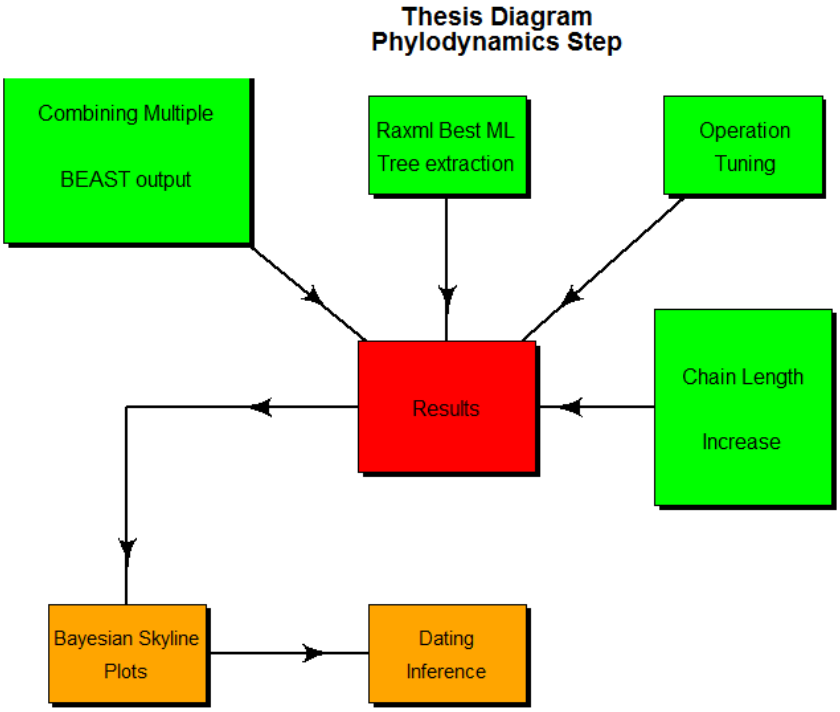
Τα ευρήματά μας συμβαδίζουν πλήρως με τη βιβλιογραφία για τη διασπορά του HIV και την θεωρία ότι ήταν ήδη πανδημία κατά την πρώτη ταύτισή του ως αιτιολογικός παράγοντας του AIDS, όπως παρουσιάζονται από το Bayesian Skyline Plot.

Appendix

The diagram below shows the work that was done.

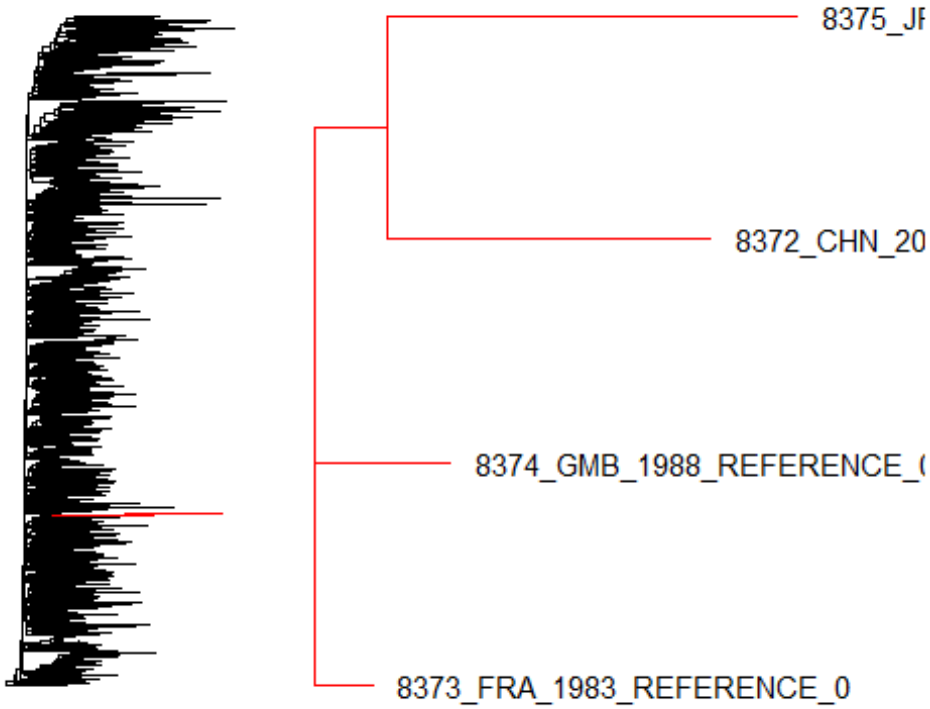
Thesis Diagram
1st Step



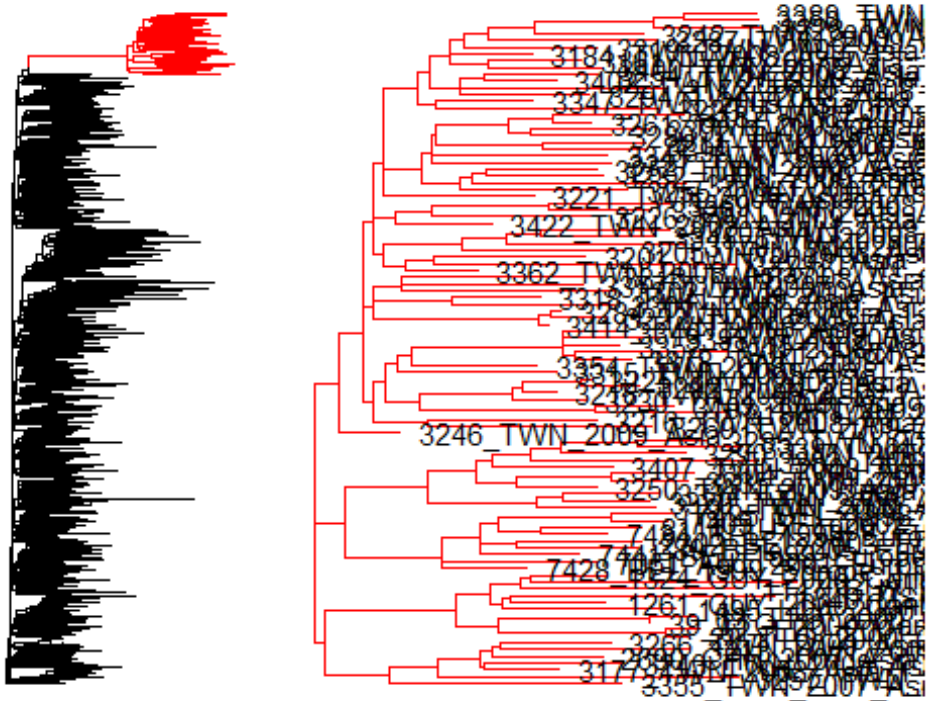


This way we checked the 24 clusters if they are monophyletic. The results are shown below

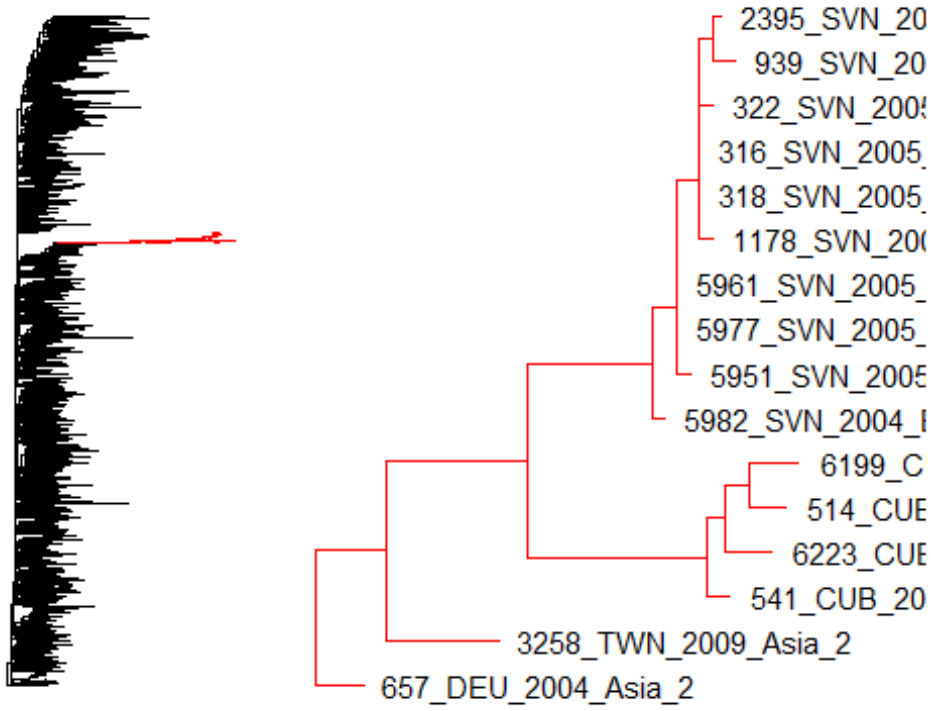
```
## [1] "Cluster 0"
```



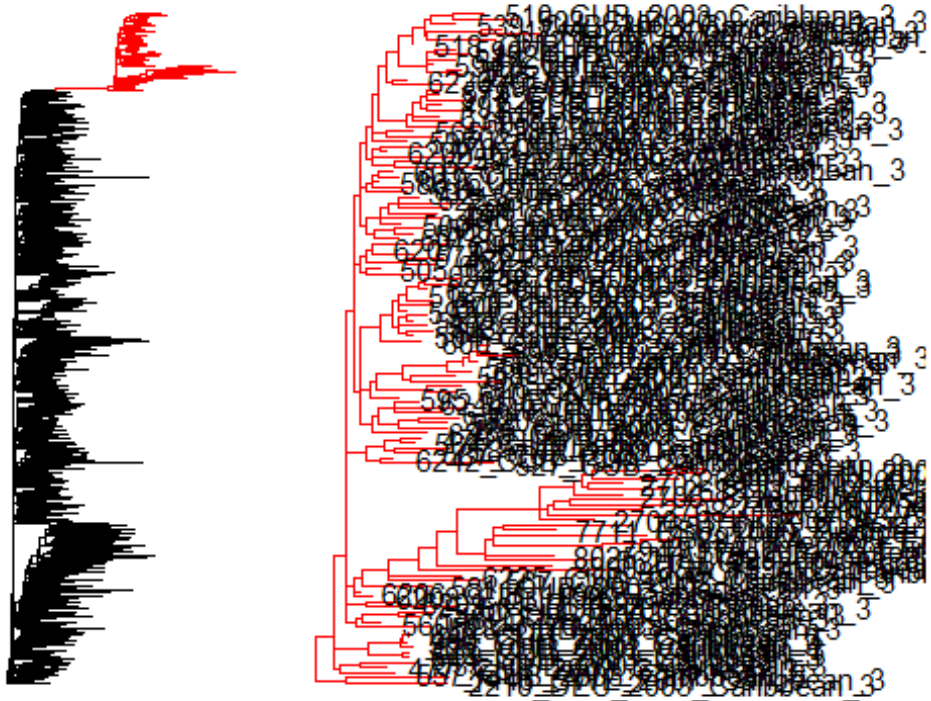

```
## [1] "Cluster 1"
```



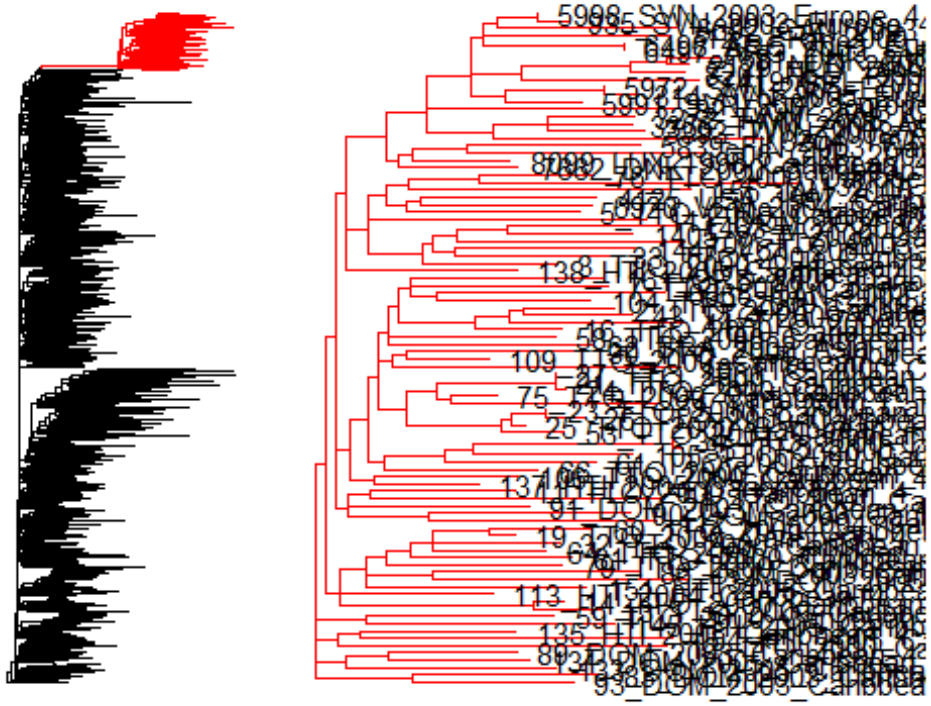

```
## [1] "Cluster 2"
```



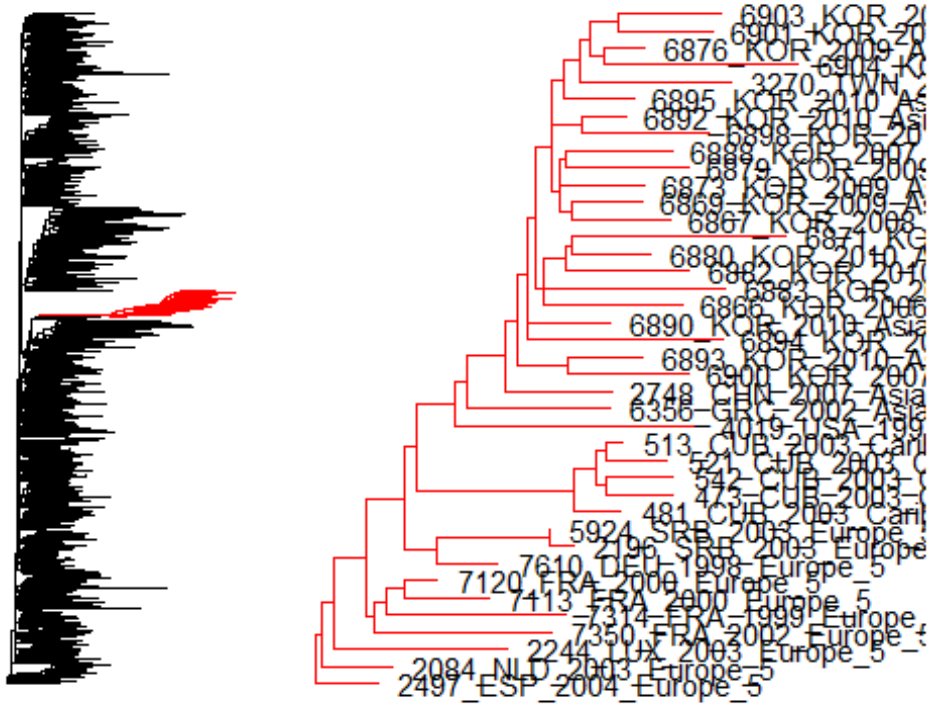

```
## [1] "Cluster 3"
```



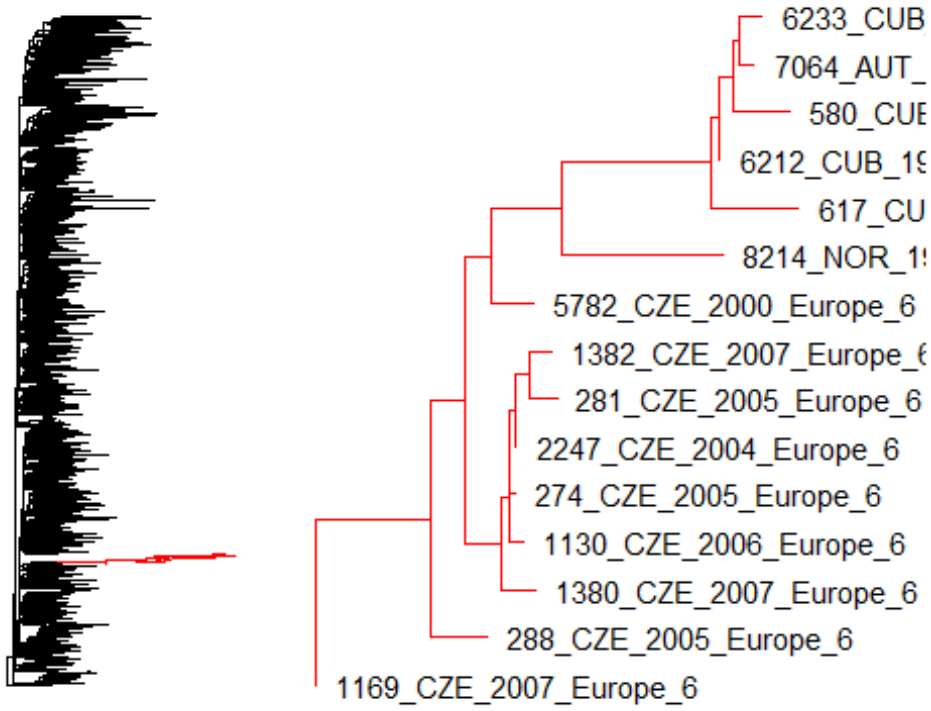
```
## [1] "Cluster 4"
```



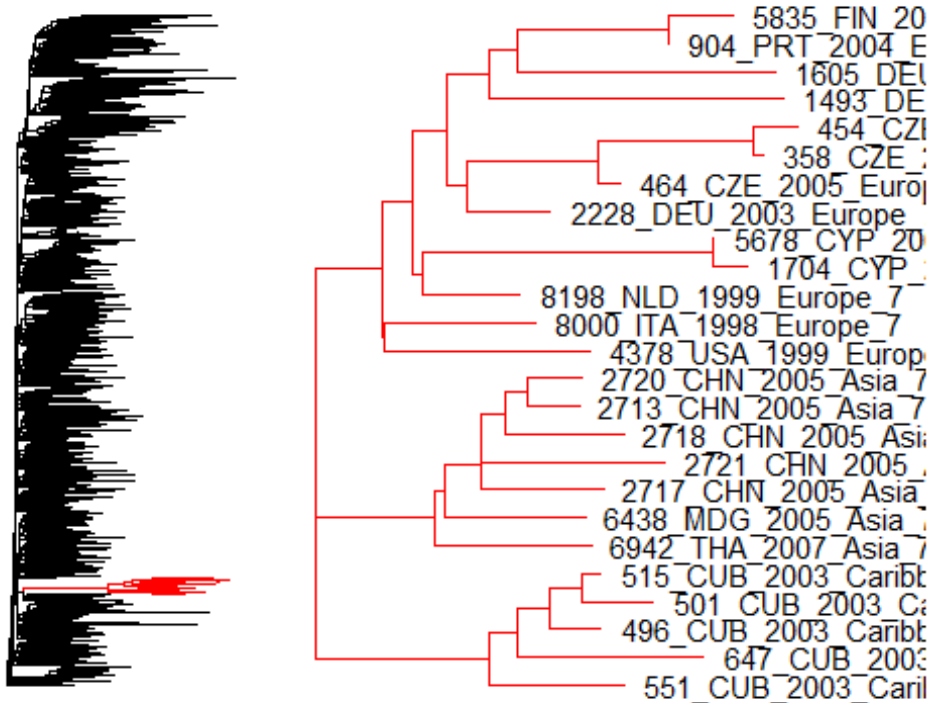
```
## [1] "Cluster 5"
```



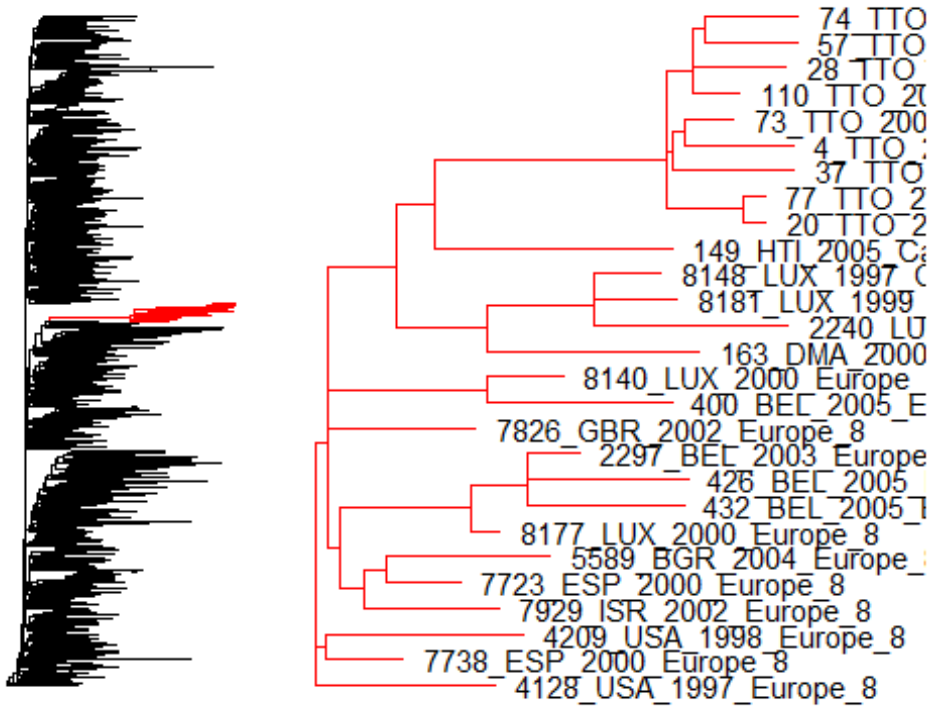
```
## [1] "Cluster 6"
```



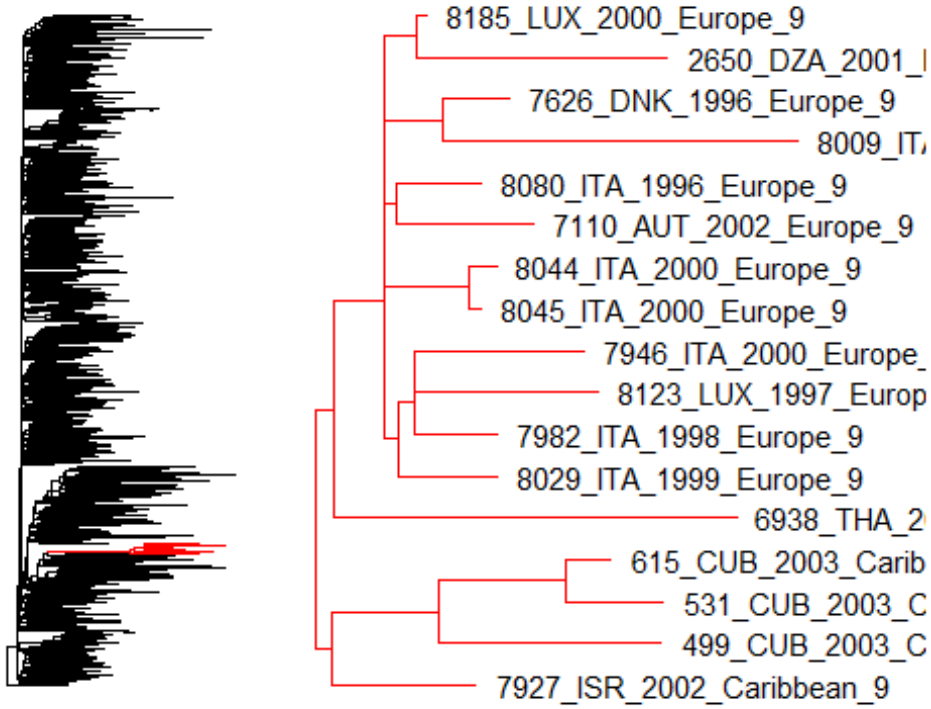

```
## [1] "Cluster 7"
```



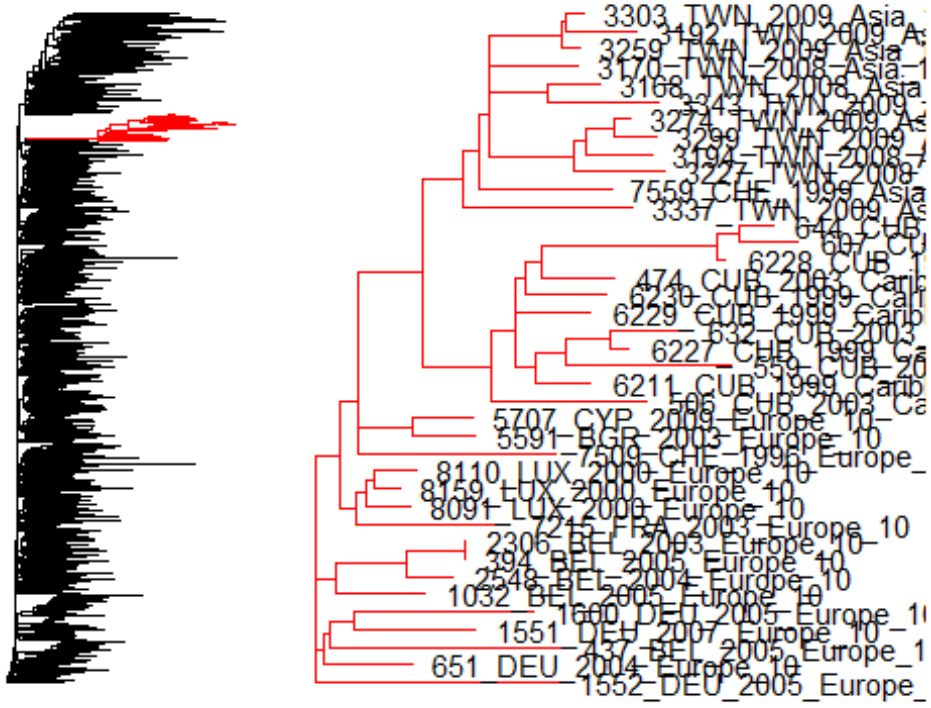

```
## [1] "Cluster 8"
```



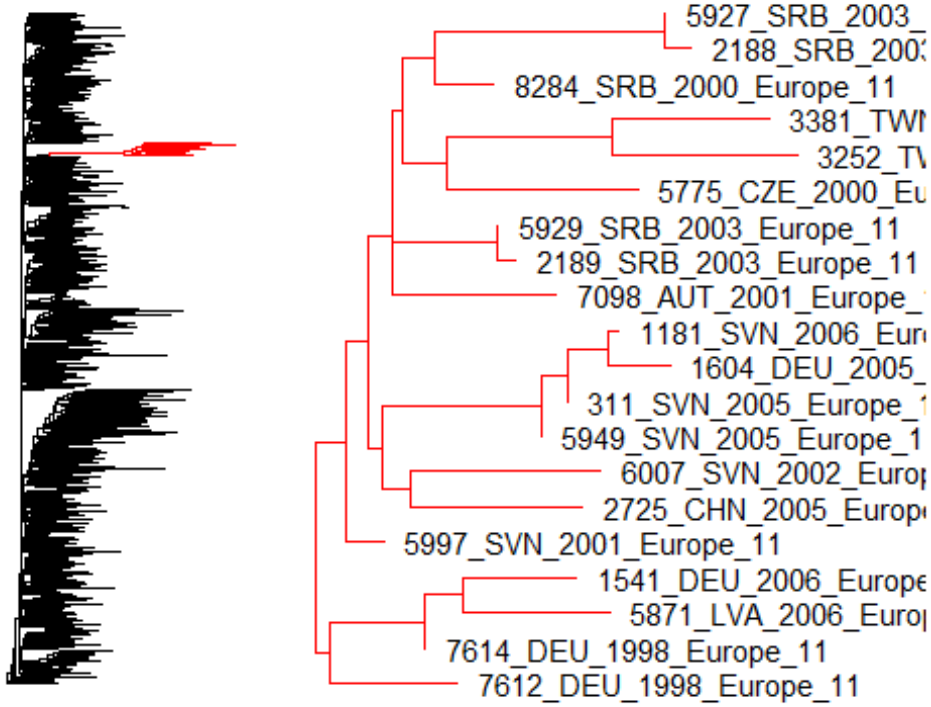

```
## [1] "Cluster 9"
```



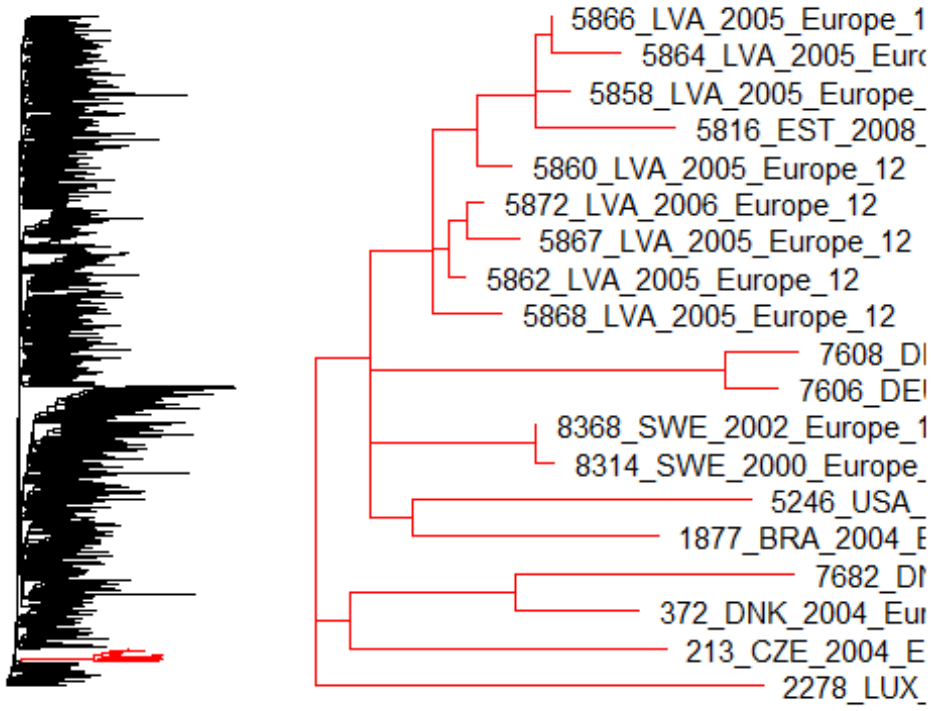

```
## [1] "Cluster 10"
```



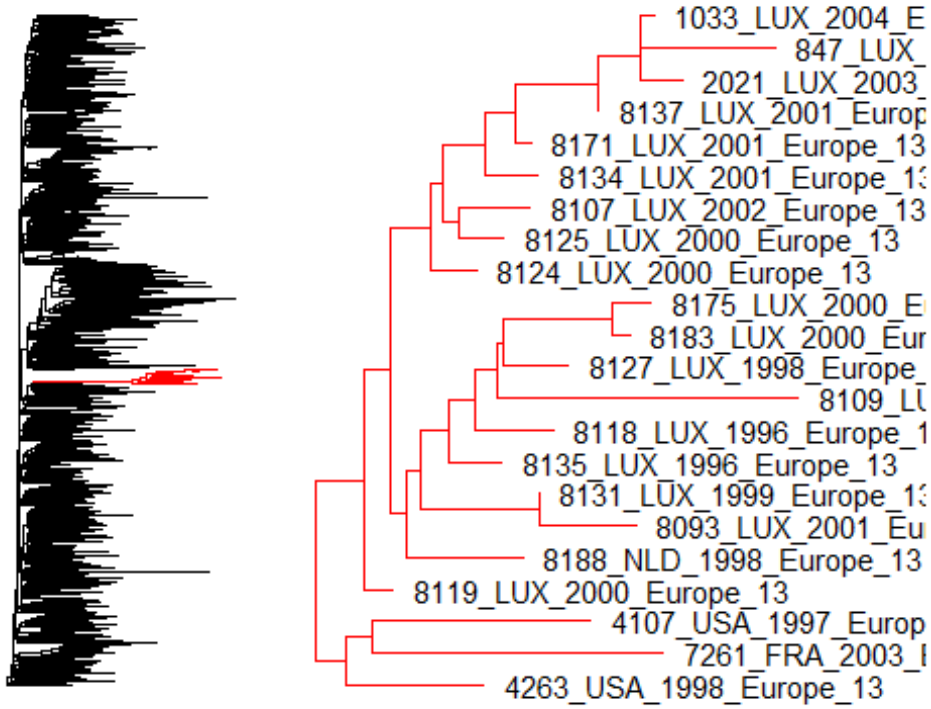

```
## [1] "Cluster 11"
```



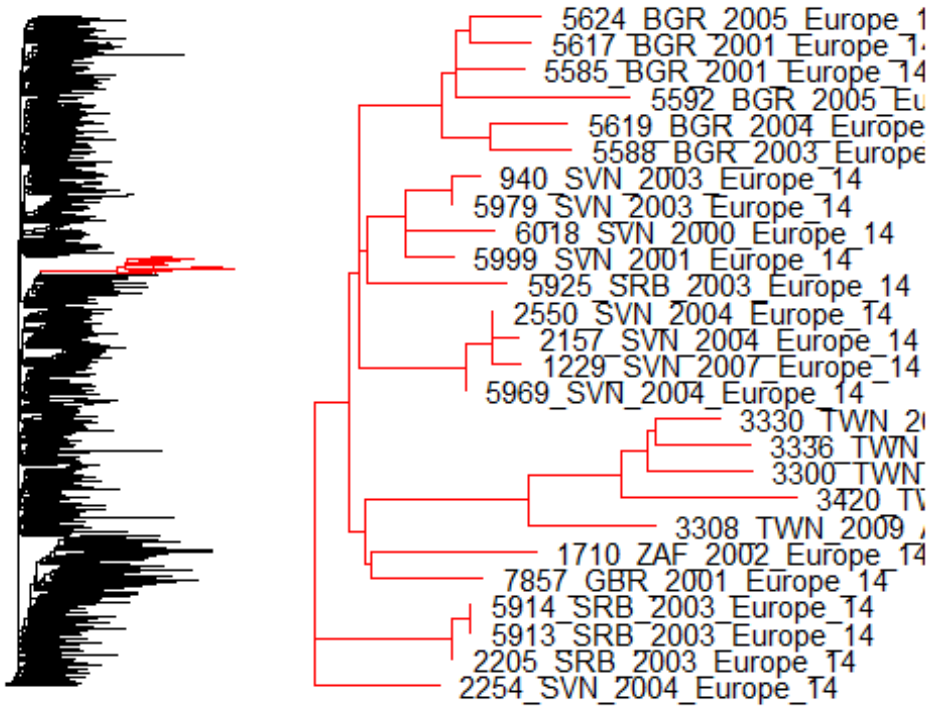

```
## [1] "Cluster 12"
```



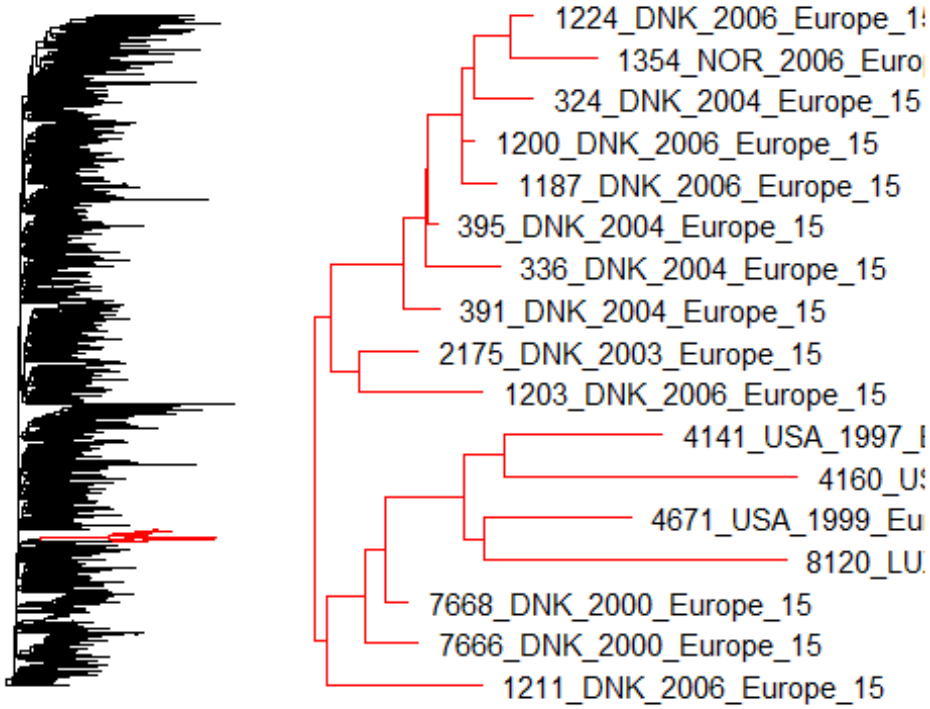

```
## [1] "Cluster 13"
```



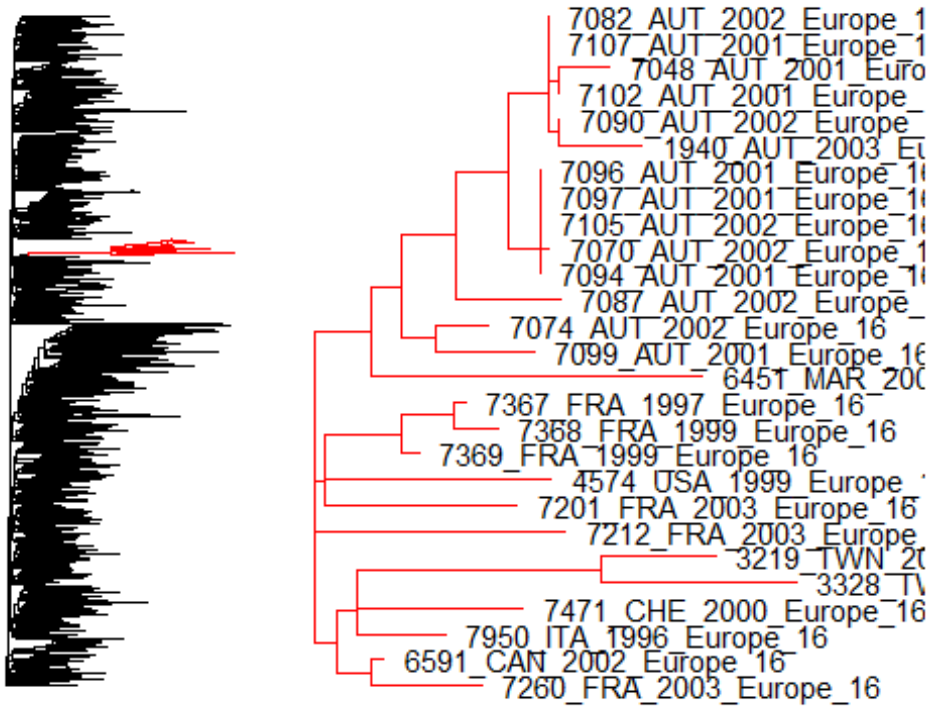

```
## [1] "Cluster 14"
```



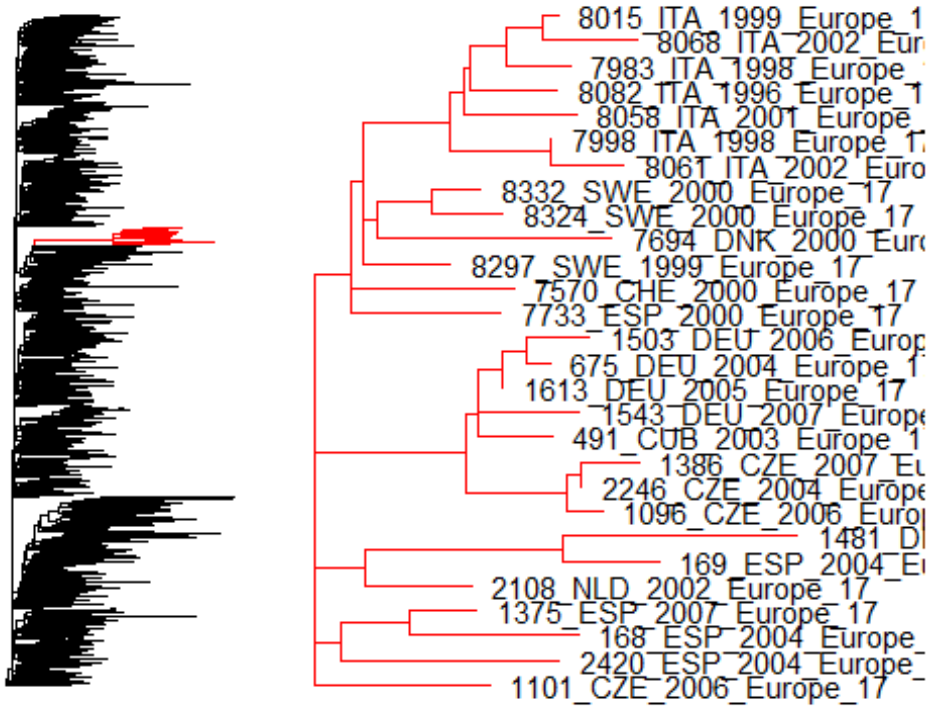

```
## [1] "Cluster 15"
```



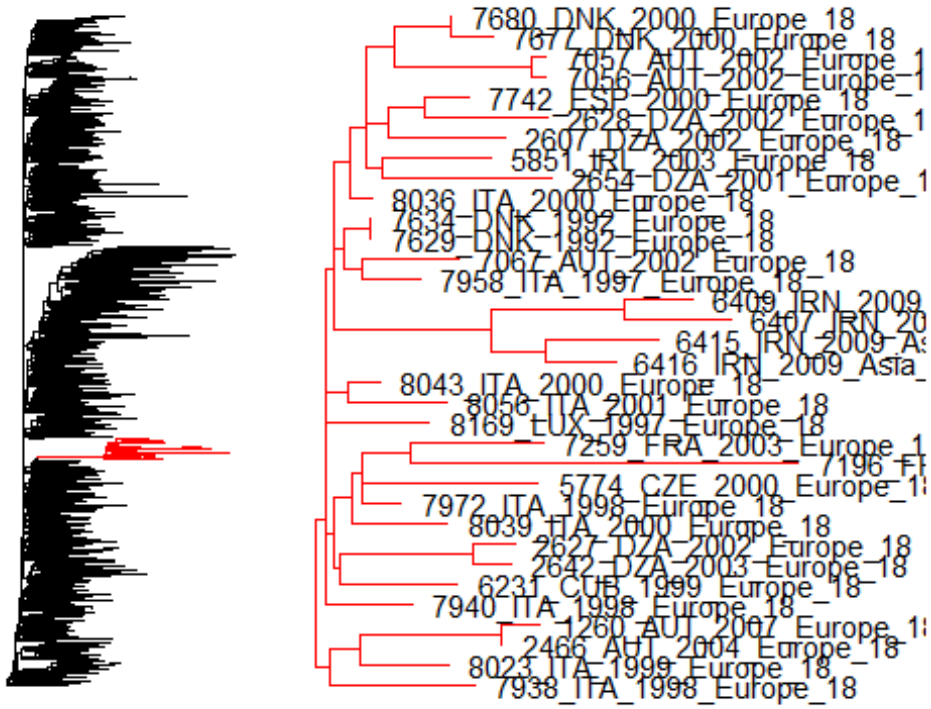

```
## [1] "Cluster 16"
```



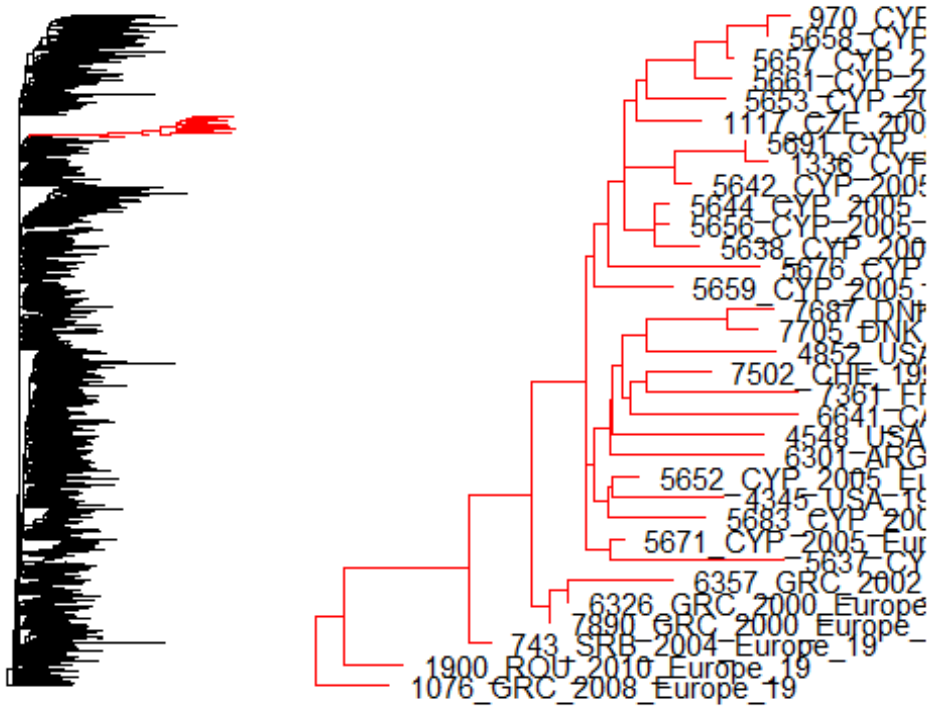

```
## [1] "Cluster 17"
```



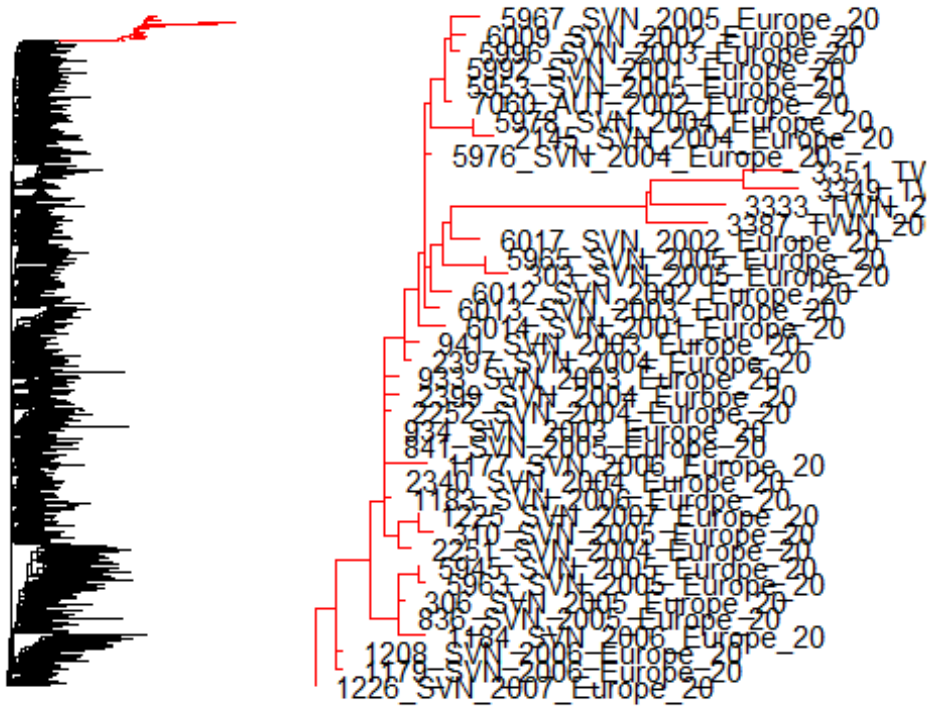

```
## [1] "Cluster 18"
```



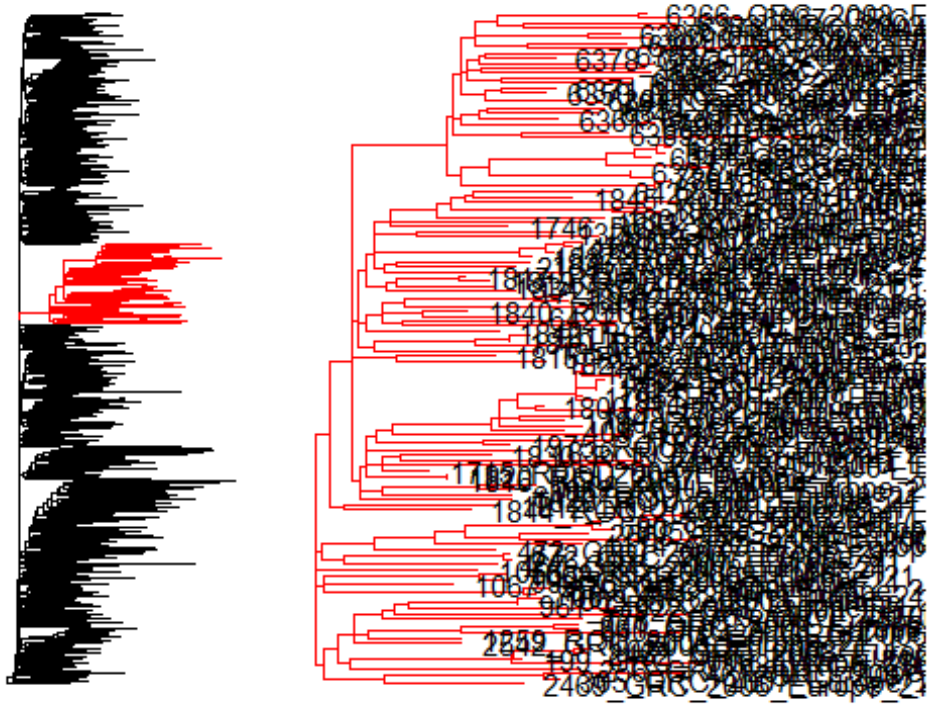

```
## [1] "Cluster 19"
```



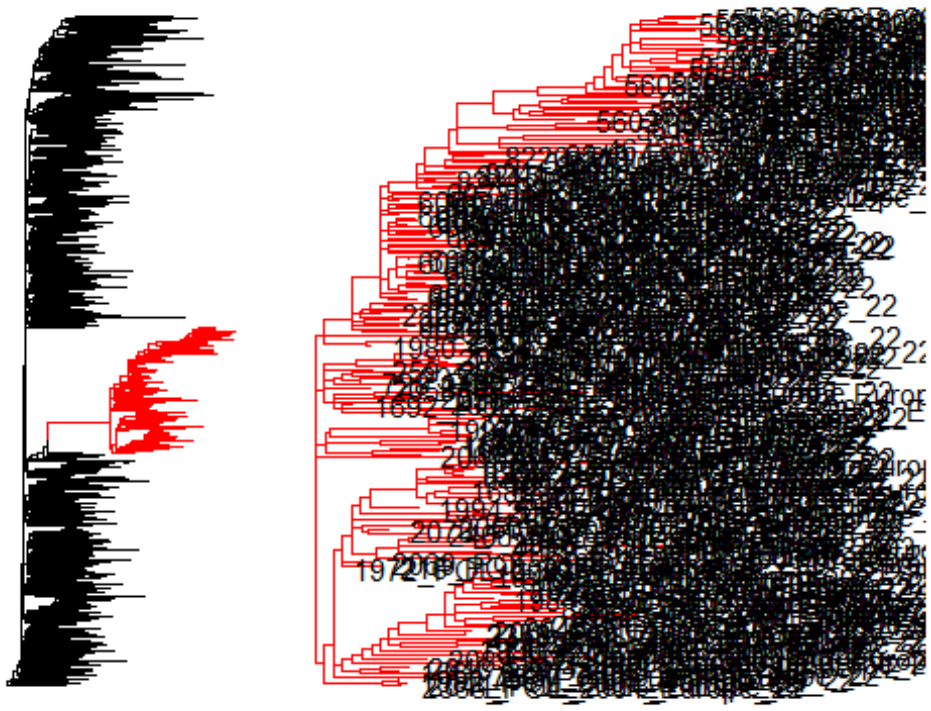

```
## [1] "Cluster 20"
```



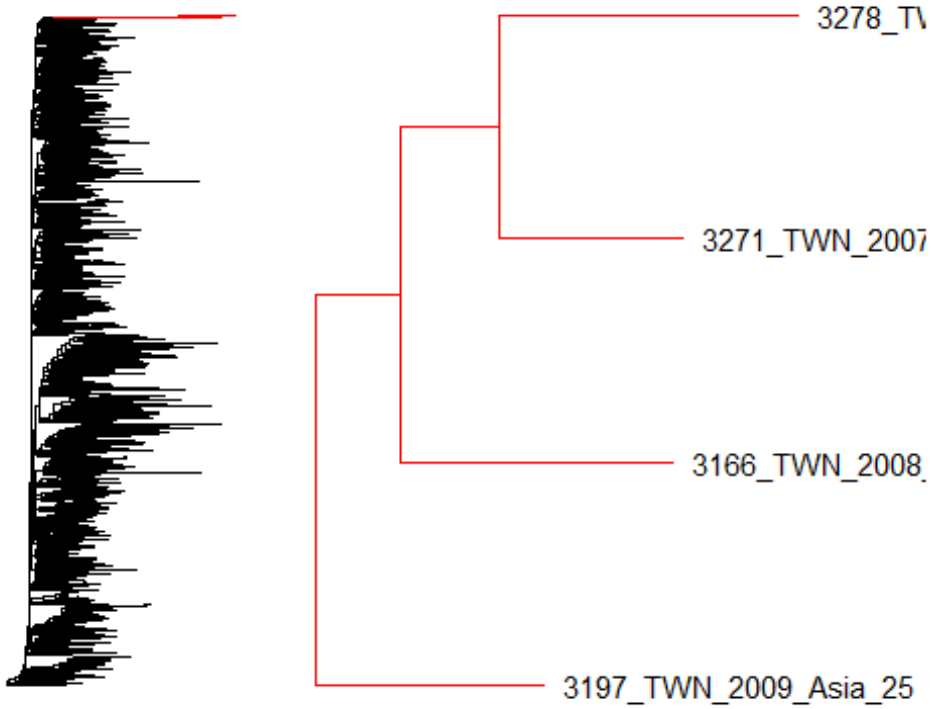

```
## [1] "Cluster 21"
```



```
## [1] "Cluster 22"
```



```
## [1] "Cluster 25"
```



References

Abecasis, A.B., Wensing, A.M., Paraskevis, D., Vercauteren, J., Theys, K., Vijver, D.A.V. de, Albert, J., Asjö, B., Balotta, C., Beshkov, D., Camacho, R.J., Clotet, B., Gascun, C.D., Griskevicius, A., Grossman, Z., Hamouda, O., Horban, A., Kolupajeva, T., Korn, K., Kostrikis, L.G., Kücherer, C., Liitsola, K., Linka, M., Nielsen, C., Otelea, D., Paredes, R., Poljak, M., Puchhammer-Stöckl, E., Schmit, J.-C., Sönnerborg, A., Stanekova, D., Stanojevic, M., Struck, D., Boucher, C.A., Vandamme, A.-M., 2013. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in europe suggest highly compartmentalized epidemics. Retrovirology 10, 7.

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., Hyndman, R., 2016. Rmarkdown: Dynamic documents for r.

Ariën, K.K., Vanham, G., Arts, E.J., 2007. Is HIV-1 evolving to a less virulent form in humans? Nature Reviews Microbiology 5, 141–151.

Barouch, D.H., 2008. Challenges in the development of an HIV-1 vaccine. Nature 455, 613–619.

Barre-Sinoussi, F., Chermann, J., Rey, F., Nugeyre, M., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., Rozenbaum, W., Montagnier, L., 1983. Isolation of a t-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220, 868–871.

Bette Korber, B.F.H., Christian Brander, Watkins, D.I. (Eds.), n.d. HIV molecular

immunology database 1999. Theoretical Biology and Biophysics.

Beyrer, C., Sullivan, P., Sanchez, J., Baral, S.D., Collins, C., Wirtz, A.L., Altman, D., Trapence, G., Mayer, K., 2013. The increase in global HIV epidemics in MSM. AIDS 27, 2665–2678.

Buonaguro, L., Tornesello, M.L., Buonaguro, F.M., 2007. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: Pathogenetic and therapeutic implications. Journal of Virology 81, 10209–10219.

Camacho R., C., 2006. The significance of subtype-related genetic variability: Controversies and unanswered questions. Mediscript.

Cohen, M.S., Shaw, G.M., McMichael, A.J., Haynes, B.F., 2011. Acute HIV-1 infection. New England Journal of Medicine 364, 1943–1954.

D. L. Swofford, P.J.W., G. J. Olsen, 1996. Phylogenetic inference in molecular systematics (2nd ed.) 407–514.

Drosopoulos, W.C., Rezende, L.F., Wainberg, M.A., Prasad, V.R., 1998. Virtues of being faithful: Can we limit the genetic variation in human immunodeficiency virus? Journal of Molecular Medicine 76, 604–612.

Drummond, A.J., Bouckaert, R.R., n.d. Bayesian evolutionary analysis by sampling trees. In: Bayesian Evolutionary Analysis with BEAST. Cambridge University Press (CUP), pp. 79–96.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology 4, e88.

Drummond, A.J., Suchard, M.A., 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biology 8, 114.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22: 1185–1192

Faria, N.R., Rambaut, A., Suchard, M.A., Baele, G., Bedford, T., Ward, M.J., Tatem, A.J., Sousa, J.D., Arinaminpathy, N., Pepin, J., Posada, D., Peeters, M., Pybus, O.G., Lemey, P., 2014. The early spread and epidemic ignition of HIV-1 in human populations. Science 346, 56–61.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17, 368–376.

Felsenstein, J., 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. Journal of Molecular Evolution 53, 447–455.

Foley, B.T., Leitner, T.K., Apetrei, C., Hahn, B., Mizrachi, I., Mullins, J., Rambaut, A., Wolinsky, S., Korber, B.T.M., 2015. HIV sequence compendium 2015. Office of Scientific; Technical Information (OSTI).

Foxman, B., 2001. Molecular epidemiology: Focus on infection. American Journal of Epidemiology 153, 1135–1141.

Gallo, R., Salahuddin, S., Popovic, M., Shearer, G., Kaplan, M., Haynes, B., Palker, T., Redfield, R., Oleske, J., Safai, B., et, 1984. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS.

Science 224, 500–503.

Gibbs, A.J., Mcintyre, G.A., 1970. The diagram, a method for comparing sequences. its use with amino acid and nucleotide sequences. Eur J Biochem 16, 1–11.

Gilbert, M.T.P., Rambaut, A., Wlasiuk, G., Spira, T.J., Pitchenik, A.E., Worobey, M., 2007. The emergence of HIV/AIDS in the americas and beyond. Proceedings of the National Academy of Sciences 104, 18566–18570.

Greene, W.C., 2007. A history of AIDS: Looking back to see ahead. European Journal of Immunology 37, S94–S102.

Hasegawa, M., Kishino, H., Yano, T.-a., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22, 160–174.

Heath, T.A., Holder, M.T., Huelsenbeck, J.P., 2011. A dirichlet process prior for estimating lineage-specific substitution rates. Molecular Biology and Evolution 29, 939–955.

Heibl, C., 2008 onwards. PHYLOCH: R language tree plotting tools and interface to diverse phylogenetic software packages.
http://www.christophheibl.de/Rpackages.html.

Hemelaar, J., Gouws, E., Ghys, P.D., Osmanov, S., 2011. Global trends in molecular epidemiology of HIV-1 during 20002007. AIDS 25, 679–689.

Hladik, F., McElrath, M.J., 2008. Setting the stage: Host invasion by HIV. Nat Re Immunol 8, 447–457.

Ho, S.Y.W., 2013. Molecular clocks, relaxed variant. In: Encyclopedia of Scientific Dating Methods. Springer Science Business Media, pp. 1–5.

Huelsenbeck, L., J. P., 2000. A compound poisson process for relaxing the molecular clock. In: Methods in Enzymology. Genetics, pp. 1879–1892.

Huson, D.H., Scornavacca, C., 2012. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Systematic Biology 61, 1061–1067.

Jansson, J., 2015. Phylogenetic tree construction from a distance matrix. In: Encyclopedia of Algorithms. Springer Science Business Media, pp. 1–4.

JUKES, T.H., CANTOR, C.R., 1969. Evolution of protein molecules. In: Mammalian Protein Metabolism. Elsevier BV, pp. 21–132.

Junqueira, D.M., Matos Almeida, S.E. de, 2016. HIV-1 subtype b: Traces of a pandemic. Virology 495, 173–184.

K, T., M, N., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial dNA in humans and chimpanzees. J Mol Evol 512 26.

Katoh, K., 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Research 30, 3059–3066.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16, 111–120.

Kingman, J., 1982. The coalescent. Stochastic Processes and their Applications 13,

235–248.

Kingman, J.F.C., 1982. On the genealogy of large populations. Journal of Applied Probability 19, 27.

Kishino, H., Hasegawa, M., 1990. [34] converting distance to time: Application to human evolution. In: Methods in Enzymology. Elsevier BV, pp. 550–570.

Kishino, H., Thorne, J.L., Bruno, W.J., 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Molecular Biology and Evolution 18, 352–361.

Lang, D.T., CRAN Team, 2015. XML: Tools for parsing and generating xML within r and s-plus.

Leisch, F., Peng, R.D., submitted. Knitr: A comprehensive tool for reproducible research in r. Methods in Ecology and Evolution.

Lepage, T., Bryant, D., Philippe, H., Lartillot, N., 2007. A general comparison of relaxed molecular clock models. Molecular Biology and Evolution 24, 2669–2680.

Lepage, T., Lawi, S., Tupper, P., Bryant, D., 2006. Continuous and tractable models for the variation of evolutionary rates. Mathematical Biosciences 199, 216–233.

M. Nei, Kumar, S., 2000. Molecular evolution and phylogenetics.

Magiorkinis, K.A., G., 2009. The global spread of hIV-1 subtype b epidemic: A phylogeographic meta-analysis.

Mau, B., Newton, M.A., 1997. Phylogenetic inference for binary data on dendograms using markov chain monte carlo. Journal of Computational and

Graphical Statistics 6, 122.

Merson, M.H.,.Malley, J.O., Serwadda, D., Apisuk, C., 2008. The history and challenge of HIV prevention. The Lancet 372, 475–488.

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES science gateway for inference of large phylogenetic trees. In: 2010 Gateway Computing Environments Workshop (GCE). Institute of Electrical & Electronics Engineers (IEEE).

Mount, D.W., 2008. Maximum parsimony method for phylogenetic prediction. Cold Spring Harbor Protocols 2008, pdb.top32–pdb.top32.

Nei M., S.N. &, 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4, 406–425.

Nordborg, M., n.d. Coalescent theory. In: Handbook of Statistical Genetics. Wiley Blackwell, pp. 843–877.

Ooms, J., James, D., DebRoy, S., Wickham, H., Horner, J., 2015. RMySQL: Database interface and 'mySQL' driver for r.

Osmanov, S., Pattou, C., Walker, N., Schwardländer, B., Esparza, J., 2002. Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. JAIDS Journal of Acquired Immune Deficiency Syndromes 29, 184–190.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

Paraskevis, D., Paraschiv, S., Sypsa, V., Nikolopoulos, G., Tsiara, C., Magiorkinis, G.,

Psichogiou, M., Flampouris, A., Mardarescu, M., Niculescu, I., Batan, I., Malliori, M., Otelea, D., Hatzakis, A., 2015. Enhanced HIV-1 surveillance using molecular epidemiology to study and monitor HIV-1 outbreaks among intravenous drug users (IDUs) in athens and bucharest. Infection, Genetics and Evolution 35, 109 121.

Paraskevis, D., Pybus, O., Magiorkinis, G., Hatzakis, A., Wensing, A.M., Vijver, D.A. van de, Albert, J., Angarano, G., Asjo, B., Balotta, C., Boeri, E., Camacho, R., Chaix, M.-L., Coughlan, S., Costagliola, D., Luca, A.D., Mendoza, C. de, Derdelinckx, I., Grossman, Z., Hamouda, O., Hoepelman, I.M., Horban, A., Korn, K., Kuecherer, C., Leitner, T., Loveday, C., Macrae, E., Maljkovic, I., Meyer, L., Nielsen, C., Coul, E.L.O. de, Ormaasen, V., Perrin, L., Puchhammer-Stockl, E., Ruiz, L., Salminen, M., Schmit, J.-C., Schuurman, R., Soriano, V., Stanczak, J., Stanojevic, M., Struck, D., Laethem, K.V., Violin, M., Yerly, S., Zazzi, M., Boucher, C.A., Vandamme, A.-M., Programme, S., 2009. Tracing the HIV-1 subtype b mobility in europe: A phylogeographic approach. Retrovirology 6, 49.

Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R., Moret, B.M., Stamatakis, A., 2010. How many bootstrap replicates are necessary? Journal of Computational Biology 17, 337–354.

Peeters, M., 2001. Recombinant hIV sequences: Their role in the global epidemic. Theoretical Biology; Biophysics Group.

Popovic, M., Sarngadharan, M., Read, E., Gallo, R., 1984. Detection, isolation, and

continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. Science 224, 497–500.

Posada, D., Crandall, K.A., 2001. Selecting models of nucleotide substitution: An application to human immunodeficiency virus 1 (HIV-1). Molecular Biology and Evolution 18, 897–906.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

R., H.R., 1991. Oxford surveys of evolutionary biology. In: Handbook of Statistical Genetics. Oxford University Press, Oxford, pp. 1–44.

Rambaut A, X.D.&.D.A., Suchard MA, 2014. Tracer v1.6. In:. Oxford University Press (OUP).

Rambaut, A., Bromham, L., 1998. Estimating divergence dates from molecular sequences. Molecular Biology and Evolution 15, 442–448.

Rannala, B., Yang, Z., 2007. Inferring speciation times under an episodic molecular clock. Systematic Biol. 56, 453–466.

Revell, L.J., 2012. Phytools: An r package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3, 217–223.

Richman, D.D., Margolis, D.M., Delaney, M., Greene, W.C., Hazuda, D., Pomerantz, R.J., 2009. The challenge of finding a cure for HIV infection. Science 323, 1304 1307.

Rohlf, F.J., 2005. J. felsenstein, inferring phylogenies, sinauer assoc. Journal of

Classification 22, 139–142.

Romero, P., 2004. Bioinformatics: Sequence and genome analysis. Briefings in Bioinformatics 5, 393–396.

RStudio Team, 2015. RStudio: Integrated development environment for r. RStudio, Inc., Boston, MA.

S, T., 1986. Some probabilistic and statistical problems in the analysis of dNA sequence. Lectures on Mathematics in the Life Sciences (American Mathematical Society) 57–86.

Schliep, K., 2011. Phangorn: Phylogenetic analysis in r. Bioinformatics 27, 592 593.

Scott A. Chamberlain, L.J.R. &, 2013. Rphylip: An r interface for pHYLIP.

Sharp, P.M., Hahn, B.H., 2011. Origins of HIV and the AIDS pandemic. Cold Spring Harbor Perspectives in Medicine 1, a006841–a006841.

South, A., 2011. Rworldmap: A new r package for mapping global data. The R Journal 3, 35–43.

Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post analysis of large phylogenies. Bioinformatics 30, 1312–1313.

Tajima, F., 1983. Evolutionary relationship of dNA sequences in finite populations. Genetics 105, 437–460.

Thorne, J.L., Kishino, H., 2002. Divergence time and evolutionary rate estimation with multilocus data. Systematic Biology 51, 689–702.

Thorne, J.L., Kishino, H., Painter, I.S., 1998. Estimating the rate of evolution of the rate of molecular evolution. Molecular Biology and Evolution 15, 1647–1657.

Urbanek, S., 2013. Png: Read and write pNG images.

Vercauteren, J., Wensing, A.M.J., Vijver, D.A.M.C. van de, Albert, J., Balotta, C., Hamouda, O., Kücherer, C., Struck, D., Schmit, J.-C., Åsjö, B., Bruckova, M., Camacho, R.J., Clotet, B., Coughlan, S., Grossman, Z., Horban, A., Korn, K., Kostrikis, L., Nielsen, C., Paraskevis, D., Poljak, M., Puchhammer-Stöckl, E., Riva, C., Ruiz, L., Salminen, M., Schuurman, R., Sonnerborg, A., Stanekova, D., Stanojevic, M., Vandamme, A.-M., Boucher, C.A.B., 2009. Transmission of drug resistant HIV-1 is stabilizing in europe. The Journal of Infectious Diseases 200, 1503–1508.

Weiss, R., 1993. How does hIV cause aIDS? 260, 1273–1279.

Wensing, A.M.J., Vijver, D.A. van de, Angarano, G., Åsjö, B., Balotta, C., Boeri, E., Camacho, R., Chaix, M.-L., Costagliola, D., Luca, A.D., Derdelinckx, I., Grossman, Z., Hamouda, O., Hatzakis, A., Hemmer, R., Hoepelman, A., Horban, A., Korn, K., Kücherer, C., Leitner, T., Loveday, C., MacRae, E., Maljkovic, I., Mendoza, C. de, Meyer, L., Nielsen, C., Coul, E.L.O. de, Ormaasen, V., Paraskevis, D., Perrin, L., Puchhammer-Stöckl, E., Ruiz, L., Salminen, M., Schmit, J.-C., Schneider, F., Schuurman, R., Soriano, V., Stanczak, G., Stanojevic, M., Vandamme, A.-M., Laethem, K.V., Violin, M., Wilbe, K., Yerly, S., Zazzi, M., Boucher, C.A., 2005. Prevalence of drug-resistant HIV-1 variants in untreated individuals in europe:

Implications for clinical management. The Journal of Infectious Diseases 192, 958 966.

Wensing, J., A.M., 2008. Transmission of drug-resistant HIV-1 in europe remains limited to single classes. AIDS 22, 625–635.

Wickham, H., 2009. Ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

World Health Organization, 2016. Epidemiology- definition.

Yang, Z., Rannala, B., 1997. Bayesian phylogenetic inference using DNA sequences: A markov chain monte carlo method. Molecular Biology and Evolution 14, 717–724.

Yang, Z., Yoder, A.D., 2003. Comparison of likelihood and bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Systematic Biology 52, 705–716.

Yu, G., Smith, D., Zhu, H., Guan, Y., Lam, T.T.-Y., submitted. Ggtree: An r package for visualization and annotation of phylogenetic tree with different types of meta data. Methods in Ecology and Evolution.

Zuckerkandl E, P.L., 1962. Molecular disease, evolution, and genetic heterogeneity. Horizons in Biochemistry. Academic Press 189–225.