SIMPLIFICACIÓN DE SECUENCIACIÓN MASIVA EN UNA ÚNICA SECUENCIA PARA SU UTILIZACIÓN EN ESTUDIOS FILOGENÉTICOS

JOSE ÁNGEL FERNÁNDEZ-CABALLERO RICO
MASTER BIOINFORMÁTICA Y BIOESTADÍSTICA

Consultor: LAURA PEDRO PUJIBET

30/05/2016

FICHA DEL TRABAJO FINAL

Título del trabajo:	Simplificación de secuenciación masiva en una única secuencia para su utilización en estudios filogenéticos
Nombre del autor:	Jose Ángel Fernández-Caballero Rico
Nombre del consultor:	Laura Pedro Pujibet
Fecha de entrega (mm/aaaa):	05/2016
Área del Trabajo Final:	Bioinformática
Titulación:	*Máster Bioinformática y Bioestadística*

Resumen del Trabajo (máximo 250 palabras):

Objetivo

En este trabajo mostramos como generar una secuencia consenso a partir de los datos de secuenciación masiva obtenidos en estudios de resistencias a antirretrovirales, que sea representativa de la secuencia Sanger, y que sirva para estudios de epidemiología molecular.

Métodos

En 62 pacientes se obtuvo la secuencia de Transcriptasa Reversa-Proteasa, mediante Sanger (Trugene-Siemens), y NGS (454GSJunior-Roche). Las secuencias consenso NGS se generaron con Mesquite, seleccionando umbrales 10%, 15% y 20%. Para el estudio filogenético se empleó MEGA.

Resultados

Utilizando el umbral 10%, 17/62 pacientes presentaron secuencias pareadas NGS-Sanger, con una mediana de *bootstrap* de 88% (IQR 83,5-95,5). La asociación aumenta a 36/62 pacientes y el *bootstrap* a 94% (IQR 85,5-98), y alcanza el máximo al 20% 61/62 pacientes, *bootstrap* 99% (IQR 98-100).

Conclusiones

Mostramos un método seguro y sencillo para generar secuencias consenso a partir de secuencias NGS, de fácil uso y aplicación en los servicios de microbiología clínica.

Abstract (in English, 250 words or less):

Aim
Here we show how to generate a consensus sequence from the information of massive parallel sequences (NGS) obtained from routine HIV antiretroviral resistance studies, suitable for molecular epidemiology studies.

Methods
Paired Sanger (Trugene-Siemens) and NGS (454 GSJunior-Roche) HIV RT & Protease sequences from 62 patients were studied. NGS consensus sequences were generated using Mesquite, using 10 %, 15 % and 20 % thresholds. For phylogenetic studies we used MEGA.

Results
At a 10%, NGS-Sanger sequences from 17/62 patients were phylogenetically related, with a median bootstrap-value of 88 % (IQR 83,5-95,5). Association increased to 36/62 sequences, median bootstrap 94 % (IQR 85,5-98)], using a 15% threshold. Maximum association was at the 20% threshold, with 61/62 sequences associated, and a median bootstrap value of 99% (IQR 98-100).

Conclusion
We show an easy and safe method to generate consensus sequences from HIVNGS data, that will prove usefull for molecular HIV epidemiological studies.

Palabras clave (entre 4 y 8):

VIH, filogenia, NGS, umbrales, Sanger.

ÍNDICE

1. RESUMEN TRABAJO FINAL DE MASTER

1.1 Contexto y justificación del Trabajo

Desde su introducción en la década de los 90 la secuenciación Sanger ha sido la técnica principal utilizada para el estudio clínico de pacientes VIH. A día de hoy se están introduciendo nuevas técnicas en la rutina asistencial para la determinación de resistencias, como son las técnicas de secuenciación masiva (UDS o NGS) [1,2], esta implementación revoluciono entre otros los estudios filogenéticos [3-5]. El potencial de NGS para detectar variantes virales VIH de baja frecuencia se ha determinado en varios estudios [6], en cambio la secuenciación tradicional Sanger presenta un umbral de detección de variantes alrededor del 20% [7,8]. (Figura 1)

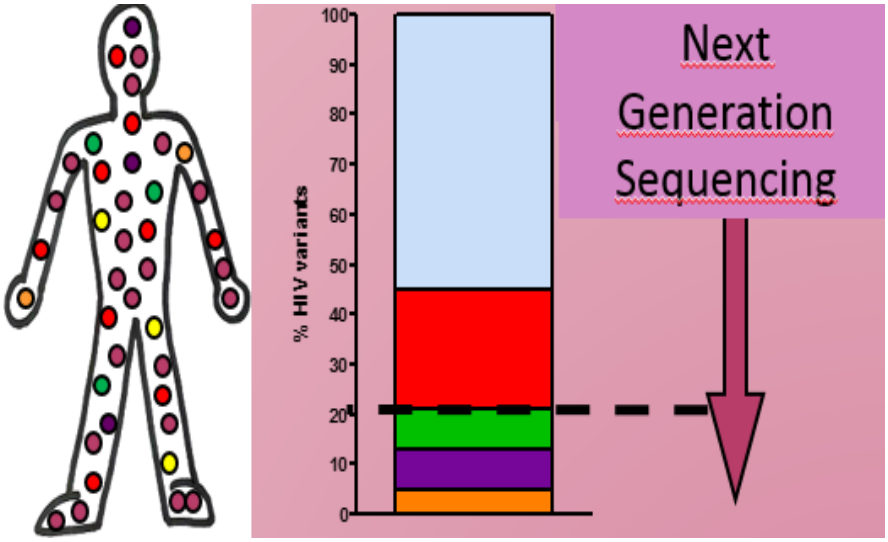


Figura 1. Detección de variables minoritarias mediante secuenciación masiva.

Existen varias técnicas UDS para secuenciar VIH [9,10], siendo capaces de generar de tres a cuatro órdenes de magnitud más de información que la secuenciación Sanger [11]. Esto hace que para los estudios filogenéticos mediante secuenciación UDS esté presente la barrera de la bioinformática, requiriéndose tanto de una formación especial para el procesado de secuencias, como de unos ordenadores de gran potencia para procesar el gran volumen de datos [12]. Esta problemática ha hecho que se estudien

nuevos algoritmos filogenéticos que consigan disminuir el tiempo de procesamiento y aumentar la cantidad o longitud de secuencias a utilizar [13,14]. Una alternativa es generar una única secuencia consenso UDS, pero algunos estudios no son claros u omiten el método utilizado para generar la secuencia consenso UDS [15], otros recurren a programación "python" para generar una única secuencia consenso UDS [5], siendo necesario altos niveles de bioinformática.

Hoy en día tienden a coexistir la secuenciación Sanger con UDS, por lo que ciertos estudios presentaran datos conjuntos de estas dos metodologías, siendo necesaria su unificación. Además en ciertos estudios filogenéticos de gran escala, donde se maneja una elevada cantidad de secuencias sería necesario la simplificación mediante una única secuencia consenso. Actualmente no existen datos que avalen si generar una secuencia consenso UDS puede o no representar adecuadamente la secuencia Sanger.

Por lo tanto este tema es muy relevante ya que cada vez se está utilizando mas la secuenciación masiva, por lo que se hace necesaria la simplificación de esta para su correcto manejo. El resultado esperado será la simplificación de las miles de secuencias obtenidas mediante secuenciación masiva en una única similar a su homóloga Sanger.

1.2 Objetivos del Trabajo

1- Determinar cuál es el mejor umbral de corte para la obtención de una secuencia consenso NGS, que sea representativa de la secuencia tipo Sanger, y que pueda ser utilizada en estudios de epidemiología molecular.

2- Aplicación de las secuencias obtenidas en estudios de epidemiología molecular.

1.3 Enfoque y método seguido

1) Se podría seleccionar todas las secuencias obtenidas por secuenciación masiva y ver cuál de todas se aproxima más a la secuencia Sanger.

2) Se podría hacer un alineamiento de la secuencia Sanger y la secuencia simplificada masiva para comprobar cuantas bases difieren y saber si se aproximan.

3) Comparar la secuencia Sanger con la Secuencia simplificada masiva mediante estudios filogenéticos.

El inconveniente de la secuenciación masiva es que obtenemos miles de secuencias del gen *pol* en cuatro fragmentos de 350 pares de bases (pb) cada una, mientras que la secuencia Sanger se procesa como un único fragmento de 950 pb aproximadamente. Por lo tanto el enfoque 1 no sería correcto porque estaríamos sesgando los resultados.

El segundo enfoque no sería correcto, ya que un alineamiento no tiene un valor establecido para dichos estudios, este punto podría ser una ayuda para comprobar en cuantas bases nucleotídicas difieren dichas secuencias sin más.

El tercer enfoque con estudio filogenético sería correcto ya que mediante estos estudios podemos comparar secuencias y ver como de próximas son entre sí mediante un valor medible (*bootstrap*). Ademes los métodos filogenéticos están bien comprobados y tienen una base sólida científica.

1.4 Planificación del Trabajo

1) Realización y obtención de secuencias (Sanger y NGS). (2 semanas)

2) Analizar y asegurar la calidad de las secuencias obtenidas. (1 semana)

3) Tratamiento de las secuencias mediante programas bioinformáticos para eliminar posibles errores de secuenciación. (2 semanas)

4) Utilización de programa para simplificación de secuenciación masiva con distintos umbrales de corte. (1 semana)

5) Estudio filogenético de las secuencias obtenidas, comprobando cual es el punto de corte NGS que más se aproxima a las secuencias Sanger. (2 semanas)

6) Estudio de cluster de transmisión VIH en Andalucía Oriental mediante filogenia y utilizando la simplificación del trabajo expuesto en TFM. (2 semanas)

FECHA	TAREA
21/03/2016 - 03/04/2016	Realización y obtención de secuencias
04/04/2016 - 10/04/2016	Analizar y asegurar la calidad de las secuencias obtenidas
11/04/2016 – 24/04/2016	Eliminación de posibles errores en la secuenciación.
25/04/2016 – 01/05/2016	Simplificación de secuenciación masiva con distintos umbrales de corte
02/05/2016 – 15/05/2016	Estudio filogenético de secuencias simplificadas NGS vs Sanger.
16/05/2016 – 29/05/2016	Estudio de cluster de transmisión VIH en Andalucía Oriental.

1.5 Breve sumario de productos obtenidos

Se ha obtenido la simplificación de la secuenciación masiva, mediante distintos arboles filogenéticos. Además se ha efectuado estudio de pacientes VIH en Andalucía Oriental y en pacientes procedentes de la cárcel. Para finalizar se procederá a la elaboración de un artículo científico.

1.6 Breve descripción de los otros capítulos de la memoria

1. Introducción: Introducción al Virus VIH. En el trabajo global servirá para contextualizar el trabajo, saber la estructura del VIH y el genoma que vamos a secuenciar, epidemia global, conocer la importancia de los distintos subtipos que observaremos en los arboles filogenéticos y una pequeña introducción a los métodos filogenéticos.

2. Objetivos: Conocer los objetivos del trabajo.

3. Materiales y métodos: Observar los distintos materiales y métodos utilizados en el trabajo de laboratorio. Importante para su posterior reproducibilidad en otros laboratorios.

4. Resultados: Obtención de los distintos resultados a través de los materiales y métodos. Aquí observaremos nuestra producción científica y veremos si hemos logrado los objetivos propuestos.

5. Discusión: Breve discusión sobre los resultados obtenidos.

2.

INTRODUCCIÓN

2.1. EL VIRUS DE LA INMUNODEFICIENCIA HUMANA

2.1.1. Historia VIH

El Virus de la Inmunodeficiencia Humana (VIH) es el agente patógeno causal del Síndrome de la Inmunodeficiencia Adquirida (SIDA).

La enfermedad del virus de la inmunodeficiencia humana fue descrita por primera vez en 1981 por dos grupos, uno en San Francisco y otro en Nueva York, asociada a una deficiencia inmunológica grave debido a neumonía por *Pneumocystis jirovecci* y Sarcoma de Kaposi agresivo [16]. El virus VIH en sí no fue identificado hasta dos años después, cuando los franceses Barré-Sinoussi y Montagnier consiguieron aislarlo [17]; durante ese período, otras causas fueron consideradas, incluyendo factores como el estilo de vida, el abuso crónico de drogas y otros agentes infecciosos [18]. Pronto se evidenció que presentaba el mismo perfil epidemiológico de hepatitis B en cuatro grupos bien diferenciados: homosexuales, hemofílicos, hemoperfundidos y heroinómanos. La similitud hizo pensar que la causa podría ser un agente infeccioso, probablemente viral, con transmisión sexual y sanguínea [19]. En 1986, el Comité Internacional de Taxonomía de Virus (ICTV) aceptó el nombre definitivo de VIH para este nuevo agente [20].

Desde estos primeros hallazgos hasta la actualidad han transcurrido tres décadas. Hoy sabemos que el VIH-1 infecta células del sistema inmune (principalmente linfocitos CD4+ y macrófagos), causando su muerte o alterando su funcionalidad, con el consiguiente deterioro progresivo de la capacidad del sistema inmune para combatir las infecciones, y que en las etapas más avanzadas de la infección sobreviene el SIDA, caracterizado biológicamente por un profundo deterioro de la inmunidad celular y una severa depleción de los linfocitos T CD4+ y clínicamente por la presencia de algunas infecciones oportunistas o tipos de cáncer [21-23]. También se sabe que el virus se transmite por contacto sexual (penil-vaginal y penil-anal), la transfusión de sangre o productos sanguíneos contaminados, el uso compartido de agujas, jeringuillas u otros instrumentos punzantes o cortantes y de la madre al hijo durante el embarazo, el parto y

la lactancia [22-25]. La eficiencia de la transmisión sanguínea depende de múltiples factores, como el número de partículas virales, volumen de sangre y el estado inmune del receptor. El contacto sexual (homosexual y heterosexual) es el principal modo de transmisión, siendo la transmisión heterosexual la predominante a escala global, aumentando el riesgo de transmisión con las infecciones genitales concomitantes causadas por otros patógenos (Herpes, Clamidia y otros) [24]. La probabilidad de transmisión de la madre al hijo en ausencia de tratamiento antirretroviral durante el embarazo es del 15-30% [22].

La epidemia de VIH/SIDA constituye en la actualidad uno de los más graves problemas de salud pública, con grandes repercusiones demográficas, sociales y económicas a nivel mundial, pero particularmente en los países en vías de desarrollo. El Programa Conjunto de *World Health Organization VIH/AIDS* estimó que a finales de 2014 había 36,9 millones de personas infectadas por el VIH-1 en todo el mundo, que corresponde al 92.95% de los adultos entre 15-49 años de edad, e incluyen 2,6 millones de niños menores de 15 años, estimándose en 2 millones las nuevas infecciones por VIH y en 1,2 millones las defunciones por SIDA en ese año [26] (figura 2).

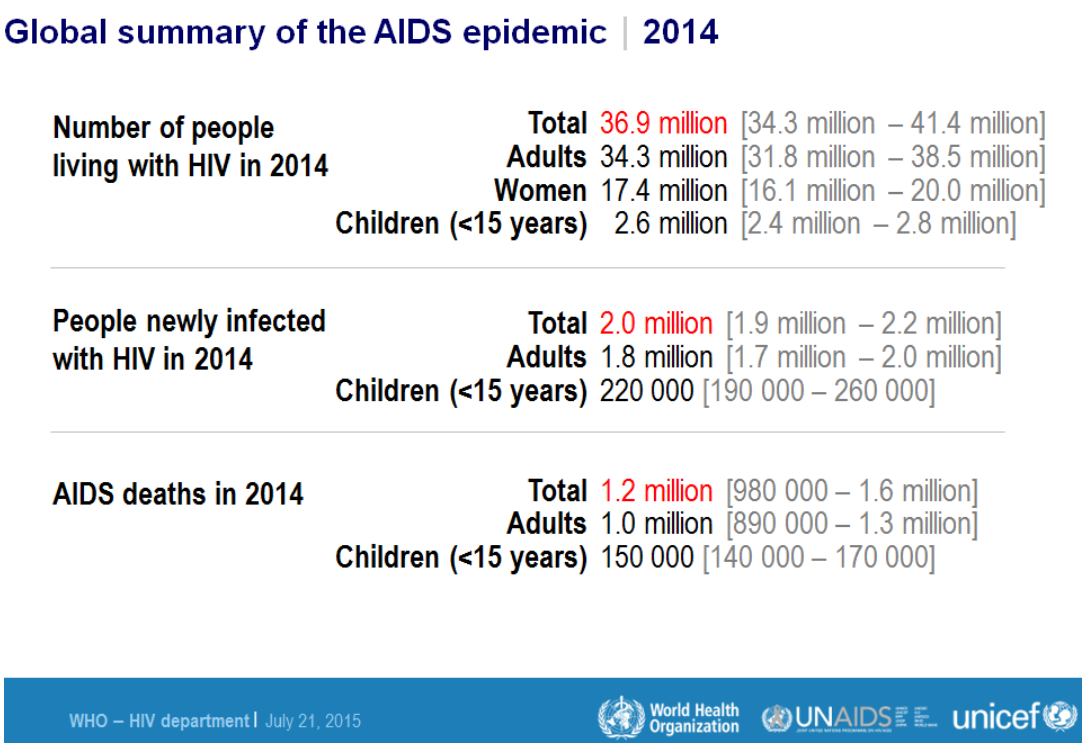


Figura 2. Estadística VIH 2014. Desde que en 1996 se introdujo el tratamiento antirretroviral de gran actividad (TARGA), el arsenal terapéutico frente al VIH se reducía a la combinación de tres familias de antirretrovirales: Los Inhibidores análogos de la Transcriptasa Reversa (ITIAN), inhibidores no análogos de la Transcriptasa Reversa (ITINAN) y los inhibidores de la Proteasa (IP). [26]

2.1.2. Origen y evolución del VIH

Con el paso del tiempo se ha descrito que los virus causantes del SIDA son lentivirus. Sus parientes más cercanos, se han encontrado infectando a otros primates, y son conocidos como Virus de la Inmunodeficiencia del Simio (SIV), aunque hasta donde se conoce estos virus no causan enfermedad en sus hospedadores naturales. Se sabe que más de 20 especies de primates (procedentes del África sub-sahariana), albergan estos virus y algunas de ellas exhiben altas tasas de infección, con virus diversos pero especie específicos. Además, numerosas especies de monos estrechamente relacionadas son infectadas por virus SIV muy cercanos filogenéticamente, lo que sugiere que llevan siendo infectados desde hace mucho tiempo [27,28]. Por el contrario, la infección humana por el VIH-1 y 2 aparece como un fenómeno relativamente reciente. Ambos

casos (VIH-1 y 2) presentan una mayor diversidad de cepas virales, así como tasas más altas de infección en el África sub-sahariana [29]. Estos datos proporcionan la evidencia de que el VIH surgió en África a través de una transmisión entre especies desde los primates. Análisis filogenéticos indican que VIH-1 y 2 proceden de dos linajes de SIV muy diferentes y por lo tanto tienen orígenes diferentes (figura 3) [30]. Si observamos la historia del VIH vemos que el VIH-2 se encuentra principalmente en la zona de África occidental, pero, han sido descritos casos en Europa y Estados Unidos. Filogenéticamente se puede observar que VIH-2 es muy cercano a $SIV_{sm,}$ perteneciente a una especie de mono de África occidental. Si ahora estudiamos la historia evolutiva del VIH-1 observamos que es más complicada, ya que su antecedente infeccioso proviene del chimpancé *Pan troglodytes troglodytes,* siendo su hábitat poco accesible en el sur de Camerún [31].

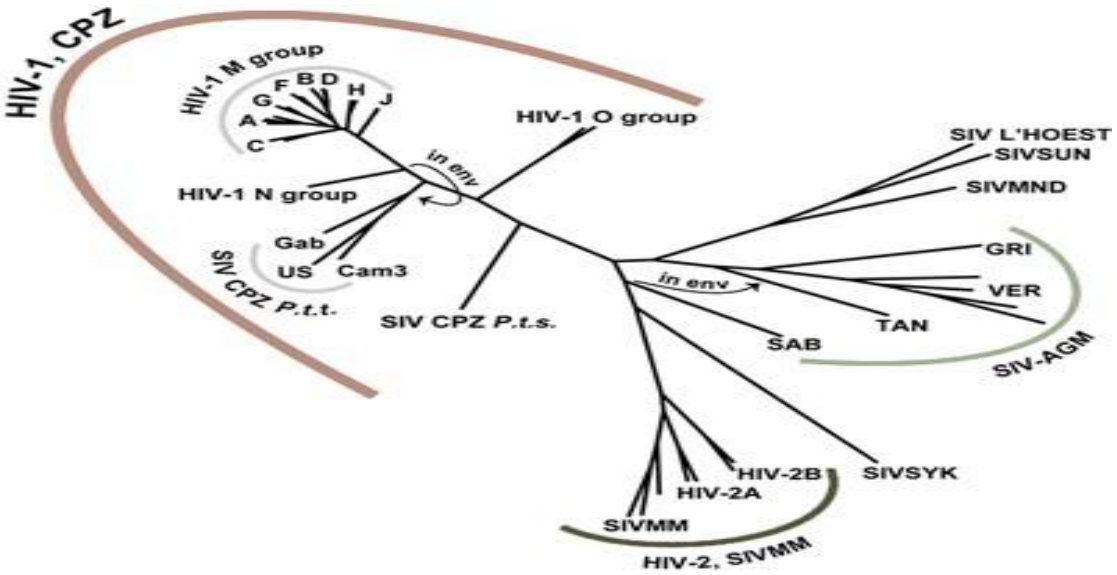


Figura 3. Árbol filogenético historia evolutiva VIH. El virus de la inmunodeficiencia de los simios viene determinado por SIV. *cpz, chimpanzee; lhoest, l´Hoest monkey; sun, sun-tailed monkey; mnd, mandrill; agm, African green monkey; ver, vervet; gri, grivet; Tan, tantalus; mac, macaque; sm, sooty mangabey; syk, Syke´s monkey*. [32]

Una de las explicaciones más simples y plausibles para la transmisión inter-específica del SIV para dar lugar al VIH es debido a la exposición directa de los humanos a la sangre de los chimpancés (así como los *sooty mangabeys* y algunos primates), a sus secreciones mucosas como resultado de la caza, u otras actividades como el consumo de

la carne sin cocinar, que anteriormente ha podido ser vendida en mercados denominados como “bushmeat” [33]. Sin embargo, también se han propuestos modos iatrogénicos (inducidos por humanos) para la transmisión inter-especie. Entre ellas, la idea de que los virus del VIH fueran transferidos desde sus hospedadores naturales hasta el humano vía preparaciones contaminadas de la vacuna oral contra la polio. En particular, la causa de la pandemia causada por la variedad M del VIH-1 se ha sugerido que fue la vacuna de la polio desarrollada por un equipo de investigadores liderado por el Dr. Hilary Koprowski y que fue administrada a cerca de un millón de personas en el Congo Belga, Rwanda y Burundi a finales de los años 50. Sin embargo, hay varias líneas de evidencia que demuestran que durante la producción de la vacuna el SIV-VIH hubieran perdido su viabilidad [30].

Considerando las relaciones filogenéticos entre el VIH-1 y las cepas de SIV_{cpz} se observa claramente que cada grupo del VIH-1 representa un salto entre especies diferente. Si estimásemos la edad del ancestro común a todos los grupos, podríamos emplazar un límite en la escala de tiempo, donde la transmisión entre especies ocurrió. Los intentos para datar estos eventos se han centrado en el grupo M del VIH-1.Después de todos los cálculos realizados se ha sugerido que el ancestro común del grupo M existía alrededor de 1960 [34]. Este dato parece coherente con los conocimientos sobre el comienzo de la epidemia, y con el suero positivo más temprano obtenido de un individuo de Léopoldville (hoy en día Kinshasa) en 1959 [35]. Sin embargo análisis más completos (teniendo en cuenta modelos más recientes) analizando los genes de la cubierta del VIH-1 aislados de más de 150 individuos, y utilizando el modelo de vecindad-máxima de la evolución de las secuencias (permitiendo a sitios individuales evolucionar a diferentes tasas), ha estimado que el ancestro común del grupo M procede de 1931, con un intervalo de confianza comprendido de 1915 a 1941 [36].

Con los datos recabados desde hace años podemos suponer con certeza que la infección VIH se quedo aislada en poblaciones remotas, hasta que con el paso de los años y el crecimiento de las poblaciones alcanzo la ciudad de Kinshasa alrededor de 1930-40. En la ciudad de Kinshasa podemos observar una gran variedad de subtipos VIH, debido a que como ya se ha comentado, en esta ciudad se produjo la expansión de la infección VIH.

Posteriormente se produjo la expansión en EE.UU a través de la inmigración sufrida por la población de Haití [31].

2.1.3. Hipótesis sobre la transmisión del VIH

Existen tres hipótesis sobre el evento y tiempo de transmisión del VIH-1 de *Pan troglodytes* al humano [37]:

1. Transmisión temprana. Propone que el virus fue transmitido a los humanos a finales del siglo XIX o principios del siglo XX a través de la caza e ingesta de carne de chimpancés. El virus permaneció aislado en un grupo pequeño de humanos hasta cerca de 1930, a partir de ese año comenzó su propagación y diversificación en otras poblaciones humanas en África y eventualmente al resto del mundo.

2. Transmisión por causas epidémicas. Según esta hipótesis, el virus fue transmitido al humano alrededor de 1930 e inmediatamente comenzó su diversificación y expansión geográfica.

3. Transmisión paralela tardía. Hipótesis que sugiere que diversas cepas de VIS fueron transmitidas al humano entre los años 1940 y 1950 a través de la vacuna de polio virus contaminada con VIS debido a su cultivo en células de riñón del chimpancé.

Los árboles filogenéticos del VIH-1 indican que la propagación del virus fue inicialmente muy lenta. La explosión de la infección en los años 1950 y 1960, coincide con el término del gobierno colonial en África, varias guerras civiles, la introducción de los programas de vacunación (con la reutilización de las agujas), el crecimiento de las ciudades africanas, la revolución sexual y el incremento de viajeros desde y hacia África [37].

Existe una coincidencia con el periodo de los años 1970s en que los síntomas del SIDA comenzaron a ser predominantes en individuos infectados en Estados Unidos y Europa y el periodo aproximado de 10 años necesarios desde la infección con VIH hasta la progresión al SIDA [37,38].

2.2. CARACTERÍSTICAS GENERALES DEL VIH

2.2.1. Estructura VIH

La partícula viral tiene forma esférica icosaédrica y mide 80-120 nm de diámetro (figura 4), presenta una estructura en tres capas:

Está rodeada por una membrana lipídica que deriva de la célula hospedadora, por lo que contiene proteínas celulares, como HLA clase I y II y proteínas de adhesión como CAM-1. Además, se integran 72 complejos de glicoproteína (gp) viral formados por trímeros de gp120 y gp41, esenciales para la interacción con la célula diana. La matriz está formada por la proteína p17 que está insertada en la superficie interna de la membrana lipídica.

La cápside está formada por la proteína p24 (p26 en VIH-2).La proteína p24 es el antígeno más fácil de detectar y son los anticuerpos contra él, los que se utilizan para el diagnostico de infección por VIH por medio de ELISA.

En la estructura interna o nucleoide se encuentran el genoma viral (ARNmc con polaridad positiva) y estabilizado por la nucleoproteínas p7 y todas las enzimas necesarias para su replicación (la Transcriptasa Reversa, la Integrasa (IN), la Proteasa (PR) y las proteínas reguladoras y accesorias) [38].

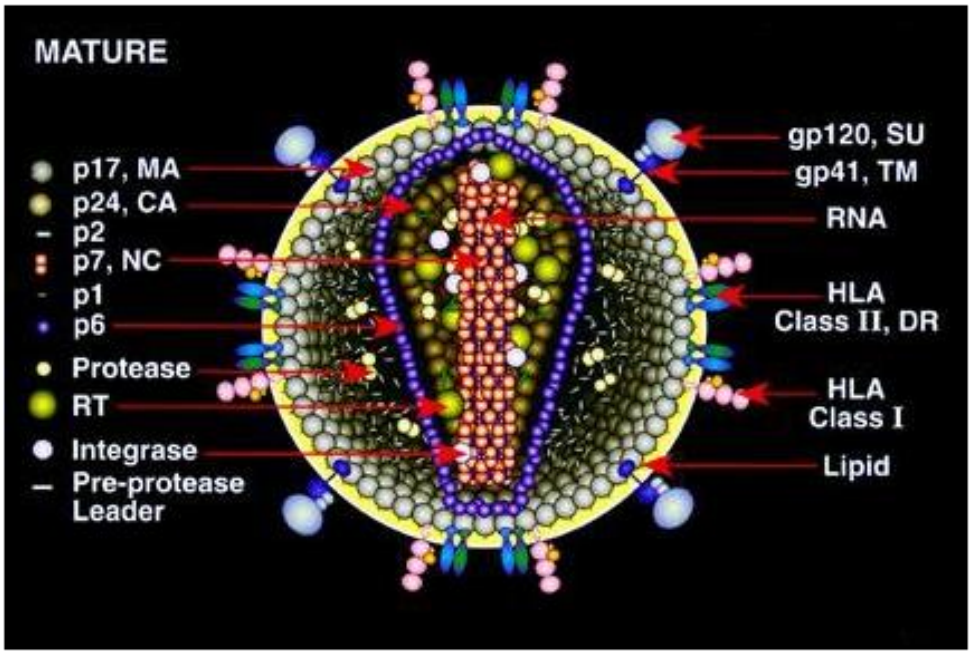


Figura 4. Estructura partícula vírica VIH. [39]

El genoma del virus es un ARN de cadena única formado por dos hebras idénticas de 9,8 Kb de polaridad positiva. Emplea la enzima Transcriptasa Reversa presente en el virón para replicarse, dando lugar al ADN proviral que se integra en el genoma de la célula huésped gracias a la enzima Integrasa. El genoma viral VIH está compuesto de tres regiones genéticas: 1) *gag*, que codifica proteínas estructurales del centro viral o *core*, 2) *env* que codifica glicoproteínas de la envoltura y 3) *pol* que contiene secuencias que codifican la enzima Transcriptasa Reversa, endonucleasa y Proteasas virales necesarias para la replicación del VIH. Además, estas regiones comentadas anteriormente contienen marcos de lectura adyacentes que corresponden a genes codificadores de proteínas no estructurales, importantes para la funcionalidad biológica del virus (crecimiento, ensamblaje y replicación). Hay que aclarar que el VIH-2 también contiene estos genes, excepto el *vpu*, además contiene otro gen, el *vpx* que está ausente en el VIH-1 [40].

La región *pol* tiene un tamaño de 3kb, flanqueado por la región *gag* y *vif* (figura 5). Esta región presenta tres enzimas que son necesarias para el ciclo replicativo. La primera que encontramos es la proteína Proteasa que se encarga de la rotura proteolítica para dar proteínas maduras. A continuación la proteína Retrotransciptasa Reversa que es la enzima que transcribe el paso de ARN a ADN. Para finalizar se encuentra la Integrasa que es la proteína encargada de integrar el ADN vírico en el ADN de la célula huésped

mediante transferencia de cadenas. Los fármacos antirretrovirales se han centrado en detener la actividad de estas tres enzimas encargadas de la replicación VIH. De este modo tenemos la familia antirretroviral de inhibidores de la proteasa, Integrasa y los inhibidores de la Retrotranscriptasa Reversa (Análogos y No análogos).

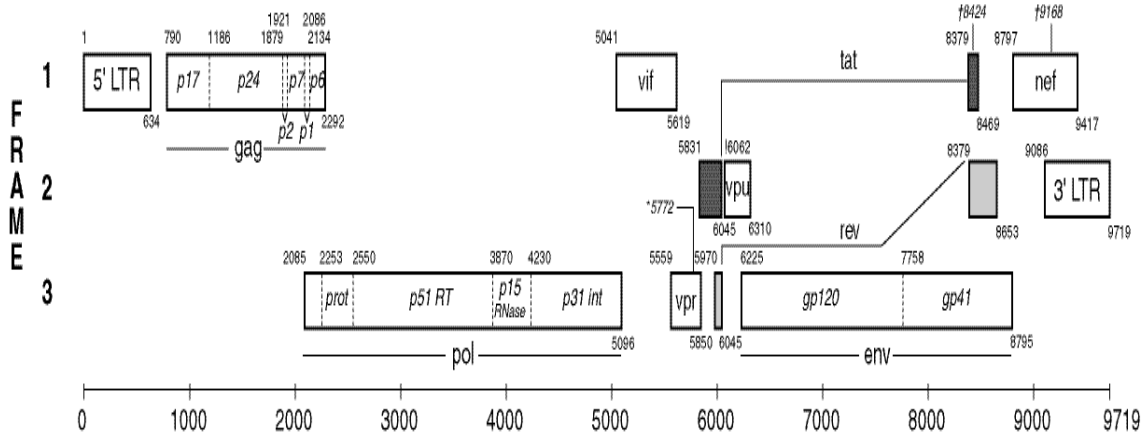


Figura 5. Representación del genoma VIH.

Además, en su forma de provirus, el genoma viral contiene unas secuencias repetidas (LTR) que permitirían su integración en el genoma de la célula huésped. Estas regiones además contendrían los elementos reguladores de la iniciación de la transcripción viral [41,42].

2.3. DIVERSIDAD GENÉTICA DEL VIH

La diversidad genética del VIH se debe a su alta tasa de variabilidad e inestabilidad genética, que veremos a continuación.

El VIH presenta una tasa de error y una generación de nuevas partículas infectivas muy altas, por lo que hace que el virus evolucione muy rápido, escapándose del sistema defensivo del organismo [43].

2.3.1. El concepto de cuasiespecie

La población VIH sufre variaciones genéticas en el organismo a medida que se va replicando, lo que hace que existan virus distintos pero relacionados genéticamente entre si. Estos virus con pequeñas variaciones son denominadas cuasiespecies. Estas poblaciones siguen un sistema evolutivo basado en la teoría de la evolución por selección natural propuesta por Charles Darwin, por lo que las variantes con mayor *fitness* tendrán mayor replicación. Por lo tanto, en la población vírica tendremos una secuencia maestra que será el genoma vírico predominante, comprendiendo las mutaciones más predominantes.

Las poblaciones víricas están expuestas a presiones selectivas en el organismo, como medicación, ejerciendo como cuellos de botella que producen una selección de las cuasiespecies víricas que tengan mayor capacidad para replicarse [44] (figura 6).

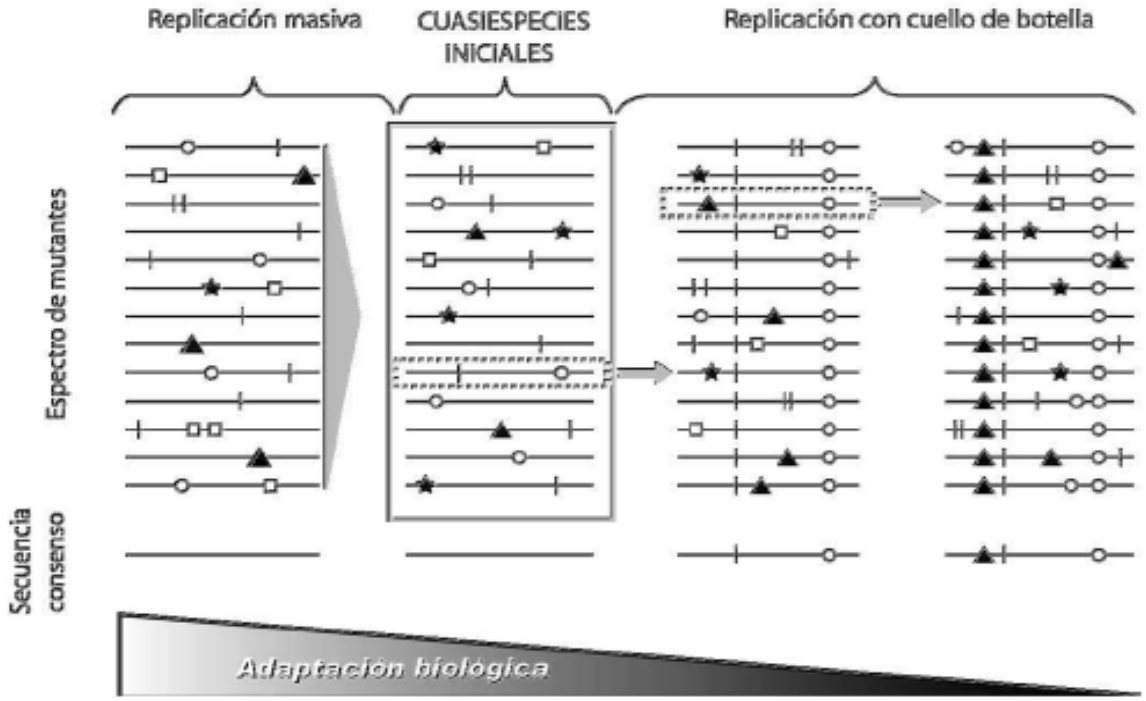
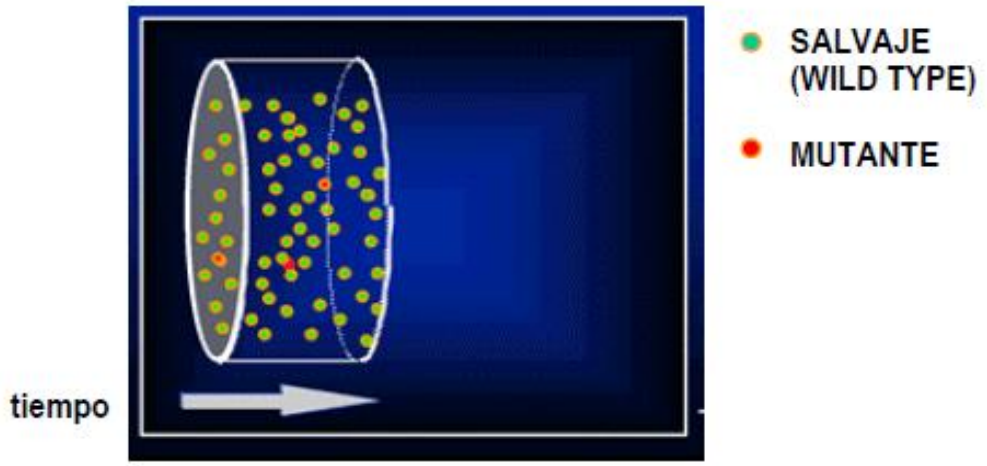


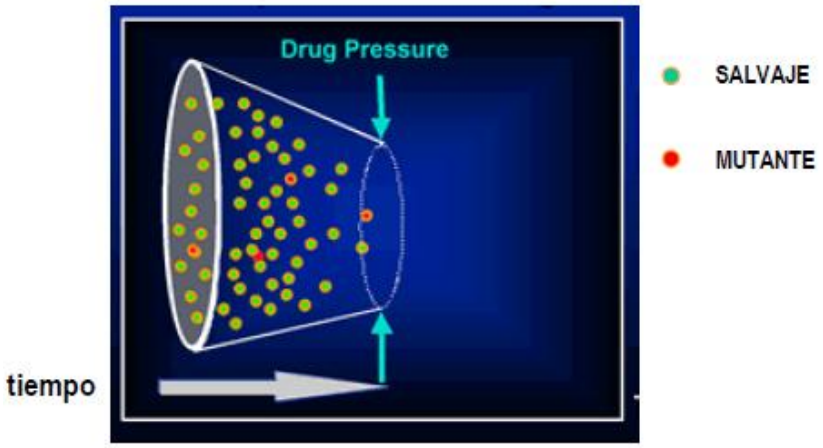
Figura 6. Representación esquemática de la dinámica en cuasiespecie. [45]

Si ahora nos fijamos en el efecto del tratamiento antirretroviral veremos que este ejerce una presión selectiva, seleccionando aquellas poblaciones víricas que presentan mutaciones que confieren resistencia a dicho medicamento (figura 7). Debido a la alta producción de partículas víricas (10^{10} partículas víricas/día), puede producirse por azar mutaciones que confieren resistencia, pero que permanecerán en niveles bajos en la población global hasta que se efectué la presión selectiva del tratamiento antirretroviral. Este hecho hace que la presión selectiva que ejerce un fármaco haga que las poblaciones se deslicen desde las que son sensibles hacia las que tienen mutaciones de resistencia, pero las poblaciones sensibles permanecerán en reservorios, con lo cual al cesar la presión farmacológica volverá haber un deslizamiento de poblaciones [44].

En las siguientes imágenes vemos la dinámica de las cuasiespecies sin ningún tipo de presión selectiva [44].



Si continua la presión reducirá el número de virus pero algunas partículas persisten.



En el caso de que continuase la presión las poblaciones resistentes serian predominantes debido a su mayor *fitness*.

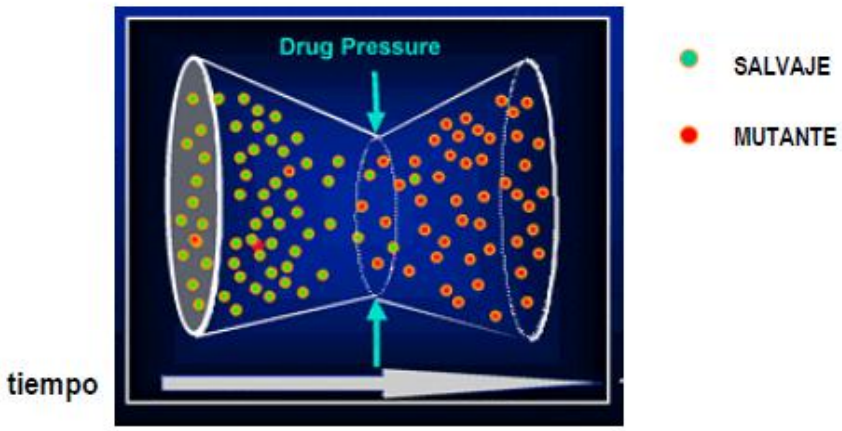


Figura 7**.** Selección de cuasiespecies. [44]

2.4. IMPLICACIONES BIOLÓGICAS Y CLÍNICAS DE LA DIVERSIDAD GENÉTICA DE VIH-1

La circulación de diferentes subtipos virales tiene una repercusión importante evolutiva, ya que favorece la dinámica del proceso de diversificación del VIH. Esta diversidad tiene importantes implicaciones como veremos a continuación.

2.4.1. Desarrollo de vacunas

La variabilidad global del VIH-1 plantea un reto para el desarrollo de una vacuna efectiva. Debido a los diversos subtipos es imposible crear una única vacuna que pueda proteger contra la mayoría de subtipos y formas recombinantes, el diseño tendría que ser mediante vacunas especificas para cada región.

2.4.2. Utilización de correceptores

Como se ha descrito anteriormente, es necesaria la presencia de receptores celulares secundarios para la entrada de VIH en el huésped. Estos correceptores, en particular los receptores de quimiocinas de tipo 5 (CCR5) y receptor tipo 4 (CXCR4), han sido objeto de una amplia investigación de intentar dilucidar el mecanismo de entrada viral y progresión de la enfermedad. Basado en el uso de correceptor, las cepas de VIH-1 se clasifican como R5 (utilizando correceptor CCR5) y X4 (correceptor CXCR4), aunque también hay cepas que pueden usar los dos tipos de receptores (mixtas). Los estudios sobre el subtipo B habían demostrado que en la etapa posterior de la enfermedad, casi el 50% de las cepas del subtipo B son X4 [46-50]. También se ha demostrado que los virus de subtipo D son en su mayoría X4 o con tropismo mixto en todo el curso de la infección en comparación con el subtipo A. Por el contrario, las cepas del subtipo C utilizan con frecuencia en su totalidad el tropismo CCR5, incluso en etapas posteriores a la infección aguda. [51-53].

2.4.3. Transmisión

Un número creciente de estudios científicos sugieren diferentes eficiencias de transmisión según los diferentes subtipos VIH-1. Un estudio en Tanzania mostro una alta velocidad de transmisión en la localización uterina para el subtipo C en comparación al subtipo A y D. Por otro lado, un estudio relazado en Tailandia entre UDI mostro un aumento de la probabilidad de transmisión de CRF01_AE en comparación con el subtipo B. [54-58]

2.5. INTRODUCCIÓN A LOS MÉTODOS DE ANÁLISIS FILOGENÉTICO

La filogenia es la ciencia que estudia cómo han evolucionado los organismos. Si hablamos de filogenia molecular, estudia la evolución a partir de la homogeneidad de las secuencias ADN o proteínas, de esta forma, cuanto más homogéneas sean las secuencias más parentesco común presentaran y estarán emparentadas filogenéticamente compartiendo un ancestro común. Estos estudios filogenéticos se basan en la creación de arboles filogenéticos que representan en ramas mediante distancia genética la relación evolutiva entre varios organismos [59]. Para establecer esa homología entre las secuencias es necesario que estas estén alineadas, para ello existen muchos programas bioinformaticos (CLUSTAW, MUSCLE…).

3. OBJETIVOS

Nuestro propósito con este estudio es:

1- Simplificación de secuencias *Next Generation Sequencing* (NGS). Determinando cuál es el mejor umbral de corte para la obtención de una secuencia consenso *Ultra Deep Sequencing* (UDS), representativa de la secuencia tipo Sanger.

2- Aplicación de la metodología propuesta en estudios de epidemiología molecular mediante base de datos de pacientes (2014-2016), observando si existen *cluster* definidos en la región de Andalucía Oriental.

Como objetivos secundarios se plantean:

1- Manejo de distintos software para analizar y procesar los resultados de secuenciación masiva y secuenciación Sanger.

2- Manejo y aprendizaje de software Mesquite, asi como software MEGA.

4. MATERIALES Y MÉTODOS

4.1. POBLACIÓN DE ESTUDIO

Se trata de un estudio piloto y retrospectivo, en el ámbito del Hospital Universitario San Cecilio de Granada.

Se han seleccionado un total de 62 pacientes *naive* infectados por VIH-1, desde el año 2014 a 2015. De cada paciente se realizó un estudio genotípico de la región *pol* viral mediante secuenciación Sanger y posteriormente con secuenciación masiva (NGS), mediante 454 GS Junior. Para el ensayo se parte de sangre total, de donde se extraerá ARN vírico. La conservación de la sangre total tras la extracción consistió en mantenerla en viales que contengan algún agente anticoagulante, ya que sería imposible la extracción en el caso de producirse la formación de coagulo.

4.2. MÉTODOS

- Extracción de ARN VIH

1) Secuenciación Sanger

- Obtención de ADN desde ARN
- Amplificación de ADN
- Secuenciación en electroforesis en gel de poliacrilamida de los amplificados
- Alineamiento de las secuencias e interpretación de los resultados

2) NGS

- Amplificación de ADN extraído
- Purificación de amplicones
- Cuantificación de amplicones
- Creación de la librería
- PCR en emulsión (emPCR)
- Secuenciación masiva
- Alineamiento de las secuencias e interpretación de resultados

4.2.1. Extracción de ARN VIH desde plasma sanguíneo

Normalmente al laboratorio llegan las muestras de plasma sanguíneo, pero en algunos casos nos encontraremos con muestra de sangre total. En este último caso tendremos que centrifugar la muestra durante 10 minutos a 2500 rpm, de esta forma conseguiremos separar la fracción de plasma en la sangre total.

Para la extracción se procedió con 1000 ml de plasma sanguíneo, contando con la maquina MagNA Pure Compact (Roche) (figura 8), mediante el programa de cartuchos de extracción MagNA Pure Compact Nucleic Acid Isolation Kit I- Large Volume, eluyendo en un volumen final de 50ul.

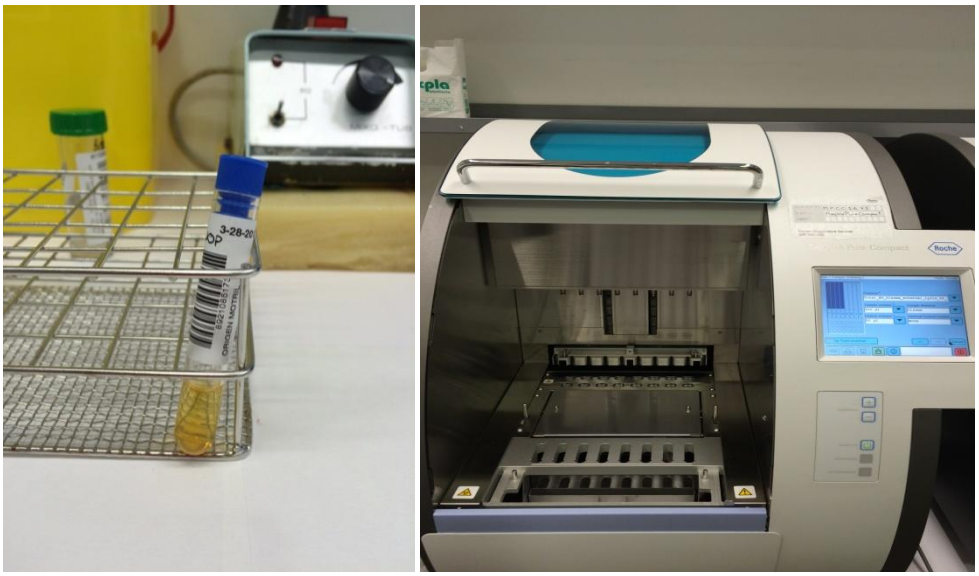


Figura 8. A la izquierda muestra de plasma de paciente VIH. A la derecha MagNA Pure (Roche).

Una cosa a tener en cuenta es que se partió del mismo extraído para la realización de los dos tipos de secuenciación: Sanger y NGS.

4.2.2. Secuenciación Sanger

4.2.2.1. Obtención de ADN desde ARN

Tras extraer el ARN llega el momento de obtener el ADN mediante una Retrotranscripcion (RT-PCR), para posteriormente amplificar la región *pol*. Este fragmento contiene tres proteínas, pero nosotros amplificamos dos de ellas; la región Proteasa y Transcriptasa Reversa, ya que la tercera enzima (Integrasa) no se hace a todos los pacientes. Por lo tanto nuestra región comprendió de la posición 4-99 de Proteasa y 38-247 de la Trancriptasa Reversa, siendo amplificado mediante el kit TruGene VIH type 1 (VIH-1) genotyping kit/OpenGene ADN sequencing system (SIEMENS).

La preparación de la RT-PCR se muestra en la tabla 1, donde se puede consultar también la preparación final que fue llevada al termociclador.

Componente	**µL de componente (1 muestra)**
RT-PCR Primer	7
DTT	1,2
dNTPs (10 mM)	1,8
RNasa-Inhibitor	0,6
El volumen final para la reacción de PCR sería de 26µL (9µL MMix + 17µL de ARN extraído)	

Tabla 1. Preparación de la master mix (MMix) de reacción RT-PCR del kit TRUGENE para VIH-1.

En la tabla 2 se muestra los tiempos y las temperaturas para la realización de la reacción en cadena de la polimerasa.

Temperatura ºC	Tiempo	Ciclos
90	2’	1
50	60’	
94	2’	
94	30“	20
57	30“	
72	1‘ 30“	
94	30”	15
60	30”	
70	2’	
70	7’	1
4	∞	

Tabla 2. Programa termociclador para la amplificación de la región a estudiar.

A los cinco minutos de poner el programa fue parado para adicionar la segunda MMix (tabla 3).

Componente	µL de componente (1 muestra)
RT-PCR Buffer	11,7
RNase-Inhibitor	0,6
RT-Enzime	1,2
ADN-Polimerasa	2,9
Añadiremos 14µl de MMX-II a cada muestra, posteriormente continuara con el mismo programa de RT-PCR.	

Tabla 3. Preparación de la MMix II reacción RT-PCR del kit TRUGENE para VIH-1.

Tras realizar la amplificación, se comprobó el resultado mediante electroforesis en gel de agarosa (figura 9), en la cual se puede apreciar una banda fluorescente con el tamaño deseado. Hay que remarcar que el gel se puso sin peso molecular, debido a que en ese momento no había en el laboratorio. El producto amplificado tiene que tener un peso molecular de 1000 pares de bases.

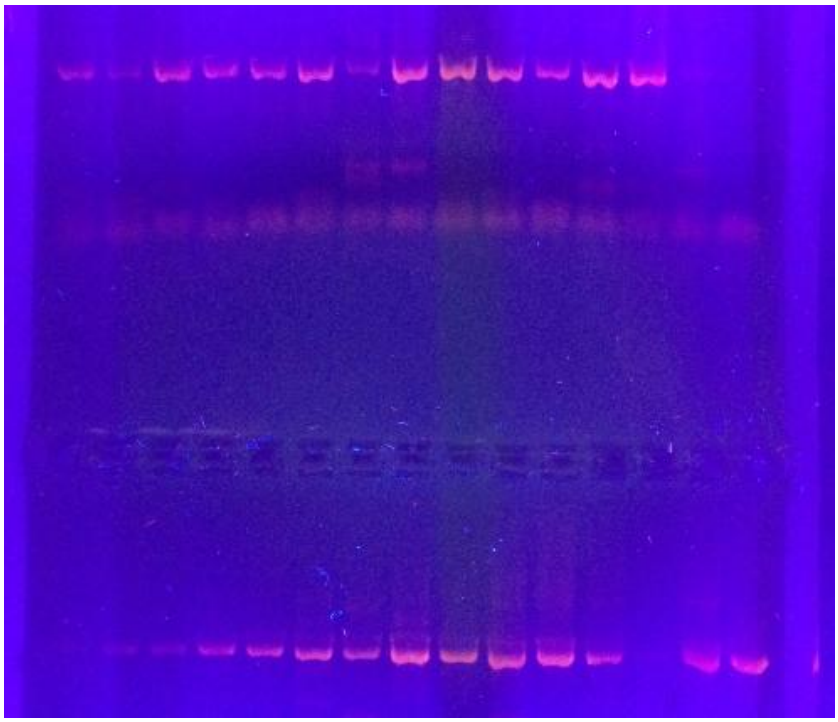


Figura 9. Gel de agarosa al 1%, amplicones producidos en la PCR con luz UV.

4.2.2.2. Amplificación de ADN

La reacción de secuenciación realizada se basa en la secuenciación bidireccional conocida como CLIP^{TM}, la cual no es más que una modificación del método de Sanger que emplea terminadores de la secuencia de ADN en forma de didesoxinucleótidos (ddNTP). Los ddNTPs carecen del grupo –OH en el carbono 3 del azúcar que es el responsable de la unión de los nucleótidos entre sí, a través del grupo fosfato para incorporase a la cadena del ácido nucleico. Por lo tanto, si un enzima al elongar una cadena de ADN coloca en vez de una Adenina en forma de dATP una ddATP, la cadena que se está formando se trunca y no puede seguir siendo elongada.

En este caso los terminadores no están marcados, son los cebadores los que se marcan. En concreto, un primer incorpora Cy5.5 como fluoróforo en 5´, y el otro cebador incorpora Cy5.0 en 3´. Estos dos fluoróforos tienen la característica de emitir luz a distintas longitudes de onda (690 y 670 nm respectivamente) de modo que podremos diferenciar con un láser la emisión de los productos de secuenciación iniciados con cada uno de estos dos cebadores y, de esta manera, conoceremos en un mismo tubo la secuencia sentido y antisentido (*forward* y *reverse*). La electroforesis en gel de

poliacrilamida la realizamos en un secuenciador semiautomático (Long-Read TowerTM) previa polimerización del gel de acrilamida. El secuenciador dispone de un sistema láser que hace lecturas cada 30 segundos de manera que va registrando las señales que emiten los fluoróforos: el producto de secuenciación en la dirección 5´ que incorpora Cy5.5 y emite a 550 nm y en dirección 3´ marcado con Cy5.0, que emite a 670 nm. El secuenciador tiene un sistema de registro de ambas longitudes de onda de modo que lee a la vez la secuencia de la dirección 5´ y 3´ (secuenciación bidireccional).

La reacción de secuenciación CLIP^{TM} para VIH se realizó siguiendo el protocolo indicado TruGene VIH type 1 (VIH-1) genotyping kit/OpenGene ADN sequencing system (SIEMENS).

Para llevar a cabo la secuenciación, se necesitan 4 tubos (A, C, G, T). En cada uno de ellos se añaden los *primers* de secuenciación marcados con Cy5.5 y Cy5.0, el didesoxinucleótido correspondiente (ddATP, ddCTP, ddGTP y ddTTP) y los dNTPs.

Preparamos la MMix con los siguientes reactivos en el orden y proporciones que se especifican en el protocolo del kit (tabla 4).

Componente	**µL de componente (1 muestra)**
Agua	60,5
CLIP Buffer	15,3
CLIP Enzime	2,9
El volumen final para la reacción de CLIP sería de 10µL (5µL MMix + 5µL de ADN amplificado en RT-PCR), en los cuatro tubos por muestra.	

Tabla 4. Preparación de la MMix de la reacción CLIP del kit TRUGENE para VIH-1.

Precalentamos el termociclador a 94ºC, colocamos los tubos y ejecutamos el programa de secuenciación (tabla 5).

Temperatura ºC	Tiempo	Ciclos
94	5’	
94	20”	30
62	20”	
70	2’	
70	5’	
4	∞	

Tabla 5. Programa de secuenciación CLIP.

Una vez finalizado, añadimos 10 µl de solución de parada (formamida + colorante) por tubo. En este momento se pueden conservar las muestras a 4ºC hasta el día siguiente.
Antes de realizar la electroforesis de las muestras, desnaturalizamos a 90ºC durante cinco minutos en termociclador. Pasamos las muestras inmediatamente a hielo para evitar que las hebras separadas hibriden de nuevo y puedan ser secuenciadas.

4.2.2.3. Secuenciación en electroforesis en gel de poliacrilamida de los amplificados

Los fragmentos de la secuenciación fueron separados mediante electroforesis en gel de acrilamida (Surefill 500®) utilizando el sistema Trugene Tower (Siemens®) a 60°C, 1600V, 50% potencia de láser, 0.5 segundos de *sampling rate* durante 60’ (figura 10).

Figura 10. Estación de secuenciado TRUGENE.

4.2.2.4. Alineamiento de las secuencias e interpretación de los resultados

El paso final para la secuenciación consiste en el alineamiento de las secuencias (figura 11). Mediante este procedimiento el software integra los 2 cromatogramas obtenidos para la región secuenciada (*pol*) y para cada dirección (5´ y 3´). Asimismo, enfrentamos las secuencias obtenidas con una cepa de referencia (HXB2) en la que no existen mutaciones de resistencia (cepa salvaje, *wild type* o WT). A continuación, revisamos la secuencia completa, viendo las bases en las que no exista concordancia entre las direcciones 5´ y 3´ o con la secuencia *wild type* y revisando todas las posiciones de resistencia. En estos pasos debemos ser extremadamente cautelosos ya que somos nosotros los que vamos a decidir si estamos o no de acuerdo con la interpretación que el *software* realiza. Las secuencias obtenidas son transformadas a formato de texto para ser así exportarlas y poder trabajar con ellas.

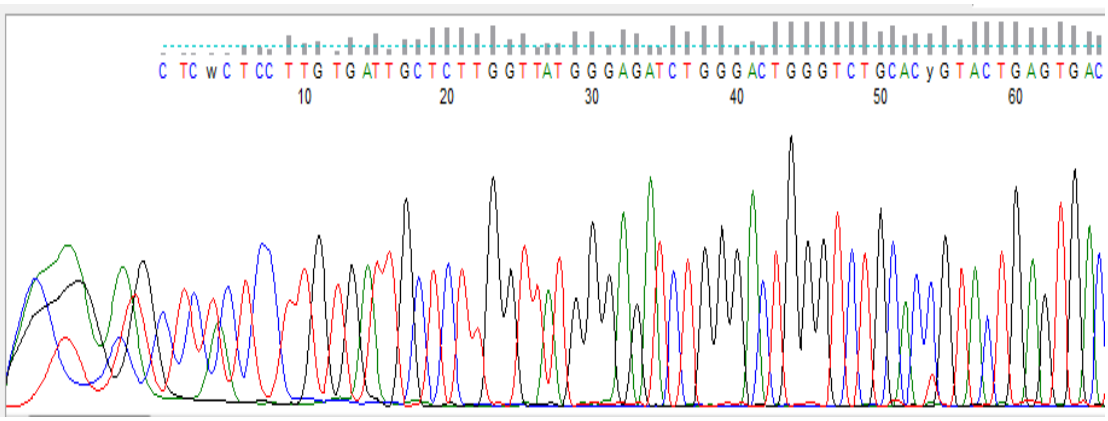


Figura 11. Proceso de alineamiento e interpretación de la secuencia. Cromatograma en el cual cada base nitrogenada corresponde con un color y se alinean para dar la secuencia.

4.2.2.5. Análisis de las secuencias obtenidas

Las secuencias fueron observadas mediante un programa de visionada de secuencias como es Chromas, mediante este programa revisamos la secuencia para comprobar que las posibles mezclas de bases nucleotídicas en una posición esta correctamente, además de eliminar posibles inserciones-delecciones.

Las secuencias fueron extraídas de la estación de trabajo en formato FASTA, a continuación se observa una secuencia con el siguiente aspecto:

>|050780282426|pol201406090940vd||Siemens nucleotide|
CCTCGTCACAATAAAGATAGGGGGGCAACTAAAGGAAGCTCTATTAGATAC
AGGAGCAGATGATACAGTATTAGAAGAAATGAGTTTGCCAGGAAGATGGA
AACCAAAAATGATAGGGGGAATTGGAGGTTTTATCAAAGTAAGACAGTATG
ATCAGATACTCATAGAAATCTGTGGACATAAAGCTATAGGTACAGTATTAG
TAGGACCTACACCTGTCAACATAATTGGAAGAAATCTGTTGACTCAGATTG
GTTGCACTTTAAATTTTCCCATTAGCCCTATTGAGACTGTACCAGTAAAATT
AAAGCCAGGAATGGATGGCCCAAAAGTTAAACAATGGCCATTGA…

Estas secuencias se introdujeron posteriormente en la base de datos Stanford VIH (http://VIHdb.stanford.edu/), la cual nos indica si existe alguna mutación, las resistencias que estas conllevan, los porcentajes de similitud entre nuestra secuencia y la del virus salvaje y el tamaño de nuestro fragmento amplificado.

4.2.3. NGS

4.2.3.1. Amplificación de ADN extraído

En este tipo de secuenciación también procederemos con el paso de ARN extraído a ADN. Para ello seguiremos el protocolo de 4 Plate VIH-Drug Resistance Assay Manual (Roche). Este tipo de protocolo nos permite trabajar en unas placas donde presentan los *primers* liofilizados en sus pocillos.

Preparamos la siguiente MMix (tabla 6):

Componente	**µl de componente (1 muestra)**
Transcriptor RT Reaction Buffer (x5)	4
dNTP	2
Protector RNase inhibitor	0,5
Transcriptor reverse Transcriptase	0,5
El volumen final para la reacción será de 20.5µL (7µL MMix + 13.5 µL de ARN extraído.	

Tabla 6. Preparación de la MMix en la reacción de retrotranscripción.

Se realizo la reaccion con el siguiente programa (tabla 7):

Temperatura ºC	Tiempo
50	60’
85	5’
4	∞

Tabla 7. Programa retrotranscripción.

Posteriormente, hicimos una PCR con el ADN obtenido. Para ello Roche tiene unas placas diseñadas con oligos específicos para amplificar la región Proteasa y Trancriptasa Reversa, junto con un fragmento llamado MID que corresponde a cada paciente. Como la región de la Transcriptasa Reversa es muy extensa debemos procesarla en tres pocillos distintos, obteniendo por tanto tres fragmentos que posteriormente fueron ensamblados. Realizamos la siguiente MMix (tabla 8):

Componente	**µL de componente (1 muestra)**
Agua	16,5
FastStart High Fidelity Reaction Buffer	2,5
$MgCl_2$	2,25
dNTP	0,5
FastStart High Fidelity Enzyme Blend	0,25
El volumen final para la reacción será de 25 µL (22µL MMix + 7 µL de ADN.	

Tabla 8. Preparación de la MMix de la reacción PCR.

El programa (tabla 9) para realizar esta reacción fue:

Temperatura ºC	Tiempo	Ciclos
95	3’	
95	30”	43
55	20”	
72	50”	
72	8’	
4	∞	

Tabla 9. Programa de secuenciación PCR.

4.2.3.2. Purificación de amplicones

Finalizada la PCR y teniendo nuestros productos amplificados tendremos que purificarlos, ya que así evitaremos posibles errores y eliminaremos secuencias cortas. Para ello utilizamos las AMPure XP, éstas son unas bolas magnéticas que se unen al ADN, lo que permitirá separar el ADN en buen estado mediante un imán.

Añadimos 22,5 µl de nuestro ADN junto a 45 µl de AMPure, se dejo transcurrir 10 minutos a temperatura ambiente para que se produjera la unión del ADN a las bolas magnéticas. A continuación se pusieron las muestras en un imán y retiramos el sobrenadante, recuperando nuestro producto de ADN purificado con la adición de 20 µl de *Buffer* TE.

4.2.3.3. Cuantificación de amplicones

Fue necesario cuantificar en ng/µl nuestros productos amplificados para la generación de una buena librería. El propósito de la cuantificación fue que todas las muestras llegasen a tener los mimos moles y poder amplificar todas las muestras por igual. Para ello utilizamos una hoja de cálculo, donde haremos diluciones para conseguir que todas las muestras estén en equidad.

Para cuantificar los amplicones utilizamos NanoDrop One (figura 12), el cual nos informo de la cantidad de ADN que presenta nuestra muestra con tan solo 1 µl de muestra.

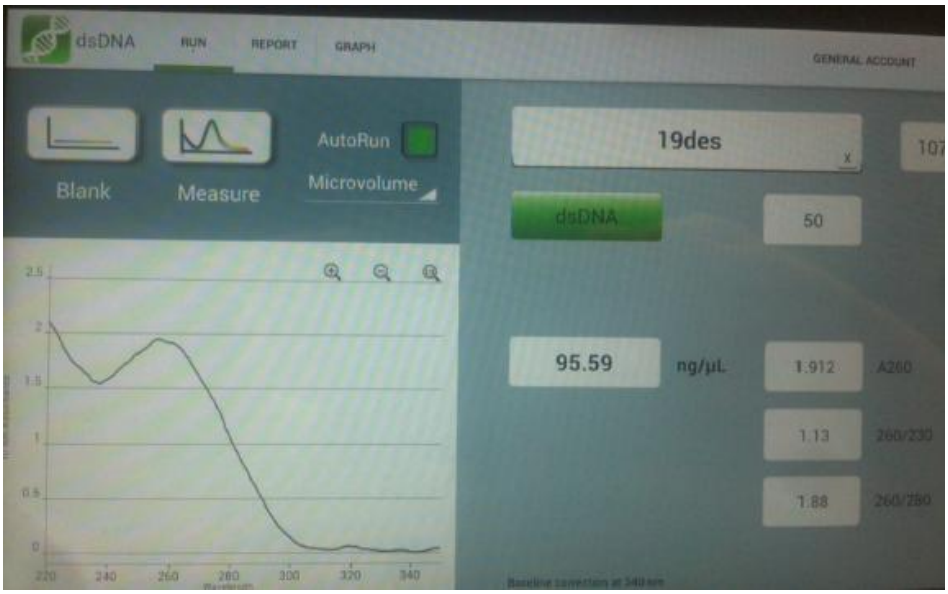


Figura 12. Cuantificación de muestra purificada VIH mediante NanoDropOne.

4.2.3.4. Creación de la librería

Esta técnica se realiza *in vitro*. Una vez estandarizados todos nuestros amplicones a la misma concentración se procedió a la creación de librería de nuestras muestras. Esta creación de librería no será más que el proceso de fragmentación mediante "nebulización". Una vez obtenido el ADN en fragmentos de 200-800 pares de bases pudimos añadir dos *primers* que contienen regiones en los extremos (adaptadores), estos adaptadores permitieron la amplificación y secuenciación, el adaptador A fue utilizado en dirección *Forward*, mientras que el B *Reverse*. Además estos adaptadores contienen una secuencia característica TCAG que dará inicio a la secuenciación.

4.2.3.5. emPCR

La emPCR es una PCR en emulsión, mediante un aceite conseguimos crear unas esferas, donde se introdujeron nuestros reactivos y un único amplicon, produciendo una PCR monoclonal.

4.2.3.6. Secuenciación masiva

La emPCR anterior fue tratada para eliminar las esferas que no hayan obtenido una buena amplificación de fragmentos. Posteriormente se añadieron cebadores complementarios a los adaptadores y enzimas necesarias para la pirosecuenciacion. La muestra fue cargada en un chip llamado "*PicoTiterPlate*", que fue introducido en el secuenciador.

La pirosecuenciacion es la forma de secuenciar de la plataforma 454-GS Junior, la cual consiste en destellos de luz producidos por una fuente luminosa que hace que las muestras produzcan destellos de luz gracias a la enzima luciferasa. Estos destellos de luz fueron captados por una cámara que los traducirá a nucleótidos en la secuencia (figura 13).

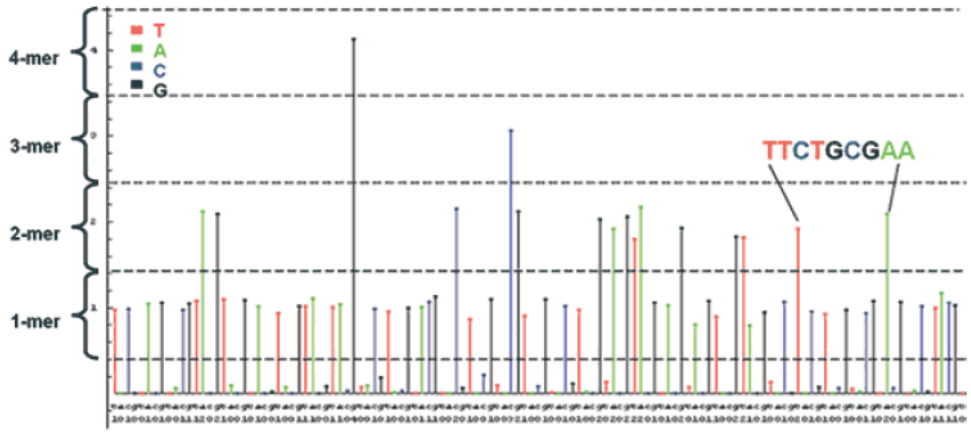


Figura 13. Resultado de la secuenciación.

4.2.3.7. Alineamiento de las secuencias e interpretación de resultados

Los resultados obtenidos tras la secuenciación de librerías de amplicones se emplearon principalmente para identificar y cuantificar variantes de ADN tanto conocidas como nuevas. Para ello se dispuso de la herramienta informática "GS Amplicon Variant Analyzer" (AVA), que se describe a continuación.

El programa AVA alinea las secuencias de la librería de amplicones obtenidas en el secuenciador, e identifica diferencias entre dichos resultados y una o más secuencias de referencia. Las diferencias se muestran de dos formas:

1. Gráficamente, mediante un histograma que muestra las variaciones en posiciones específicas.

2. Textualmente, mediante un alineamiento en colores que enfatiza las regiones y bases que varían respecto a la secuencia de referencia.

Cabe destacar que el programa muestra las variantes identificadas en una tabla resumen, permitiendo la detección y cuantificación tanto de posibles nuevas variantes como de las ya conocidas. Además, es capaz de detectar variantes de baja frecuencia (< 1%) en mezclas complejas, tales como mutaciones somáticas y cuasiespecies virales.

En nuestro estudio utilizamos la tecnología AVA para alinear las secuencias obtenidas por pirosecuenciación, exportandolas a formato FASTA para posteriormente trabajar con ellas.

Por último, diversas herramientas permiten al usuario examinar los alineamientos (figura 14) en detalle para evaluar si las variantes identificadas por el programa son viables.

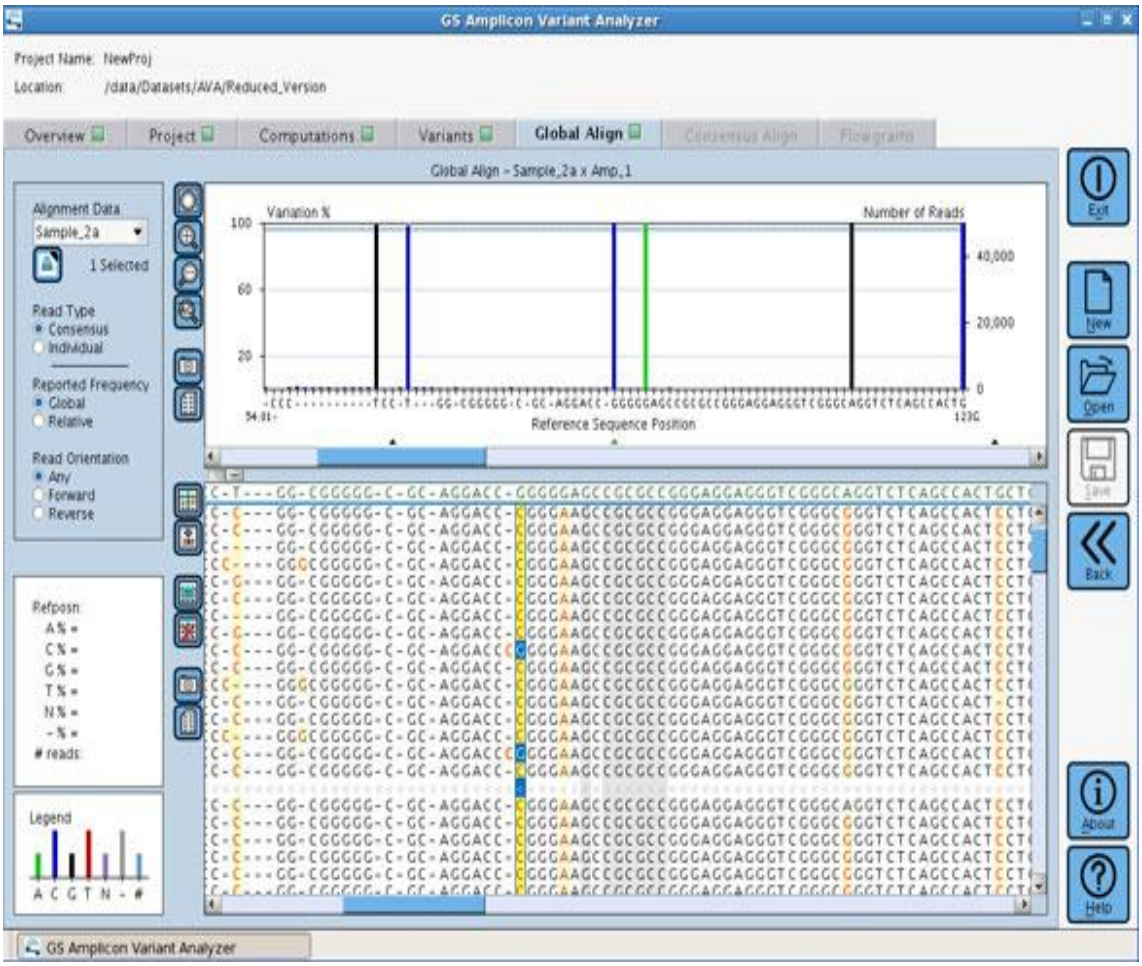


Figura 14. Captura de pantalla en la que se observa un cambio en una base comparado con una secuencia de referencia.

Un resumen esquematizado de la técnica NGS se describe en la siguiente figura 15.

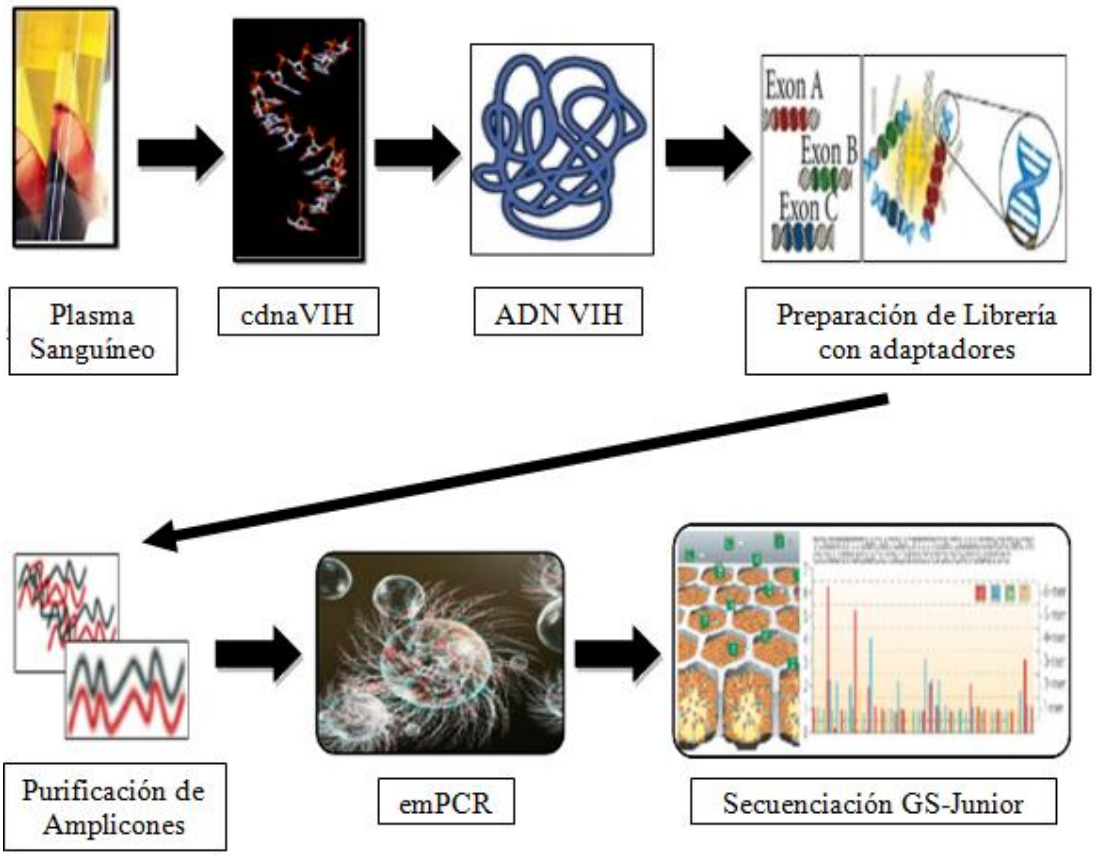


Figura 15. Descripción esquemática del sistema de secuenciación NGS mediante 454 GS-Junior.

4.2.3.8. Comprobación de la calidad de secuencias.

Al finalizar la técnica el aparato 454-GS Junior nos devolvio un archivo fasta y otro de calidad (*quality*, q.).

Si abrimos el archivo *quality* (figura 16) se puede observar la calidad de cada uno de nuestros nucleótidos, ordenados desde el principio al final. Estos valores van desde 0 a 40, por lo que un valor medio de calidad será entre 20-25, y valores de 30-40 son muy buenos.

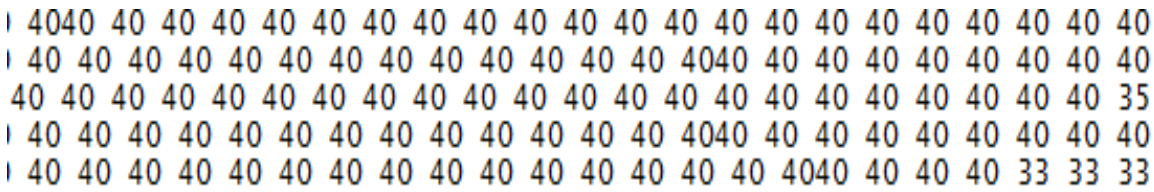


Figura 16. Archivo *quality*. Los números que aparecen es la calidad numérica de la base nucleotídica.

Mediante el servidor Galaxy unificamos los archivos fasta y *quality* para obtener un único archivos fastq. Mediante estos archivos efectuamos un filtrado de secuencias, mediante un umbral de calidad y de longitud de secuencias, para efectuar esta labor fuimos a la opción *Filter* FASTQ. Pusimos valores adecuados para eliminar las secuencias cortas (0-100pb) y secuencias que no cumplían una calidad buena (<20-25).

Aquí podemos ver un ejemplo de una muestra, mediante la opción boxplot (figura 17). Podemos observar como aumenta de calidad nuestro archivo, sobre todo en la parte final de la secuenciación, donde por motivos de la plataforma disminuye la calidad de las secuencias.

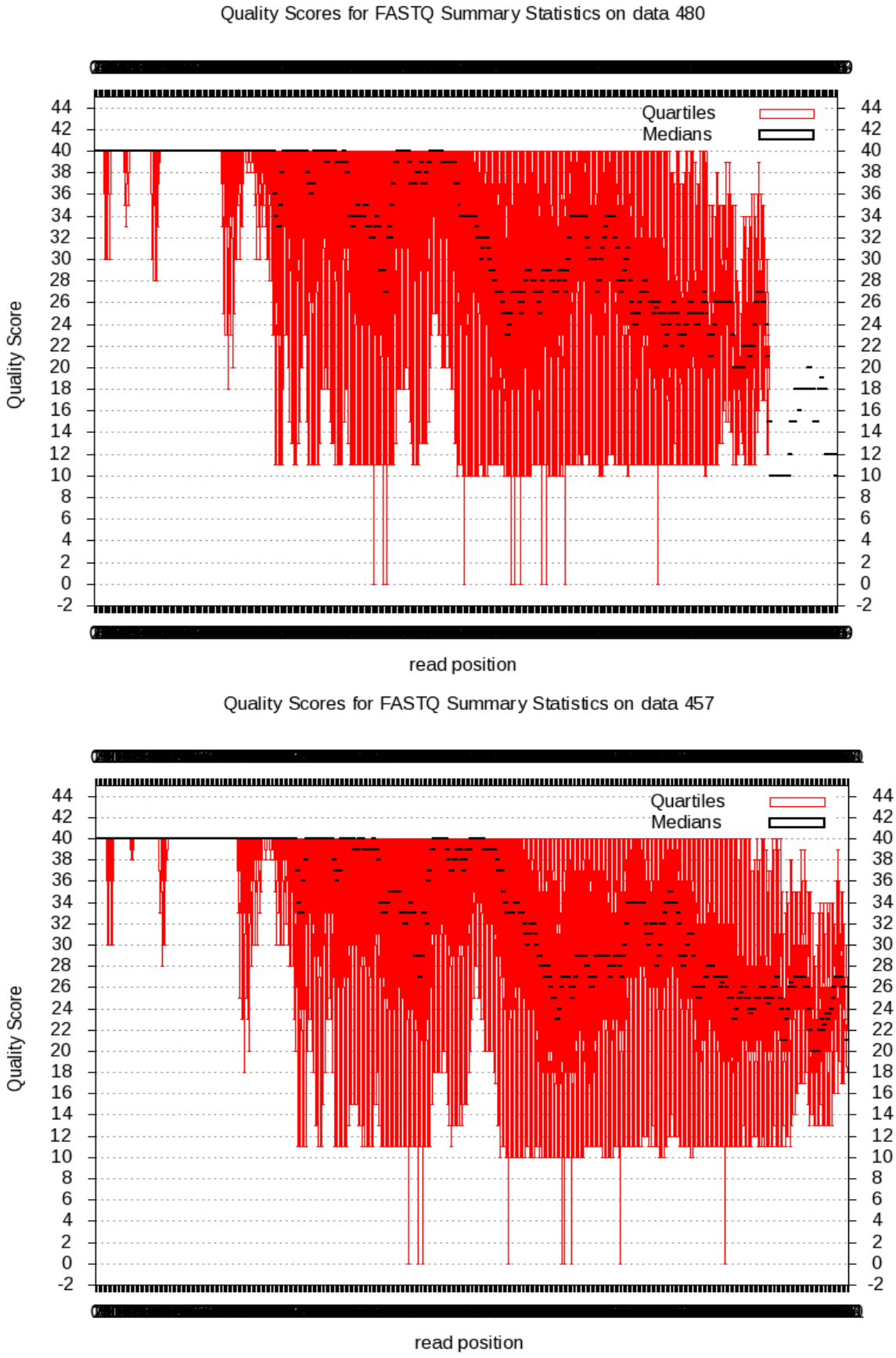


Figura 17. Boxplot de calidad de secuencias.

4.2.3.9. Eliminación de posibles errores

Después de este primer paso en el que observamos la calidad de secuencias necesitaremos observar las secuencias en un segundo paso, para asegurarnos que las secuencias consenso obtenidas no tienen errores. Para ello utilizamos un software para observar secuencias, en este caso se utiliza Jalview (figura 18). Mediante este programa eliminaremos los *gaps*, inserciones o secuencias que no cumplan una calidad óptima, todo ello mediante una secuencia VIH de referencia.

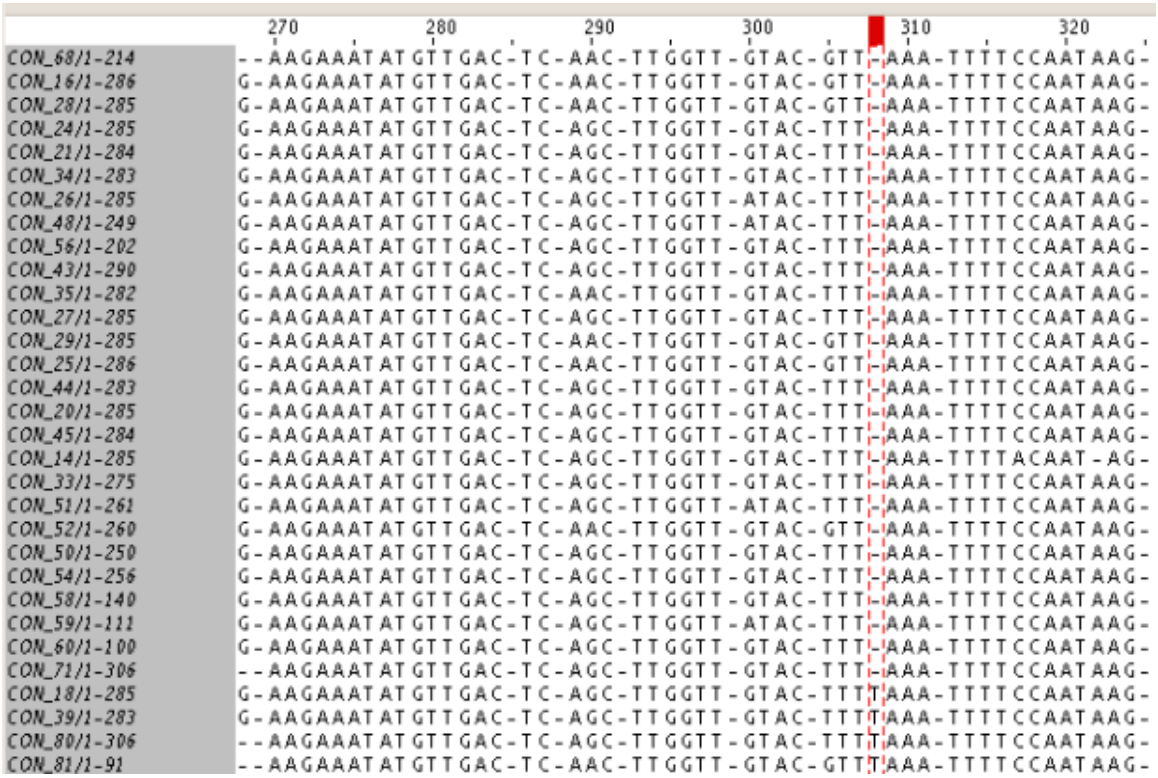


Figura 18. Se observa las distintas secuencias consenso obtenidas VIH. La posición marcada se observa una inserción con nucleótido T (Timina).

4.2.3.10. Obtención de secuencia completa región *pol*

Como ya se ha descrito en el caso de la secuenciación masiva obtenemos secuencias consenso para la región Proteasa, pero el problema es que tendremos tres amplicones para cubrir toda la región de la Transcriptasa Reversa (figura 19). Por lo tanto necesitamos un programa informático que nos pueda generar una única secuencia

consenso tanto de la región Proteasa como de la región Transcriptasa Reversa.

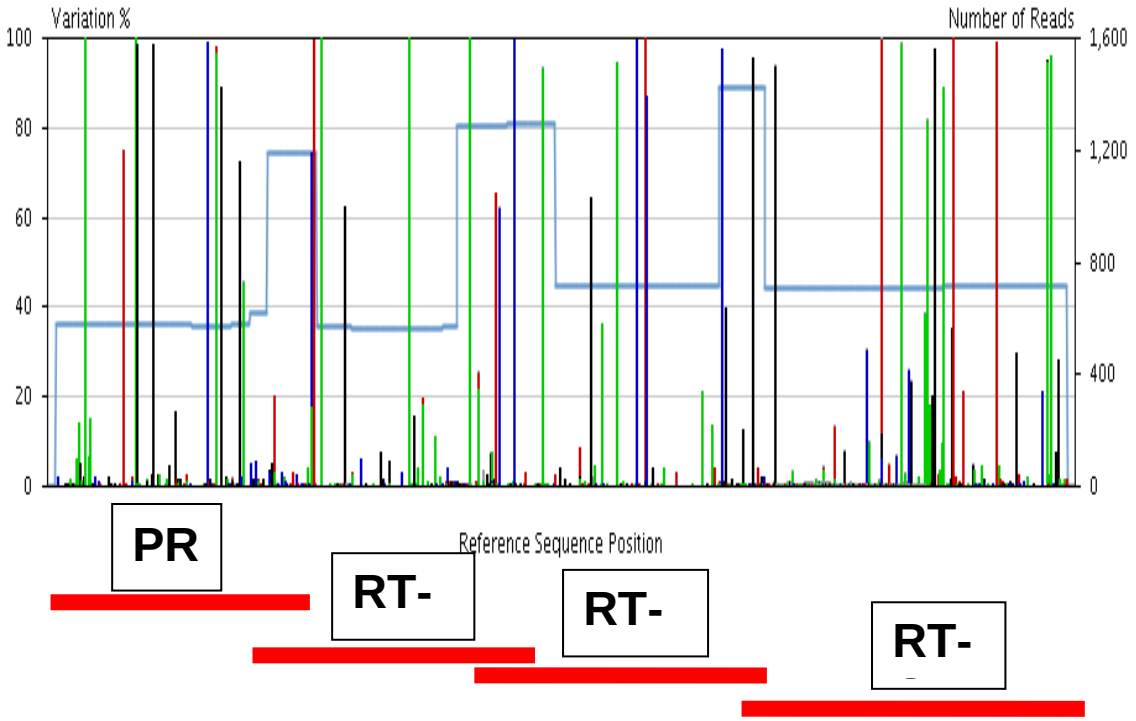
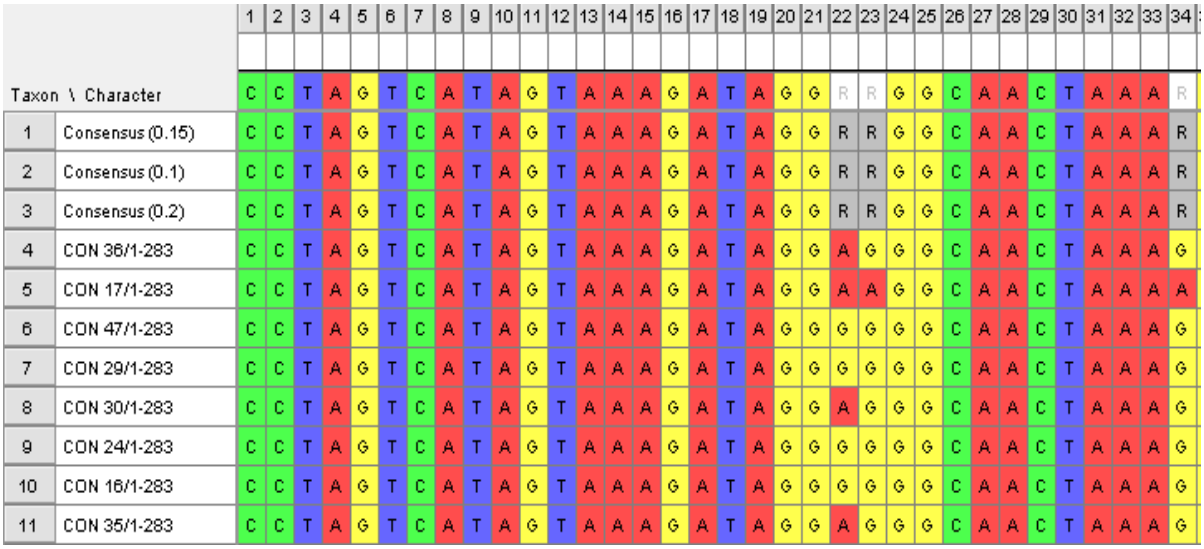


Figura 19. Cobertura de la región *pol*, se observan los distintos fragmentos de amplicones.

Haciendo una búsqueda exhaustiva obtuvimos un software llamado Mesquite, proporcionando una secuencia única teniendo en cuenta las secuencias consenso de cada fragmento. Para la utilización de este software únicamente necesitamos el archivo de nuestras secuencias en formato PFAM. Para generar esta secuencia única *pol* utilizamos los umbrales de corte al 10%, 15% y 20%, para observar cómo afectan en la obtención de la secuencia única (figura 20).



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Taxon \ Character		C	C	T	A	G	T	C	A	T	A	G	T	A	A	A	G	A	T	A	G	G	R	R	G	G	C	A	A	C	T	A	A	A	R
1	Consensus (0.15)	C	C	T	A	G	T	C	A	T	A	G	T	A	A	A	G	A	T	A	G	G	R	R	G	G	C	A	A	C	T	A	A	A	R
2	Consensus (0.1)	C	C	T	A	G	T	C	A	T	A	G	T	A	A	A	G	A	T	A	G	G	R	R	G	G	C	A	A	C	T	A	A	A	R
3	Consensus (0.2)	C	C	T	A	G	T	C	A	T	A	G	T	A	A	A	G	A	T	A	G	G	R	R	G	G	C	A	A	C	T	A	A	A	R
4	CON 36/1-283	C	C	T	A	G	T	C	A	T	A	G	T	A	A	A	G	A	T	A	G	G	A	G	G	G	C	A	A	C	T	A	A	A	G
5	CON 17/1-283	C	C	T	A	G	T	C	A	T	A	G	T	A	A	A	G	A	T	A	G	G	A	A	G	G	C	A	A	C	T	A	A	A	A
6	CON 47/1-283	C	C	T	A	G	T	C	A	T	A	G	T	A	A	A	G	A	T	A	G	G	G	G	G	G	C	A	A	C	T	A	A	A	G
7	CON 29/1-283	C	C	T	A	G	T	C	A	T	A	G	T	A	A	A	G	A	T	A	G	G	G	G	G	G	C	A	A	C	T	A	A	A	G
8	CON 30/1-283	C	C	T	A	G	T	C	A	T	A	G	T	A	A	A	G	A	T	A	G	G	A	G	G	G	C	A	A	C	T	A	A	A	G
9	CON 24/1-283	C	C	T	A	G	T	C	A	T	A	G	T	A	A	A	G	A	T	A	G	G	G	G	G	G	C	A	A	C	T	A	A	A	G
10	CON 16/1-283	C	C	T	A	G	T	C	A	T	A	G	T	A	A	A	G	A	T	A	G	G	G	G	G	G	C	A	A	C	T	A	A	A	G
11	CON 35/1-283	C	C	T	A	G	T	C	A	T	A	G	T	A	A	A	G	A	T	A	G	G	A	G	G	G	C	A	A	C	T	A	A	A	G

Figura 20. Alineamiento mediante software Mesquite. Se observa las bases según colores y letras de las distintas secuencias consenso.

4.2.4. Estudio filogenético

Los arboles filogenéticos fueron generados en MEGA 6.06, mediante el método de Máxima Verosimilitud, utilizado el modelo General Time Reversible (GTR) para el cálculo de las distancias evolutivas ,con una distribución gamma equivalente a 1,89, obtenido con FindModel DNA y utilizando remuestreo de *bootstrap* con 1000 replicas para construir los arboles filogenéticos consenso. Para definir una relación entre secuencias se tuvo en cuenta solo ramas pertenecientes a *clusters* con un valor de *bootstrap* superior al 70%. Finalmente los arboles fueron procesados en FigTree v. 1.4.2 (figura 21).

El análisis del subtipo viral se realizo utilizando el software REGA HIV-1Subtyping Tool v. 3.0.

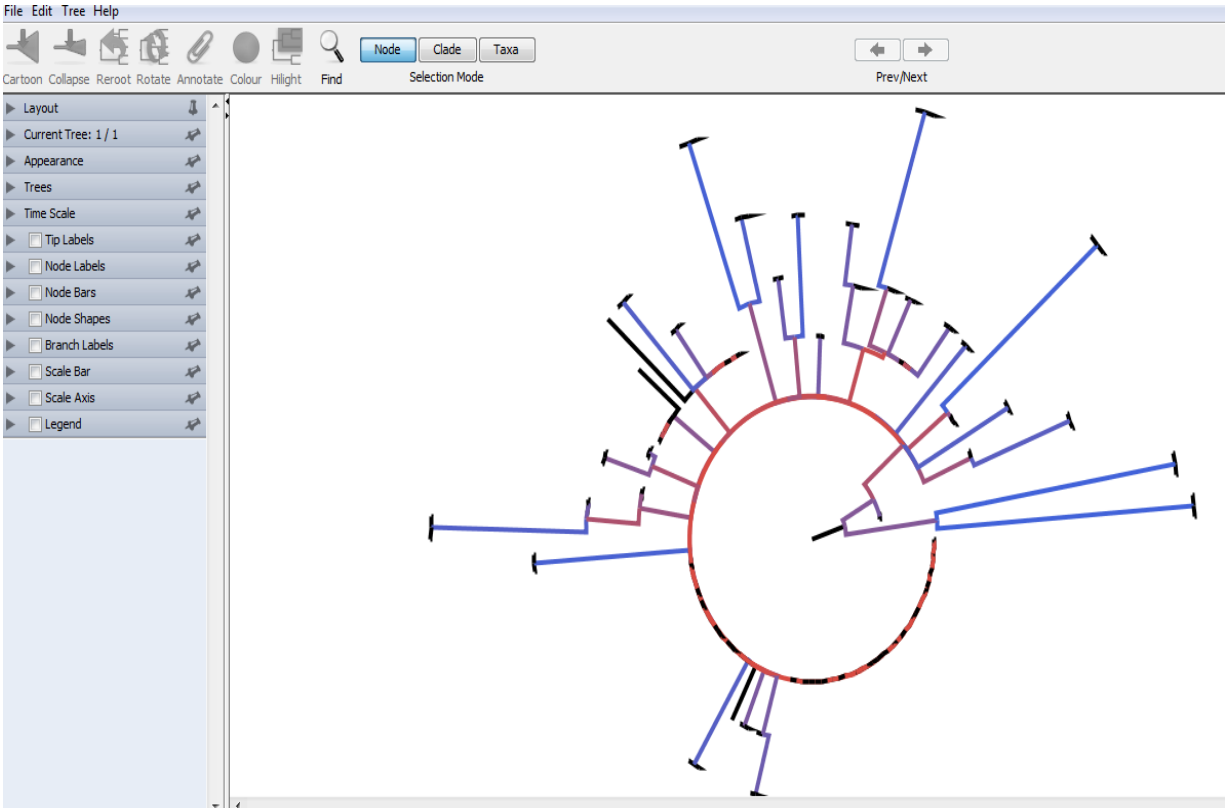


Figura 21. Interfaz del software Figtree.

5. RESULTADOS

Utilizando un umbral de corte al 10% para la creación de la secuencia consenso UDS para el gen *pol*, se observa que solo en 17/62 casos las secuencias Sanger están pareadas con UDS en la misma muestra, presentando un valor de *bootstrap* mayor de 80%, con una mediana (IQR) de *bootstrap* de 88% (83,5-95,5) (figura 22 y 23A). Aumentando el umbral consenso UDS al 15% estos valores ascienden hasta 36/62, relacionándose los dos tipos de secuencias con una mediana de *bootstrap* de 94% (85,5-98) (figura 23B). Por último, al utilizar un umbral consenso UDS al 20% se observaron que 61/62 casos presentaban las secuencias pareadas UDS-Sanger, con una mediana de *bootstrap* de 99% (98-100) (figura 23C). En este último caso, no se relaciono la secuencia UDS con su Sanger en una muestra, debido a la multitud de diferencias entre algunas bases nucleotídicas en los dos tipos de secuencias, impidiendo así establecer una relación fiable. En ningún caso hubo agrupaciones de secuencia formando *cluster* de transmisión con valores de *bootstrap* mayores a 40%.

Respecto al subtipado de secuencias, 52 muestras (83,87%) fueron subtipo B para toda la región *pol*. Hubo algunas diferencias en el subtipado, según el umbral de corte de la secuencia consenso UDS y su Sanger. Utilizando secuencias consenso UDS con umbral 10% y 15%, se observo en tres muestras una discrepancia de subtipo frente al subtipado de las secuencias Sanger, en un caso desde subtipo A1-B (UDS) a A1 (Sanger) (figura 24A-C) y en dos muestras desde subtipo B (UDS) a CRF02_AG (Sanger). Sin embargo, estas diferencias no se observaron al utilizar las secuencias consenso UDS al umbral 20%. Un ejemplo de esta discrepancia en el subtipado se puede observar en la figura 25, donde se muestra un alineamiento entre la secuencia Sanger y las secuencias consenso UDS en los distintos umbrales. La diferencia entre las secuencias UDS con distintos umbrales esta en el numero de bases ambiguas, conforme aumentamos el umbral de corte éstas disminuyen, siendo por tanto más semejante a su secuencia Sanger (figura 26).

Posteriormente se aplico esta metodología a la base de datos de pacientes VIH del año 2014- Mayo 2016 procedentes de Andalucia Oriental, efectuando la metodología anteriormente descrita de obtención de secuencias únicas UDS. Por un lado se realizo el estudio a 29 pacientes que cumplen sentencia en la cárcel de Granada (Albolote) (figura

27). Por otro lado a los pacientes que van a centros hospitalarios de Andalucía Oriental (Granada, Jaén y Almería) (figura 28). Se procesaron un total de 493 pacientes, de los cuales 198 fueron pacientes diagnosticados como *naive* (sin experiencia a tratamientos antirretrovirales).

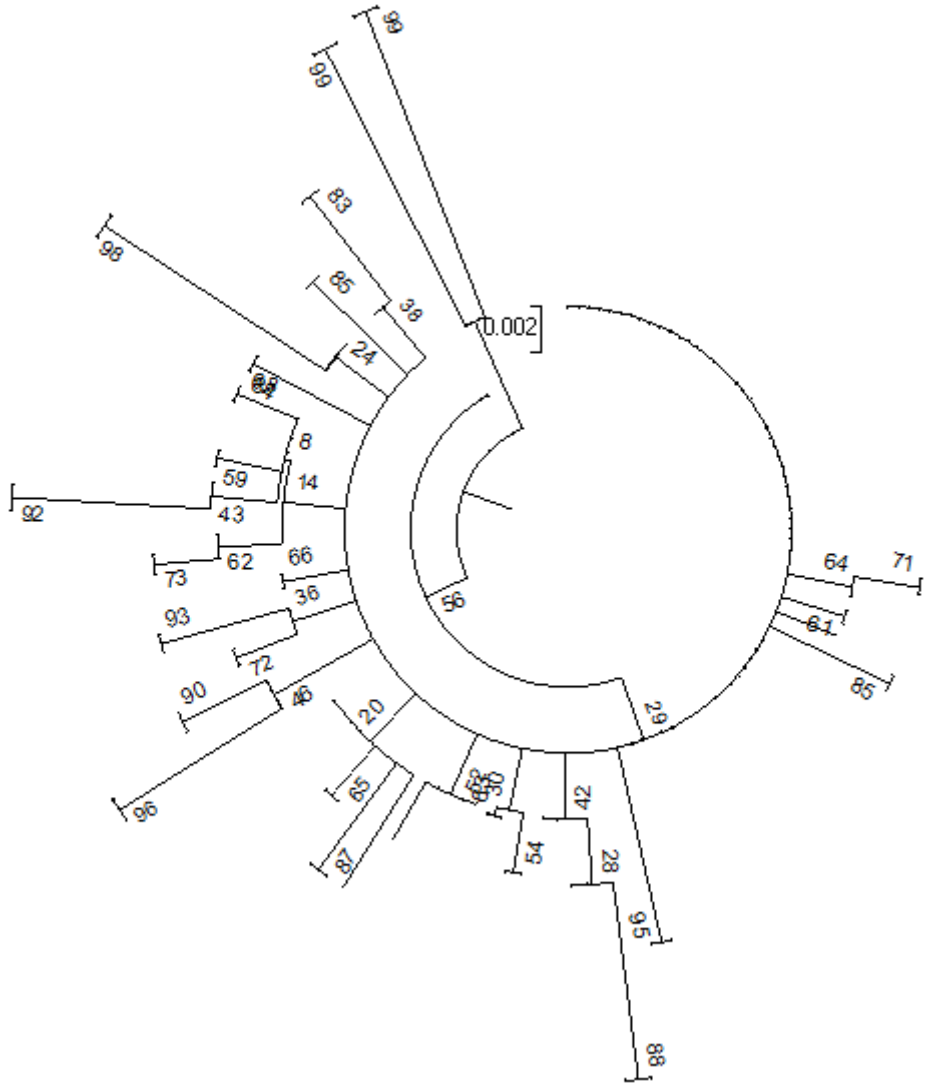


Figura 22. Árbol filogenético obtenido mediante MEGA con secuencias consenso UDS 10% y secuencias Sanger. Al final de las ramas se indican los valores de *bootstrap*.

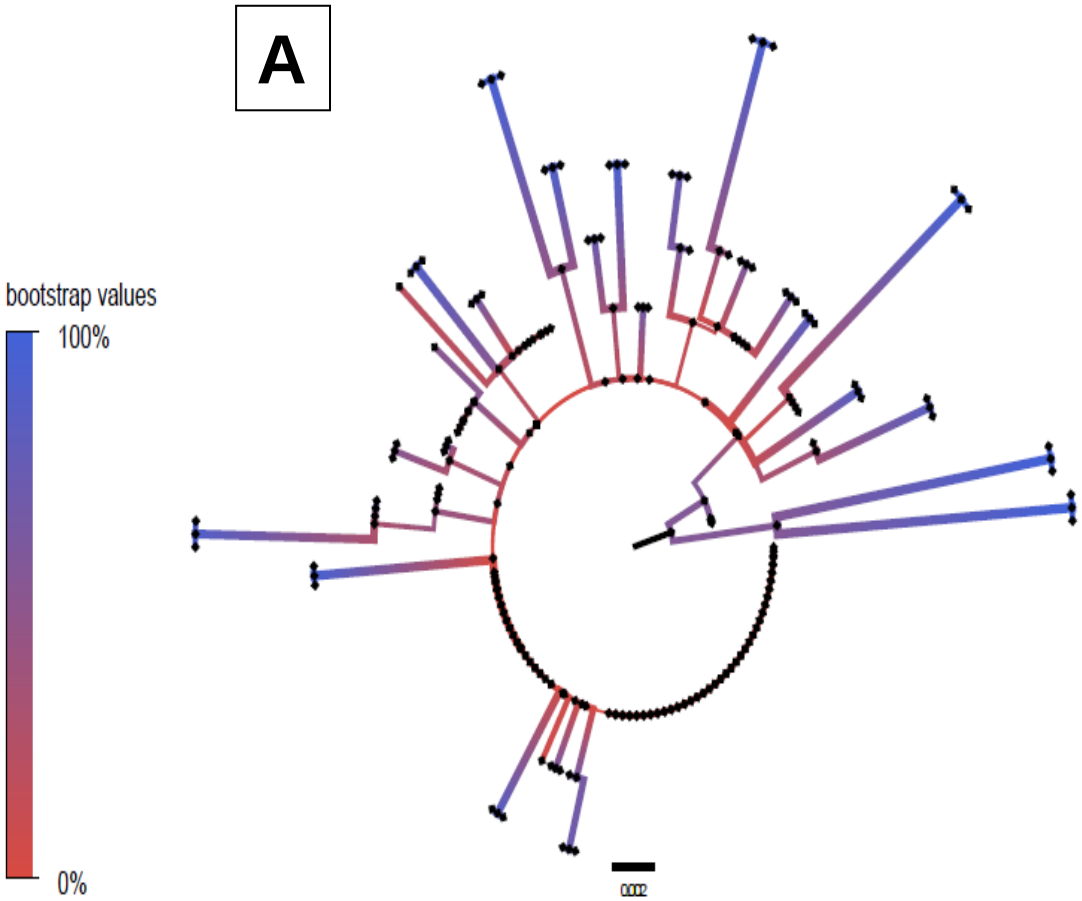
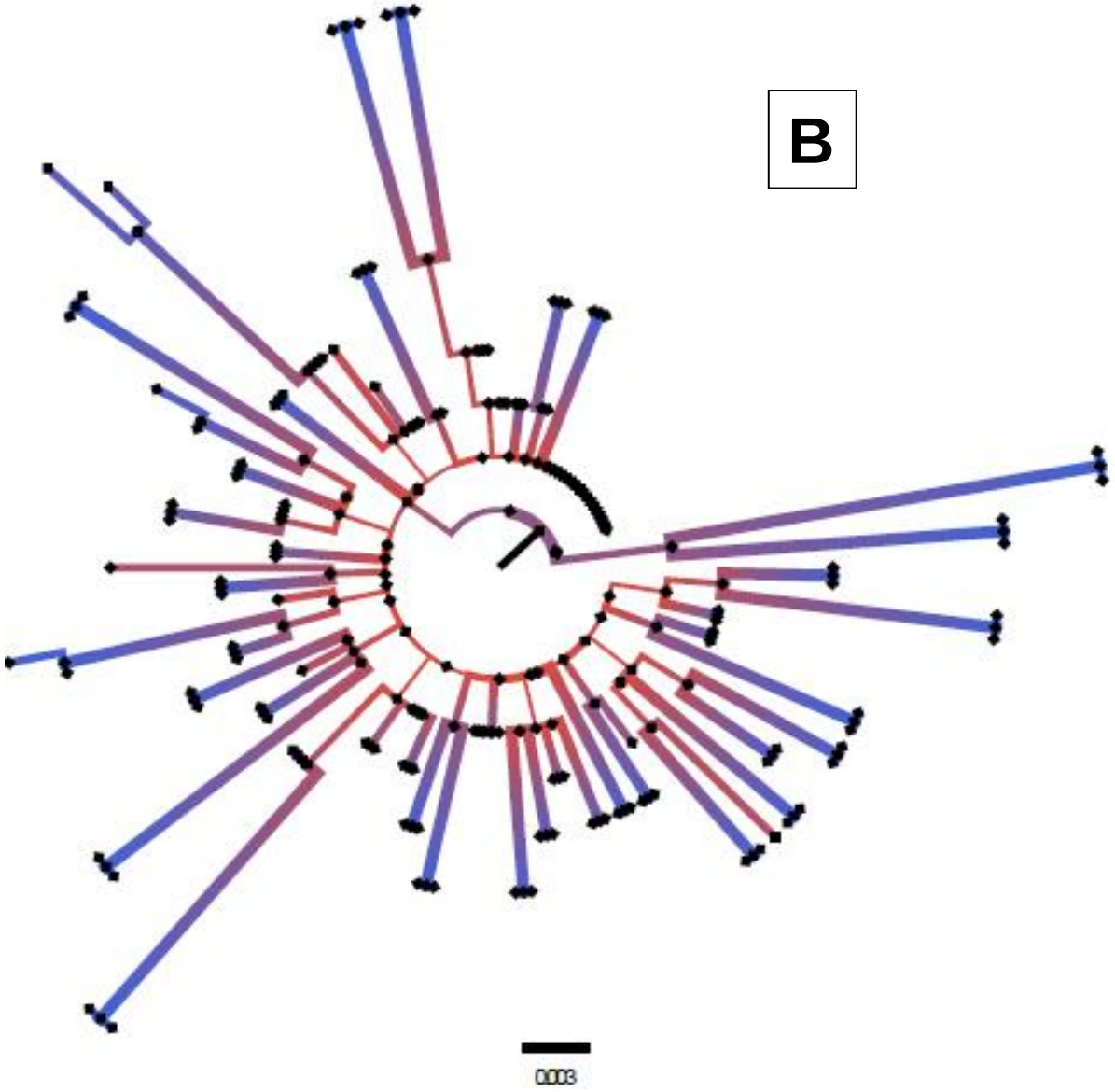
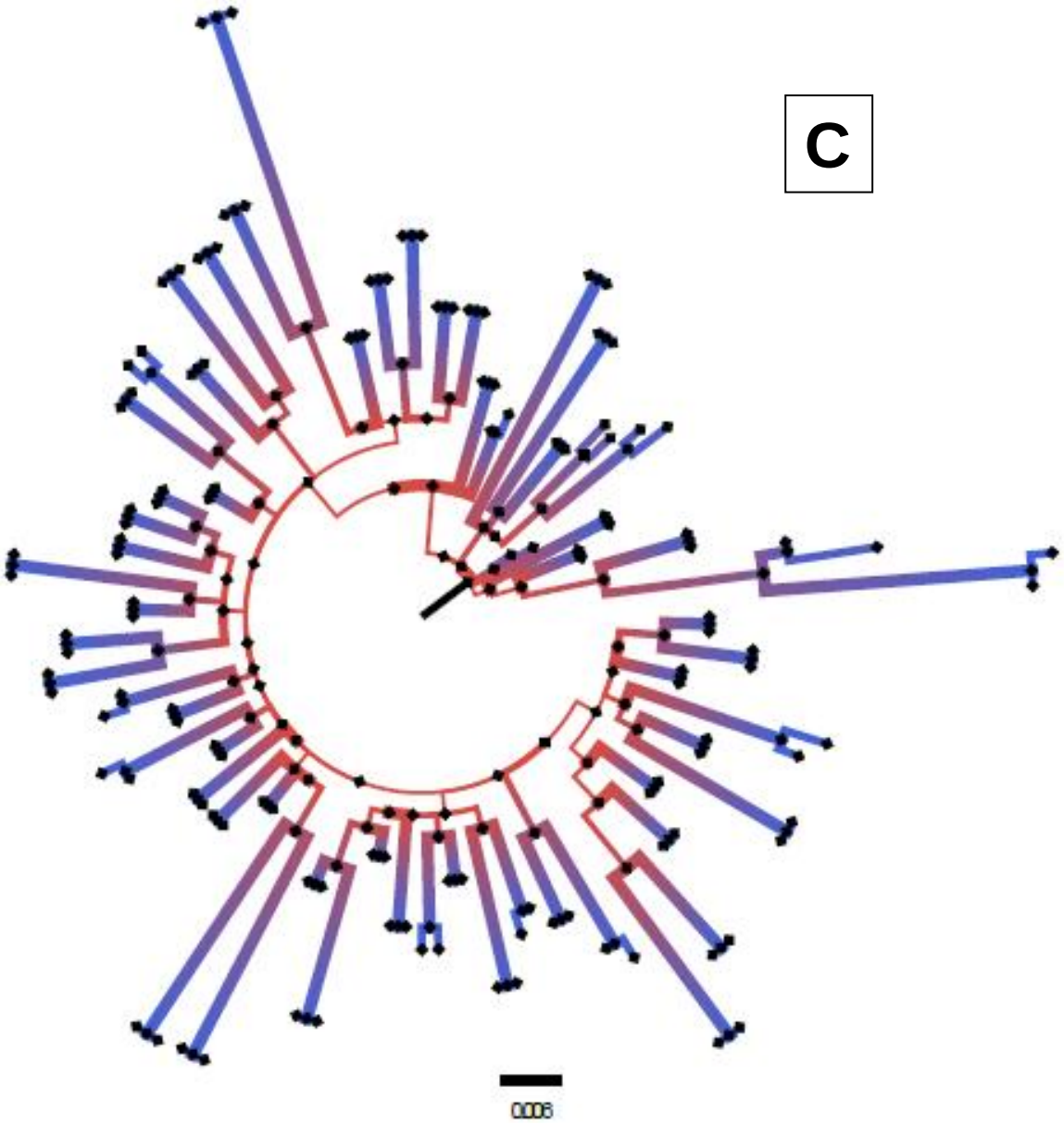


Figura 23. Representación de los arboles filogenéticos en FigTree v. 1.4.2, formados por las secuencias Sanger y secuencias UDS a los distintos umbrales, (A) UDS-10%, (B) UDS-15% y (C) UDS-20%. Los valores de *bootstrap* están asociados según color de la grafica, siendo una buena relación a partir de 70%.


B
0.003


C
0.006

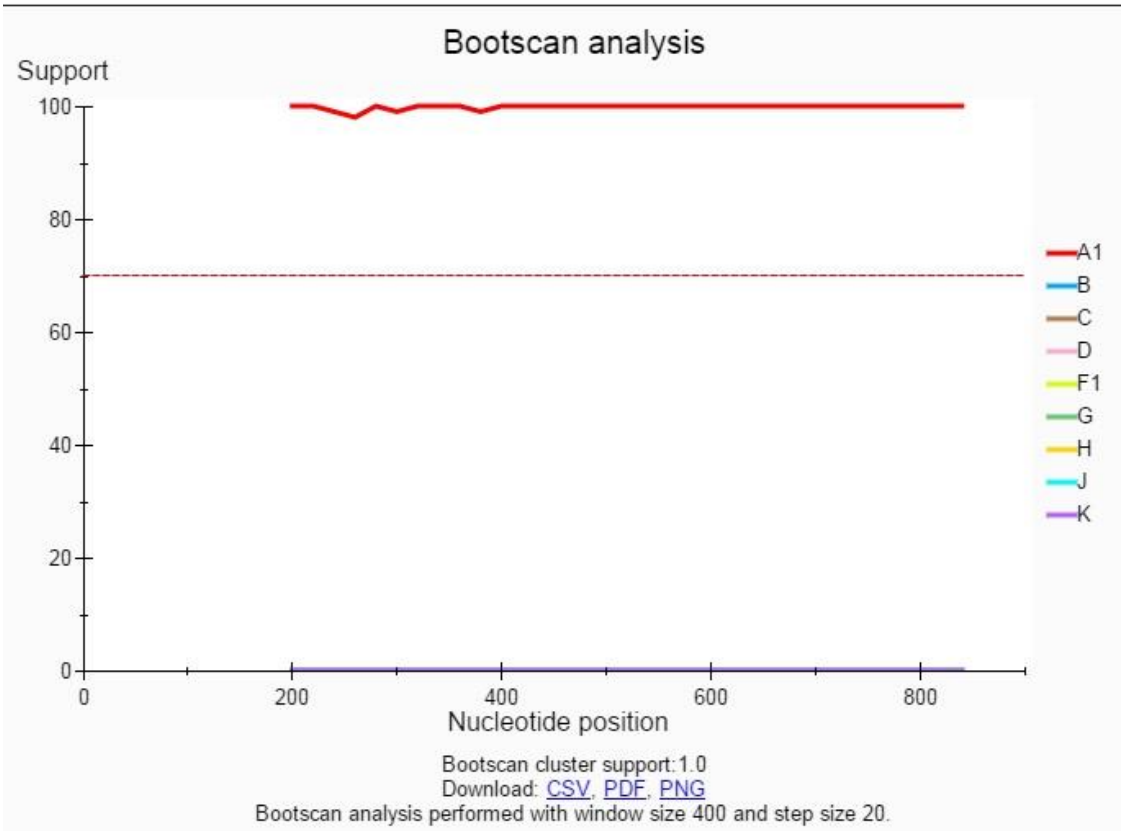


Figura 24A. *Bootscan* de secuencia Sanger. Se observa que el *bootscan* da un valor de subtipado A1.

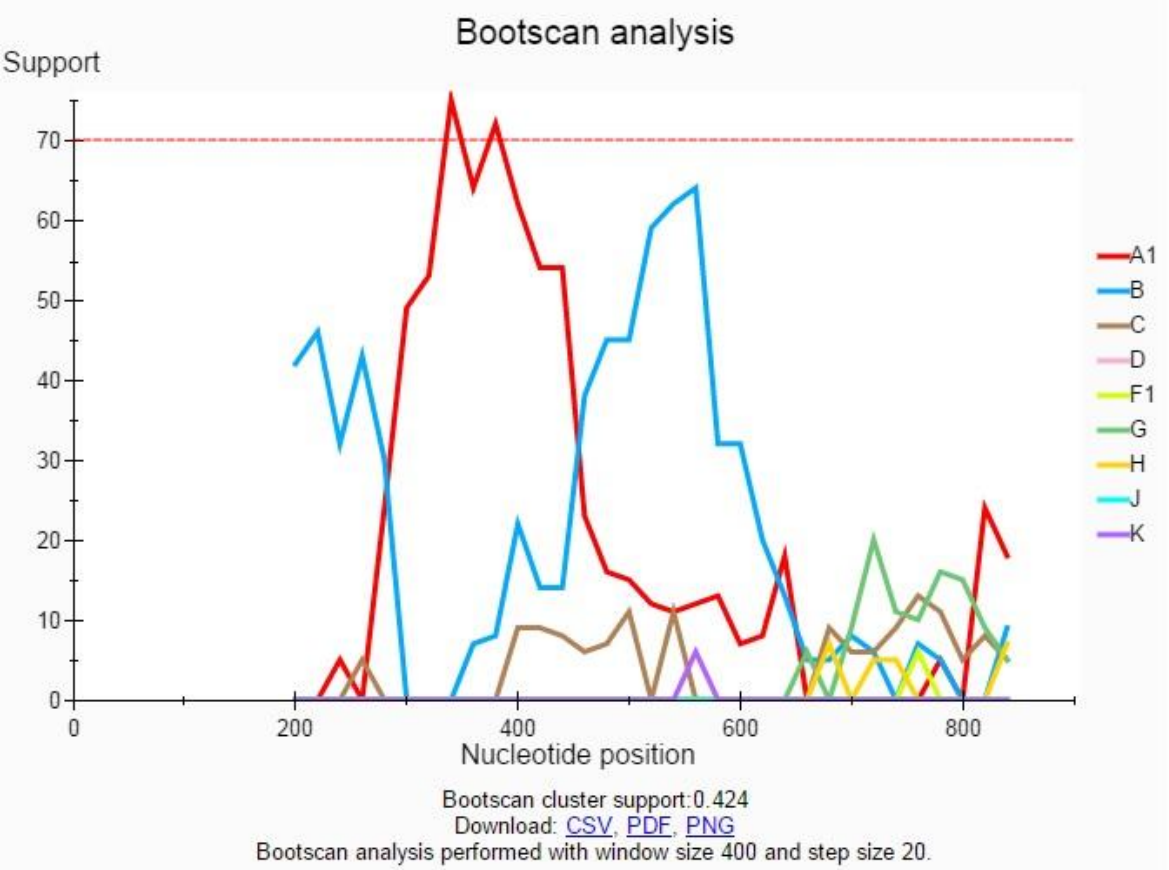


Figura 24B. *Bootscan* de secuencia UDS-10%. Se observa que el *bootscan* da un valor de subtipado recombinante A1-B.

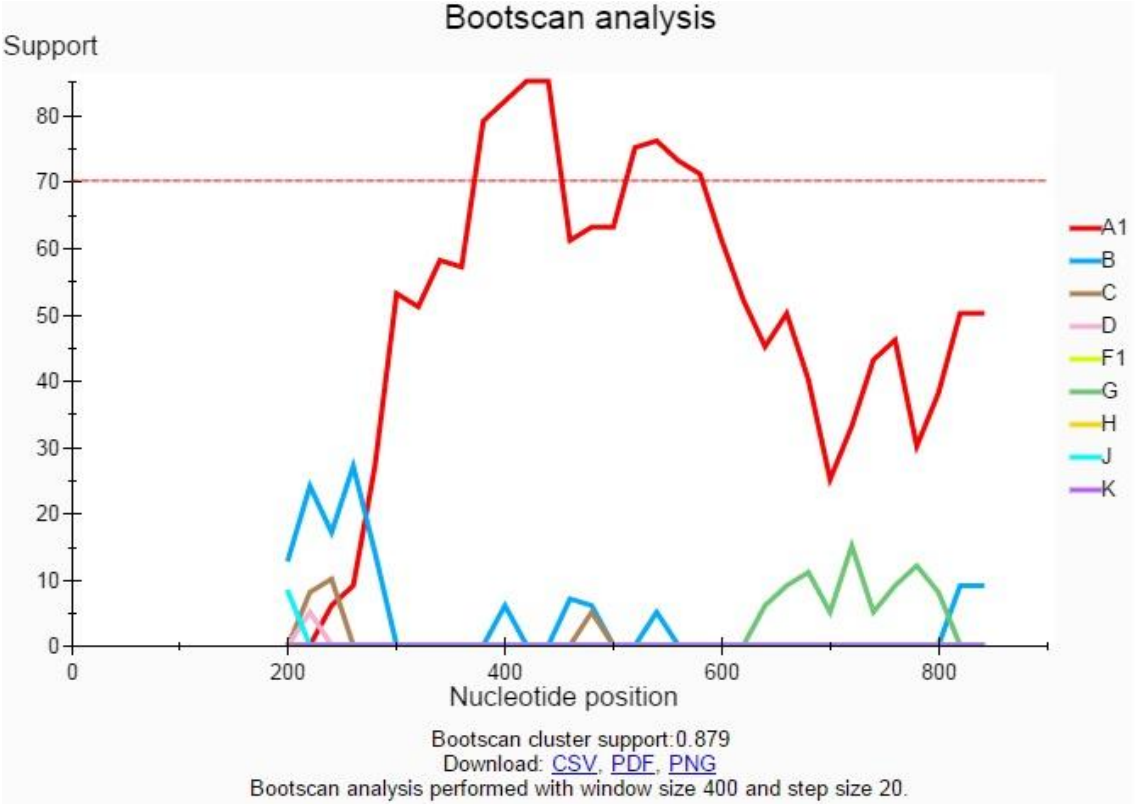


Figura 24C. *Bootscan* de secuencia UDS-20%. Se observa que el *bootscan* da un valor de subtipado A1, al igual que secuenciación Sanger para dicha muestra.

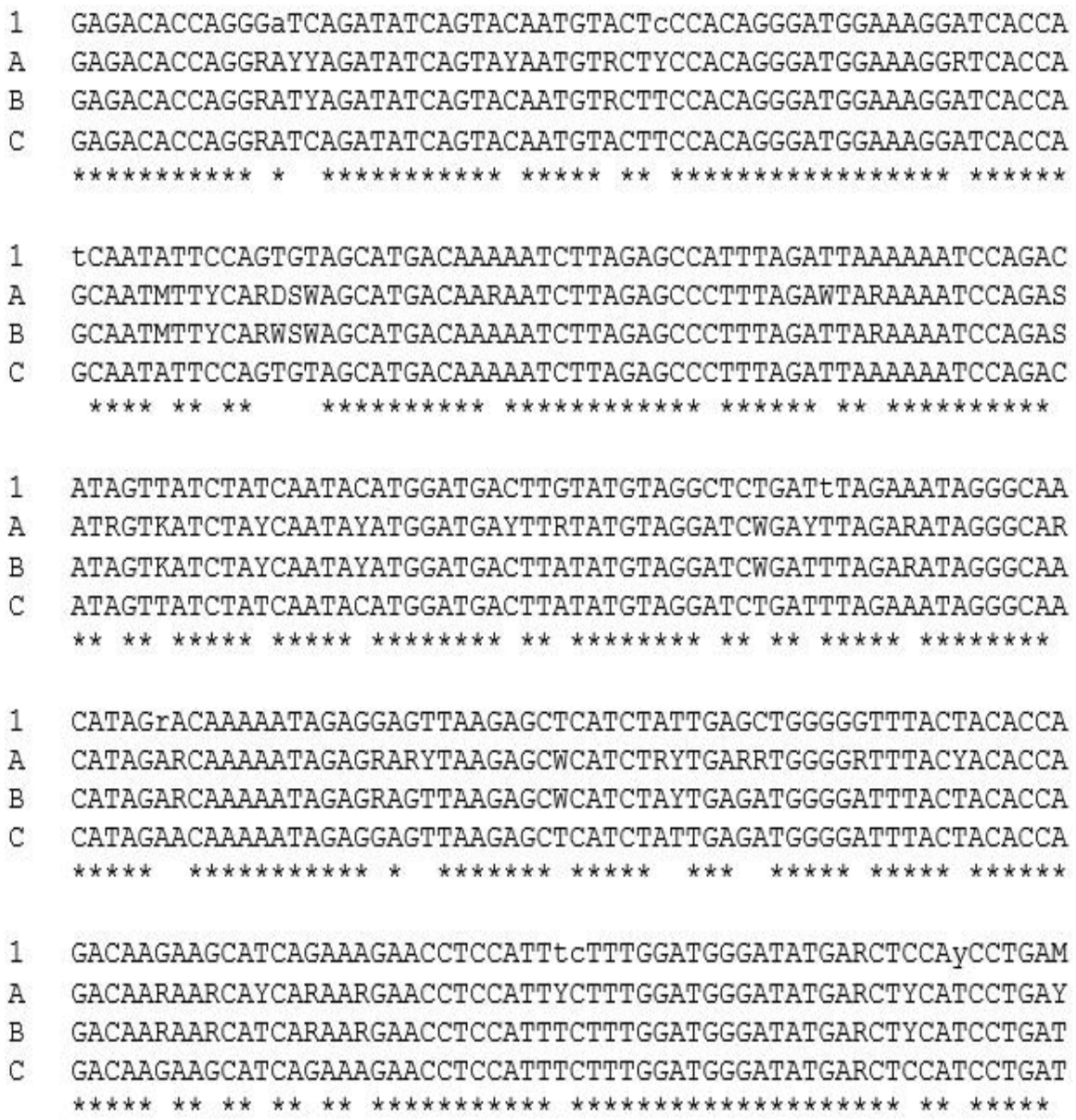


Figura 25. Alineamiento de un fragmento de región RT con múltiples secuencias (CLUSTAL O (1.2.1)), en muestra donde existe disparidad en subtipado. Numero 1 corresponde a la secuencia Sanger, (A) consenso UDS umbral 10%, (B) consenso UDS umbral 15% y (C) consenso UDS umbral 20%. Las posiciones que mantienen el mismo nucleótido están marcadas con asteriscos, por el contrario, en las que hay una disparidad no presentan asteriscos. Marcado en azul vemos las bases que con discrepancias mayores, donde existe un cambio de nucleótido entre ambos tipos de secuencia. En color amarillo están marcadas las posiciones en las que Sanger detecta mezcla y UDS no.

1	Consensus (0.1)	Y	S	T	H	G	T	C	W	H	A	R	T	A	A	R	R	R	T	A	G	R
2	Consensus (0.15)	C	S	T	H	G	T	C	W	M	A	R	T	A	A	A	R	R	T	A	G	G
3	Consensus (0.2)	C	C	T	M	G	T	C	W	C	A	R	T	A	A	A	R	A	T	A	G	G

Figura 26. Alineamiento de las secuencias consenso UDS en distintos umbrales. Se observa como a medida que aumenta el umbral disminuye las bases nucleotídicas con mezclas o ambiguas.

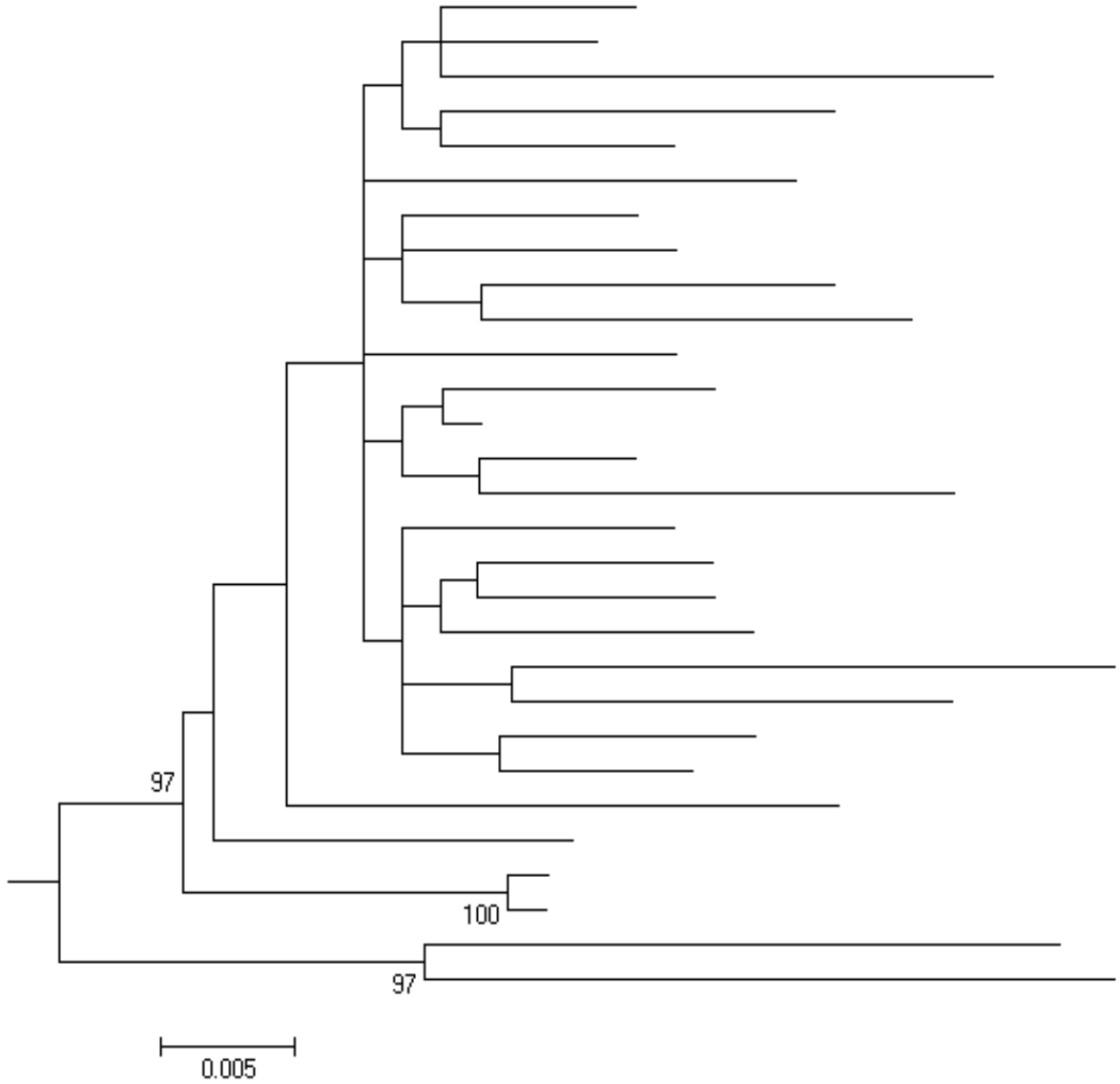


Figura 27. Pacientes VIH procedentes de la cárcel de Albolote. Se observan 29 pacientes. Hay un *cluster* con 97% de *bootstrap* formado por secuencias de subtipo B. Abajo se observan cuatro muestras relacionadas dos a dos con *bootstrap* de 100% y 97% respectivamente de subtipos no B.

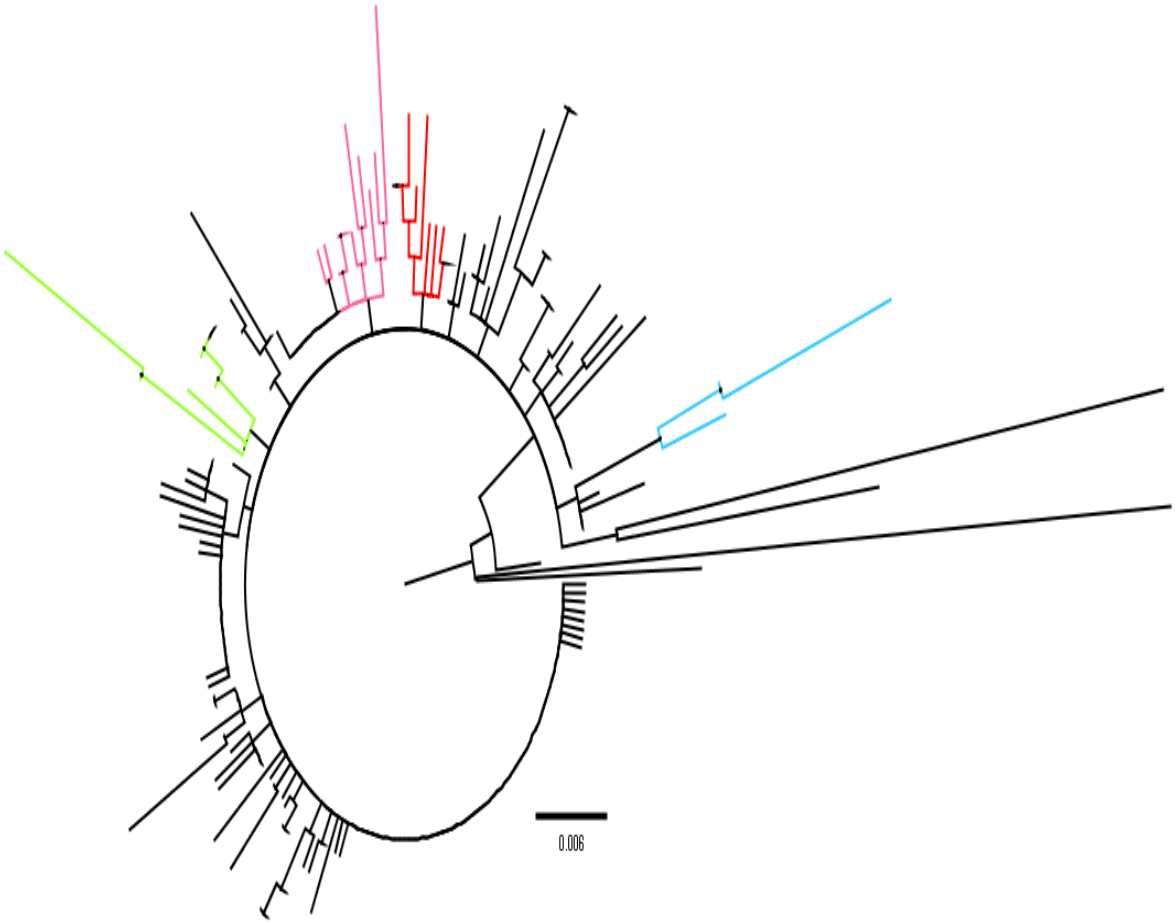


Figura 28. Filogenia de pacientes VIH Andalucía Oriental periodo 2014-2016 pacientes *naive*. Se observan cuatro *cluster* marcados en distintos colores con *bootstrap* mayores de 70%. Árbol filogenético efectuado con Máxima Verosimilitud.

6. DISCUSIÓN

La plataforma 454 Roche ha sido adoptada por muchos laboratorios en la investigación VIH, sustituyendo a la secuenciación Sanger como el principal método utilizado, donde UDS es capaz de detectar variantes de baja frecuencia en la población viral y generar una información por paciente de tres a cuatro órdenes de magnitud. Como resultado, también se han beneficiado los estudios moleculares de filogenia. A pesar de la potencia experimental de NGS, estudios a gran escala sobre filogenia podrían tener una limitación bioinformática, debido al tiempo de procesamiento computacional de miles de datos, especialmente en entornos con recursos limitados. Generalmente los estudios filogenéticos VIH [60.61], en concreto los estudios de parentesco o dinámica de la epidemia VIH, tienen como finalidad la asignación de redes sexuales en base a la comparación de una única secuencia global del gen pol de VIH, obtenidos de pacientes individuales.

Ciertos estudios utilizan toda la información obtenida mediante UDS [62], lo que hace que los arboles resultantes no se observen bien y tengan que explicarse con un pie de imagen muy extenso. Esta limitación se intenta solventar con la creación de una única secuencia consenso a partir de las secuencias consenso generadas por un mismo paciente. Algunos estudios tienden a generar estas secuencias mediante comandos informáticos complejos. En este caso la utilización de MESQUITE [63] es intuitiva, de fácil manejo y sin necesidad de comandos, simplificando la obtención de la secuencia consenso. Para su utilización es necesario exportar en formato pfam las secuencias obtenidas en la plataforma GS Junior 454 y posteriormente se selecciona el umbral de corte para la creación de la secuencia consenso. La ambigüedad en las bases nucleotídicas de la secuencia consenso vendrá dada por el umbral de corte. Este umbral permite especificar la frecuencia de detección de una base, es decir, las bases que estén por encima de dicha frecuencia serán incluidas en la consenso. Si en dicha posición hay dos o más bases por encima de dicho umbral se creara una base ambigua. Por lo tanto, conforme aumentamos el umbral de corte en la creación de secuencias consenso disminuye el porcentaje de bases ambiguas, presentando mayor homología con la secuencia Sanger y favoreciendo el análisis filogenético. Previo a la utilización de MESQUITE, se aconseja efectuar un filtrado de las secuencias obtenidas, ya que ha sido ampliamente descrito [64] que la tecnología 454 Roche genera ciertos errores en el proceso de amplificación, pudiendo trasladarse a la secuencia consenso creada.

En este estudio se ha conseguido adecuar los datos producidos UDS a una única secuencia consenso tipo Sanger, para su correcta utilización en estudios filogenéticos. Se obtuvo un incremento de muestras pareadas o diadas [65] Sanger-UDS al ir aumentando el umbral UDS consenso. Finalmente todas las secuencias formaron diadas, con la utilización de secuencias consenso UDS umbral 20%. Una única muestra no formo diada con valor de bootstrap alto, debido a las muchas discrepancias entre las bases nucleotídicas en ambas secuencias. Estas discrepancias fueron debidas a errores en algún tipo de secuenciación.

Una parte importante en los estudios filogenéticos es el proceso de alineación de secuencias, teniendo como objetivo aproximar posiciones homologas en base a la verdadera historia evolutiva de las secuencias [66.67]. Por lo tanto, el éxito de la inferencia filogenética dependerá entre otras medidas de la exactitud de los datos. Ciertas regiones presentan una incertidumbre sustancial por la presencia de bases ambiguas o regiones con indel. En la mayoría de los estudios, estas regiones ambiguas son eliminadas antes de llevar a cabo el análisis. La ambigüedad de alineación puede socavar los métodos de inferencia bioinformáticos basados en la estimación secuencial, evitando la robustez en los análisis filogenéticos y otros parámetros evolutivos, obteniendo como resultado arboles filogenéticos que no se corresponden a lo esperado [68,69]. Hoy en día no existe en el software MEGA algoritmos de procesamiento que tengan en cuenta las bases ambiguas [70]. Por lo tanto, los estudios filogenéticos se ven mermados por la utilización de secuencias con altos niveles de bases ambiguas, siendo necesario el desarrollo de dichos algoritmos.

En cuanto al subtipado de muestras, también se vio afectado en la utilización de distintos umbrales UDS. Estas discrepancias también fueron debidas a la multitud de bases ambiguas generadas en la secuencia consenso UDS umbral 10%, impidiendo así el correcto subtipado.

Los pacientes estudiados fueron mayoritariamente hombres jóvenes (mediana 40 años), que mantienen sexo con hombres (48,5%), nativos españoles (80,5%), que residen en la zona de Granada (53,7%) o Almería (38,3%) y que están infectados por subtipos B (68,2%). De las 198 secuencias utilizadas para generar el árbol filogenético, 32 se agruparon en 4 *clusters* diferentes, con un valor medio de *bootstrap* de 92,54%.

En cambio para el estudio de pacientes en la cárcel se observo que existían cuatro pacientes relacionados, pero al revisar los nombres de las muestras se observo que estos dos *cluster* de dos muestras cada uno eran debido a dos pacientes distintos que tenían muestras procesadas en años distintos.

En resumen, este trabajo presenta datos que demuestran que es posible la simplificación de multitud de secuencias UDS en una única secuencia tipo Sanger. Para el correcto uso de las secuencias generadas en estudios de epidemiología molecular, es necesario efectuar un procesado de las secuencias y utilizar puntos de corte superiores al 20% para obtener la secuencia consenso.

Teniendo en cuenta la importancia de los estudios filogenéticos y de evolución, será necesario la creación e implementación de nuevos algoritmos que tengan en cuenta posiciones de bases ambiguas para el análisis bioinformático.

7. VALORACIÓN ECONÓMICA DEL TRABAJO

Este trabajo no tiene ningún beneficio económico. Respecto al gasto de la elaboración tampoco tiene ninguno asociado, ya que los reactivos utilizados, equipo técnico, fungibles son proporcionados en la rutina asistencial hospitalaria.

8.

CONCLUSIONES

Mediante este trabajo he podido perfeccionar mi manejo en el ámbito de la Biología Molecular efectuando los dos tipos de secuenciación. Además he podido aprender y mejorar programas bioinformáticos utilizados en el máster o nuevos. Hay que decir que este trabajo se ha llevado a cabo solamente por el alumno, ha sido difícil guiarse solo por este camino, pero al mismo tiempo esto ha hecho que pueda aprender más y desenvolverme sin problemas.

Los objetivos planteados fueron altos, pero se ha conseguido el objetivo principal. El factor tiempo para desarrollar los objetivos ha hecho complicada la elaboración de este trabajo, faltando unos días para su elaboración. Los objetivos secundarios también se llevaron a cabo.

El seguimiento de la planificación se llevo a cabo salvo algún pequeño retraso por falta de tiempo. La metodología prevista fue la adecuada. No se ha introducido ningún cambio en la metodología propuesta desde el principio.

Las líneas futuras de trabajo será seguir incorporando pacientes a la base de datos para seguir efectuando trabajos de epidemiologia. Solamente queda pendiente la elaboración detallada de los *clustes* de transmisión en Andalucía Oriental, teniendo datos sociodemográficos y clínicos de los pacientes.

9. GLOSARIO

VIH: Virus de la Inmunodeficiencia Humana.

ADN: Acido desoxirribonucleico.

ARN: Acido ribonucleico.

UDS (*Ultra Deep Sequencing*): Secuenciación masiva.

NGS (*Next Generation Sequencing*): Secuenciación masiva.

PRC: Reacción en Cadena de la Polimerasa.

10.

BIBLIOGRAFÍA

[1] Richard M. Gibson & Christine L. Schmotzer & Miguel E. Quiñones-Mateu. Next-Generation Sequencing to Help Monitor Patients Infected with VIH: Ready for Clinical Use? Curr Infect Dis Rep (2014) 16:401.

[2] Miguel E. Quiñones-Mateu, Santiago Avila, Gustavo Reyes-Teran, and Miguel A. Martinez. Deep Sequencing: Becoming a Critical Tool in Clinical Virology.*J Clin Virol*. 2014 September ; 61(1): 9–19.

[3] Sara Gianella,Wayne Delport,Mary E. Pacold, Jason A. et al. Detection of Minority Resistance during Early VIH-1 Infection: Natural Variation and Spurious Detection rather than Transmission and Evolution of Multiple Viral Variants.JOURNAL OF VIROLOGY, Aug. 2011, p. 8359–8367.

[4] Mary Pacold, Davey Smith, Susan Little, Pok Man Cheng, Parris Jordan et al.Comparison of Methods to Detect VIH Dual Infection.AIDS RESEARCH AND HUMAN RETROVIRUSES. 2010; Volume 26, Number 12.

[5] Art F.Y. Poon, Luke C. Swenson, Winnie W.Y. Dong, Wenjie Deng et al. Phylogenetic Analysis of Population-Based and Deep Sequencing Data to Identify Coevolving Sites in the *nef* Gene of VIH-1.*Mol. Biol. Evol.* 27(4):819–832. 2010.

[6] Binhua Liang1, Ma Luo, Joel Scott-Herridge1, Christina Semeniuk1, Mark Mendoza1. A Comparison of Parallel Pyrosequencing and Sanger Clone-Based Sequencing and Its Impact on the Characterization of the Genetic Diversity of VIH-1.PLoS ONE .October 2011 | Volume 6 | Issue 10.

[7] Osvaldo Zagordi, Rolf Klein, Martin Da¨ umer and Niko Beerenwinkel. Error correction of next-generation sequencing data and reliable estimation of VIH quasispecies.Nucleic Acids Research, 2010, Vol. 38, No. 21.

[8] Sofiane Mohameda,d, Guillaume Penarandaa, Dimitri Gonzalezb, Claire Camusa, Hace`ne Khiria et al.. Comparison of ultra-deep versus Sanger sequencing detection of minority mutations on the VIH-1 drug resistance interpretations after virological failure. AIDS 2014, 28:000–000.

[9] Jay Shendure1 & Hanlee Ji. Next-generation ADN sequencing.nature biotechnology volume 26 number 10 OCTOBER 2008.

[10] Watson SJ, Welkers MRA, Depledge DP, Coulter E, Breuer JM, de Jong MD, Kellam P. 2013 Viral population analysis and minority-variant detection using short read next-generation sequencing. Phil Trans R Soc B 368: 20120205. .February 2, 2016.

[11] Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nature biotechnology. 2012.

[12] Zhang J, Chiodini R, Badr A, Zhang G.. The impact of nextgeneration sequencing on genomics. J. Genet. Genomics 2011; 38:95–109.

[13] Matthew J. Brauer, Mark T. Holder, Laurie A. Dries, Derrick J. Zwickl, Paul O. Lewis and David M. Hillis. Genetic Algorithms and Parallel Processing in Maximum-Likelihood Phylogeny Inference. *Mol. Biol. Evol.* 19(10):1717–1726. 2002.

[14] Ari Löytynoja1, Albert J. Vilella1 and Nick Goldman1. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm.*BIOINFORMATICS. Vol. 28 no. 13 2012, pages 1684–1691*.

[15] Luk K-C, Berg MG, Naccache SN, Kabre B, Federman S, Mbanya D, et al. Utility of Metagenomic Next-Generation Sequencing for Characterization of VIH and Human Pegivirus Diversity. PLoS ONE 2015; 10(11): e0141723.

[16] Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men New York City and California. Morb Mortal Wkly Rep. Jul 3 1981; 30(25):305-308.

[17] Barré-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science. 1983; 220 (4599):868-871.

[18] Ascher MS, Sheppard HW, Winkelstein W Jr, Vittinghoff E. Does drug use cause AIDS? Nature. Mar 11 1993; 362(6416):103-104.

[19] Gallo, R. C., Sarin, P. S., Gelmann, E. P., Robert-Guroff, M.,Richardson et al.. Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). Science 1983; 220(4599): 865-867.

[20] International Committee on Taxonomy of Viruses (ICTV) www.ictvonline.org/ Accedido el 01/05/16.

[21] Gallo, R. C., and Montagnier, L. 2003. The discovery of VIH as the cause of AIDS. N ENGL J MED 2003; 349:24:2283-2285.

[22] Montagnier., L. 1999. Human Immunodeficiency viruses (Retroviridae).En: Encyclopedia of Virology. Second Edition. Volume Two. Allan Cranoff y Robert Webster. Academic Press, San Diego, California, USA.

[23] Van Regenmortel, M. H. V., Fauquet, C.M., Bishop, D.H.L., Carstens, E.B. et al.. Virus Taxonomy. Seventh Report of the International Committee on Taxonomy of Viruses. Academic Press. United States of America.2000

[24] Cohen, M., Hellmann, N., Levy, J., DeCock, K., and Lange, J. The spread, treatment, and prevention of VIH-1: evolution of a global pandemic. The Journal of Clinical Investigation 2008;. Vol 118, Nº 4:1244-1254.

[25] McCutchan, F. Global epidemiology of VIH. Journal of Medical Virology. 2006; 78:S7-S12.

[26] www.who.int/VIH/data/en/ . Accedido el 01/05/16.

[27] Hahn B. H, Shaw G. M, De Cock K. M. & Sharp P. M. AIDS as a zoonosis: scientific and public health implications. Science. 2000;287:607-614.

[28] Beer B. E, Bailes E, Sharp P. M. & Hirsch V. M. Diversity and evolution of primate lentiviruses. In human retroviruses and AIDS: a compilation and analysis of nucleic acid and amino acid secuences (ed. Kuiken C. L, Foley B, Hahn B.), pp. 460-474. Los Alamos, NM: Los Alamos National Laboratory.

[29] Peeters M & Sharp P. M. genetic diversity of VIH-1: the moving target. AIDS. 2000;14:129-140.

[30] Sharp P. M, Bailes E, Chaudhuri R. R, Rodenburg C. M, Santiago M. O. & Hahn B. H. The origins of acquired immune deficiency syndrome viruses: where and when?. Phil. Trans. R. Soc. Lond. B. 2001;356:867-876.

[31] Delgado R. Características virológicas del VIH. Enferm Infecc. Microbiol Clin. 2011;29(1):58-65.

[32] www.VIH.lanl.gov/ Accedido el 01/05/16.

[33] Robinson J. G, Redford K. H. & Bennett E. L. Wildlife harvest in logged tropical forest. Science. 1999;284:595-596.

[34] Nahmias A. J. et al. Evidence for human infection with an HTLV III/LAV-like virus in central Africa, 1959. The lancet. 1986;i:1279-1280.

[35] Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn B. H, Wolinsky S. & Bhattacharya T. Timing the ancestor of the VIH-1 pandemic strains. Science.2000;288:1789-1796.

[36] Hills DM. Origins of VIH. Science. 2000; 288:1757-59.

[37] Korber B, Muldoon M, Theiler J. Timing the ancestor of the VIH-1 pandemic strains. Science 2000; 288:1789-96.

[38]Alcamí, J. (2004). Avances en la inmunopatología de la infección por el VIH. Enferm Infecc Microbiol Clin; 22 (8): 486-496.

[39] www.fundacionio.blogspot.com.es/2013/05/descrubriendo-al-vih-el-genoma.html Accedido el 01/05/16.

[40] Stratov I., DeRose R., Purcell D. F. J., Kent S. J.: Vaccines and vaccine strategies against VIH. Curr. Drug Targets 2004; 05: 71-88.

[41]Birch MR, Learmont JC, Dyer WB, Deacon NJ, Zaunders JJ, Saksena N. An examination of signs of disease progression in survivors of the Sydney Blood Bank Cohort (SBBC). J Clin Virol. Oct 2001; 22(3):263-70.

[42]Dyer WB, Geczy AF, Kent SJ, McIntyre LB, Blasdall SA, Learmont JC. Lymphoproliferative immune function in the Sydney Blood Bank Cohort, infected with natural nef/long terminal repeat mutants, and in other long-term survivors of transfusion-acquired VIH-1 infection. AIDS. Nov 1997; 11(13):1565-74.

[43] Hoffmann C., Kamps S. (2003): VIH Medicine 2003.

[44] Perelson,A.S., Neumann,A.U., Markowitz,M., Leonard,J.M., and Ho,D.D.. VIH-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science 1996; 271, 1582-1586.

[45] María del Carmen Casañas Carrillo. Modelos de interpretación de la resistencia del virus de la inmunodeficiencia humana a los fármacos antirretrovirales. Valoracion de la capacidad predictora de la respuesta virológica. Tesis doctoral. 2008.

[46] Jose Maximiliano Medina Ramirez. Busqueda de respuesta humoral neutralizante en oacientes VIH-1 con niveles indetectables de viremia. Tesis doctoral. 2012

[47] Moore, J. P., S. G. Kitchen, P. Pugach, and J. A. Zack. 2004. The CCR5 and CXCR4 coreceptors--central to understanding the transmission and pathogenesis of human immunodeficiency virus type 1 infection. AIDS Res Hum Retroviruses 20:111-126.

[48] Tersmette, M., de Goede, R. E., Al, B. J., Winkel, I. N., Gruters, R. A., Cuypers, H. T. et al. Differential syncytium-inducing capacity of human immunodeficiency virus isolates: frequent detection of syncytiuminducing isolates in patients with acquired immunodeficiency syndrome (AIDS) and AIDS-related complex. 1988; J Virol 62:2026-2032.

[49] Peeters, M., Vincent, R., Perret, J. L., Lasky, M., Patrel, D., Liegeois, F., et al.. Evidence for differences in MT2 cell tropism according to genetic subtypes of VIH-1: syncytium-inducing variants seem rare among subtype C VIH-1 viruses. J Acquir Immune Defic Syndr Hum Retrovirol 1999; 20:115-121.

[50] Tscherning, C., Alaeus, A., Fredriksson, R., Bjorndal, A., Deng, H., Littman, D. R., et al.. Differences in chemokine coreceptor usage between genetic subtypes of VIH-1. Virology 1998; 241:181-188.

[51] Zhang, L., Y. Huang, T. He, Y. Cao, and D. D. Ho. 1996. VIH-1 subtype and second-receptor use. Nature 383:768.

[52] Abebe, A., Demissie, D., Goudsmith, J., Brouwer, M., Kuiken, CL., Pollakis, G., et al.. VIH subtype C syncytium – and non-syncytium inducing phenotypes and coreceptor usage among Ethiopian patients with AIDS 1999; 13:1305-1311.

[53] Cecilia, D., S. S. Kulkarni, S. P. Tripathy, R. R. Gangakhedkar, R. S. Paranjape, and D. A. Gadkari.. Absence of coreceptor switch with disease progression in human immunodeficiency virus infections in India. Virology 2000; 271:253-258.

[54] Ping, L. H., Nelson, J. A., Hoffman, I. F., Schock, J., Lamers, S. L., Goodman, M., et al.. Characterization of V3 sequence heterogeneity in subtype C human immunodeficiency virus type 1 isolates from Malawi: underrepresentation of X4 variants. J Virol 1999; 73:6271-6281.

[55] Adjorlolo-Johnson, G., De Cock, K. M., Ekpini, E., Vetter, K. M., Sibailly, T., Brattegaard, K., et al. Prospective comparison of mother-to-child transmission of VIH-1 and VIH-2 in Abidjan, Ivory Coast. Jama 1994; 272:462-466.

[56] O'Donovan, D., Ariyoshi, K., Milligan, P., Ota, M., Yamuah, L., Sarge-Njie, R., and Whittle, H.. Maternal plasma viral RNA levels determine marked differences in mother-to-child transmission rates of VIH-1 and VIH-2 in The Gambia. MRC/Gambia Government/University College London Medical School working group on mother-child transmission of VIH. Aids 2000;14:441-448.

[57] Kiwanuka, N., Laeyendecker, O., Quinn, T. C., Wawer, M. J., Shepherd, J., Robb, M., et al. VIH-1 subtypes and differences in heterosexual VIH transmission among VIH-discordant couples in Rakai, Uganda. Aids 2009; 23:2479-2484.

[58] Conroy, S. A., Laeyendecker, O., Redd, A. D., Collinson-Streng, A., Kong, X., Makumbi, F., et al.. Changes in the distribution of VIH type 1 subtypes D and A in Rakai District, Uganda between 1994 and 2002. AIDS Res Hum Retroviruses 2010; 26:1087-1091.

[59] Blackard, J. T., Renjifo, B., Fawzi, W., Hertzmark, E., Msamanga, G., Mwakagile, D., et al.. VIH-1 LTR subtype and perinatal transmission. Virology 2001; 287:261-265.

[60] Ronald J. Lubelchek, MD,Sarah C. Hoehnen, MD et al. Transmission Clustering Among Newly Diagnosed HIV Patients in Chicago, 2008 to 2011:

Using Phylogenetics to Expand Knowledge of Regional HIV Transmission Patterns.J Acquir Immune Defic Syndr 2015; Volume 68, Number 1.

[61] Eduardo Castro-Nallara,Marcos Pérez-Losadab,Gregory F. Burtonc,Keith A. Crandalla. The evolution of HIV: Inferences using phylogenetics.Molecular Phylogenetics and Evolution .Volume 62, Issue 2, February 2012, Pages 777–792.

[62] Susan H. Eshleman, Sarah E. Hudelson, Andrew D. Redd, Lei Wang, Rachel Debes, Ying Q. et al.. .Analysis of Genetic Linkage of HIV From Couples Enrolled in the HIV Prevention Trials Network 052 Trial. Journal of Infectious Diseases Advance Access published November 1, 2011.

[63] Maddison, W. P. and D.R. Maddison. 2015. Mesquite: a modular system for evolutionary analysis. Version 2.75 http://mesquiteproject.org

[64] Wei Shao, Valerie F Boltz, Jonathan E Spindler, Mary F Kearney et al. Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of Low-frequency drug resistance mutations in HIV-1 DNA.Shao et al. Retrovirology 2013, 10:18.

[65] Novitsky V, Bussmann H, Logan A, Moyo S, van Widenfelt E, okui L, et al. Phylogenetic relatedness of circulating HIV-1C variants in Mochudi, Botswana. PLoSOne. 2013;8:e80589.21.

[66] Lutzoni F, Wagner P, Reeb V, Zoller S. .Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology.Syst Biol. 2000 Dec;49(4):628-51.

[67] Andreas D. Baxevanis,B. F. Francis Ouellette . Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Willey 3 ed.November 2004.

[68] BENJAMIN D. REDELINGS AND MARC A. SUCHARD .Joint Bayesian Estimation of Alignment and Phylogeny .Syst. Biol. 54(3):401–418, 2005.

[69] Pasquier, N. Millot, R. Njouom, K. Sandres, M. Cazabat, J. Puel, J. Izopet .HIV-1 subtyping using phylogenetic analysis of *pol* gene sequences.Journal of Virological Methods 94 (2001) 45–54.

[70] Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30:27259.