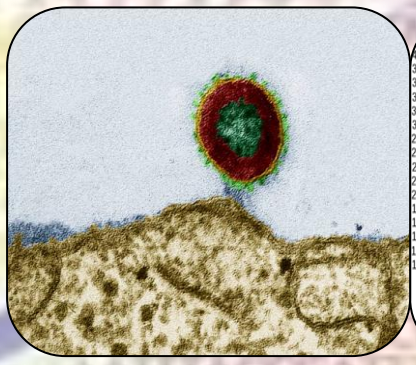
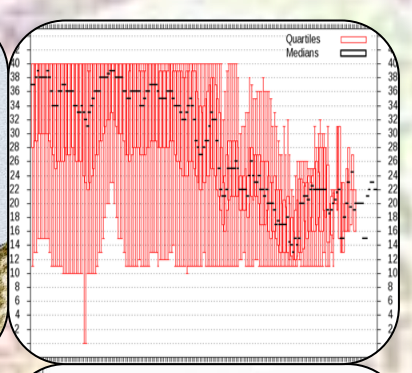
Universidad de Granada

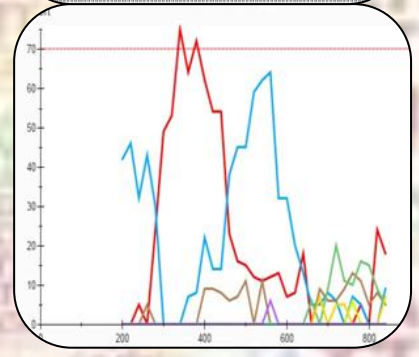
TESIS DOCTORAL

Julio 2017

Resistencias a antirretrovirales: nuevos métodos de detección, nuevas aplicaciones







Director de Tesis

Federico García García

Doctorando

Jose Ángel Fernández-Caballero Rico

Editor: Universidad de Granada. Tesis Doctorales
Autor: José Ángel Fernández-Caballero Rico
ISBN: 978-84-9163-422-5
URI: http://hdl.handle.net/10481/47977

AGRADECIMIENTOS

En primer lugar quiero dar las gracias a Federico García García por haberse arriesgado aquel día que fui a proponerle realizar el TFM en el servicio de microbiología molecular y posteriormente contando conmigo al finalizar para formar parte del equipo de investigación. Gracias por el gran apoyo y esfuerzo que me ha brindado en la realización de esta tesis.

No puedo olvidarme de los pacientes, siempre dispuestos a ayudar y apoyar la investigación. Gracias al Hospital Universitario San Cecilio por sus magnificas instalaciones que facilitan el trabajo y a la Red de SIDA de la que formo parte, por su gran financiación y labor en investigación. También agradecer al Hospital Clinic de Barcelona por su colaboración en mi investigación.

Quería dar las gracias a mi familia, a mis padres por haberme dado una educación ejemplar y haber hecho el esfuerzo para que pudiera estudiar. Sois los mejores padres y sin vosotros no hubiese llegado hasta aquí. A mis hermanos que siempre tengo su apoyo y por los buenos momentos que compartimos juntos, no podría tener unos hermanos mejores. A mi tía Encarni que siempre ha seguido de cerca mis logros y sé que está orgullosa de mi. A mi primo Juan que tanto me cuido de pequeño y para mi es más que un primo. A mis tíos María Jesús y Jose Antonio que les tengo muy presentes siempre por su cariño incondicional. A mis primas Maite y Angelines que me han visto crecer y cada vez que nos juntamos es una alegría. A mi otra familia José, Mari Carmen, Jessica y Yasmina, gracias por vuestro apoyo, por compartir mi felicidad y formar parte de mi vida.

A mi pareja y jardinero favorito Santi por su ayuda, cariño, por su paciencia y por ser tan bueno conmigo, sé que mis logros son los tuyos. Espero tenerte siempre en mi vida. No puedo olvidarme de mi fiel amigo Luis que para mí es un hermano mas, gracias por no soltarme nunca la mano, por estar en lo malo y en lo bueno sin esperar nada a cambio, sabes que te quiero mucho y mi vida no sería lo mismo sin ti. A mi amiga María Jesús por su amistad fiel, su siempre ayuda, escucharme cuando lo necesito y por esos desayunos que disfrutamos. A mis amigas de carrera Sonia, Lamia, Asun, Encarni, Ana María, por cada risa, por compartir tantos buenos momentos A mi fiel amigo Andrés y Stephan que aunque estéis en Paris os siento muy cerca. A Sergio por ser el mejor compañero de piso de los fines de semana, por haberme cuidado en mi gripazo y

por nuestras buenas conversaciones. A Jaime que es mi “yo mejorado”, siempre sacas una sonrisa aunque el día sea malo, se que te sientes orgulloso de mis logros. A Jose Dom, Arturo, Miguel Ángel, Jorge y Juanma por acogerme tan bien y darme vuestra amistad. A mis amigas TEL; Angeles, Mari, Silvia, Gloria y Emilio, gracias por haberme ayudado tanto y seguir a mi lado.

Este doctorado me ha dado la oportunidad de conocer a gente extraordinaria, ofreciéndome apoyo y mucho cariño: Vicente, Cándido, Soraya, Ana Sapiña, Isabel Viciana, Isabel García, Berta, Toñi Moreno, Alba.

Por último pero no menos importante a mi otra familia, formada por el servicio de microbiología. Marta que me enseño tanto en mi llegada al mundo del VIH, por su apoyo incondicional, paciencia, transmitirme sus conocimientos sin esperar nada a cambio y por tantos desayunos en la facultad con los que la mañana tenía otro empezar. Natalia mi fiel confidente y amiga, sin ti no habría logrado todo esto, gracias por enseñarme tantas cosas desde el cariño y la bondad, se que siempre tendré tu apoyo cuando lo necesite. María Dolores que siempre me ha cuidado, preocupándose por mí en cada momento, gracias por las confidencias, por las risas y hacer las mañanas más llevaderas, sin olvidar esas tartas de queso. José Antonio por su sentido común y buenos consejos, por hacerme ver la cara sensata y buena de las cosas. María Ángeles por tu alegría, cantos y refranes que amenizan las mañanas. Carmen Pérez por esos desayunos tan alegres y sus historias tan bien contadas. Ana Belén por escucharme y entenderme, los congresos no hubiesen sido lo mismo sin tu compañía, ahora la siguiente eres tu ¡animo y a por ello que tu puedes! Alex Peña por su cercanía y ser tan buena persona, eres una de esas personas que sin tener tanta relación les tienes un inmenso cariño. A la enfermera Pilar por ser tan buena gente y devolverme siempre una sonrisa. Emi, Carmen Ortiz, Paqui López por vuestra alegría y buenos momentos, María Dolores Valero por ser tan cercana y buena gente. A Vivi por esos buenos ratos en la mañana, en los que podemos hablar de cualquier cosa desde la confianza. Al resto de compañeros del servicio que ocupáis un sitio en mi corazón.

Este doctorado no es solo mío, es parte de todos vosotros. Muchas gracias a cada uno por formar parte de mi vida, se que si me caigo vosotros me levantareis.

ÍNDICE

ABREVIATURAS

ADN: Acido desoxirribonucleico

ARN: Acido ribonucleico

ARNmc: Acido ribonucleico monocatenario

ARNt: Acido ribonucleico de transferencia

CAM-1: Molécula de adhesión celular

CD4: Linfocito CD4

CDD: Dispositivo de carga acoplada

CCR5: Receptor de quimiocinas CCR5

CXCR4: Receptor de quimiocinas CXCR4

DTG: Dolutegravir

E: Acido Glutámico

ELISA: Ensayo inmunoabsorción ligado a enzimas

emPCR: Reacción en cadena de la polimerasa en emulsión

EVG: Elvitegravir

FDA: Federación de alimentos y medicamentos

gp: Glicoproteína

HLA: Antígeno leucocitario humano

ICTV: Comité internacional de taxonomía de virus

IN: Integrasa

Kb: Kilobyte

LTR: Repetición terminal larga

Mg: Magnesio

MHC: Complejo mayor de histocompatibilidad

MID: Identificador multiplex

ML: Máxima verosimilitud

N's: Base desconocida

NF-kB: Factor nuclear potenciador de cadenas ligeras kappa

NGS: *Next Generation Sequencing* ó secuenciación de segunda generación

NJ: *Neighbor joining* ó union de vecinos

nm: Nanómetro

NNRTI: Inhibidor no nucleósidos de la Transcriptasa Reversa

NRTI: Inhibidor nucleósido de la Transcriptasa Reversa

INSTI: Inhibidor de transferencia de la cadena de la Integrasa

N: Asparagina

PBMC: Células mononucleares de sangre periférica

PCR: Reacción en cadena de la polimerasa

PR: Proteasa

Q: Glutamina

RAL: Raltegravir

RSV: Virus del sarcoma de Rous

RT: Transcriptasa Reversa

S: Serina

SIDA: Síndrome de inmunodeficiencia adquirida

SNP: Polimorfismo de un solo nucleotido

SP1: Factor de transcripción SP1

T: Treonina

UDS: *Ultra Deep Sequencing* ó secuenciación ultra profunda

VIH: Virus de la inmunodeficiencia humana

Y: Tirosina

Zn: Zinc

RESUMEN

En 1987 se produjo la entrada del primer fármaco antirretroviral VIH, con el paso de los años estos fármacos han mejorado tanto en efectos secundarios como en eficacia, sin embargo sigue preocupando la aparición de resistencias frente a los distintos fármacos. El estudio de resistencia antirretrovirales en pacientes VIH-1 se realiza mayormente mediante secuenciación poblacional o Sanger del gen *pol*, incluyendo las regiones Transcriptasa Reversa, Proteasa e Integrasa. Una limitación conocida de las pruebas de susceptibilidad a antirretrovirales es la falta de fiabilidad en la detección de variantes minoritarias de VIH-1 en la población del virus infectante. Esta limitación es particularmente importante en pacientes con antecedentes de tratamientos complejos y que han fracasado a una o varias líneas de fármacos antirretrovirales. Como alternativa, el análisis del ADN proviral puede tener interesantes aplicaciones, ya que permite detectar las mutaciones de resistencias archivadas en las poblaciones víricas, incluso tras la supresión del tratamiento. La detección de estas mutaciones de resistencia tiene gran relevancia clínica, como es la planificación de tratamiento antirretroviral en pacientes previamente tratados y con una limitada o nula disponibilidad de informes de resistencias previos.

Durante los últimos seis años, Raltegravir (RAL), el primer inhibidor aprobado de transferencia de hebra de la Integrasa (INSTIs), ha tenido un papel importante en el tratamiento de la infección por VIH, tanto en pacientes *naïves* como en pretratados. Hoy en día, existen otros dos fármacos disponibles de la misma familia (Elvitegravir y Dolutegravir (DTG)), ampliando las opciones de utilización. Elvitegravir es un fármaco de primera generación, compartiendo el perfil de resistencias con RAL, por el contrario Dolutegravir, de segunda generación presenta una elevada potencia antiviral, con excelente perfil de eficacia y seguridad. Aunque con elevada barrera genética a la resistencia, este fármaco puede verse afectado por la combinación de algunas mutaciones de resistencia que se acumulan en el fracaso a RAL. DTG presenta una larga vida media plasmática, lo que le permite dosificarlo una vez al día. Esta dosificación es adecuada para los pacientes sin resistencia preexistente a INSTIs, por el contrario, en pacientes con mutaciones de resistencia documentadas en la Integrasa se debe administrar dos veces al día.

En el capítulo 1, mostramos los resultados del estudio de 7 pacientes con historia de fracaso previo a un régimen que incluyera RAL y que posteriormente al fracaso

consiguieran supresión virológica. La población a estudiar fue en su totalidad subtipo B, con una mediana de edad de 55 años. La mediana de carga viral y recuento de CD4 fue de 1,3 $\log_{10}$ cp/ml y de 765,5 cel/ml. La mediana (IQR) de tiempo entre el momento del fracaso y el estudio en ADN proviral fue de 48 meses (29-53). La mutación primaria N155H fue la más prevalente en el momento del fracaso, manteniéndose en el ADN proviral, al igual que las mutaciones accesorias, excepto en un paciente en el que hubo que recurrir a la secuenciación masiva, detectándose N155H en una proporción del 9,77%. En conclusión, a pesar de las limitaciones de nuestro estudio, que no deja de ser un estudio piloto que deberá ser corroborado en estudios posteriores, el estudio de las resistencias en la IN utilizando el ADN proviral de una muestra actual es un fiel reflejo de lo que ocurrió en el fracaso. De confirmarse estos resultados en otras series, esta herramienta podrá ser de gran utilidad para la toma de decisiones en la simplificación a DTG de este tipo de pacientes.

A finales del año 2013 se incorporo la tecnología de secuenciación masiva (NGS) en algunos centros, mejorando la detección de variantes minoritarias. Existen ciertas diferencias entre la secuenciación tradicional Sanger y la secuenciación masiva, como es el umbral de detección de mutaciones: alrededor de 20% para Sanger y del <1% para masiva, además del volumen de secuencias obtenidos: una única secuencia Sanger frente a las miles de secuencias que ofrece la secuenciación masiva. El gran volumen de datos presentes en secuenciación masiva dificulta el trabajo bioinformático, necesitando potentes ordenadores, por lo que la simplificación hacia una única secuencia en ciertos casos, como estudios filogenéticos podría mejorar el flujo de trabajo. Además la convivencia entre tecnología Sanger y masiva hace necesaria una estandarización o unificación.

En el capítulo 2 mostramos un método que permite simplificar los datos NGS a una secuencia consenso única, comparable a la secuenciación Sanger. Para el estudio se utilizaron secuencias de 62 pacientes *naïve* del periodo 2014-2015, nuevos diagnósticos VIH-1 referidos para estudios de resistencias a antirretrovirales en los servicios de Microbiología Clínica de el Hospital San Cecilio de Granada y del Hospital Clinic de Barcelona. Las secuencias tipo Sanger se obtuvieron utilizando el kit Trugene HIV-1. Para NGS se partió del mismo ARN, utilizando el kit GS V Type HIV-1 para 454 GS Junior. Las secuencias consenso NGS se generaron mediante el software Mesquite,

seleccionando umbrales de corte del 10%, 15% y 20%. A continuación se construyeron arboles filogenéticos con Mega, con las distintas secuencias consenso NGS y sus respectivas Sanger, teniendo en cuenta valores de *bootstrap* mayores de 70% para definir una relación entre secuencias. Utilizando un punto de corte de 10% en las secuencias NGS, sólo en 17/62 pacientes las secuencias pareadas NGS-Sanger presentaron valores de *bootstrap* >70%, con una mediana (IQR) de los valores de *bootstrap* de 88% (83,5-95,5). Aumentando el umbral consenso NGS al 15%, estos valores ascienden hasta 36/62, relacionando las dos secuencias con un *bootstrap* >70%, mediana (IQR) de valores *bootstrap* de 94% (85.5-98). Por último, al utilizar un umbral de 20%, en 61/62 casos las secuencias pareadas NGS-Sanger se relacionan con *bootstrap* >70%, con una mediana (IQR) de los valores de *bootstrap* de 99% (98-100). En este último caso, en una única muestra no se relaciono la secuencia NGS con su Sanger debido a la multitud de diferencias entre algunas bases nucleotídicas entre ambas secuencias. Al 10%, las diferencias NGS-Sanger alcanzaron diferencias para influir en el subtipado en 3 pacientes, cambiando en un paciente desde subtipo B (NGS) a CRF01_AE (Sanger) y en dos pacientes desde subtipo B (NGS) a CRF02_AG (Sanger) respectivamente. Sin embargo, estas diferencias no se observaron al utilizar las secuencias consenso NGS al umbral 20%. Por lo tanto, presentamos una metodología que permite generar secuencias consenso que son representativas de la secuencia Sanger, para su uso en estudios de epidemiología molecular, siendo necesario efectuar un procesado de las secuencias y utilizar puntos de corte de al menos el 20%.

Finalmente, varias investigaciones han estudiado la precisión de la tecnología NGS basada en pirosecuenciación, determinando la principal fuente de errores las regiones homopoliméricas, siendo mayoritarias inserciones y deleciones (*indel*). Además su curva de secuenciación presenta una disminución de la calidad al final del proceso. Por otra parte, la secuenciación VIH depende de la transcripción inversa y de la utilización de distintas PCR con el fin de aumentar la cantidad de genoma vírico, contribuyendo al aumento de los posibles errores en la secuenciación genómica. Estos artefactos producidos en PCR son bien conocidos, pudiendo minimizarlos mediante la optimización en las condiciones de amplificación y el uso de polimerasas de alta fidelidad. De otra forma, estos posibles errores pueden provocar un impacto significativo en diversos análisis, tales como montaje de secuencias, identificación de polimorfismos, identificación de subtipado vírico, estudios de mutaciones de resistencia

y estudios de expresión génica. La mayoría de programas disponibles en las distintas plataformas proporcionan una tubería de control de calidad en el proceso de secuenciación, con el fin de filtrar la salida de secuencias, aun así, permanecen varios artefactos en el conjunto de datos. Por lo tanto, es aconsejable llevar a cabo un control de calidad final a nivel usuario, mediante filtrado de secuencias de alta calidad. Para ello se hizo una revisión sistemática sobre la eliminación de posibles errores en la plataforma 454 GS Junior. Identificamos siete estudios bajo los criterios de búsqueda, tres estudios de modificación de protocolo de amplificación y cuatro sobre programas bioinformáticos. Los estudios de modificación de protocolo en la PCR demostraron una disminución en el porcentaje de recombinación de hasta dos órdenes de magnitud para una PCR optimizada. Mediante los programas bioinformáticos se consiguió disminuir también los errores, el porcentaje de tasa de error disminuyo en un orden de magnitud, reducción de 93-98% de *indel* y una sensibilidad y especificidad en *SNP variant calling* cercanos a 1. La aplicación de estos métodos de corrección permitirá sin duda alguna ayudar a disminuir los posibles errores, asegurando las mutaciones observadas mediante NGS para el estudio de resistencias en pacientes VIH-1. Es necesaria la implantación de un flujo de trabajo que permita la eliminación de posibles errores.

INTRODUCCIÓN GENERAL

1. EL VIRUS DE LA INMUNODEFICIENCIA HUMANA

1.1. Epidemiología 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 [1]. El virus VIH en sí no fue identificado hasta dos años después, cuando los franceses Barré-Sinoussi y Montagnier consiguieron aislarlo [2]; durante ese período, otras causas fueron consideradas, incluyendo factores como el estilo de vida, el abuso crónico de drogas y otros agentes infecciosos [3]. 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 [4]. En 1986, el Comité Internacional de Taxonomía de Virus (ICTV) aceptó el nombre definitivo de VIH para este nuevo agente [5].

Desde estos primeros hallazgos hasta la actualidad han transcurrido tres décadas. Hoy sabemos que el VIH 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 [6]. También se sabe que el virus se transmite por contacto sexual, 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 [7]. 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) [7].

La epidemia de VIH/SIDA constituye en la actualidad uno de los más graves problemas de salud pública después de haberse cobrado más de 35 millones de vidas, 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 2015 más de 70 millones de personas habían sido infectadas por el virus, actualmente unas 36,7 millones de personas viven infectadas, e incluyen 1,8 millones de niños menores de 15 años, estimándose en 2,1 millones las nuevas infecciones por VIH y en 1,1 millones las defunciones por SIDA en ese año [8] (Figura 1).

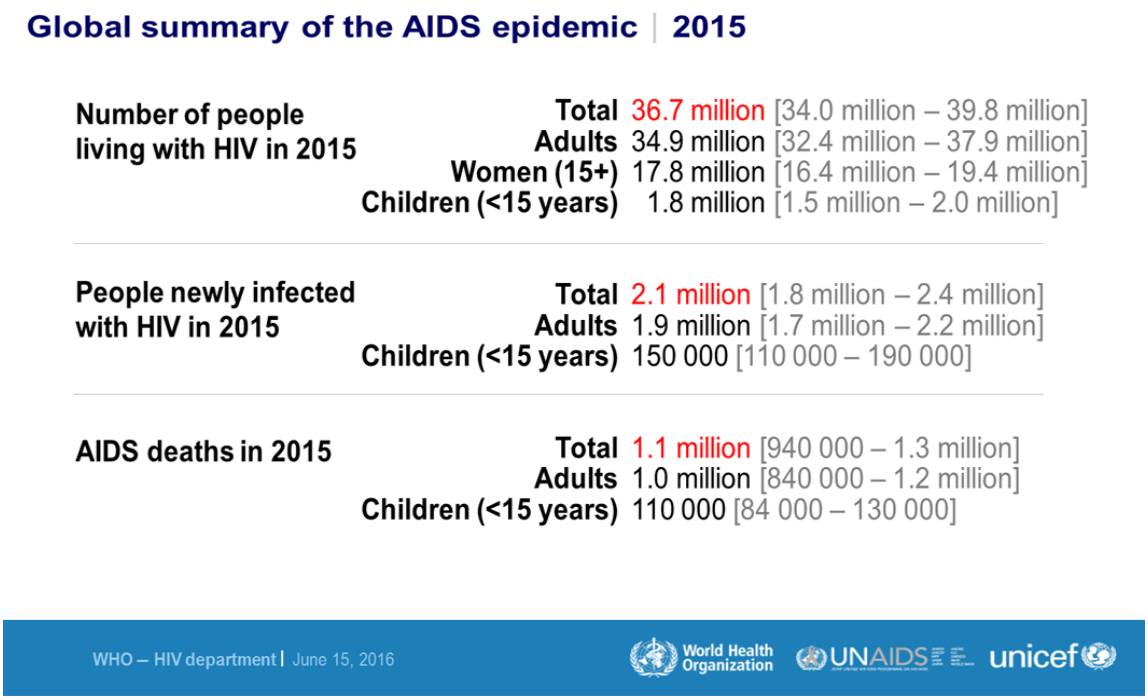


Figura 1. Estadística VIH del año 2015 [8].

2. CARACTERÍSTICAS GENERALES VIH-1

2.1. Estructura VIH-1

La partícula viral tiene forma esférica icosaédrica y mide 80-120 nm de diámetro (Figura 2), 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-1 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 (RT), la Integrasa (IN), la Proteasa (PR) y las proteínas reguladoras y accesorias) [9].

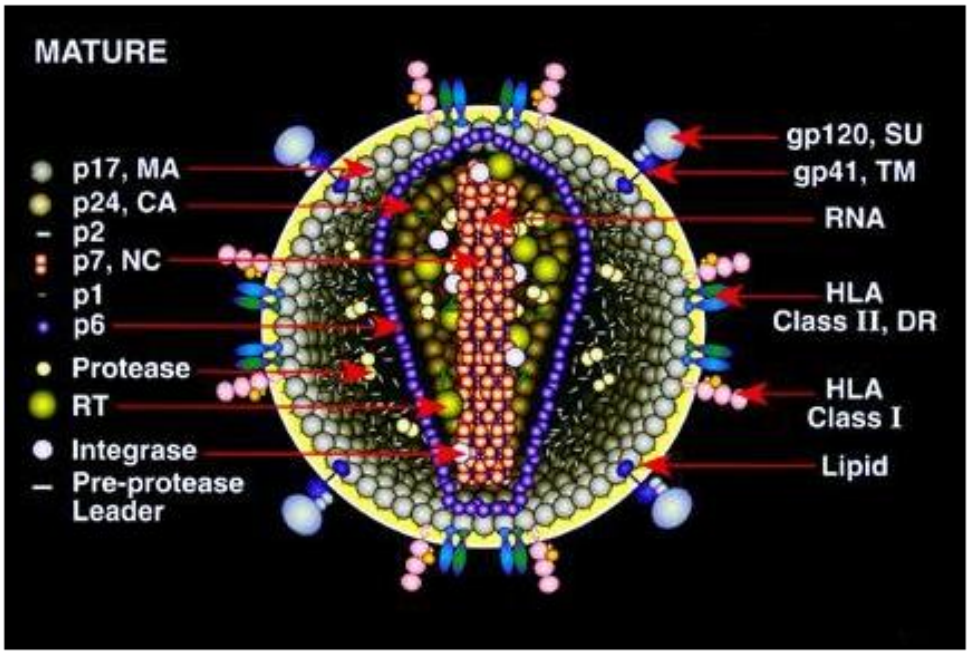


Figura 2. Estructura partícula vírica VIH [10].

El genoma del virus es un ARN de cadena única formado por dos hebras idénticas de 9,8 Kb de polaridad positiva. El virón emplea la enzima RT para replicarse, dando lugar al ADN proviral que se integra en el genoma de la célula huésped gracias a la enzima IN. El genoma viral VIH-1 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 RT, endonucleasa y PR virales necesarias para la replicación del VIH-1. 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) [10].

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 [11].

3. CICLO BIOLÓGICO VIH-1

La infección de las células por el VIH-1 comienza cuando la glucoproteína de la cubierta de una partícula vírica se une a CD4 y a un correceptor de quimiocinas; CCR5 y/o CXCR4 [12]. El gen env es un complejo formado por una subunidad transmembana (gp41) y una subunidad externa (gp120) asociada no covalentemente, estas subunidades se producen por escisión proteolítica de un precursor (gp160). Este complejo media un proceso en múltiples pasos de fusión de la cubierta del virión con la membrana de la célula diana (Figura 3). El primer paso de este proceso es la unión de las subunidades gp120 a las moléculas de CD4, lo que induce un cambio conformacional que favorece la unión secundaria de de gp120 a un receptor de quimiocinas que actúa como correceptor. Esta unión al correceptor induce un cambio conformacional de gp41, lo que expone una región hidrófoba, denominada péptido de fusión, que se inserta en la membrana celular y permite que la membrana del virus se fusione con la membrana de la célula diana [12].

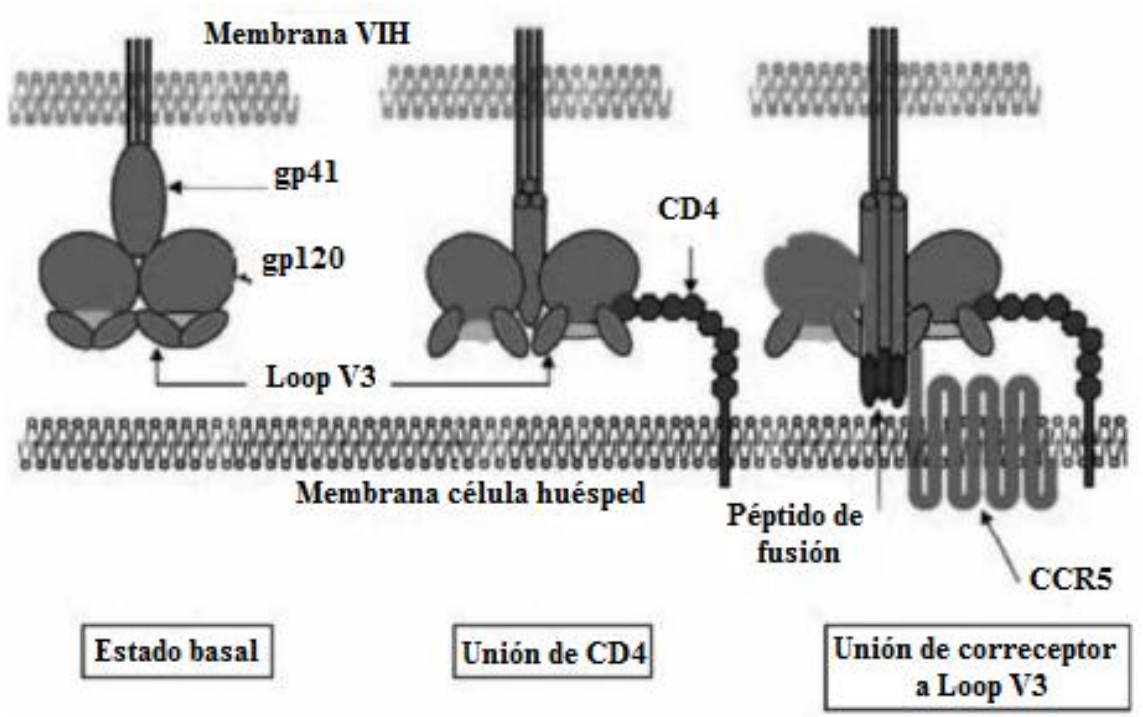


Figura 3. Interacción de gp120 con los receptores celulares y cambio de conformación [13].

Una vez que el virión del VIH-1 entra en la célula, las enzimas del complejo de la nucleoproteína se activan y comienzan el ciclo reproductor del virus. El núcleo nucleoproteínico del virus se rompe, el genoma de ARN del VIH-1 se transcribe en forma de ADN bicatenario por la RT vírica, y el ADN del virus entra en el núcleo. Al igual que la RT, la IN vírica también entra en el núcleo y cataliza la integración del ADN vírico en el genoma de la célula huésped (en el siguiente punto se verá detenidamente el mecanismo de integración). El ADN del VIH-1 integrado se denomina provirus, se ha visto que el provirus puede no transcribir durante meses o años, con una producción escasa o nula de nuevas proteínas séricas o viriones, y de esta forma la infección por el VIH-1 de una célula individual puede ser latente [13].

La transcripción de los genes del provirus (ADN integrado) es regulada por los LTR en dirección 5´ a los genes estructurales del virus, mientras que las citocinas u otros estímulos fisiológicos de los linfocitos T y/o los macrófagos potencian la transcripción de los genes del virus. Los LTR contienen secuencias señal; la secuencia promotora TATA, NF-κB y SP1 (sitios de unión para dos factores de transcripción de las células del huésped). El inicio de la transcripción genética del VIH en los linfocitos T se asocia con la activación de éstos por antígenos o citocinas.

La síntesis de partículas víricas infecciosas maduras comienza después de la producción de transcritos de ARN vírico completos y de que los genes víricos se hayan expresado como proteínas [14]. La expresión génica del VIH-1 se puede dividir en una fase temprana, durante la que se expresan los genes reguladores, y una fase tardía, durante la cual se expresan los genes estructurales y se empaquetan los genomas víricos completos. En lo que respecta a la expresión de los genes tempranos: la transcripción se inicia después de que los factores de transcripción celulares se unan al promotor y a la secuencia potenciadota en la región U3 situada corriente arriba de la secuencia LTR. Durante la fase temprana las proteínas que son producidas son; *nef*, *tat* y *rev*. La proteina *nef* estimula la replicación VIH-1, se observó que en las células infectadas alteraba el tráfico de los endosomas en la célula, reduciendo la expresión del receptor CD4 y de los MHC de clase I y II, por lo que protege al VIH de la respuesta del sistema inmunológico. También se ha visto que la proteína *tat* juega un papel muy importante potenciando la transcripción. Una señal de localización nuclear dirige la proteína *tat* al núcleo, donde se une a la secuencia TAR (elemento de respuesta a la transactivación) en

el extremo 5´ del transcrito naciente. Además de tat, se unen proteínas celulares a la secuencia TAR, entre las cuales se encuentra una proteína quinasa, la cual fosforila algunos componentes del complejo ARN polimerasa. De manera que la proteína *tat* funciona como un factor de transcripción, pero inusual, porque se une al ARN y no al ADN. La otra proteína de expresión temprana es *rev*, la acumulación de esta en el núcleo produce el cambio de la síntesis de proteínas tempranas a proteínas tardías mediante su unión al elemento de respuesta a *rev* (RRE) en el ARN del virus. RRE está presente en los transcritos sin reorganizar y en los que lo han sido solo una vez, pero no en los reorganizados múltiples veces. Por lo tanto, los genes de expresión tardía son traducidos a partir de de transcritos del tamaño del genoma o reordenados una sola vez, pero estos ARNs no son exportados del núcleo hasta que tienen unidas múltiples copias de la proteína Rev. En el caso de las proteínas tardías, *gag* y *gag-pol* son traducidas a partir de transcritos sin reordenar, siendo traducido *gag-pol* cuando se produce un desplazamiento del marco de lectura en los ribosomas.

El resto de proteínas del virus (*vif*, *vpr*, *vpu* y *env*) son traducidas desde transcritos reordenados una sola vez. *vpu* y *env* son traducidas en el retículo endoplasmático rugoso desde un transcrito bicistrónico, seguidamente *env* es glicosilado, formando Trímeros de *env*, siendo posteriormente cortados para formar las proteínas de cubierta gp41 y gp120; esta escisión es llevada a cabo por una proteasa celular localizada en el aparato de Golgi. Por su parte, *vpu* es una proteína asociada a la membrana y es necesaria para una gemación eficiente de los viriones [15].

Algunos retrovirus forman partículas inmaduras en el citoplasma que luego son transportadas a la membrana plasmática, pero la mayoría de retrovirus ensamblan sus componentes en la superficie interna de la membrana plasmática. El grupo amino terminal de las proteínas *gag-pol* se anclan a la membrana plasmática, además los dominios de la nucleocápside de *gag* y *gag-pol* unen las poliproteínas al ARN del virus mediando la formación del dímero genómico. Las proteínas se unen primero a la señal de empaquetado cercana al extremo 5´ de cada molécula de ARN, y un ARNt se une a la secuencia PBS también situada en el mismo extremo. A continuación, el ARN es envuelto en varias copias de la proteína *gag* y unas pocas de *gag-pol*. De esta forma, el virión inmaduro adquiere su envuelta mediante gemación de la superficie celular. Durante o después del proceso de gemación las poliproteínas *gag* y *gag-pol* son

escindidas por la proteasa viral, de esta forma, los productos de la escisión de *gag* forman la matriz, la cápside y los componentes proteicos de la nucleocápside, mientras que los productos de la escisión de Pol son las enzimas del virión (RT, IN y PR) [15] (Figura 4).

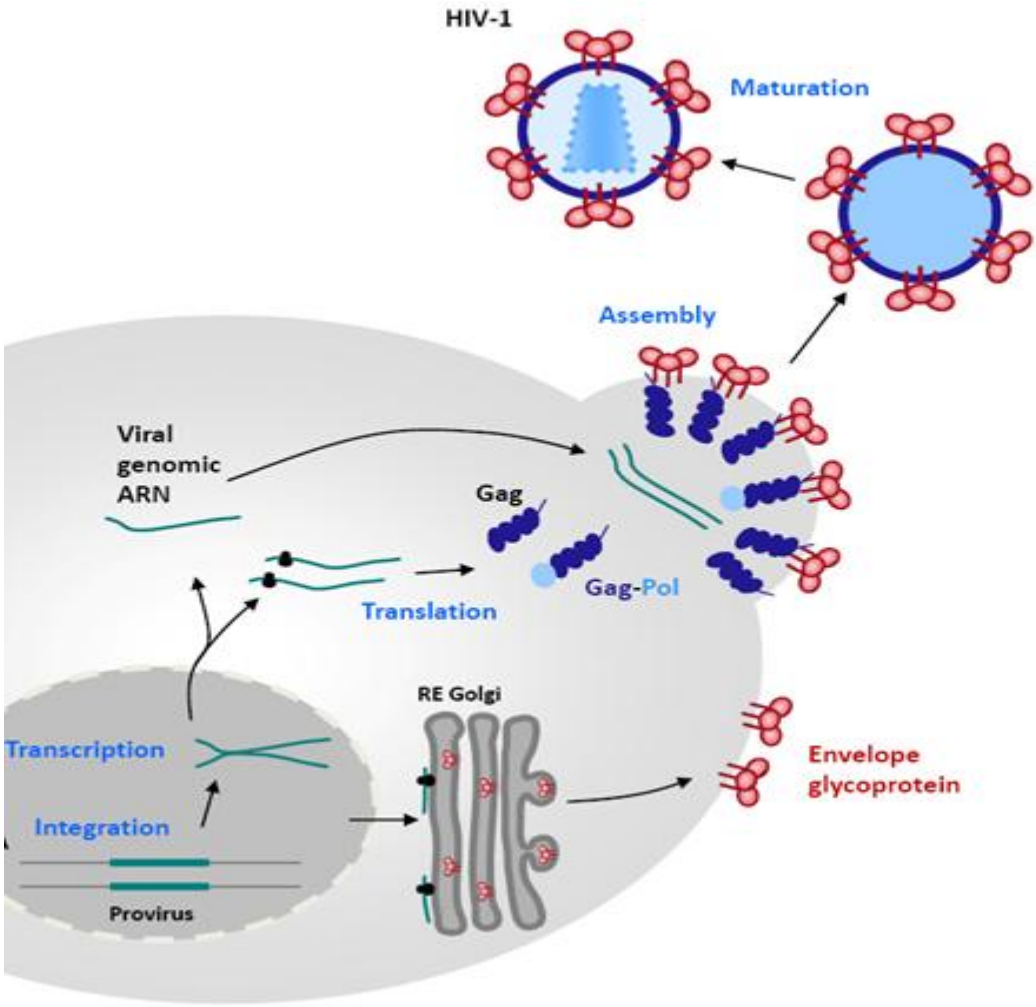


Figura 4. Los últimos pasos del ciclo de replicación del VIH corresponden con la síntesis, el ensamblaje de las proteínas virales con el genoma del virus y la producción de las partículas virales infectivas maduras [15].

4. INTEGRASA

La importancia del proceso de Integración se reconoció con el descubrimiento del virus del Sarcoma de Rous (RSV), identificándose genoma homologo de dicho provirus en el ADN de la célula huésped [16]. Posteriormente se demostró que las mutaciones selectivas en el gen IN detenían la integración viral y su consiguiente reproducción viral [17]. Estos hechos hicieron que la IN fuera una diana terapéutica en los antirretrovirales.

La IN viral es una proteína de 288 aminoácidos (32 kDa) codificada hacia el final del gen *pol*. Es producida como parte del polipéptido precursor *gag-pol*, de donde es escindida por la enzima proteasa [18]. La proteína posee tres dominios independientes: A) El dominio N-terminal (aminoácidos del 1 al 49) que porta el dominio HHCC análogo a los dedos de zinc (de hecho une Zn^{2+}), lo que posiblemente favorece la multimerización, un proceso clave en la integración [19]. B) El dominio central o dominio catalítico (aminoácidos del 50 al 212). Toda actividad IN requiere de la presencia de un cofactor catiónico (el Mg2+) que es unido en el centro catalítico. C) El dominio C-terminal (aminoácidos del 213 al 288) que une ADN de forma inespecífica, de manera que su función fundamental es estabilizar todo el complejo (enzima-ADN) [18].

4.1. Proceso de integración

Para la integración covalente del ADN proviral en el cromosoma del huésped son necesarias dos reacciones. En primer lugar la IN se une a la secuencia corta LTR y cataliza la escisión de nucleótidos conocida como procesamiento en 3´, donde un dinucleótido es eliminado de cada extremo del ADN vírico. El ADN resultante es entonces utilizado como sustrato para la integración covalente del ADN del virus en el interior del genoma de la célula infectada. Esta segunda reacción ocurre simultáneamente en ambos extremos de la molécula de ADN del virus [18] (Figura 5).

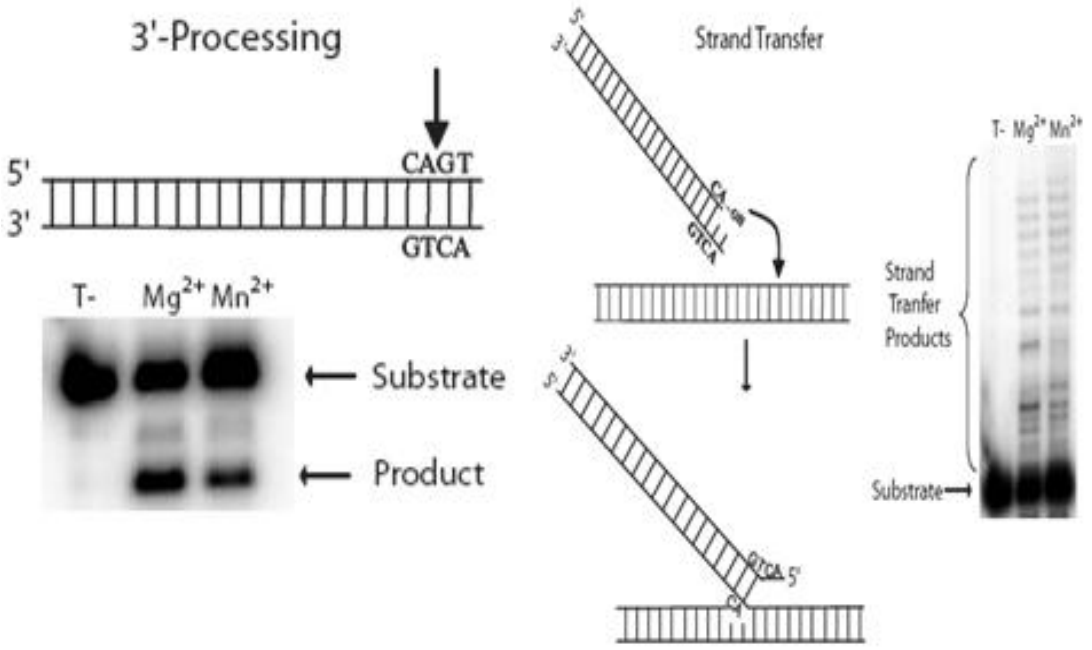


Figura 5. Actividad catalítica de IN: se puede observar el procesado 3' y la reacción de transferencia de hebra [18].

Estas dos reacciones ocurren de una manera secuencial (*in vivo*), y son energéticamente independientes. En ambos casos se trata de una reacción de trans-esterificación en un único paso que consiste en la rotura del enlace fosfodiester por un ataque nucleofílico [18]. Numerosas líneas de evidencia soportan que la integración no ocurre de forma aleatoria sino que esta ocurre de forma preferente dentro de unidades de transcripción [20]. La integración está dirigida fundamentalmente por las interacciones entre el complejo pre-integración (ADN del virus unido a la IN) y la cromatina de la célula del huésped. Desde el punto de vista de la secuencia de ADN, recientemente se ha demostrado que la integración ocurre de forma preferente entre secuencias simétricas [21].

5. MECANISMOS DE RESISTENCIA A LOS FARMACOS ANTIRRETROVIRALES

Hay más de 30 fármacos antirretrovirales aprobados repartidos en 6 clases diferentes, según la diana terapéutica: 1) inhibidores de RT análogos de nucleósidos (NRTIs), 2) inhibidores de RT no análogos de nucleósidos (NNRTIs), 3) Inhibidores PR, 4) Inhibidores IN, 5) Inhibidores de fusión y 6) Inhibidores de entrada. Pero no todo está ganado con el amplio arsenal farmacológico, para una eficacia absoluta se requiere adherencia, tolerabilidad y que no tengan interacciones con otros medicamentos.

Cada vez más pacientes infectados por el VIH tienen acceso a los tratamientos antirretrovirales combinados (generalmente tres fármacos). Estos tratamientos generalmente funcionan bien, manteniendo al virus suprimido y sano al paciente. Sin embargo, el tratamiento solo funciona si el virus no es resistente a los fármacos utilizados. Hoy en día, algunos pacientes son tratados durante años sin tener ningún problema de resistencias, mientras que para otros, las resistencias siguen siendo una amenaza para su salud [22].

Con la aparición de nuevos regímenes mejor tolerados y dosificaciones de una única toma, la adherencia a mejorado y los fracasos han disminuido [22]. También hay factores farmacológicos que influyen en el desarrollo de resistencias. En general, los fármacos antirretrovirales son bien absorbidos y generan altos niveles del medicamento capaces de inhibir la replicación del VIH.

5.1. Fundamentos biológicos de la creación de resistencias

La creación de resistencias aparece como consecuencia de la replicación residual de cuasiespecies víricas mutantes que han podido perdurar debido a regímenes supresivos incompletos [23].

El VIH-1 tiene una tasa de mutación muy alta, acumulando casi una mutación por ciclo

de replicación [24]. Aunque los individuos normalmente están infectados por uno o pocos clones originalmente [25], una cantidad aproximada de 10^{10} viriones son producidos diariamente en individuos sin tratar, lo que resulta en innumerables variantes del virus, conocidas comúnmente como cuasiespecies [26]. La complejidad de las cuasiespecies del VIH-1 se ve incrementada por la alta tasa de recombinación que ocurre cuando más de una variante del virus infecta la misma célula [27]. La capacidad de crear nuevas variantes rápidamente permite al VIH-1 evadir al sistema inmunitario y fomentar el desarrollo de resistencias ante los antirretrovirales. Las mutaciones de resistencia se pueden adquirir mediante presión selectiva (resistencia adquirida) o transmitirse de persona a persona (resistencia transmitida) [28]. Entre el 7 y el 17% de los pacientes de reciente infección portan al menos una mutación mayor de resistencia [22].

Para algunos agentes antirretrovirales son necesarias múltiples mutaciones de resistencia para causar un descenso en la susceptibilidad del virus, sin embargo otros requieren tan solo una. El número de mutaciones necesarias para la adquisición de resistencia, y lo fácil o frecuente que esta mutación la desarrolla, contribuyen a la barrera genética (cuantas más mutaciones sean necesarias para el desarrollo de resistencias más alta es la barrera genética) de un agente antirretroviral. Las mutaciones de resistencia se pueden dividir entre mutaciones primarias, las cuales disminuyen la susceptibilidad del virus al agente antirretroviral, o accesorias, que mejoran el *fitness* del virus y también disminuyen su susceptibilidad. En la figura 6 se ilustra la barrera genética y la potencia de algunos antirretrovirales representativos.

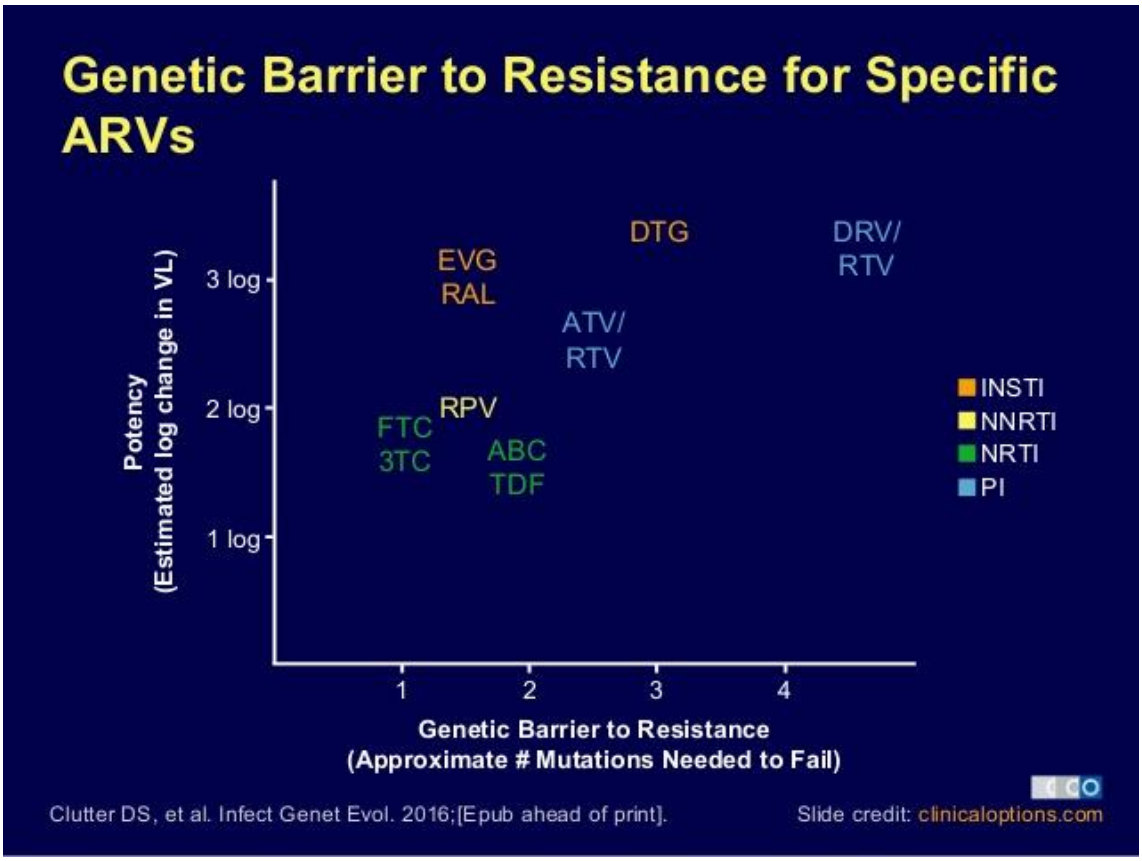


Figura 6. Esta grafica muestra las barreras genéticas relativas y la potencia de los retrovirales [29].

5.2. Inhibidores de transferencia de hebra de integración

Todos los inhibidores de IN clínicamente disponibles se dirigen específicamente a la reacción de transferencia de hebra de integración, y por lo tanto se clasifican como INSTIs. Tres INSTIs han sido aprobados por el comité para que puedan ser prescritos a los pacientes: Raltegravir (RAL), Elvitegravir (EVG) y Dolutegravir (DTG). INSTIs se unen dentro del sitio activo de IN y actúan desplazando el residuo terminal desoxiadenilato del complejo IN-vDNA [30]. INSTIs contiene un farmacóforo de unión a metal para secuestrar los cofactores de metal del sitio activo y un grupo hidrofóbico para ayudar en la interacción vDNA con IN [30]. El primer agente inhibidor de la IN fue RAL (Isentres®), el cual fue desarrollado a partir de los estudios de Hazuda *et al*

[31], siendo aprobado en el otoño de 2007. Sus valores de IC_{50} están en el rango nanomolar tanto *in vitro* como *in vivo* con un alto índice terapéutico [32]. Por su parte, DTG fue el último INSTIs aprobado por la FDA en agosto de 2013.

5.3. Resistencias ante los inhibidores de la Integrasa viral

Los virus resistentes a INSTIs se presentan con mayor frecuencia a través de una alteración primaria de los aminoácidos que influyen en la coordinación de los cofactores metálicos esenciales que se unen dentro del sitio activo (N155 y Q148) [30]. Independientemente de las mutaciones de resistencia primarias y secundarias, la resistencia cruzada entre INSTIs se ha estudiado [33,34]. RAL y EVG seleccionan mutaciones de resistencia similares [35]. Se ha pensado en el uso secuencial de estos fármacos ante el fracaso, pero es improbable que RAL y EVG puedan ser usados de este modo ya que poseen perfiles de resistencia muy similares (Tabla 1) [36]. Las resistencias clínicas a RAL están asociadas con 3 mutaciones genéticas primarias (en la Tabla 1 se pueden observar las distintas sustituciones): Y143, Q148 o N115. Para EVG las resistencias se asocian a las mutaciones: Q148 o N115, así como a T66, E92, T97 o S147 [37-40]

Cons	66 T	92 E	143 Y	147 S	148 Q	155 N
RAL	A	Q	RCH	G	HRK	H
EVG	IAK	Q		G	HRK	H
DTG		Q			HRK	

Tabla 1. Resumen mutaciones que crean resistencias a INSTIs. Las mutaciones están representadas por la posición numérica y el cambio de aminoácido [36].

DTG al ser un INSTIs de segunda generación, tiene una barrera genética mayor que RAL y EVG [40]. Se ha propuesto que el perfil característico de resistencia de DTG se

debe a la mayor cinética de disociación de DTG (71h) para el complejo IN-vDNA, en comparación con RAL (8,8h) y EVG (2,7h) [41]. Varias mutaciones de resistencia a RAL y EVG muestran mutaciones de bajo nivel a DTG, aun así DTG se puede utilizar eficazmente para tratar a los pacientes que han fracasado con regímenes RAL/EVG [42].

El uso de DTG para rescatar pacientes que han desarrollado resistencia a RAL ha sido estudiado y documentado [43]. En la mayoría de los casos, aparentemente se puede obtener algún grado de beneficio al utilizar DTG para tratar a individuos que ya hayan desarrollado resistencia ante RAL o EVG. El peligro de retrasar el uso de DTG es que un número significativo de individuos, los cuales hayan desarrollado resistencia a RAL y/o EVG pueden en ese momento haber perdido la capacidad de responder total y eficazmente a un tratamiento con DTG [44]. Por ejemplo, los resultados del estudio VIKING ponen de manifiesto los problemas de durabilidad y respuesta de un régimen de segunda línea basado en DTG una vez que ya hay mutaciones relevantes contra RAL y EVG [43].

DTG ha demostrado una gran eficacia conteniendo a virus portadores de las mutaciones Y143 o N155, y una respuesta mucho más limitada ante los virus con la mutación Q148 junto con otras mutaciones secundarias adicionales [43]. La actividad antirretroviral de DTG contra virus con una sola mutación permanece comparable a su actividad frente al tipo salvaje del virus [45]. Esta actividad ante mutantes con una sola de las mutaciones se corresponde con la mayor barrera genética para el desarrollo de resistencias (en comparación con RAL y EVG) y sugiere que para adquirir resistencia el virus necesitaría más de una mutación [40,43].

6. 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 inmune del organismo [26].

6.1. El concepto de cuasiespecie

Como hemos visto anteriormente, la población VIH sufre variaciones genéticas en el organismo a medida que se va replicando, debido a la incorporación de nucleótidos erróneos, lo que hace que existan virus distintos pero relacionados genéticamente entre sí. Estos virus con pequeñas variaciones son denominados cuasiespecies [46]. Estas poblaciones siguen un sistema evolutivo basado en Darwin, por lo que las variantes con mayor *fitness* tendrán mayor replicación [47]. Por lo tanto, en la población vírica tendremos unas cuasiespecies más homogéneas, que en conjunto proporcionaran una secuencia maestra, siendo esta secuencia el genoma vírico predominante que comprende las mutaciones más predominantes [46].

Las poblaciones víricas están expuestas a presiones selectivas en el organismo, como la medicación, ejerciendo como cuellos de botella, produciendo una selección de las cuasiespecies víricas, que tendran mayor capacidad para replicarse [48] (Figura 7).

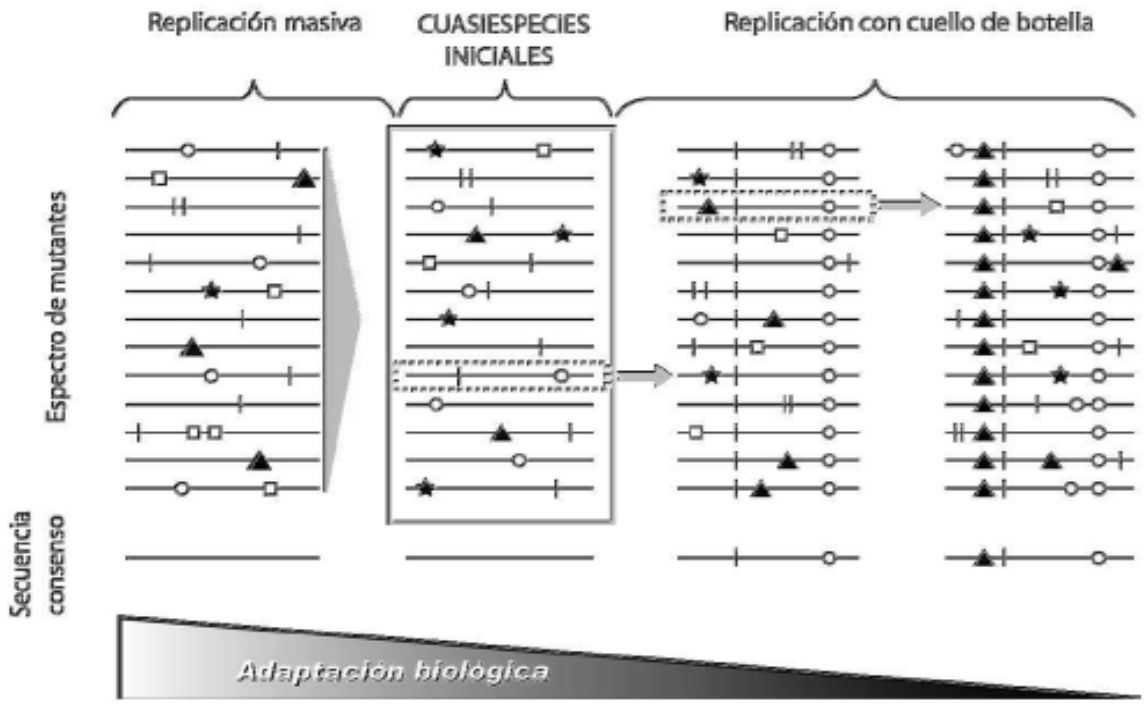
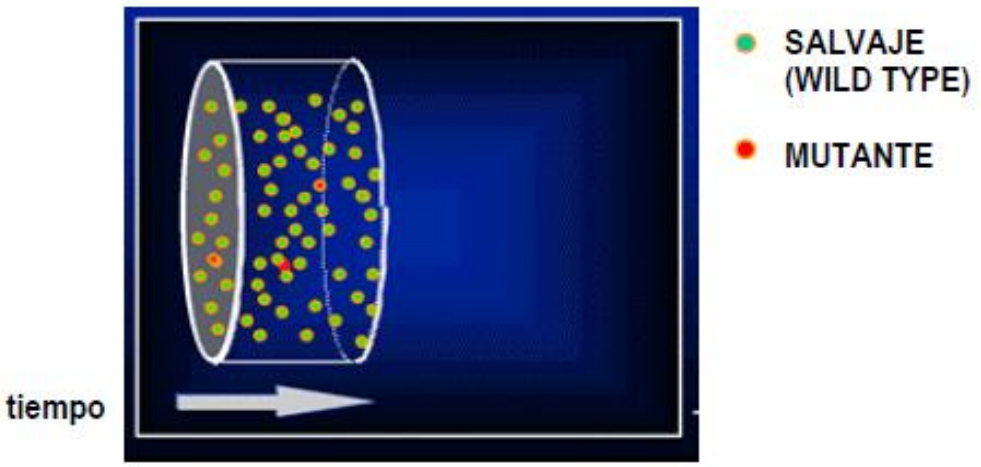


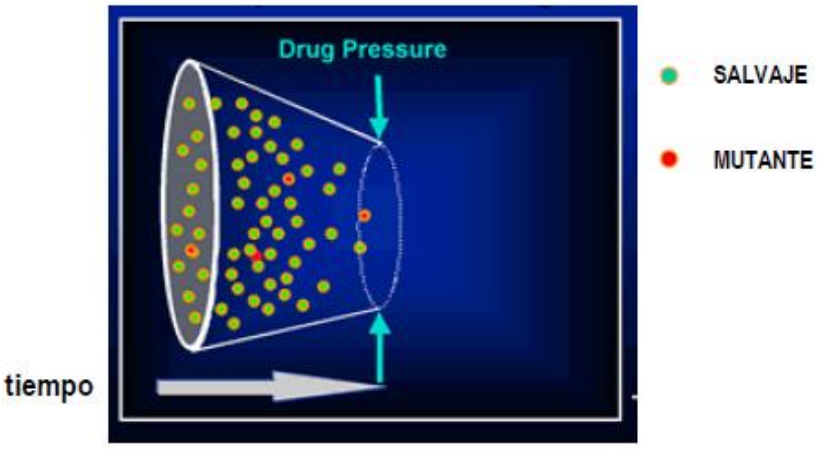
Figura 7. Representación esquemática de la dinámica en cuasiespecie. [48]

Si nos centramos 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 8). 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 [48].

En las siguientes imágenes vemos la dinámica de las cuasiespecies sin ningún tipo de presión selectiva [49].



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*.

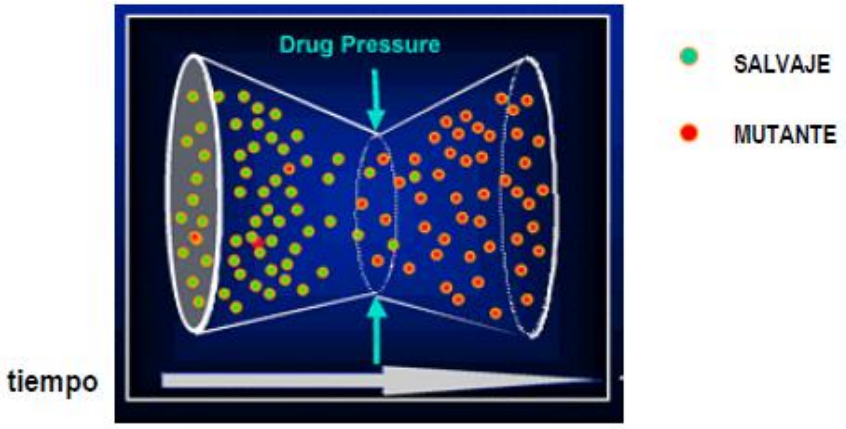


Figura 8**.** Selección de cuasiespecies. [49]

7. INTRODUCCIÓN A LOS MÉTODOS DE ANÁLISIS FILOGENÉTICO

La filogenia es la ciencia que estudia como 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 presentaran y aumentara la probabilidad de compartir un ancestro común. Los estudios filogenéticos se basan en la creación de arboles, que relacionan mediante ramas la distancia genética y la relación evolutiva entre varios organismos. Posteriormente, estos análisis también se han aplicado al estudio del VIH, permitiendo establecer su origen, clasificación y la relación evolutiva entre los distintos virus [50].

7.1 Árboles filogenéticos

El primer paso para realizar un análisis filogenético es el alineamiento de las secuencias, de tal forma que los nucleótidos homólogos queden posicionados en la misma columna [51], para ello, existen programas bioinformáticos que realizan alineamientos múltiples [52]. Una vez que se ha generado el alineamiento de secuencias y se ha seleccionado el modelo apropiado de evolución de secuencia, se puede inferir un árbol filogenético. Un árbol filogenético sigue una estructura determinada dada por un algoritmo matemático, que puede usarse para representar en el caso de VIH relaciones de parentesco entre distintos pacientes. La estructura de un árbol filogenético se compone de ramas, nodos y raíz. Cada rama corresponde con una secuencia genética, dichas ramas conectan los nodos y un nodo es el punto donde el cual dos o más ramas divergen. Las longitudes de las ramas que conectan los nodos representan la cantidad de cambios que ocurren entre cada nodo. Estas longitudes se calculan directamente desde la alineación y pueden variar dependiendo del modelo de evolución que se use. Por lo tanto, a mayor longitud de rama mayor divergencia y evolución de la secuencia analizada [53]. Por otro lado, las ramas y nodos pueden ser internos o externos, los nodos externos corresponden a las secuencias a partir de las cuales deriva el árbol y los nodos internos corresponden a hipotéticos ancestros comunes [53,54]. Llamamos *cluster* filogenético al conjunto de

secuencias que divergen de un solo nodo, lo que significa que dicho conjunto se originaron a partir de un ancestro común [53].

La reproducibilidad o estabilidad de la topología de los árboles filogenéticos generados debe analizarse mediante métodos estadísticos. El más utilizado se denomina *bootstrapping* no paramétrico, que asigna un valor (*bootstrap*) a cada nodo, expresado en forma de porcentaje. Por regla general se considera que un valor de *bootstrap* >70% representa una probabilidad aproximada del 95% de que la relación filogenética entre las secuencias sea real [53, 51].

Durante el curso de esta tesis se utilizo el métodos de máxima verosimilitud (ML) [55], por lo que a continuación se explica dicho método.

7.2 Método de Máxima Verosimilitud

En el enfoque ML el árbol seleccionado es el que da mayor probabilidad de producir la alineación de múltiples secuencias de entrada [56]. Este enfoque es más lento que la metodología Neighbor joining (NJ), ya que deben ser examinadas muchas topologías usando una heurística apropiada. El algoritmo de Felsenstein se puede utilizar para los cálculos de puntuaciones de probabilidad para los árboles individuales [57]. Esto se repite para todas las topologías de árboles posibles y el árbol con la puntuación de probabilidad más alta es el árbol con la mayor probabilidad de producir la alineación de entrada. Por lo tanto, el número de posibles topologías de árboles aumenta rápidamente con el número de taxones. En general, los árboles filogenéticos están bifurcados [56] y el numero de topologías de posibles arboles para n taxones esta dado por (2n - 3)! [58], lo que da lugar a la posibilidad de un gran número de árboles filogenéticos.

Por lo tanto, el método ML optimiza el ajuste entre el árbol y los datos, a medida que busca la topología de árbol que más probablemente dio origen al conjunto de datos a estudiar. La máxima verosimilitud es un método de inferencia ampliamente utilizado y se considera que produce el resultado más preciso [55].

8. SECUENCIACION GENÓMICA VIH-1

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-1. 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) [59,60], esta implementación revoluciono entre otros los estudios filogenéticos [61,62]. El potencial de NGS para detectar variantes virales VIH-1 de baja frecuencia (<1%) se ha determinado en varios estudios [63], en cambio la secuenciación tradicional Sanger presenta un umbral de detección de variantes alrededor del 20% [64, 65] (Figura 9).

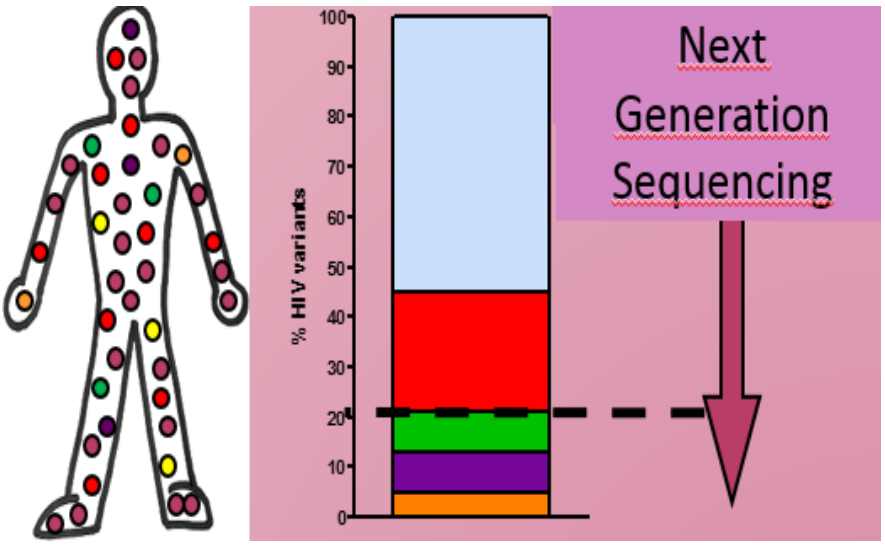


Figura 9. Detección de variables minoritarias mediante secuenciación masiva.

Existen varias técnicas NGS para secuenciar VIH-1 [66,67], siendo capaces de generar de tres a cuatro órdenes de magnitud más de información que la secuenciación Sanger [68]. Esto hace que para los estudios filogenéticos mediante secuenciación NGS 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 [69]. 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 [70,71]. Una alternativa es

generar una única secuencia consenso NGS, pero algunos estudios no son claros u omiten el método utilizado para generar la secuencia consenso NGS [72], otros recurren a programación tipo Python para generar una única secuencia consenso NGS [73], siendo necesario altos niveles de bioinformática.

Hoy en día tienden a coexistir la secuenciación Sanger con NGS, 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.

En el presente doctorado se ha utilizado la tecnología NGS mediante la plataforma 454-GS Junior por lo que detallaremos esta tecnología a continuación.

8.1 Secuenciación NGS VIH-1

Durante muchos años la secuenciación Sanger ha sido la utilizada para secuenciar el gen *pol* en el estudio de resistencias de pacientes infectados VIH-1. Esta secuenciación daba como resultado una única secuencia consenso, en la que cada posición de nucleótido era resultado de la base más frecuente en la población vírica [74]. En 2013 se empezó a introducir la tecnología NGS en ciertos laboratorios con fines clínicos, como es el estudio de resistencias antirretrovirales. Actualmente la compañía Roche Life Sciences es la que tiene validado el kit de resistencias VIH-1 para la plataforma 454 GS Junior utilizado en 2013-2017 y desde 2017 hizo una adaptación a la plataforma Illumina.

8.1.1 Creación de la librería

Se comenzara realizando las PCR con unos *primers* específicos de la región *pol* a estudiar, añadiendo a los *primers* un pequeño fragmento de nucleótidos que serán identificativos de cada paciente (MIDs). Una vez estandarizados todos nuestros

amplicones a la misma concentración podemos hacer la librería de nuestras muestras. Esta creación de librería no será más que el proceso de fragmentación mediante un proceso llamado "nebulización". Cuando tengamos el ADN en fragmentos de 200-800 pares de bases podemos añadir dos *primers* que contienen regiones en los extremos (adaptadores), permitiendo la amplificación y secuenciación, el adaptador A serán la dirección *forward*, mientras que el B el *reverse*. Además estos adaptadores contienen una secuencia característica TCAG que dará inicio a la secuenciación.

8.1.2 emPCR

La emPCR es una PCR en emulsión, mediante un aceite conseguiremos crear unas esferas, donde se introducirán nuestros reactivos y un único fragmento de ADN que producirá una PCR monoclonal.

8.1.3 Pirosecuenciación

La emPCR anterior será tratada para eliminar las esferas que no hayan obtenido un buen rendimiento de amplificación. Posteriormente se añadirán cebadores complementarios a los adaptadores y enzimas necesarios para la pirosecuenciación. La muestra será cargada en un chip llamado *PicoTiterPlate*, que será introducido en el secuenciador.

La pirosecuenciación 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 que aprovecha el pirofosfato liberado al añadirse nucleótidos a la cadena secuenciada. Estos destellos de luz serán captados por una cámara CDD (*charge-coupled device*) que los traducirá a nucleótidos en la secuencia.

Un resumen esquematizado de la técnica NGS para plataforma 454 GS-Junior describe en la siguiente figura 10.

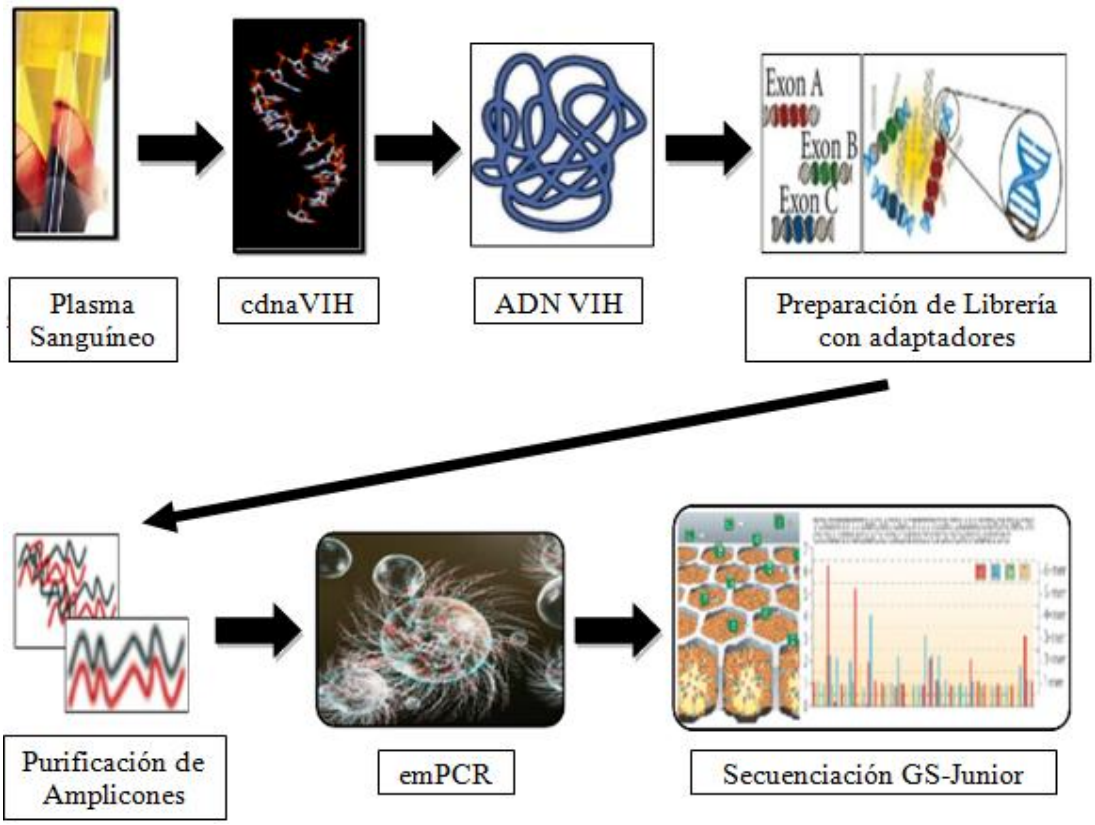


Figura 10. Descripción esquemática del sistema de secuenciación NGS mediante 454 GS-Junior.

9. CORRECCION DE ERRORES NGS

La secuenciación NGS genera hasta millones de lecturas que pueden incluir variantes de baja frecuencia o errores de secuenciación, por ello es crucial distinguir con alta sensibilidad y especificidad variantes biológicas de los errores del proceso [75, 76].

Hay cuatro tipos de errores básicos de secuenciación: inserción, delección, desajuste y bases desconocidas (N's) [79]. El error más frecuente en la plataforma 454 GS Junior son los *indels* (inserción/delección), seguido de las bases desconocidas [77, 78, 79]. Estos tipos de errores pueden ser introducidos en las etapas de pre-secuenciación, es decir antes en el proceso de preparación de la librería, y/o en el proceso de pirosecuenciación (en el caso de 454) [80]. Se ha estimado que 454 tiene una tasa de error media aproximada del 1% [68]. Ha sido ampliamente estudiado que dicha plataforma pierde precisión a medida que transcurre la secuenciación, disminuyendo la calidad, además pierde exactitud en las regiones homopolímeras, aumentando el error con la longitud de estas regiones. Este hecho es debido a que en la reacción de secuenciación se solapan las distribuciones de intensidades de luz en los ciclos de flujo [81, 82].

Podemos observar en casi todas las publicaciones un apartado donde incluyen la metodología de corrección de errores [83-86]. Además varios estudios demuestran la necesidad de corrección de errores para generar unos datos fieles para el posterior ensamblaje [87]. En el caso del estudio de resistencia frente a antirretrovirales VIH, las secuencias son alineadas frente a un genoma de referencia, por lo tanto los errores serán traducidos en mutaciones (pudiendo ser primarias) falsas [88], por lo tanto se hace evidente la necesidad de un filtrado de secuencias, ya que los posibles errores afectaran en la detección de los polimorfismos de un solo nucleótido (SNPs).

9.1 Apoyo tecnológico

Para el proceso de filtrado hace falta partir del archivo fastq generado en el proceso de secuenciación. El archivo fastq es la conjunción de las distintas base nucleotídicas junto a su calidad de secuenciación, 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 buenos (Figura 11). Tras finalizar el proceso de secuenciación, la plataforma 454 devuelve un archivo fasta y otro archivo de calidad (q), por lo que deberemos producir el archivo fastq mediante script.

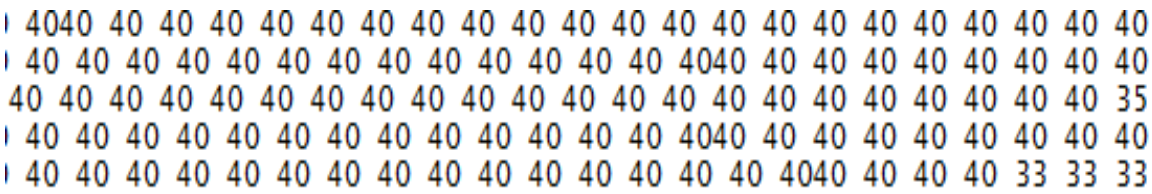


Figura 11. Archivo *quality*. Los números que aparecen es la calidad numérica de la base nucleotídica.

Hay varias herramientas de software que manejan múltiples tecnologías que pueden hacer frente a todo tipo de errores [89, 90, 91, 92]. Estos programas tienen distintos algoritmos y parámetros para dar soporte para los diferentes tipos de errores comentados anteriormente. Estas herramientas pueden tener distintas metodologías que veremos a continuación.

Espectro *k*

El enfoque más utilizado por la mayoría de software de corrección está basado en el espectro *k*. Un *k-mer* es un segmento de una lectura con k-bases [93]. El conjunto de todos los *k-mers* de una lectura se genera utilizando una ventana deslizante de dimensión *k*. En el proceso de comprobación de errores, dicha ventana se desplaza por un elemento y se añade el segmento visible de la lectura al conjunto de espectro. El algoritmo se basa en la alineación espectral [93,94], por lo que, si un *k-mer* aparece

menor que un umbral se considera débil, descartando estas secuencias que podrían contener posibles errores.

Alineación basada en secuencia múltiple

Los softwares basados en la alineación de secuencias múltiples se centran en alinear las lecturas para identificar la superposición entre ellas. Los métodos utilizan diferentes algoritmos [95] para construir secuencias consenso más probables a partir del conjunto de secuencias. Las secuencias que contengan *k-mers* erróneos fallarán en el proceso de alineación, provocando la eliminación de esas secuencias. Posteriormente, estas secuencias se enfrentaran a una secuencia de referencia y mediante unos parámetros de calidad establecidos (profundidad de secuenciación, longitud de secuencia, calidad) determinara si dichos SNPs presentes son anomalías.

Modelos basados en probabilística

Los modelos basados en probabilística utilizan el algoritmo de maximización de expectativas para determinar la base correcta en cada posición calculando la probabilidad de las variantes existentes en esa posición específica.

HIPÓTESIS

La utilización de la secuenciación masiva en muestras de ADN proviral y plasma para la determinación de resistencias antirretrovirales, con los adecuados métodos bioinformáticos de corrección permiten aumentar la calidad de los resultados y de la información para su utilización en la clínica.

OBJETIVOS

El empleo NGS da múltiples ventajas, al producir millones de secuencias permite la detección de mutaciones por debajo del umbral Sanger (aproximadamente 20%), por lo que mejora la toma de decisiones a la hora de seleccionar el tratamiento, tanto en pacientes *naïve*, como en pacientes con fallos en varias líneas de tratamiento. Dicho estudio de resistencias se puede realizar en plasma sanguíneo, pero en ciertas ocasiones (viremias muy bajas o indetectabilidad de carga viral) también puede realizarse en ADN proviral, observando mutaciones que han quedado de forma latente. Estudios en ADN proviral pueden ser útiles en casos de simplificación de tratamiento, siendo un caso importante la simplificación en pacientes que contienen RAL hacia una o dos tomas al día de DTG, dependiendo de la presencia o no de determinadas mutaciones. Por ello nos propusimos analizar la presencia de mutaciones de la región IN en ADN proviral, observando la posibilidad de simplificar la línea de tratamiento RAL hacia DTG.

A día de hoy, siguen conviviendo la tecnología Sanger y NGS en el estudio de resistencias antirretrovirales. Este hecho hace que en estudios multicéntricos se tenga que estandarizar los resultados NGS hacia Sanger, incluso en ciertas bases de datos VIH europeas o mundiales es necesaria la introducción de una secuencia consenso tipo Sanger. Todo ello hace necesario la simplificación de las miles de secuencias NGS hacia una única secuencia consenso, pero no existen datos sobre como puede representar esta secuencia consenso a la secuencia tipo Sanger. Por otra parte, la tecnología NGS puede producir ciertos errores, tanto en el proceso de secuenciación como en el procesado de muestras. Por lo tanto, es necesaria la revisión de los datos obtenidos, eliminando los posibles errores. Dicho paso nos aumentara la exactitud de nuestros resultados en varios estudios, como son; el estudio de resistencia, obtención de secuencia consenso, estudios filogenéticos, etc. Por lo tanto, en este doctorado se pretende obtener el máximo rendimiento de las herramientas bioinformáticas para el manejo de datos NGS.

Todos estos interrogantes han sido los que nos han inducido a realizar el presente trabajo, cuyos objetivos se exponen a continuación.

Objetivo principal

1) Aplicación de NGS al manejo de las resistencias a antirretrovirales y a la epidemiologia molecular en la infección por VIH.

Objetivos específicos

1) Estudiar a partir de muestras de células mononucleares de sangre periférica (PBMC) si DTG se puede utilizar como fármaco de rescate ante situaciones de fracaso a RAL, y cuál debe ser su pauta de dosificación.

2) Comparar los resultados obtenidos mediante secuenciación poblacional con NGS.

3) Analizar los resultados NGS con diferentes herramientas bioinformáticas, mejorando la detección de mutaciones y la eliminación de posibles errores.

4) Establecer cuál es el mejor umbral de corte para la obtención de una secuencia consenso de las obtenidas por NGS, representativa de la secuencia tipo Sanger.

RESULTADOS: ARTÍCULOS CIENTÍFICOS

Usefulness of Integrase resistance testing in proviral HIV-1 DNA in patients with Raltegravir prior failure

1

BMC Infectious Diseases. 2016:16;197

Jose Ángel Fernández-Caballero[1,*], Natalia Chueca[1], Marta Álvarez[1], María Dolores Mérida[1], Josefa López[1], José Antonio Sánchez[1], David Vinuesa[1], María Ángeles Martínez[1], José Hernández[1] and Federico García[1]

* Correspondence: jose.angel.fernandez.caballero@gmail.com
1. Complejo Hospitalario Universitario Granada. Servicio de Microbiología, Hospital Universitario San Cecilio. Instituto de Investigación Ibs. Granada, Av. Del Conocimiento s/n, 18016 Granada (Andalucía), Spain

Abstract

Background: In our study, we have hypothesized that proviral DNA may show the history of mutations that emerged at previous failures to a Raltegravir containing regimen, in patients who are currently undetectable and candidates to simplification to a Dolutegravir containing regimen, in order to decide on once a day or twice a day dosing.

Methods: We have performed a pilot, observational, retrospective, non interventional study, including 7 patients infected by HIV-1, all with a history of previous failure to a RAL containing regimen, that were successfully salvaged and had reached viral suppression. A genotypic viral Integrase region study was available for each patient at the moment of RAL failure. After an average (IQR) time of 48 months (29–53) Integrase resistance mutations in proviral DNA were studied.

Results: All the patients were infected by HIV-1 B subtypes, with a mean age of 55 (range 43 to 56), originating from Spain, and 4 were women. Median viral load (log) and CD4 count at the moment of the study on proviral DNA was of 1.3 log cp/ml (range 0–1.47) and 765.5 cells/μL (range; 436.75–1023.75). The median time (IQR) between previous failure to RAL and the study on proviral DNA was 48 (29–53) months. At Raltegravir failure, N155H was detected in four patients, and other secondary mutations were detected in five patients (71.4 %). In proviral DNA, N155H was detected by population sequencing in three patients (42.8 %), and UDS demonstrated a 9.77 % relative abundance of N155H in the remaining patient. Sanger sequencing correctly identified all the secondary mutations.

Conclusion: This is a pilot study that demonstrates the possibility of properly identifying N155H and some secondary mutations 29–53 months after failure.

Keywords: HIV, Integrase, Proviral DNA, Raltegravir, Dolutegravir

Background

For the last 6 years, Raltegravir (RAL), the first strand Integrase transfer inhibitor (INIs) approved, has played a relevant role on the treatment of HIV infection, both for naive and pretreated patients [1].

Nowadays, Elvitegravir (EVG) and Dolutegravir (DTG), two new INIs have become available. EVG is a first generation drug and shares its resistance profile with RAL, whereas DTG is a second generation INI with a high antiviral effect, and excellent efficacy and safety profile. Although DTG has a high genetic barrier to resistance, its activity may be limited by certain combination of resistance mutations that may accumulate during failure to RAL [2]. The ocurrence of these mutations is not frequent [3] and the use of DTG to salvage patients who have developed resistance to RAL has been well studied and documented [4, 5]. DTG has a long plasmatic half-life, so it can be dosed once a day (QD) on patients without preexistent resistance to INIs; if the patient to be treated is infected by viruses carrying any resistance mutation to INIs, then DTG has to be dosed twice a day [6, 7] (BID). If the resistance profile at failure was not documented, patients with a prior failure to RAL need to be treated BID, as this is the safest approach.

In our study, we have hypothesized that proviral DNA may show the history of mutations that emerged at failure to RAL, and we attempted to provide the proof of concept that testing proviral DNA may be used as a sentinel for RAL mutations, in patients who are currently undetectable and candidates to simplification to a DTG containing regimen.

Methods

We have performed a pilot, observational, retrospective, non interventional study, including 7 patients infected by HIV-1, all with a history of previous failure to a RAL containing regimen, that were successfully salvaged and had reached viral suppression. A genotypic viral Integrase region study was available for each patient at the moment of

RAL failure. After an average (IQR) time of 48 months (29–53), Integrase resistance mutations in proviral DNA were studied.

Peripheral Blood Mononuclear Cells (PBMC) were separated by Ficoll gradient centrifugation, washed, counted and pelleted to 5x106 leukocytes aliquots, that were used for DNA extraction with QIAamp DSP DNA Blood Mini Kit (QIAGEN). DNA was quantified by spectrophotometry (NanoDrop, ThermoScientific).

Integrase was amplified using a nested-PCR. For the first round of amplification, we used outer primers IN1F (5′-GGAAAAGGTCTACCTGTCATGGGT-3′) and IN1R (5′-GGAGAAAGAGACTGGCATTTGG-3′) and the following PCR profile: 94 °C/2′; (94 °C/15″; 60 °C/30″; 72 ° C/45″) x35 cycles; 72 °C/7′; 4 °C. This profile was used for the nested PCR, using inner primers INR2F (5′-TGGAGGAAATGAACAAGTAGATAAATT-3′) and (INT2R: 5′-GGGTCTGCATACAGGAGAAA-3′).

For Sanger sequencing, we used a bidirectional sequencing protocol and the Thermo Sequenase Dye Primer Manual Cycle Sequencing Kit (Affymetrix USB), which was optimized for the TruGene/OpenGene DNA sequencing system (SIEMENS); sequencing primers were P1F:5′- Cy5.5-GTAGCCAGCTGTGATAAATGTC-3′) and P2R (5′- Cy5-CTGCCATTTGTACTGCTGTCT-3′), allowing to sequence 414 nucleotides in the Integrase region, covering positions 40 to 178. The amplification profile was: 94 °C/5′; (94 °C/20″; 62 °C/20″; 72 °C/2′) x30 cicles; 70 ° C/5′; 4 °C. The obtained sequences were aligned and combined using OpenGene Geneobjects™ software, and fasta sequences were interpreted Stanford HIV Database Versión 7.0 [8].

Ultra Deep Sequencing (UDS) was done using an HIV-1 UDS prototype (Roche diagnostics) and the 454 GS Junior (Roche 454 Life Sciences Branford, CT). As template for the UDS PCR, we used the same amplification products as for standard Sanger sequencing following from this point the manufacturer's recommendations. Once UDS was performed, sequences were exported by AVA (GS Amplicon Variant Analyzer, Roche) software and interpreted by DeepChek®- VIH (ABL, SA) software, using for interpretation the same Stanford version as for Sanger sequencing.

This was a retrospective, non-interventional study, and patient information was anonymized and de-identified prior to analysis.

Results

All the patients were infected by HIV-1 B subtypes, with a mean age of 55 (range 43 to 56), originating from Spain, and 4 were women. Median viral load (log) and CD4 count at the moment of the study on proviral DNA was 1.3 log cp/ml (range 0–1.47) and 765.5 cells/ml (range; 436.75-1023.75). The median time (IQR) between previous failure to RAL and the study on proviral DNA was 48 (29–53) months. Detailed information on baseline characteristics is shown in Table 1.

Table 1 Socio-demographic and clinical characteristics of patients

Patient	Date of failure	Age	CD4	VL (log)	Time (monghts) among samples
1	15/10/2009	56	NR	ND	21
2	20/10/2009	56	484	1,47	53
3	01/10/2009	59	670	1,30	52
4	17/02/2011	39	1365	1,47	36
5	02/03/2010	51	910	1,30	48
6	26/10/2012	43	295	2,74	29
7	13/8/2009	55	861	ND	56

Viral load (VL) is in the actual moment. (*NR* Not recorded, *ND* Not detectable)

Table 2 shows the correlation on resistance mutations (Sanger sequencing) detected at RAL failure in plasma and on proviral DNA, after a median period of 48 months of being undetectable. At failure, N155H was detected in four patients, and other secondary mutations were detected in five patients (71.4%). In proviral DNA, N155H was detected by population sequencing in three patients (42.8%), and UDS demonstrated a 9.77% relative abundance of N155H in the remaining patient. Sanger sequencing correctly identified all the secondary mutations. We observed that proviral

DNA and plasma RNA drug resistance mutations and polymorphisms were highly concordant.

Table 2 Primary and secondary resistance mutations in the Integrase by Sanger population sequencing

Patient	Major resistance mutation	Accesory mutation	Polymorphism mutation
1F	N155H	___	C56S, E85EG, L101I, S119P, T122I, H171Q, K173EK
1A	N155H	___	C56S, L101I, S119P, T122I, H171Q
2F	N155H	___	M50I, L68R, V71I, L101I, S119P, H171Q
2A	N155H	___	M50I, V71I, P90PS, L101I, S119P
3F	___	L74I	E96D, K111T, K160KT
3A	___	L74I	E96D, K111T, G123RS
4F	___	G163GR	L101I, I113V, G134E, V150AV
4A	___	G163GR	M50IM, L101I, I113V, V150A
5F	___	L74IM	M50V, V72I, K103R, K111T, A124T
5A	___	L74I	M50V, V72I, K103R, K111T
6F	N155H	T97A	D55Y, V72I, K111T, I113V, S119R, G123S, A124N, T125A
6A	___	___	V72I, K111T, I113V, S119R, G123S, A124N, T125A
6 UDS	N155H (9.77 %)	T97A (12.42 %)	V72I(37.44 %), Y99C(4.65 %), T122I(12.56 %), K156N(14.35 %), E157A(15.35 %), K111T(39.53 %), I113V(29.3 %), S119R(37.44 %), G123S(97.21 %), A124N (43.26 %), T125A(45.58 %)
7F	N155H	V151I	I113V, S119P, T122I, A124N, C130Y
7A	N155H	V151I	G52P, S119PR, T122I, I161X

Patients are indicated with the numbers 1 to 7; F relates to the time point of therapeutic failure (plasma RNA), A to the proviral DNA studies after virological suppression, and UDS to massive sequencing data

Discussion

Dolutegravir has shown excellent efficacy and safety in individuals infected by HIV both in naïve [9], and patients with prior exposure to RAL [10]. Only the accumulation of Q148H/R/K, together with other secondary mutations broadens DTG activity [11]. The VIKING study evaluated Dolutegravir dosing, demonstrating a higher efficacy, tolerability and safety when dosing DTG 50 mg [10] twice a day (BID) for patients with resistance mutations in the Integrase. While BID is the safest approach, DTG is only recommended 50 mg once a day (QD) for patients with no resistance against Integrase inhibitors. For some patients who have not been tested for Integrase resistance at failure, and have been effectively suppressed with a new antiretroviral regimen, BID remains

the safest dosing strategy, but QD could possibly play a role, reducing the cost of the new regimen.

Proviral DNA may be a useful tool to investigate the presence of resistance mutations [12–14], especially in patients who as a consequence of antiretroviral therapy are virologically suppressed.

In our study, using Sanger sequencing of the Integrase region of proviral DNA, we could correctly identify failing selected mutations in 6/7 patients. Although for the remaining patient we could not demonstrate the failing mutation with Sanger sequencing, using a more sensitive test [15], resulted in the correct identification of the failing mutations [N155H (9.7 %) and T97A (12.42 %)] suggesting that, given the superiority of massive parallel sequencing, this should be the tool recommended for testing proviral DNA in virologically suppressed patients, although at present it is an expensive tool that may not be feasible in some laboratories.

Although the sampling time in patients 2, 3 & 7 exceeded the half-life of the HIV-1 reservoir, this did not interfere in the correlation between the failing sample and proviral DNA testing. Despite some studies have demonstrated that the latent viral reservoirs half-life is from four to six months in patients who start therapy in the acute infection stage, there are other studies in chronically infected patients who have shown a half-time of 44 months [16, 17].

Our study has certain limitations. First, only subtype B patients have been included, so the methodology needs to be validated for other subtypes. Secondly, only the N155H pathway was confirmed in some patients and it is possible that resistance pathways other than N155H, that could have emerged before N155H was established, may have been archived in the proviral DNA of the patients compromising DTG activity.

Conclusions

In summary, despite the limitations of our study, which is just a pilot study that should be confirmed in further studies, we have shown the proof of concept that for patients who failed a Raltegravir containing regimen in the past, who are currently virologically suppressed, and lack the resistance information at failure, studying Integrase resistance in the proviral DNA accurately reflects the possibility of properly identifying N155H and some secondary mutations 29–53 months after failure.

References

1. Rockstroh JK, DeJesus E, Lennox JL, Yazdanpanah Y, Saag MS, Wan H, et al. Durableefficacy and safety of raltegravir versus efavirenz when combined with tenofovir/emtricitabine in treatment-naive HIV-1-infected patients: final 5-year results from STARTMRK. J Acquir Immune Defic Syndr. 2013;63(1):77–85.

2. Carganico A, Dupke S, Ehret R, Berg T, Baumgarten A, Obermeier M, et al. New dolutegravir resistance pattern identified in a patient failing antiretroviral therapy. J Int AIDS Soc. 2014;17(4 Suppl 3):19749.

3. Codoñer FM, Pou C, Thielen A, García F, Delgado R, Dalmau D, et al. Dynamic escape of pre-existing raltegravir-resistant HIV-1 from raltegravir selection pressure. Antiviral Research. 2010;88(3):281–6.

4. Blanco JL, Martinez-Picado J. HIV Integrase inhibitors in ART-experienced patients. Curr Opin HIV AIDS. 2012;7(5):415–21.

5. Garrido C, Soriano V, Geretti AM, Zahonero N, Garcia S, Booth C, et al. Resistance associated mutations to dolutegravir (S/GSK1349572) in HIVinfected patients-Impact of HIV subtypes and prior raltegravir experience. Antiviral Res. 2011;90(3):164–7.

6. Katlama C, Murphy R. Dolutegravir for the treatment of HIV. Expert Opin Investig Drugs. 2012;21(4):523–30.

7. Rathbun RC, Lockhart SM, Miller MM, Liedtke MD. Dolutegravir, a Second-Generation Integrase Inhibitor for the Treatment of HIV-1 Infection. Ann Pharmacother. 2014;48(3):395–403.

8. Stanford HIV Database Version 7.0. http://hivdb.stanford.edu/. Accessed 05 Dec 2015.

9. Van Lunzen J, Maggiolo F, Arribas JR, Rakhmanova A, Yeni P, Young B, et al. Once daily dolutegravir (S/GSK1349572) in combination therapy in antiretroviralnaïve adults with HIV: planned interim 48 week results from SPRING-1, a doseranging, randomised, phase 2b trial. Lancet Infect Dis. 2012;12(2):111–8.

10. Eron JJ, Clotet B, Durant J, Katlama C, Kumar P, Lazzarin A, et al. Safety and efficacy of dolutegravir in treatment-experienced subjects with raltegravirresistant HIV type 1 infection: 24-week results of the VIKING Study. J Infect Dis. 2013;207(5):740-8.

11. Castagna A, Maggiolo F, Penco G, Wright D, Mills A, Grossberg R, et al. Dolutegravir in antiretroviral-experienced patients with raltegravir.and/or elvitegravir-resistant HIV-1: 24-week results of the phase III VIKING-3 study. J Infect Dis. 2014;210(3):354–62.

12. Chew CB, Potter SJ, Wang B, Wang YM, Shaw CO, Dwyer DE, et al. Assessment of drug resistance mutations in plasma and peripheral blood mononuclear cells at different plasma viral loads in patients receiving HAART. J Clin Virol. 2005;33(3):206–16.

13. Coovadia A, Hunt G, Abrams EJ, Sherman G, Meyers T, Barry G, et al. Persistent minority K103N mutations among women exposed to single-dose nevirapine and virologic response to nonnucleoside reverse-transcriptase inhibitor-based therapy. Clin Infect Dis. 2009;48(4):462–72.

14. MacLeod IJ, Rowley CF, Thior I, Wester C, Makhema J, Essex M, et al. Minor resistant variants in nevirapine-exposed infants may predict virologic failure on nevirapine-containing ART. J Clin Virol. 2010;48(3):162–7.

15. García F, Álvarez M, Bernal C, Chueca N, Guillot V. Laboratory diagnosis of HIV infection, viral tropism and resistance to antiretrovirals. Enferm Infecc Microbiol Clin. 2011;29(4):297–307.

16. Quercia R, Dam E, Perez-Bercoff D, Clavel F. Selective-advantage profile of human immunodeficiency virus tipe 1 Integrase mutants explains in vivo evolution of raltegravir resistance genotypes. J Virol. 2009;83(19):10245–9.

17. Zhang L, Ramratnam B, Tenner-Racz K, He Y, Vesanen M, Lewin S, et al. Quantifying residual HIV-1 replication in patients receiving combination antiretroviral therapy. N Engl J Med. 1999;340(21):1605–13.

Validación de un método seguro y sencillo para la elaboración de secuencias consenso del virus de la inmunodeficiencia humana a partir de los datos de secuenciación masiva 454

2

Enfermedades infecciosas y microbiología clinica. 2016 Oct 3. Doi: 10.1016/j.eimc.2016.08.008

Jose Ángel Fernández-Caballero-Ricoa[a], Natalia Chueca-Porcuna[a], Marta Álvarez-Estévez[a], María del Mar Mosquera-Gutiérrez[b], María Ángeles Marcos-Maeso[b] y Federico García[a]

a. Servicio de Microbiología Clínica, Hospital Universitario San Cecilio, Complejo Hospitalario Universitario Granada e Instituto de Investigación IBS, Granada, España
b. Servicio de Microbiología Clínica, Centro de Diagnóstico Biomédico, Hospital Clínic, Universidad de Barcelona, Barcelona, España

Abstract

Objective: To show how to generate a consensus sequence from the information of massive parallel sequences data obtained from routine HIV anti-retroviral resistance studies, and that may be suitable for molecular epidemiology studies.

Material and methods: Paired Sanger (Trugene-Siemens) and next-generation sequencing (NGS) (454 GSJunior-Roche) HIV RT and protease sequences from 62 patients were studied.NGSconsensus sequences were generated using Mesquite, using 10%, 15%, and 20% thresholds. Molecular evolutionary genetics analysis (MEGA) was used for phylogenetic studies.

Results: At a 10% threshold, 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: A safe method is presented to generate consensus sequences from HIV-NGS data at 20% threshold, which will prove useful for molecular epidemiological studies.

Keywords: Human immunodeficiency virus, Phylogeny, Next generation sequencing, Thresholds

Introducción

Recientemente, un buen número de servicios de microbiología clínica han adoptado las técnicas de secuenciación masiva (next generation sequencing [NGS]) para los estudios de resistencias a antirretrovirales en pacientes VIH. La capacidad de NGS en la detección de variantes virales de baja frecuencia se ha determinado en varios estudios[1], disminuyendo la sensibilidad en la detección de mutaciones de resistencia hasta niveles del 1% (variantes minoritarias), lo que proporciona ventajas para la elección de la mejor línea de tratamiento y evitar el fracaso al tratamiento[2,3]. En nuestro país, uno de los motivos de la instauración de NGS para la detección de resistencias a antirretrovirales ha sido la discontinuación de los métodos de secuenciación Sanger comerciales por alguno de los proveedores.

Las secuencias de proteasa (PR) y transcriptasa reversa (RT) obtenidas de los ensayos para determinar resistencias se utilizan a menudo por parte de investigadores en estudios de epidemiología molecular, mediante el empleo de técnicas de filogenética y filodinámica[4]. Con la introducción de las técnicas de NGS, esta información se puede perder debido a que el manejo y el almacenamiento de las secuencias para este tipo de estudios son complejos; además, si las secuencias de NGS no se tratan apropiadamente, pueden aportar resultados equivocados. Para emplear las secuencias de NGS en estudios filogenéticos se requiere tanto una formación especial para el procesado de secuencias, como de ordenadores de gran potencia para procesar el gran volumen de datos obtenidos[5]. Para los estudios de epidemiología molecular, una alternativa es generar una única secuencia consenso de NGS, pero algunos estudios no son claros u omiten el método utilizado para generarla[6]; además, no conocemos con certeza cuál es la representatividad de esta consenso de NGS de la secuencia obtenida por Sanger, y cómo influyen los puntos de corte que utilicemos para generar dicho consenso.

El objetivo de nuestro trabajo ha sido 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.

Métodos

Para nuestro estudio hemos utilizado secuencias de 62 pacientes naïve del periodo 2014-2015, nuevos diagnósticos VIH, referidos para estudios de resistencias a antirretrovirales. Las secuencias tipo Sanger se obtuvieron utilizando Trugene®HIV-1 Genotyping (Siemens-[NAD]). Para NGS utilizamos el kit GS V Type HIV-1 Drug Resistance Primer (Roche) para 454 GS-Junior, partiendo del mismo ARN. Las secuencias consenso de NGS se generan mediante el software Mesquite v. 2.75, seleccionando umbrales de corte del 10, del 15 y del 20%. Previo a la utilización de Mesquite se efectúa un filtrado de las secuencias, utilizando los comandos fastq filter del software Usearch según longitud deseada de amplicón y calidad de secuencia (> 30 Q). Mesquite[7] es un programa que funciona mediante iconos y pestañas, siendo intuitivo. Para su utilización es necesario exportar las secuencias filtradas en formato pfam y seleccionar el umbral de corte para la creación de la secuencia consenso, exportándola en formato fasta. Posteriormente, las secuencias del gen pol (PR 4-99; RT 38-247) se procesan, alinean mediante MUSCLE en MEGA 6.06 y se generan árboles filogenéticos mediante el método de máxima verosimilitud, utilizando el modelo General Time Reversible (GTR) para el cálculo de las distancias evolutivas, con una distribución gamma equivalente a 1,89, obtenido con Find-Model DNA y utilizando remuestreo de bootstrap con 1.000 réplicas para construir los árboles filogenéticos consenso. Para definir una relación entre secuencias se tienen en cuenta solo las ramas pertenecientes a clusters con un valor de bootstrap superior al 75%.

Finalmente, los árboles son procesados en FigTree v. 1.4.2. El análisis del subtipo viral se realizó utilizando REGA HIV-1 Subtyping Tool v. 3.0.

Resultados

Nuestro estudio ha incluido 62 pacientes VIH-1, naïve, mediana de edad de 37 años (IQR 30-45), carga viral (mediana) 74.900 cp/ml (IQR 20.715-176.250), recuento de CD4 (mediana) 430 células/ml (IQR 48,5-567,78); el 82% eran hombres.

Para evaluar la concordancia entre las secuencias consenso de NGS con diferentes umbrales y la secuencia original de Sanger hemos analizado el número de secuencias

que se asocian por pares entre sí, y los valores de bootstrap entre los pares. Utilizando un umbral de corte al 10% se observa que solo en 17/62 (27%) pacientes las secuencias Sanger están pareadas con NGS de la misma muestra, y en estas, la mediana de bootstrap fue del 88% (IQR 83,5-95,5). Aumentando el umbral al 15%, las secuencias se asocian por pares en 36/62 (58%) pacientes, con una mediana de bootstrap del 94% (IQR 85,5-98). Al 20%, esto sucede en 61/62 pacientes con una mediana de bootstrap del 99% (IQR 98-100) (fig. 1); para el caso en que la secuencia de NGS no se asocia con la secuencia de Sanger, detectamos un gran número de diferencias entre bases.

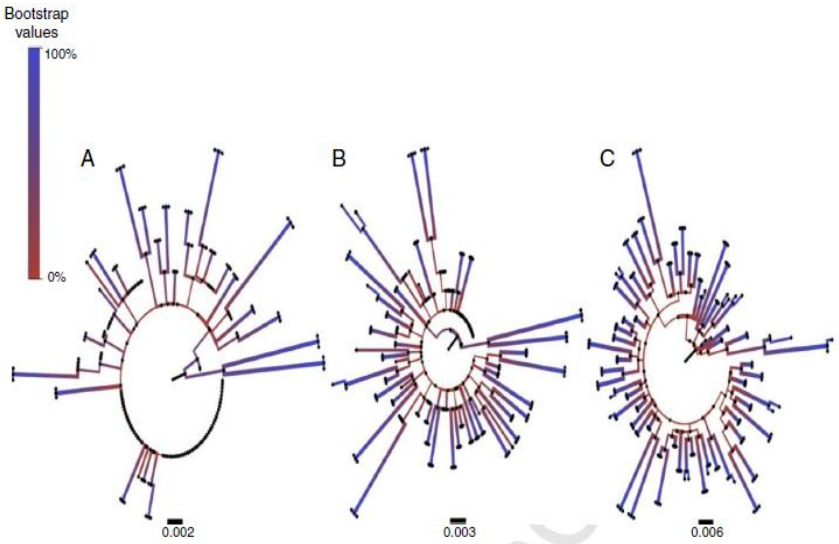


Figura 1. Representación de los árboles filogenéticos en FigTree v. 1.4.2, formados por las secuencias Sanger y secuencias NGS a los distintos umbrales: A) NGS-10%; B) NGS-15%, y C) NGS-20%. Los valores de *bootstrap* están asociados según el color de la gráfica, siendo una buena relación a partir de 70%.

La mayoría de los pacientes estaban infectados por subtipo B (77,4%), seguido de CRF02 AG (12,9%), A y F (3,2%) y C y G (1,6%). Utilizando consenso NGS umbral 10% y 15% se observan 2 casos discordantes respecto al subtipo Sanger: un caso subtipo B-NGS y A1-Sanger, y otro desde subtipo CRF03 AB-NGS y A1-Sanger. Estas diferencias desaparecen al utilizar las secuencias consenso NGS umbral 20% (tabla 1).

Tabla 1
Distribución de subtipos virales HIV según las secuencias analizadas Sanger y secuencia consenso NGS a los distintos umbrales, mediante REGA HIV-1*Subtyping Tool* v. 3.0

	Subtipo HIV						
	B	G	F	C	A	crf02_AG	crf03_AB
Sanger	48	1	2	1	2	8	0
NGS-10%	49	1	2	1	0	8	1
NGS-15%	49	1	2	1	0	8	1
NGS-20%	48	1	2	1	2	8	0

La figura 2 muestra la gráfica bootscan del subtipado en la segunda muestra discordante.

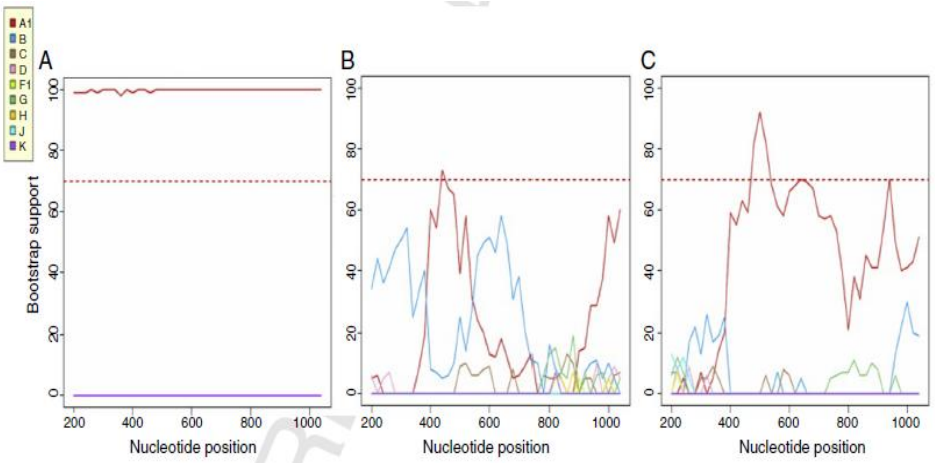


Figura 2. *Bootscan* de secuencia Sanger (A), secuencia consenso NGS-10% (B) y NGS-20% (C), mediante REGA HIV-1*Subtyping Tool* v. 3.0. El *bootscan* ofrece un mismo valor de subtipado HIV A para la secuencia Sanger y secuencia consenso NGS-20%, sin embargo se observa un subtipado CRF03_AB para secuencia consenso NGS-10%.

Discusión

Los estudios filogenéticos en VIH[8,9], en concreto los estudios de parentesco, dinámica de la epidemia VIH y de subtipado molecular utilizando la secuencia del gen pol, se han utilizado entre otros fines para conocer redes y nodos de transmisión del VIH, así como redes migratorias de los diferentes subtipos. Para estos objetivos la mayoría de los

estudios publicados, a nivel internacional[10] y a nivel local[11,12], han utilizado la secuenciación de tipo Sanger. Algunos de estos estudios han utilizado toda la información obtenida mediante NGS[13], pero por lo general se intenta generar una única secuencia consenso, habitualmente mediante comandos informáticos complejos. La transición desde la secuenciación Sanger a NGS para el estudio del gen pol en el análisis de mutaciones de resistencia ha provocando un cambio en el tipo de secuencias que manejamos en los servicios de microbiología clínica y paradójicamente puede suponer un freno para los estudios locales de epidemiología molecular de VIH en nuestro país. En nuestro trabajo proponemos la utilización de Mesquite, un software intuitivo, de fácil manejo y sin necesidad de comandos, que simplifica la obtención de la secuencia consenso a partir de secuencias obtenidas mediante NGS, demostrando que utilizando un umbral del 20% para generar esta consenso obtenemos una información segura, fiable y de idénticas características que la secuencia Sanger, que puede ser utilizada en estudios de epidemiología molecular, obviando la problemática actual de las secuencias obtenidas mediante NGS.

Como podemos observar en nuestro estudio, para poder utilizar con seguridad las secuencias consenso en estudios de epidemiología molecular en VIH, y para que la secuencia sea representativa de la secuencia tipo Sanger, debemos elevar el umbral de corte hasta el 20%. Solo así hemos conseguido una mediana de bootstrap del 99% (IQR 98-100) entre las secuencias consenso de NGS y la tipo Sanger. Con umbrales del 10 o del 15% el porcentaje de secuencias que se asocian por pares NGS-Sanger y la mediana de bootstrap son insuficientes. Además, la variabilidad llega a ser tal que hasta en la asignación del subtipo viral se cometen errores, hecho que se corrige con el consenso al 20%. Estas discrepancias son debidas a la multitud de bases ambiguas generadas con umbral 10 y 15%, con una disminución del soporte estadístico para la correcta adjudicación del subtipo viral.

Una parte importante en los estudios de epidemiología molecular es el proceso de alineación de secuencias, teniendo como objetivo aproximar posiciones homólogas en base a la verdadera historia evolutiva de las secuencias[14]. El problema de la utilización de secuencias consenso NGS 10% y 15% en tales estudios radica en la presencia de regiones ambiguas, presentando una incertidumbre sustancial, evitando la robustez de

análisis estadísticos tanto filogenéticos[15] como de subtipado, obteniendo resultados que no se corresponden a lo esperado.

Es importante indicar que la metodología que presentamos aquí es apropiada para obtener secuencias consenso para su uso en estudios de epidemiología molecular de VIH, pero no para el análisis de mutaciones de resistencia. La mayor sensibilidad de NGS para detectar variantes minoritarias y su utilidad clínica han sido estudiadas con detalle[1-4]. NGS proporciona una información muy valiosa respecto de la proporción relativa de una mutación con respecto al total de virus circulantes, información que se perdería al obtener la secuencia consenso.

En resumen, en nuestro trabajo presentamos una metodología que permite generar secuencias consenso que son representativas de la secuencia Sanger para su uso en estudios de epidemiología molecular, siendo necesario efectuar un procesado de las secuencias y utilizar puntos de corte de al menos el 20%.

Bibliografía

1. Liang B, Luo M, Scott-Herridge J, Semeniuk C, Mendoza M, Capina R, et al. A comparison of parallel pyrosequencing and Sanger clone-based sequencing and its impact on the characterization of the genetic diversity of HIV-1. PLoS One. 2011;6:e26745.

2. Pou C, Noguera-Julian M, Pérez-Álvarez S, García F, Delgado R, Dalmau D, et al. Improved prediction of salvage antiretroviral therapy outcomes using ultrasensitive HIV-1 drug resistance testing. Clin Infect Dis. 2014;59:578–88.

3. Simen BB, Simons JF, Hullsiek KH, Novak RM, Macarthur RD, Baxter JD, et al. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. J Infect Dis. 2009;199:93–701.

4. Perez-Parra S, Chueca-Porcuna N, Alvarez-Estevez M, Pasquau J,OmarM, Collado A, et al. Study of human immunodeficiency virus transmission chains in Andalusia: Analysis from baseline antiretroviral resistance sequences. Enferm Infecc Microbiol Clin. 2015;33:603–8.

5. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. J Genet Genomics. 2011;38:95–109.

6. Luk KC, Berg MG, Naccache SN, Kabre B, Federman S, Mbanya D, et al. Utility of metagenomic next-generation sequencing for characterization of HIV and human pegivirus diversity. PLoS One. 2015;10:e0141723.

7. Maddison W.P., Maddison D.R. 2009. Mesquite: A modular system for evolutionary analysis. Version 2.75. Disponible en: http://mesquiteproject.org

8. Lubelchek RJ, Hoehnen SC, Hotton AL, Kincaid SL, Barker DE, French 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;68:46–54.

9. Castro-Nallara E, Pérez-Losada M, Burtonc GF, Crandall KA. The evolution of HIV: Inferences using phylogenetics. Mol Phylogenet Evol. 2012;62:777–92.

10. Hofstra LM, Sauvageot N, Albert J, Alexiev I, García F, Struck D, et al. Transmission of HIV drug resistance and the predicted effect on current first-line regimens in Europe. Clin Infect Dis. 2016;62:655–63.

11. Monge S, Díez M, Alvarez M, Guillot V, Iribarren JA, Palacios R, et al. Use of cohort data to estimate national prevalence of transmitted drug resistance to antiretroviral drugs in Spain (2007-2012). Clin Microbiol Infect. 2015;21:105.e1–5.

12. García F, Pérez-Cachafeiro S, Alvarez M, Pérez-Romero P, Pérez-Elias MJ, Viciana I, et al. Transmission of HIV drug resistance and non-B subtype distribution in the Spanish cohort of antiretroviral treatment naïve HIV-infected individuals (CoRIS). Antiviral Res. 2011;91:150–3.

13. Eshleman SH, Hudelson SE, Redd AD, Wang L, Debes R, Chen YQ, et al. Analysis of genetic linkage of HIV from couples enrolled in the HIV Prevention Trials Network 052 trial. J Infect Dis. 2011;204:1918–26.

14. Pasquier C, Millot N, Njouom R, Sandres K, Cazabat M, Puel J, et al. HIV-1 subtyping using phylogenetic analysis of pol gene sequences. J Virol Methods. 2001;94:45–54.

15. Lutzoni F, Wagner P, Reeb V, Zoller S. Integrating ambiguously aligned regions of DNAsequences in phylogenetic analyses without violating positional homology. Syst Biol. 2000;49:628–51.

Minimizing Next-Generation Sequencing Errors for HIV Drug Resistance Testing

3

AIDS Review. 2017:19(2)

José A. Fernández-Caballero[1], Natalia Chueca[1], Eva Poveda[2], and Federico García[1]

1. Department of Clinical Microbiology, Hospital Universitario San Cecilio de Granada,Instituto de Investigacion Biosanitaria(IBIS),Granada,Spain.
2. Division of Clinical Virology, Instituto de Investigación Biomédica de A Coruña (INIBIC)-Complexo Hospitalario Universitario de A Coruña (CHUAC), Sergas. Universidade da Coruña (UDC), La Coruña,Spain.

Abstract

Next-generation sequencing prototypes for the routine diagnosis of resistance to antiretrovirals approved for the treatment of HIV infection are now being used in many clinical diagnostic laboratories. As some of the next-generation sequencing platforms may be a source of errors, it is necessary to improve the currently available protocols and implement bioinformatic tools that may help to correctly identify the presence of resistance mutations with clinical impact. Several studies have addressed these issues in recent years. Some of them are mainly focused on improving protocols for decreasing the magnitude of errors during the polymerase change reaction. Other studies propose specific bioinformatic tools, able to reach both a 93-98% reduction of indels (insertions/deletions) and a sensitivity and specificity close to 100% in single nucleotide polymorphism variant calling. The implementation of new protocols and bioinformatic tools improving the accuracy of next-generation sequencing results must be considered for a correct analysis of HIV resistance mutations for making clinical decisions. This review summarizes the most relevant data available for the optimization of next-generation sequencing applied to HIV resistance testing.

Key words: HIV. NGS. 454. Denoising. Filter. Indel.

Introduction

Next-generation sequencing (NGS) has revolutionized the studies of genomics characterization applied to virology[1-4]. This technology allows generating a large amount of information in short periods of time and at a relatively low cost[5]. In RNA viruses, such as HIV, high rates of error-prone mutations during viral replication generate a multitude of genetically diverse but similar variants known as quasispecies[6]. The NGS technologies are very useful for the genetic characterization of HIV infection, allowing an in depth analysis of the spectrum of variants present in an HIV-infected patient[7-9]. NGS has a technical limit of detection of 0.01% and a clinical threshold above 1% for mutations present on the viral population compared with the 15-20% obtained using Sanger (population) sequencing. Therefore[10-12], NGS provides both technical and clinical improvements for the detection of resistance to antiretroviral drugs against HIV infection[13]. More specifically, NGS improvements are especially for choosing first-line antiretroviral regimens including drugs with low genetic barrier for resistance[14-16]. This is the case for the non-nucleoside reverse transcriptase inhibitors (NNRTI) (i.e. efavirenz)[17], and potentially also for integrase inhibitors (INI) in the next future. In addition, NGS has also proven to be of value for salvage therapy[14] and as a surrogate of the cumulative resistance in the failing patient.

The 454 GS Junior platform (454 Life Sciences/ Roche Diagnostics) has been widely used across some countries for testing HIV resistance mutations in the clinical setting[18-20] in the last two years. Roche Diagnostics has distributed a prototype allowing sequencing reverse transcriptase (RT), protease and integrase fragments, with an appropriate amplicon length of 300-500 base pairs (bp), which include all resistance associated codons described to date[21].

Although technically sound, 454 technologies have several handicaps. As the method is based on pyrosequencing, homopolymeric regions serve as an error source leading frequently to insertions and deletions (indels)[22]. In addition, the quality of the reads decreases at the end of the reaction[23]. Finally, an additional source of error for HIV resistance analysis is that the protocol is based on a reverse transcribed polymerase chain reaction (RT-PCR). These errors may highly impact the quality of the sequences,

with a significant impact on sequence assembly, polymorphism detection, HIV subtype assignment, and resistance analysis[24].

Most NGS platforms, including 454, incorporate quality control (QC) pipelines in their sequencing protocol to filter the final results. However, as these QC pipelines may be insufficient, it is reasonable to run an own user level additional QC based on high-quality control filtering. This software has been widely used for microbiome analysis based on the 454 platforms, aiming to improve the quality of results[25-28].

To date, there is no systematic review on the use of methods that can correct sequencing errors on the 454 platforms with protocols used for the characterization of HIV resistance in the clinical routine. In this study, we have reviewed all studies dealing with software or methods aiming to decrease these errors, which have been published during the period 2006-2016.

Selection criteria and systematic review

We considered as bioinformatic strategies, software aiming to delete or detect sequencing errors, and as protocol improvements those changes in PCR temperature profiles and/or reagent concentration aiming to minimize sequencing errors.

Our systematic review was performed according to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidance[29]. We used a combination of non-MeSH and MeSH terms related with error correction and NGS sequence filtering. We searched PubMed, Medline (Ovid) and Embase (Ovid). The term “HIV” was combined with “NGS” and “Next generation sequencing” and “Error rates” and “454” and “Artifact recombination” and “Denoising” and “Filter sequence” and “Insertion-deletion” and “Carry forward correction”.

All abstracts of papers available through January 2006 and June 2016 were reviewed. We used an iterative search strategy, as all references in the papers that were selected for reviewing were also studied. We limited our review to original articles that evaluated error correction of HIV NGS sequences obtained on the 454 platforms.

Studies had to provide enough information on how sequence filtering was performed, and which modification before/after processing was performed. Abstracts, comments, and letters to the Editor were excluded as they lacked information on sequence processing.

Data analysis and risk of bias

JA.F reviewed the studies, excluding those with irrelevant titles and references; N.C and JA.F independently read studies that were selected, paying attention to abstracts, descriptive terms, and titles, and identified potentially eligible papers. Both authors finally reviewed the full text of the papers, and applied inclusion criteria. There were no discrepancies among both authors' selection. JA.F and N.C independently extracted the data from the selected studies, providing them in a separate standardized file. Again, no discrepancies between authors were found. When provided, crude 95% confidence intervals (95% CI) were extracted; they were calculated when not available and enough data were provided in the papers. However, due to the high heterogeneity of most studies and different methodologies, meta-analysis could not be performed.

Seven studies were selected, and all were estimated to have no overall risk of bias (supplementary data). Only one study[30] had a high risk of bias, due to the scarce information provided by the authors for all the parameters we evaluated.

JA.F and N.C evaluated the overall quality of the studies, pointing out their strengths and weakness, according to STROBE recommendations[31]. Three categories were given for the quality parameter (high, medium, low). Here, several discrepancies were found between the two reviewers. When this happened, F.G was asked to re-evaluate and discrepancies were solved accordingly.

Selection of the studies

Our search selected 611 studies. After removing all the studies that were not research studies and by the date in which NGS was introduced, 161 papers were removed. After irrelevant studies were removed, only 36 studies remained eligible. From these, 29 papers were removed, 15 because they focused only on detection cut-offs, five because they did not deal with 454 NGS errors, two due to a poor description of the methodology used, and seven for some other reasons (non plasma samples, non-HIV studies, and not showing the results). We finally selected seven papers that met all the eligibility criteria (Fig. 1). Three of them dealt with protocol modifications to avoid or diminish sequencing errors[32-34]. These studies reported data on the number of reads, the percentage of recombination, and error rates (Table 1). The other four studies dealt with bioinformatics aiming to eliminate errors[30,35-37]. In this case, the studies provided data on the number of reads, error rate, indels, single nucleotide polymorphism (SNP) variant calling, and software characteristics (Table 1).

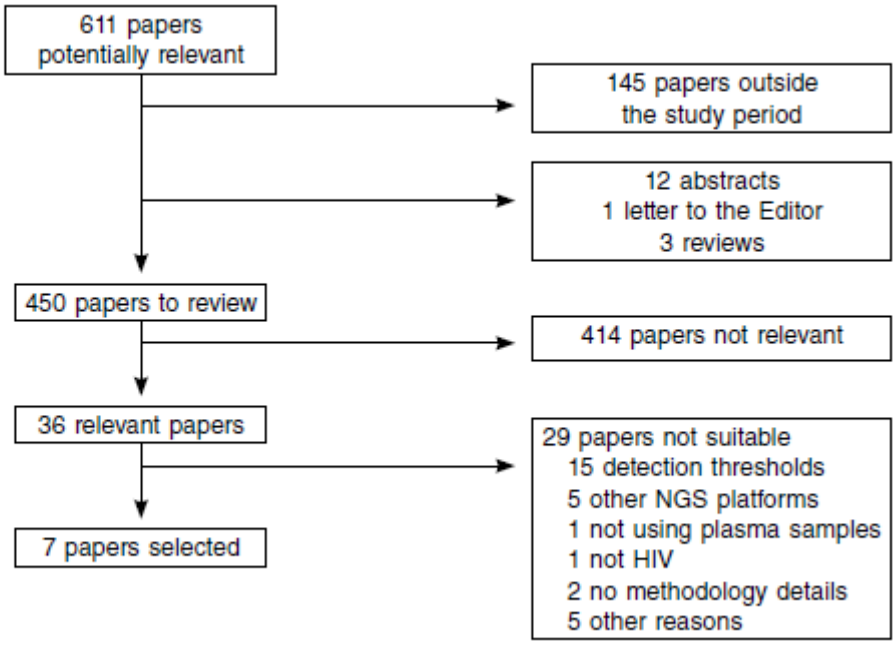


Figure 1. *Flowchart for study selection. NGS: next-generation sequencing.*

Table 1. Studies based on modifications to the amplification protocol or bioinformatics solutions

Study	Year	NGS platform	Region	Design	Procedure
Di Giallonardo, et al.[32]	2013	454 FLX	Protease (271 bp)	Protocol modification; Use of several HIV-1 strains to allow recombination. PCR reagent optimization	1) Standard PCR: 40 cycles (94ºC 15"/55ºC 30"/72ºC 30") plus 72º/8'. Reagents: Primers (uM) 0.4; dNTPs (mM) 0.2; FastStart High Fidelity DNA polymerase (U) 1.25. 2) Optimized PCR: 35 cycles (94ºC 30"/55ºC 60"/72ºC 60"). Primers (uM) 1; dNTPs (mM) 0.4; FastStart High Fidelity DNA polymerase (U) 3
Shao, et al.[33]	2013	454 FLX	RT (pol)	Protocol modification; Use of several HIV-1 strains to allow recombination. PCR reagent optimization	1) Standard PCR: 95ºC 2'; 45 cycles (95ºC 30"/50ºC 30"/72ºC 30"). Reagents; Primers (Nm) 400; dNTPs (uM) 200, $MgSO_4$ (mM) 4; HI Fidelity Platinum Taq (U) 2.5. 2) Optimized PCR: 95ºC 15'; 25 cycles (95ºC 15"/51ºC 30"/68ºC 1'30"). Primers (uM) 1; dNTPs (uM) 200; $MgCl_2$ (mM) 2.3; Taq Gold (U) 5
Waugh, et al.[34]	2015	454 FLX	Gag	Protocol modification; Use of several HIV-1 strains to allow recombination. Changes in PCR cycle number and RNA input	Two PCRs in parallel for each of 3 RNA inputs (ng): 160, 1600, 3990. 1st PCR: 27 cycles; 2nd PCR: 35 cycles; 98ºC 30"/98ºC 10"/72ºC 1'. Primers (Nm) 400; dNTP (uM) 200; Phusion DNA polymerase (U) 0.3
Iyer, et al.[35]	2013	454 FLX	gag/env/nef (1,500 bp/2550/681)	Software: "CorQ". Data learning for massive sequencing data generated from a mixture of strains, and from artificial SNP errors	CorQ that utilizes a multiple sequence alignment to map base qualities to the positions within the alignment. Reads bases according to coverage, quality between adjacent bases, and the base in question
Deng, et al.[36]	2013	454 FLX	gag and pol (269 bp-443 bp)	Software: "Indel and Carryforward Correction (ICC)" Data learning for massive sequencing data generated from a mixture of strains	Unique sequences ranking by their abundance, and align the sequences taking into account the abundances of the aligned unique sequences and scoring parameters; match, mismatch, gap, penalty. Reads are filtered based on parameters; ambiguous bases, length, and average quality
Brodin, et al.[37]	2013	454 FLX	pol (167 bp)	Software: "BioPerl" using error correction scripts. Data learning from env SG3Δ plasmid subjected to nested PCR (30+30 cycles) using Faststart high fidelity	Filtering reads with: less than 80% similarity to a user reference sequence, taking ambiguous nucleotide calls, indels and stop codons into consideration
Kijak, et al.[30]	2013	454 FLX	Pol	Software: "Nautilus"	Using as an input an alignment determines the nucleotide base frequency and read depth at each position and computes the haplotype frequencies based on the linkage among polymorphisms. Also computes the frequency of the variants in the setting of their sequence context and mapping orientation

Bp: base pairs; dNTP: deoxynucleotide; NGS: next-generation sequencing; PCR: polymerase chain reaction; SNP: single nucleoside polymorphism.

Modifications to the amplification protocol

Some previous reports have shown how error rates, as well as chimera and indel formation, increase after 30 cycles of amplification, whatever the fidelity of the polymerase may be[38-40]. Three of the studies we have included in this systematic review aimed to minimize sequencing errors by modifying the amplification protocol[32-34] (Table 2). Two of the studies we have evaluated optimized the PCR by lowering the number of PCR cycles and doubling PCR reagent concentrations, with no variation on RNA input. One of the studies estimated that median (95% CI) recombination was 48.7% (53.6-43.9) for standard PCR, compared to 0.8% (0.07-2.1) after PCR optimization[32]. The error rate for both, standard and optimized PCR, was also analyzed; a drop from 23.2% (95% CI: 18.8-27.6) for standard PCR to 2% (95% CI: 0.6-3.1) was described. A second paper also observed a reduction in median values of recombination from 12% (standard PCR) to 0.8% (optimized PCR)[33]. Finally, the last study used different RNA inputs, changing PCR conditions, but with the same reagent concentrations[34]. A higher number of PCR cycles always resulted in a higher degree of recombination. This study demonstrates that a higher RNA input at first stages of amplification, together with a reduction in the number of cycles, minimizes sequencing errors.

Table 2. Standard vs. optimized polymerase chain reaction: differences in the number of reads and other quality variables

Study	Reads median, (95% CI)		Recombination: %, (95% CI)	Error rate; %, (95% CI)
Di Giallonardo, et al.[32*]	S; 52,389 (84,645-20,133) O; 7,827.5(23,15.6-14,548.7)		S; 48.7(53.6-43.9) O; 0.8 (0.07-2.1)	S; 23.2 (18.8-27.6) O; 2 (0.6-3.1)
Shao, et al.[33]	S; 12,2327 O; 62,437		S; 12 O; 0.8	
Waugh, et al.[34]		27 cycles	RNA 160 ng 0.006 (0-0.01) RNA 1600 ng 0.1 (0.08-0.1) RNA 3,990 ng 0.9 (0.4-1.8)	
		35 cycles	RNA 160 ng 2.9 (2.8-3.2) RNA 1,600 ng 4.6 (4.5-4.8) RNA 3,990 ng 1.1 (0.8-1.4)	

*Not provided in the study (calculated).
O: optimized; PCR: polymerase chain reaction; S: standard.

Bioinformatics-based solutions

Four studies evaluated the use of bioinformatics to locate and diminish/eliminate sequencing errors[30,35-37] (Table 3). Two of them tested for different variables before and after they were used, and describe less numbers of errors after their usage[36,37]. Deng, et al., used the indel and Carryforward Correction (ICC) software, finding that error rates decreased from a median value (95% CI) of 0.3% (0-1) for both gag and pol regions, to 0.02% (0-0.04) for gag and 0.01% (0-0.04), after using ICC[36]. Indels were reduced in 98-99%, and sensitivity and specificity for SNP variant calling was 100 and 98%, respectively. Brodin, et al., used BioPerl for sequence processing: error rates decreased from a median value (95% CI) of 0.2% (0.008-0.4) before processing to 0.06% (0.05-0.08); using the software resulted in the filtering of approximately 2,000 reads per sample (from 8,749.5 reads [95% CI: 4,137.1-11,760.5], to 5,394 reads (95% CI: 1,958.4- 8,951.6])37. Iyer, et al. used CorQ that resulted in a reduction on indels of 93-97% for gag/env/nef regions, with a sensitivity and specificity of 99 and 88%, respectively, for SNP variant calling[35]. The last study, performed by Kijak, et al., used Nautilus software, which is an observational tool that helps to discern between errors; in contrast to the other three, this study did not show any data on the parameters that were analyzed[30].

Table 3. Main characteristics of bioinformatics tools available to eliminate next-generation sequencing errors

Study	Reads median, (95% CI)	Error rate; %, (95% CI)		Indel	SNP variant calling	Software characteristics					
		Prior to filtering	Post filtering			Chim	Trim	Var L	(n)	Qual.	k
Iyer, et al.[35*]	26,620 gag 48,927 env 21,963 nef			Reduction 93-97%	Sensitivity: 99% Specificity: 88%	√	?	√	√	√	√
Deng, et al.[36*]	12,617 (8,247-16,987) gag 17,228.5 (12,001-22,456) pol	0.3 (0-1)	gag; 0.02 (0-0.04) pol; 0.01 (0-0.04)	Reduction 98-99%	Sensitivity: 100% Specificity: 98%	√	√	√	√	√	√
Brodin, et al.[37*]	Prior to filtering; 8,749.5 (4,137.1-11,760.5) Post filtering; 5,394 (1,958.4-8,951.6)	0.2 (0.08-0.4)	0.06 (0.05-0.08)			?	?	?	?	?	?
Kijak, et al.[30]						?	?	?	?	?	?

*Not provided in the study (calculated).
Chim: chimeric; k: k-mer; n: number of reads; trim: trimming; qual: quality scores; SNP: single-nucleotide polymorphism; var L: variation lengths.

Overall, bioinformatics software have shown to be highly efficacious to remove sequences that contained sequencing errors, with a 93-99% reduction in indels, which is highly relevant, specially for 454-based NGS[35-37]. In a similar way, using these programs also decreased error rate, allowing a more accurate estimation of variants, especially when they are present at a low relative proportion (1-5%). Filtering may result in a significant loss in the number of reads, and may enable NGS to go down to 1% for minor variant detection. All the software that have been reviewed in this paper did not incur a high loss in the number of reads; therefore, this was not seen to be a problem. All bioinformatics software also showed excellent results for SNP variant calling, with sensitivity and specificity values near 100%.

The GS Reference Mapper in another tool for bioinformatic analysis. This software filters the sequences based on quality parameters, length of the sequences, expected deep and variant coverage. The GS Reference Mapper allowed improving the quality of sequences, especially at their terminal end, improving the Q values from 10 to 25 (Table 4). The Q value is an index of the probability of a given base of being a sequencing error; Q values > 30 are optimal, as the probability error is between 1 and 1000[41]. This software has been used for the analysis of 73 samples from HIV-1-infected patients (84.2% subtype B) that were sequenced (RT and protease) on a 454 GS Junior. After denoising, most of the resistance mutations that were corrected by GS Reference Mapper (43/68) had been detected at very low relative prevalence (1-2%), and all were below 5%. In addition, after filtering, only a small number of sequences are deleted (mean 95% CI: 30; 22-37), and it shows a high sensitivity (99%) and specificity (98%) for SNP variant calling. This software was able to eliminate HIV resistance mutations that were artificially detected at low levels, ranging 1-5% of the whole quasispecies population that was infecting the patients.

Table 4. GS Reference Mapper characteristics

Region	Reads median (95% CI)		Error rate; %, (95% CI)		Indel	SNP variant calling	Software characteristics					
	Prior to filtering	Post filtering	Prior to filtering	Post filtering			Chim	Trim	Var L	(n)	Qual	k
pol (533 bp)	7,214 (5,321-9,106)	7,184 (5,299-9,069)	0.3 (0.18-0.42)	0.03 (0.02-0.05)	Reduction 99%	Sensitivity: 99% Specificity: 98%	√	√	√	√	√	√

Bp: base pair; Chim: chimeric; k: k-mer; n: number of reads; trim: trimming; qual: quality scores; SNP: single-nucleotide polymorphism; var L: variation lengths.

Recommendations and conclusions

This is the first systematic review evaluating the benefits of sequence processing after massive parallel sequencing for the characterization of HIV drug resistance. There are specific technical recommendations and several bioinformatics software that can be very useful to improve the results obtained using 454 GS Junior NGS. The recommendations proposed in this review will certainly help NGS users to be really confident that HIV resistance mutations detected using 454 protocols are true mutations and not test-associated artifacts, especially if they are at low proportions (1- 5%) in the whole viral population. Implementing these recommendations will certainly be of benefit and will improve patient care.

Considering all these data, there are some specific technical recommendations that might help for the minimization of the errors generated using the 454 GS Junior NGS platform (Table 5). Some of them include; lowering the number of cycles during PCR, using high-fidelity polymerases (Phusion High-Fidelity DNA Polymerase or pfu DNA polymerase) and using bioinformatic software to filter short and low quality sequences, including all bases with a Q value < 25, and sequence depth 10, might help for the minimization of the errors generated using the 454 GS Junior NGS platform.

Table 5. Technical recommendations to minimize errors

What	How
Lower the number of cycles during PCR	Use 5-7 cycles less
Use high fidelity polymerases	Phusion High-Fidelity DNA Polymerase or pfu DNA polymerase
Use bioinformatic software to filter short and low quality sequences	GS Reference Mapper or CorQ
Filter all bases with a Q value < 25, and sequence depth 10	Delete indels that can affect the final results

PCR: polymerase chain reaction.

References

1. Visser M, Bester R, Burger JT, Maree HJ. Next-generation sequencing for virus detection: covering all the bases. Virol J. 2016;13:85.

2. McGinnis J, Laplante J, Shudt M, George KS. Next generation sequencing for whole genome analysis and surveillance of influenza A viruses. J Clin Virol. 2016;79:44-50.

3. Thorburn F, Bennett S, Mosha S, Murdoch D, Gunson R, Murcia PR. The use of next generation sequencing in the diagnosis and typing of respiratory infections. J Clin Virol. 2015;69:96-100.

4. Quan PL, Wagner TA, Briese T, et al. Astrovirus encephalitis in boy with X-linked agammaglobulinemia. Emerg Infect Dis. 2010;16:918-25.

5. Dudley DM, Chin EN, Bimber BN, et al. Low-cost ultra-wide genotyping using Roche/454 pyrosequencing for surveillance of HIV drug resistance. PLoS One. 2012;7:e36494.

6. Biebricher CK, Eigen M. What is a quasispecies? Curr Top Microbiol Immunol. 2006;299:1-31.

7. Zukurov JP, do Nascimento-Brito S, Volpini AC, Oliveira GC, Janini LM, Antoneli F. Estimation of genetic diversity in viral populations from next generation sequencing data with extremely deep coverage. Algorithms Mol Biol. 2016;11:2.

8. Nicot F, Sauné K, Raymond S, et al. Minority resistant HIV-1 variants and the response to first-line NNRTI therapy. J Clin Virol. 2015;62:20-4.

9. Fisher RG, Smith DM, Murrel B et al. Next generation sequencing improves detection of drug resistance mutations in infants after PMTCT failure. J Clin Virol. 2015;62:48-53.

10. Ram D, Leshkowitz D, Gonzalez D, et al. Evaluation of GS Junior and MiSeq next-generation sequencing technologies as an alternative to Trugene population sequencing in the clinical HIV laboratory. J Virol Methods. 2015;212:12-16.

11. Mohamed S, Penaranda G, Gonzalez D, et al. Comparison of ultra-deep versus Sanger sequencing detection of minority mutations on the HIV-1 drug resistance interpretations after virological failure. AIDS. 2014;28:1315-24.

12. Stelzl E, Pröll J, Bizon B, et al. Human immunodeficiency virus type 1 drug resistance testing: Evaluation of a new ultra-deep sequencingbased protocol and comparison with the TRUGENE HIV-1 Genotyping kit. J Virol Methods. 2011;178:94-7.

13. Liang B, Luo M, Scott-Herridge J, et al. A comparison of parallel pyrosequencing and sanger clone-based sequencing and its impact on the characterization of the genetic diversity of HIV-1. PLoS One. 2011;6:e26745.

14. Pou C, Noguera-Julian M, Pérez-Álvarez S, et al. Improved prediction of salvage antirretroviral therapy outcomes using ultrasensitive HIV-1 drug resistance testing. Clin Infect Dis. 2014;59:578-88.

15. Simen BB, Simons JF, Hullsiek KH, et al. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naïve patients significantly impact treatment outcomes. J Infect Dis. 2009;199:693-701.

16. Messiaen P, Verhofstede C, Vandenbroucke I, et al. Ultra-deep sequencing of HIV-1 reverse transcriptase before start of an NNRTI based regimen in treatment-naive patients. Virology. 2012;426;7-11.

17. Wang J, Zhang G, Bambara RA, et al. Nonnucleoside reverse transcriptase inhibitor-resistant HIV is stimulated by efavirenz during early stages of infection. J Virol. 2011;85:10861-73.

18. Chen X, Zou X, He J, Zheng J, Chiarella J, Kozal MJ. HIV drug resistance mutations (DRMs) detected by deep sequencing in virologic failure subjects on therapy from Hunan Province, China. PLoS One. 2016;11:e0149215.

19. Fernández-Caballero JÁ, Chueca N, Alvarez M, et al. Usefulness of Integrase resistance testing in proviral HIV-1 DNA in patients with Raltegravir prior failure. BMC Infect Dis. 2016;16:197.

20. Dauwe K, Staelens D, Vancoillie L, Mortier V, Verhofstede C. Deep sequencing of HIV-1 RNA and DNA in newly diagnosed patients with baseline drug resistance showed no indications for hidden resistance and is biased by strong interference of hypermutation. J Clin Microbiol. 2016;54:1605-15.

21. Gall A, Ferns B, Morris C, et al. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. J Clin Microbiol. 2012;50:3838-44.

22. Loman NJ, Misra RV, Dallman TJ, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012;30:434-9.

23. Patel RK, Jain M. NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. Plos One. 2012;7:e30619.

24. Fernández-Caballero Rico JA, Chueca Porcuna N, Alvarez Estévez M, Mosquera Gutiérrez MD, Marcos Maeso MA, García F. [A safe and easy method for building consensus HIV sequences from 454 massively parallel sequencing data]. Enferm Infecc Microbiol Clin. [Epub ahead of print].

25. Reeder J, Knight R. Rapid denoising of pyrosequencing amplicon data: exploiting the rank-abundance distribution. Nat Methods. 2010; 7:668-9.

26. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D. Experimental and analytical tools for studying the human microbiome. Nat Rev Genet. 2012;13:47-58.

27. Gibson J, Shokralla S, Porter TM, et al. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proc Natl Acad Sci USA. 2014; 111:8007-12.

28. Aagaard K, Riehle K, Ma J, et al. A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. PLoS One. 2012;7:e36466.

29. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. 2015;4:1.

30. Kijak GH, Pham P, Sanders-Buell E, et al. Nautilus: a bioinformatics package for the analysis of HIV type 1 targeted deep sequencing data. AIDS Res Hum Retroviruses. 2013;29:1361-4.

31. Von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Int J Surg. 2014;12:1495-9.

32. Di Giallonardo F, Zagordi O, Duport Y, et al. Next-generation sequencing of HIV-1 RNA genomes: determination of error rates and minimizing artificial recombination. Plos One. 2013;8:e74249.

33. Shao W, Boltz VF, Spindler JE, 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. Retrovirology. 2013;10:18.

34. Waugh C, Cromer D, Grimm A, et al. A general method to eliminate laboratory induced recombinants during massive, parallel sequencing of cDNA library. Virol J. 2015;12:55.

35. Iyer S, Bouzek H, Deng W, Larsen B, Casey E, Mullins JI. Quality score based identification and correction of pyrosequencing errors. Plos One. 2013;8:e73015.

36. Deng W, Maust BS, Westfall DH, et al. Indel and Carryforward Correction (ICC): a new analysis approach for processing 454 pyrosequencing data. Bioinformatics. 2013;29:2402-9.

37. Brodin J, Mild M, Hedskog C, et al. PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. Plos One. 2013;8:e70338.

38. Smyth RP, Schlub TE, Grimm A et al. Reducing chimera formation during PCR amplification to ensure accurate genoryping. Gene. 2010;469:45-51.

39. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl Environ Microbiol. 2005;71:8966-9.

40. Sipos R, Szekely AJ, Palatinszky M, Revesz S, Marialigeti K, Nikolausz M. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiol Ecol. 2007;60:341-50.

41. Brockman W, Alvarez P, Young S, et al. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res. 2008;18:763-70.

DISCUSIÓN GENERAL

Cuando se produce el fracaso terapéutico, el virus puede desarrollar resistencia a alguno de los fármacos antirretrovirales y se impone la necesidad de un cambio de terapia. Esto puede ocurrir de forma natural durante el curso de la terapia, puesto que la infección por el VIH es muy productiva y como ya sabemos tiene una alta tasa de mutación, o por la aparición de resistencias debidas a la falta de adherencia al tratamiento por parte del paciente, lo cual facilita que el virus que ya ha estado sometido a la presión por el fármaco las desarrolle [96,97]. En los casos en los que esto ocurre se puede observar como al abandonar totalmente el tratamiento la situación de resistencia del virus revierte, presentando en ese momento un perfil de resistencias sensible al fármaco en cuestión. Pero lo que ocurre en realidad es que el virus salvaje (sin mutaciones) vuelve a surgir al retirar la presión selectiva y se impone puesto que tiene un mayor *fitness* [49]. Aunque no podamos (en dicho momento) detectar los mutantes resistentes, no significa que hayan desaparecido por completo, si no que se encuentran acantonados en el interior de las células que infecta el virus y en mucha menor medida en el plasma sanguíneo (solo detectables con técnicas de secuenciación que puedan detectar variantes víricas en proporciones de alrededor de un 1%).

En el caso de esta tesis doctoral nos hemos centrado en el fármaco DTG, puede servir como quimioterápico de rescate ante situaciones de fracaso de RAL. En los ensayos clínicos, se han observado tres vías de adquisición de mutaciones primarias que confieren un alto nivel de resistencia a RAL:

1) N155H en combinación con L74M, E92Q o G163R
2) Q148H/R/K con E138K o G140S/A
3) Y143R/C más otras mutaciones

El estudio VIKING evaluó la dosificación de DTG en una toma única o dos tomas al día [98], presentando mayor nivel de eficacia, tolerabilidad y seguridad la toma dos veces al día DTG 50 mg [42], en los casos de pacientes en los que se encontraron mutaciones de resistencia. Por esta razón, la ficha técnica de DTG sólo recomienda usar DTG 50 mg una vez al día en aquellos pacientes en los que no se documenta resistencia previa a INSTIs. El estudio VIKING llega a la conclusión de que la mutación mayor Q148 en conjunción con otras secundarias confiere resistencia frente a ambos fármacos. Sin embargo la presencia de la mutación N155 determina resistencia frente a RAL, en

cambio, resistencia parcial frente a DTG [43] (en este caso sería posible su uso en dos tomas diarias en lugar de la habitual toma única).

La mutación N155H suele ser la primera mutación en aparecer en pacientes bajo tratamiento con RAL, pero, bajo continuada presión farmacológica es reemplazada por mutaciones de la ruta Q148 [99]. Respecto a los polimorfismos, se observo un mayor número de estos en la muestra de fracaso (correspondiente a muestra de plasma) frente a la muestra de ADN proviral, además de encontrarse mutaciones secundarias en ADN proviral no presentes en muestra de plasma. Este hecho no es de extrañar, debido al ciclo biológico del VIH, ya que después de replicarse en la célula saldrían al torrente sanguíneo, presentando aquí miles de partículas con distintos polimorfismos, por esta razón. Por el contrario, efectuando un estudio de ADN proviral, observaríamos las mutaciones que han quedado integradas en el ADN de la célula, siendo estas menores.

Por otro lado, existen dos estudios [32, 100] que demuestran que la vida media de los reservorios latentes virales es de 4 a 6 meses en los pacientes que inician la terapia en la fase aguda de la infección, otros estudios enfocados en sujetos en la fase crónica de la infección (siempre con terapia antirretroviral) colocan la vida media de los reservorios virales en 44 meses [101, 102]. En nuestro estudio, varios pacientes excedían del tiempo de vida medio de los reservorios, pero esto no interfirió en la concordancia de mutaciones entre la muestra en el momento del fracaso virológico y la muestra ADN proviral. Para un paciente no fue posible detectar mutaciones primarias mediante secuenciación poblacional y hubo que recurrir a técnicas NGS del ADN proviral. Estas mutaciones no se detectaron en este paciente mediante secuenciación poblacional debido a que estaban por debajo de su umbral de detección [103]; N155H (9,7%) y T97A (12,42%).

La plataforma 454 GS Junior ha sido adoptada por muchos laboratorios en la investigación VIH, sustituyendo a la secuenciación Sanger como el principal método utilizado. 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 [104-105], 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.

Esta limitación se intenta solventar con la creación de una única secuencia consenso a partir de las secuencias generadas por un mismo paciente. Algunos estudios tienden a generar estas secuencias mediante comandos informáticos complejos. En este caso la utilización de MESQUITE [106] es intuitiva, de fácil manejo y sin necesidad de comandos, simplificando la obtención de la secuencia consenso. La ambigüedad en las bases nucleotídicas de la secuencia consenso vendrá dada por el umbral de corte. 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 [107] que la tecnología 454 Roche genera ciertos errores en el proceso de amplificación, pudiendo trasladarse a la secuencia consenso creada.

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 [108-109]. 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 [110,111]. Hoy en día no existe en el software MEGA algoritmos de procesamiento que tengan en cuenta las bases ambiguas [112]. 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 ve 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.

En este doctorado presentamos la primera revisión sistemática que evalúa los beneficios del procesado de secuencias en pacientes VIH, mediante secuenciación masiva para la eliminación o disminución de posibles errores.

La introducción de las técnicas NGS hace necesaria la utilización de herramientas bioinformáticas para el correcto manejo de secuencias, de otro modo pueden aportar resultados equivocados o distorsionados. Actualmente no existe un protocolo para manejar datos de pacientes VIH, el cual elimine las secuencias defectuosas obtenidas, sin embargo, existen protocolos de estudios basados en microbiota [83-86] a partir de los cuales eliminan el ruido, quimeras y posibles errores que puedan existir en el conjunto de datos, efectuando un filtrado de secuencias mediante calidad y longitud de secuencias, contando con multitud de programas para tal propósito (QIIME, MG-RAST, etc).

El uso de programas bioinformáticos han demostrado alta eficacia en la eliminación de secuencias con errores, obteniendo una reducción en el porcentaje de *indel* del 93-99%, este dato tiene relevancia, ya que la plataforma de secuenciación 454 GS-Junior presenta una alta tasa de error *indel* [90-92]. En cuanto a la corrección de tasa de error se consigue disminuir en más de una unidad, lo que nos permite asegurar con mas fiabilidad el diagnostico clínico, sobre todo en aquellas mutaciones minoritarias que presentan una prevalencia muy baja (1-5%).

Los cambios en el protocolo de amplificación [107, 113, 114] son fundamentalmente la disminución en el número de ciclos de programa PCR y el aumento de concentraciones de reactivos para la PCR, de esta forma se consigue disminuir sustancialmente la tasa de error y recombinación. Varios estudios [115-117] han demostrado que tras 30 ciclos de amplificación PCR se produce un aumento en la tasa de error, producción de quimeras e *indel*, independientemente de la enzima utilizada.

Teniendo en cuenta la revisión realizada podemos minimizar la generación de errores en 454 GS Junior mediante las siguientes recomendaciones; 1) Acortar los ciclo de amplificación PCR, ya que al aumentar los ciclos estamos aumentando el riesgo de errores o bias. 2) Utilización de enzima polimerasa que mantenga una alta fidelidad. 3) Utilización de programas bioinformáticos para el filtrado de secuencias que no cumplan

una buena calidad y longitud. Nuestras recomendaciones es que se utilicen programas bioinformáticos que tengan en cuenta la calidad según cada base, con profundidad de secuenciación (ej: GS Reference Mapper, CorQ) y no la calidad global de la secuencia. Nosotros recomendamos una calidad de base > Q25 y una profundidad de secuencias >10, además de eliminación de secuencias que no cumplan con la longitud deseada del amplicon a estudiar, ya que estas contienen *indel* que pueden interferir con los resultados finales.

CONCLUSIONES

1) La investigación de mutaciones en la Integrasa en ADN proviral, en pacientes indetectables que fracasaron previamente a INSTIs de primera generación, en los que no se dispone de estudio de resistencias en el momento del fracaso, puede resultar de utilidad para la toma de decisiones en la pauta de dosificación de Dolutegravir.

2) Para el estudio de resistencias en la Integrasa viral, el ADN proviral de una muestra actual es un fiel reflejo de lo que ocurrió en el fracaso.

3) Las secuencias obtenidas por NGS son aptas para estudios de epidemiología molecular, siendo necesario efectuar un pre-procesado de las secuencias y utilizar puntos de corte de al menos el 20%.

4) La utilización de programas bioinformáticos u optimización de protocolo de amplificación disminuyen los posibles errores en las secuencias obtenidas, disminuyendo la incertidumbre de mutaciones relevantes de baja prevalencia (1%-5%) y mejorando los posteriores análisis moleculares.

5) La implantación de un flujo de trabajo para la eliminación de posibles errores en la obtención de secuencias NGS es necesaria. De lo contrario podrían verse afectados los resultados de subtipado, filogenia y los datos de resistencias.

BIBLIOGRAFÍA

[1] Centers for disease control. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men New York City and California. Morb Mortal Wkly Rep. 1981: 30(25);305-8.

[2] 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);868-71.

[3] Ascher MS, Sheppard HW, Winkelstein W Jr, Vittinghoff E. Does drug use cause AIDS? Nature. 1993:362(6416);103-4.

[4] Gallo RC, Sarin PS, Gelmann EP, Robert-Guroff M, Richardson E, Kalyanarman VS et al. Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). Science 1983:220(4599);865-67.

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

[6] Gallo RC, Montagnier L. The discovery of VIH as the cause of AIDS. N ENGL J MED. 2003:349(24);2283-5.

[7] Cohen M, Hellmann N, Levy J, DeCock K, Lange J. The spread, treatment, and prevention of VIH-1: evolution of a global pandemic. J Clin Invest. 2008:118(4);1244-54.

[8] www.who.int/VIH/data/en/ Accedido el 01/06/17

[9] Alcamí J. Avances en la inmunopatología de la infección por el VIH. Enferm Infecc Microbiol Clin. 2004:22(8);486-96.

[10] www.fundacionio.blogspot.com.es/2013/05/descrubriendo-al-vih-el-genoma.html Accedido el 01/06/17.

[11] 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. 1997:11(13);1565-74.

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

[13] Ferrer P, Rodríguez C, Tordecilla R, Guzmán M, Afani A. Antagonistas de CCR5 en la infección por virus de inmunodeficiencia humana (VIH): aspectos generales y tropismo viral . Rev Hosp Clín Univ Chile. 2012:23;346-53.

[14] Abbas AK, Lichtman AH. & Pillai S. Cap. 20: Inmunodeficiencias congénitas y adquiridas. Inmunología celular y molecular. Travessera de Gràcia, Barcelona, España: Elsevier España, S.L. 2008. 463-489. ISBN: 978-84-8086-311-7.

[15] Carter J, Saunders V. Cap. 16: Retroviruses. Virology principles and applications. The Atrium, Southern Gate, Chichester, West Sussex, England: John Wiley and Sons, Ltd. 2007. 185-196. ISBN: 978-0-470-02386-0.

[16] Temin HM. Homology between RNA from Rous Sarcoma Virus and DNA from Rous Sarcoma Virus-Infected Cells. PNAS. 1964:52;323-9.

[17] Schwartzberg P, Colicelli J, Go SP. Construction and analysis of deletion mutations in the pol gene of moloney murine leukemia virus: A new viral function required for productive infection. Cell. 1984:37;1043-52.

[18] Delelis O, Carayon K, Saïb A, Deprez E, Mouscadet JF. Integrase and integration: biochemical activities of HIV-1 integrase. Retrovirology. 2008:5;114-26.

[19] Lee SP, Xiao J, Knutson JR, Lewis MS, Han MK. Zn^{2+} promotes the self-association of human immunodeficiency virus type-1 integrase *in vitro*. Biochemistry. 1997:36;173-80.

[20] Marshall HM, Ronen K, Berry C, Llano M, Sutherland H, Saenz D et al. Role of PSIP/LEDGF/p75 in lentiviral infectivity and integration targeting. PLoS One. 2007:2;e1340.

[21] Grandgenett DP. Symmetrical recognition of cellular DNA target sequences during retroviral integration. Proc Natl Acad Sci USA. 2005:102;5903-4.

[22] Pennings PS. HIV drug resistance: problems and perspectives. Inf. Dis. Rep. 2013:5;s1e5.

[23] Buendia P, Cadwallader B, DeGruttola V. A phylogenetic and Markov model approach for the reconstruction of mutational pathways of drug resistance. Bioinformatics. 2009:25(19);2522-9.

[24] Abram ME, Ferris AL, Shao W, Alvord WG, Hughes SH. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. J Virol. 2010:84(19);9864-78.

[25] Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG *et al*. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci USA. 2008:105(21);7552-7.

[26] Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. Science. 1996:271(5255);1582-6.

[27] Levy DN, Aldrovandi GM, Kutsch O, Shaw GM. Dynamics of HIV-1 recombination in its natural target cells. Proc Natl Acad Sci USA. 2004:101 (12);4204-9.

[28] Tang MW, Shafer RW. HIV-1 antiretroviral resistance; scientific principles and clinical applications. Drugs. 2012:72(9);e1-e25.

[29] Clutter DS, Jordan MR, Bertagnolio S, Shafer RW. HIV-1 drug resistance and resistance testing. Infect Genet Evol. 2016:46;292-307.

[30] Hare S, Gupta SS, Valkov E, Engelman A, Cherepanov P .Retroviral intasome assembly and inhibition of DNA strand transfer. Nature. 2010:464(7286);232-6.

[31] Hazuda DJ, Felock P, Witmer M, Wolfe A, Stillmock K, Grobler JA et al. Inhibitors of strand transfer that prevent integration and inhibit HIV-1 replication in cells. Science. 2000:287;646-50.

[32] Malet I, Delelis O, Valantin MA, Montes B, Soulie C, Wirden M,et al. Mutations associated with failure of raltegravir treatment affect integrase sensitivity to the inhibitor *in vitro*. Antimicrob Agents Chemother. 2008:52;1351-8.

[33] Goethals O, Clayton R, Van Ginderen M, Vereycken I, Wagemans E, Geluykens P et al. Mutations in human immunodeficiency virus type 1 integrase selected with elvitegravir confer reduced susceptibility to a wide range of integrase inhibitors. J virol. 2008:82(21);10366-74.

[34] Marinello J, Marchand C, Mott BT, Bain A, Thomas CJ, Pommier Y. Comparison of raltegravir and elvitegravir on HIV-1 integrase catalytic reactions and on a series of drug-resistant integrase mutants. Biochemistry. 2008:47(36);9345-54.

[35] Metifiot M, Vandegraaff N, Maddali K, Naumova A, Zhang X, Rhodes D et al. Elvitegravir overcomes resistance to raltegravir induced by integrase mutation Y143. AIDS. 2011:25(9);1175-8.

[36] Blanco JL, Varghese V, Rhee SY, Gatell JM, Shafer RW. HIV-1 integrase inhibitor resistance and its clinical implications. J Infect Dis. 2011:203;1204-14.

[37] Abram ME, Hluhanich RM, Goodman DD, Andreatta KN, Margot NA Ye L et al. Impact of primary elvitegravir resistance-associated mutations in HIV-1 integrase on drug susceptibility and viral replication fitness. Antimicrob Agents Chemother. 2013:57;2654-63.

[38] Cooper DA, Steigbigel RT, Gatell JM, Rockstroh JK, Katlama C, Yeni P et al. Subgroup and resistance analyses of raltegravir for resistant HIV-1 infection. N Engl J Med. 2008:59;355-65.

[39] Goethals O, Van Ginderen M, Vos A, Cummings MD, Van Der Borght K, Van Wesenbeeck L et al. Resistance to raltegravir highlights integrase mutations at codon 148 in conferring cross-resistance to a second-generation HIV-1 integrase inhibitor. Antiviral Res. 2011:91;167-76.

[40] Kobayashi M, Yoshinaga T, Seki T, Wakasa-Morimoto C, Brown KW, Ferris R et al. In vitro antiretroviral properties of S/GSK1349572, a next- generation HIV integrase inhibitor. Antimicrob Agents Chemother. 2011: 55;813-21.

[41] Hightower KE, Wang R, Deanda F, Johns BA, Weaver K, Shen Y et al. Dolutegravir (S/GSK1349572) exhibits significantly slower dissociation than raltegravir and elvitegravir from wild-type and integrase inhibitor-resistant HIV-1 integrase-DNA complexes. Antimicrob Agents Chemother. 2011:55(10);4552-9.

[42] Eron JJ, Clotet B, Durant J, Katlama C, Kumar P, Lazzarin A et al.Safety and efficacy of dolutegravir in treatment-experienced subjects with raltegravir-resistant HIV type 1 infection: 24-week results of the VIKING Study. J Infect Dis. 2013:207(5);740-8.

[43] Mesplède T, Wainberg MA. Integrase strand transfer inhibitors in HIV therapy. Infect Dis Ther. 2013:2;83-93.

[44] DeAnda F, Hightower KE, Nolte RT, Hattori K, Yoshinaga T, Kawasuji T et al. Dolutegravir interactions with HIV-1 integrase-DNA: structural rationale for drug resistance and dissociation kinetics. PLoS One. 2013:8(10);e77448.

[45] Domingo E, Martin V, Perales C, Grande-Pérez A, García-Arriaza J, Arias A. Viruses as quasispecies: biological implications. Curr Top Microbiol Immunol. 2006: 299;51-82.

[46] Quiñones-Mateu M. Is HIV-1 evolving to a less virulent (pathogenic) virus? AIDS. 2005:19;1689-90.

[47] Jose Maximiliano Medina Ramírez. Búsqueda de respuesta humoral neutralizante en pacientes VIH-1 con niveles indetectables de viremia. Tesis doctoral. 2012.

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

[49] Hungnes O, Jonassen TO, Jonassen CM, Grinde B. Molecular epidemiology of viral infections. APMIS. 2000:108(2);81-97.

[50] Hall BG, Barlow M. Phylogenetic analysis as a tool in molecular epidemiology of infectious diseases. Ann Epidemiol. 2006:16;157-69.

[51] Larkin MA., Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al*. Clustal W and Clustal X version 2.0. Bioinformatics. 2007:23;2947-8.

[52] Baldauf, S. Phylogeny for the faint of heart: a tutorial. Trends Genet. 2003:19(6);345-51.

[53] Stamatakis A, Hoover P, Rougemont, J. A Rapid Bootstrap Algorithm for the RAxML Web-Servers. Syst Biol. 2008:75(5);758-71.

[54] Steel M, Penny D. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol Biol Evol. 2000:17(6);839-50.

[55] Nei M, Kumar S. Molecular Evolution and Phylogenetics. Ed. 2000: Oxford University Press. Cap 8.

[56] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol, 1981:17(6);368-76.

[57] Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. 1998.Cambridge University Press.

[58] Gibson RM, Schmotzer CL, Quiñones-Mateu ME. Next Generation Sequencing to Help Monitor Patients Infected with HIV: Ready for Clinical Use? Curr Infect Dis Rep. 2014:16(4);401.

[59] Quiñones-Mateu ME, Avila S, Reyes Teran G, Martinez MA. Deep Sequencing: Becoming a Critical Tool in Clinical Virology.J Clin Virol. 2014:61(1);9–19.

[60] Gianella S, Delport W, Pacold ME, Young JA, Choi JY, Little SJ et al. Detection of Minority Resistance during Early HIV-1 Infection: Natural Variation and Spurious Detection rather than Transmission and Evolution of Multiple Viral Variants. J Virol. 2011:85(16);8359-67.

[61] Pacold M, Smith D, Little S, Cheng PM, Jordan P, Ignacio C et al .Comparison of Methods to Detect HIV Dual Infection.AIDS Res Hum Retroviruses. 2010:26(12);1291-8.

[62] Liang B, Luo M, Scott-Herridge J, Semeniuk C, Mendoza M. A Comparison of Parallel Pyrosequencing and Sanger Clone-Based Sequencing and Its Impact on the Characterization of the Genetic Diversity of HIV-1.PLoS One. 2011:6(10);e26745.

[63] Zagordi O, Klein R, Daümer M, Beerenwinkel N. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Res.2010:38(21):7400-9.

[64] Mohamed S, Penaranda G, Gonzalez D, Camus C, Khiri H, Boulmé R et al. Comparison of ultra-deep versus Sanger sequencing detection of minority mutations on

the HIV-1 drug resistance interpretations after virological failure. AIDS. 2014:28(9);1315-24.

[65] Shendure J, Ji H. Next-generation ADN sequencing. Nat biotechnol.2008:26(10);1135-45.

[66] Watson SJ, Welkers MR, Depledge DP, Coulter E, Breuer JM, Jong MD. Viral population analysis and minority-variant detection using short read next-generation sequencing. Philos Trans R Soc Lond B Biol Sci. 2013:368(1614);20120205.

[67] Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012:30(5);434-9.

[68] Zhang J, Chiodini R, Badr A, Zhang G. The impact of next generation sequencing on genomics. J Genet Genomics. 2011:38(3);95–109.

[69] Brauer MJ, Holder MT, Dries LA, Zwickl DJ, Lewis PO, Hillis DM. Genetic Algorithms and Parallel Processing in Maximum-Likelihood Phylogeny Inference. Mol Biol Evol. 2002:19(10);1717–26.

[70] Löytynoja A, Vilella AJ, Goldman N. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. Bioinformatics.2012:28(13);1684-91

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

[72] Poon AF, Swenson LC, Dong WW, Deng W, Kosakovsky Pond SL, Brumme ZL et al. Phylogenetic Analysis of Population-Based and Deep Sequencing Data to Identify Coevolving Sites in the *nef* Gene of HIV-1.Mol. Biol. Evol. 2010:27(4);819-32.

[73] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977:74(12);5463-7.

[74] Zagordi O, Klein R, Daumer M, Beerenwinkel N. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Res. 2010:38(21);7400-9.

[75] Beerenwinkel N, Zagordi O. Ultra-deep sequencing for the analysis of viral populations. Curr Opin Virol. 2011:1(5);413-8.

[76] Schulz MH, Weese D, Holtgrewe M, Dimitrova V, Niu S, Reinert K et al. Fiona: a parallel and automatic strategy for read error correction. Bioinformatics. 2014:30(17);i356-63.

[77] Wirawan A, Harris RS, Liu Y, Schmidt B, Schröder J. HECTOR: a parallel multistage homopolymer spectrum based error corrector for 454 sequencing data. BMC bioinformatics. 2014:15(1);131.

[78] Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14(5):R51.

[79] Wang XV, Blades N, Ding J, Sultana R, Parmigiani G. Estimation of sequencing error rates in short reads. BMC bioinformatics. 2012:13(1);185.

[80] Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 2015:43(6);e37.

[81] Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PloS one. 2012:7(2);e30087.

[82] Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin JF. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics. 2011:12(1);245.

[83] Reeder J, Knight R. Rapid denoising of pyrosequencing amplicon data: exploiting the rank-abundance distribution. Nat Meth. 2010:7,668-9.

[84] Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D. Experimental and analytical tools for studying the human microbiome. Nat Rev Genet. 2012:13;47-58.

[85] Gibson J, Shokralla S, Porter TM, King I, Van Konynenburg S, Janzen DH et al. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proc Natl Acad Sci USA. 2014:111 (22);8007-12.

[86] Aagaard K, Riehle K, Ma J, Segata N, Mistretta TA, Coarfa C et al. A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. PLoS One. 2012:7(6);e36466.

[87] Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 2012:22(3);557-67.

[88] Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 2010:11(11);R116.

[89] Kijak GH, Pham P, Sanders-Buell E, Harbolick EA, Eller LA, Robb ML et al. Nautilus: a bioinformatics package for the analysis of HIV type 1 targeted deep sequencing data. AIDS Res Hum Retroviruses. 2013:29(10);1361-4.

[90] Iyer S, Bouzek H, Deng W, Larsen B, Casey E, Mullins JI. Quality score based identification and correction of pyrosequencing errors. Plos One. 2013:8(9);e73015.

[91] Deng W, Maust BS, Westfall DH, Chen L, Zhao H, Larsen BB et al. Indel and Carryforward Correction (ICC): a new analysis approach for processing 454 pyrosequencing data. Bioinformatics. 2013:29(19);2402-9.

[92] Brodin J, Mild M, Hedskog C, Sherwood E, Leitner T, Andersson B et al. PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. Plos One. 2013:8(7);e70338.

[93] Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci USA. 2001:98(17);9748-53.

[94] Chaisson M, Pevzner P, Tang H. Fragment assembly with short reads. Bioinformatics. 2004:20(13);2067-74.

[95] Salmela L, Schröder J. Correcting errors in short reads by multiple alignments. Bioinformatics. 2011:27(11);1455-61.

[96] Levy DN, Aldrovandi GM, Kutsch O, Shaw GM. From the cover: dynamics of HIV-1 recombination in its natural target cells. Proc Natl Acad Sci USA. 2004:101(12);4204-9.

[97] Jung A, Maier R, Vartanian JP, Bocharov G, Jung V, Fischer U et al. Recombination: multiply infected spleen cells in HIV patients. Nature. 2002:418(6894);144.

[98] Lovern M, Underwood M, Nichols G, Song I. PK/PD modeling supports dose escalation decision in VIKING. J Int AIDS Soc. 2010;13(4):182.

[99] Robinson JG, Redford KH, Bennett EL. Wildlife harvest in logged tropical forest. Science. 1999;284(5414):595-6.

[100] Molina JM, Lamarca A, Andrade-Villanueva J, Clotet B, Clumeck N, Liu YP et al. Efficacy and safety of once daily elvitegravir versus twice daily raltegravir in treatment experienced patients with HIV-1 receiving a ritonavir-boosted protease inhibitor: randomised, double blind, phase 3, non-inferiority study. Lancet Infect Dis. 2012 Jan:12(1);27-35.

[101] Quercia R, Dam E, Perez-Bercoff D, Clavel F. Selective-advantage profile of human immunodeficiency virus tipe 1 integrase mutants explains in vivo evolution of raltegravir resistance genotypes. J Virol. 2009:83(19);10245-9.

[102] Zhang L, Ramratnam B, Tenner-Racz K, He Y, Vesanen M, Lewin S et al. Quantifying residual HIV-1 replication in patients receiving combination antiretroviral therapy. N Engl J Med.1999:340(21);1605-13.

[103] García F, Álvarez M, Bernal C, Chueca N, Guillot V. Laboratory diagnosis of HIV infection, viral tropism and resistance to antiretrovirals. Enferm Infecc Microbiol Clin. 2011:29(4);297-307.

[104] Lubelchek RJ, Hoehnen SC, Hotton SL, Kincaid SL, Barker DE, French 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. 2015:68(1);46-54.

[105] Castro-Nallar E, Pérez-Losada M, Burton GF, Crandall KA. The evolution of

HIV: Inferences using phylogenetics. Mol Phylogenet Evol. 2012:62(2);777-92.

[106] Maddison WP, Maddison DR. Mesquite: a modular system for evolutionary analysis. 2015. Version 2.75 http://mesquiteproject.org

[107] Shao W, Boltz VF, Spindler JE, Kearney MF, Malderelli F, Mellors JW 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. Retrovirology. 2013:10;18.

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

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

[110] Redelings BD, Suchard MA .Joint Bayesian Estimation of Alignment and Phylogeny. Syst. Biol. 2005:54(3);401-18

[111] Pasquier C, Millot N, Njouom R, Sandres K, Cazabat M, Puel J et al. HIV-1 subtyping using phylogenetic analysis of *pol* gene sequences. J Virol Methods. 2001:94(1-2);45-54.

[112] Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013:30(12);2725-9.

[113] Di Giallonardo F, Zagordi O, Duport Y, Leemann C, Joos B, Künzli-Gontarczyk M et al. Next-generation sequencing of HIV-1 RNA genomes: determination of error rates and minimizing artificial recombination. PLoS One. 2013:8(9);e74249.

[114] Waugh C, Cromer D, Grimm A, Chopra A, Mallal S Davenport M et al. A general method to eliminate laboratory induced recombinants during massive, parallel sequencing of cDNA library. Virol J. 2015:12;55.

[115] Smyth RP, Schlub TE, Grimm A, Venturi V, Chopra A, Mallal S et al. Reducing chimera formation during PCR amplification to ensure accurate genoryping. Gene. 2010:469(1-2):45-51.

[116] Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl Environ Microbiol. 2005:71(12):8966-9.

[117] Sipos R, Szekely AJ, Palatinszky M, Revesz S, Marialigeti K, Nikolausz M. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiol Ecol. 2007:60(2);341-50.