UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO

COLEGIO DE CIENCIA Y TECNOLOGÍA

POSGRADO DE CIENCIAS GENÓMICAS

Efectos de la modularidad en la diversidad y evolución del Virus de Inmunodeficiencia Humana tipo 1

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS GENÓMICAS

PRESENTA:

IBT. JOSE MANUEL CABALLERO CONTRERAS

DIRECTORA

DRA. CLAUDIA SELENE ZÁRATE GUERRA

Ciudad de México, enero de 2021

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

CA Coordinación Académica

SISTEMA BIBLIOTECARIO DE INFORMACIÓN Y DOCUMENTACIÓN

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO
COORDINACIÓN ACADÉMICA

RESTRICCIONES DE USO PARA LAS TESIS DIGITALES

DERECHOS RESERVADOS©

INTEGRACIÓN DEL JURADO:

Presidente: Dra. Rosa Martha Eugenia Yocupicio Monroy, UACM

Secretario: Dra. Claudia Selene Zárate Guerra, UACM

Vocal: Dr. José Alberto Campillo Balderas, UNAM

Plantel de adscripción:

PLANTEL DEL VALLE, UACM.

DIRECTOR

Dra. Claudia Selene Zárate Guerra

Universidad Autónoma de la Ciudad de México

Agradecimientos

Agradezco profundamente al Consejo Nacional de Ciencia y Tecnología (CONACyT), por haberme otorgado la beca 928146 para llevar a cabo mis estudios.

Al Laboratorio de Bioinformática a cargo de la Dra. Claudia Selene Zárate Guerra del Posgrado en Ciencias Genómicas de la Universidad Autónoma de la Ciudad de México, por haberme permitido realizar mi proyecto de tesis y guiarme en todo momento.

A la Dra. Martha Rosa Eugenia Yocupicio Monroy por su constante asesoramiento para el proyecto, por su disponibilidad y apoyo incondicional.

Al Dr. José Alberto Campillo Balderas por compartir su conocimiento, entusiasmo y amor por la ciencia. También por su asesoramiento, disponibilidad e inclusión en su laboratorio de trabajo.

Al Laboratorio del Origen de la Vida, específicamente al Dr. José Alberto Campillo Balderas y al Dr. Alejandro Rodrigo Jácome Ramírez, de la Facultad de Ciencias de la Universidad Nacional Autónoma de México, por contribuir en mi formación en los aspectos evolutivos y de virología.

A la Dra. Lilia López Cánovas por ayudarme en la corrección de la escritura de la tesis y apoyar a concluir este proyecto.

Dedicatoria

A mis padres, Miguel Caballero Mejía y Verónica Contreras León, por su amor incondicional, apoyo, confianza y por siempre creer en mí. A mis hermanos, Abraham y Jesús Alexander, por ser mi fuerza para continuar siempre. A mis abuelos paternos, Juan y Rosa, porque en vida siempre me amaron y cuidaron. A mis abuelos maternos, Miguel e Irma, por ser mi ejemplo a seguir y cada uno de sus consejos. A mi tío Carlos, por enseñarme que a pesar de todo lo malo, siempre se debe dar lo mejor de sí mismo.

A Mariannita, por su amor incondicional, apoyo y ánimo. Por ayudarme a superarme día a día, ayudarme a ser una mejor persona y siempre estar aún en los peores momentos.

A mis mejores amigos, Leonardo, Enrique, Sergio, Juan Manuel, Lupita, Laura, Rocío, Gisela y Nelly, por su valiosa amistad, por su apoyo incondicional y siempre tener fe en mí. A Migue, que donde quiera que te encuentres, gracias por cada momento compartido, te recuerdo siempre.

Índice general

Índice de tablas 8
Índice de figuras 9
Abreviaturas 11
Resumen 12
Abstract 13
1. INTRODUCCIÓN 14
1.1. Origen y diversificación del VIH 14
1.2. Clasificación del grupo M del VIH-1 16
1.3. Estructura del genoma del VIH-1 17
1.4. Ciclo de replicación del VIH-1 21
a) Fase temprana: 21
b) Fase tardía: 22
1.5. Mecanismos de evolución en virus de ARN 23
1.5.1. Mutación 24
1.5.2. Recombinación y reordenamiento génico 24
1.5.3. Selección natural y deriva génica 25
1.5.4. Cuasiespecies 25
1.5.5. Tamaño de la población 26
1.6. Modularidad 27
1.7. Robustez mutacional 31
2. ANTECEDENTES PARTICULARES 33
2.1. Mecanismos de evolución del VIH-1 33
2.1.1. Tasa de mutación del VIH-1 33
2.1.2. Dinámica de replicación del VIH-1 33
2.2. El ancestro del VIH-1 tiene un origen recombinante 34
2.3. La recombinación y su efecto sobre la evolución del VIH-1 35
2.3.1. Transcripción reversa en el VIH-1 39
2.3.2. Proceso de recombinación en el VIH-1 41
2.3.3. Mecanismos de recombinación 43
2.4. La selección natural y su implicación en el VIH-1 45

2.5. Patrones de "mosaicismo recombinante" en VIH-1 46
3. JUSTIFICACIÓN 49
4. HIPÓTESIS 50
5. OBJETIVOS 50
5.1. Objetivo general 50
5.2. Objetivos particulares 50
6. ESTRATEGIA EXPERIMENTAL 50
7. MATERIALES Y MÉTODOS 53
7.1. Datos de secuencias 53
7.2. Alineamiento de genomas 56
7.3. Detección de recombinación 57
7.4. Identificación de módulos en alineamientos 57
7.5. Detección de diversidad 58
7.5.1. Diversidad presente en módulos 59
7.5.2. Análisis estadísticos de los módulos 59
7.5.3. Generación de filogenias 60
7.5.4. Reconstrucción de estados ancestrales 62
7.6. Detección de selección natural 62
7.6.1. Detección de selección natural mediante FEL 63
7.6.2. Identificación de sitios bajo selección en módulos 63
7.6.3. Asignación de sitios bajo selección en genes 64
7.6.4. Análisis de coevolución 65
8. RESULTADOS 66
8.1. Identificación de puntos de recombinación en las CRFs 66
8.2. Identificación de módulos en las CRFs 66
8.3. Detección de diversidad presente en los genomas de las CRFs y de los subtipos puros que las constituyen 69
8.3.1. Comparación de la diversidad de los genomas completos entre las CRFs y los subtipos puros que las constituyen 69
8.3.2. Comparación de la diversidad de los módulos entre las CRFs y los subtipos puros que las constituyen 71
8.3.3. Reconstrucción de estados ancestrales de los polimorfismos 73
8.4. Efecto de la selección natural en la evolución modular de los genomas de las CRFs y de los subtipos puros que las constituyen 74
8.4.1. Sitios bajo selección en los módulos de las CRFs y de los subtipos puros que las constituyen 74

8.4.2. Sitios bajo selección en los genes de las CRFs y de los subtipos puros que las constituyen 76

8.4.3. Sitios bajo coevolución en los genomas de las CRFs y de los subtipos puros que las constituyen 81

9. DISCUSIÓN 84

9.1. Identificación de módulos en las CRFs 84

9.2. Diversidad presente tanto en el genoma completo como en los módulos de las CRFs y de los subtipos puros que las constituyen 86

9.3. Efecto de la selección natural en la evolución de las CRFs y de los subtipos puros que las constituyen 87

9.3.1. Efecto de la selección natural en los módulos 87

9.3.2. Efecto de la selección natural en los genes 88

9.3.3. Efecto de la coevolución 90

10. CONCLUSIONES 92

11. BIBLIOGRAFÍA 93

12. ANEXOS 100

12.1. Congresos 100

12.2. Glosario 101

12.3. Código 103

12.4. Figuras adicionales 117

12.4.1. Módulos identificados en las CRFs 117

12.4.2. Diversidad del genoma completo de las CRFs y de los subtipos puros que las conforman 119

12.4.3. Polimorfismos compartidos entre las CRFs y sus subtipos puros 122

12.4.4. Gráficas de comparación de la diversidad genética presente en los módulos . 126

Índice de tablas

Tabla I. Proteínas del VIH-1 y su función. 19
Tabla II. Distribución geográfica de las secuencias obtenidas para este estudio. 53
Tabla III. Comparación para la prueba Wilcoxon. 60
Tabla IV. Modelos evolutivos usados para construir la filogenia bayesiana. 61
Tabla V. Reglas para la asignación de módulos en los genes. 63
Tabla VI. Reglas para la asignación de coordenadas de los genes. 64
Tabla VII. Posición y longitud de los módulos de los CRFs. 67
Tabla VIII. Sitios bajo selección en los módulos de las CRFs y de los subtipos puros que las constituyen. 75
Tabla IX. Sitios bajo selección en las proteínas de la **CRF-06_cpx(A1,G,J,K)** y en sus subtipos puros que la constituyen. 77
Tabla X. Sitios bajo coevolución en las CRFs y los subtipos puros que las constituyen. 81

Índice de figuras

Figura 1. Diversificación y clasificación del VIH. 15
Figura 2. Clasificación del grupo M del VIH-1..... 17
Figura 3. Estructura del genoma del VIH-1. 18
Figura 4. Ciclo de vida del VIH-1. 23
Figura 5. Mecanismos de evolución viral. 27
Figura 6. Los múltiples grados de modularidad en sistemas biológicos. 28
Figura 7. Construcciones recombinantes del MSV y su adecuación biológica relativa. 31
Figura 8. Supervivencia del más plano. 32
Figura 9. Filogenia de máxima verosimilitud de secuencias de Pol y Env de SIVs. 35
Figura 10. Consecuencias evolutivas de la recombinación. 36
Figura 11. Formación de genomas recombinados. 37
Figura 12. Transcripción reversa en el VIH-1. 40
Figura 13. Proceso de Template Switching. 41
Figura 14. Proceso de Copy Choice. 42
Figura 15. Modelos de recombinación en VIH-1. 44
Figura 16. Estructuras recombinantes. 46
Figura 17. Patrones de mosaicismo recombinante. 47
Figura 18. Puntos de recombinación en la CRF90_BF1. 48
Figura 19. Estrategia experimental llevada a cabo en este estudio. 52
Figura 20. Distribución geográfica de las secuencias de genomas obtenidas para este estudio. 55
Figura 21. Puntos de recombinación en la **CRF-02_A1G**. 66
Figura 22. Módulos de la **CRF-02_A1G**. 67
Figura 23. Diversidad del genoma completo. 70
Figura 24. Número de sitios polimórficos presentes en los genomas de la **CRF-02_A1G** y de sus subtipos puros. 71

Figura 25. Comparación de la diversidad genética presente en los módulos de la **CRF-02_A1G** y sus subtipos puros..... 72
Figura 26. Reconstrucción de estados ancestrales. 73
Figura 27. Sitios bajo selección positiva en gp120..... 78
Figura 28. Sitios bajo selección negativa en la RT. 79
Figura 29. Sitios bajo selección negativa en gp120 del **subtipo A1**. 80
Figura 30. Sitios bajo selección positiva en la RT del **subtipo A1**. 80
Figura 31. Sitios bajo coevolución en las proteínas de las CRFs y los subtipos puros que los constituyen. 83
Figura 32. Distribución de puntos de recombinación a través del genoma del VIH-1.85
Figura 33. Patrones de recombinación de las 15 CRF_BF que circulan a nivel mundial. 86
Figura 34. Sitios bajo selección en la estructura primaria de gp120. 89
Figura 35. Vista del sitio activo del dominio de polimerasa de la RT del VIH-1 en un complejo con ADN bicatenario. 90

Abreviaturas

ADN: Ácido Desoxirribonucleico.

ARN: Ácido Ribonucleico.

CRFs: *Circulating Recombinant Forms*, formas recombinantes circulantes.

FEL: *Fixed Effects Likelihood*, probabilidad de efectos fijos.

GARD: *Genetic Algorithm for Recombination Detection*, algoritmo genético para la detección de recombinación.

kb: Kilobase, es una unidad de medida en biología molecular equivalente a 1,000 pares de bases de ADN o ARN.

LANL: *Los Alamos National Laboratory*, Laboratorio Nacional Los Álamos.

LTRs: *Long Terminal Repeats*, repeticiones terminales largas.

nt: Nucleótido.

ORF: *Open Reading Frame*, marco de lectura abierto.

RdRp: *RNA dependent RNA polymerase*, ARN polimerasa dependiente de ARN.

RT: *Reverse Transcriptase*, transcriptasa reversa.

SIDA: Síndrome de Inmunodeficiencia Adquirida.

SIV: *Simian Immunodeficiency Virus*, virus de inmunodeficiencia de simios.

URFs: *Unique Recombinant Forms*, formas recombinantes únicas.

VIH-1: Virus de Inmunodeficiencia Humana tipo 1.

Resumen

Los virus no han evolucionado de un único antepasado común; sin embargo, algunos de ellos forman una densa red evolutiva en la que los genomas están vinculados a través de diferentes genes compartidos. Este tipo de relación evolutiva resulta de un amplio intercambio de genes y módulos genéticos (Koonin *et al.*, 2015). El objetivo de este trabajo fue evaluar el efecto de la modularidad en la diversidad y evolución en los genomas del grupo M del VIH-1.

Se analizaron 306 secuencias de genoma completo provenientes de la base de datos del Laboratorio Nacional de Los Álamos (LANL), tanto de formas recombinantes circulantes (CRFs) como de subtipos puros. Utilizando el algoritmo GARD implementado en el servidor evolutivo de DataMonkey, identificamos los puntos de recombinación en las CRFs; posteriormente, usando la herramienta SNP de la base de datos de ViPR, identificamos una mayor diversidad genética en los subtipos puros respecto a sus CRFs.

Utilizando el algoritmo FEL implementado en DataMonkey y el programa Chimera, encontramos que independientemente si es una CRF o un subtipo puro, en el gen que codifica para gp120 predominó la selección positiva mientras que para el gen que codifica para la RT predominó la selección negativa. También, usando SpiderMonkey a través de Hyphy, identificamos un mayor número de relaciones co-evolutivas en los subtipos puros respecto a sus CRFs; un dato de gran relevancia que observamos fue que en las CRFs que contienen al subtipo B, sólo se identificó una relación co-evolutiva. De estas relaciones co-evolutivas, la que se presentó mayoritariamente fue la de gp120-gp120.

Los resultados de nuestro trabajo nos permiten afirmar que la modularidad en los genomas del grupo M del VIH-1 puede presentarse en cualquier parte del genoma, mantiene una restricción en cuanto a diversidad genética, permite las marcas de selección y elimina las interacciones co-evolutivas presentes en los subtipos puros.

Abstract

Viruses have not evolved from a single common ancestor; however, these elements form a dense evolutionary network in which genomes are linked through different shared genes. This type of evolutionary relationship results from extensive exchange of genes and gene modules (Koonin *et al.*, 2015). The objective of this work was to evaluate the effect of modularity on diversity and evolution in the genomes of group M of HIV-1.

306 whole genome sequences from the Los Alamos National Laboratory (LANL) database were analyzed, both from circulating recombinant forms (CRFs) and from pure subtypes. Using the GARD algorithm implemented in the DataMonkey evolutionary server, we identified the recombination points in the CRFs; later, using the SNP tool from the ViPR database, we identified a greater genetic diversity in the pure subtypes with respect to their CRFs.

Using the FEL algorithm implemented in DataMonkey and the Chimera program, we found that regardless of whether it is a CRF or a pure subtype, positive selection predominated in the gene coding for gp120, while negative selection predominated for the gene coding for RT. Also, using SpiderMonkey through Hyphy, we identified a greater number of co-evolutionary relationships in the pure subtypes with respect to their CRFs; A highly relevant data that we observed was that in the CRFs containing subtype B, only a co-evolutionary relationship was identified. Of these co-evolutionary relationships, the one that was presented the most was that of gp120-gp120.

The results of our work allow us to affirm that the modularity in the genomes of the HIV-1 group M can occur in any part of the genome, maintains a restriction regarding genetic diversity, allows selection marks and eliminates co-evolutionary interactions present in the pure subtypes.

1. INTRODUCCIÓN

1.1. Origen y diversificación del VIH

El Síndrome de Inmunodeficiencia Adquirida (SIDA) fue reconocido como una nueva enfermedad en la década de los 80´s y su agente causal de dicha enfermedad fue identificado como un retrovirus, ahora conocido como Virus de Inmunodeficiencia Humana tipo 1 (VIH-1) (Sharp & Hahn, 2011).

El VIH se originó de múltiples transmisiones zoonóticas del Virus de Inmunodeficiencia de Simios (SIV) de primates a humanos en África occidental y central (Hemelaar, 2013). A la fecha, han sido identificados SIVs en al menos 45 especies diferentes de primates no humanos de África y en general, cada especie es infectada con un linaje específico del virus (Peeters *et al.*, 2013).

Los eventos independientes de transmisión zoonótica de primates a humanos han generado 2 tipos de VIH, como se muestra en la Figura 1, el tipo 1 (VIH-1) clasificado en 4 grupos: M, N, O y P, y el tipo 2 (VIH-2) clasificado en 8 grupos: A-H (Hemelaar, 2013).

El VIH-1 está más relacionado filogenéticamente al SIV_{cpz} y al SIV_{gor}, los cuales infectan al chimpancé (*Pan troglodytes troglodytes*) y al gorila (*Gorilla gorilla*), respectivamente. Los chimpancés y gorilas son especies simpátricas y comparten los mismos hábitats en ciertas áreas de África Central. El SIV_{cpz} está mayormente relacionado a los grupos M y N del VIH-1, mientras que el SIV_{gor} lo está a los grupos O y P (Peeters *et al.*, 2013); por otro lado, los grupos A, B, C, G y H del VIH-2 están más relacionados filogenéticamente al SIV_{sm}, el cual infecta a los mangabeys tiznados (*Cercocebus atys*) (Kerina *et al.*, 2013; Peeters *et al.*, 2013). El VIH-1 y VIH-2 son virus relacionados filogenéticamente con una similitud de secuencia de nucleótidos del 58%, 59% y del 39% en los genes *gag*, *pol* y *env*, respectivamente (Kerina *et al.*, 2013).

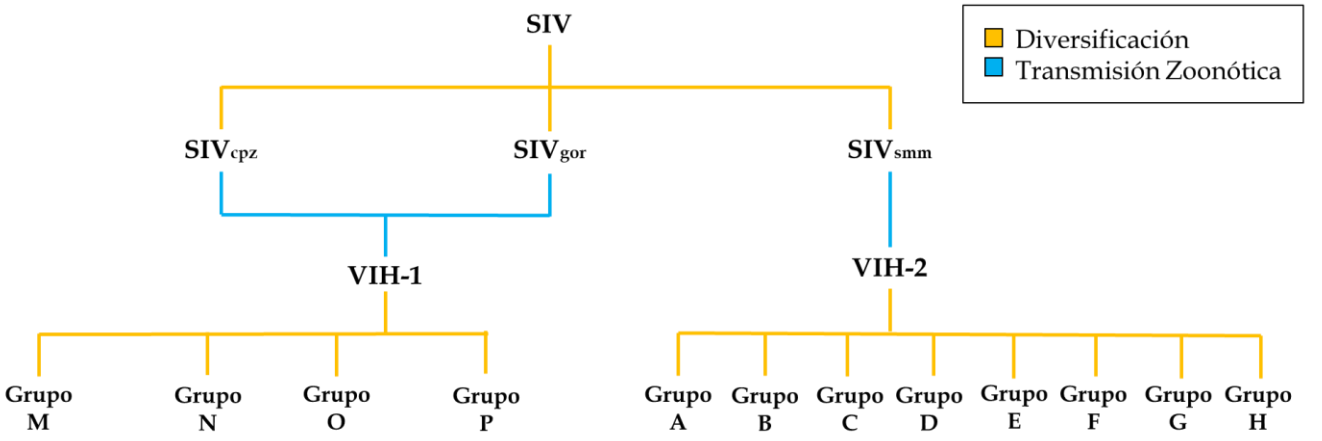

Figura 1. Diversificación y clasificación de las variantes de VIH.

Los cuatro grupos del VIH-1 son claramente el resultado de cuatro transmisiones independientes inter-especies de chimpancés y gorilas al humano en África Central y del Oeste, pero sólo uno, el grupo M ("*major*"), descubierto en 1983, ha sido distribuido a través de África y los demás continentes. El grupo O ("*outlier*"), descrito en 1990, permanece restringido a África Central y del Oeste, y la prevalencia más alta se ha reportado en Camerún donde representa actualmente el 1% de las infecciones por VIH-1. Las infecciones por los grupos N ("*non-M/non-O*") y P ("*putative group*"), descritos en 1998 y en 2009 respectivamente, solamente han sido observadas en unos pocos pacientes cameruneses, <20 pacientes para el grupo N y dos pacientes para el grupo P (Kerina *et al.*, 2013; Peeters *et al.*, 2013).

De acuerdo a la clasificación propuesta por Baltimore (Baltimore, 1971), el VIH-1 pertenece al grupo VI, ya que su genoma es de ARN de cadena sencilla con polaridad positiva que transcribe a un ADN intermediario gracias a su transcriptasa reversa. Los virus de ARN tienen un alto potencial para generar diversidad genética en niveles intra- e inter-hospedero; lo cual facilita la **adaptación** viral a nuevos ambientes. Las consecuencias de esta plasticidad tienen repercusiones clínicas importantes en la progresión de la enfermedad, infectividad, transmisibilidad y resistencia a tratamientos antivirales. La gran diversidad de las poblaciones del VIH-1 son el resultado de una alta tasa de mutación derivada del error inherente de su transcriptasa

reversa, de su alta tasa de recombinación, aunadas a su rápida replicación y el gran tamaño de la población (Rojas-Sánchez et al., 2017).

1.2. Clasificación del grupo M del VIH-1

La diseminación global del grupo M del VIH-1 durante la segunda mitad del siglo XX ha generado variantes genéticas clasificada en subtipos, sub-subtipos y formas recombinantes (Delatorre & Bello, 2013). Estas últimas se han clasificado en formas recombinantes circulantes (CRFs) cuando tres o más pacientes no relacionados epidemiológicamente presentan virus con patrones idénticos de mosaicismo, mientras que los virus con una estructura de mosaico única o presente en uno o dos pacientes no relacionados epidemiológicamente se denominan formas recombinantes únicas (URFs) (Reis et al., 2017).

El grupo M del VIH-1, como se muestra en la Figura 2, se ha diversificado en nueve subtipos genéticos (denominados A-D, F-H, J y K) (Hemelaar, 2013); esta clasificación ha sido basada en el análisis filogenético de las secuencias de los genes *env* y *gag* (Subbarao y Schochetman, 1996). La variación inter-subtipos está alrededor del 30% con respecto a la secuencia del gen env, mientras que para la secuencia del gen *gag* es alrededor del 15% (Kerina *et al.*, 2013).

Los subtipos A y F han sido además divididos en dos sub-subtipos (A1, A2, F1 y F2) (Galetto & Negroni, 2005). Sin embargo, Désiré y colaboradores en 2018, propusieron una nueva clasificación para los sub-subtipos A y D, ya que, utilizando análisis filogenéticos mediante el método de máxima verosimilitud, identificaron patrones de divergencia significativa que no habían sido incluidos en la primera clasificación entre los dos subtipos. La propuesta fue una ligera modificación a la clasificación del subtipo A, dividiéndolo en seis sub-subtipos, nombrados A1, A2, A3, A4, A6 y A7; mientras que con el subtipo D, los análisis reportaron tres sub-subtipos, nombrados D1-D3 (Désiré et al., 2018). Actualmente, los recombinantes incluyen una lista extensa de 102 CRFs (www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html).

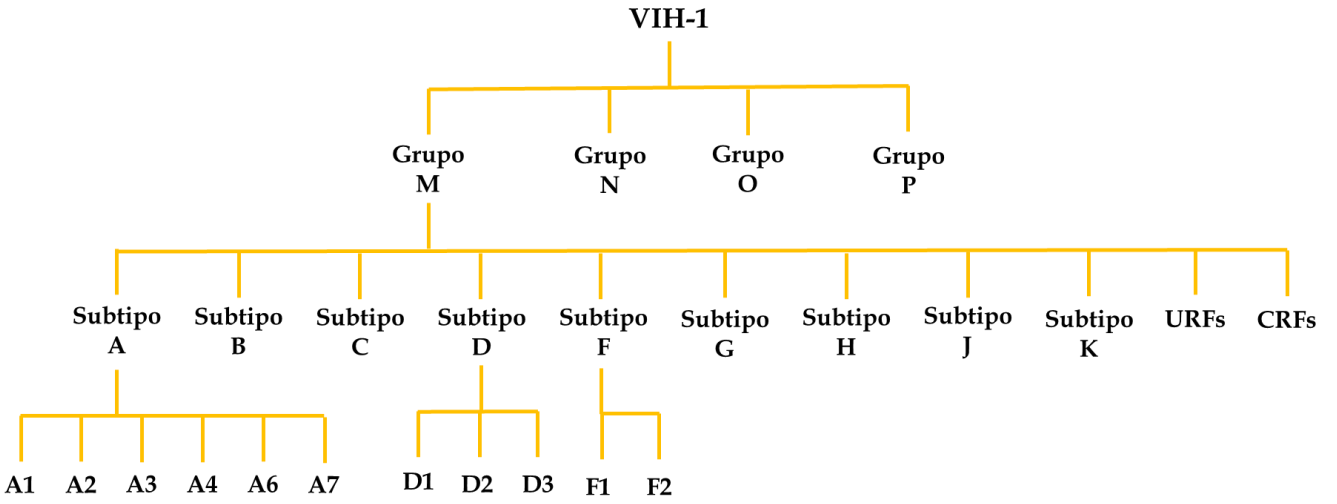

Figura 2. Clasificación del grupo M del VIH-1; se añade la modificación propuesta por Désiré *et al.*, 2018.

La distribución global de los subtipos y sub-subtipos del VIH-1 es altamente heterogénea. En 2004, el subtipo C fue el más prevalente globalmente, atribuyéndosele hasta el 50% de todas las infecciones, seguido por el subtipo A y B, con un 12% y 10%, respectivamente. En cuanto a su distribución geográfica, el subtipo A predomina en países de África Central y del Este (Kenia, Uganda, Tanzania y Rwanda), al igual que en países de Europa del Este. El subtipo B es el responsable del mayor número de casos en Europa Central y del Oeste, América del Norte (EEUU, Canadá y México), así como en varios países en América Central y del Sur, el Caribe, Australia y en el noreste y el medio este de África (Beloukas et al., 2016).

1.3. Estructura del genoma del VIH-1

El genoma del VIH-1 está formado por una molécula de ARN de cadena sencilla de polaridad positiva (Sakuragi *et al.*, 2016) de aproximadamente 9,700 nt de longitud. En cada partícula viral se encapsidan dos copias del genoma que se asocian como dímeros, por lo que se considera que es una entidad biológica diploide (Lever & Jeang, 2006).

Se ha especulado que el empaquetamiento de dos copias de ARN viral en los retrovirus les confiere ventajas, porque aumenta su diversidad genética mediante el proceso de

recombinación (Nikolaitchik *et al.*, 2015). El material genético está recubierto por cerca de 1,800 copias de la proteína de nucleocápside (NC), que constituye el componente interior de la partícula viral. Esta estructura además contiene las enzimas virales retrotranscriptasa (RT) e integrasa (IN) y ARNt celulares, principalmente el $ARNt^{Lys}$ que funciona como cebador de la transcripción reversa (Cimarelli & Darlix, 2014).

Como se muestra en la Figura 3, el genoma del VIH-1 contiene los elementos comunes de todos los retrovirus, incluyendo los genes estructurales *gag*, *pol* y *env* que están flanqueados en sus extremos 5´ y 3´ por una secuencia de repetidos (R) y una secuencia única (U), que luego del proceso de transcripción reversa se denominan repetidos terminales largos (LTR) en el orden 5´-LTR-*gag-pol-env*-LTR-3´. El gen *gag* codifica para las proteínas p17 (matriz [MA]), p24 (cápside [CA]) y p15 (nucleocápside [NC]), las cuales son sintetizadas como una poliproteína (p55); el gen *pol* codifica para la proteasa (PR), la transcriptasa reversa (RT) y la endonucleasa/integrasa (IN); mientras que el gen *env* codifica para las proteínas de la envoltura viral (SU) y transmembranal (TM), gp120 y gp41, respectivamente (Gonda, 1988).

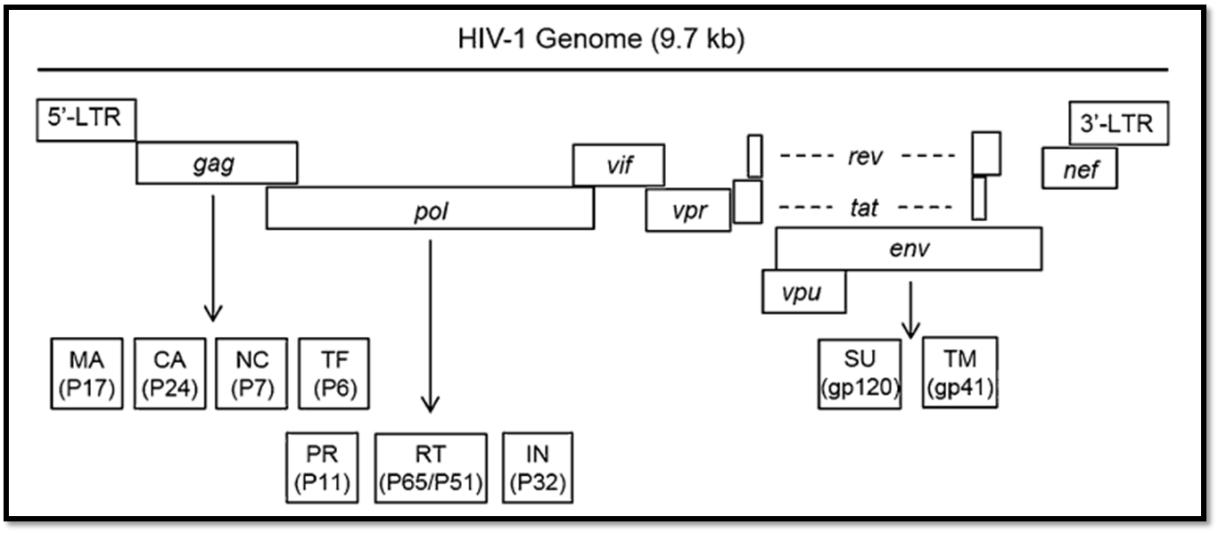

Figura 3. Estructura del genoma del VIH-1; tomada de Nkeze *et al.*, 2015.

El VIH-1 codifica seis proteínas adicionales, de las cuales, cuatro son proteínas accesorias (Frankel y Young, 1998): Nef, Vif, Vpu y Vpr, que participan en la evasión

(y manipulación) tanto de la inmunidad innata como la adaptativa (Malim y Emerman, 2008), y dos proteínas reguladoras: Tat y Rev, que llevan a cabo funciones esenciales en la regulación de la expresión génica. Por lo tanto, el VIH-1 puede ser considerado como una entidad molecular que consiste de 15 proteínas y un ARN (Frankel y Young, 1998). En la Tabla I se muestran las funciones que lleva a cabo cada una de las proteínas del VIH-1.

Tabla I. Proteínas del VIH-1 y su función.

Gen	**Proteína**	**# copias por virión**	**Interacciones con otros factores virales**	**Función**
gag	Matriz (p17)	~5000	TM	Se dirige a la membrana plasmática para el ensamblaje de viriones; incorporación de Env; eventos post-entrada
gag	Cápside (p24)	~5000	Desconocidas	Estructura principal del virión y ensamblado
gag	Nucleocápside (p7)	~5000	ARN genómico del virus	Empaquetamiento del ARN viral; ensamblado del virión
gag	Factor de transferencia (p6)	~5000	Vpr	Promueve la salida del virión
pol	Proteasa	~250	Gag, Pol	Procesamiento proteolítico de las poliproteínas Gag y Gag-Pol
pol	Transcriptasa reversa (p51-p66)	~250	ARN genómico, IN	Síntesis del ADN_c; el dominio de RNAsa H degrada el ARN
pol	Integrasa	~250	ADN_c viral, RT	Inserción del ADN_c viral en el ADN celular

env	Glicoproteína de superficie (gp120)	4-35 trímeros	TM	Unión del virus a los receptores de la superficie celular; media la unión y entrada del virus
env	Glicoproteína de transmembrana (gp41)	4-35 trímeros	gp120, MA	Contiene el péptido de fusión; facilita la fusión de la membrana y la entrada del virus
tat	Trans-activador de la transcripción (Tat)	Ninguna	ARN viral	Activador potente de la elongación de la transcripción viral
rev	Regulador de la expresión de las proteínas virales (Rev)	Ninguna	ARN viral con intrones, vía Elemento de Respuesta a Rev (RRE)	Induce la exportación nuclear de ARNs virales; bloquea el splicing
nef	Factor Negativo (Nef)	Escindido por PR	Desconocidas	Regulación negativa de CD4 y HLA; activación de células T; potencia la infectividad viral; bloquea la apoptosis; determinante de patogenicidad
vif	Factor de infectividad del virión (Vif)	1-150	Desconocidas	Suprime los factores de restricción de la infección del hospedero (APOBEC)
vpr	Proteína R viral (Vpr)	~700	p6	Potencia la infectividad post-entrada; arresta el ciclo celular
vpu	Proteína U viral (Vpu)	Ninguna	Desconocidas	Regulación negativa de CD4 y HLA; induce la liberación del virión de la superficie celular del hospedero

CA: Cápside, **IN:** Integrasa, **MA:** Matriz, **NC:** Nucleocápside, **PR:** Proteasa, **RT:** Transcriptasa reversa, **SU:** Glicoproteína de superficie, **TM:** Transmembrana. Adaptada de De Goede *et al.*, 2015.

1.4. Ciclo de replicación del VIH-1

El ciclo de replicación del VIH-1, como se muestra en la Figura 4, puede ser dividido en dos fases: el *estadio temprano*, que ocurre desde la entrada en la célula hospedadora hasta la integración en su genoma, y la *fase tardía*, la cual ocurre posterior a la integración del provirus hasta la replicación viral completa. De acuerdo con esto, dos tipos de latencia viral pueden ser diferenciados: *latencia de pre-integración* se refiere a la generación de diferentes formas de ADN viral antes de la integración, mientras que la *latencia post-integración* se refiere a la falta de replicación después de la inserción del ADN viral en el genoma hospedador (Coiras *et al.*, 2009).

Por lo tanto, el ciclo de replicación del VIH-1 puede resumirse en 11 pasos generales:

a) Fase temprana:

1) La entrada viral involucra la fusión de la membrana viral y celular a través de interacciones sucesivas de la proteína gp120 con el receptor celular CD4 y los correceptores CXCR4 o CCR5. **2)** La nucleocápside del VIH-1 que contiene dos copias del ARN genómico, oligonucleótidos de ARN de transferencia (ARN_t), proteasa viral, transcriptasa reversa e integrasa, son liberados en el citosol. Este conjunto intracelular es denominado *complejo de transcripción reversa*, ya que el ARN genómico es retro-transcrito por la RT para generar una molécula de ADN de doble cadena lineal (ADN_{dcc}) con extremos repetidos directos. **3)** Una vez en el citoplasma, el complejo de transcripción reversa se desensambla progresivamente para formar el *complejo de pre-integración* (PIC), el cual está compuesto de un ADN_{dcc} lineal, IN, MA, RT, Vpr y de varias proteínas del hospedador. **4)** El tráfico viral es mediado por el transporte retrógrado usando microtúbulos y dineína para moverse hacia el complejo del poro nuclear (NPC). En el complejo del poro nuclear, el PIC es capaz de transportar el ADN_{dcc} a través de la membrana nuclear. **5)** El ADN_{dcc} lineal puede integrarse en cualquier cromosoma del hospedero o circularizarse como uno o dos círculos que contienen repetidos terminales largos (LTRs). Algunos factores del hospedero como la emerina y el factor de crecimiento derivado del epitelio ocular (LEDGF) son moléculas clave que facilitan la integración del ADN_{dcc} del VIH-1 (Coiras *et al.*, 2009).

b) Fase tardía:

6) Después de la integración, el provirus flanqueado con los LTRs se comporta como un gen celular: el extremo 5´-LTR funciona como un promotor de células eucariotas y el extremo 3´-LTR actúa como un sitio de poliadenilación y de término. La activación de las células T CD4+ induce la unión de elementos activadores de la transcripción en el extremo 5´-LTR del virus. **7)** Este complejo reúne factores transcripcionales esenciales del hospedero, tales como el factor nuclear κB (NF-κB), factor nuclear de células T activadas (NFAT) y la proteína específica 1 (SP1). Estas proteínas transmiten señales de activación a los factores basales que pertenecen a la maquinaria de transcripción general y promueven la unión de la ARN polimerasa II (ARNPII) a la caja TATA para iniciar la transcripción del ARN_m. Una estructura de tallo-burbuja de 59 nt de longitud denominada región de respuesta al transactivador (TAR) se forma en el extremo 5´ del transcrito viral naciente, creando un sitio de unión para el transactivador viral Tat. La interacción Tat-TAR promueve una elongación eficiente de los transcritos virales por el reclutamiento de factores celulares que incrementan la capacidad funcional de la ARNPII. **8)** La elongación eficiente de los transcritos virales permite la síntesis del ARN_m, el cual será procesado con ayuda de la proteína regulatoria Rev. Rev es un factor viral de unión al ARN_m que regula el transporte núcleo-citoplasma y el splicing de los ARN_m virales. **9)** Una vez en el citoplasma, los ARN_m del VIH-1 son traducidos como poliproteínas. Primero, se procesa la glicoproteína gp160, para dar lugar a las proteínas de la envoltura (gp120 y gp41) que son transportadas hacia la membrana plasmática, en donde **10)** ocurre el ensamblaje de las partículas virales con la asociación de las poliproteínas *gag*, *gag-pol* y el ARN genómico, **11)** las partículas inmaduras son liberadas de la célula y finalmente la proteasa viral procesará a las poliproteínas *gag* y al híbrido *gag-pol*, para convertirse en una partícula viral madura e infecciosa (Coiras *et al.*, 2009).

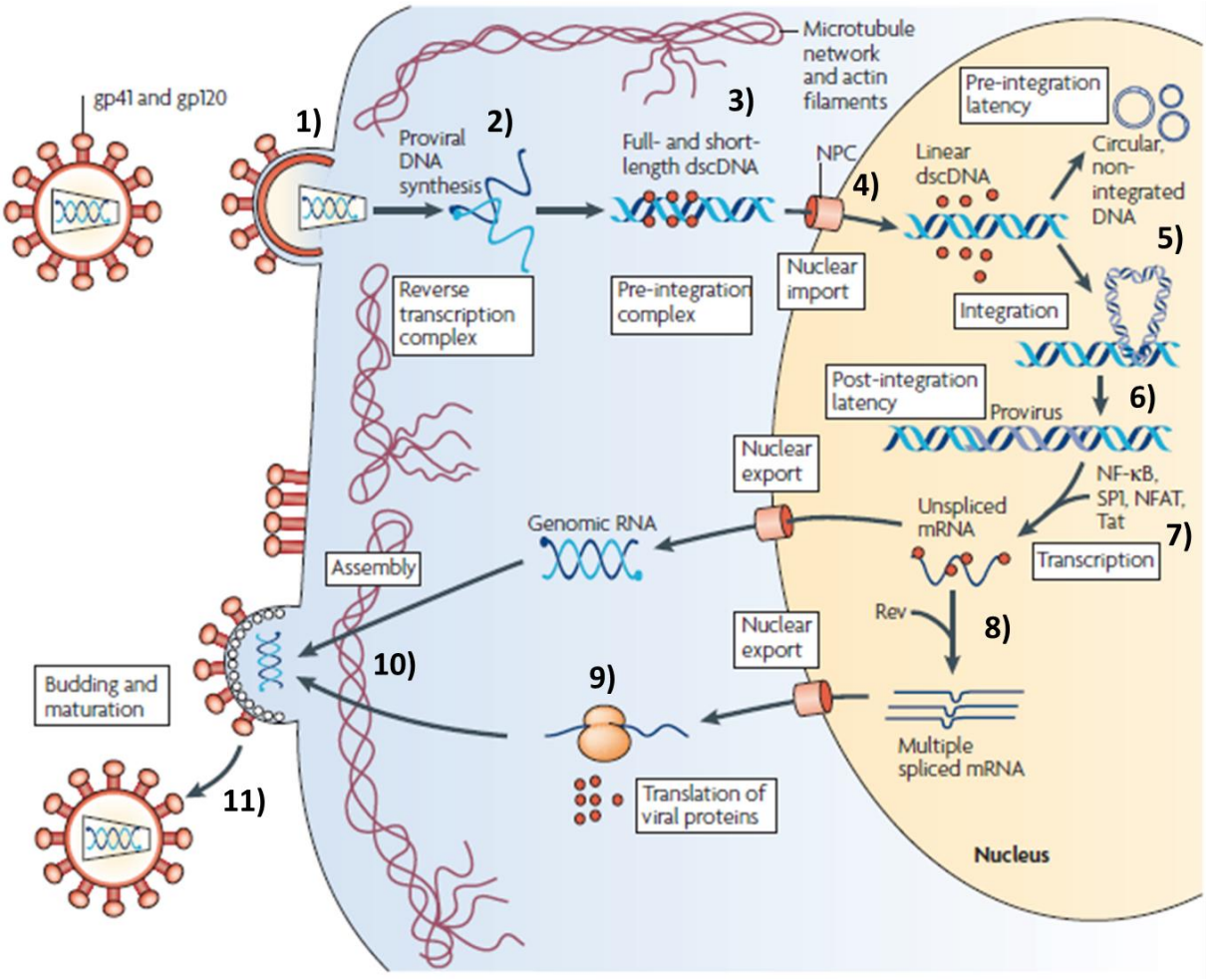

Figura 4. Ciclo de vida del VIH-1; cada uno de los 11 pasos generales es explicado en el texto. Modificado de Coiras *et al.*, 2009.

1.5. Mecanismos de evolución en virus de ARN

Los virus de ARN, se replican por medio de una polimerasa dependiente de ARN (RdRp) y exhiben las tasas de mutación más altas en la naturaleza. Constituyen diversas clases de virus que infectan hospederos en los tres dominios de la vida: eukarya, bacteria y archaea. Las poblaciones virales evolucionan por la acción de la mutación y la recombinación, y están sujetos a las mismas fuerzas evolutivas como todos los organismos, incluidas la **deriva génica** y la **selección natural**. Las historias de vida de los virus están caracterizadas por episodios de selección purificadora fuerte que conlleva a una rápida **evolución**, y frecuentemente, las poblaciones sufren cuellos

de botella en los que actúa la deriva génica (Dolan *et al.*, 2016). En la Figura 5 se muestran los mecanismos que permiten la evolución en los virus de ARN.

1.5.1. Mutación

La mutación es la fuente de la diversidad genotípica dentro de una población. Las RdRps de los virus de ARN son especialmente propensas a errores, exhibiendo las tasas de mutación más altas (mutaciones por sitio por generación) en la naturaleza (Dolan *et al.*, 2016). Esta característica es debida a la ausencia de "*proofreading*" o mecanismos de reparación (Sanjuán, 2012). Las tasas más altas de mutación de los virus de ARN corresponden a genomas de longitudes cortas, usualmente en el orden de los 10 kb, consistente con una correlación inversa general entre la longitud del genoma y la tasa de mutación. Además de las mutaciones adquiridas a través de los errores de replicación por la RdRp, las tasas de mutación de los virus están influenciadas por enzimas del hospedero, incluido el complejo de edición de la apoliproteína B (APOBEC) y la adenosina desaminasa específica de ARN (ADAR), por ejemplo (Dolan *et al.*, 2016).

1.5.2. Recombinación y reordenamiento génico

La recombinación en los virus de ARN ocurre a través de un mecanismo conocido como "*copy choice*", donde una RdRp asociada con un transcrito naciente se disocia de una cadena y se asocia con otra. Cuando esto ocurre en el mismo sitio, la recombinación es homóloga; mientras que cuando la recombinación ocurre entre sitios diferentes (o con ARN celular), la recombinación es no homóloga (Dolan *et al.*, 2016).

El reordenamiento ocurre en virus de ARN con genomas segmentados (Dolan *et al.*, 2016). Durante el cual se intercambian genes completos (o conjuntos de genes) por el intercambio de segmentos (McDonald *et al.*, 2016) cuando múltiples genotipos infectan una célula, así, la progenie hereda una mezcla de segmentos de genes de los virus parentales. El reordenamiento crea un medio eficiente para explorar combinaciones de genotipos, al mismo tiempo que preserva las interacciones epistáticas dentro de los segmentos de genes (Dolan *et al.*, 2016).

A diferencia del reordenamiento, la recombinación puede ocurrir en casi cualquier parte del genoma, incluso en medio de un gen. Por lo tanto, la recombinación puede dar lugar a la formación de proteínas quiméricas no funcionales, mientras que el reordenamiento no lo puede hacer. En otras palabras, el reordenamiento es un mecanismo que mantiene el marco de lectura abierto (ORF) de un gen y, por lo tanto, mantiene la integridad de la proteína; mientras que la recombinación puede introducir cambios en los ORFs y sus proteínas codificadas (McDonald *et al.*, 2016)

1.5.3. Selección natural y deriva génica

Cada mutación tiene un efecto característico en el éxito de replicación o **adecuación biológica** del virus, conocido como efecto de adecuación mutacional. La selección natural actúa sobre la diversidad fenotípica de genomas mutantes en la población para conducirla hacía el incremento de la adecuación biológica. Mutaciones benéficas, podrían aumentar su frecuencia hasta la fijación a través de la **selección positiva**. Mutaciones deletéreas, las cuales resultan en una reducción de la adecuación, son eliminadas de la población mediante la **selección negativa** o purificadora. Como tal, la selección natural tiene una influencia determinista en la evolución de las poblaciones virales (Dolan *et al.*, 2016).

En poblaciones finitas, los procesos evolutivos son llevados a cabo no sólo por la influencia determinista de la selección, sino también por la influencia estocástica de la deriva genética. La deriva genética es la fluctuación estocástica de las frecuencias alélicas en una población, y ocurre en todas las poblaciones finitas como resultado de un error de muestreo de una generación a otra. El tamaño de la población podría ser reducido dramáticamente durante la transmisión de virus a nuevos hospederos o especialmente, a nuevas especies; por lo tanto, mutaciones neutrales y deletéreas pueden ser fijadas en la población por la deriva génica, además de que pueden fijarse a través de su ligamiento con mutaciones adaptativas (Dolan *et al.*, 2016).

1.5.4. Cuasiespecies

Las poblaciones de virus de ARN existen como una larga colección de genotipos relacionados rodeando una "*secuencia maestra*", a menudo denominada como un

"enjambre mutante" o *"cuasiespecies"*. En un ambiente dado, el efecto determinista de la selección actúa en el efecto de la adecuación mutacional de los genotipos en la población dando lugar a poblaciones de equilibrio, con los genotipos de mayor adecuación que conformarán la mayoría de las secuencias. A medida que cambia el entorno, puede cambiar rápidamente la frecuencia de las variantes dentro de las cuasiespecies, cambiando la estructura genética de la población (Dolan *et al.*, 2016).

1.5.5. Tamaño de la población

El tamaño de población es un parámetro crítico en la evolución de las poblaciones, y es uno de los que fluctúa ampliamente durante los ciclos de infección viral. El término *"tamaño de población efectiva"* se refiere al tamaño de la población en el que una población modelo exhibiría la misma diversidad. En algunos virus se ha observado un mecanismo que contrarresta la influencia negativa de la deriva génica en la transmisión de poblaciones pequeñas que se denomina **"transmisión en bloque"**. Cuando el número de partículas infecciosas es limitado, como al comienzo de una infección o después de un cuello de botella en la población, la transmisión en bloque puede inclinar la balanza hacia co-infecciones más frecuentes, lo que resulta en un entorno selectivo más tolerante y, por lo tanto, mejora la robustez de una población viral (Dolan *et al.*, 2016).

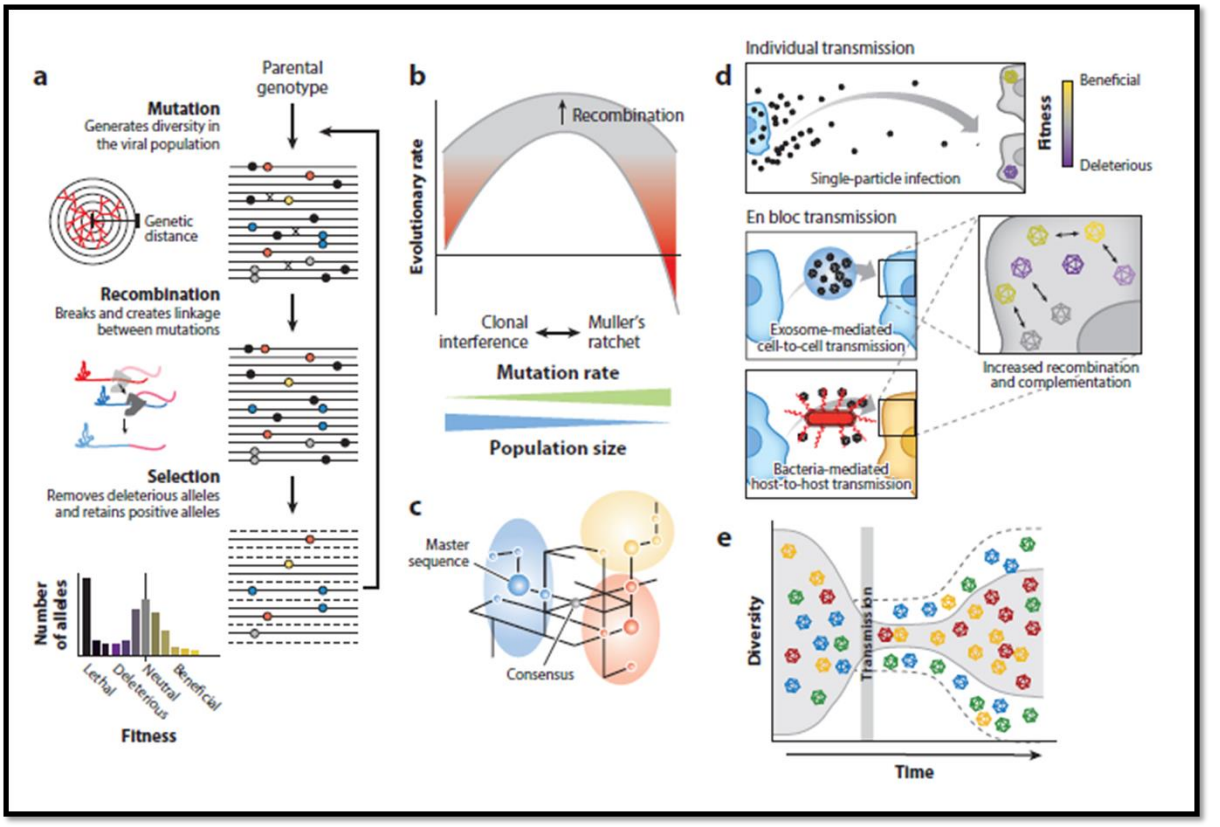

Figura 5. Mecanismos de evolución viral. **(a)** La evolución ocurre a través del proceso de mutación, selección y deriva. Estos procesos generan y eliminan la diversidad en la población, lo que lleva a un cambio evolutivo en las poblaciones. **(b)** La recombinación aumenta la tasa de adaptación aliviando el efecto de la deriva génica en poblaciones de tamaño pequeñas y altas tasas de mutación. **(c)** Las poblaciones virales forman un enjambre de genotipos mutantes que rodean una secuencia maestra, o secuencias (grandes nodos en la red), conectadas por una red de mutaciones individuales. Las subpoblaciones individuales (grupos de nodos coloreados) pueden interactuar dentro de la población viral a través de interacciones antagónicas y cooperativas. **(d)** La transmisión de múltiples virus como una sola unidad infecciosa es más probable que produzca células coinfectadas. **(e)** A medida que una población viral se transmite a través de un hospedero, o entre hospederos, se encuentra con múltiples cuellos de botella que reducen el tamaño de la población (área gris). Tomada de Dolan *et al.*, 2016.

1.6. Modularidad

El término **"modularidad"** tiene sus orígenes en el diseño industrial, donde es referido a la técnica que permite construir sistemas grandes por la combinación de subsistemas. Sin embargo, en ciencias biológicas, el término es a menudo utilizado para designar la existencia de "bloques funcionales" en organismos, como se muestra en la Figura 6; pero el significado exacto puede variar dependiendo de la disciplina o incluso del autor (Porcar et al., 2013).

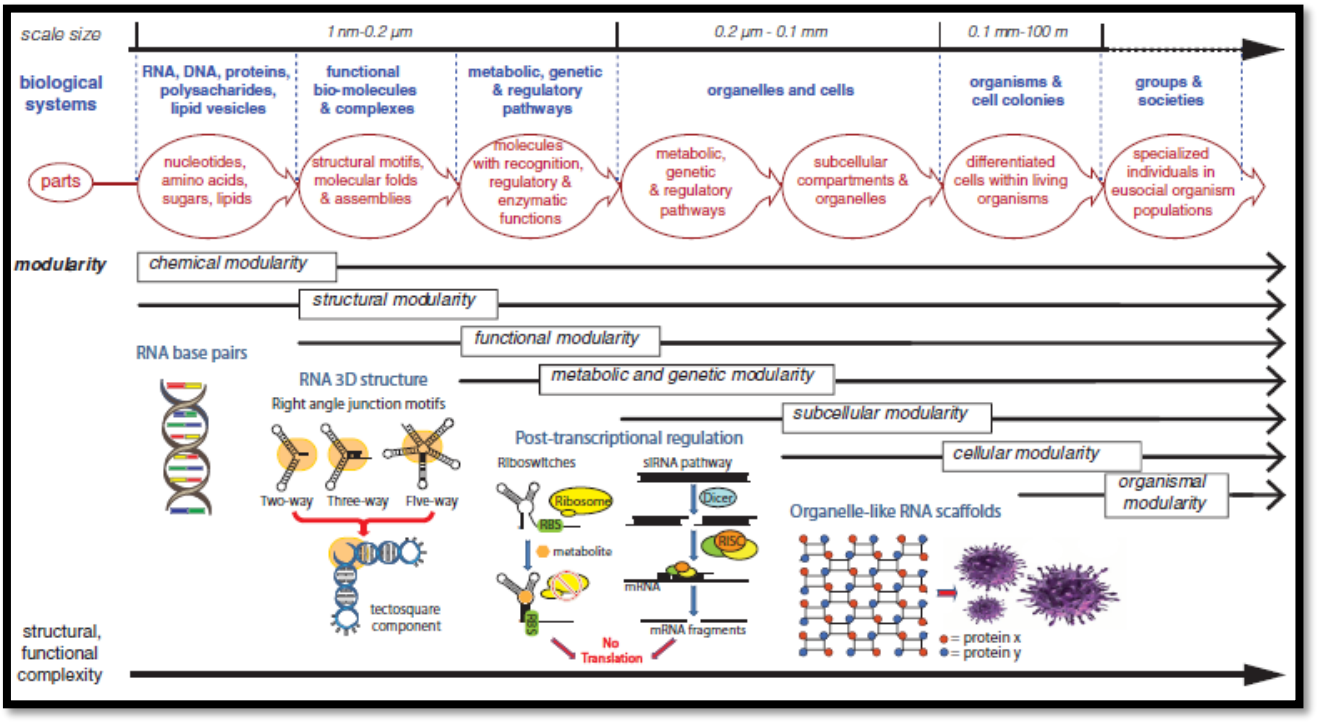

Figura 6. Los múltiples grados de modularidad en sistemas biológicos. Como ejemplo, el ARN es un módulo químico, estructural y funcional que puede ser integrado al nivel de múltiples rutas metabólicas, genéticas y regulatorias, que son partes de componentes subcelulares o unidades celulares. También, el ARN forma circuitos regulatorios que están involucrados en mecanismos de desarrollo que guían a la especialización de organismos individuales. Tomada de Grabow y Jaeger, 2013.

En biología, el concepto de modularidad se refiere a un conjunto de características (por ejemplo, genes) que interaccionan entre sí en unidades discretas, a los que podemos denominar **módulos**, y éstos a su vez interaccionan entre ellos en un proceso determinado (Bonner, 1988). Un cambio genético en uno de los módulos puede ocurrir sin afectar a los demás, lo que permite al organismo permanecer viable frente al cambio genético; la modularidad limita la **pleiotropía** y provee una manera en la cual los organismos puedan evolucionar (Wagner & Altenberg, 1996).

A través del curso de la evolución, el genoma de un organismo y su modularidad funcional se mantienen, los cambios en los módulos pueden permitirle responder y adaptarse mejor a las perturbaciones en el ambiente. La **evolucionabilidad** (*evolvability*) está relacionada con la diversidad presente en las poblaciones naturales para facilitar su adaptación a cambios ambientales a través de la selección natural. Un

gran número de factores juegan un papel importante en determinar la evolucionabilidad; uno de ellos es la modularidad, o independencia funcional de múltiples subunidades en un ensamblaje más grande (Mol *et al.*, 2018).

Un ejemplo de evolución por modularidad es la organización de dominios proteicos que involucra la combinación y la división de dominios para crear, reconstruir o potenciar las funciones biológicas que son necesarias para el desarrollo de la complejidad en los organismos vivos (Wang & Caetano-Anollés, 2009). La modularidad es de gran importancia en la mayoría de las entidades funcionales complejas, ya que permite la innovación y mejora en una subunidad, sin perder su funcionalidad. Por lo tanto, la modularidad podría ser crucial para la evolucionabilidad, facilitando una arquitectura genómica robusta. El proceso natural de evolución, por la vía de la arquitectura genómica modular, potencia la complejidad de los sistemas biológicos (Mol *et al.*, 2018).

La modularidad también ha jugado un papel importante en la evolución de los virus, a pesar de que estos no han evolucionado de un único antepasado común; sin embargo, algunos de ellos forman una densa red evolutiva en la que los genomas están vinculados a través de diferentes genes compartidos. Este tipo de relación evolutiva resulta de un amplio intercambio de genes y módulos genéticos (Koonin *et al.*, 2015).

El descubrimiento del intercambio de módulos genéticos entre diversos virus con poca o ninguna similitud tiene implicaciones en su relación evolutiva. Este intercambio está ilustrado por las proteínas de cápside con plegamiento de "*jelly-roll*", una proteína que representa la subunidad principal de la cápside de los viriones con una estructura icosahédrica; la cápside "*jelly-roll*" está presente tanto en virus de ARN como en los de ADN, incluidos virus como los picornavirus (virus de ARN de una sola cadena y polaridad positiva [ssRNA+]), birnavirus (virus de ARN de doble cadena [dsRNA]), herpesvirus (virus de ADN de doble cadena [dsDNA]), y algunos bacteriófagos de ADN (Holmes, 2011).

También, se ha encontrado que una arquitectura de cápside altamente conservada está incluida en el linaje del "adenovirus PRD1", caracterizado por un doble pliegue de barril β que se encuentra en virus de dsDNA tan diversos como el fago PRD1 y el adenovirus humano, así como en una variedad de virus de archaeas. Dentro del virión, es notable una estructura proteica que se compone de cuatro cadenas de hojas anti-paralelas y dos α-hélices, que es compartida tanto en algunas polimerasas dependientes de ARN como en las dependientes de ADN (Holmes, 2011).

Un ejemplo de modularidad aplicada al estudio de los virus, es el estudio llevado a cabo por Martin y colaboradores en 2005, en el cual utilizan como modelo experimental el Virus de la Vena del Maíz (*MSV: Maize Streak Virus*) el cual consta de tres proteínas: MP, proteína de matriz; CP, proteína de cápside; y Rep, una proteína asociada a la replicación; y dos regiones intergénicas, una larga (*LIR*) y una corta (*SIR*).

Ellos utilizaron cinco cepas de este virus, que como se muestra en la Figura 7, se identifican con diferentes colores. Mediante ingeniería genética, para cada una de las cepas, construyeron diferentes variaciones del genoma, recombinando genes de una cepa con otra, en la figura se muestra el genoma recombinante con diferentes fragmentos de colores.

Posteriormente, infectaron plantas de maíz tanto con los genomas puros como con las construcciones recombinantes, y evaluaron su adecuación biológica mediante la coloración amarillenta de áreas de la hoja (*ICLA: Induced Chlorotic Leaf Areas*). Al analizar la adecuación biológica de las cepas puras respecto de las cepas con una construcción recombinante, se observó que algunas de las recombinantes podían aumentar el nivel de adecuación, sin embargo, algunas otras lo disminuían.

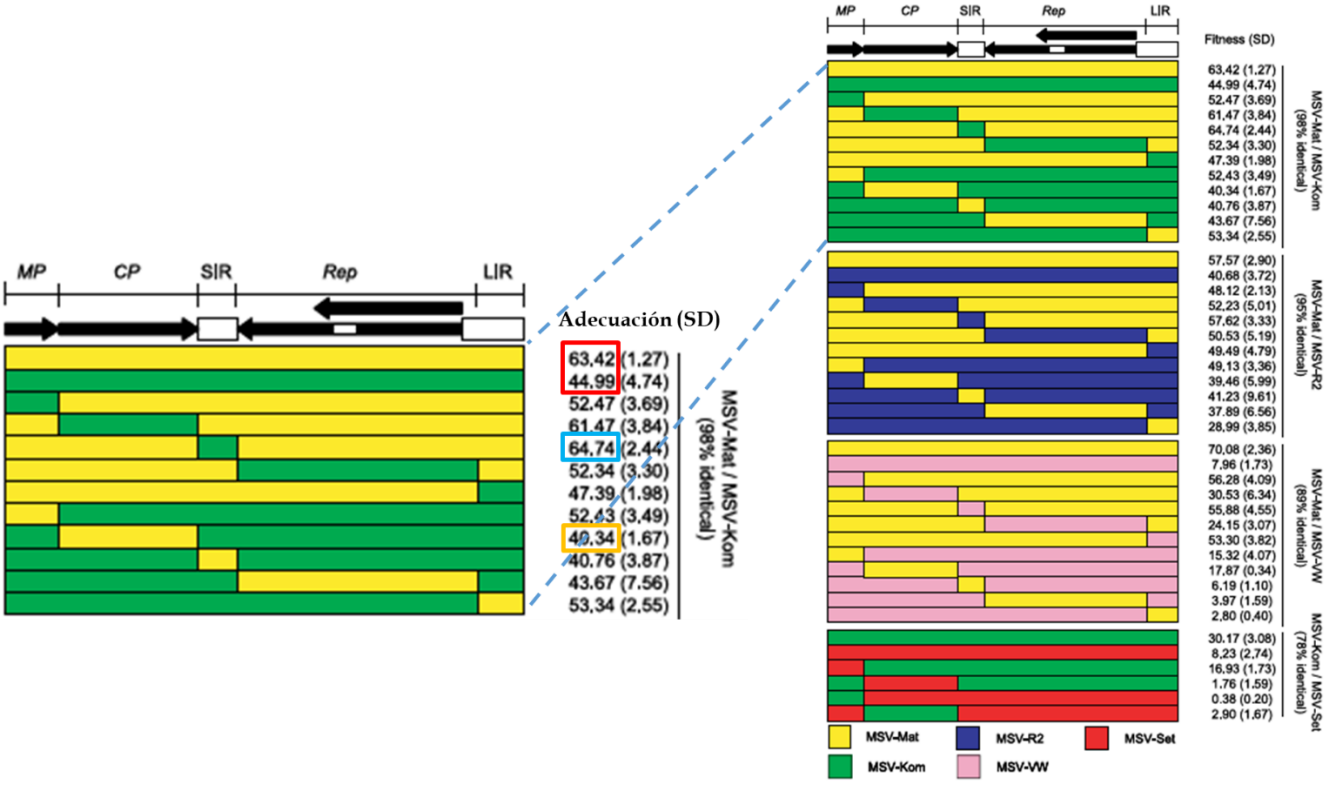

Figura 7. Construcciones recombinantes del MSV y su adecuación biológica relativa. En recuadro color rojo se muestran los valores de la adecuación biológica de los genomas de las cepas de MSV puros, mientras que en el recuadro color azul se indica el valor más alto de adecuación respecto al valor de las cepas puras, de una construcción recombinante de ambas cepas; y en recuadro color naranja, el valor más bajo. MP: proteína de matriz, CP: proteína de cápside, Rep, proteína asociada a la replicación, SIR: región intergénica corta, LIR: región intergénica larga, SD: desviación estándar. Modificada de Martin *et al.*, 2005.

1.7. Robustez mutacional

Todos los sistemas biológicos son resistentes a la variación genética y a los cambios ambientales, una propiedad conocida como "*robustez*". En un sistema robusto, muchas variantes de un gen pueden ser toleradas mientras mantengan el mismo fenotipo (por ejemplo, variaciones genéticas neutrales). Por lo tanto, la "**robustez mutacional**" es la medida en la que la adecuación de un organismo permanece constante a pesar de que ocurran mutaciones en su genotipo (Fares, 2015).

Los virus de ARN exhiben tasas de mutación extremadamente altas, en varios órdenes de magnitud mayor que las de la mayoría de las formas de vida basadas en ADN. Se estima que los virus de ARN generan de 10^{-4} a 10^{-6} errores por nucleótido (Sanjuán *et*

al., 2010), el cual es equivalente a aproximadamente una mutación por genoma, por ciclo de replicación. La tasa de mutación en poblaciones de virus de ARN está peligrosamente cerca de la tasa de error máxima tolerable. La tolerancia mutacional de un virus determinará el tipo (por ejemplo, variación en proteínas estructurales y no estructurales) y la extensión de la diversidad genética que podría ser mantenida en la población. Por lo tanto, la diversidad de la población viral resulta tanto de la generación como de la tolerancia de mutaciones; estos dos factores impulsan la adaptación y la evolución viral (Lauring *et al.*, 2013).

En un estudio realizado por Sardanyés y colaboradores en 2008, ellos modelaron dos poblaciones virales: aquellas que se encuentran en el pico de su adecuación biológica y las que se encuentran en una meseta, 1 y 2, respectivamente, como se muestra en la Figura 8. En el escenario A, donde la tasa de mutaciones (μ) es menor a la tasa de mutaciones crítica (μ_{crit}), en la cual el 95% de los individuos de una población pierde su habilidad para permanecer en el pico, ambas poblaciones se mantienen constantes. Mientras tanto, en el escenario B, donde μ es mayor que la μ_{crit}, la población 1 se desploma, mientras que la población 2 permanece constante en su adecuación, esto es debido a la robustez mutacional. Este proceso es conocido como "*supervivencia del más plano*" (*survival of the flatness*).

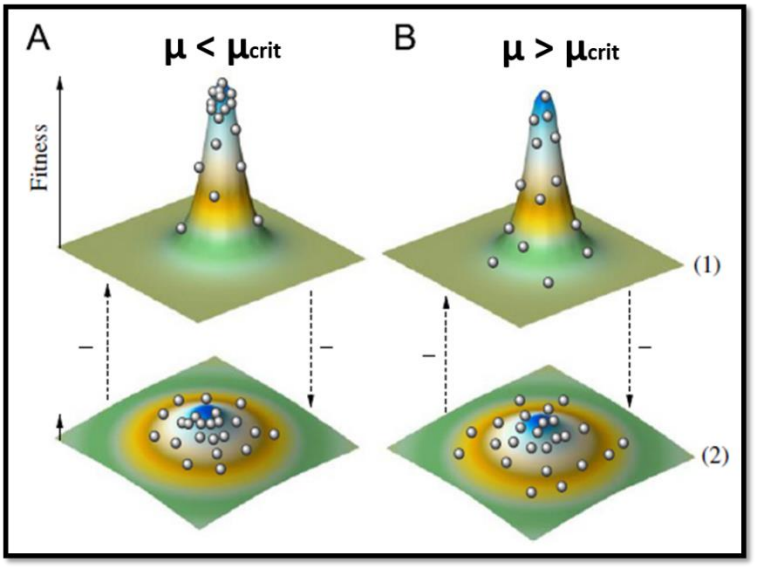

Figura 8. Supervivencia del más plano. La imagen es explicada en el texto. Las flechas punteadas indican las cuasiespecies entre las poblaciones. Modificada de Sardanyés *et al.*, 2008.

2. ANTECEDENTES PARTICULARES

2.1. Mecanismos de evolución del VIH-1

El VIH-1 es uno de los organismos que evoluciona más rápido. Tal evolución rápida es el resultado de una combinación de factores. Primero, el virus experimenta una alta tasa de mutación, ya que la transcriptasa reversa produce ~0.2 errores por genoma durante cada ciclo de replicación, y otros errores que ocurren durante la transcripción del ADN por la polimerasa Pol II. En segundo lugar, el VIH-1 tiene un tiempo de generación viral de ~2.5 días y produce ~10^{10}-10^{12} viriones nuevos cada día. Finalmente, la recombinación frecuente eleva aún más su tasa de cambio evolutivo (Rambaut *et al.*, 2004).

2.1.1. Tasa de mutación del VIH-1

Al igual que otros virus de ARN, la gran diversidad del VIH-1 proviene de su alta tasa de mutación espontánea. La ausencia de *proofreading* de la RT del VIH-1 resulta en una tasa de error estimada en el orden de $3x10^{-5}$ por base por ronda de copiado. Sin embargo, estos estimados podrían subestimar el proceso de mutación en pacientes con VIH-1, ya que factores celulares, como los niveles de dNTPs pueden afectar la frecuencia y el tipo de mutaciones producidas. Además, el VIH-1 está sujeto a la edición por enzimas celulares, incluido el complejo de edición de la apoliproteína B (APOBEC), que media la edición de citosina a uracilo en la cadena negativa del ADN viral, resultando en sustituciones de G → A en el ARN genómico viral (Cuevas et al., 2015).

2.1.2. Dinámica de replicación del VIH-1

Luego de la infección con VIH-1, los pacientes típicos entran en un período asintomático que dura varios años durante el cual el virus está presente en niveles bajos a medios, típicamente una carga viral de <50 a $3x10^4$ copias por mL de plasma. Sin embargo, el número de viriones producido y eliminado es estimado en alrededor del 10^{10} por día, con un tiempo de generación de un poco menos de dos días (Neher & Leitner, 2010).

La capacidad replicativa (RC) del virus está correlacionada con su virulencia. La RC es definida como el número promedio de células infectadas por una célula infectada típica (directamente o por la producción viral e infección) en presencia de un alto número de células blanco y en ausencia de respuestas inmunes efectivas. Las mutaciones encontradas en virus con adecuación biológica baja incluyen tanto mutaciones de resistencia a fármacos como mutaciones de escape a la respuesta de linfocitos T citotóxicos (CTL), en particular a algunos alelos como el HLA-B57, el cual se ha asociado con baja progresión al SIDA. Es probable que las mutaciones de escape con RC baja evolucionen cuando la presión de selección ejercida por el sistema inmune del hospedero es más fuerte que la selección intrínseca para aumentar la RC (Fraser et al., 2014).

El aumento de la carga viral que ocurre de manera progresiva durante la fase crónica conlleva al desarrollo de una población de virus diversa y divergente. Por ejemplo, las mutaciones de escape pueden resultar un costo replicativo, con implicaciones para la carga viral y la transmisibilidad. Aun así, la adecuación viral puede ser restablecida por las interacciones epistáticas y las mutaciones compensatorias (Theys *et al.*, 2018).

2.2. El ancestro del VIH-1 tiene un origen recombinante

El ancestro del VIH-1 ha sido trazado al SIV_{cpz}, el cual infecta chimpancés (*Pan troglodytes*) en África occidental y central, pero el origen del SIV_{cpz} aún es desconocido (Bailes *et al.*, 2003). El estudio de Bailes y colaboradores en 2003, mostró que el SIV_{cpz} tiene un origen híbrido; como se muestra en la Figura 9, se construyeron árboles filogenéticos utilizando secuencias de los genes *pol* y *env* de varios genomas de SIVs que infectan diferentes especies de primates. En el árbol filogenético construido con secuencias del gen *pol*, el SIV_{cpz} se agrupa con el SIV que infecta a los mangabeys de tapa roja (*Cercocebus torquatus*), SIV_{rcm}, mientras que en el árbol construido con secuencias del gen *env* se agrupa con el SIV que infecta a los monos mayores de nariz plana (*Cercopithecus nictitans*), SIV_{gsn}, infiriendo que el SIV_{cpz} podría ser un virus recombinante.

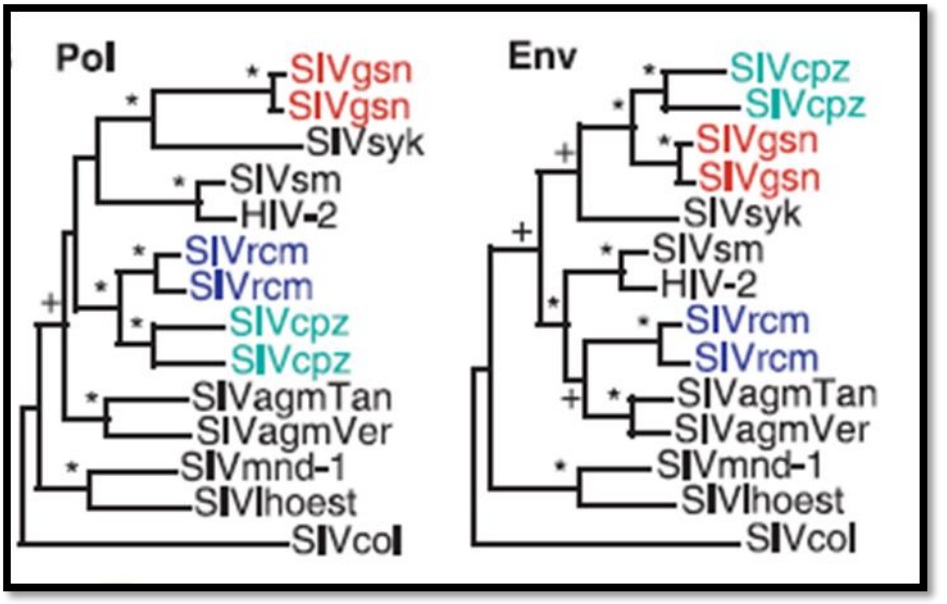

Figura 9. Filogenia de máxima verosimilitud de secuencias de Pol y Env de SIVs. Ramas internas encontradas en al menos el 70% y el 95% de las réplicas de bootstrap son indicadas con + y *, respectivamente. Tomada de Bailes *et al.*, 2003.

La cantidad de variación genética que constituye una población está determinada por la tasa de recombinación, el desequilibrio de ligamiento y los caracteres con patrones de dominancia. En el VIH-1, la tasa de recombinación disminuye a medida que aumenta la diversidad de secuencias, por lo tanto, la recombinación intra-subtipos es más frecuente; por ejemplo, en la región del gen *pol*, la frecuencia de recombinación entre dos variantes B (que comparten un 96% de similitud en su secuencia de nucleótidos) es aproximadamente un 30% mayor que entre los subtipos B y F (con un 90% de similitud en su secuencia de nucleótidos). La recombinación entre subtipos de HIV-1 ha sido identificado como un mecanismo importante para la diversificación del grupo M (Vuilleumier & Bonhoeffer, 2015).

2.3. La recombinación y su efecto sobre la evolución del VIH-1

Como se muestra en la Figura 10, el proceso de recombinación es conocido como un mecanismo que mantiene la diversidad genética, limita el acumulamiento de mutaciones deletéreas y permite la asociación de mutaciones benéficas (Vuilleumier &

Bonhoeffer, 2015; Simon-Loriere & Holmes, 2011). La naturaleza combinatoria de la recombinación puede proporcionar a los organismos muchas más opciones evolutivas de las que están disponibles sólo a través de la mutación (Martin *et al.*, 2005). La recombinación produce nuevas combinaciones de alelos de la variación genética existente y aleatoriza la distribución de los genotipos (Neher & Leitner, 2010).

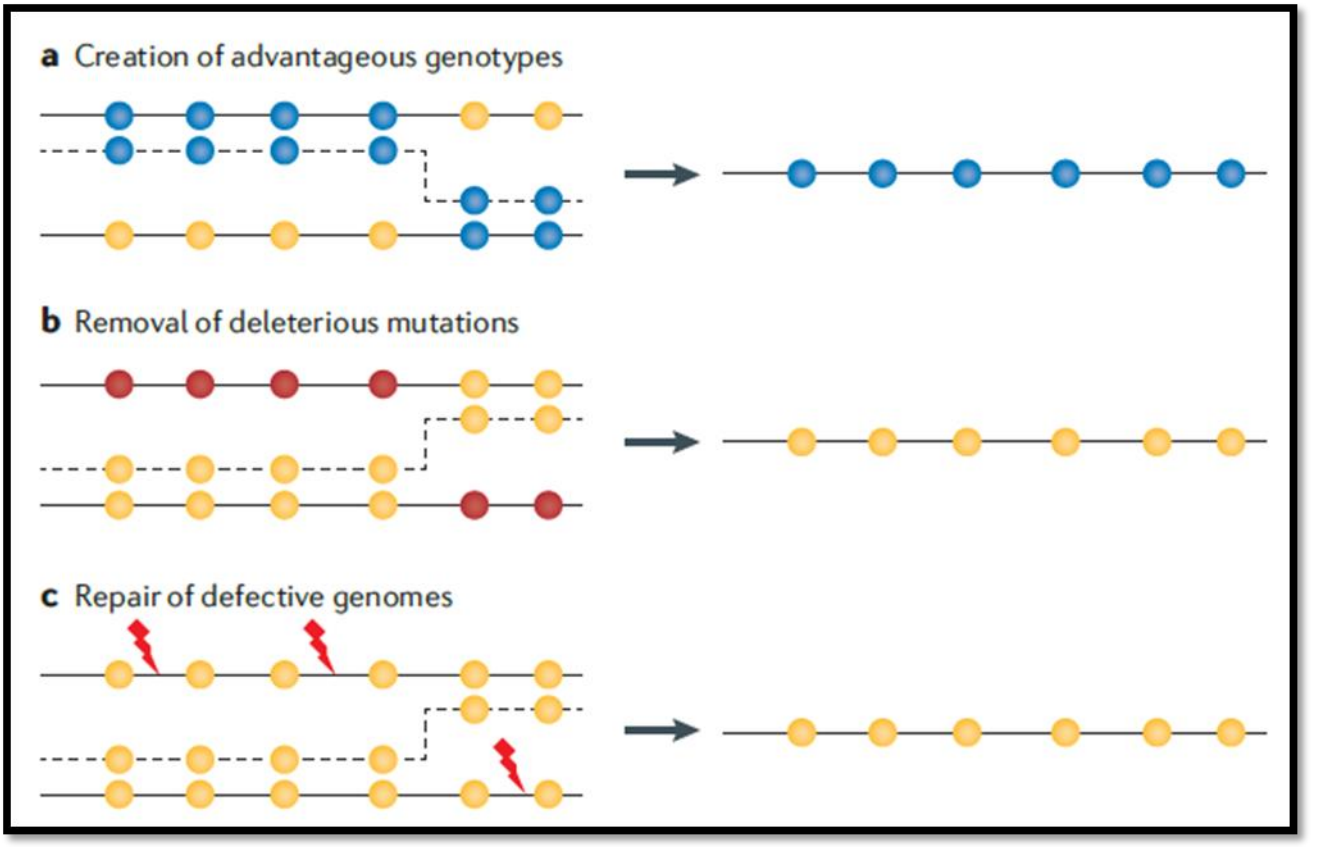

Figura 10. Consecuencias evolutivas de la recombinación. **(a)** La recombinación puede crear combinaciones ventajosas de mutaciones (círculos azules) que incrementan la tasa de evolución. **(b)** La recombinación puede eliminar mutaciones deletéreas (círculos rojos) y reestablecer el genotipo original. **(c)** La recombinación también puede generar un genoma funcional a partir de moléculas parentales dañadas. El daño genético (como roturas de filamentos o modificaciones oxidativas de la base) está representado por símbolos de rayos rojos. Tomada de Simon-Loriere & Holmes, 2011.

La Figura 11 trata de mostrar que la relevancia efectiva de la tasa de recombinación depende de la probabilidad de co-infección de una célula con más de un virus (Neher & Leitner, 2010). La recombinación en el VIH-1 podría haber tenido un papel importante en su historia evolutiva y en las epidemias globales; además, la

recombinación pudo haber homogenizado la composición genómica de la población del VIH-1 ya que es un mecanismo efectivo que desencadena la evolucionabilidad (Vuilleumier & Bonhoeffer, 2015). Los procesos de evolución rápida son desencadenados cuando en las poblaciones la tasa de mutaciones es más baja que la tasa de recombinación, ya que, en poblaciones grandes, la adaptación puede ser impulsada por selección debido al cambio ambiental en lugar de por nuevas mutaciones (Weissman & Hallatschek, 2014).

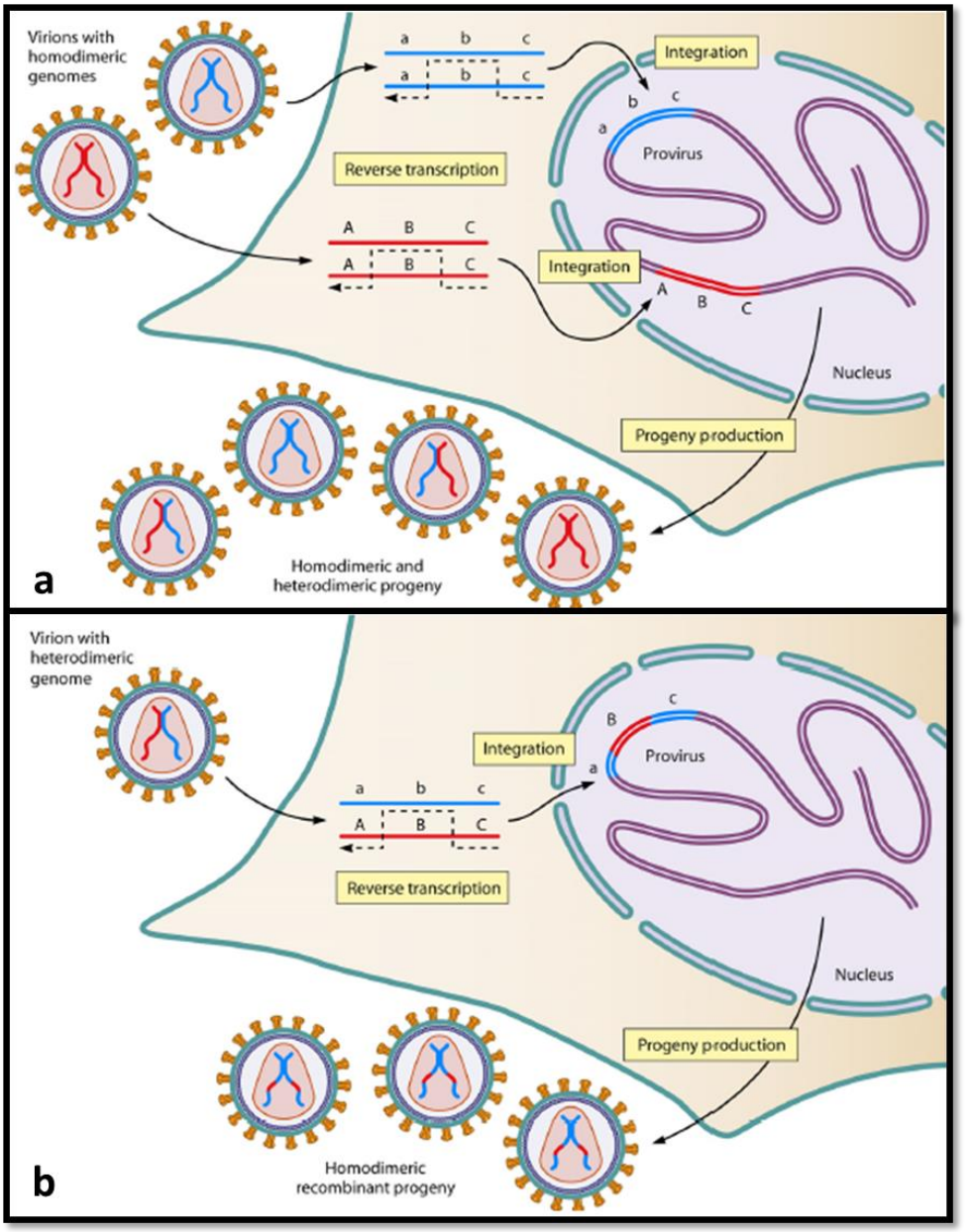

Figura 11. Formación de genomas recombinados. **(a)** La co-infección con dos virus genéticamente distintos no produce recombinantes. Sin embargo, una célula debe co-infectarse con estos dos virus

(mostrados aquí como partículas virales con dos moléculas de ARN de color azul o rojo) para producir partículas virales con ARNs heterodiméricos (partículas que contienen una cadena de ARN roja y una azul). **(b)** La recombinación se observa en células infectadas con viriones heterodiméricos. El cambio de molde durante la transcripción reversa puede generar un provirus recombinante. Tomada de Onafuwa & Telesnitsky, 2009.

El genoma del VIH-1 experimenta 1.51 eventos de recombinación por cada 1000 nucleótidos, mientras que, bajo las mismas condiciones ocurren 0.12 eventos de mutación (Vuilleumier & Bonhoeffer, 2015). Las regiones con baja recombinación podrían tener un papel importante en el mantenimiento de la integridad del genoma y podrían acumular divergencia adaptativa, ya que la interacción entre dos fuerzas evolutivas comunes, la selección natural y el flujo génico, podrían estar ligadas a que ocurra la adaptación de estas regiones del genoma (Samuk *et al.*, 2017). Las regiones de baja recombinación incrementan divergencia entre subtipos, pero la decrecen dentro de ellos; estos patrones podrían explicar las tasas de restricción de recombinación inter-subtipos en regiones a lo largo del genoma del VIH-1 (Vuilleumier & Bonhoeffer, 2015).

Los segmentos de subtipos específicos parecen estar localizados preferentemente en regiones genómicas recombinantes definidas, sugiriendo que algunos segmentos del genoma del VIH-1 son incorporados preferencialmente por el proceso de introgresión por algunos subtipos puros durante la recombinación (Vuilleumier & Bonhoeffer, 2015).

La recombinación puede ocurrir en todo el genoma viral; sin embargo, se ha propuesto que varias regiones del genoma del VIH-1 son "*hot spots*" de recombinación. Uno de los *hot spots* mejor definidos se encuentra en el gen de la envoltura (*env*). Se ha demostrado que una estructura secundaria de ARN en particular promueve la recombinación en esta región, y la destrucción de dicha estructura reduce la tasa de recombinación. Recientemente, se ha propuesto que una región cercana al tracto central de poli-purina (cPPT) sea un *hot spot* de recombinación. Se hipotetizó que una región rica en guaninas (Gs) puede formar un cuarteto G intermolecular, facilitando así la dimerización y la recombinación del ARN (Nikolaitchik *et al.*, 2015).

2.3.1. Transcripción reversa en el VIH-1

Cuando un virión maduro del VIH-1 infecta una célula blanco susceptible, las interacciones entre las glicoproteínas de la envoltura con el receptor y correceptor en la superficie de la célula inducen la fusión de las membranas, tanto de la célula del hospedero como la del virión. Esta fusión introduce el contenido del virión dentro del citoplasma de la célula, preparando el escenario para la transcripción reversa (Hu & Hughes, 2012).

En la Figura 12 se muestra el proceso de transcripción reversa, el cual se puede llevar a cabo en cinco pasos:

1) El proceso es iniciado por la síntesis de la cadena negativa (mostrada en color azul en la figura) de ADN [(-) ssDNA] catalizada por la RT usando un ARNt del hospedero como cebador, $ARNt^{Lys}$, el cual se une al sitio de unión del cebador (PB) localizado aproximadamente 180 nt desde el extremo 5´ del genoma de ARN (en color rojo) (Basu *et al.*, 2008). La síntesis del ADN crea un heterodúplex ARN-ADN, el cual es sustrato para el dominio de RNAsa H de la RT, llevando a cabo la degradación del ARN genómico (Hu & Hughes, 2012). Al llegar la RT al extremo 5´ del genoma, ha generado la cadena negativa intermedia de ADN (Basu *et al.*, 2008).

2) Los extremos del ARN viral son repetidos directos, denominados "R". Estos repetidos actúan como un puente que permite la transferencia de la síntesis al segmento R del extremo 3´ y así poder sintetizar toda la cadena negativa de ADN (Hu & Hughes, 2012). Simultáneo con la síntesis de la cadena negativa, la actividad de RNAsa H de la RT escinde el genoma de ARN recién copiado en fragmentos cortos. Un fragmento de ARN, que es resistente a la escisión de la RNAsa H porque posee una secuencia única del tracto de polipurina (PP), permanece asociado con el ADN naciente (Basu *et al.*, 2008).

3) Este oligómero de ARN sirve como cebador para la síntesis de ADN de cadena positiva (en color verde) (Basu *et al.*, 2008). Cuando la RT genera el ADN de cadena positiva, no sólo copia el ADN de cadena negativa, sino que también los primeros 18

nt del cebador ARNt^{Lys}. Una vez que el cebador ha sido copiado en ADN, se convierte en sustrato para la RNAsa H (Hu & Hughes, 2012).

4) El extremo 3´ del ssDNA(+) debe transferirse al extremo 3´ de la cadena negativa para completar la replicación viral. Este proceso se basa en la homología entre las regiones PB y también la escisión por medio de la RNAsa H tanto del PP como del cebador (Basu *et al.*, 2008).

5) Una vez que ambas secuencias se alinean, la síntesis de ADN por parte de la RT extiende las cadenas positiva y negativa hasta el extremo 3′ de ambas (en color negro) (Hu & Hughes, 2012). La finalización de la síntesis da como resultado la generación de ADN bicatenario con extremos de repetidos terminales largos (LTR) duplicados (Basu *et al.*, 2008).

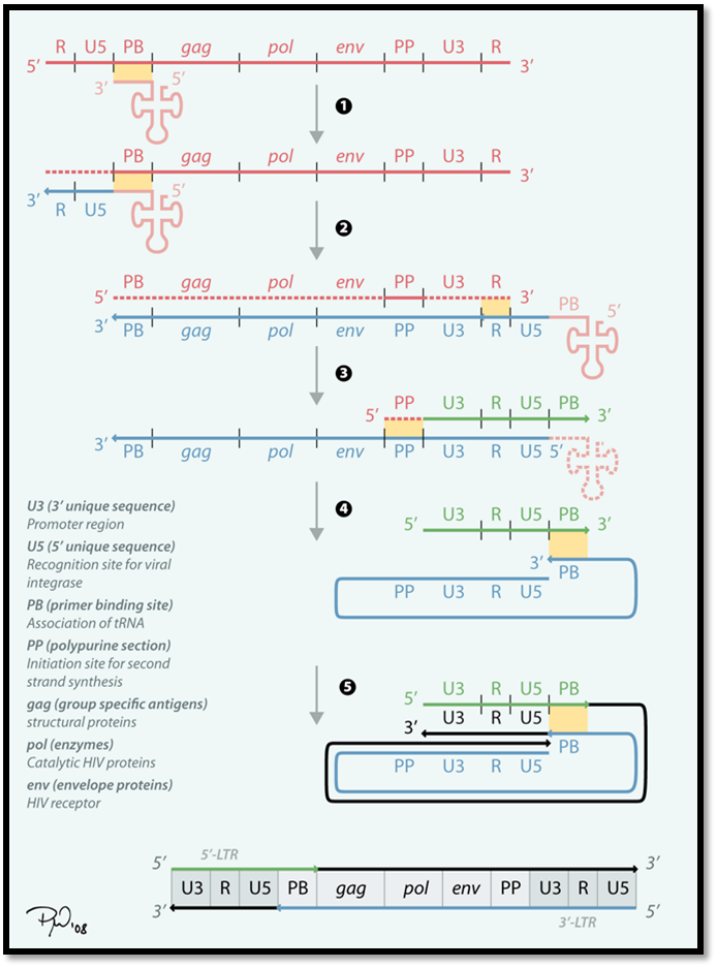

Figura 12. Transcripción reversa en el VIH-1. Cada uno de los pasos es explicado en el texto. Color amarillo: sitios con complementariedad. Tomada de Modrow *et al.*, 2003.

2.3.2. Proceso de recombinación en el VIH-1

La recombinación genética es una parte integral del ciclo de replicación del VIH-1 (Rambaut *et al.*, 2004). Después de la infección de una célula blanco, la transcripción reversa generará una molécula de ADN de doble cadena que será integrada en el genoma del hospedero; un proceso que ha sido observado extremadamente frecuente en el VIH-1 es el "*template switching*" entre las dos copias del ARN genómico durante la transcripción reversa (Galetto *et al.*, 2005). Este proceso, como se muestra en la Figura 13, es cuando la RT cambia del molde donador (cadena de ARN que es copiada antes del cambio, color rojo en la figura) al molde aceptor (cadena de ARN que es transferida después del cambio, color azul) (Onafuwa & Telesnitsky, 2009).

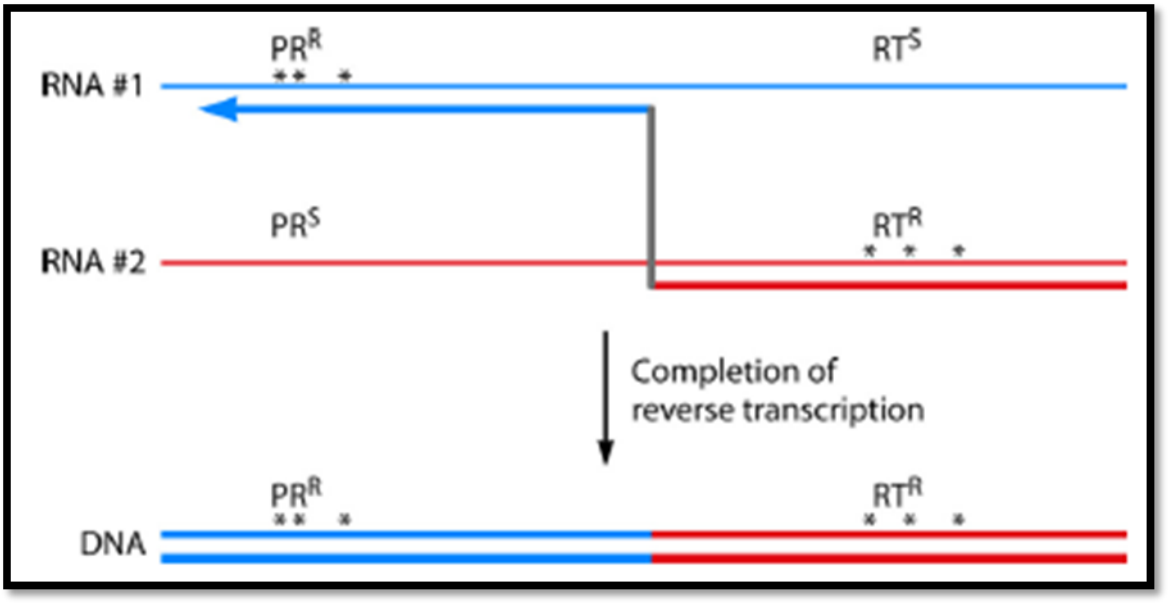

Figura 13. Proceso de "*Template Switching*". Las líneas delgadas (azul y rojo) representan ARNs genómicos genéticamente distintos; las líneas gruesas representan el ADN viral. La flecha muestra la dirección de la síntesis de ADN, y los asteriscos representan sitios de mutaciones que confieren resistencia a inhibidores de RT o PR. En este caso, el cambio de molde genera un provirus recombinante que es resistente a ambos tipos de inhibidores. Tomada de Onafuwa & Telesnitsky, 2009.

La recombinación ocurre cuando la RT cambia de molde entre las cadenas alternas de ARN genómico durante la replicación, en un proceso conocido como recombinación por "*copy choice*" (Rambaut *et al.*, 2004). Se ha propuesto que las moléculas de ADN recombinantes se generan durante la síntesis tanto de la primera cadena de ADN (ADN

de polaridad negativa) como de la segunda (ADN de polaridad positiva) mediante el proceso de "*copy choice*" y esto es posible gracias al cambio de molde durante la transcripción reversa (Galetto & Negroni, 2005).

La Figura 14 muestra el proceso de "*copy choice*", el cual consta de cinco pasos: **1)** Las ARN polimerasas dependientes de ARN (RdRp) inician la síntesis de hebras nacientes en el extremo 3´ del ARN genómico (molde donante de ARN, en color azul). **2)** Si la RdRp se detiene durante el alargamiento, **3)** ésta puede disociarse del molde donante y re-asociarse con otro, llamado "molde aceptor de ARN" (en color rojo). **4)** Finalmente, la RdRp reanuda la síntesis de la cadena naciente en la plantilla del aceptor. **5)** Obteniendo como resultado una cadena de ADN recombinante (Sztuba-Solińska *et al.*, 2011).

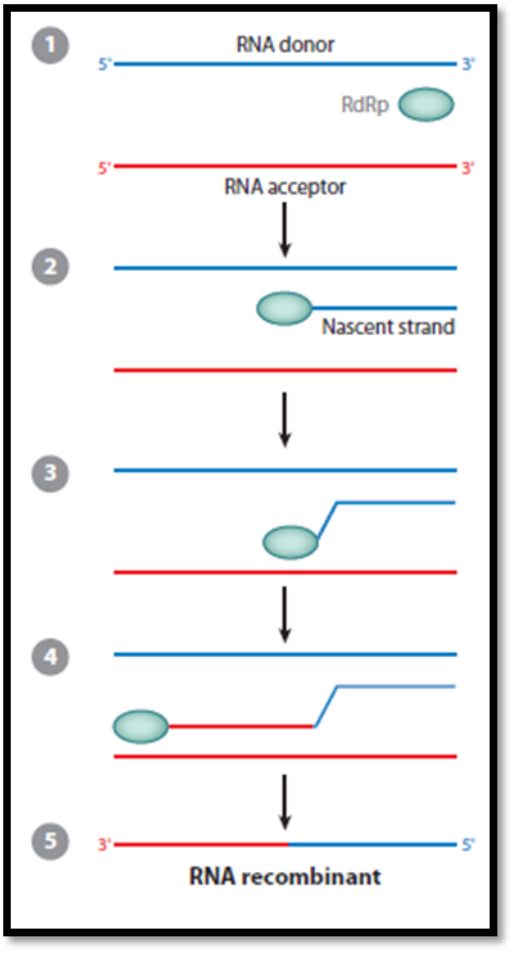

Figura 14. Proceso de "*Copy Choice*". Cada uno de los pasos es explicado en el texto. Tomada de Sztuba-Solińska *et al.*, 2011.

Durante la síntesis de la cadena de ADN de polaridad negativa, la RT puede usar segmentos de cada cadena de ARN como molde para generar un ADN quimérico que contiene fragmentos de cada uno de los dos ARN genómicos (Hu & Hughes, 2012). Aunque también, la recombinación puede ocurrir en partículas homocigotas, que contienen dos copias de ARN derivadas del mismo provirus, los recombinantes resultantes tienen el mismo genotipo que el virus parental y es difícil identificarlos (Nikolaitchik *et al.*, 2015).

2.3.3. Mecanismos de recombinación

Se han propuesto varios mecanismos para explicar el proceso de "*copy choice*", pero la idea más aceptada es que ocurre un alineamiento entre el ADN naciente y el ARN aceptor, a este paso se le conoce como "*docking*". Posteriormente ocurre la transferencia del extremo 3´ del ADN naciente al aceptor de ARN. La presencia de la actividad de RNAsa H de la RT hace que el extremo 3' del ADN naciente se encuentre en forma de cadena sencilla lo que le permite alinearse con la otra copia del ARN genómico (proceso de "*docking*"). Este evento podría generar una estructura de ADN ramificado que posteriormente podría ser resuelto tanto por la ligasa de ADN como las nucleasas celulares. Las evidencias obtenidas sugieren que la recombinación durante la síntesis de la primera cadena de ADN es el mecanismo predominante en la recombinación del VIH-1 (Galetto & Negroni, 2005).

Por lo tanto, se han propuesto tres modelos para el mecanismo de recombinación del VIH-1, los cuales se resumen en la Figura 15:

a) Modelo de opción de copia forzada

La presencia de "*breaks*" en el ARN genómico fue propuesto como hipótesis para explicar la alta frecuencia de eventos de recombinación que ocurren en el VIH-1. Se ha propuesto que cada "*break*" en el molde de ARN podría forzar a que la maquinaria de síntesis de ADN sea transferida a la otra copia del ARN genómico. Estas pausas en el proceso de transcripción reversa debidas a los "*breaks*" en el ARN molde podría permitir una degradación extensiva del molde de ARN por la actividad de RNAsa H de la RT (Galetto & Negroni, 2005).

b) Modelo de pausa impulsada

Este modelo indica que el cambio de molde durante la transcripción reversa podría ocurrir eficientemente en ausencia de "*breaks*" en el molde de ARN. Esto estaría mediado por la presencia de un sitio de pausa fuerte en la región del molde de ARN, debido a la degradación de éste por la actividad de RNAsa H de la RT; lo anterior sugiere que dichas pausas podrían constituir un detonante para el cambio de molde (Galetto & Negroni, 2005).

c) Modelo impulsado por la estructura secundaria del ARN

La presencia de regiones con estructuras secundarias en el molde de ARN, como la secuencia de iniciación del dímero (DIS), la estructura de "*hairpin*" de TAR en la secuencia R, y la región codificante para la porción C2 de gp120, son regiones propensas para promover la recombinación (Galetto & Negroni, 2005).

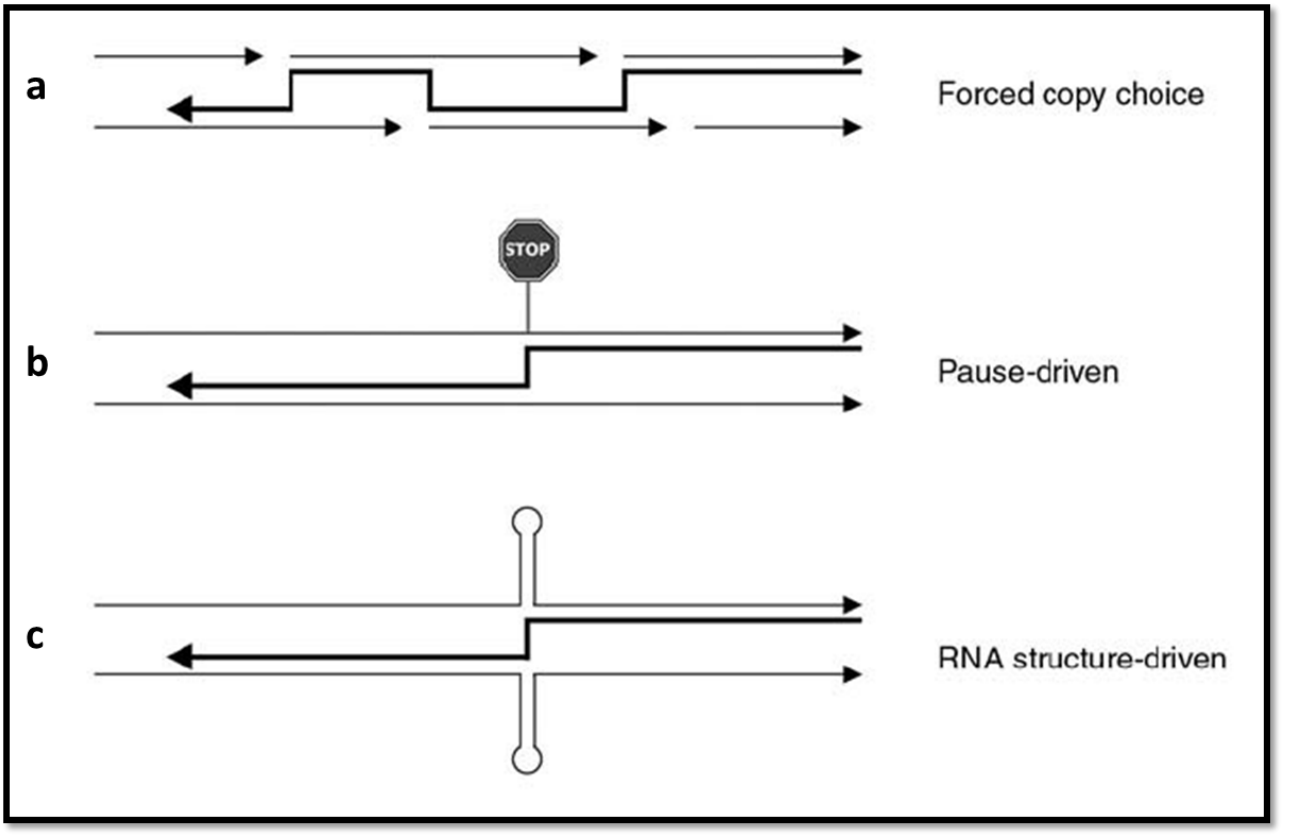

Figura 15. Modelos de recombinación en VIH-1. **(a)** Modelo de opción de copia forzada, mecanismo principal. **(b)** Modelo de pausa impulsada, en ausencia del modelo anterior. **(c)** Modelo impulsado por la estructura secundaria del ARN, llevado a cabo en regiones importantes para la activación de la transcripción y la modulación del splicing. Modificada de Galetto & Negroni, 2005.

2.4. La selección natural y su implicación en el VIH-1

La mayoría de los estudios sobre la evolución intra-hospedero del VIH-1 indican que la respuesta inmune del hospedero genera una presión selectiva fuerte sobre el virus, medida en la tasa de sitios de **sustituciones no sinónimas** (dN, cambio de aminoácido) entre **sustituciones sinónimas** (dS, cambio silencioso) (dN/dS); por lo tanto, en pacientes con fases asintomáticas más largas se observa mayor presencia de selección positiva (Rambaut *et al.*, 2004). Una observación de dN/dS >1 indica selección positiva para la variante nueva, mientras que un dN/dS <1 indica que la secuencia de aminoácidos cambia mucho más lento que la secuencia de nucleótidos, indicando una limitación funcional al nivel de la proteína (Neher & Leitner, 2010).

Evidencia de la importancia de la selección natural en la evolución del VIH-1 se encuentra en estudios de ambos, tanto del hospedero como del virus. Del lado del hospedero, está bien establecido que la respuesta inmune en contra de la infección del VIH-1 es principalmente orquestada por los linfocitos T, dentro de los cuales, las células T CD8+ citotóxicos (CTLs) juegan un papel vital en el reconocimiento de epítopos presentados por las moléculas del MHC clase I. Del lado del virus, hay evidencia que el VIH-1 es capaz de escapar del reconocimiento de los CTLs durante una infección. Varios reportes sugieren que el VIH-1 puede responder a presiones de selección impuestas por los CTLs, al fijar mutaciones o eliminar aminoácidos (Zanotto et al., 1999). La amplia diversidad genética del VIH-1 en un individuo infectado es, al menos en parte, el resultado de la selección positiva mediada por inmunidad (Ross & Rodrigo, 2002).

La acción de la selección natural en el gen *env* es evidencia de patrones de sustituciones sinónimas y no sinónimas en esta secuencia. La tasa de sustituciones no sinónimas es más alta que la tasa de sustituciones sinónimas en algunas regiones de *env*, lo cual es una clara indicación de selección positiva. Además, entre individuos infectados, la fuerza de selección y la frecuencia relativa de mutaciones adaptativas están asociadas con el tiempo de progresión al SIDA. La fuerza selectiva primaria actuando en el gen *env* es la evasión de la respuesta inmune (Williamson, 2003).

2.5. Patrones de "mosaicismo recombinante" en VIH-1

Actualmente, no existe ningún estudio que aborde específicamente el tema de la modularidad en el VIH-1, sin embargo, hay algunos trabajos que mencionan que este virus tiene patrones de *"mosaicismo recombinante"*, haciendo referencia a que contiene módulos tanto de un subtipo como de otro, incluso, hasta fragmentos de algunas CRFs.

Uno de los primeros estudios que abordan este tema, es el estudio realizado por Wang y colaboradores en 2015, en el cual utilizaron 140 muestras de suero de personas VIH-1+ de la provincia de Yunnan, China. La amplificación y secuenciación de las regiones *gag*-RT y del gen de la RT, y su análisis usando la herramienta RIP (*Recombinant Identification Program*) de la base de datos de Los Álamos les permitió identificar el subtipo o CRF al que pertenece cada fragmento, como se muestran en la Figura 16. A esta característica, la denominaron "*estructuras recombinantes*" del VIH-1.

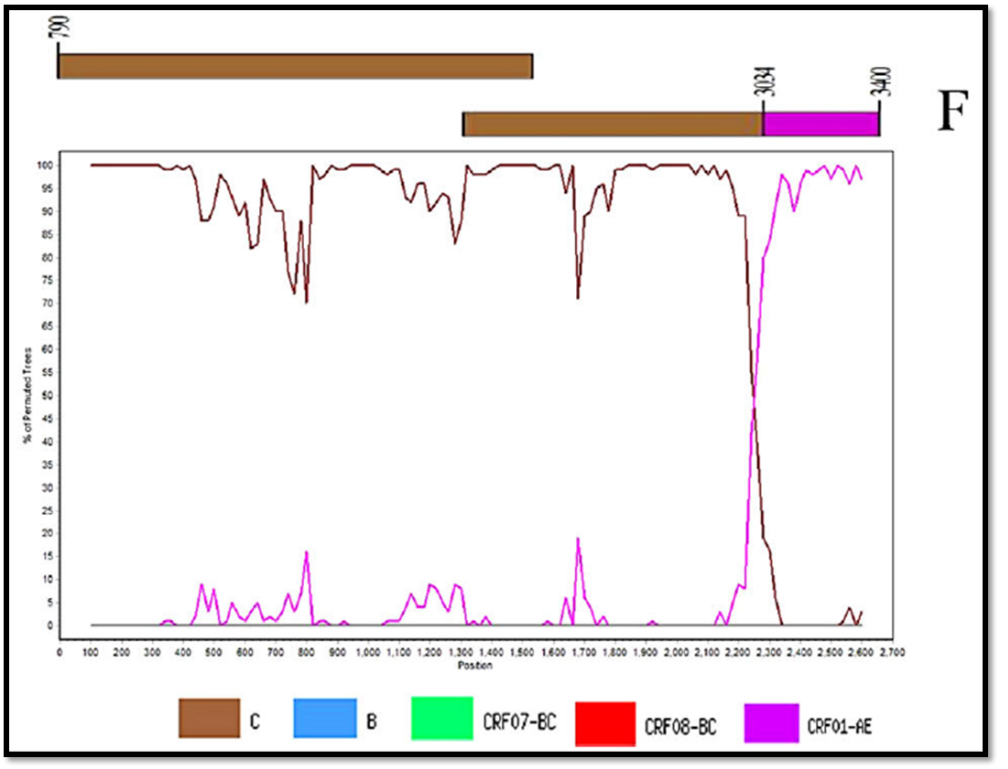

Figura 16. Estructuras recombinantes. En el eje horizontal se muestra la longitud del fragmento en nucleótidos; mientras que en el eje vertical se muestra la similitud de la secuencia blanco con las secuencias de entrada. Tomada de Wang *et al.*, 2015.

Mientras que, un segundo estudio, desarrollado por Delatorre y colaboradores en 2017, se secuenció el gen *pol* de 140 muestras de pacientes VIH-1+ de la región noroeste de Brasil. Mediante una herramienta de subtipificación del VIH-1, REGA v3.0, generaron gráficos como los que se observan en la Figura 17, donde muestran el subtipo al que pertenece cada fragmento del gen *pol* para cada secuencia blanco. A tal característica la denominaron "*patrones de mosaicismo recombinante*" del VIH-1.

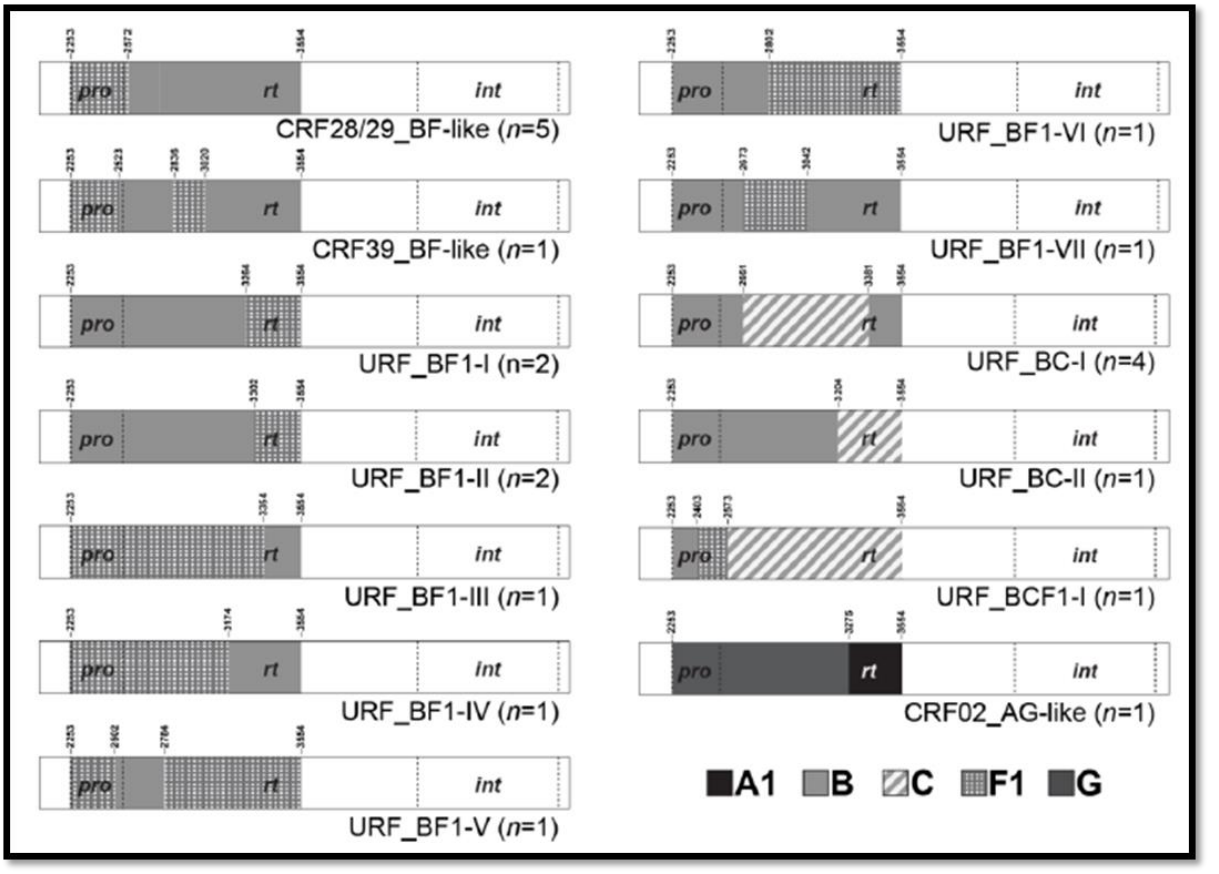

Figura 17. Patrones de mosaicismo recombinante. Los segmentos están coloreados de acuerdo a el subtipo asignado como se describe en el código inferior. Fragmentos no secuenciados fueron representados en color blanco. Líneas punteadas delimitan las regiones de la proteasa (pro), transcriptasa reversa (rt) e integrasa (int). Tomada de Delatorre *et al.*, 2017.

Finalmente, un estudio realizado por Reis y colaboradores en 2017, donde secuenciaron el genoma completo de 828 muestras de pacientes VIH-1+ del oeste central de Brasil, identificaron los puntos de recombinación presentes en éstos. El uso del método de bootscan implementado en el software Simplot v3.5.1, les permitió caracterizar una

nueva forma recombinante, la CRF90_BF1. En este trabajo, los autores encontraron 7 regiones no recombinantes en la CRF90_BF1, como se muestra en la Figura 18.

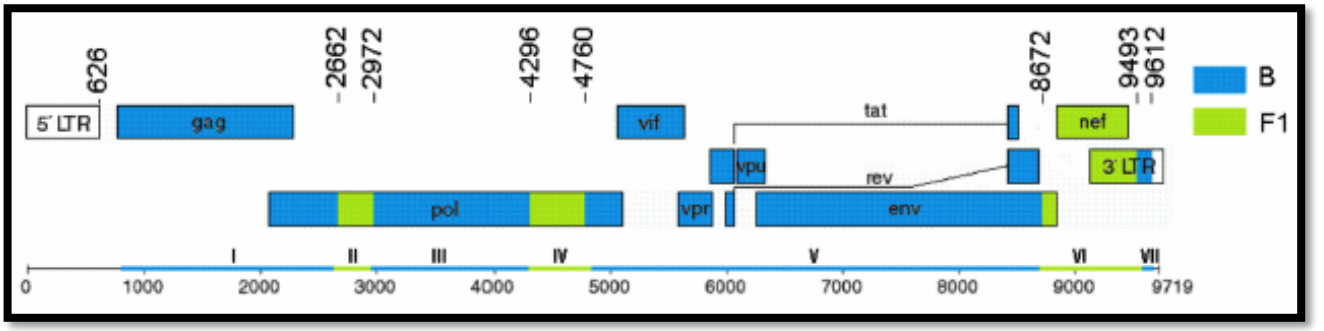

Figura 18. Puntos de recombinación en la CRF90_BF1. Se muestra en números verticales la posición de cada uno de los 8 puntos de recombinación, mientras que, en color azul y verde se muestran cada una de las "*regiones*", posteriormente subtipificadas, del subtipo B y el subtipo F1, respectivamente. Tomada de Reis *et al.*, 2017.

3. JUSTIFICACIÓN

El grupo M, también denominado "forma pandémica del VIH-1" (Sharp y Hahn, 2011) ha infectado aproximadamente a 75.7 millones de personas y causado al menos 32.7 millones de muertes (UNAIDS, 2020). Actualmente, se estima que las formas recombinantes son las responsables de al menos el 20% de las infecciones de VIH-1 a nivel mundial (Hemelaar, 2013).

Comprender los mecanismos que gobiernan la evolución del VIH-1 es crucial para reconstruir su origen, descifrar su interacción con el sistema inmune y poder desarrollar estrategias de control efectivas; de igual manera, el proceso de recombinación tiene implicaciones importantes en la comprensión de la epidemia del VIH-1. La recombinación interactúa con la selección y la deriva génica para producir dinámicas de población complejas, y proporciona un mecanismo eficiente para que el virus escape de la acumulación de mutaciones deletéreas o salte entre picos adaptativos. Específicamente, la recombinación podría acelerar la progresión hacia el SIDA y proporcionar un mecanismo eficaz (junto con la mutación) para evadir la terapia farmacológica, el tratamiento de vacunas o el sistema inmunitario (Rambaut *et al.*, 2004). En casos de súper-infección con varios subtipos de VIH-1, la recombinación puede dar lugar a formas nuevas que pasan a formar parte de la epidemia mundial (Neher & Leitner, 2010). La dificultad en la identificación de recombinantes intra-subtipo como también el bajo número de genomas completos secuenciados de VIH-1 en intra-pacientes, sugiere que el impacto de la recombinación en la epidemia del VIH-1 aún permanece subestimada (Galetto & Negroni, 2005).

Comprender el papel de la recombinación en la acumulación de diversidad genética intra e inter-hospedero en poblaciones de VIH-1 y otros virus de ARN es necesario para entender la dinámica de muchas enfermedades virales (Rojas-Sánchez *et al.*, 2017). El presente estudio será de gran importancia ya que pretende ayudar al entendimiento del efecto de la recombinación en la diversidad genética y evolución del VIH-1, y esto podrá ser de utilidad en el diseño de estrategias de control de la infección.

4. HIPÓTESIS

El proceso de recombinación en el VIH-1 genera módulos que acumulan diversidad genética de manera diferencial si pertenecen a un subtipo puro o a una forma recombinante.

5. OBJETIVOS

5.1. Objetivo general

Evaluar el efecto de la modularidad en la diversidad y evolución de los genomas del grupo M del VIH-1.

5.2. Objetivos particulares

1. Identificación de módulos en genomas del grupo M del VIH-1 en grupos de datos inter-pacientes infectados con CRFs o subtipos puros.
2. Comparar el nivel de diversidad presente en los módulos de las CRFs y de los subtipos puros que las constituyen.
3. Comparar el efecto de la selección natural en la evolución de las regiones modulares entre genomas recombinados y de los subtipos puros que los constituyen.

6. ESTRATEGIA EXPERIMENTAL

Como se muestra en la Figura 19, la estrategia que se siguió en este estudio fue la siguiente: se descargaron secuencias de genoma completo tanto de subtipos puros como de CRFs del grupo M del VIH-1, posteriormente se realizaron alineamientos múltiples y globales en cada uno de los conjuntos de datos mediante el algoritmo MUSCLE (Edgar, 2004) implementado en el programa AliView v1.26 (Larsson, 2014).

Para abordar el objetivo 1, cada uno de los alineamientos de los CRFs se subió al servidor DataMonkey para poder identificar los puntos de recombinación utilizando el algoritmo GARD (Kosakovsky-Pond *et al.*, 2006), posteriormente mediante el lenguaje de programación R (https://www.r-project.org/), se hizo la identificación de los módulos en cada uno de los alineamientos, tanto de las CRFs como en los subtipos puros que las constituyen.

Para el objetivo 2, tanto los alineamientos de los subtipos puros como el de las CRFs se subió al servidor ViPR (*Virus Pathogen Database and Analysis Resource*, https://www.viprbrc.org) para llevar a cabo los análisis de diversidad; posteriormente, en R se realizaron análisis estadísticos para obtener la significancia de cada uno de los módulos. También, para comparar los resultados obtenidos de diversidad, se generaron filogenias para cada uno de las CRFs como de los subtipos puros mediante la ejecución de MrBayes (Huelsenbeck & Ronquist, 2001) a través de la terminal de Ubuntu, y se realizó la reconstrucción de estados ancestrales de aquellos polimorfismos que tuvieran una puntuación de diversidad >=85, y que se encontraran tanto en la CRF como en los subtipos puros que la constituyen.

Finalmente, para cumplir con el objetivo 3, los alineamientos tanto de las secuencias de genomas de subtipos puros como los de las CRFs se subieron al servidor DataMonkey para llevar a cabo la detección de selección natural mediante el algoritmo FEL (Kosakovsky-Pond & Frost, 2005); posteriormente, se utilizó R para la identificación de los sitios bajo selección en cada uno de los módulos tanto de los CRFs como de los subtipos puros. Luego, se identificaron los sitios bajo selección presentes en cada uno de los genes que codifican a las proteínas estructurales, tanto para las CRFs como para los subtipos puros, y utilizando Chimera v1.131 (https://www.cgl.ucsf.edu/chimera/), se mapearon los sitios en la estructura terciaria de las proteínas de interés. Por último, se llevaron a cabo análisis de coevolución utilizando SpiderMonkey (Poon *et al.*, 2008) a través de Hyphy (Pond *et al.*, 2004).

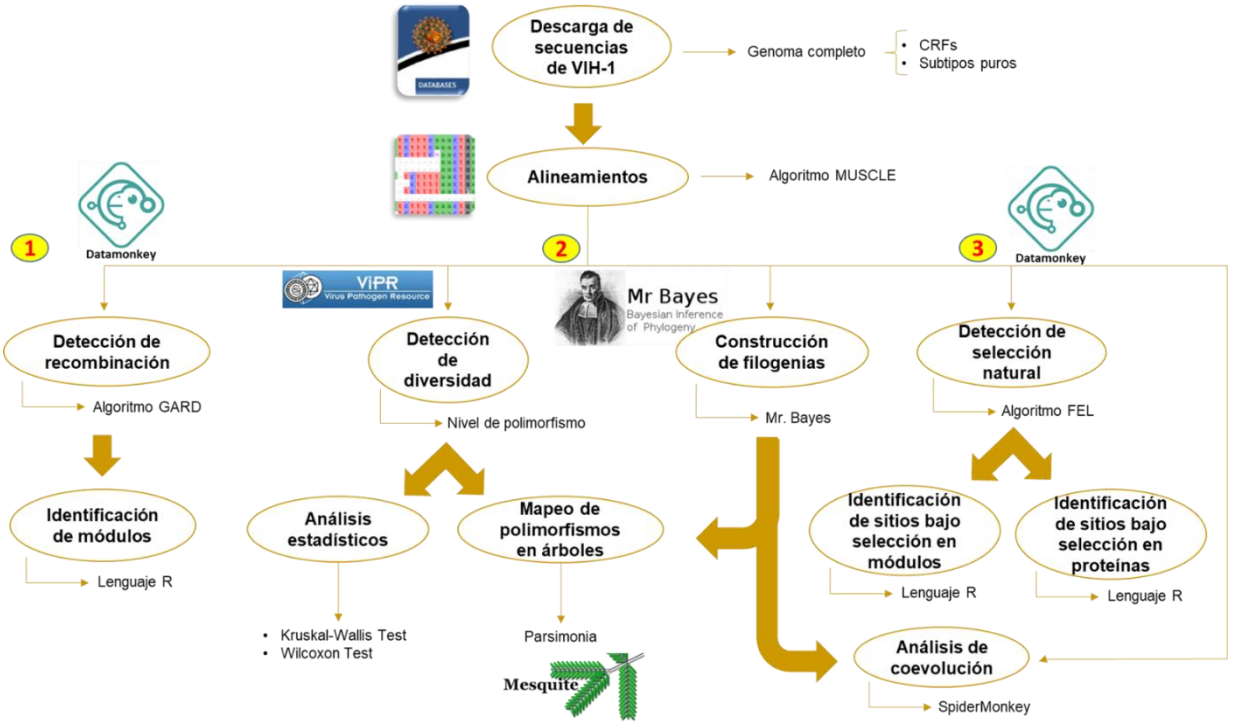

Figura 19. Estrategia experimental llevada a cabo en este estudio.

7. MATERIALES Y MÉTODOS

7.1. Datos de secuencias

Se descargaron 8 conjuntos de secuencias de genoma completo de CRFs, siendo los siguientes: **CRF-02_A1G** con 20 secuencias, **CRF-06_cpx(A1,G,J,K)** con 14 secuencias, **CRF-07_BC** con 20 secuencias, **CRF-11_cpx(A1,E,G,J,U)** con 20 secuencias, **CRF-14_BG** con 13 secuencias, **CRF-35_A1D** con 20 secuencias, **CRF-42_BF** con 17 secuencias y **CRF-63_02A6** con 11 secuencias; mientras que de los subtipos puros que las constituyen se obtuvieron 10 conjuntos de secuencias: **subtipo A1**, **A6**, **B**, **C**, **D**, **F1**, **G** y **U**, con 20 secuencias cada uno, **subtipo J** con 7 secuencias y **subtipo K** con 3 secuencias, ya que estos subtipos son muy raros.

Todas las secuencias de los genomas se descargaron de la base de datos del Laboratorio Nacional de Los Álamos (LANL, https://www.hiv.lanl.gov/content/index); en la Tabla II se muestra el número de secuencias obtenidas por país y por continente; nuestro objetivo fue que el muestreo fuera aleatorio, por lo tanto, en la Figura 20 se puede observar de una manera visual los datos presentados en la Tabla II. Las secuencias de los genomas de las CRFs fueron analizadas para identificar los puntos de recombinación de sus genomas y definir los módulos presentes en cada uno de ellos; mientras que las secuencias de los genomas de subtipos puros fueron utilizadas para realizar las comparaciones de los análisis de diversidad y de selección natural.

Tabla II. Distribución geográfica de las secuencias obtenidas para este estudio.

Continente	País	Total por país	Total por continente
África	Ghana	4	99
	Liberia	1	
	Cameron	23	
	Senegal	5	
	Guinea-Bissau	2	
	Costa de Marfil	1	
	Nigeria	9	
	Angola	3	
	Uganda	5	
	República Democrática del Congo	18	
	Burkina Faso	1	

	Rwanda	2	
	Tanzania	4	
	Kenia	6	
	África del Sur	6	
	Gabón	1	
	Etiopia	1	
	Somalia	1	
	Zambia	1	
	Malawi	1	
	Botsuana	1	
	Chad	1	
	Mali	2	
América	USA	5	18
	Canadá	3	
	Argentina	3	
	Cuba	1	
	Brasil	6	
Asia	Pakistán	4	81
	South Corea	3	
	Federación de Rusia	19	
	China	20	
	Taiwán	3	
	Afganistán	11	
	Irán	9	
	India	2	
	Georgia	2	
	Kazakstán	1	
	Uzbekistán	1	
	Tailandia	1	
	Indonesia	1	
	Israel	1	
	Yemen	2	
	Nepal	1	
Europa	Chipre	13	104
	Alemania	5	
	Suecia	8	
	Francia	6	
	Reino Unido	12	
	España	21	
	Estonia	1	
	Grecia	2	

	Portugal	2	
	Bulgaria	2	
	Ucrania	4	
	Bielorrusia	1	
	Italia	1	
	Países Bajos	5	
	Finlandia	1	
	Bélgica	2	
	Romania	1	
	Luxemburgo	17	
Oceanía	Australia	3	3

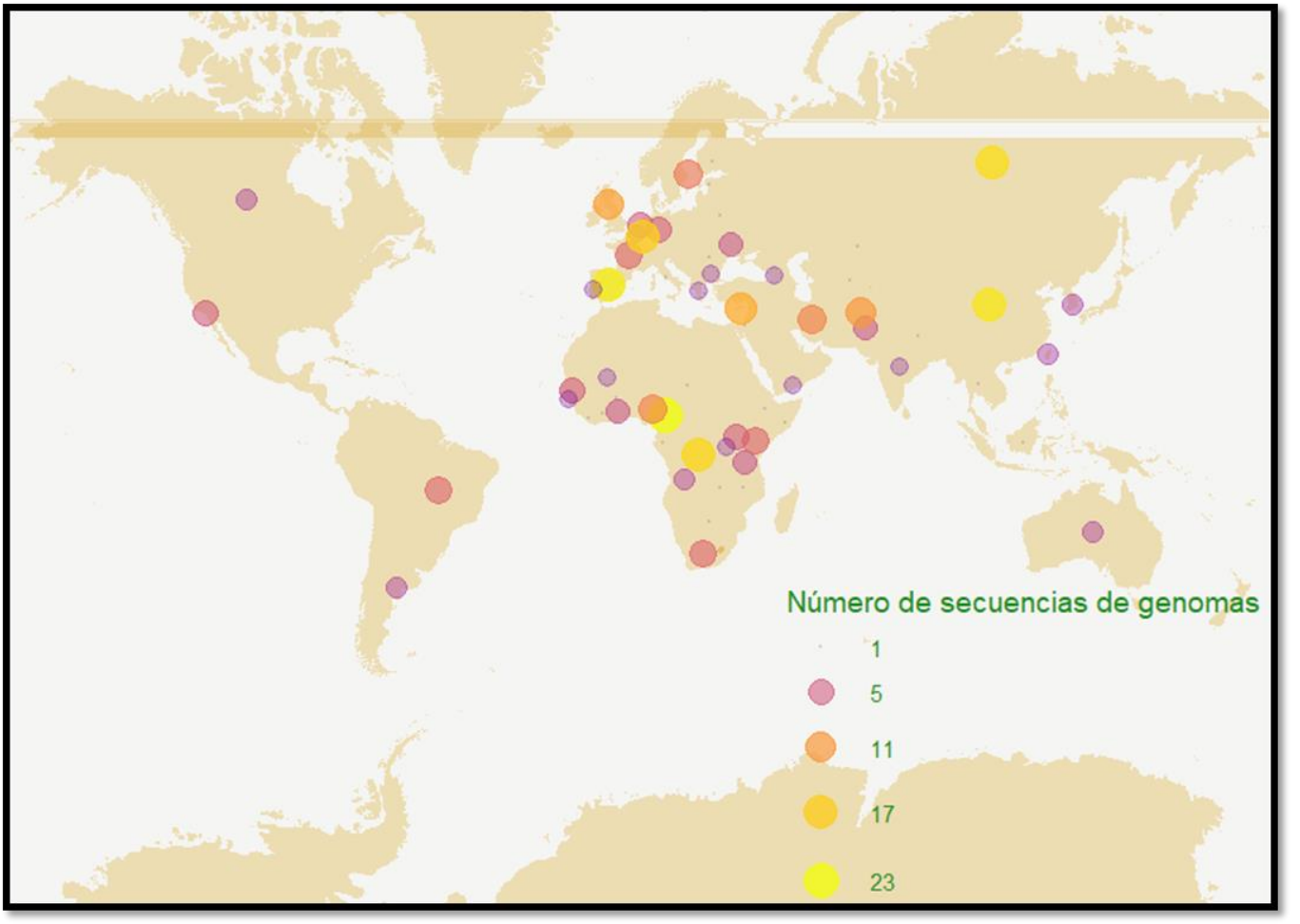

Figura 20. Distribución geográfica de las secuencias de genomas obtenidas para este estudio.

7.2. Alineamiento de genomas

Para cada uno de los conjuntos de datos tanto de las CRFs como de los subtipos puros que las constituyen, se generó un alineamiento múltiple y global con las secuencias de nucleótidos de genoma completo utilizando el algoritmo MUSCLE (*Multiple Sequence Comparison by log-Expectation*) ya que es un algoritmo iterativo que utiliza dos medidas de distancia para un par de secuencias: una distancia *kmer* (para un par no alineado) y la distancia *Kimura* (para un par alineado). Un kmer es una sub-secuencia contigua de longitud *k*, también conocida como *k-tupla*. Las secuencias relacionadas tienden a tener más kmers en común de lo esperado por azar. Dado un par de secuencias alineadas, se calcula la identidad por pares y se convierte en una estimación de distancia aditiva, aplicando la corrección Kimura para múltiples sustituciones en un solo sitio. Las matrices de distancia se agrupan utilizando el algoritmo UPGMA (Edgar, 2004). Se utilizaron los parámetros predeterminados tanto para realizar el cálculo de las distancias como para el del agrupamiento. Los alineamientos fueron construidos y visualizados con el programa AliView v1.26 (Larsson, 2014). En un alineamiento global, se intenta que el alineamiento cubra las secuencias completamente introduciendo los gaps que sean necesarios, ya que sirve para comparar secuencias que se supone son similares a lo largo de toda su longitud (Peris & Marzal, 2014).

Para realizar los análisis de selección, los alineamientos de secuencias de nucleótidos del genoma completo fueron cortados en sus tres genes principales: *gag*, *pol* y *env*, utilizando la secuencia más antigua como secuencia de referencia para realizar el corte. Esto se realizó en cada uno de los conjuntos de datos, utilizando el Script 1 (ver sub-apartado Código, en el apartado Anexos) construido en Python v2.0 y ejecutado a través de la terminal de Ubuntu 16.04.7 LTS (*Xenial Xerus*).

Posteriormente, se eliminaron todos los gaps de cada uno de los alineamientos de genes y se re-alinearon como aminoácidos traducidos. Finalmente, los codones de paro que surgieron, se eliminaron y se añadió en su lugar el aminoácido consenso en esa posición del alineamiento.

7.3. Detección de recombinación

Se hizo uso del servidor de evolución adaptativa DataMonkey para realizar los análisis de recombinación. Cada uno de los alineamientos de secuencias de nucleótidos del genoma completo de las CRFs se subió al servidor. DataMonkey utiliza el algoritmo GARD para inferir los puntos de recombinación. Dicho algoritmo está diseñado para buscar evidencia de filogenias específicas del segmento; dado un número máximo de puntos de recombinación (B), el método buscará en el espacio de todas las ubicaciones posibles B o menos puntos de recombinación en el alineamiento, infiriendo filogenias para cada fragmento putativo no recombinante, y evalúa por un criterio basado en la información derivada de un modelo de máxima verosimilitud adecuado a cada segmento (Kosakovsky-Pond *et al.*, 2006).

El algoritmo GARD funciona inicialmente buscando el número y la colocación de puntos de recombinación que produce el mejor criterio de información de Akaike (AICc); una medida de las propiedades del ajuste del modelo a los datos. Sin embargo, cualquier mejora en el ajuste del modelo podría deberse a una serie de factores (por ejemplo, variación de la velocidad espacial o **heterotaquia**) que no sea un cambio en la topología de árbol, que sería la firma principal para la recombinación. Posteriormente, GARD comprueba la congruencia de los árboles a ambos lados del punto de recombinación putativo utilizando la prueba Kishino-Hasegawa (KH) (Thompson *et al.*, 2014). Si existe una **inconsistencia filogenética**, GARD indica que en esa posición existe un punto de recombinación.

7.4. Identificación de módulos en alineamientos

A partir de los resultados obtenidos de los puntos de recombinación por GARD, se utilizó la herramienta de dibujo del LANL (https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html) para generar esquemas de los genomas de las CRFs para poder identificar los módulos de una manera visual. También, se crearon scripts bajo el entorno del lenguaje de programación R, para identificar los segmentos no recombinantes (**módulos**) en cada

alineamiento de secuencias de nucleótidos del genoma completo, tanto de las CRFs como de los subtipos puros que las constituyen. Se utilizó R porque es un lenguaje y ambiente de cómputo estadístico y gráfico; además, R provee una vasta variedad de técnicas estadísticas (modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de tiempo-series, clasificación, agrupamiento, etc.).

7.5. Detección de diversidad

Se utilizó la base de datos ViPR (*Virus Pathogen Database and Analysis Resource*, https://www.viprbrc.org) la cual contiene una herramienta para la detección de polimorfismos de un solo nucleótido (SNPs). Los alineamientos de secuencias de nucleótidos del genoma completo tanto de las CRFs como de los subtipos puros que las constituyen se utilizaron para alimentar esta base de datos y llevar a cabo los análisis de diversidad.

Para evaluar la diversidad presente en las secuencias del VIH-1, se calculó el grado de polimorfismo presente en cada una de las posiciones de los alineamientos. Primero, la herramienta crea una secuencia consenso por la "*regla de la mayoría*"; en cada posición, el consenso es el alelo con una frecuencia superior al 50%, independientemente de la cobertura. Sí ningún alelo supera el 50%, se asigna la letra N (cualquier nucleótido) indicando ambigüedad en esa posición. Las secuencias en el alineamiento se comparan con la secuencia consenso para identificar polimorfismos. Para asignar un puntaje al polimorfismo en cada posición, se utiliza una fórmula modificada a partir de la citada en Crooks y colaboradores en 2004:

$$\mathbf{S} = -100 * \Sigma(Pi * \log_2 Pi)$$, donde Pi es la frecuencia del -iésimo alelo

La puntuación es la entropía normalizada de la distribución del alelo observado. Para secuencias de nucleótidos, las puntuaciones pueden oscilar entre 0 (sin polimorfismo) y 232 (4 alelos y un indel, 20% de frecuencia cada uno).

ViPR genera un archivo con extensión ".csv" (valores separados por comas) con los resultados del puntaje del polimorfismo para cada una de las posiciones del

alineamiento. Este fue utilizado para generar una gráfica con la diversidad del genoma completo para cada uno de los conjuntos de datos, utilizando el Script 2, implementado en R.

También, se generaron diagramas de Venn para identificar el número de posiciones de los polimorfismos que se compartían entre las CRFs y los subtipos puros que las constituyen; que posteriormente se utilizaron estos sitos para la reconstrucción de estados ancestrales. Se utilizó el Script 3 para generar el diagrama de Venn.

7.5.1. Diversidad presente en módulos

Se construyeron gráficas para comparar la diversidad presente en cada uno de los módulos de las CRFs respecto a los subtipos puros que las constituyen, graficando los sitios con una puntuación >=85 en diversidad.

7.5.2. Análisis estadísticos de los módulos

Se llevaron a cabo pruebas de Kruskal-Wallis porque nuestros grupos de datos son independientes, y creemos que siguen una misma distribución ya que provienen de una misma población; inferimos que los módulos presentes en las CRFs provienen de alguno de los subtipos puros que las constituyen. Usando la prueba Kruskal-Wallis, se puede decidir sí las distribuciones de la población son idénticas sin suponer que sigan una distribución normal (Yau, 2014). Posteriormente, a los resultados estadísticamente significativos se les realizó una segunda prueba no paramétrica, utilizando la prueba de Wilcoxon, la cual es generalmente usada para comparar dos grupos de datos que no siguen una distribución normal, asumiendo la independencia de las muestras (Rosner & Glynn, 2010).

Los archivos .csv generados por ViPR, tomando el valor del puntaje del polimorfismo, se utilizaron para realizar cada una de las pruebas; las posiciones conservadas en el genoma del VIH-1 tienen un puntaje de polimorfismo igual a cero, éstas fueron eliminados para realizar dichas pruebas. El Script 4 se utilizó para ejecutar la prueba Kruskal-Wallis.

Mientras que la prueba Kruskal-Wallis puede realizar la comparación con *n* grupos, la prueba Wilcoxon sólo puede comparar dos grupos. Por lo tanto, se utilizó la herramienta RIP (*Recombinant Identification Program*, https://www.hiv.lanl.gov/content/sequence/RIP/RIP.html) del LANL, la cual permite identificar a que subtipo pertenece un fragmento de un CRF, para poder decidir con qué subtipo debíamos comparar el módulo de la CRF. La Tabla III muestra la comparación que se llevó a cabo en cada uno de los CRFs. El Script 5 se empleó para realizar la prueba de Wilcoxon.

Tabla III. Comparación para la prueba Wilcoxon.

CRF	Módulos								
	1	2	3	4	5	6	7	8	9
02_A1G	A1		G	A1		G	-	-	-
06_cpx(A1,G,J,K)	J	J	K	A1	A1	J	A1	J	J
07_BC	C					B	C	C	C
11_cpx(A1,E,G,J,U)	J	A1	G	A1	A1	A1	J	-	-
14_BG	G	B	B	G	G	G	-	-	-
35_A1D	A1	D		D	D	D	A1	-	-
42_BF	F1		B	B	B	F1	F1		
63_02A6	CRF-02	A6	CRF-02	A6	A6	CRF-02	CRF-02	-	-

Los módulos vacíos son los que no fueron estadísticamente significativos en la prueba Kruskal-Wallis. Los módulos que no están presentes en el CRF se identifican con el signo "-".

7.5.3. Generación de filogenias

Para la construcción de una filogenia bayesiana, se necesitan archivos nexus para ejecutar MrBayes, por lo tanto, la conversión de archivo fasta a nexus se realizó mediante AliView v1.26. También, se necesita de un modelo evolutivo con el que se pueda guiar la inferencia bayesiana, por lo que se usó el programa jModelTest v2.1.10 (Posada, 2008), el Script 6 se utilizó para encontrar los modelos evolutivos.

Para comprobar que el modelo fuese el indicado, también se utilizó el programa MEGA v7.0.26 (Kumar *et al.*, 2016) para encontrar el modelo evolutivo. Cuando se encontró una discrepancia entre los resultados de ambos programas, se eligió el modelo más robusto. La Tabla IV muestra el modelo evolutivo utilizado para generar la filogenia, tanto para cada una de las CRFs como para los subtipos puros.

Tabla IV. Modelos evolutivos usados para construir la filogenia bayesiana.

Genotipo	jModelTest	MEGA	Utilizado
CRF-02_A1G	GTR+I+G	GTR+I+G	GTR+I+G
CRF-06_cpx(A1,G,J,K)	GTR+I+G	GTR+I+G	GTR+I+G
CRF-07_BC	GTR+I+G	GTR+I+G	GTR+I+G
CRF-11_cpx(A1,E,G,J,U)	GTR+I+G	GTR+I+G	GTR+I+G
CRF-14_BG	GTR+I+G	GTR+I+G	GTR+I+G
CRF-35_A1D	GTR+I+G	GTR+I+G	GTR+I+G
CRF-42_BF	GTR+I+G	HKY+G+I	GTR+I+G
CRF-63_02A6	GTR+I+G	GTR+I+G	GTR+I+G
Subtipo A1	GTR+I+G	GTR+I+G	GTR+I+G
Subtipo A6	GTR+I+G	GTR+I+G	GTR+I+G
Subtipo B	GTR+I+G	GTR+I+G	GTR+I+G
Subtipo C	GTR+I+G	GTR+I+G	GTR+I+G
Subtipo D	GTR+I+G	GTR+I+G	GTR+I+G
Subtipo F1	GTR+I+G	GTR+I+G	GTR+I+G
Subtipo G	GTR+I+G	GTR+I+G	GTR+I+G
Subtipo J	GTR+I+G	GTR+G	GTR+I+G
Subtipo K	GTR+I	HKY+G	GTR+I
Subtipo U	GTR+I+G	GTR+G	GTR+I+G

GTR: *General Time Reversible*; I: *Inverse*; G: *Gamma distribution*.

En un análisis bayesiano, la inferencia de la filogenia está basada sobre las probabilidades posteriores de los árboles filogenéticos. MrBayes usa una Cadena de Markov Monte Carlo (MCMC) para aproximar las probabilidades posteriores de los árboles. MCMC es un método para tomar muestras válidas, aunque dependientes, de

la distribución de probabilidad de interés (en este caso, las probabilidades posteriores de los árboles filogenéticos) (Huelsenbeck & Ronquist, 2001).

Para llevar a cabo la construcción de las filogenias, se utilizó MrBayes v3.2.6 ejecutado desde la terminal de Ubuntu 16.04.7 LTS (*Xenial Xerus*). El Script 7 se utilizó para generar las filogenias.

Los comandos "*nst*" y "*rates*" se determinaron dependiendo del modelo evolutivo de cada uno de los alineamientos. Para poder visualizar los árboles generados, se utilizó el programa FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

7.5.4. Reconstrucción de estados ancestrales

La reconstrucción de estados ancestrales de los polimorfismos se realizó con el programa Mesquite v3.61 (https://www.mesquiteproject.org/) mediante el método de parsimonia (Mount, 2008). El mapeo de caracteres en las filogenias provee una estructura histórica para comprender la evolución de las características morfológicas y moleculares (Bollback, 2006).

Para llevar a cabo la reconstrucción de estados ancestrales, Mesquite requiere de una matriz y un árbol filogenético. La matriz se construyó a partir de los sitios con un puntaje >=85 en diversidad que se encontraron tanto en las CRFs como en los subtipos puros que las constituyen. Mientras que el árbol que se utilizó fue el del módulo en que se encontraba el polimorfismo. El Script 8 se ejecutó para crear la matriz.

7.6. Detección de selección natural

Las tasas de sustitución están ligadas a la identificación de selección natural y a la adaptación molecular basada en sitios. Se considera que los cambios nucleotídicos que no generan un cambio en la composición de aminoácidos (sustitución sinónima; dS) tienden a ser neutrales, mientras que los cambios que sí afectan la composición se encuentran bajo selección (sustitución no sinónima; dN) (Castro-Nallar *et al.*, 2012).

7.6.1. Detección de selección natural mediante FEL

Los alineamientos de secuencias de codones de los tres genes estructurales (*gag*, *pol* y *env*) de cada uno de las CRFs y de los subtipos puros que las constituyen se subieron al servidor DataMonkey.

Se utilizó el algoritmo FEL (*Fixed Effects Likelihood*), ya que realiza la detección de selección penetrante en cada codón utilizando un enfoque de máxima verosimilitud para inferir la tasa de sustituciones no sinónimas (dN) y sinónimas (dS) por sitio de un alineamiento dado; este método asume que la presión de selección para cada sitio es constante a lo largo de la filogenia completa (Kosakovsky-Pond & Frost, 2005).

FEL genera un archivo con extensión ".csv" el cual contiene valores para el dS, dN, Ω y el LRT (*Likelihood Ratio Test*) los cuales se utilizaron para identificar los sitios bajo selección, tanto positiva como negativa. El Script 9 se utilizó para realizar esta actividad.

7.6.2. Identificación de sitios bajo selección en módulos

Para identificar los sitios bajo selección en los módulos de cada uno de los genes, se crearon reglas para asignar el módulo correspondiente en cada gen. La Tabla V muestra las reglas que se siguieron para cada CRF y sus subtipos puros que las constituyen. Para generar un archivo con extensión ".txt" con las posiciones bajo selección en cada módulo, se utilizó el Script 10.

Tabla V. Reglas para la asignación de módulos en los genes.

CRF	Genes		
	gag	***pol***	***env***
02_A1G	**M1:** <=143; **M2:** >= 144 & <=361; **M3:** >=362	**M3:** <=207; **M4:** >=208 & <=501; **M5:** >=502 & <=718; **M6:** >=719	**M6:** >=1
06_cpx(A1,G,J,K)	**M1:** <=165; **M2:** >=166	**M2:** >=1	**M3:** <=128; **M4:** >=129 & <=169; **M5:** >=170 & <=289; **M6:** >=290 & <=386; **M7:** >=387 & <=491; **M8:** >=492 & <=645; **M9:** >=646

07_BC	**M1:** <=509	**M1:** <=738; **M2:** >=739	**M5:** <=98; **M6:** >=99 & <=130; **M7:** >=131 & <=518; **M8:** >=519 & <=666; **M9:** >=667
11_cpx(A1,E,G,J,U)	**M1:** <=172; **M2:** >=173 & <=304; **M3:** >=305 & <=427; **M4:** >=428	**M4:** <=182; **M5:** >=183 & <=332; **M6:** >=333 & <=734; **M7:** >=735	**M7:** >=1
14_BG	**M1:** <=497	**M1:** <=835; **M2:** >=836	**M4:** <=248; **M5:** >=249
35_A1D	**M1:** <=490; **M2:** >=491	**M2:** <=1220; **M3:** >=1221	**M5:** <=480; **M6:** >=481 & <=863; **M7:** >=864
42_BF	**M1:** <=338; **M2:** >=339 & <=568; **M3:** >=569	**M3:** >=1	**M6:** <=458; **M7:** >=459 & <=573; **M8:** >=574 & <=897; **M9:** >=898
63_02A6	**M3:** >=1	**M3:** >=1	**M5:** <=710; **M6:** >=711 & <=962; **M7:** >=963

M: Módulo.

7.6.3. Asignación de sitios bajo selección en genes

Para realizar la identificación de sitios bajo selección en cada uno de los genes principales del VIH-1, se generaron reglas para asignar la coordenada de cada gen. La Tabla VI muestra las reglas que se utilizaron tanto para las CRFs como para los subtipos puros que las constituyen. El Script 11 se utilizó para generar un archivo con extensión ".txt" con las posiciones bajo selección en cada gen.

Tabla VI. Reglas para la asignación de coordenadas de los genes.

CRF	Genes		
	gag	*pol*	*env*
02_A1G	**MA:** <=141; **CA:** >=142 & <=394; **NC:** >=409 & <=466	**PR:** >=74 & <=189; **RT:** >=190 & <=810; **IN:** >=851 & <=1020	**gp120:** >=54 & <=555; **gp41:** >=632 & <=741
06_cpx(A1,G,J,K)	**MA:** <=141; **CA:** >=142 & <=372; **NC:** >=389 & <=443	**PR:** >=74 & <=172; **RT:** >=173 & <=732; **IN:** >=775 & <=942	**gp120:** >=48 & <=518; **gp41:** >=574 & <=683
07_BC	**MA:** <=132; **CA:** >=133 & <=363; **NC:** >=376 & <=430	**PR:** >=68 & <=166; **RT:** >=167 & <=726; **IN:** >=769 & <=936	**gp120:** >=50 & <=533; **gp41:** >=597 & <=708
11_cpx(A1,E,G,J,U)	**MA:** <=134; **CA:** >=135 & <=365; **NC:** >=381 & <=435	**PR:** >=88 & <=186; **RT:** >=187 & <=746; **IN:** >=789 & <=956	**gp120:** >=48 & <=532; **gp41:** >=588 & <=700

14_BG	**MA:** <=131; **CA:** >=132 & <=362; **NC:** >=376 & <=430	**PR:** >=58 & <=156; **RT:** >=157 & <=716; **IN:** >=759 & <=926	**gp120:** >=48 & <=500; **gp41:** >=555 & <=664
35_A1D	**MA:** <=137; **CA:** >=138 & <=368; **NC:** >=382 & <=435	**PR:** >=57 & <=155; **RT:** >=156 & <=715; **IN:** >=758 & <=925	**gp120:** >=48 & <=525; **gp41:** >=580 & <=689
42_BF	**MA:** <=132; **CA:** >=133 & <=363; **NC:** >=377 & <=431	**PR:** >=61 & <=159; **RT:** >=160 & <=719; **IN:** >=762 & <=929	**gp120:** >=48 & <=609; **gp41:** >=667 & <=789
63_02A6	**MA:** <=141; **CA:** >=142 & <=372; **NC:** >=386 & <=439	**PR:** >=69 & <=167; **RT:** >=168 & <=727; **IN:** >=770 & <=937	**gp120:** >=48 & <=518; **gp41:** >=575 & <=684

MA: proteína de matriz, CA: proteína de cápside, NC: proteína de nucleocápside, PR: proteasa, RT: transcriptasa reversa, IN: integrasa.

7.6.4. Análisis de coevolución

Para identificar que sitios tanto intra- como inter-módulo están coevolucionando, se realizaron análisis de coevolución mediante SpiderMonkey (Poon *et al.*, 2008) a través de Hyphy (Pond *et al.*, 2004). SpiderMonkey reconstruye la historia de sustituciones de un alineamiento por métodos filogenéticos basados en máxima verosimilitud, y luego analiza la distribución de los eventos de sustitución usando un modelo gráfico bayesiano para identificar los sitios que acumulan mutaciones en el mismo conjunto de ramas (Poon *et al.*, 2008).

Primero, los alineamientos de los genes *gag*, *pol* y *env* se concatenaron en uno solo, utilizando el Script 12. Posteriormente, utilizando este alineamiento y la filogenia de cada uno de las CRFs como la de los subtipos puros que las constituyen, se corrió SpiderMonkey utilizando el Script 13.

8. RESULTADOS

8.1. Identificación de puntos de recombinación en las CRFs

El algoritmo GARD identificó cinco puntos de recombinación tanto en la **CRF-02_A1G** como en la **CRF-14_BG**, seis puntos de recombinación en las **CRF-11_cpx(A1,E,G,J,U)**, **CRF-35_A1D** y **CRF-63_02A6**, y ocho puntos de recombinación en las **CRF-06_cpx(A1,G,J,K)**, **CRF-07_BC** y **CRF-42_BF**. La Figura 21 nos muestra una de las gráficas generadas por el algoritmo GARD al identificar los puntos de recombinación, donde en el eje horizontal se muestra la longitud del genoma, mientras que en el eje vertical se observa el soporte para cada uno de los puntos de recombinación.

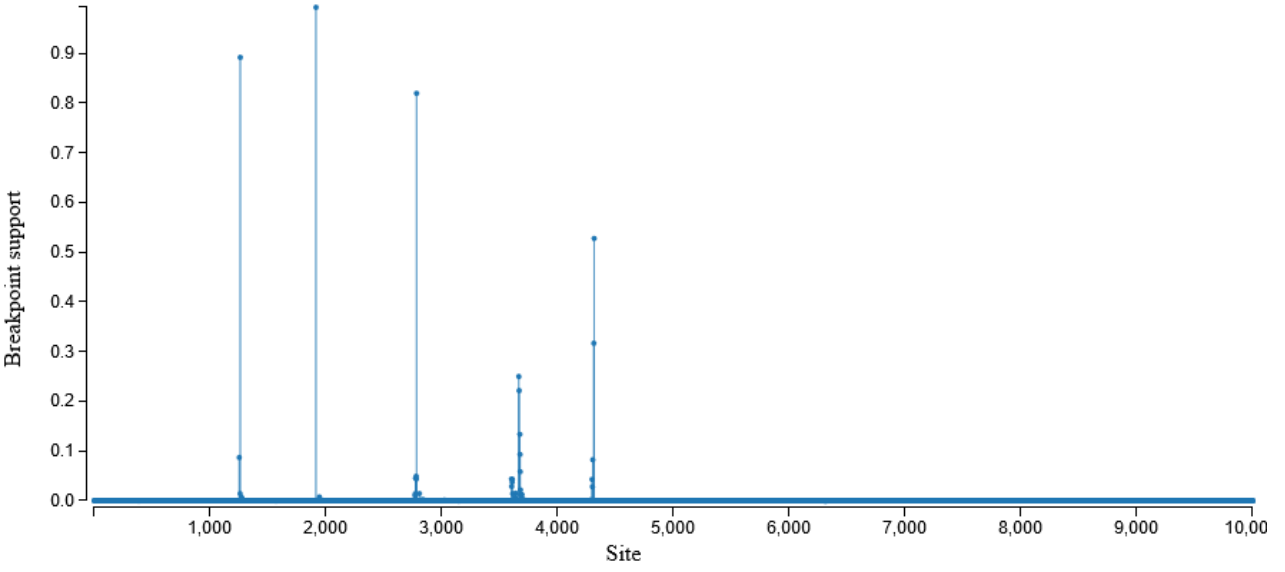

Figura 21. Puntos de recombinación en la **CRF-02_A1G**. El algoritmo GARD identificó cinco puntos de recombinación para esta forma recombinante. En esta forma recombinante, el gen *pol* podría ser una región "*hot spot*", ya que se identificaron tres puntos de recombinación ahí.

8.2. Identificación de módulos en las CRFs

Con la ayuda de los puntos de recombinación encontrados por el algoritmo GARD, se identificaron seis módulos tanto en la **CRF-02_A1G** como en la **CRF-14_BG**, siete módulos en las **CRF-11_cpx(A1,E,G,J,U)**, **CRF-35_A1D** y **CRF-63_02A6**, y nueve módulos en las **CRF-06_cpx(A1,G,J,K)**, **CRF-07_BC** y **CRF-42_BF**. Utilizando la herramienta de dibujo de la base de datos del LANL, se generaron esquemas de los genomas de las CRFs con los módulos que se identificaron, la Figura 22 muestra los

resultados obtenidos para la **CRF-02_A1G**, donde se observan los seis módulos que conforman a esta forma recombinante. La longitud y las posiciones de inicio y término de los módulos de cada uno de las CRFs se registraron en la Tabla VII.

Se observaron algunos patrones en algunos módulos de las CRFs que comparten algún subtipo puro. El primero de ellos, fue el último módulo presente en la **CRF-02_A1G** y la **CRF-11_cpx(A1,E,G,J,U)**; otro patrón identificado fue el primer módulo de la **CRF-06_cpx(A1,G,J,K)**, **CRF-11_cpx(A1,E,G,J,U)** y la **CRF-35_A1D**; y finalmente, el primer módulo de la **CRF-07_BC** y la **CRF-14_BG**, ya que en todos los casos la longitud de los módulos son parecidos. Por lo tanto, nuestros resultados nos indican que el proceso de recombinación puede llevarse a cabo en diferentes partes del genoma y la forma recombinante observada será aquella que soporte las presiones de selección a las que se enfrente.

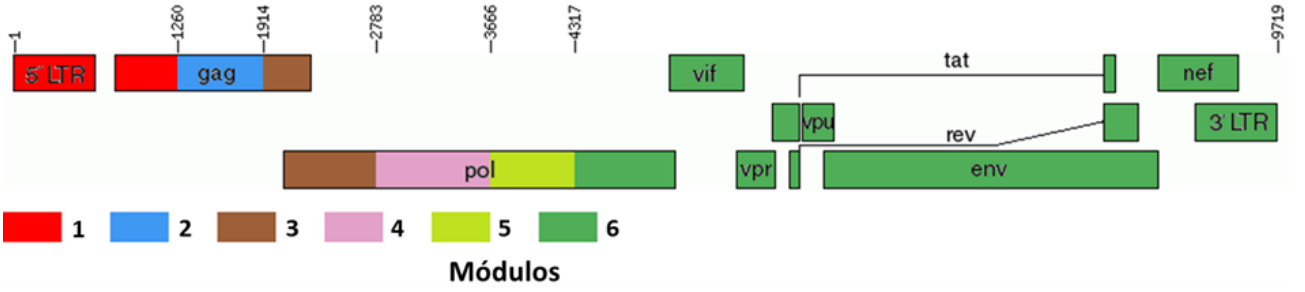

Figura 22. Módulos de la **CRF-02_A1G**. Con la ayuda de los puntos de recombinación, se identificaron seis módulos en esta forma recombinante. Se observa que el gen *pol* está compuesto por la mayoría de los módulos, mientras que el gen *env* sólo está formado por un módulo.

Tabla VII. Posición y longitud de los módulos de los CRFs.

CRF		Módulos								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
02_A1G	Posición	1-1259	1260-1913	1914-2782	2783-3665	3666-4316	4317-10007	-	-	-
	Longitud (bases)	1259	654	869	883	651	5691	-	-	-

06_cpx(A1,G,J,K)	Posición	1-1328	1329-5888	5889-6741	6742-6864	6865-7224	7225-7517	7518-7830	7831-8293	8294-9988
	Longitud (bases)	1328	4560	853	123	293	315	313	463	1695
07_BC	Posición	1-4300	4301-5174	5175-5532	5533-6118	6119-6361	6362-6653	6654-7817	7818-8262	8263-9999
	Longitud (bases)	4300	874	358	586	243	292	1164	445	1737
11_cpx(A1,E,G,J,U)	Posición	1-1344	1345-1740	1741-2109	2110-2680	2681-3119	3120-4334	4335-10205	-	-
	Longitud (bases)	1344	396	369	571	439	1215	5871	-	-
14_BG	Posición	1-4066	4067-4794	4795-5376	5377-6460	6461-8547	8548-9174	-	-	-
	Longitud (bases)	4066	728	582	1084	2087	627	-	-	-
35_A1D	Posición	1-1475	1476-4970	4971-5866	5867-5982	5983-6922	6923-8070	8071-8958	-	-
	Longitud (bases)	1475	3495	896	116	940	1148	888	-	-
42_BF	Posición	1-1483	1484-2040	2041-5184	5185-5470	5471-6142	6143-7152	7153-7497	7498-8469	8470-9256
	Longitud (bases)	1483	557	3144	286	672	1010	345	972	787
63_02A6	Posición	1-364	365-589	590-5321	5322-5680	5681-7996	7997-8752	8753-9439	-	-
	Longitud (bases)	364	225	4732	359	2316	756	687	-	-

Los módulos que no están presentes en la CRF se identifican con el signo "-".

8.3. Detección de diversidad presente en los genomas de las CRFs y de los subtipos puros que las constituyen

8.3.1. Comparación de la diversidad de los genomas completos entre las CRFs y los subtipos puros que las constituyen

Utilizando los puntajes de polimorfismo para cada una de las posiciones del alineamiento de genoma completo que la herramienta SNP de la base de datos de ViPR calculó, y mediante el lenguaje de programación R, generamos gráficas como las que se muestran en la Figura 23, en donde, en el eje horizontal se muestran las posiciones del genoma y en el eje vertical la puntuación de diversidad para cada una de las posiciones del genoma.

El objetivo de este análisis fue identificar la diversidad presente tanto en las CRFs como en sus subtipos puros que las constituyen, para posteriormente hacer una comparación. Como se observa en la Figura 23, en las tres gráficas a partir del puntaje de 85 se comienza a visualizar una mayor distribución de la diversidad, también, tanto en las CRFs como en sus subtipos puros que las constituyen, se aprecian dos líneas bien marcadas a la altura de las posiciones 7000-8000, las cuales corresponden con el patrón hipervariable de las regiones variables V1/V2 y V3 de gp120, que se encuentran en el gen *env*. Estos patrones se observaron en todas las CRFs analizadas y en todos sus subtipos puros que las constituyen.

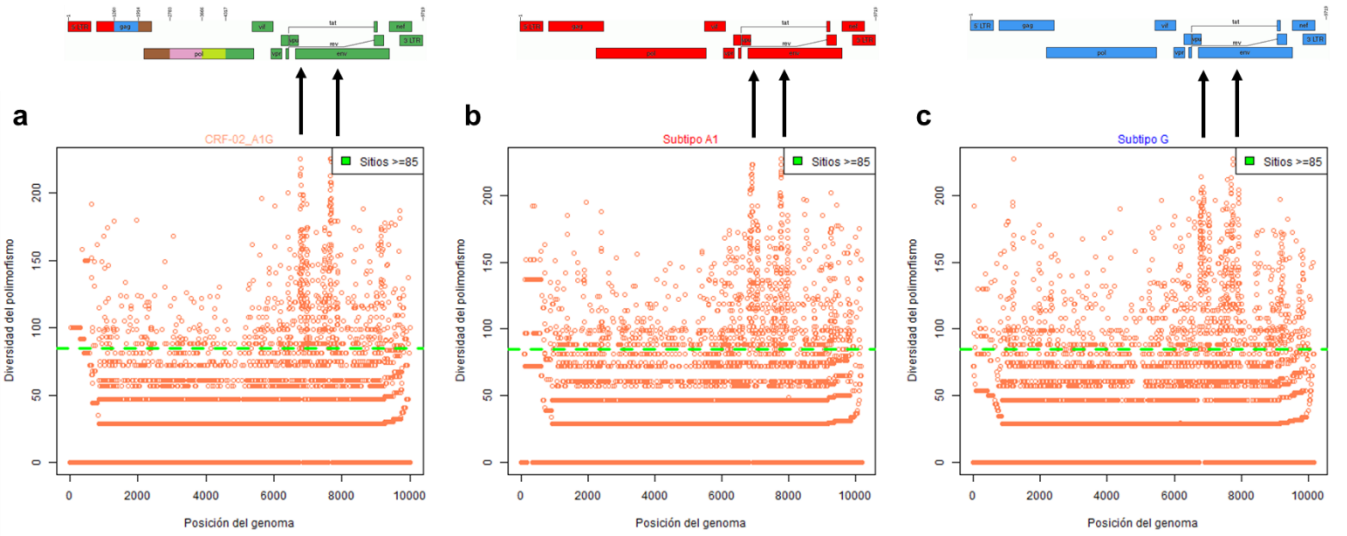

Figura 23. Diversidad del genoma completo. **a) CRF-02_A1G**. **b) Subtipo A1**. **c) Subtipo G**. En las gráficas se señala con una línea punteada color verde un umbral en el puntaje de 85, que es donde se observa que comienza una mayor distribución de la diversidad; también, se identifican con flechas dos líneas pronunciadas en la gráfica que corresponden a las regiones V1/V2 y V3 de gp120, que se encuentran en el gen *env*.

Por lo tanto, al analizar las gráficas se puede observar que las CRFs contienen un menor número de polimorfismos con un puntaje >=85, respecto a sus subtipos puros. En el apartado de Anexos se muestran las gráficas de las demás CRFs y de los subtipos puros que las constituyen.

Posteriormente, generamos diagramas de Venn para identificar el número de sitios polimórficos que se encontraban solamente en la CRF y aquellos que compartía con sus subtipos puros. Como se muestra en la Figura 24, se identificaron 1054 sitios con un puntaje >=85 en el **subtipo A1**, mientras que para el **subtipo G** se identificaron 1020 sitios, sin embargo, 865 sitios fueron identificados en la **CRF-02_A1G**; de todos los sitios polimórficos, sólo 29 están presentes tanto en la forma recombinante como en sus subtipos puros. Este patrón de que en los subtipos puros hay un mayor número de sitios polimórficos respecto a su CRF, se presentó en todos los conjuntos analizados; al igual que el número de sitios que están presentes en la CRF y en sus subtipos puros, son muy pocos.

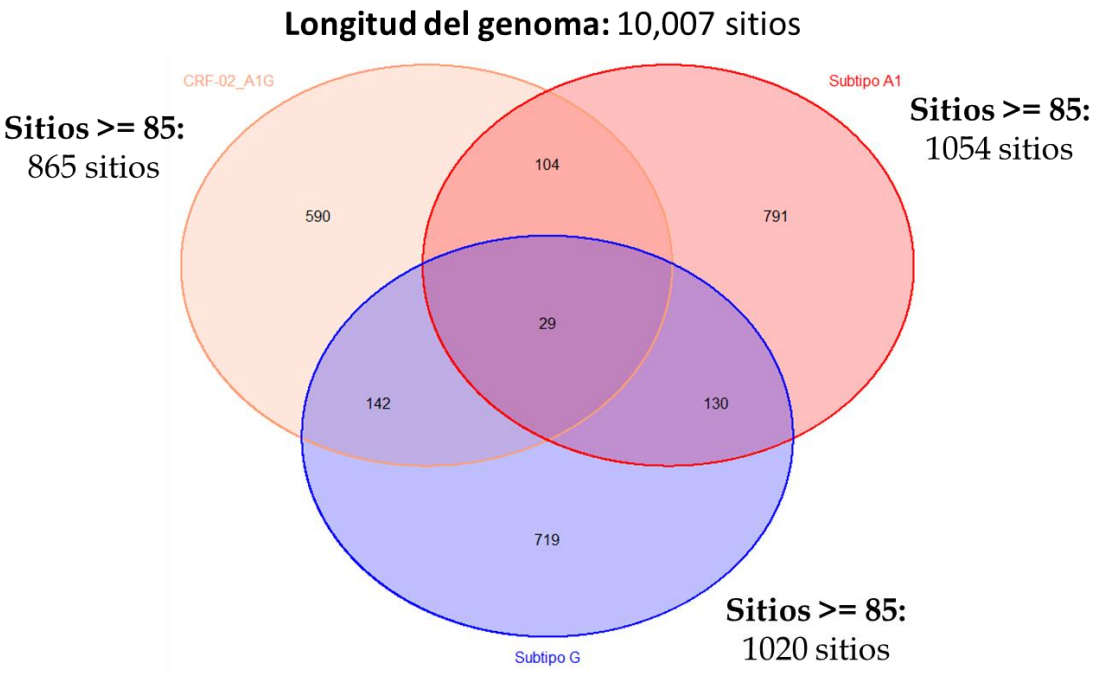

Figura 24. Número de sitios polimórficos presentes en los genomas de la **CRF-02_A1G** y de sus subtipos puros. Se observa que la forma recombinante contiene un menor número de sitios polimórficos respecto a sus subtipos puros. El número de sitios polimórficos que comparte la CRF con sus subtipos puros es menor al 4%.

8.3.2. Comparación de la diversidad de los módulos entre las CRFs y los subtipos puros que las constituyen

Para evaluar el nivel de diversidad presente en los módulos, se realizaron gráficas como la que se muestra en la Figura 25, donde se compara cada uno de los módulos de la CRF respecto a sus subtipos puros. En la figura, para el módulo 1 el **subtipo G** es el que tiene menor diversidad, pero para los demás módulos es **CRF-02_A1G** la que presenta menor diversidad. Este patrón se presenta de igual manera en las demás CRFs.

También, se identificaron con un asterisco en color amarillo aquellos módulos que resultaron estadísticamente significativos para la prueba Kruskal-Wallis, mientras que en color rosa, aquellos que resultaron para la prueba Wilcoxon. De los 60 módulos analizados, 50 resultaron con un *p-value* <0.05 para la prueba Kruskal-Wallis, mientras que 42 lo fueron para la prueba Wilcoxon.

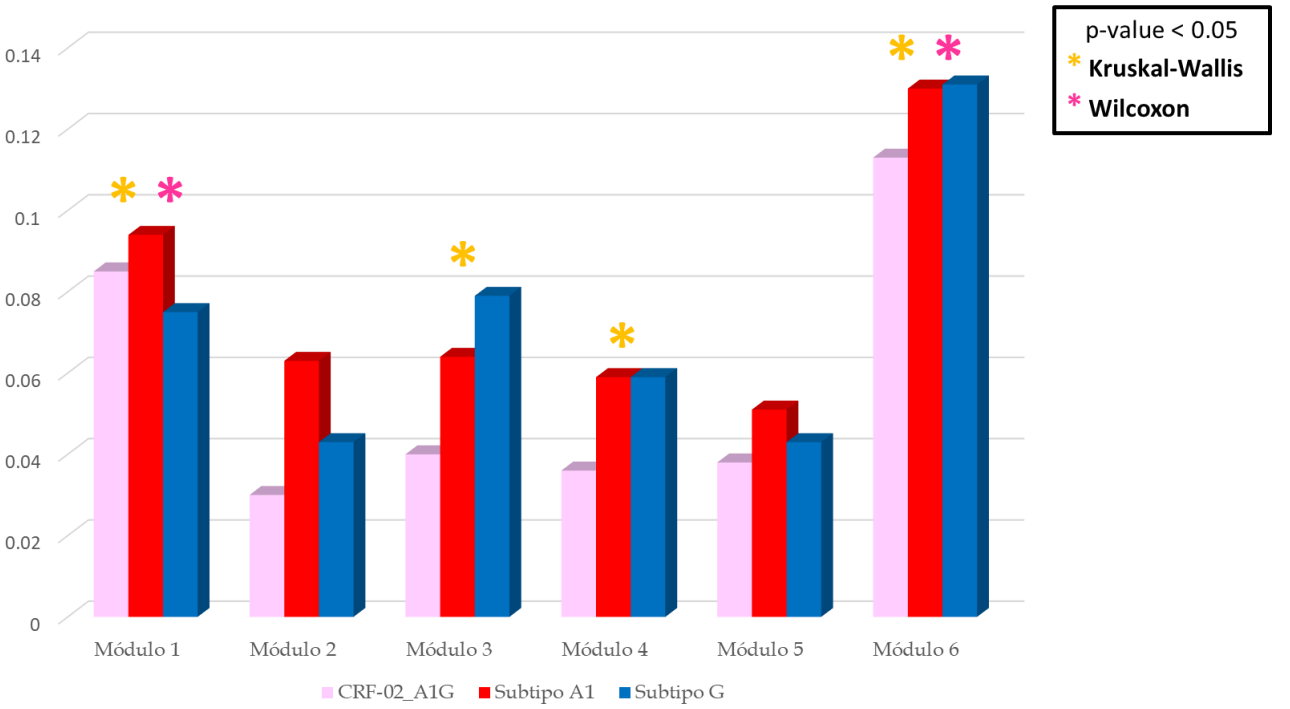

Figura 25. Comparación de la diversidad genética presente en los módulos de la **CRF-02_A1G** y sus subtipos puros. Sólo en el módulo 1 el **subtipo G** contiene menor diversidad, sin embargo, en los demás módulos, la **CRF-02_A1G** es la que contiene una menor diversidad respecto a sus subtipos puros. De los seis módulos, se encontró una diferencia estadísticamente significativa en cuatro de ellos para la prueba Kruskal-Wallis, mientras que de esos, sólo dos lo fueron para la prueba Wilcoxon.

Por lo tanto, de los 60 módulos analizados en total, 50 resultaron con un *p-value* <0.05 para la prueba Kruskal-Wallis, mientras que 42 lo fueron para la prueba de Wilcoxon. Esto nos permite afirmar que en 50 módulos se identificó una menor diversidad en las CRFs, mientras que en los 10 módulos restantes, algunos subtipos puros contienen menor diversidad respecto a su forma recombinante.

Nuestros resultados nos permitieron identificar una mayor diversidad genética en los genomas de subtipos puros respecto a su forma recombinante, e inferir que esto podría ser debido a su característica de robustez mutacional, ya que la diversidad genética de un subtipo puro, podría permitir un mayor número de mutaciones en su genotipo sin afectar su adecuación biológica.

8.3.3. Reconstrucción de estados ancestrales de los polimorfismos

Para determinar si la diversidad observada era debido a eventos de homoplasia (cambio evolutivo paralelo) y no de estados de carácter sinapomórficos (heredados de un ancestro común), se llevó a cabo la reconstrucción de estados ancestrales de aquellos polimorfismos que se encontraban tanto en las CRFs como en los subtipos puros que las constituyen.

Como se muestra en la Figura 26, en el panel *a* se observa la reconstrucción de estados ancestrales para el polimorfismo 6806 de la **CRF-02_A1G**, donde cada uno de los estados está representado por colores, azul (timina), amarillo (guanina), verde (citosina) y rojo (adenina), para este polimorfismo su estado ancestral es una guanina; mientras que en el panel *b*, el estado ancestral del polimorfismo 7606 de la **CRF-07_BC** es una adenina. El primer caso es evidencia de homoplasia, sin embargo, en el segundo caso son estados de carácter sinapomórficos.

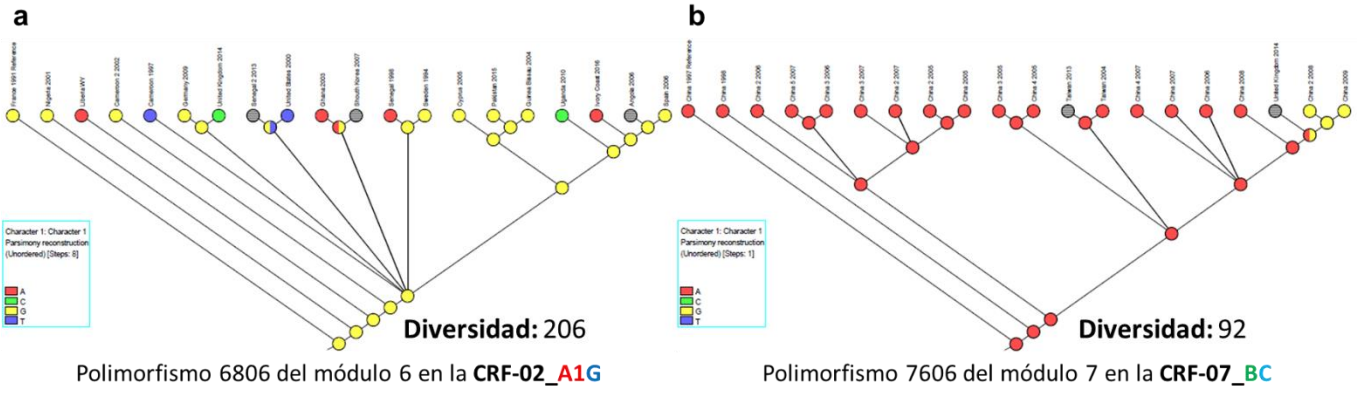

Figura 26. Reconstrucción de estados ancestrales. **a)** Polimorfismo 6806, presente en la **CRF-02_A1G** y en los subtipos que la constituyen, se observa que los estados del carácter son evidencia de homoplasia; el estado ancestral es una guanina. **b)** Polimorfismo 7606, presente en la **CRF-07_BC** y en los subtipos que la constituyen, estos estados de carácter son sinapomórficos; el estado ancestral es una adenina.

En total, se realizó la reconstrucción de 83 estados de carácter, obteniendo que en el 37% de ellos resultó una adenina (A) como estado de carácter ancestral, infiriendo que ésta podría ser un estado de carácter plesiomórfico en los polimorfismos de las CRFs del grupo M del VIH-1.

8.4. Efecto de la selección natural en la evolución modular de los genomas de las CRFs y de los subtipos puros que las constituyen

8.4.1. Sitios bajo selección en los módulos de las CRFs y de los subtipos puros que las constituyen

Con los resultados obtenidos del algoritmo FEL, se realizó la identificación de los sitios bajo selección tanto positiva como negativa en los módulos de cada una de las CRFs y de los subtipos puros que las constituyen. En la Tabla VIII se muestran el número de sitios para cada uno de los módulos presentes en cada CRF y sus subtipos puros que las constituyen. Se sombreó aquel genotipo que contiene un mayor número de sitios para cada uno de los módulos, tanto para selección positiva como para la negativa, en color azul y rojo, respectivamente.

Como se observa en la Tabla VIII, se identificó un patrón para la selección positiva, donde de los 60 módulos analizados, 45 de ellos contienen un mayor número de sitios bajo selección positiva en los subtipos puros, mientras que las CRFs sólo contienen cuatro módulos. Por lo tanto, identificamos que la selección natural positiva predomina en los módulos de los subtipos puros. Y esto es congruente con nuestros resultados anteriores, ya que los subtipos puros contienen un mayor número de sitios polimórficos, por ende, estas mutaciones podrían estar ocasionando un aumento en el número de sustituciones no sinónimas que la selección positiva está fijando.

Por otro lado, 44 módulos contienen un mayor número de sitios bajo selección negativa en los subtipos puros, mientras que sólo cinco módulos presentan las CRFs. Por lo tanto, no se identificó el patrón que se esperaba para las CRFs, ya que al presentar una menor diversidad genética, estos tenderían a presentar un mayor número de sitios bajo selección negativa. Lo anterior nos conduce a proponer que las mutaciones que ocurren en una CRF son neutrales.

Tabla VIII. Sitios bajo selección en los módulos de las CRFs y de los subtipos puros que las constituyen.

Genotipo	Selección	Módulos								
		1	2	3	4	5	6	7	8	9
CRF-02_A1G	+	8	4	18	2	7	58	-	-	-
	-	29	59	79	52	61	279	-	-	-
Subtipo A1	+	6	4	21	38	20	83	-	-	-
	-	29	84	60	4	18	224	-	-	-
Subtipo G	+	5	0	13	9	1	51	-	-	-
	-	28	74	76	73	76	269	-	-	-
CRF-06_cpx(A1,G,J,K)	+	3	14	4	1	3	11	5	3	11
	-	22	269	21	6	20	13	16	34	31
Subtipo A1	+	8	105	11	4	5	7	6	8	18
	-	39	164	35	6	31	19	27	30	68
Subtipo G	+	5	25	7	1	9	9	4	7	12
	-	38	372	33	6	17	22	25	25	58
Subtipo J	+	1	3	5	0	0	5	2	0	1
	-	29	193	19	4	18	14	21	14	24
Subtipo K	+	0	1	0	0	0	0	1	0	0
	-	20	141	4	3	12	11	9	12	21
07_BC	+	14	1	*	*	3	0	27	1	14
	-	234	46	*	*	10	5	45	9	22
Subtipo B	+	27	1	*	*	11	0	35	5	15
	-	313	75	*	*	11	9	59	32	46
Subtipo C	+	26	4	*	*	5	0	12	6	13
	-	396	102	*	*	15	12	77	21	44
CRF-11_cpx(A1,E,G,J,U)	+	7	2	4	12	2	5	52	-	-
	-	49	41	41	56	71	152	292	-	-
Subtipo A1	+	8	0	5	17	19	43	80	-	-
	-	41	51	43	37	3	23	222	-	-
Subtipo G	+	5	0	1	11	5	6	51	-	-
	-	39	44	43	44	30	133	263	-	-
Subtipo J	+	1	0	1	1	0	1	13	-	-
	-	30	29	22	31	27	52	145	-	-
Subtipo U	+	2	4	4	11	0	2	46	-	-
	-	51	60	51	65	78	180	367	-	-
14_BG	+	2	0	*	3	13	*	-	-	-

	-	93	5	*	13	45	*	-	-	-
Subtipo B	+	25	0	*	23	43	*	-	-	-
	-	347	40	*	34	123	*	-	-	-
Subtipo G	+	29	1	*	16	33	*	-	-	-
	-	359	51	*	49	137	*	-	-	-
CRF-35_A1D	+	12	10	0	*	35	22	1	-	-
	-	82	198	0	*	69	58	6	-	-
Subtipo A1	+	16	97	0	*	31	23	5	-	-
	-	156	47	0	*	113	86	17	-	-
Subtipo D	+	12	16	0	*	31	15	4	-	-
	-	163	347	0	*	109	66	13	-	-
CRF-42_BF	+	1	1	0	*	*	6	3	2	2
	-	12	7	63	*	*	29	10	15	0
Subtipo B	+	9	7	13	*	*	45	4	14	3
	-	83	35	272	*	*	69	21	63	4
Subtipo F1	+	8	4	19	*	*	18	5	14	0
	-	86	43	325	*	*	87	21	66	0
CRF-63_02A6	+	*	*	0	*	21	3	0	-	-
	-	*	*	49	*	30	14	0	-	-
CRF-02_A1G	+	*	*	41	*	40	14	2	-	-
	-	*	*	399	*	112	48	0	-	-
Subtipo A6	+	*	*	19	*	50	13	0	-	-
	-	*	*	292	*	89	24	0	-	-

Los módulos que no están presentes en las CRFs se identifican con el signo "-", mientras que aquellos que están fuera de los genes estructurales, se identifican con "*".

8.4.2. Sitios bajo selección en los genes de las CRFs y de los subtipos puros que las constituyen

Posteriormente, al no encontrar un patrón de la selección natural negativa en los módulos de las CRFs, se llevó a cabo la identificación de los sitios bajo selección en los genes tanto de las CRFs como de los subtipos puros que las constituyen, para analizar si la modularidad podría estar rompiendo las huellas de selección existentes en los subtipos puros parentales.

La Tabla IX muestra los porcentajes de selección positiva y negativa en cada uno de los genes, el cual se obtuvo del cociente entre el número de sitios que se identificaron en ese gen dividido por el número total de sitios identificados en el genotipo y multiplicado por 100. El número total de sitios bajo selección positiva fue de 37, 132, 52, 13 y 2 sitios para el **CRF-06_cpx(A1,G,J,K)**, **subtipo A1**, **subtipo G**, **subtipo J** y el **subtipo K**, respectivamente, mientras que el número total de sitios bajo selección negativa fue de 352, 309, 469, 280 y 192.

Asimismo, se muestra en sombreado amarillo, los genes con un mayor número de sitios bajo selección siendo el gen que codifica para gp120 el que contenía un mayor número de sitios bajo selección positiva; mientras que, para la selección negativa, la mayoría de sitios se centró en el gen que codifica para la RT.

Tabla IX. Sitios bajo selección en las proteínas de la **CRF-06_cpx(A1,G,J,K)** y en sus subtipos puros que la constituyen.

Genotipo	Selección positiva								Selección negativa							
	MA	CA	NC	PR	RT	IN	gp120	gp41	MA	CA	NC	PR	RT	IN	gp120	gp41
CRF-06_cpx(A1,G,J,K)	9.091	9.091	0.000	3.030	12.121	3.030	51.515	12.121	5.000	16.875	4.063	5.313	41.563	7.500	17.813	1.875
Subtipo A1	4.545	3.030	0.758	4.545	47.727	12.879	21.212	5.303	9.385	28.155	7.120	1.294	8.414	1.294	36.893	7.443
Subtipo G	9.615	1.923	0.000	5.769	21.154	3.846	50.000	7.692	5.970	16.418	4.051	4.904	35.181	10.021	20.256	3.198
Subtipo J	7.692	7.692	0.000	0.000	7.692	0.000	76.923	0.000	8.214	16.429	3.571	3.929	28.929	6.429	27.857	4.643
Subtipo K	0.000	0.000	0.000	0.000	0.000	50.000	50.000	0.000	8.333	10.417	3.646	3.646	38.021	8.854	23.438	3.646

Este mismo patrón se presentó en la mayoría tanto de las CRFs como de los subtipos puros. Sin embargo, sólo el **subtipo A1** mostró patrones inversos, ya que la mayoría de los sitios bajo selección positiva recae en el gen de la RT, mientras que para la selección negativa, recae en el gen de gp120. También, la **CRF-63_02A6** tanto para la selección positiva como para la negativa, el mayor número de sitios se presentó en el gen de gp120.

Para visualizar las regiones de las proteínas en las que se presentaban los sitios bajo selección y relacionarlos con su importancia funcional, se mapearon en la estructura terciaria tanto de gp120 como de la RT. Las Figuras 27a y 29a son el modelo del monómero de gp120 que se utilizó como referencia para ilustrar los dominios que conforman a la proteína; mientras que las Figuras 28a y 30a son el modelo de la RT, que de igual forma se utilizó como referencia.

Como se muestra en la Figura 27b, la mayoría de los sitios que se encuentran bajo selección positiva en gp120 corresponden con las regiones variables 4 y 5, donde se podría esperar un mayor número de sustituciones no sinónimas que faciliten al virus escapar del sistema inmunitario del hospedero. Mientras que en la Figura 28b, se observa que la distribución de los sitios bajo selección negativa en la RT es constante tanto en el dominio de RNAsa H como en los subdominios "*palm*", "*fingers*", "*thumb*" y "*connection*" que pertenecen al dominio de polimerasa, lo cual concuerda con la importancia biológica de esta proteína que es fundamental para la replicación del virus.

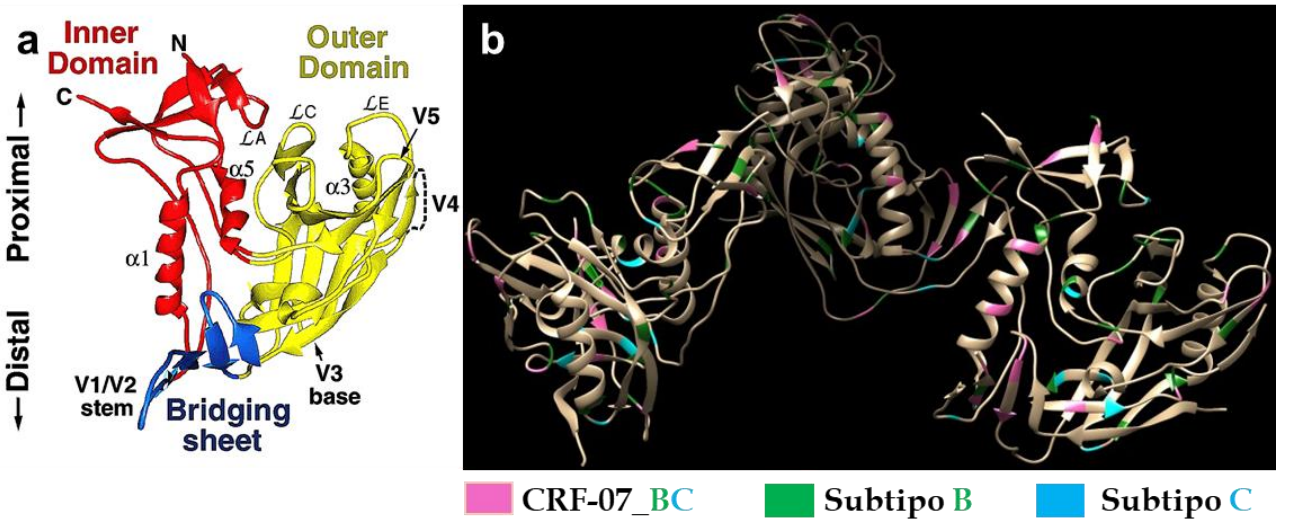

Figura 27. Sitios bajo selección positiva en gp120. **a)** Modelo del monómero de gp120. **b)** Sitios bajo selección positiva en el trímero de gp120 de la **CRF-07_BC** y de sus subtipos puros que la constituyen.

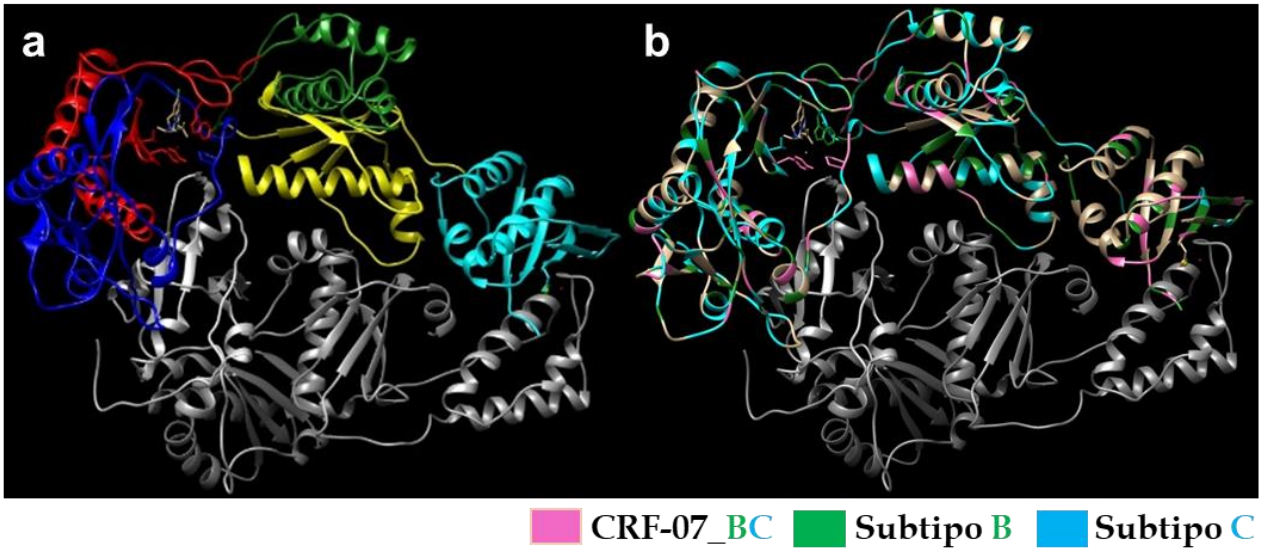

Figura 28. Sitios bajo selección negativa en la RT. **a)** Modelo de la RT. Azul: *fingers*, rojo: *palm*, verde: *thumb*, amarillo: *connection*, cian: RNAsa H, gris: subunidad p51. **b)** Sitios bajo selección negativa en la RT de la **CRF-07_BC** y de sus subtipos puros que la constituyen.

En las Figuras 29 y 30 se muestran los sitios bajo selección en las proteínas del **subtipo A1**. En la Figura 29b se muestran los sitios bajo selección negativa en gp120, de los cuales resaltan los que se encuentran en las regiones variables 4 y 5, y las del puente de hojas beta, que se han descrito como sitios críticos para la unión tanto de CD4+ y CCR5, como residuos de cisteína para la formación de puentes disulfuro (Yamaguchi-Kabata & Gojobori, 2000). Por otro lado, el mismo estudio destaca que las posiciones 103 y 106 están bajo selección negativa, ya que son residuos fundamentales para formar una interfaz trimérica, mientras que en nuestro caso las posiciones 104 y 106 están bajo selección negativa.

Por otra parte, la Figura 30b muestra los sitios bajo selección positiva en la RT. De los cuales, los sitios 185 y 186 han sido asociados por Coffin y colaboradores (1997) al centro catalítico de "*palm*", se trata de tres ácidos aspárticos en los residuos 110, 185 y 186. Buscando en la base de datos de la Universidad de Stanford (la cual contiene todas las mutaciones reportadas en secuencias de VIH-1 con resistencia a retrovirales, tanto para la PR, RT e IN), encontramos que están reportadas cuatro mutaciones para el residuo 185 mientras que para el residuo 186 son cinco, ambas mutaciones se han

reportado en virus de pacientes bajo tratamiento tanto con inhibidores de la RT análogos a nucleósidos (NRTI) como no análogos (NNRTI).

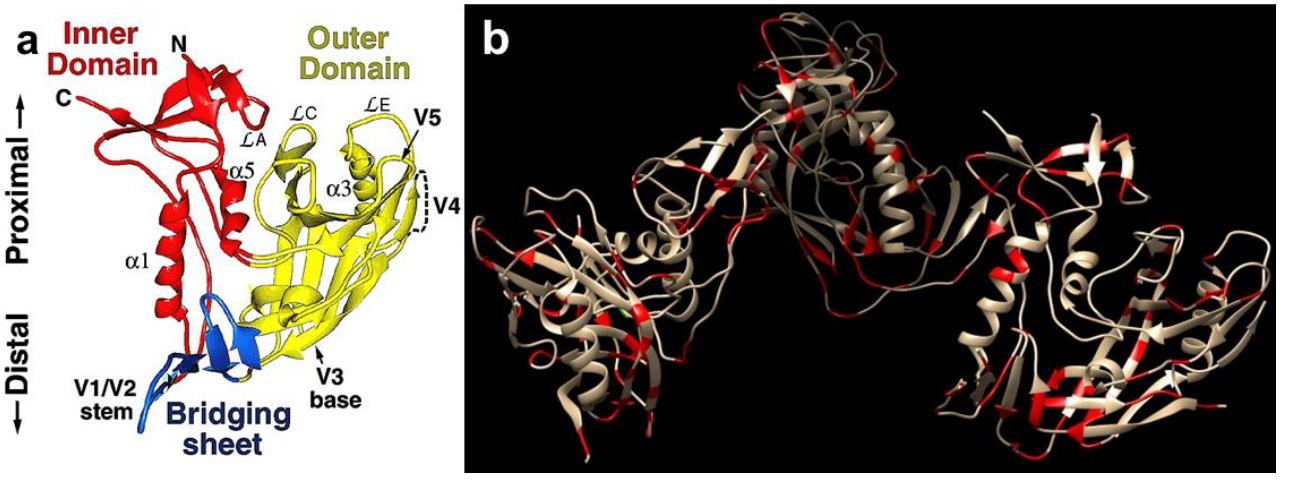

Figura 29. Sitios bajo selección negativa en gp120 del **subtipo A1**. **a)** Modelo del monómero de gp120. **b)** Sitios bajo selección negativa en el trímero de gp120 del **subtipo A1**.

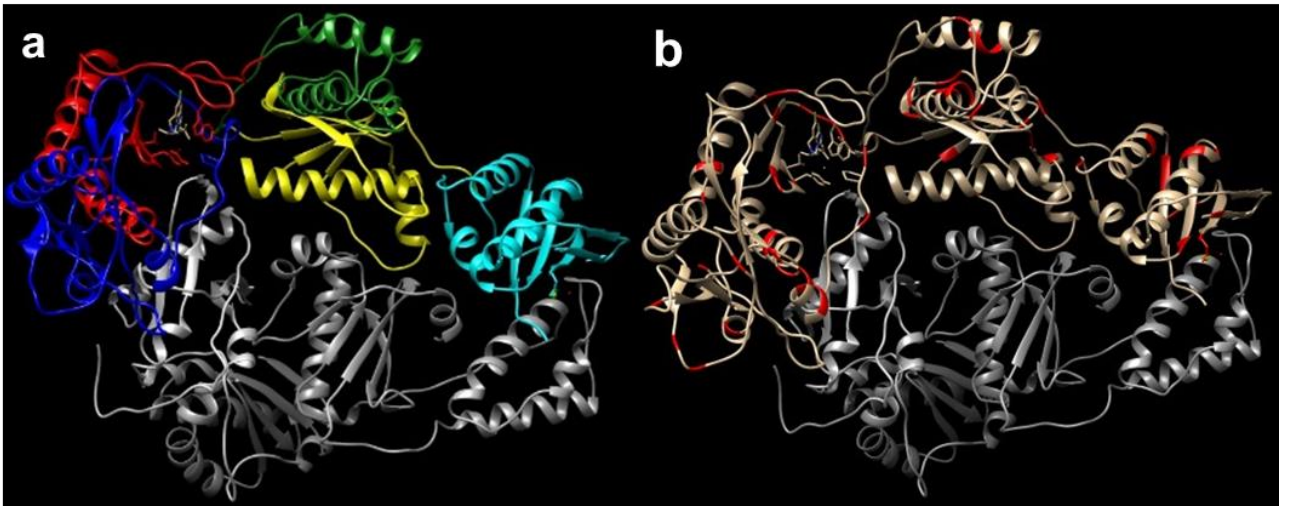

Figura 30. Sitios bajo selección positiva en la RT del **subtipo A1**. **a)** Modelo de la RT. Azul: *fingers*, rojo: *palm*, verde: *thumb*, amarillo: *connection*, cian: RNAsa H, gris: subunidad p51. **b)** Sitios bajo selección positiva en la RT del **subtipo A1**.

8.4.3. Sitios bajo coevolución en los genomas de las CRFs y de los subtipos puros que las constituyen

En la Tabla X se muestran el número de sitios bajo coevolución tanto en las CRFs como en los subtipos puros que las constituyen. Se observa que en todas las CRFs se encuentra un menor número de sitios bajo coevolución respecto a los subtipos puros que las constituyen. Inferimos que esto podría ser debido al tiempo de la historia evolutiva, ya que como menciona Tebit y Arts en el año 2011, el ancestro común del grupo M del VIH-1 data desde el año 1908, mientras que la primera CRF, la CRF-01_AE se identificó en 1980; teniendo así un mayor tiempo para la generación de redes evolutivas dentro y entre sus proteínas los subtipos puros.

Asimismo, se puede observar que la **CRF-07_BC**, **CRF-14_BG** y la **CRF-42_BF** sólo contienen un sitio bajo coevolución. Igualmente, creemos que esto es debido al tiempo de la historia evolutiva del **subtipo B**, ya que como menciona Vega-Sanabria en 2016, la propagación de este subtipo se originó desde el año de 1960, cuando el **subtipo B** emigró del Congo a Haití, posteriormente a Estados Unidos y de ahí a varios países del mundo; por lo cual, se considera que el **subtipo B** es un subtipo joven y sus CRFs tienen menos tiempo evolucionando.

Tabla X. Sitios bajo coevolución en las CRFs y los subtipos puros que las constituyen.

Genotipo	Sitios bajo coevolución	Intra-módulo	Inter-módulo
CRF-02_A1G	22	59.10%	40.90%
Subtipo A1	37	35.14%	64.86%
Subtipo G	22	68.18%	31.82%
CRF-06_cpx(A1,G,J,K)	2	0%	100%
Subtipo A1	37	54.05%	45.95%
Subtipo G	22	54.54%	45.46%
Subtipo J	0	0%	0%
Subtipo K			
CRF-07_BC	1	0%	100%
Subtipo B	30	30%	70%
Subtipo C	29	37.93%	62.07%
CRF-11_cpx(A1,E,G,J,U)	26	53.85%	46.15%

Subtipo A1	37	37.84%	62.16%
Subtipo G	22	68.18%	31.82%
Subtipo J	0	0%	0%
Subtipo U	61	60.66%	39.34%
CRF-14_BG	1	100%	0%
Subtipo B	30	46.66%	53.34%
Subtipo G	22	31.82%	61.18%
CRF-35_A1D	11	36.36%	63.64%
Subtipo A1	37	35.14%	64.86%
Subtipo D	24	29.17%	70.83%
CRF-42_BF	1	100%	0%
Subtipo B	30	23.33%	76.67%
Subtipo F1	24	37.50%	62.50%
CRF-63_02A6	0	0%	0%
CRF-02_A1G	22	40.90%	59.10%
Subtipo A6	37	43.24%	56.76%

En sombreado color rojo se encuentra el **subtipo K**, ya que sólo hay 3 secuencias, no se pudo llevar a cabo la filogenia bayesiana.

Finalmente, al asociar los sitios bajo coevolución en las coordenadas de sus proteínas, de los 86 sitios bajo coevolución identificados en las CRFs, 24 sitios se identificaron en la interacción de gp120-gp120 y 14 sitios de la interacción gp120-gp41; mientras que de los 501 sitios bajo coevolución detectados en los subtipos puros, 116 y 74 sitios se identificaron en estas mismas interacciones.

En la Figura 31 se muestran ocho gráficas, donde en la parte superior están las gráficas correspondientes a los sitios bajo coevolución en las proteínas de las CRFs, tanto intra- como inter-módulos y las interacciones tanto de una proteína consigo misma como con otras; en la parte media se presentan de igual manera las gráficas de los sitios bajo coevolución en las proteínas de los subtipos puros; mientras que en la parte inferior se muestran las leyendas para la identificación de las relaciones. Se puede observar que en todas las gráficas está presente la relación de gp120-gp120 y de gp120-gp41, indicada en color café. Nuestros resultados muestran que estas relaciones son de gran relevancia para el virus.

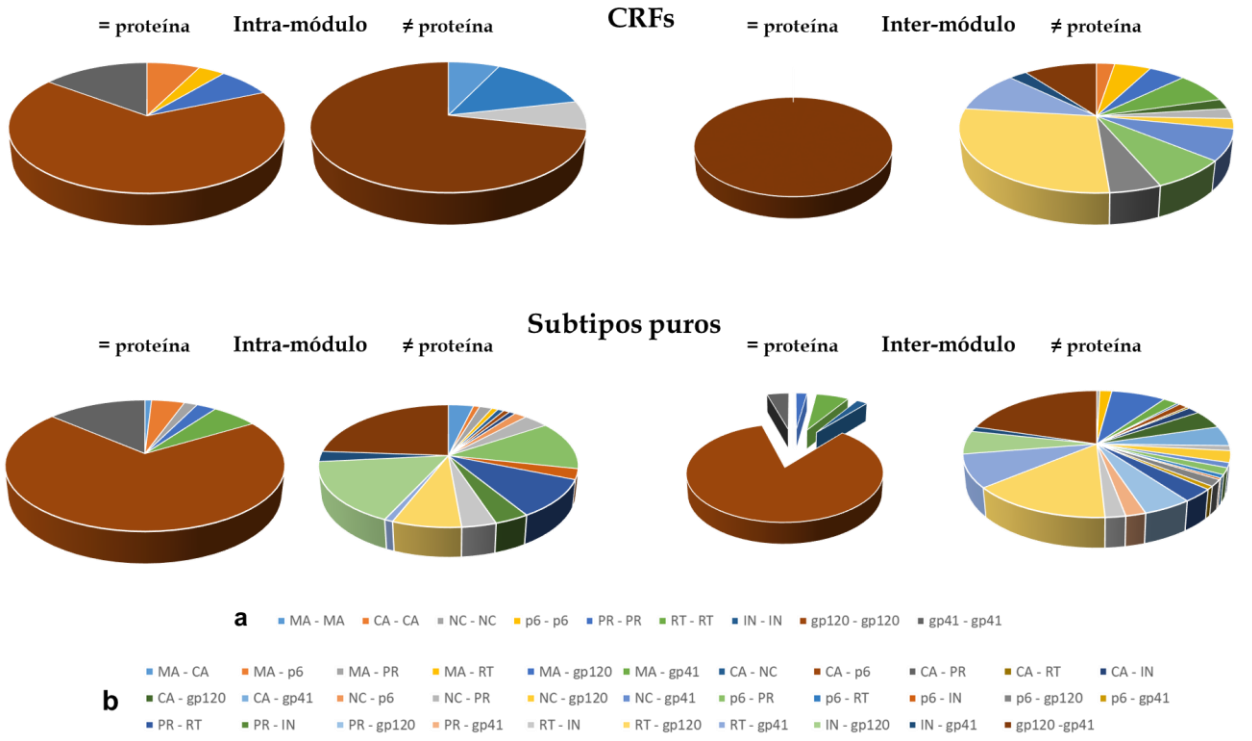

Figura 31. Sitios bajo coevolución en las proteínas de las CRFs y los subtipos puros que los constituyen. **a)** Leyendas para las interacciones de una misma proteína. **b)** Leyendas para las interacciones entre diferentes proteínas. Se observa que tanto en las CRFs como en los subtipos puros, la interacción de gp120-gp120 y gp120-gp41 es la que predomina.

9. DISCUSIÓN

9.1. Identificación de módulos en las CRFs

Todos los puntos de recombinación identificados en la **CRF-02_A1G** y en la **CRF-11_cpx(A1,G,J,U)** se encuentran en los genes *gag* y *pol*, mientras que en el resto de las CRFs analizadas, la mayoría de los puntos de recombinación se encuentran a lo largo del gen *env*. Estos resultados correlacionan con los puntos de recombinación reportados por Fan y colaboradores en 2007, donde, como se muestra en la Figura 32, señalan el número de puntos de recombinación identificados a lo largo del genoma del VIH-1. En la figura se muestran los "*hot spots*" y los "*cold spots*", que son regiones con una mayor probabilidad de que ocurra un punto de recombinación y regiones que participan con una menor asiduidad que un segmento en promedio (Watson, 2006), respectivamente.

Por lo tanto, en seis de las ocho CRFs, la mayoría de los puntos de recombinación se encuentran en el gen *env*, aunque como menciona Fan y colaboradores en 2007, no existe una explicación mecanicista para la acumulación de puntos de recombinación en esa región, parece ser una ventaja selectiva en el proceso de recombinación del VIH-1.

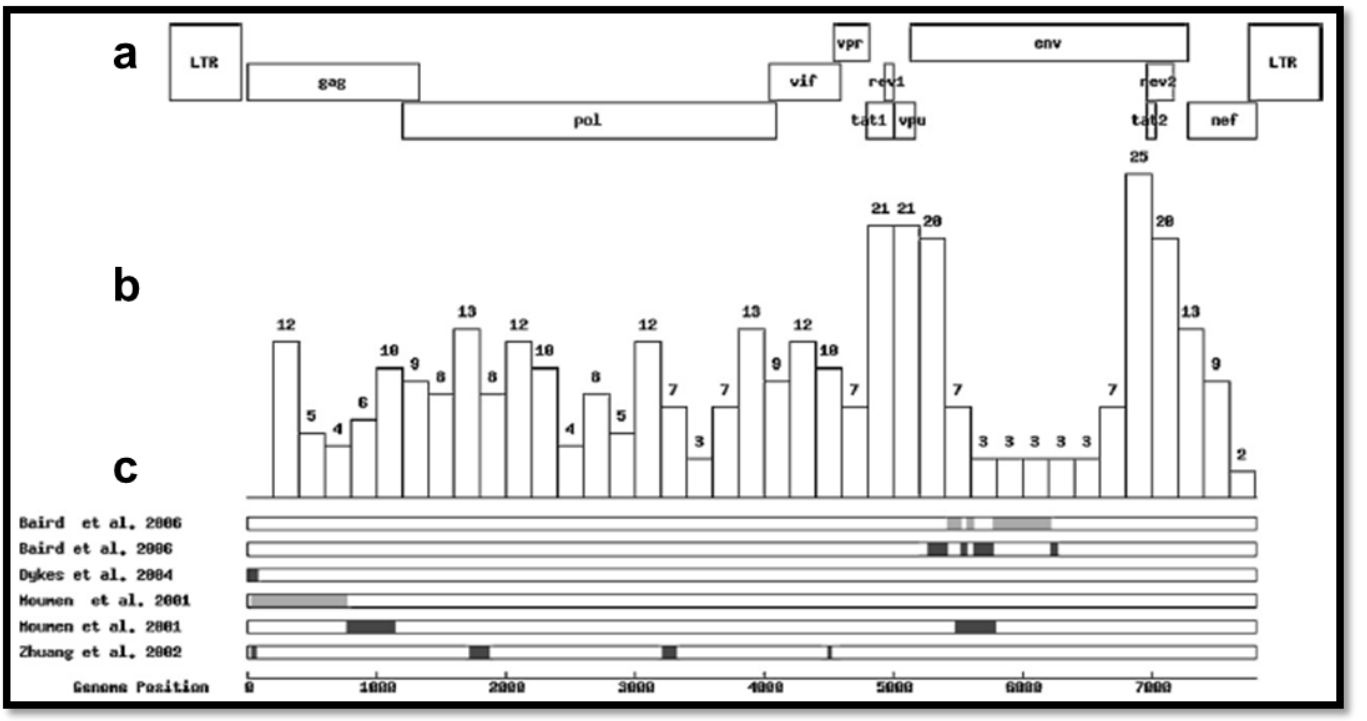

Figura 32. Distribución de puntos de recombinación a través del genoma del VIH-1. **a)** Representación esquemática del genoma del VIH-1. **b)** Número de puntos de recombinación detectados en el estudio de Fan *et al.*, 2007. **c)** Regiones analizadas en estudios experimentales; se infiere que son "*cold spots*" y "*hot spots*" las barras grises claras y oscuras, respectivamente. La posición del genoma se indica en la parte inferior. Modificada de Fan *et al.*, 2007.

De los 60 módulos identificados en las ocho CRFs, un 15% de ellos pertenecen al gen *gag*, mientras que un 26.7% al gen *pol*, un 45% al gen *env* y un 13.3% restante, que se identificó en genes que codifican proteínas accesorias y en el 5´-LTR. Estos resultados demuestran que la acumulación de módulos en el gen *env* es mayoritaria, probablemente debido a su importancia biológica, ya que juega un papel importante en los procesos de entrada y escape del sistema inmunitario del hospedero.

Souza y colaboradores en 2019 llevaron a cabo la caracterización de 15 CRFs que contienen al subtipo B y F en Brasil, donde identificaron los puntos de recombinación presentes en cada una de ellas y observaron el patrón que se muestra en la Figura 33. En esta figura se puede notar claramente que cada una de las CRFs, aunque estén formadas por los mismos subtipos, estará conformada de los módulos que le confieran una ventaja selectiva respecto a otras CRFs en la misma población viral, es por ello que

no se observa un patrón estandarizado para la formación de módulos en algún gen en específico.

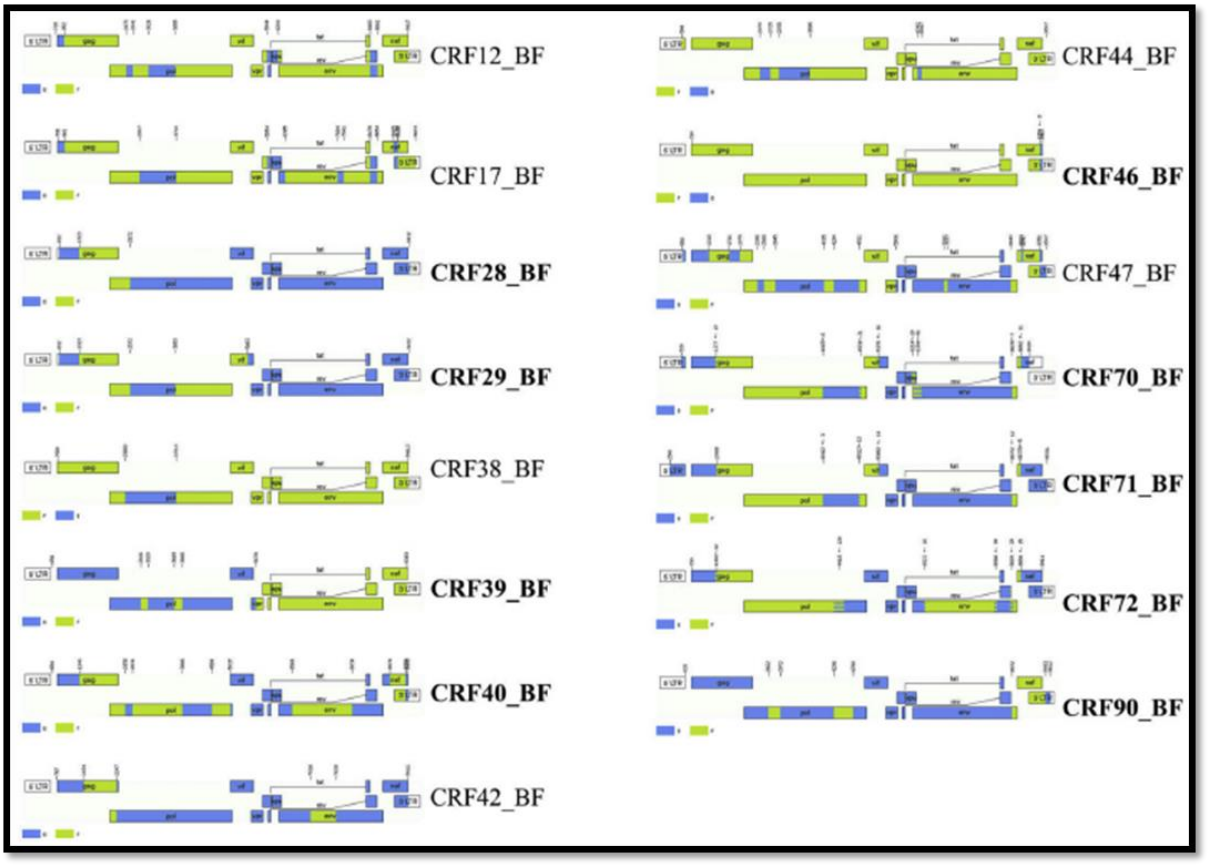

Figura 33. Patrones de recombinación de las 15 CRF_BF que circulan a nivel mundial. Las CRFs descritas en Brasil se identifican en negrita. Tomada de Souza *et al.*, 2019.

9.2. Diversidad presente tanto en el genoma completo como en los módulos de las CRFs y de los subtipos puros que las constituyen

La mutación y la recombinación son dos fuerzas principales que generan variación genética, el material crudo con el que la selección natural se alimenta. No obstante, una fracción pequeña de la variación generada por mutación y recombinación produce innovaciones evolutivas, la mayor parte de esta variación puede ser deletérea (Martin & Wagner, 2009). Aunque la recombinación se lleve a cabo intra-subtipos, ésta no puede ser detectada, ya que tienen el mismo fondo genético (Nikolaitchik *et al.*, 2015).

Nuestros resultados demostraron que los subtipos puros contienen una mayor diversidad genética respecto a sus CRFs, e inferimos que esto es debido a la característica de robustez mutacional, ya que un subtipo puro tiene la flexibilidad de poder recibir mutaciones a lo largo de su genoma sin afectar su adecuación biológica; sin embargo, la modularidad restringe esa flexibilidad en una CRF, ya que proviene de 2 o más subtipos parentales. De acuerdo con Masel y Trotter en 2010, la robustez mutacional puede incrementar la evolucionabilidad al permitir a la población visitar un número mayor de genotipos durante un período de tiempo dado.

Gibson y Dworkin en 2004 definen la **variación genética críptica** como la variación genética permanente que no contribuye al rango normal de fenotipos observados en una población, pero que está disponible para modificar un fenotipo que surge después del cambio ambiental. Un término relativamente nuevo es el de **capacitancia evolutiva**, que se utiliza para describir el escondite y la liberación de la variación genética críptica (Masel & Trotter, 2010). Esto concuerda con nuestros resultados, ya que al asociarlos con estos conceptos, un subtipo puro podría considerarse un "capacitor" completo, que podría explicar la abundancia de su diversidad genética; mientras que una CRF, al estar compuesta de varios fragmentos de diferentes subtipos, podría romper la estructura de cada uno de los "capacitores" que la constituyen, y por lo tanto, su variación genética es menor.

9.3. Efecto de la selección natural en la evolución de las CRFs y de los subtipos puros que las constituyen

9.3.1. Efecto de la selección natural en los módulos

Basándonos en nuestros resultados anteriores, inferimos que los subtipos puros deberían tener un mayor número de sitios bajo selección positiva, ya que como menciona Nielsen en 2005, cuando un sitio es seleccionado positivamente es porque existe un exceso en el número de sustituciones no sinónimas por sitio no sinónimo (dN) respecto al número de sustituciones sinónimas por sitio sinónimo (dS), esto debido a que contienen una mayor variación genética en la mayoría de los módulos; mientras

que las CRFs, al tener una menor diversidad en los módulos, supusimos que deberían tener un mayor número de sitios bajo selección negativa.

Sin embargo, sólo observamos el patrón para los subtipos puros, teniendo éstos un mayor número de sitios bajo selección positiva y también bajo selección negativa. A este respecto, Martin y colaboradores en 2005 demostraron experimentalmente que los fragmentos de material genético (módulos) del Virus de la Vena del Maíz (MSV) solo funcionan de manera óptima si residen dentro de genomas similares a aquellos en los que evolucionaron, por lo tanto, posiblemente la modularidad podría estar eliminando los sitios bajo selección negativa en las CRFs, aumentando así su evolucionabilidad.

9.3.2. Efecto de la selección natural en los genes

Nuestros resultados acerca de la identificación de los sitios bajo selección natural en los genes correlacionan con varios hallazgos reportados. Dos de ellos, el estudio realizado por Ross y Rodrigo en 2002 y otro estudio llevado a cabo por Canducci y colaboradores en 2009, donde identificaron huellas de selección positiva en las regiones variables V1-V5 en gp120 y en la superficie del sitio de unión al receptor CD4+, respectivamente; como ellos mencionan, se postula que estos sitios ayudan al virus a escapar del sistema inmune del hospedero. Mientras que el estudio realizado por Seibert y colaboradores en 1995, mostró que la transcriptasa reversa está bajo selección negativa; probablemente debido a su importante papel en el ciclo de replicación del virus. Ambas proteínas resultaron concordar con dicho patrón, tanto en la mayoría de los subtipos puros como al igual que en las CRFs analizadas.

Sin embargo, el **subtipo A1**, presentó un patrón inverso, siendo el gen que codifica para la transcriptasa reversa el que contiene un mayor número de sitios bajo selección positiva; mientras que, para la selección negativa, lo fue el gen que codifica para gp120. También, se observó que la **CRF-63_02A6**, tanto para la selección positiva como negativa, gp120 presentó un mayor número de sitios bajo selección.

De acuerdo con Yamaguchi-Kabata & Gojobori en el año 2000, existen varios sitios bajo selección negativa en gp120, entre los que se identificaron sitios críticos en las regiones variables V4 y V5 para la unión al receptor CD4+ y a los correceptores CCR5 y CXCR4,

al igual que sitios en las hojas beta del puente entre los dominios interno y externo de gp120, siendo residuos de cisteína que ayudan en la formación de los puentes disulfuro. Por otro lado, detectaron dos sitios bajo selección negativa en las posiciones 103 y 106 que se encuentran en el dominio interno, ambos son fundamentales para formar la interfaz del trímero de gp120. Por último, como se muestra en la Figura 34, en la estructura primaria de gp120 se identifican los sitios bajo selección negativa (color azul) de la posición 211 a 259, los cuales en su mayoría correlacionan con nuestros resultados.

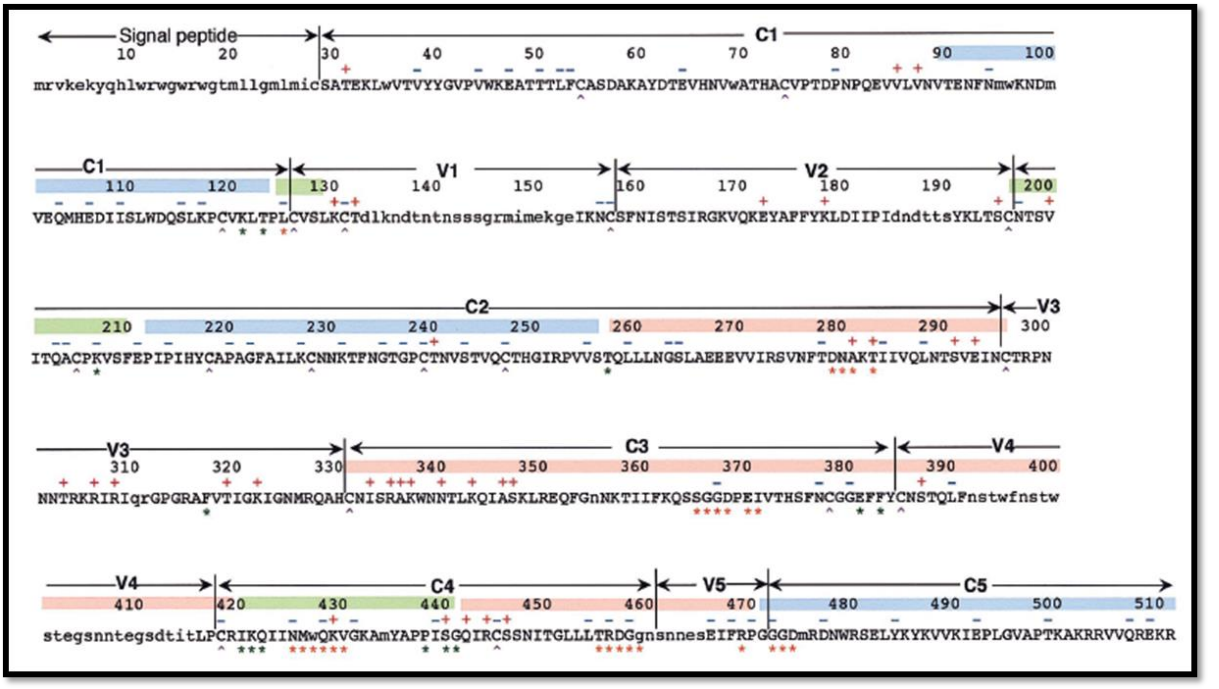

Figura 34. Sitios bajo selección en la estructura primaria de gp120. Los sitios bajo selección positiva se muestran en color rojo, mientras que los sitios bajo selección negativa en color azul; en color verde se muestran los sitios neutrales. Tomada de Yamaguchi-Kabata & Gojobori, 2000.

Chen y colaboradores en 2004 llevaron a cabo la identificación de sitios bajo selección positiva en la RT, los cuales se encontraron en sitios ligados a mutaciones de resistencia a fármacos, que se encuentran reportados particularmente en los subdominios "*fingers*" y "*palm*", de este último los codones 86 a 117 y 156 a 237. Lo anterior concuerda con nuestros resultados, ya que, de los sitios encontrados bajo selección positiva, destacan las posiciones 185 y 186; que de acuerdo con Coffin y colaboradores en 1997, son parte

del centro catalítico de "*palm*" compuesto por tres ácidos aspárticos en los residuos 110, 185 y 186, como se muestra en la Figura 35. Por lo tanto, estas dos últimas posiciones podrían estar jugando un gran papel evolutivo en el **subtipo A1**, debido a mutaciones de escape claves.

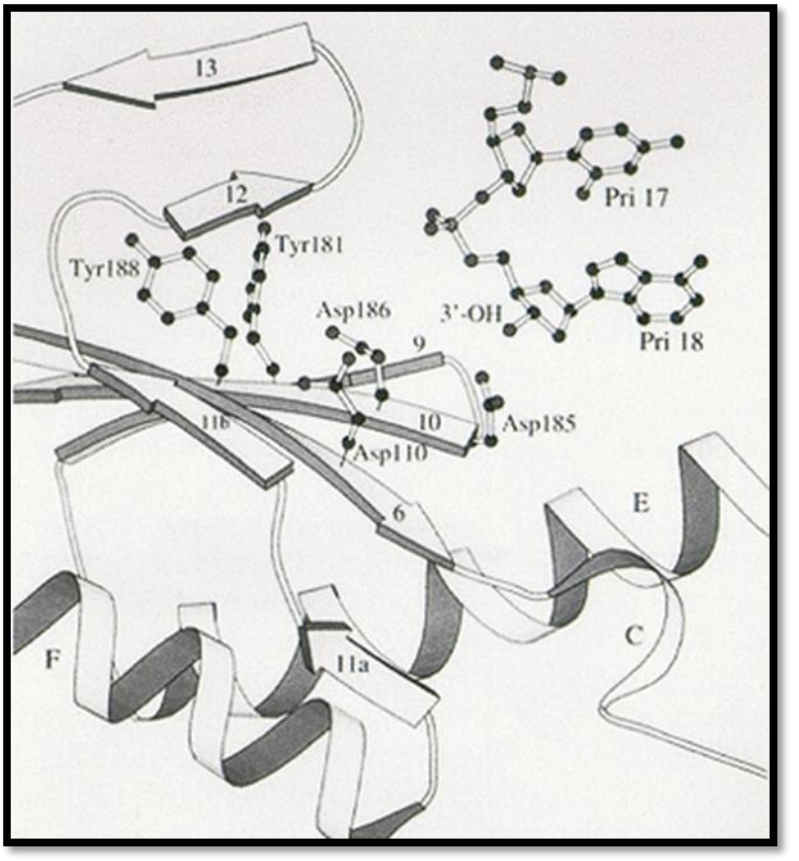

Figura 35. Vista del sitio activo del dominio de polimerasa de la RT del VIH-1 en un complejo con ADN bicatenario. Se observa que el centro catalítico está conformado por tres residuos de ácido aspártico en las posiciones 110, 185 y 186. Tomada de Coffin *et al.*, 1997.

9.3.3. Efecto de la coevolución

La identificación de las proteínas que coevolucionan tanto en las CRFs como en los subtipos puros nos permitió observar que existe una interacción mayoritaria de gp120 consigo misma y con gp41, tanto en el mismo módulo como entre módulos. Como menciona Fraser en 2006, se puede determinar que dos proteínas coevolucionan si se demuestra que existen interacciones físicas entre ambas dentro de una especie. Por lo tanto, esto concuerda con nuestros resultados, ya que tanto gp120 como gp41, son las

proteínas de la envoltura del virión e interaccionan entre ellas para llevar a cabo la unión con sus células blanco.

Travers y colaboradores en 2007 identificaron coevolución entre residuos de los dominios C2, V3 y C3 de gp120, proponiendo que esto ocurre con la finalidad de mantener las propiedades estructurales generales para la función óptima de la proteína. También, identificaron coevolución entre residuos de gp120 con el dominio citoplasmático de gp41, específicamente con una secuencia denominada "Kennedy", para la cual el sistema inmune del hospedero produce anticuerpos. Estos autores concluyen que la coevolución identificada entre residuos de aminoácidos a través del gen *env* refleja la codependencia funcional del trímero de gp120-gp41.

Cabe destacar, que en las **CRF-07_BC**, **CRF-14_BG** y **CRF-42_BF** que sólo contienen un sitio bajo coevolución, que indica la interacción de gp120 consigo misma, sugiriendo que esta relación podría ser la primera en desarrollarse, ya que como mencionan Tebit y Arts en 2011, en un individuo con infección reciente, la mayoría de los virus usan el correceptor CCR5, pero a medida que la infección progresa, las variantes generadas pueden usar el correceptor CXCR4, o ambos, pudiendo así emerger y dominar la población de VIH-1; la habilidad de usar tanto un correceptor como otro, está mediada por la variación genética en la región variable 3 (V3) de gp120.

10. CONCLUSIONES

1. La modularidad en el VIH-1 se puede presentar en cualquier parte de su genoma, dependerá de las diferentes presiones de selección a las que se encuentre la población viral.
2. La modularidad restringe la diversidad genética, debido a que existe una disminución en la robustez mutacional de la CRF resultante.
3. La modularidad permite las marcas de selección provenientes de los subtipos puros parentales.
4. La modularidad elimina las interacciones co-evolutivas presentes en los subtipos puros parentales.

11. BIBLIOGRAFÍA

1. **Bailes**, E., Gao, F., Bibollet-Ruche, F., Courgnaud, V., Peeters, M., Marx, P. A., Hahn, B. H. & Sharp, P. M. (**2003**). Hybrid Origin of SIV in Chimpanzees. *Science, 300*(5626): 1713-1713.
2. **Baltimore**, D. (**1971**). Expression of Animal Virus Genomes. *Bacteriological Reviews, 35*(3): 235-41.
3. **Basu**, V. P., Song, M., Gao, L., Rigby, S. T., Hanson, M. N., & Bambara, R. A. (**2008**). Strand transfer events during HIV-1 reverse transcription. *Virus Research, 134*(1-2): 19-38. doi: 10.1016/j.virusres.2007.12.017.
4. **Beloukas**, A., Psarris, A., Giannelou, P., Kostaki, E., Hatzakis, A., & Paraskevis, D. (**2016**). Molecular epidemiology of HIV-1 infection in Europe: An overview. *Infection, Genetics and Evolution, 46*: 180-189.
5. **Bollback**, J. (**2006**). SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics, 7*(1): 88.
6. **Bonner**, J. T. (**1988**). The Evolution of Complexity by Means of Natural Selection. 1st Edition, Princeton University Press, Princeton, New Jersey: 174-175.
7. **Canducci**, F., Marinozzi, M., Sampaolo, M., Berrè, S., Bagnarelli, P., Degano, M., … Clementi, M. (**2009**). Dynamic features of the selective pressure on the human immunodeficiency virus type 1 (HIV-1) gp120 CD4-binding site in a group of long term non progressor (LTNP) subjects. *Retrovirology, 6*(1): 4. doi:10.1186/1742-4690-6-4.
8. **Castro-Nallar**, E., Pérez-Losada, M., Burton, G. F., & Crandall, K. A. (**2012**). The evolution of HIV: Inferences using phylogenetics. *Molecular Phylogenetics and Evolution, 62*(2): 777-792.
9. **Chen**, L., Perlina, A., & Lee, C. J. (**2004**). Positive Selection Detection in 40,000 Human Immunodeficiency Virus (HIV) Type 1 Sequences Automatically Identifies Drug Resistance and Positive Fitness Mutations in HIV Protease and Reverse Transcriptase. *Journal of Virology, 78*(7): 3722-3732. doi:10.1128/jvi.78.7.3722-3732.2004.
10. **Cimarelli**, A., & Darlix, J.-L. (**2014**). HIV-1 Reverse Transcription. Human Retroviruses, 55–70. doi:10.1007/978-1-62703-670-2_6.
11. **Coffin**, J. M., Hughes, S. H. & Varmus, H. E. (**1997**). Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.
12. **Coiras**, M., López-Huertas, M. R., Pérez-Olmeda, M., & Alcamí, J. (**2009**). Understanding HIV-1 latency provides clues for the eradication of long-term reservoirs. *Nature Reviews Microbiology, 7*(11): 798-812.
13. **Crooks**, G. E. (**2004**). WebLogo: A Sequence Logo Generator. *Genome Research, 14*(6): 1188-1190.
14. **Cuevas**, J. M., Geller, R., Garijo, R., López-Aldeguer, J., & Sanjuán, R. (**2015**). Extremely High Mutation Rate of HIV-1 In Vivo. *PLOS Biology, 13*(9): e1002251. doi: 10.1371/journal.pbio.1002251.

15. **De Goede**, A. L., Vulto, A. G., Osterhaus, A. D. M. E., & Gruters, R. A. (**2015**). Understanding HIV infection for the design of a therapeutic vaccine. Part I: Epidemiology and pathogenesis of HIV infection. *Annales Pharmaceutiques Françaises, 73*(2): 87-99.
16. **Delatorre**, E., & Bello, G. (**2013**). Phylodynamics of the HIV-1 Epidemic in Cuba. *PLoS ONE, 8*(9): e72448.
17. **Delatorre**, E., Couto-Fernandez, J. C., & Bello, G. (**2017**). HIV-1 Genetic Diversity in Northeastern Brazil: High Prevalence of Non-B Subtypes. *AIDS Research and Human Retroviruses, 33*(7): 639-647.
18. **Désiré**, N., Cerutti, L., Le-Hingrat, Q., Perrier, M., Emler, S., Calvez, V., Descamps, D., Marcelin, A.G., Hué, S. & Visseaux B. (**2018**). Characterization update of HIV-1 M subtypes diversity and proposal for subtypes A and D sub-subtypes reclassification. *Retrovirology, 15*(1): 80.
19. **Dolan**, P. T., Whitfield, Z. J., & Andino, R. (**2016**). Mechanisms and Concepts in RNA Virus Population Dynamics and Evolution. *Annual Review of Virology, 5*(1). doi:10.1146/annurev-virology-101416-041718.
20. **Edgar**, R. C. (**2004**). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*(5): 1792-1797.
21. **Fan**, J., Negroni, M., & Robertson, D. L. (**2007**). The distribution of HIV-1 recombination breakpoints. *Infection, Genetics and Evolution, 7*(6): 717-723. doi:10.1016/j.meegid.2007.07.012
22. **Fares,** M. A. (**2015**). The origins of mutational robustness. *Trends in Genetics, 31*(7): 373-381.
23. **Frankel**, A. D., & Young, J. A. T. (**1998**). HIV-1: Fifteen Proteins and an RNA. *Annual Review of Biochemistry, 67*(1): 1-25.
24. **Fraser**, C., Lythgoe, K., Leventhal, G. E., Shirreff, G., Hollingsworth, T. D., Alizon, S., & Bonhoeffer, S. (**2014**). Virulence and Pathogenesis of HIV-1 Infection: An Evolutionary Perspective. *Science, 343*(6177): 1243727-1243727. doi: 10.1126/science.1243727.
25. **Fraser**, H. B. (**2006**). Coevolution, modularity and human disease. *Current Opinion in Genetics & Development, 16*(6): 637-644. doi:10.1016/j.gde.2006.09.001.
26. **Galetto**, R. & Negroni, M. (**2005**). Mechanistic features of recombination in HIV. *AIDS Rev., 7*(2): 92-102.
27. **Galetto**, R., Giacomoni, V., Véron, M., & Negroni, M. (**2005**). Dissection of a Circumscribed Recombination Hot Spot in HIV-1 after a Single Infectious Cycle. *Journal of Biological Chemistry, 281*(5): 2711-2720.
28. **Gibson**, G., & Dworkin, I. (**2004**). Uncovering cryptic genetic variation. *Nature Reviews Genetics, 5*(9): 681-690. doi:10.1038/nrg1426.
29. **Gonda**, M. A. (**1988**). Molecular genetics and structure of the human immunodeficiency virus. *Journal of Electron Microscopy Technique, 8*(1): 17-40.

30. **Grabow**, W., & Jaeger, L. (**2013**). RNA modularity for synthetic biology. *F1000Prime Reports*, 5. doi:10.12703/p5-46
31. **Hemelaar**, J. (**2013**). Implications of HIV diversity for the HIV-1 pandemic. *Journal of Infection, 66*(5): 391-400.
32. **Holmes**, E. C. (**2011**). What Does Virus Evolution Tell Us about Virus Origins? *Journal of Virology, 85*(11): 5247-5251.
33. **Hu**, W. S., & Hughes, S. H. (**2012**). HIV-1 Reverse Transcription. *Cold Spring Harbor Perspectives in Medicine, 2*(10): a006882–a006882.
34. **Huelsenbeck**, J. P., & Ronquist, F. (**2001**). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics, 17*(8): 754-755. doi: 10.1093/bioinformatics/17.8.754.
35. **Kerina**, D., Stray-Pedersen, B. & Müller, F. (**2013**). HIV Diversity and Classification, Role in Transmission. *Advances in Infectious Diseases, 3*(1): 146-156.
36. **Koonin**, E. V., Dolja, V. V., & Krupovic, M. (**2015**). Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology, 479*(480): 2-25.
37. **Kosakovsky-Pond**, S. L. (**2016**). Quantifying Natural Selection in Coding Sequences [diapositivas de PowerPoint]. Recuperado de: https://www.dropbox.com/s/adbizki8q3kso66/Molecular-adaptation.pdf?dl=0
38. **Kosakovsky-Pond**, S. L., & Frost, S. D. W. (**2005**). Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Molecular Biology and Evolution, 22*(5): 1208-1222.
39. **Kosakovsky-Pond**, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., & Frost, S. D. W. (**2006**). Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm. *Molecular Biology and Evolution, 23*(10): 1891-1901.
40. **Kumar**, S., Stecher, G., & Tamura, K. (**2016**). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution, 33*(7): 1870-1874. doi:10.1093/molbev/msw054
41. **Larsson**, A. (**2014**). AliView: a fast and lightweight alignment viewer and editor for large data sets. *Bioinformatics, 30*(22): 3276-3278.
42. **Lauring**, A. S., Frydman, J., & Andino, R. (**2013**). The role of mutational robustness in RNA virus evolution. *Nature Reviews Microbiology, 11*(5): 327-336.
43. **Lenski**, R. E. (**2017**). What is adaptation by natural selection? Perspectives of an experimental microbiologist. *PLOS Genetics, 13*(4): e1006668.
44. **Lever**, A., & Jeang, K.-T. (**2006**). Replication of Human Immunodeficiency Virus Type 1 from Entry to Exit. *International Journal of Hematology, 84*(1): 23-30.
45. **Malim**, M. H., & Emerman, M. (**2008**). HIV-1 Accessory Proteins—Ensuring Viral Survival in a Hostile Environment. *Cell Host & Microbe, 3*(6): 388-398.

46. **Martin**, D. P., van der Walt, E., Posada, D., & Rybicki, E. P. (**2005**). The Evolutionary Value of Recombination Is Constrained by Genome Modularity. *PLoS Genetics, 1*(4): e51. doi: 10.1371/journal.pgen.0010051.
47. **Martin**, O. C., & Wagner, A. (**2009**). Effects of Recombination on Complex Regulatory Circuits. *Genetics, 183*(2): 673-684. doi:10.1534/genetics.109.104174.
48. **Masel**, J., & Trotter, M. V. (**2010**). Robustness and Evolvability. *Trends in Genetics, 26*(9): 406-414. doi:10.1016/j.tig.2010.06.002.
49. **McDonald**, S. M., Nelson, M. I., Turner, P. E., & Patton, J. T. (**2016**). Reassortment in segmented RNA viruses: mechanisms and outcomes. *Nature Reviews Microbiology, 14*(7): 448-460.
50. **Modrow**, S., Falke, D., & Truyen, U. (**2003**). Molekulare Virologie, 2. Auflage. Spektrum Akademischer Verlag. Heidelberg, Berlin.
51. **Mol**, M., Kabra, R., & Singh, S. (**2018**). Genome modularity and synthetic biology: Engineering systems. *Progress in Biophysics and Molecular Biology, 132*: 43-51.
52. **Mount**, D. W. (**2008**). Maximum Parsimony Method for Phylogenetic Prediction. *Cold Spring Harbor Protocols, 3*(4): 1-6. doi:10.1101/pdb.top32.
53. **Neher**, R. A., & Leitner, T. (**2010**). Recombination Rate and Selection Strength in HIV Intra-Patient Evolution. *PLoS Comput Biol 6*(1): e1000660.
54. **Nielsen**, R. (**2005**). Molecular Signatures of Natural Selection. *Annual Review of Genetics, 39*(1): 197-218. doi:10.1146/annurev.genet.39.073003.112420.
55. **Nikolaitchik**, O., Keele, B., Gorelick, R., Alvord, W. G., Mazurov, D., Pathak, V. K., & Hu, W.-S. (**2015**). High recombination potential of subtype A HIV-1. *Virology, 484*(1): 334-340.
56. **Nkeze**, J., Li, L., Benko, Z., Li, G., & Zhao, R. Y. (**2015**). Molecular characterization of HIV-1 genome in fission yeast *Schizosaccharomyces pombe*. *Cell & Bioscience, 5*(1). doi:10.1186/s13578-015-0037-7
57. **Onafuwa-Nuga**, A., & Telesnitsky, A. (**2009**). The Remarkable Frequency of Human Immunodeficiency Virus Type 1 Genetic Recombination. *Microbiology and Molecular Biology Reviews, 73*(3): 451-480.
58. **Peeters**, M., Jung, M., & Ayouba, A. (**2013**). The origin and molecular epidemiology of HIV. *Expert Review of Anti-Infective Therapy, 11*(9): 885-896.
59. **Peris**, G., & Marzal, A. (**2014**). Statistical Significance of Normalized Global Alignment. *Journal of Computational Biology, 21*(3): 257-268.
60. **Pond**, S. L. K., Frost, S. D. W., & Muse, S. V. (**2004**). HyPhy: hypothesis testing using phylogenies. *Bioinformatics, 21*(5): 676-679. doi:10.1093/bioinformatics/bti079.
61. **Poon**, A.F., Lewis, F.I., Frost, S.D. & Kosakovsky-Pond, S.L. (**2008**). Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. *Bioinformatics, 24*(17): 1949-1950. doi:10.1093/bioinformatics/btn313

62. **Porcar**, M., Latorre, A., & Moya, A. (**2013**). What Symbionts Teach us about Modularity. *Frontiers in Bioengineering and Biotechnology*, 1. doi:10.3389/fbioe.2013.00014
63. **Posada**, D. (**2008**). jModelTest: Phylogenetic Model Averaging. *Molecular Biology and Evolution, 25*(7): 1253-1256. doi:10.1093/molbev/msn083.
64. **Rambaut**, A., Posada, D., Crandall, K. A., & Holmes, E. C. (**2004**). The causes and consequences of HIV evolution. *Nature Reviews Genetics, 5*(1): 52-61.
65. **Reis**, M. N. da G., Bello, G., Guimarães, M. L., & Stefani, M. M. A. (**2017**). Characterization of HIV-1 CRF90_BF1 and putative novel CRFs_BF1 in Central West, North and Northeast Brazilian regions. *Plos One, 12*(6): e0178578.
66. **Rojas-Sánchez**, P., Cobos, A., Navaro, M., Ramos, J. T., Pagán, I., & Holguín, Á. (**2017**). Impact of Clinical Parameters in the Intrahost Evolution of HIV-1 Subtype B in Pediatric Patients: A Machine Learning Approach. *Genome Biology and Evolution, 9*(10): 2715-2726.
67. **Rosner**, B., & Glynn, R. J. (**2010**). Power and Sample Size Estimation for the Clustered Wilcoxon Test. *Biometrics, 67*(2): 646-653. doi:10.1111/j.1541-0420.2010.01488.x.
68. **Ross**, H. A., & Rodrigo, A. G. (**2002**). Immune-Mediated Positive Selection Drives Human Immunodeficiency Virus Type 1 Molecular Variation and Predicts Disease Duration. *Journal of Virology, 76*(22): 11715-11720. doi: 10.1128/jvi.76.22.11715-11720.2002.
69. **Sakuragi**, S., Yokoyama, M., Shioda, T., Sato, H., & Sakuragi, J. (**2016**). SL1 revisited: functional analysis of the structure and conformation of HIV-1 genome RNA. *Retrovirology, 13*(1): 1-13.
70. **Samuk**, K., Owens, G. L., Delmore, K. E., Miller, S. E., Rennison, D. J., & Schluter, D. (**2017**). Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Molecular Ecology, 26*(17): 4378-4390.
71. **Sanjuán**, R. (**2012**). From Molecular Genetics to Phylodynamics: Evolutionary Relevance of Mutation Rates Across Viruses. *PLoS Pathogens, 8*(5): e1002685. doi: 10.1371/journal.ppat.1002685.
72. **Sanjuán**, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (**2010**). Viral Mutation Rates. *Journal of Virology, 84*(19): 9733-9748. doi:10.1128/jvi.00694-10.
73. **Sardanyés**, J., Elena, S. F., & Solé, R. V. (**2008**). Simple quasispecies models for the survival-of-the-flattest effect: The role of space. *Journal of Theoretical Biology, 250*(3): 560-568.
74. **Seibert**, S. A., Howell, C. A., Hughes, M. K & and Hughes, A. L. (**1995**). Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (HIV-1). *Molecular Biology and Evolution, 12*(5): 803-813. doi:10.1093/oxfordjournals.molbev.a040257.
75. **Sharp**, P. M., & Hahn, B. H. (**2011**). Origins of HIV and the AIDS Pandemic. *Cold Spring Harbor Perspectives in Medicine, 1*(1): 1-23.
76. **Simon-Loriere**, E., & Holmes, E. C. (**2011**). Why do RNA viruses recombine? *Nature Reviews Microbiology, 9*(8): 617–626. doi: 10.1038/nrmicro2614.

77. **Som**, A. (**2014**). Causes, consequences and solutions of phylogenetic incongruence. *Briefings in Bioinformatics, 16*(3): 536–548. doi:10.1093/bib/bbu015.
78. **Souza**, J. S. M., Silva Júnior, J. J., Brites, C., & Monteiro-Cunha, J. P. (**2019**). Molecular and geographic characterization of hiv-1 bf recombinant viruses. *Virus Research, 197650*. doi:10.1016/j.virusres.2019.197650
79. **Subbarao**, S., & Schochetman, G. (**1996**). Genetic variability of HIV-1. *AIDS, 10*(Supplement): S13–24.
80. **Sztuba-Solińska**, J., Urbanowicz, A., Figlerowicz, M., & Bujarski, J. J. (**2011**). RNA-RNA Recombination in Plant Virus Replication and Evolution. *Annual Review of Phytopathology, 49*(1): 415-443.
81. **Tebit**, D. M., & Arts, E. J. (**2011**). Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *The Lancet Infectious Diseases, 11*(1): 45–56. doi:10.1016/s1473-3099(10)70186-9.
82. **Theys**, K., Libin, P., Pineda-Peña, A.-C., Nowé, A., Vandamme, A.-M., & Abecasis, A. B. (**2018**). The impact of HIV-1 within-host evolution on transmission dynamics. *Current Opinion in Virology, 28*: 92-101. doi: 10.1016/j.coviro.2017.12.001.
83. **Thompson**, J. R., Kamath, N., & Perry, K. L. (**2014**). An Evolutionary Analysis of the Secoviridae Family of Viruses. *PLoS ONE, 9*(9): e106305.
84. **Travers**, S. A. A., Tully, D. C., McCormack, G. P., & Fares, M. A. (**2007**). A Study of the Coevolutionary Patterns Operating within the env Gene of the HIV-1 Group M Subtypes. *Molecular Biology and Evolution, 24*(12): 2787-2801. doi:10.1093/molbev/msm213.
85. **UNAIDS**. (**2020**). Global HIV & AIDS statistics — 2020 fact sheet. January/12/2021, from UNAIDS Web site: https://www.unaids.org/es/resources/fact-sheet.
86. **Vega-Sanabria**, G. (**2016**). Science, stigmatization and afro-pessimism in the South African debate on AIDS. *Vibrant: Virtual Brazilian Anthropology, 13*(1): 22-51.
87. **Vitti**, J. J., Grossman, S. R., & Sabeti, P. C. (**2013**). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics, 47*(1): 97-120.
88. **Vuilleumier**, S., & Bonhoeffer, S. (**2015**). Contribution of recombination to the evolutionary history of HIV. *Current Opinion in HIV and AIDS, 10*(2): 84-89.
89. **Wagner**, G. P., & Altenberg, L. (**1996**). Perspective: Complex Adaptations and the Evolution of Evolvability. *Evolution, 50*(3): 967-976.
90. **Wang**, M., & Caetano-Anollés, G. (**2009**). The Evolutionary Mechanics of Domain Organization in Proteomes and the Rise of Modularity in the Protein World. *Structure, 17*(1): 66-78.
91. **Wang**, Y., Liang, Y., Feng, Y., Wang, B., Li, Y., Wu, Z., … Xia, X. (**2015**). HIV-1 prevalence and subtype/recombinant distribution among travelers entering China from Vietnam at the HeKou

port in the Yunnan province, China, between 2003 and 2012. *Journal of Medical Virology, 87*(9): 1500-1509.

92. **Watson**, J. D. (**2006**). Biología molecular del gen (México: Editorial Médica Panamericana).
93. **Weissman**, D. B., & Hallatschek, O. (**2014**). The Rate of Adaptation in Large Sexual Populations with Linear Chromosomes. *Genetics, 196*(4): 1167-1183.
94. **Williamson**, S. (**2003**). Adaptation in the env Gene of HIV-1 and Evolutionary Theories of Disease Progression. *Molecular Biology and Evolution, 20*(8): 1318-1325. doi: 10.1093/molbev/msg144.
95. **Yamaguchi-Kabata**, Y. & Gojobori, T. (**2000**). Reevaluation of Amino Acid Variability of the Human Immunodeficiency Virus Type 1 gp120 Envelope Glycoprotein and Prediction of New Discontinuous Epitopes. *Journal of Virology, 74*(9): 4335-4350.
96. **Yau**, C. (**2014**). R Tutorial with Bayesian Statistics Using OpenBUGS. 1st edition, California: 261-262.
97. **Zanotto**, P. M., Kallas, E. G., de Souza, R. F., & Holmes, E. C. (**1999**). Genealogical Evidence for Positive Selection in the Nef Gene of HIV-1. *Genetics, 153*(3): 1077-89.

12. ANEXOS

12.1. Congresos

Durante el desarrollo de mis estudios de maestría tuve la oportunidad de asistir al XI Congreso Nacional de Virología, el cual se llevó a cabo los días 4 a 7 de septiembre del 2019, en los Espacios Magnos de la Universidad de Guanajuato, Ciudad de Guanajuato, Guanajuato, México.

XI SIMPOSIO INTERNACIONAL DE BIOSEGURIDAD Y BIOCUSTODIA

XI CONGRESO NACIONAL DE VIROLOGÍA

BIOSEGURIDAD VIROLOGÍA

La Sociedad Mexicana de Virología y la Asociación Mexicana de Bioseguridad otorgan la presente

CONSTANCIA

a **JOSE MANUEL CABALLERO CONTRERAS**

por su valiosa asistencia como *Participante* en el

XI CONGRESO NACIONAL DE VIROLOGÍA

que tuvo lugar del 4 al 7 de septiembre de 2019 con duración de 26.5 horas
en los Espacios Magnos de la Universidad de Guanajuato.
Ciudad de Guanajuato, Guanajuato, México, a 7 de septiembre de 2019.

Dr. Ramón González García Conde
Presidente
Sociedad Mexicana de Virología AC

QFB Lissete Valenzuela Fabris
Presidente
Asociación Mexicana de Bioseguridad AC

Folio: XICNV90

En este congreso tuve la oportunidad de presentar los inicios de mi proyecto de investigación en modalidad de cartel, para el cual utilizamos cuatro conjuntos de datos de secuencias de genomas de VIH-1 de formas recombinantes (**CRF-02_AG**, **CRF-07_BC**, **CRF-35_AD** y **CRF-42_BF**) y las secuencias de los genomas de subtipos puros que las conforman; realizamos la detección de recombinación presente en cada uno de

los genomas de las formas recombinantes e identificamos la diversidad presente en los módulos, tanto en las CRFs como en los subtipos puros que las constituyen.

XI SIMPOSIO INTERNACIONAL DE BIOSEGURIDAD Y BIOCUSTODIA

XI CONGRESO NACIONAL DE VIROLOGÍA

BIOSEGURIDAD VIROLOGÍA

La Sociedad Mexicana de Virología y la Asociación Mexicana de Bioseguridad
otorgan la presente

CONSTANCIA

a *CABALLERO CONTRERAS, JOSE MANUEL and Zarate, Selene.*
por su valiosa exposición del trabajo libre en CARTEL intitulado
Effects of Modularity on the Diversity and Evolution of Human Immunodeficiency Virus 1
presentado en el **XI CONGRESO NACIONAL DE VIROLOGÍA**
que tuvo lugar del 4 al 7 de septiembre de 2019 en los Espacios Magnos de la Universidad de Guanajuato.
Ciudad de Guanajuato, Guanajuato, México, a 7 de septiembre de 2019.

Dr. Ramón González García Conde
Presidente
Sociedad Mexicana de Virología AC
Folio: XICNV-C118

QFB Lissete Valenzuela Fabris
Presidenta
Asociación Mexicana de Bioseguridad AC

SOCIEDAD MEXICANA DE VIROLOGÍA

12.2. Glosario

- **Adaptación:** Cualquier carácter, o conjunto de caracteres de un organismo, que afecta positivamente a su adecuación biológica.
- **Adecuación Biológica:** Describe como un organismo puede sobrevivir, crecer, funcionar y replicarse en un entorno determinado de una generación con respecto a la generación anterior (Kosakovsky-Pond, 2016).
- **Capacitancia evolutiva:** Se utiliza para describir el escondite y la liberación de la variación genética críptica (Masel & Trotter, 2010).

- **Catástrofe de error:** La pérdida de información genética significativa cuando una población es empujada más allá de su tasa de mutación máxima (Lauring *et al.*, 2013).
- **Deriva Génica:** Es una fuerza evolutiva que actúa junto con la selección natural modificando las características de las especies, en el tiempo, mediante el cambio aleatorio de la frecuencia de alelos de una generación a la siguiente (Dolan *et al.*, 2016).
- **Evolución Molecular:** Cambio en la frecuencia genética de una población a través de generaciones.
- **Evolucionabilidad:** Capacidad de un organismo con un genotipo particular para ganar adecuación biológica con el tiempo después de evolucionar en un entorno determinado. (Lauring *et al.*, 2013).
- **Heterotaquia:** Se refiere al hecho de que la tasa evolutiva de una posición dada de nucleótidos/aminoácidos varía a lo largo del tiempo (Som, 2014).
- **Inconsistencia filogenética:** Se dice que dos (o más) árboles filogenéticos son incongruentes cuando exhiben órdenes de ramificación en conflicto (es decir, topologías) y no se pueden superponer. Esto implica que al menos un nodo (también conocido como bipartición) presente en un árbol no se encuentra en el otro(s), donde es reemplazado por agrupaciones alternativas de taxones (Som, 2014).
- **Modularidad:** Se refiere a un conjunto de características (genes) que interaccionan entre sí en unidades discretas (**módulos**), y éstos a su vez interaccionan entre ellos en un proceso determinado (Bonner, 1988).
- **Módulo:** Gen o fragmento de éste que conforma una CRF.
- **Pleiotropía:** Tendencia en la que un genotipo afecta a múltiples fenotipos (Vitti *et al.*, 2013).
- **Robustez Mutacional:** Es la medida en la que la adecuación biológica de un organismo permanece constante a pesar de que ocurran mutaciones en su genotipo (Fares, 2015).

- **Selección Natural negativa:** Eliminación de alelos nocivos de una población por selección natural. También es llamada "*selección purificadora*" (Lauring *et al.*, 2013).
- **Selección Natural Positiva:** Selección que actúa sobre nuevas mutaciones ventajosas (Nielsen, 2005).
- **Selección Natural:** Es el proceso evolutivo que explica el emparejamiento o adecuación, entre las características de organismos y los ambientes donde viven (Lenski, 2017).
- **Sustitución No Sinónima (N):** Un cambio en la región de un gen que codifica una proteína que altera el aminoácido codificado (Vitti *et al.*, 2013).
- **Sustitución Sinónima (S):** Un cambio en la región de un gen que codifica una proteína que no cambia el aminoácido codificado (Vitti *et al.*, 2013).
- **Transmisión en bloque:** Es la transmisión de múltiples partículas virales como una sola unidad. Puede ocurrir a través de la asociación con cuerpos multivesiculares (MVB), sinapsis celulares o virológicas, unión de bacterias entéricas y/o agregación de viriones (Dolan *et al.*, 2016).
- **Variación genética críptica:** Se define como la variación genética permanente que no contribuye al rango normal de fenotipos observados en una población, pero que está disponible para modificar un fenotipo que surge después del cambio ambiental o la introducción de alelos novedosos (Gibson & Dworkin, 2004).

12.3. Código

Script 1. Generación de alineamientos de los genes principales del VIH-1.

```
##Cargar alineamiento en formato fasta##
from Bio import AlignIO
Alignment=AlignIO.read("Alignment.fasta", "fasta")
##Crear una matriz y exportarla como archivo fasta##
Gene = Alignment[:,828:2409]
from Bio import SeqIO
```

```
SeqIO.write(Gene, "Gag_gene.fasta", "fasta")
```

Script 2. Generación de gráfica de diversidad del genoma completo.

```
#Cargar archivo separado por comas

data                    <-                    read.csv(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Recombinant/06_cpx(A1,G,J,K)/Re
sults.csv", header = TRUE)

#Generar las variables para la gráfica

x <- as.integer(data$Position)

y <- as.integer(data$Score)

#Generar una figura XY y la exporta en formato PNG

png("Diversity_CRF-06_cpx(A1,G,J,K).png")

plot(x, y, main="CRF-06_cpx(A1,G,J,K)", col= "gray", xlab = "Genome Position",
ylab="Polymorphism Diversity")

abline(85, 0, col = "green")

dev.off()
```

Script 3. Generación de diagramas de Venn.

```
#Paquetes a cargar

library(dplyr)

library(futile.logger)

library(grid)

library(VennDiagram)

library(ggplot2)

#Cargar archivos con datos

data1                   <-                    read.csv(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Recombinant/42_BF/Results.csv",
header = TRUE)

data2                   <-                    read.csv(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/B_Subtype/Results.csv",
header = TRUE)

data3                   <-                    read.csv(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/F1_Subtype/Results.csv",
header = TRUE)

#Filtros para identificar sitios polimórficos

x <- filter(data1, data1$Score >= 85)
```

```
y <- filter(data2, data2$Score >= 85)

z <- filter(data3, data3$Score >= 85)

#Crear un archivo con extensión ".txt"

write.table(x, file = "Polymorphic_Sites_Genome_Completed_CRF-42_BF.txt", append =
F, quote = F, col.names = T, row.names = T)

write.table(y, file = "Polymorphic_Sites_Genome_Completed_B_Subtype.txt", append =
F, quote = F, col.names = T, row.names = T)

write.table(z, file = "Polymorphic_Sites_Genome_Completed_F1_Subtype.txt", append =
F, quote = F, col.names = T, row.names = T)

# ------------------------------------------------------------------------------
---------------------------------------------- #

#Crear diagramas de Venn

#Datos

Recombinant                                                                    <-
read.table("C:/Users/josec/Documents/Thesis/Code/Scripts/Diversity/Polymorphic_Sit
es_Genome_Completed_CRF-42_BF.txt", header = T)

First_Pure                                                                     <-
read.table("C:/Users/josec/Documents/Thesis/Code/Scripts/Diversity/Polymorphic_Sit
es_Genome_Completed_B_Subtype.txt", header = T)

Second_Pure                                                                    <-
read.table("C:/Users/josec/Documents/Thesis/Code/Scripts/Diversity/Polymorphic_Sit
es_Genome_Completed_F1_Subtype.txt", header = T)

#Variables

Area_1 <- length(Recombinant$Position)

Area_2 <- length(First_Pure$Position)

Area_3 <- length(Second_Pure$Position)

Int_12 <- length(intersect(Recombinant$Position, First_Pure$Position))

Int_23 <- length(intersect(First_Pure$Position, Second_Pure$Position))

Int_13 <- length(intersect(Recombinant$Position, Second_Pure$Position))

Int_123 <- length(intersect(intersect(Recombinant$Position, First_Pure$Position),
Second_Pure$Position))

#Realizar diagrama de Venn

plot <- draw.triple.venn(area1 = Area_1,

                         area2 = Area_2,

                         area3 = Area_3,

                         n12 = Int_12,

                         n23 = Int_23,

                         n13 = Int_13,
```

```
                        n123 = Int_123,
                        col=c("#440154ff", '#21908dff', 'Yellow'),
                        fill  =  c(alpha("#440154ff",0.5),  alpha('#21908dff',0.5),
alpha('Yellow',0.5)),
                        cex = 1,
                        fontfamily = "sans",
                        cat.cex = 1,
                        cat.default.pos = "outer",
                        cat.fontfamily = "sans",
                        cat.col = c("#440154ff", '#21908dff', 'Yellow'),
                        category = c("CRF-42_BF", "B_Subtype", "F1_Subtype"))
ggsave("Polymorphic_Sites.png", plot = plot, dpi = 300)
```

Script 4. Prueba de Kruskal-Wallis.

```
#Paquete a instalar
library("dplyr")
#Cargar archivo separado por comas
data1                       <-                      read.csv(file                       =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Recombinant/35_A1D/Results.csv"
, header = TRUE)
data2                       <-                      read.csv(file                       =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/A1_Subtype/Results.csv",
header = TRUE)
data3                       <-                      read.csv(file                       =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/D_Subtype/Results.csv",
header = TRUE)
#Pregunta al usuario por los módulos de su CRF
Input = readline("How many modules do your CRF have? ")
Input = as.numeric(Input)
#Ciclo
for(i in Input){
  sum = 1
  input = 0
  while (sum <= Input) {
    x = input + 1
    input = readline("What´s the last position of the current module? ")
```

```
    input = as.numeric(input)
    #Filtros para el módulo
    x1 <- data1[x:input, ]
    y1 <- filter(x1, Score >1)
    z1 <- y1$Score
    x2 <- data2[x:input, ]
    y2 <- filter(x2, Score >1)
    z2 <- y2$Score
    x3 <- data3[x:input, ]
    y3 <- filter(x3, Score >1)
    z3 <- y3$Score
    #Variables para el análisis Kruskal-Wallis
    a <- replicate(length(z1), "CRF")
    b <- replicate(length(z2), "First_Subtype")
    c <- replicate(length(z3), "Second_Subtype")
    w <- c(z1,z2,z3)
    p <- c(a,b,c)
    #Prueba Kruskal-Wallis
    o <- kruskal.test(w,p)
    #Crear archivo con extensión ".txt"
    trial <- matrix(c(o$method,o$data.name,o$statistic,o$parameter,o$p.value), ncol
= 1)
    rownames(trial) <- c('Method:', 'Data:', 'Chi-squared:', "Parameter:", "P-
value:")
    g <- trial.table <- as.table(trial)
    name <- "Kruskal-Wallis_Test_Module_"
    name_2 <- paste(name, sum, sep = "")
    name_3 <- paste(name_2, ".txt", sep = "")
    write.table(g, file = name_3, append = F, quote = F, col.names = F)
    sum = sum + 1
  }
}
```

Script 5. Prueba de Wilcoxon.

```
#Paquete a instalar

library("dplyr")

#Cargar archivo separado por comas

data1                    <-                    read.csv(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Recombinant/35_A1D/Results.csv"
, header = TRUE)

data2                    <-                    read.csv(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/A1_Subtype/Results.csv",
header = TRUE)

data3                    <-                    read.csv(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/D_Subtype/Results.csv",
header = TRUE)

#Pregunta al usuario por los módulos de su CRF

Input = readline("How many modules do your CRF have? ")

Input = as.numeric(Input)

#Ciclo

for(i in Input){

  sum = 1

  input = 0

  while (sum <= Input) {

    x = input + 1

    input = readline("What´s the last position of the current module? ")

    input = as.numeric(input)

    #Filtros para el módulo

    x1 <- data1[x:input, ]

    y1 <- filter(x1, Score >1)

    z1 <- y1$Score

    x2 <- data2[x:input, ]

    y2 <- filter(x2, Score >1)

    z2 <- y2$Score

    x3 <- data3[x:input, ]

    y3 <- filter(x3, Score >1)

    z3 <- y3$Score

    #Variables para el análisis Wilcoxon

    w <- z1
```

```
    p <- readline("With what pure genotype do we compare the current module? (First
Pure = 0 or Second Pure = 1) ")
    if (p==0)
    {
      e = z2
    }
    else{
      e = z3
    }
    #Prueba de Wilcoxon
    o <- wilcox.test(w,e)
    #Crear archivo con extensión ".txt"
    trial <- matrix(c(o$method,o$statistic,o$p.value), ncol = 1)
    rownames(trial) <- c('Method:', 'W:', "P-value:")
    g <- trial.table <- as.table(trial)
    name <- "Wilcoxon_Test_Module_"
    name_2 <- paste(name, sum, sep = "")
    name_3 <- paste(name_2, ".txt", sep = "")
    write.table(g, file = name_3, append = F, quote = F, col.names = F)
    sum = sum + 1
  }
}
```

Script 6. Generación de modelos evolutivos en jModelTest.

```
java –jar /mnt/c/Users/josec/Documents/Thesis/Code/Phylogeny/jModelTest/jmodeltest-
2.1.10/jModelTest.jar -d Alignment.fasta -g 4 -i -f -AIC -BIC –a
```

Script 7. Generación de filogenias bayesianas.

```
lset nst=6 rates=invgamma
prset
mcmc ngen=1000000 samplefreq=1000 printfreq=1000 diagnfreq=1000 burnin=100
relburnin=YES nchains=4 temp=0.5 stoprule=YES
sump
```

```
sumt
```

Script 8. Generación de matriz para reconstrucción de estados ancestrales.

```
##Leer un alineamiento en formato fasta##

from Bio import AlignIO

alignment=AlignIO.read("Alignment.fasta", "fasta")

##Crear una matriz y exportarla como archivo fasta##

SNP = alignment[:,378:379]

from Bio import SeqIO

SeqIO.write(SNP, "Polymorphism_379_Module_2.fasta", "fasta")
```

Script 9. Identificación de sitios bajo selección en los archivos separados por coma.

```
#Paquete a instalar

library("dplyr")

#Cargar archivo separado por comas

data                    <-                    read.csv(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Recombinant/02_A1G/Results/Data
Monkey/Selection/FEL/Env_Gene/datamonkey-table.csv", header = TRUE)

#Filtrado de valores

p <- filter(data, data$omega > 1 & data$LRT >= 2.698) #Posiciones bajo selección
positiva

n <- filter(data, data$omega < 1 & data$LRT >= 2.698) #Posiciones bajo selección
negativa

#Crear archivos con extensión ".txt"

p_2 <- p$Site

write.table(p_2, "Positive_Selection_Sites.txt", append = F, quote = F, col.names =
F)

n_2 <- n$Site

write.table(n_2, "Negative_Selection_Sites.txt", append = F, quote = F, col.names =
F)
```

Script 10. Identificación de sitios bajo selección en módulos.

```
#Paquete a instalar

library("dplyr")
```

```
#Datos

vector                    <-                    read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Recombinant/14_BG/Results/DataM
onkey/Selection/FEL/Gag_Gene/Negative_Selection_Sites.txt", header = F)

vector_2                    <-                    read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/B_Subtype/Results/DataMon
key/Selection/FEL/Gag_Gene/Negative_Selection_Sites.txt", header = F)

vector_3                    <-                    read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/G_Subtype/Results/DataMon
key/Selection/FEL/Gag_Gene/Negative_Selection_Sites.txt", header = F)

vector_P                    <-                    read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Recombinant/14_BG/Results/DataM
onkey/Selection/FEL/Pol_Gene/Negative_Selection_Sites.txt", header = F)

vector_2_P                    <-                    read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/B_Subtype/Results/DataMon
key/Selection/FEL/Pol_Gene/Negative_Selection_Sites.txt", header = F)

vector_3_P                    <-                    read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/G_Subtype/Results/DataMon
key/Selection/FEL/Pol_Gene/Negative_Selection_Sites.txt", header = F)

vector_E                    <-                    read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Recombinant/14_BG/Results/DataM
onkey/Selection/FEL/Env_Gene/Negative_Selection_Sites.txt", header = F)

vector_2_E                    <-                    read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/B_Subtype/Results/DataMon
key/Selection/FEL/Env_Gene/Negative_Selection_Sites.txt", header = F)

vector_3_E                    <-                    read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/G_Subtype/Results/DataMon
key/Selection/FEL/Env_Gene/Negative_Selection_Sites.txt", header = F)

#Gen gag

Module_1 <- filter(vector, vector$V2 <= 497)

Module_1.2 <- filter(vector_2, vector_2$V2 <= 497)

Module_1.3 <- filter(vector_3, vector_3$V2 <= 497)

#Gen pol

Module_1_P <- filter(vector_P, vector_P$V2 <= 835)

Module_1.2_P <- filter(vector_2_P, vector_2_P$V2 <= 835)

Module_1.3_P <- filter(vector_3_P, vector_3_P$V2 <= 835)

Module_2_P <- filter(vector_P, vector_P$V2 >= 836)

Module_2.2_P <- filter(vector_2_P, vector_2_P$V2 >= 836)

Module_2.3_P <- filter(vector_3_P, vector_3_P$V2 >= 836)

#Gen env

Module_4_E <- filter(vector_E, vector_E$V2 <= 248)
```

```
Module_4.2_E <- filter(vector_2_E, vector_2_E$V2 <= 248)

Module_4.3_E <- filter(vector_3_E, vector_3_E$V2 <= 248)

Module_5_E <- filter(vector_E, vector_E$V2 >= 249)

Module_5.2_E <- filter(vector_2_E, vector_2_E$V2 >= 249)

Module_5.3_E <- filter(vector_3_E, vector_3_E$V2 >= 249)

#Crear archivo con extensión “.txt”

#Gen gag

write.table(Module_1$V2,       "Negative_Selection_Sites_FEL_Module_1_Gag_Gene_CRF-
14_BG.txt", append = F, quote = F, col.names = F, row.names = T)

write.table(Module_1.2$V2,
"Negative_Selection_Sites_FEL_Module_1_Gag_Gene_B_Subtype.txt", append = F, quote =
F, col.names = F, row.names = T)

write.table(Module_1.3$V2,
"Negative_Selection_Sites_FEL_Module_1_Gag_Gene_G_Subtype.txt", append = F, quote =
F, col.names = F, row.names = T)

#Gen pol

write.table(Module_1_P$V2,     "Negative_Selection_Sites_FEL_Module_1_Pol_Gene_CRF-
14_BG.txt", append = F, quote = F, col.names = F, row.names = T)

write.table(Module_1.2_P$V2,
"Negative_Selection_Sites_FEL_Module_1_Pol_Gene_B_Subtype.txt", append = F, quote =
F, col.names = F, row.names = T)

write.table(Module_1.3_P$V2,
"Negative_Selection_Sites_FEL_Module_1_Pol_Gene_G_Subtype.txt", append = F, quote =
F, col.names = F, row.names = T)

write.table(Module_2_P$V2,     "Negative_Selection_Sites_FEL_Module_2_Pol_Gene_CRF-
14_BG.txt", append = F, quote = F, col.names = F, row.names = T)

write.table(Module_2.2_P$V2,
"Negative_Selection_Sites_FEL_Module_2_Pol_Gene_B_Subtype.txt", append = F, quote =
F, col.names = F, row.names = T)

write.table(Module_2.3_P$V2,
"Negative_Selection_Sites_FEL_Module_2_Pol_Gene_G_Subtype.txt", append = F, quote =
F, col.names = F, row.names = T)

#Gen env

write.table(Module_4_E$V2,     "Negative_Selection_Sites_FEL_Module_4_Env_Gene_CRF-
14_BG.txt", append = F, quote = F, col.names = F, row.names = T)

write.table(Module_4.2_E$V2,
"Negative_Selection_Sites_FEL_Module_4_Env_Gene_B_Subtype.txt", append = F, quote =
F, col.names = F, row.names = T)

write.table(Module_4.3_E$V2,
"Negative_Selection_Sites_FEL_Module_4_Env_Gene_G_Subtype.txt", append = F, quote =
F, col.names = F, row.names = T)

write.table(Module_5_E$V2,     "Negative_Selection_Sites_FEL_Module_5_Env_Gene_CRF-
14_BG.txt", append = F, quote = F, col.names = F, row.names = T)
```

```
write.table(Module_5.2_E$V2,
"Negative_Selection_Sites_FEL_Module_5_Env_Gene_B_Subtype.txt", append = F, quote =
F, col.names = F, row.names = T)

write.table(Module_5.3_E$V2,
"Negative_Selection_Sites_FEL_Module_5_Env_Gene_G_Subtype.txt", append = F, quote =
F, col.names = F, row.names = T)
```

Script 11. Identificación de sitios bajo selección en los genes.

```
#Paquete a instalar

library("dplyr")

#Cargar archivos con datos

vector                    <-                    read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Recombinant/02_A1G/Results/Data
Monkey/Selection/FEL/Gag_Gene/Negative_Selection_Sites.txt", header = F)

vector_2                   <-                   read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/A1_Subtype/Results/DataMo
nkey/Selection/FEL/Gag_Gene/Negative_Selection_Sites.txt", header = F)

vector_3                   <-                   read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/G_Subtype/Results/DataMon
key/Selection/FEL/Gag_Gene/Negative_Selection_Sites.txt", header = F)

vector_P                   <-                   read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Recombinant/02_A1G/Results/Data
Monkey/Selection/FEL/Pol_Gene/Negative_Selection_Sites.txt", header = F)

vector_2_P                  <-                   read.table(file                   =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/A1_Subtype/Results/DataMo
nkey/Selection/FEL/Pol_Gene/Negative_Selection_Sites.txt", header = F)

vector_3_P                  <-                   read.table(file                   =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/G_Subtype/Results/DataMon
key/Selection/FEL/Pol_Gene/Negative_Selection_Sites.txt", header = F)

vector_E                   <-                   read.table(file                    =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Recombinant/02_A1G/Results/Data
Monkey/Selection/FEL/Env_Gene/Negative_Selection_Sites.txt", header = F)

vector_2_E                  <-                   read.table(file                   =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/A1_Subtype/Results/DataMo
nkey/Selection/FEL/Env_Gene/Negative_Selection_Sites.txt", header = F)

vector_3_E                  <-                   read.table(file                   =
"C:/Users/josec/Documents/Thesis/Data/Interpatient/Pures/G_Subtype/Results/DataMon
key/Selection/FEL/Env_Gene/Negative_Selection_Sites.txt", header = F)

#Variables para asignar las coordenadas de cada proteína dentro de los genes

#Gen gag

MA <- filter(vector, vector$V2 <= 141)

MA_2 <- filter(vector_2, vector_2$V2 <= 141)
```

```
MA_3 <- filter(vector_3, vector_3$V2 <= 141)

CA <- filter(vector, vector$V2 >= 142 & vector$V2 <= 394)

CA_2 <- filter(vector_2, vector_2$V2 >= 142 & vector_2$V2 <= 394)

CA_3 <- filter(vector_3, vector_3$V2 >= 142 & vector_3$V2 <= 394)

NC <- filter(vector, vector$V2 >= 409 & vector$V2 <= 466)

NC_2 <- filter(vector_2, vector_2$V2 >= 409 & vector_2$V2 <= 466)

NC_3 <- filter(vector_3, vector_3$V2 >= 409 & vector_3$V2 <= 466)

#Gen pol

PR <- filter(vector_P, vector_P$V2 >= 74 & vector_P$V2 <= 189)

PR_2 <- filter(vector_2_P, vector_2_P$V2 >= 74 & vector_2_P$V2 <= 189)

PR_3 <- filter(vector_3_P, vector_3_P$V2 >= 74 & vector_3_P$V2 <= 189)

RT <- filter(vector_P, vector_P$V2 >= 190 & vector_P$V2 <= 810)

RT_2 <- filter(vector_2_P, vector_2_P$V2 >= 190 & vector_2_P$V2 <= 810)

RT_3 <- filter(vector_3_P, vector_3_P$V2 >= 190 & vector_3_P$V2 <= 810)

IN <- filter(vector_P, vector_P$V2 >= 851 & vector_P$V2 <= 1020)

IN_2 <- filter(vector_2_P, vector_2_P$V2 >= 851 & vector_2_P$V2 <= 1020)

IN_3 <- filter(vector_3_P, vector_3_P$V2 >= 851 & vector_3_P$V2 <= 1020)

#Gen env

gp120 <- filter(vector_E, vector_E$V2 >= 54 & vector_E$V2 <= 555)

gp120_2 <- filter(vector_2_E, vector_2_E$V2 >= 54 & vector_2_E$V2 <= 555)

gp120_3 <- filter(vector_3_E, vector_3_E$V2 >= 54 & vector_3_E$V2 <= 555)

gp41 <- filter(vector_E, vector_E$V2 >= 632 & vector_E$V2 <= 741)

gp41_2 <- filter(vector_2_E, vector_2_E$V2 >= 632 & vector_2_E$V2 <= 741)

gp41_3 <- filter(vector_3_E, vector_3_E$V2 >= 632 & vector_3_E$V2 <= 741)

#Crear archivo con extensión “.txt”

#Gen gag

write.table(MA$V2, "Negative_Selection_Sites_FEL_MA_CRF-02_A1G.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(MA_2$V2, "Negative_Selection_Sites_FEL_MA_A1_Subtype.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(MA_3$V2, "Negative_Selection_Sites_FEL_MA_G_Subtype.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(CA$V2, "Negative_Selection_Sites_FEL_CA_CRF-02_A1G.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(CA_2$V2, "Negative_Selection_Sites_FEL_CA_A1_Subtype.txt", append = F,
quote = F, col.names = F, row.names = T)
```

```
write.table(CA_3$V2, "Negative_Selection_Sites_FEL_CA_G_Subtype.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(NC$V2, "Negative_Selection_Sites_FEL_NC_CRF-02_A1G.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(NC_2$V2, "Negative_Selection_Sites_FEL_NC_A1_Subtype.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(NC_3$V2, "Negative_Selection_Sites_FEL_NC_G_Subtype.txt", append = F,
quote = F, col.names = F, row.names = T)

#Gen pol

write.table(PR$V2, "Negative_Selection_Sites_FEL_PR_CRF-02_A1G.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(PR_2$V2, "Negative_Selection_Sites_FEL_PR_A1_Subtype.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(PR_3$V2, "Negative_Selection_Sites_FEL_PR_G_Subtype.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(RT$V2, "Negative_Selection_Sites_FEL_RT_CRF-02_A1G.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(RT_2$V2, "Negative_Selection_Sites_FEL_RT_A1_Subtype.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(RT_3$V2, "Negative_Selection_Sites_FEL_RT_G_Subtype.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(IN$V2, "Negative_Selection_Sites_FEL_IN_CRF-02_A1G.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(IN_2$V2, "Negative_Selection_Sites_FEL_IN_A1_Subtype.txt", append = F,
quote = F, col.names = F, row.names = T)

write.table(IN_3$V2, "Negative_Selection_Sites_FEL_IN_G_Subtype.txt", append = F,
quote = F, col.names = F, row.names = T)

#Gen env

write.table(gp120$V2, "Negative_Selection_Sites_FEL_gp120_CRF-02_A1G.txt", append =
F, quote = F, col.names = F, row.names = T)

write.table(gp120_2$V2, "Negative_Selection_Sites_FEL_gp120_A1_Subtype.txt", append
= F, quote = F, col.names = F, row.names = T)

write.table(gp120_3$V2, "Negative_Selection_Sites_FEL_gp120_G_Subtype.txt", append
= F, quote = F, col.names = F, row.names = T)

write.table(gp41$V2, "Negative_Selection_Sites_FEL_gp41_CRF-02_A1G.txt", append =
F, quote = F, col.names = F, row.names = T)

write.table(gp41_2$V2, "Negative_Selection_Sites_FEL_gp41_A1_Subtype.txt", append =
F, quote = F, col.names = F, row.names = T)

write.table(gp41_3$V2, "Negative_Selection_Sites_FEL_gp41_G_Subtype.txt", append =
F, quote = F, col.names = F, row.names = T)
```

Script 12. Concatenación de alineamientos de genes.

```
##Script para concatenar alineamientos##
#Contador#
contador_global=0
contador_impar=$(($contador_global+1))
contador_par=$(($contador_global+2))
sed -n 1,40p gag_gene_without_stop_codons.translated.fas > CRF-02_A1G_Genome.fasta
#Ciclo#
for alignment in {1..20};
do
    contador_impar=$(($contador_global+1))
    contador_par=$(($contador_global+2))
    pol_gene=$(sed -n "$contador_par"p pol_gene_without_stop_codons.translated.fas)
    sed -i ""$contador_par"s/$/"$pol_gene"/" CRF-02_A1G_Genome.fasta
    env_gene=$(sed -n "$contador_par"p env_gene_without_stop_codons.translated.fas)
    sed -i ""$contador_par"s/$/"$env_gene"/" CRF-02_A1G_Genome.fasta
    contador_global=$(($contador_global+2))
done
```

Script 13. Generación de los análisis de coevolución.

```
Baseline substitution model: 9 (HIVBm, Specialist empirical model of protein evolution for between-host HIV sequences)
number of MCMC steps to sample: 100000
number of MCMC steps to discard as burn-in: 10000
number of steps to extract from the chain sample: 100
the maximum number of parents allowed per node: 1
the minimum number of substitutions per site to include it in the analysis: 1
```

12.4. Figuras adicionales

12.4.1. Módulos identificados en las CRFs

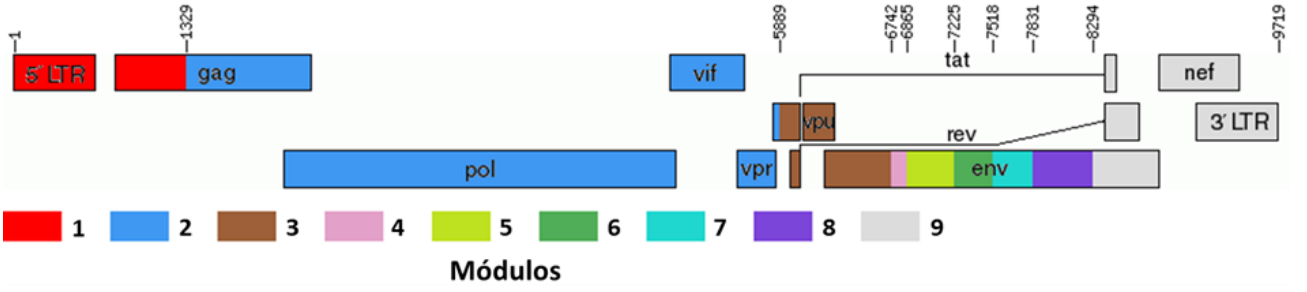

Figura 36. Módulos de la **CRF-06_cpx(A1,G,J,K)**.

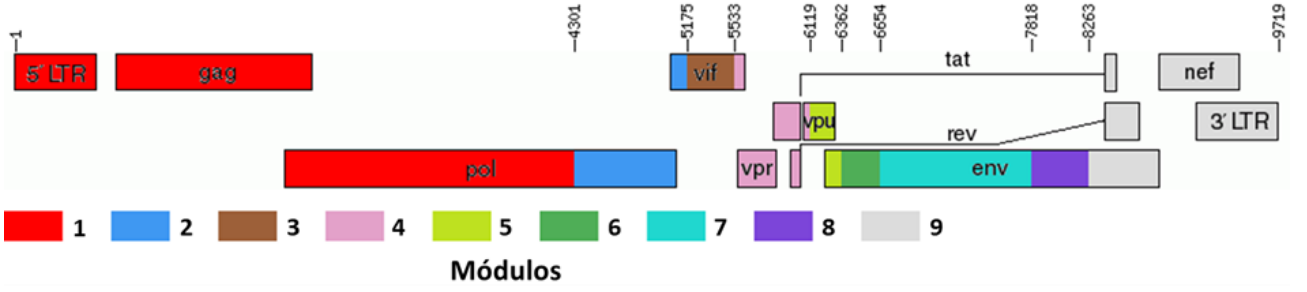

Figura 37. Módulos de la **CRF-07_BC**.

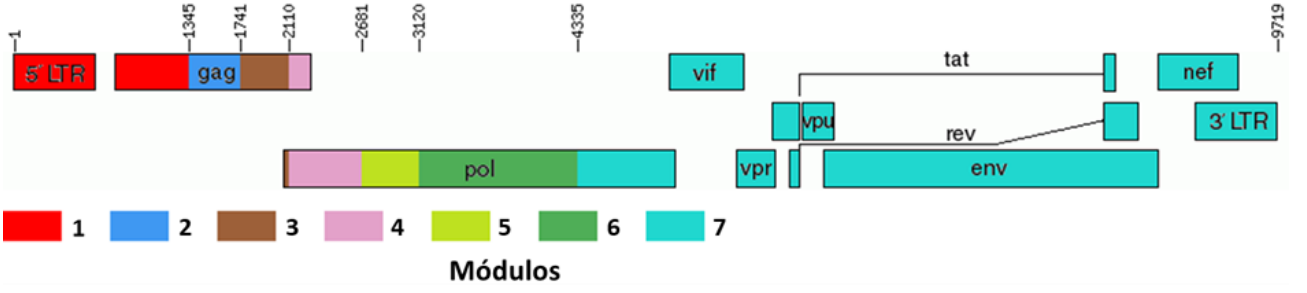

Figura 38. Módulos de la **CRF-11_cpx(A1,E,G,J,U)**.

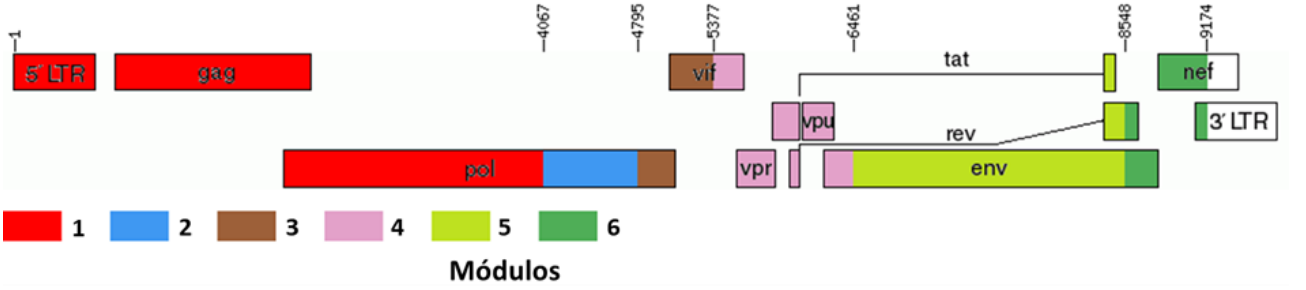

Figura 39. Módulos de la **CRF-14_BG**.

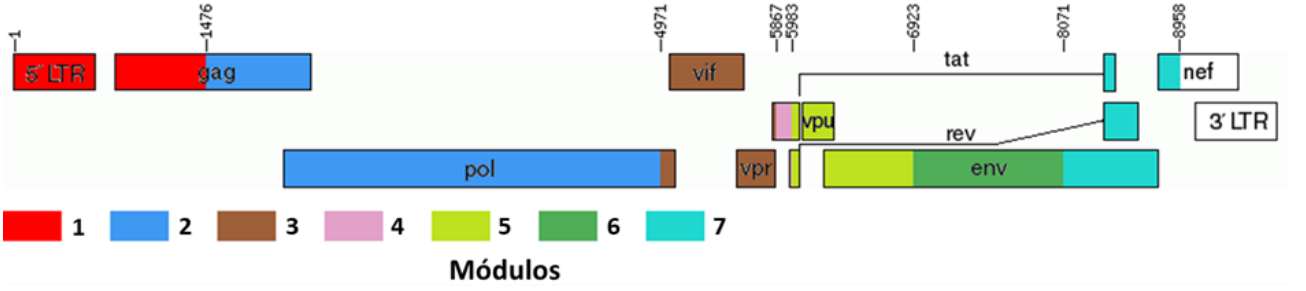

Figura 40. Módulos de la **CRF-35_A1D**.

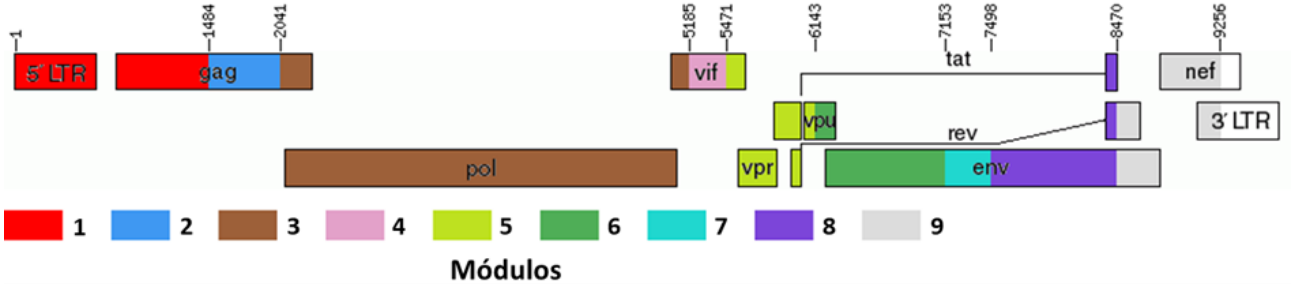

Figura 41. Módulos de la **CRF-42_BF**.

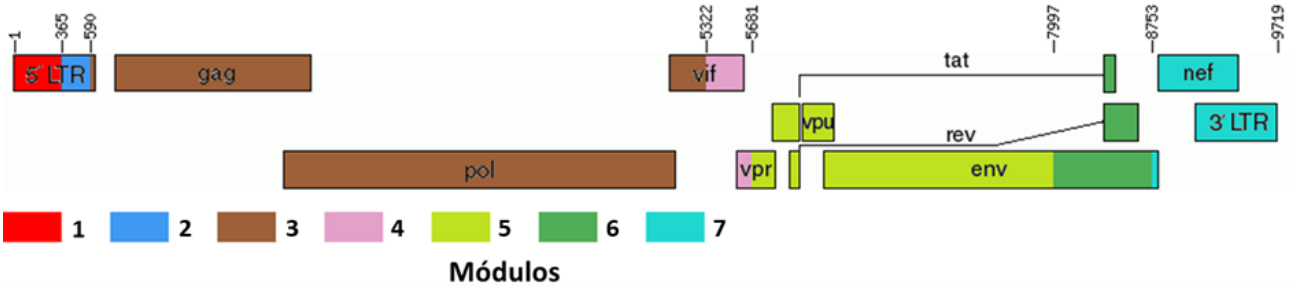

Figura 42. Módulos de la **CRF-63_02A6**.

12.4.2. Diversidad del genoma completo de las CRFs y de los subtipos puros que las conforman

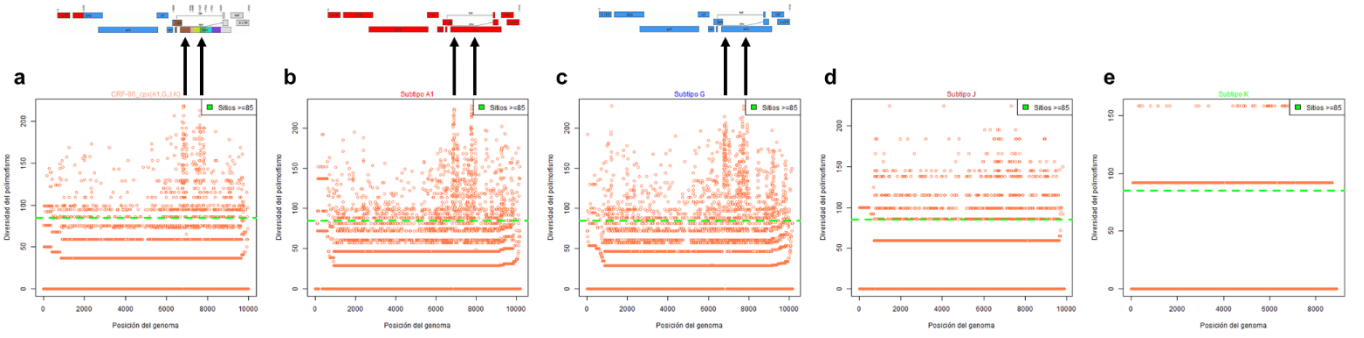

Figura 43. Diversidad del genoma completo de la **CRF-06_cpx(A1,G,J,K)** y de los subtipos puros que la constituyen.

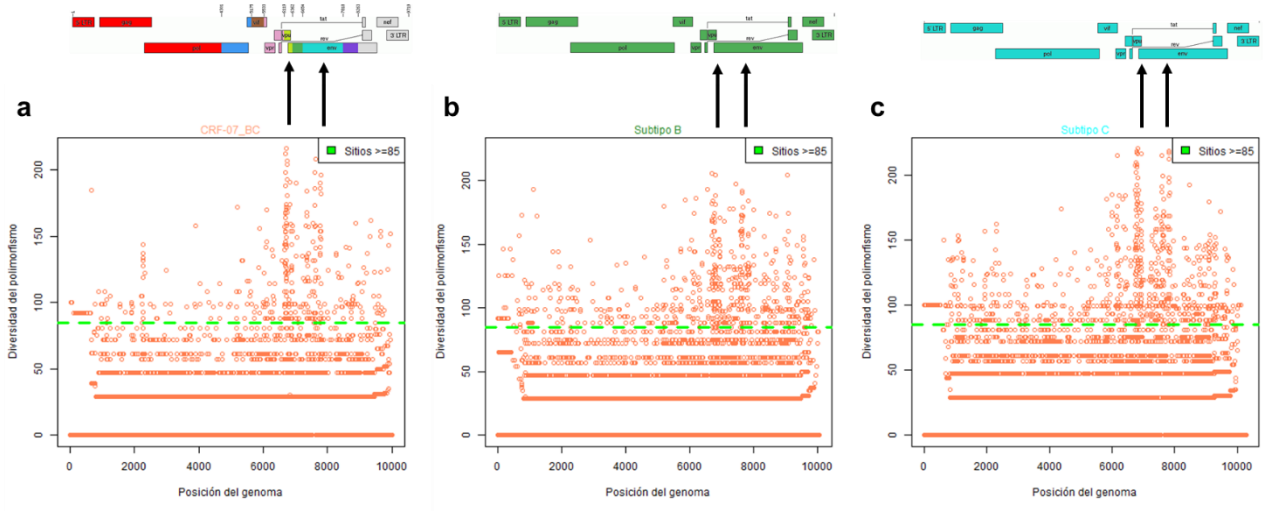

Figura 44. Diversidad del genoma completo de la **CRF-07_BC** y de los subtipos puros que la constituyen.

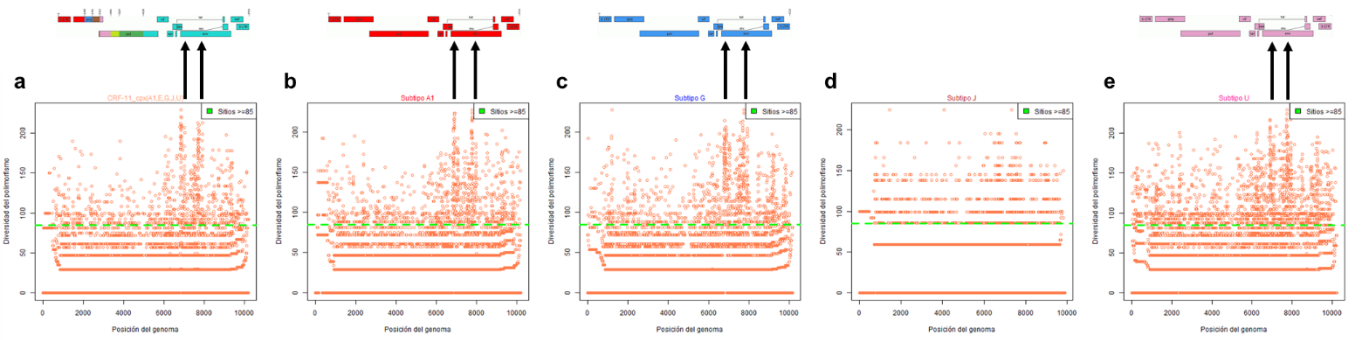

Figura 45. Diversidad del genoma completo de la **CRF-11_cpx(A1,E,G,J,U)** y de los subtipos puros que la constituyen.

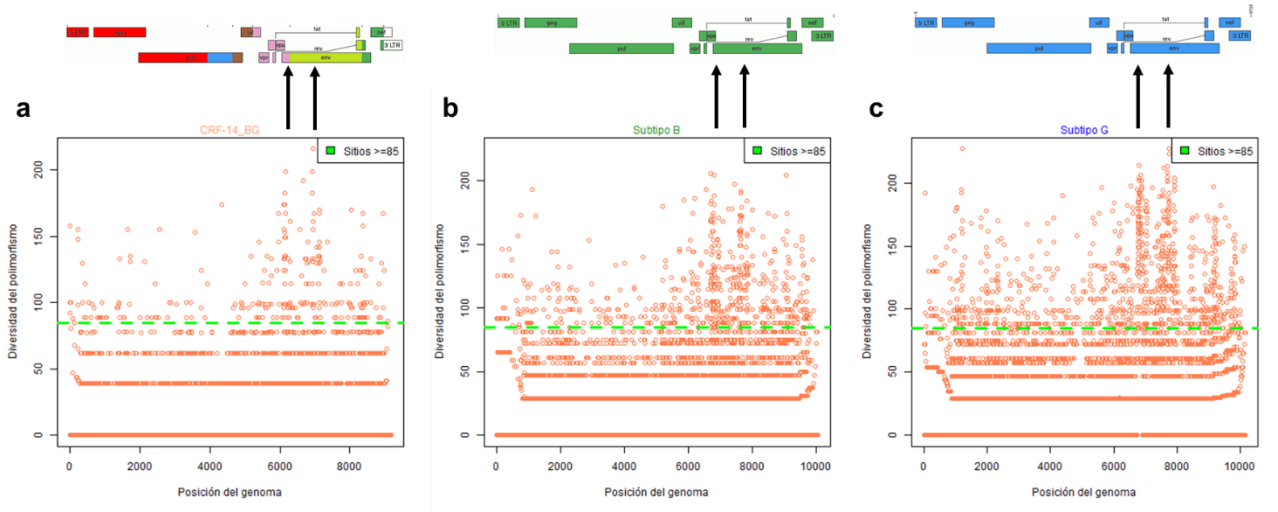

Figura 46. Diversidad del genoma completo de la **CRF-14_BG** y de los subtipos puros que la constituyen.

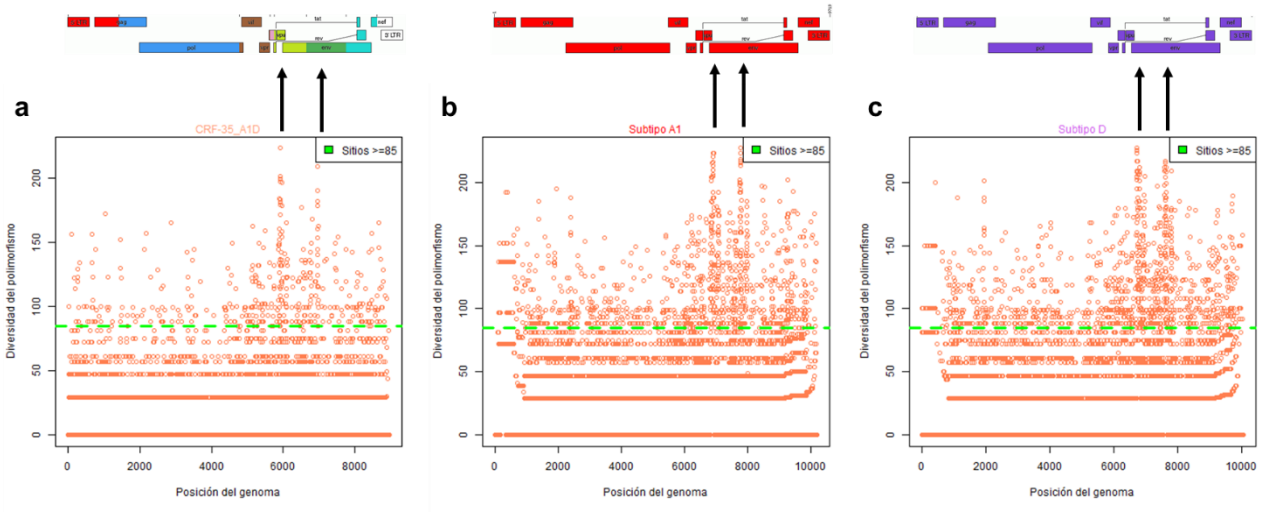

Figura 47. Diversidad del genoma completo de la **CRF-35_A1D** y de los subtipos puros que la constituyen.

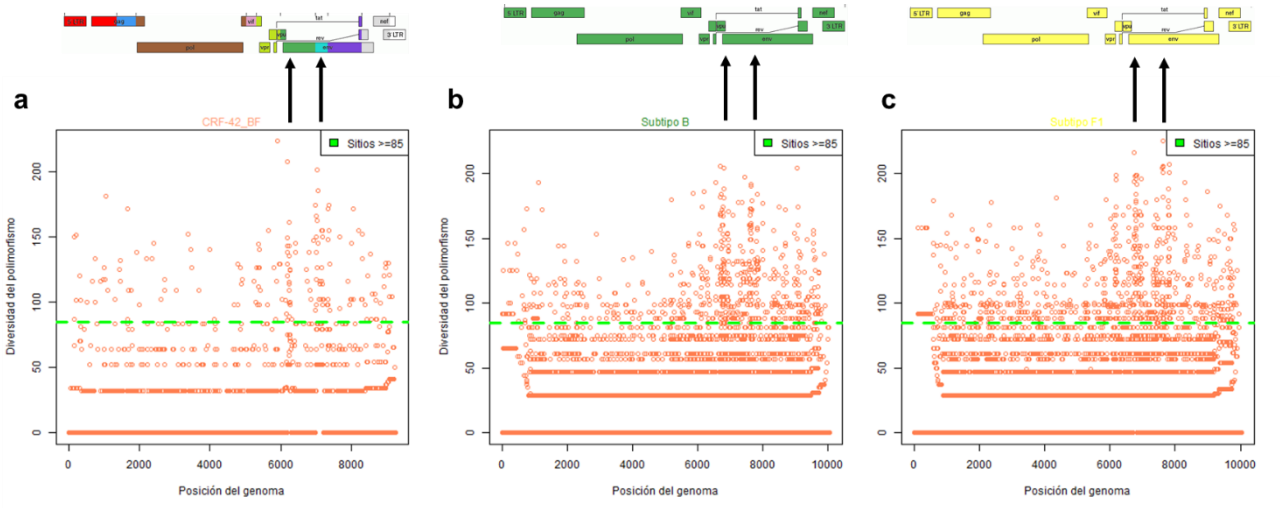

Figura 48. Diversidad del genoma completo de la **CRF-42_BF** y de los subtipos puros que la constituyen.

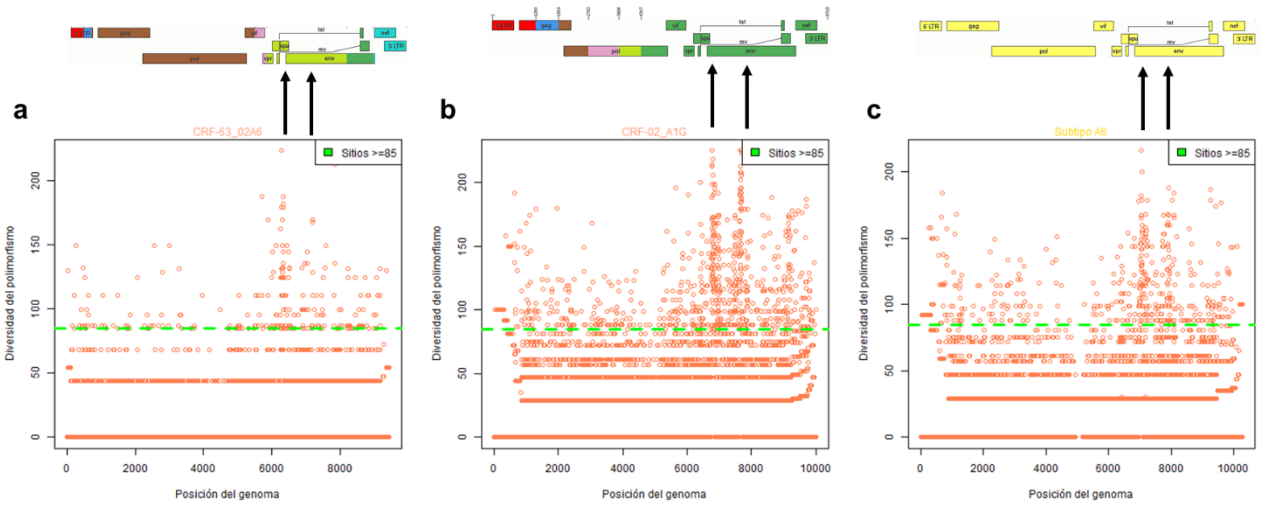

Figura 49. Diversidad del genoma completo de la **CRF-63_02A6** y de los subtipos puros que la constituyen.

12.4.3. Polimorfismos compartidos entre las CRFs y sus subtipos puros

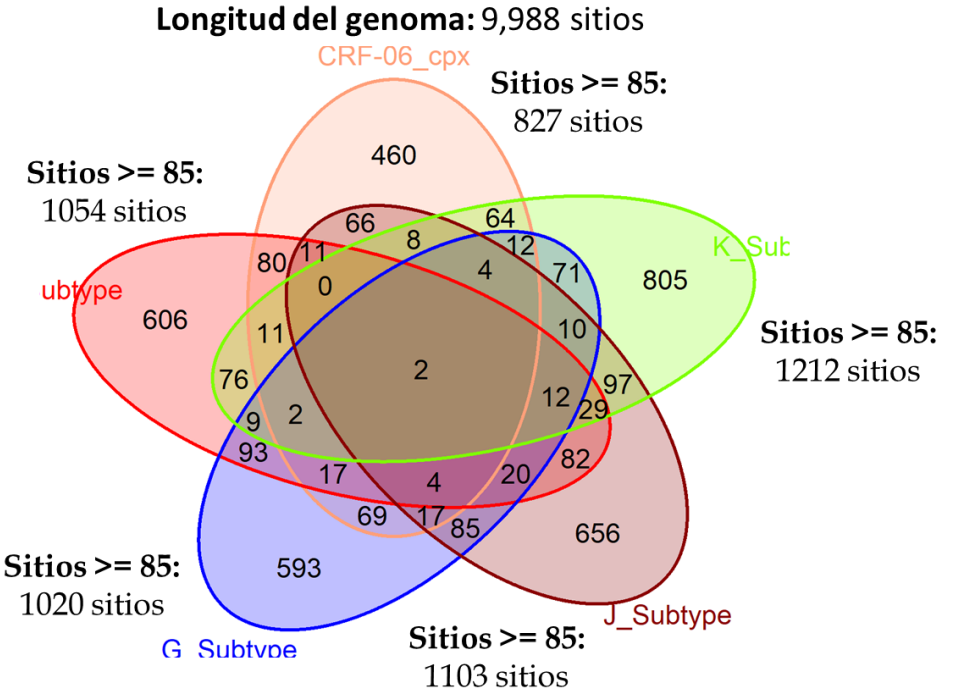

Figura 50. Número de sitios polimórficos presentes en los genomas de la **CRF-06_cpx(A1,G,J,K)** y de los subtipos puros que la constituyen.

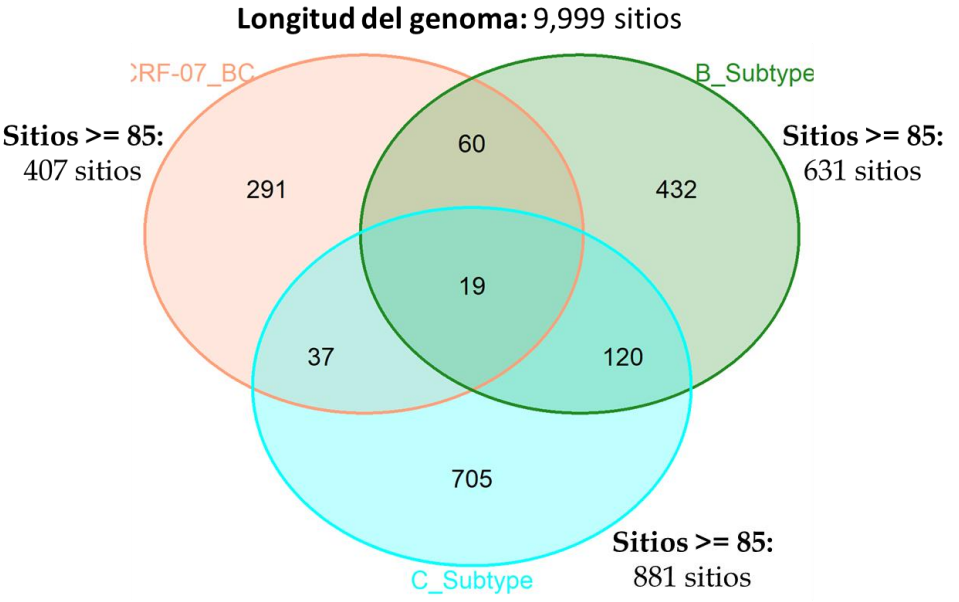

Figura 51. Número de sitios polimórficos presentes en los genomas de la **CRF-07_BC** y de los subtipos puros que la constituyen.

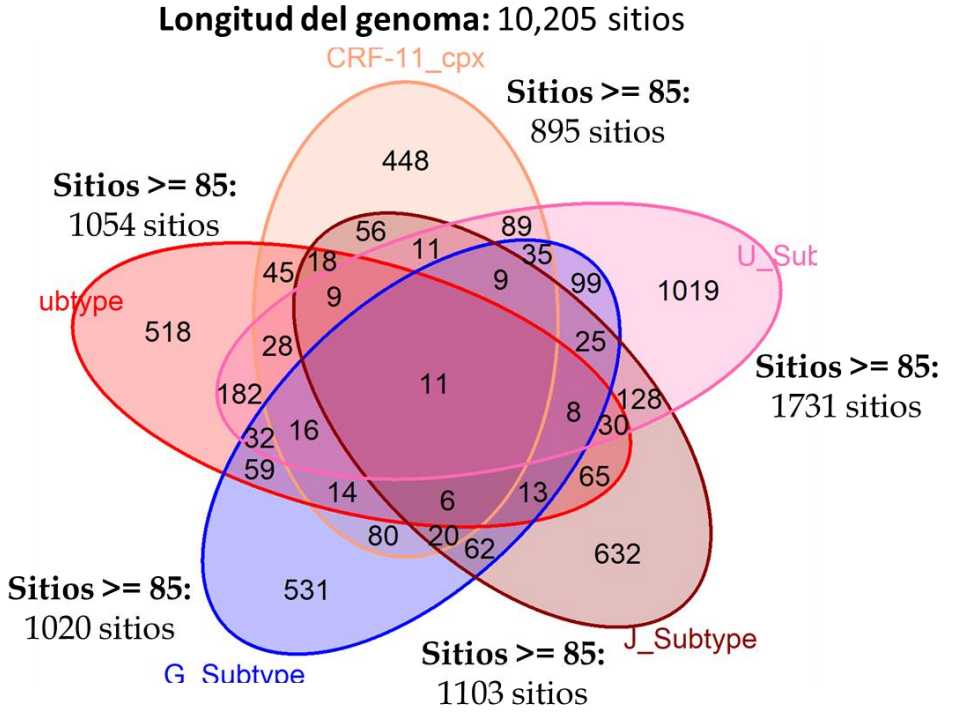

Figura 52. Número de sitios polimórficos presentes en los genomas de la **CRF-11_cpx(A1,E,G,J,U)** y de los subtipos puros que la constituyen.

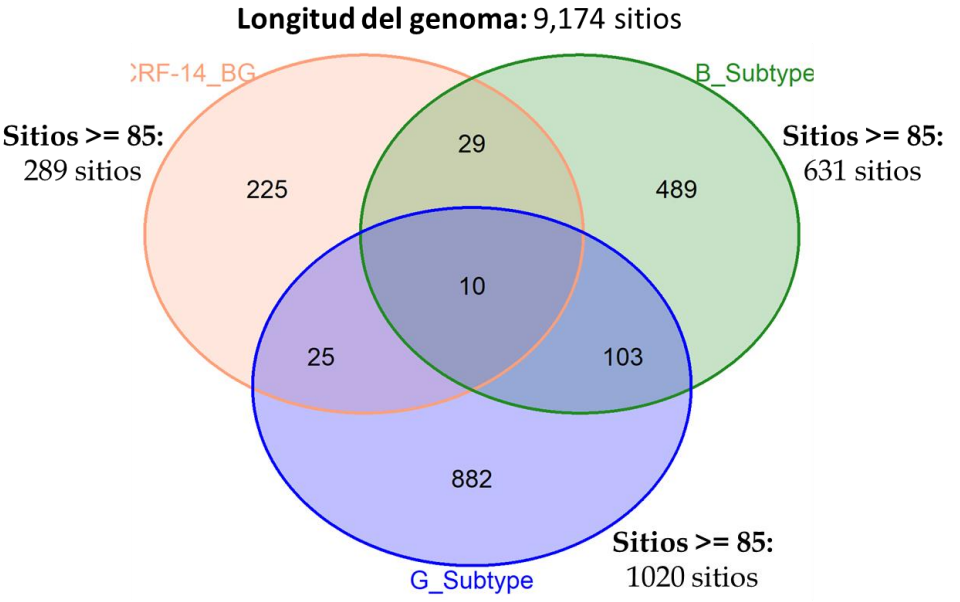

Figura 53. Número de sitios polimórficos presentes en los genomas de la **CRF-14_BG** y de los subtipos puros que la constituyen.

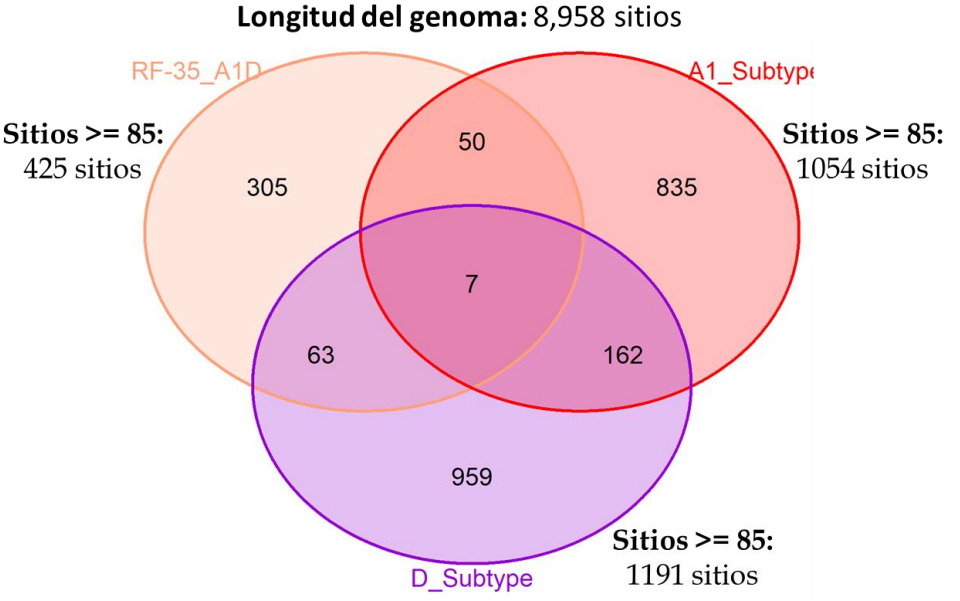

Figura 54. Número de sitios polimórficos presentes en los genomas de la **CRF-35_A1D** y de los subtipos puros que la constituyen.

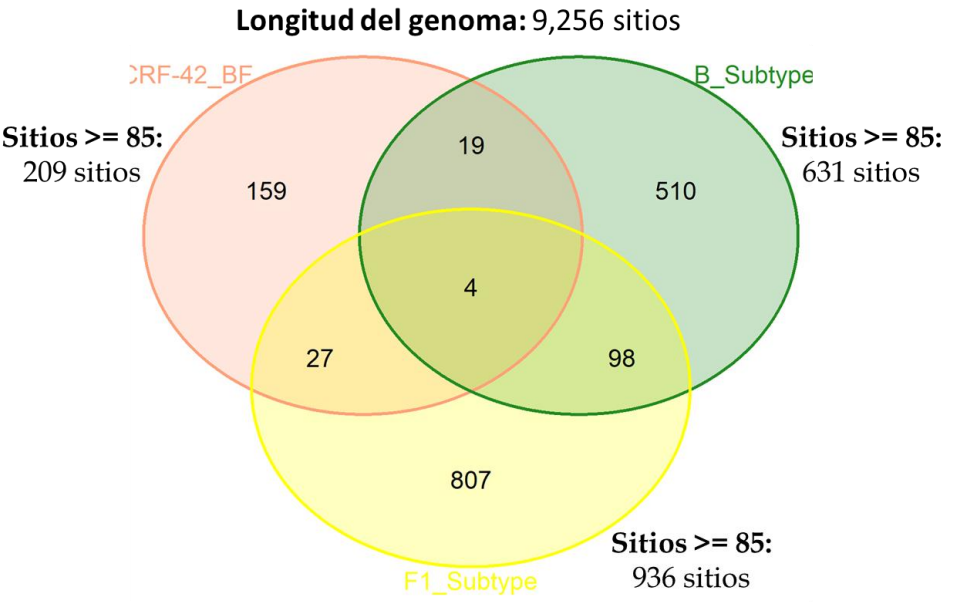

Figura 55. Número de sitios polimórficos presentes en los genomas de la **CRF-42_BF** y de los subtipos puros que la constituyen.

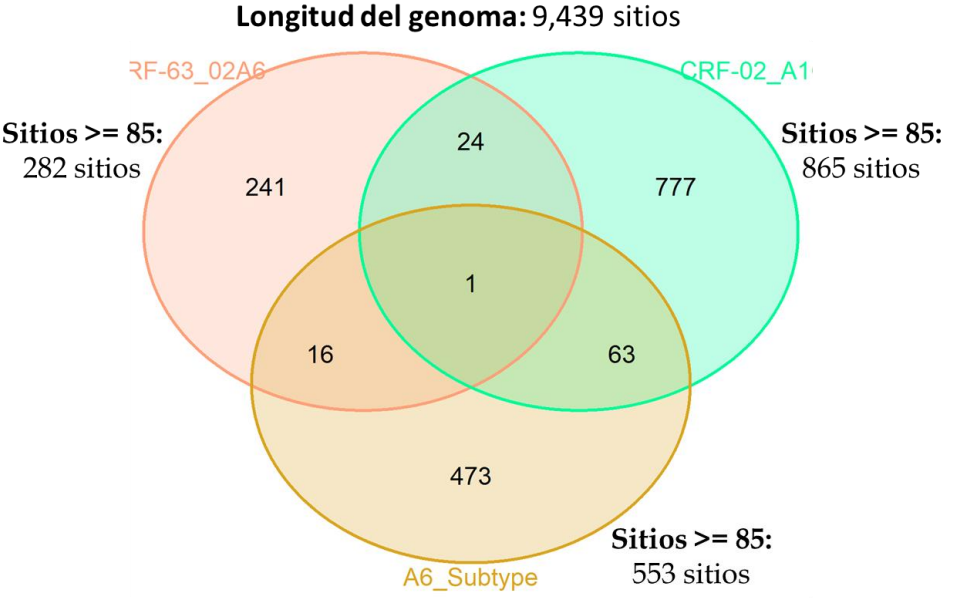

Figura 56. Número de sitios polimórficos presentes en los genomas de la **CRF-63_02A6** y de los subtipos puros que la constituyen.

12.4.4. Gráficas de comparación de la diversidad genética presente en los módulos

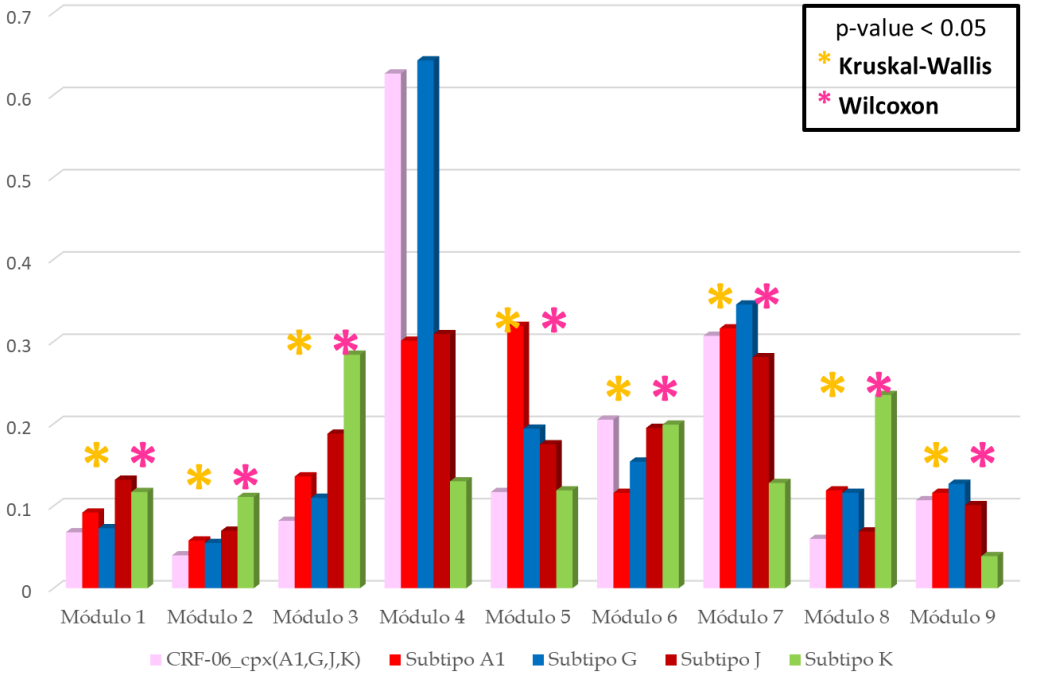

Figura 57. Comparación de la diversidad genética presente en los módulos de la **CRF-06_cpx(A1,G,J,K)** y sus subtipos puros.

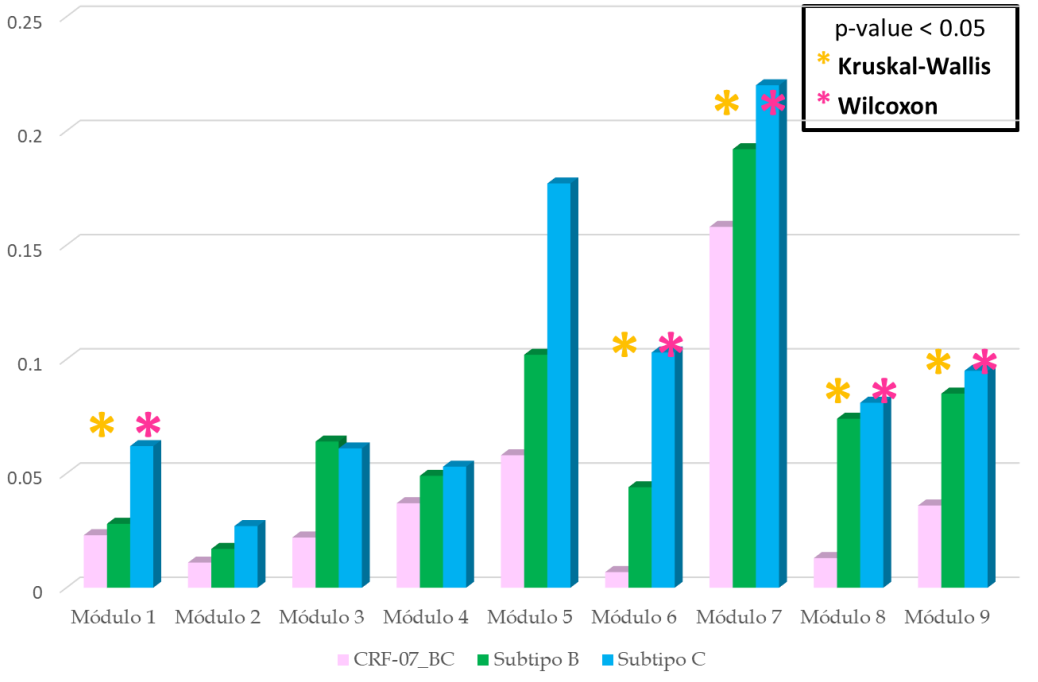

Figura 58. Comparación de la diversidad genética presente en los módulos de la **CRF-07_BC** y sus subtipos puros.

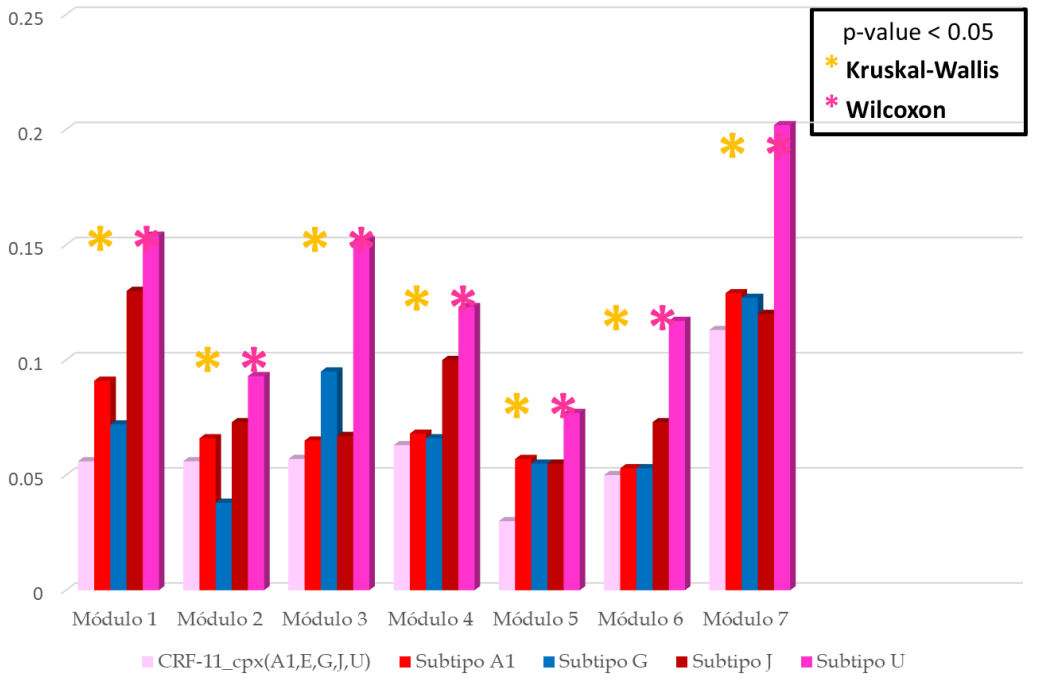

Figura 59. Comparación de la diversidad genética presente en los módulos de la **CRF-11_cpx(A1,E,G,J,U)** y sus subtipos puros.

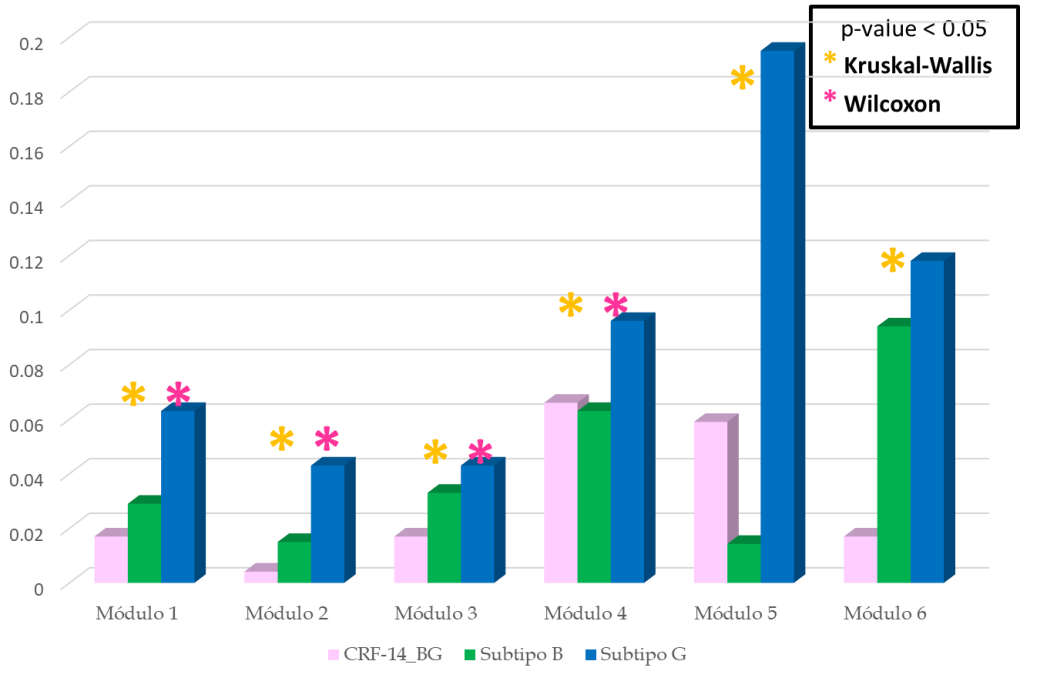

Figura 60. Comparación de la diversidad genética presente en los módulos de la **CRF-14_BG** y sus subtipos puros.

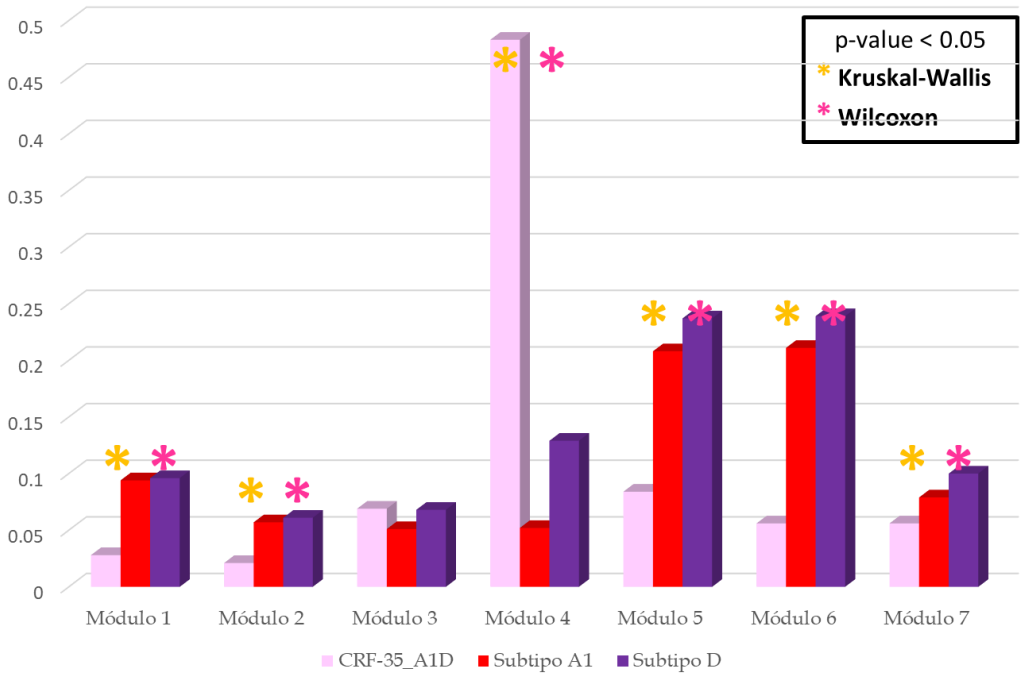

Figura 61. Comparación de la diversidad genética presente en los módulos de la **CRF-35_A1D** y sus subtipos puros.

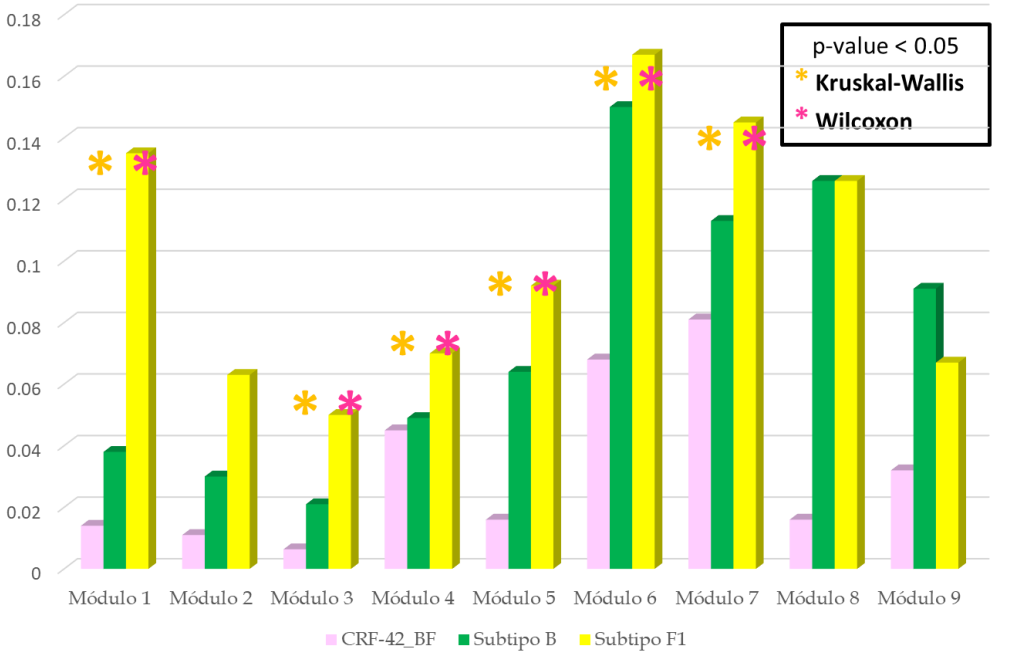

Figura 62. Comparación de la diversidad genética presente en los módulos de la **CRF-42_BF** y sus subtipos puros.

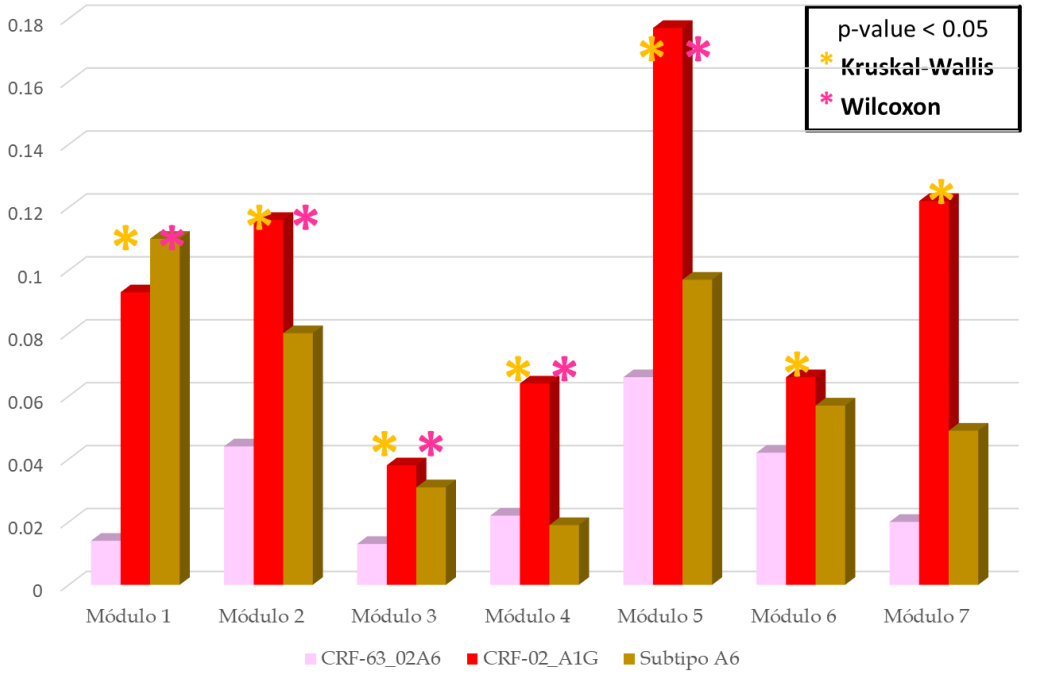

Figura 63. Comparación de la diversidad genética presente en los módulos de la **CRF-63_02A6** y sus subtipos puros.