UNIVERSIDAD DE PANAMÁ

FACULTAD DE CIENCIAS NATURALES, EXACTAS Y TECNOLOGÍA

ESCUELA DE BIOLOGÍA

“CAPACIDAD DE ANCLAJE DE LOS SITIOS DE UNIÓN DE LOS FACTORES DE TRANSCRIPCIÓN EN VIH-1 SUBTIPO B CIRCULANTE EN PANAMÁ”

Presentado por:

Ana Cristina Ortega Batista de Martínez; 8-952-191

Trabajo de graduación para optar por el título de Licenciado en Biología con orientación en Biología Molecular y Genética.

PANAMÁ, REPUBLICA DE PANAMÁ

2022

TRIBUNAL EXAMINADOR

Título:

“CAPACIDAD DE ANCLAJE DE LOS SITIOS DE UNIÓN DE LOS FACTORES DE TRANSCRIPCIÓN EN VIH-1 SUBTIPO B CIRCULANTE EN PANAMÁ”

Por:

Ana Cristina Ortega Batista de Martínez _____

8-952-191

Trabajo de Graduación presentado a consideración de la Escuela de Biología como requisito parcial para optar por el título de Licenciatura en Biología con Orientación en Biología Molecular y Genética.

Prof. Yaxelis Mendoza _____
Tutor (preside)

Prof. Carlos Ramos _____
Jurado

Prof. Ariel Magallón _____
Jurado

DEDICATORIA

A Jehová mi Dios y a mi madre.

AGRADECIMIENTO

Agradezco a Jehová mi Dios, a mi querido esposo Homaly Martínez por su incondicional apoyo y a mis padres. De igual forma, extiendo mi gratitud a la Dra. Yaxelis Mendoza por sus excelentes enseñanzas y acertadas correcciones. Agradezco a mis co-asesores, el Dr. Carlos Ramos y el Dr. Ariel Magallón por su confianza y tiempo invertido.

Además, extiendo mis agradecimientos al Instituto Conmemorativo Gorgas de Estudios de la Salud por abrirme sus puertas, y a la Secretaria Nacional de Ciencia y Tecnología por su financiamiento como becaria de pregrado en la licenciatura de Biología. Por último, agradezco a mi entrenador y maestro el Mgter. Juan Castillo Mewa por todo su apoyo a lo largo de esta investigación.

INDICE GENERAL

INDICE DE TABLAS

INDICE DE FIGURAS

LISTADO DE ABREVIATURAS

ADN	Ácido Desoxirribonucleico
ARN	Ácido Ribonucleico
CRF	Forma Recombinante Circulante
FRU	Forma Recombinante Única
FT	Factor de Transcripción
ICGES	Instituto Conmemorativo Gorgas de Estudios de la Salud
LASAGNA	Length-Aware Site Alignment Guided by Nucleotide Association
MPP	Matriz de peso de posición
NCBI	National Center for Biotechnology Information
PNC	Polimorfismo de Nucleotídeo Característico
REF	Referencia
RTL	Repetición terminal larga
SFT	Sitio de Unión del Factor de Transcripción
SIDA	Síndrome de Inmunodeficiencia Adquirida
TRANS	Transición nucleotídica
TRANV	Transversión nucleotídica
VESPA	Viral Epidemiology Signature Pattern Analysis

VIH	Virus de Inmunodeficiencia Humana
VIS	Virus de Inmunodeficiencia en Simios
VSA	Valor de Significancia de Anclaje

RESUMEN

El VIH es un retrovirus que lleva cuatro décadas de pandemia a nivel mundial. Su estructura flanqueada por las Repeticiones de Terminal Larga (RTL) poseen elementos genéticos que contienen importantes sitios de unión para los factores de transcripción (SFT). Estas regiones están encargadas de regular, modular e iniciar la transcripción por medio de factores de transcripción (FT). El objetivo de este estudio es determinar si la capacidad de anclaje de los FT a los sitios de unión depende de la diversidad genética en esa región. Para ello, se seleccionaron dos conjuntos de bases de datos del subtipo B pertenecientes a secuencias panameñas y de la base de datos super filtrada del Laboratorio de los Álamos. Luego, se realizó un análisis filogenético con métodos de Neighboor Joining y Bootstrap para subtipificar las secuencias de la base de datos de Panamá a través de MEGA7. Con un total de 55 secuencias por base de datos se identificaron los sitios de unión de la RTL 3' de los 5 factores de transcripción, con el programa bioinformático LASAGNA-Search. Una vez identificados los SFT, se determinó la diversidad genética de las regiones RTL 3' con un análisis de polimorfismo nucleotídico característico (PNC) en VESPA, y con la herramienta HIGHLIGHTER del Laboratorio de los Álamos se identificaron las transversiones y transiciones en los PNC. Luego, se predijo la capacidad de anclaje de los FT con un análisis de enriquecimiento en el programa CiiiDER. Se analizaron los resultados de las transversiones y transiciones en los PNC de los SFT con respecto al valor y significancia de la capacidad de anclaje de los FT estudiados (GATA3, SP1, USF1, NFKB1, NFATC2). Se demostró una correlación entre la cantidad de PNC y la

capacidad de anclaje del FT al sitio de unión en la secuencia del VIH-1. Se comprobó que la presencia de polimorfismos en los SFT afecta la capacidad de anclaje de los FT.

ABSTRACT

HIV is a retrovirus that has been a global pandemic for 4 decades. Its structure flanked by the Long Terminal Regions (RTL) possesses genetic elements that contain important binding sites for transcription factors (SFT). These regions oversee regulating, modulating and initiating transcription through transcription factors (FT). The objective of this study is to determine if the ability of TFs to anchor to binding sites depends on their genetic diversity in that region. For this, two sets of databases of subtype B belonging to Panamanian sequences and the super-filtered database of the Los Alamos Laboratory were selected. Then, a phylogenetic analysis was performed with Neighbor Joining and Bootstrap methods to subtype the sequences from the Panama database through MEGA7. With a total of 55 sequences per database, the 3' LTR binding sites of the 5 transcription factors were identified with the LASAGNA-Search bioinformatics program. Once the SFT were identified, the genetic diversity of the 3' RTL regions was determined with a nucleotide frequency analysis (PNC) in VESPA, and with the HIGHLIGHTER tool from the Los Alamos Laboratory, the transversions and transitions in the SNPs were identified. Then, the anchoring capacity of the TFs was predicted with an enrichment analysis in the CiiiDER program. The results of the transversions and transitions in the SNPs of the SFT were analyzed with respect to the value and significance of the anchoring capacity of the TFs studied (GATA3, SP1, USF1, NFKB1, NFATC2). A relationship was demonstrated between the number of SNPs with transversions or transitions and the ability of the FT to anchor to the binding site in the HIV-1 sequence. It was verified that the presence of polymorphisms in the SFT affects the anchoring capacity of the TFs.

INTRODUCCIÓN

El virus de la inmunodeficiencia humana es responsable 41 años de pandemia a nivel mundial (Delatorre & Bello, 2013). Su alta diversidad ha representado un reto para muchas generaciones de investigadores en todo el mundo. El VIH-1 grupo M es el responsable de la mayoría de las infecciones y ha sido clasificado en los subtipos A-D, F-H, J y K. La distribución de este retrovirus depende de varios factores como la prevalencia, la mortalidad global, los factores socioeconómicos, el acceso a tratamiento antirretroviral y el aumento de recombinación en los diferentes subtipos entre otros. Este virus, es uno de los patógenos más diversos genéticamente debido a su alta tasa de mutación, recombinación, y replicación. El rápido proceso evolutivo de este patógeno dio como resultado muchos subtipos que están distribuidos heterogéneamente a nivel mundial. El subtipo A predomina en el este de África, Rusia y la Comunidad de Estados Independientes; el subtipo B predominan el continente de América y Oceanía; subtipo C el sur de África e India; CRF01_AE prevalece en Asia y CRF02_AG en el Oeste de África. El aumento en la recombinación viral en las cepas de VIH por medio de la coinfección y super-infección se ha hecho más frecuente, lo que exige una continua vigilancia y seguimiento de la diversidad viral de este patógeno.

El estudio de la diversidad de los SFT (Sitios de Unión del Factor de Transcripción) en las RTL es el primer paso para hacer estudios que estén dirigidos al mecanismo de activación viral del provirus por medio de los sitios de unión de las RTL (Rausch & Le Grice, 2020). Este trabajo de graduación tiene como objetivo determinar la diversidad genética de los factores de transcripción de las RTL del VIH-1 que circula en Panamá del año 2011 al 2018 por medio

de un grupo de secuencias del grupo de investigación del VIH del Instituto Conmemorativo Gorgas de Estudios de la Salud.

Se intificaron los sitios de unión de los factores de transcripción a través del programa LASAGNA-Search y se determinó la diversidad genética de los sitios de unión por medio de análisis de frecuencia nucleotídica «Signature Sequence Analysis» en VESPA. Con el programa CiiiDER se determinó si los polimorfismos de estos sitios afectan el anclaje con los factores de trascripción. En este trabajo de graduación se estudió a nivel *in-silico* si la capacidad de anclaje de los factores de transcripción a los sitios de unión depende de su variabilidad genética en esa región.

JUSTIFICACIÓN

El VIH-1 es un retrovirus responsable de la muerte de 30 millones de seres humanos en las últimas 4 décadas de pandemia, y continúa siendo una preocupación de alta relevancia para la salud pública (McLaren & Fellay, 2021). Múltiples transmisiones zoonóticas de la cepa SIV a humanos ha permitido el desarrollo de distintas variantes o grupos de VIH (Smyth, Davenport, & Mak, 2012). Las altas tasas de mutación y recombinación durante la transcripción inversa del VIH crean una diversidad genética entre y dentro de los individuos afectados (Berg et al., 2016). El estudio y monitoreo de la diversidad genética del VIH ayuda a entender el desarrollo y aparición de nuevos subtipos y la presencia de nuevas variantes en una ubicación geográfica determinada (Hemelaar, 2012).

Las Repeticiones terminales largas (RTL) poseen elementos genéticos que contienen importantes sitios de unión para los factores de transcripción (Ramirez de Arellano, Soriano, & Holguin, 2005). La RTL son pares de secuencias idénticas de ADN en ambos extremos del genoma (5' y 3'), que se encargan de promover y modular la transcripción proviral por medio de sus diferentes complementos genéticos (Ramirez de Arellano et al., 2005). La variabilidad genética de las RTL del VIH-1 provocan alteraciones funcionales en los sitios de unión del factor de transcripción, lo que da como resultado una actividad promotora alterada (Burdo et al., 2004). Incluso, los cambios específicos en las RTL anulan la unión de factores de transcripción afines a su sitio de unión correspondiente y están relacionadas con las etapas graves de la enfermedad provocada por el VIH-1 (Nonnemacher, Irish, Liu, Mauger, & Wigdahl, 2004). A través de la identificación de los elementos genéticos ya mapeados en la secuencia consenso HXB2 (GenBank K03455.1) y por medio de programas bioinformáticos

como LASAGNA-Search se puede identificar elementos genéticos en la región 3' como: NF-kB1, USF1, NFATC2, SP1 y GATA3 (Gómez-Román & Soler-Claudín, 2000; Jeeninga et al., 2000). Los polimorfismos que ocurren en la RTL influyen en la patogenicidad y transmisión del VIH-1 (Singh et al., 2021). Una mayor comprensión de las RTL del VIH-1 permitirá elucidar la diversidad de estas secuencias nucleotídicas y aportará al desarrollo de estudios moleculares futuros sobre el VIH-1 (Maina, Adan, Mureithi, Muriuki, & Lwembe, 2021).

OBJETIVOS

Objetivo general

Determinar la variación a nivel de secuencia de los sitios de unión de los factores de transcripción de la RTL del grupo M del VIH-1 de secuencias publicadas en base de datos y virus circulantes en Panamá.

Objetivos específicos

- ✓ Identificar a que subtipos del VIH pertenece la matriz Panamá.
- ✓ Determinar la diversidad genética de los sitios de unión de los factores de transcripción de la RTL del VIH-1.
- ✓ Predecir la capacidad de anclaje a los factores de transcripción a los sitios de unión.
- ✓ Comparar la diversidad genética de los sitos de unión de secuencias publicadas en bases de datos con las secuencias obtenidas de virus circulantes en Panamá.

CAPÍTULO I:

ANTECEDENTES

1.1. Descubrimiento del VIH

El origen de la enfermedad provocada por el virus de inmunodeficiencia humana (VIH) se remonta a inicios del año 1920 y se cree que se propagó desde Camerún del Sur hacia la región de Kinshasa, conocida actualmente como la República Democrática del Congo (Berg et al., 2016). Desde el reconocimiento del VIH como causante del Síndrome de la inmunodeficiencia humana (SIDA) en 1984, 60 millones de personas han sido infectadas y de ellas 30 millones han muerto (Hemelaar, 2012).

En el año 1981 se reportaron los primeros casos de SIDA en cinco hombres homosexuales previamente sanos, infectados por *Pneumocystis jirovecii* (PCP), un hongo oportunista que infecta a individuos gravemente inmunosuprimidos (Gallo & Montagnier, 2003; Rodríguez, 2017). Hasta la fecha, no se comprendía la causa probable del síndrome, sin embargo, se sospechaba de retrovirus que infectan humanos como el Virus linfotrópico humano de células T y el VIH. La identificación de la causa del SIDA fue un desafío y no fue hasta el año 1894 que se produjo evidencia fundamentada de que existía una asociación entre la infección por el VIH y el SIDA (Agarwal-Jans).

Curiosamente, previo al descubrimiento del VIH, a finales del 1970, la población citadina consideraba que las enfermedades causadas por los microbios (bacterias, hongos y virus) no afectaban áreas altamente urbanizadas e industrializadas y no se conocía hasta la fecha un retrovirus capaz de infectar a humanos (White et al., 2012). Sin embargo, 15 años de estudios básicos en retroviruses leucemogénicos en animales llevaron al descubrimiento del HTLV, primer retrovirus descubierto que infectaba humanos (Gallo & Montagnier, 2003). Tal

hallazgo fue de utilidad para el descubrimiento de la cepa del VIH al sentar las bases de estudios de retroviruses en humanos.

1.2. Origen y evolución

Análisis filogenéticos han permitido identificar que el VIH se originó de procesos zoonóticos, su transmisión inició por medio del virus de la inmunodeficiencia en simios (VIS) desde primates no-humanos a humanos en el oeste de África Central (Hahn, Shaw, De, Cock, & Sharp, 2000). El VIH-1, causante de la actual pandemia, se deriva de la variante VIScpz que es el VIS encontrado en el chimpancé *Pan troglodytes troglodytes* (Smyth et al., 2012); mientras que, el pariente más cercano es el VIH-2 y se deriva del virus VISsm proveniente de monos verdes (*Cercocebus torquatus atys*) (Velasco-de-Castro et al., 2014). Por ende, el VIH-1 no se agrupa con el VIH-2, ya que ambas secuencias son parcialmente homólogas (Sharp & Hahn, 2011).

1.3. Clasificación filogenética

1.3.1. Grupos y subtipos del VIH-1

Actualmente la clasificación filogenética identifica al VIH-1 en cuatro grupos: M (mayor), O (valor atípico), N (no M, no O) y P (Berg et al., 2016). El VIH-1 grupo M se divide en nueve subtipos (A a F, H, J y K) (Liu et al., 2012). Un análisis a gran escala de 2996 secuencias genómicas de longitud completa reveló que las diversidades de nucleótidos

promedio es 37.5% entre grupos, 14.7% entre subtipos y 8.2% dentro de los subtipos (Berg et al., 2016).

1.3.2. Formas circulantes recombinantes y únicas

La diversidad genética, la coinfección y superinfección permitieron el desarrollo de "formas recombinantes circulantes" (CRF, según sus siglas en inglés), estas aportan un mayor grado de complejidad a la diversidad del virus y representan casi el 20% de las infecciones de todo el mundo (Hemelaar, 2012). La combinación de diversos factores como el rápido crecimiento de la población, las relaciones sexuales, el uso de agujas no esterilizadas, la manipulación, caza y consumo de carne de primate no-humano confluyeron a la rápida dispersión del virus, especialmente del VIH-1 grupo M y grupo O a inicios de la pandemia (Hahn et al., 2000; Hemelaar, 2012; Junqueira et al., 2011; Kirchner, 2019).

1.4. Diversidad genética del VIH en Panamá

El VIH-1 subtipo B es el de mayor prevalencia en América y ha jugado un rol importante en la historia de la epidemia del VIH (Junqueira & Almeida, 2016). A través de la rápida propagación del VIH-1 subtipo B en el continente africano, el virus se estableció en La Española conformada por República Dominicana y Haití para luego ser introducido en el Norte de América y finalmente extender la epidemia al continente de Europa, Asia, América Latina y Australia (Junqueira et al., 2011). Evidencia histórica y filogenética sugieren que el VIH-1 subtipo B se introdujo en América en el área del caribe evolucionando al linaje VIH-1 subtipo B caribeño, del cual asciende el linaje VIH-1 subtipo B pandémico que se diseminó

independientemente a otros países de América, Europa y el mundo a partir de los años 60s (Cabello, Mendoza, & Bello, 2014; Junqueira et al., 2011). Un estudio indica que aproximadamente un tercio de las infecciones del VIH-1 subtipo B pandémico en Latinoamérica, son resultado de la expansión de cepas fundadoras que fueron introducidas en una etapa temprana de la epidemia en América (Mir, Cabello, Romero, & Bello, 2015). En Panamá, coexisten ambos linajes con una mayor prevalencia del VIH-1 subtipo B pandémico, y se sugiere que la epidemia panameña del VIH-1 está mayormente impulsada por la expansión de clados VIH-1 subtipo B endémicos (Mendoza, Bello, et al., 2014).

1.5. Estructura y genoma del VIH

1.5.1. Ciclo replicativo

Según National Institute of Health, El ciclo replicativo del VIH ocurre en los siguientes pasos (figura 1). Primero, ocurre un enlace o fijación del virus a los receptores de la superficie de células T CD4+ y células del linaje de monocitos/macrófagos (Deeks, Overbaugh, Phillips, & Buchbinder, 2015). La glicoproteína gp120 de la cubierta viral se une al receptor 5 de quimiocinas CD4 y CC (CCR5) en la superficie de la célula, lo que desencadena la fusión del virus en la célula huésped, integrándose a la célula (Delgado, 2011). Segundo, el VIH ya localizado en el linfocito utiliza una enzima llamada transcriptasa inversa para convertir el ARN del virus a ADN, esta nueva cadena sintetizada entra al núcleo del linfocito y por medio de la enzima integrasa integra su ADN vírico dentro del ADN del linfocito CD4. A este punto del ciclo replicativo, se puede considerar que ha ocurrido una infección exitosa por parte del virus (Cordeiro, Taroco, & Higiene, 2008). Tercero, ocurre una integración por

parte del virus, multiplicándose dentro del linfocito, y creando largas cadenas de proteínas del virus. Estas cadenas se convierten en elementos constitutivos para la creación de las copias de VIH. Cuando hay suficientes proteínas ocurre un ensamblaje que da lugar al VIH inmaduro o no infeccioso (Goodsell, 2015). Cuarto, el VIH inmaduro se impulsa hacia el exterior de la célula y libera en su interior una enzima llamada proteasa que permite la maduración del virus y su infectividad. Este mecanismo lo utiliza el virus repetidas veces dentro del organismo infectado (Kirchhoff, 2013).

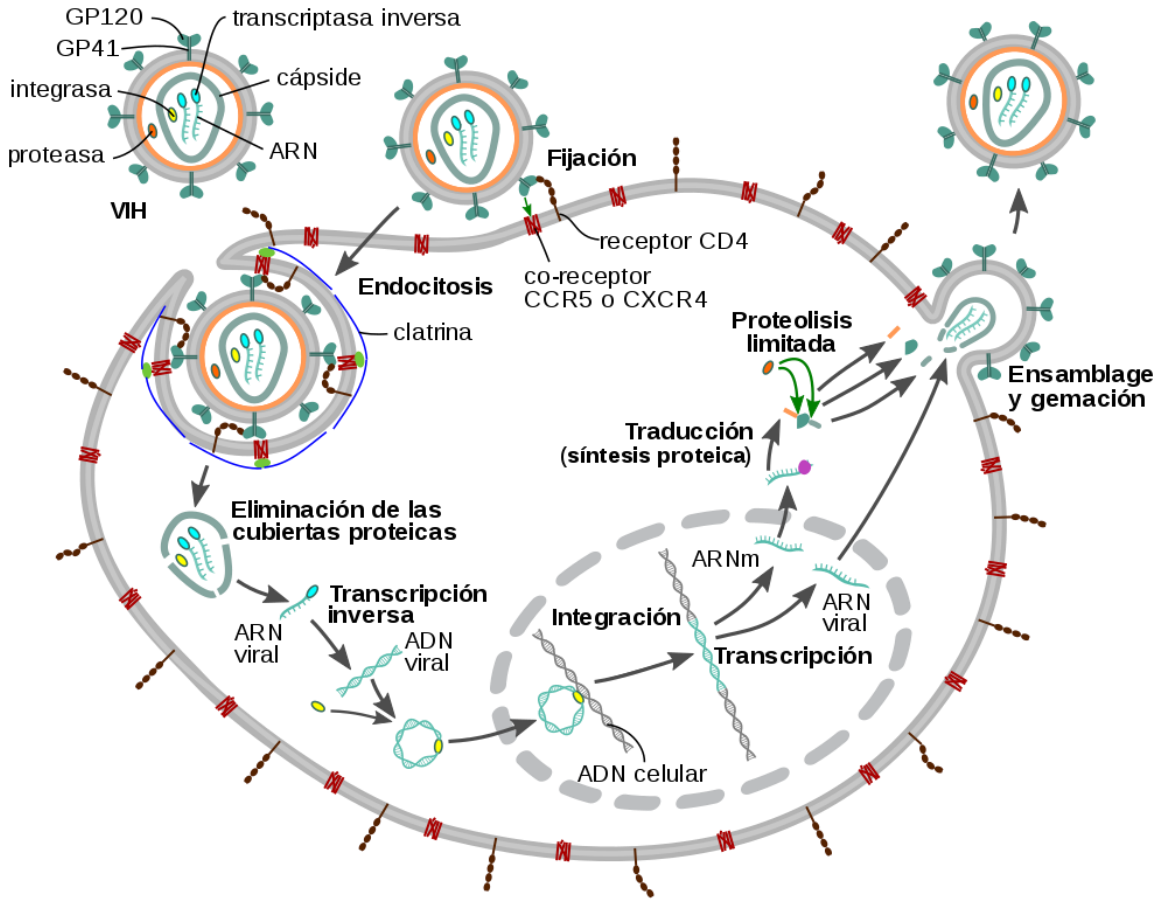


Figura 1. Ciclo replicativo del VIH. Fuente: Wikipedia Commons.

1.5.2. Genoma y composición

La longitud del genoma del VIH es de aproximadamente de 9.8 kb, a pesar de ser un genoma pequeño, su variabilidad tiene gran impacto, debido a que su tiempo de generación es corto (de 10^{10} viriones producidos por día en un individuo infectado), y produce alta tasa de errores en el proceso de transcripción inversa (generando en promedio 3.4x10-5 mutaciones por sitio por generación) (Castro-Nallar, Perez-Losada, Burton, & Crandall, 2012; Mansky & Temin, 1995).

1.6. Repeticiones de Terminal Larga del VIH

El genoma del VIH-1 está flanqueado en los extremos 5´ y 3´ por las RTL que consisten en regiones únicas 5′ (U5), únicas 3′ (U3) y repetidas (R) (Figura 2) (Roebuck & Saifuddin, 1999). Las RTL dirigen la transcripción inicial del VIH-1 proviral que está controlado por la 5´RTL y depende de los factores de transcripción de la célula huésped que se unen a una serie de elementos reguladores en cis de ADN en el promotor de RTL (Roebuck & Saifuddin, 1999). De éstas regiones, la U3 contiene las regiones promotoras, potenciadoras y moduladoras con importantes sitios de unión para proteínas celulares (Gómez-Román & Soler-Claudín, 2000). Estas regiones son consideradas esenciales para la transcripción viral, la integración y la expresión génica (C. Mbondji-Wonje et al., 2018).

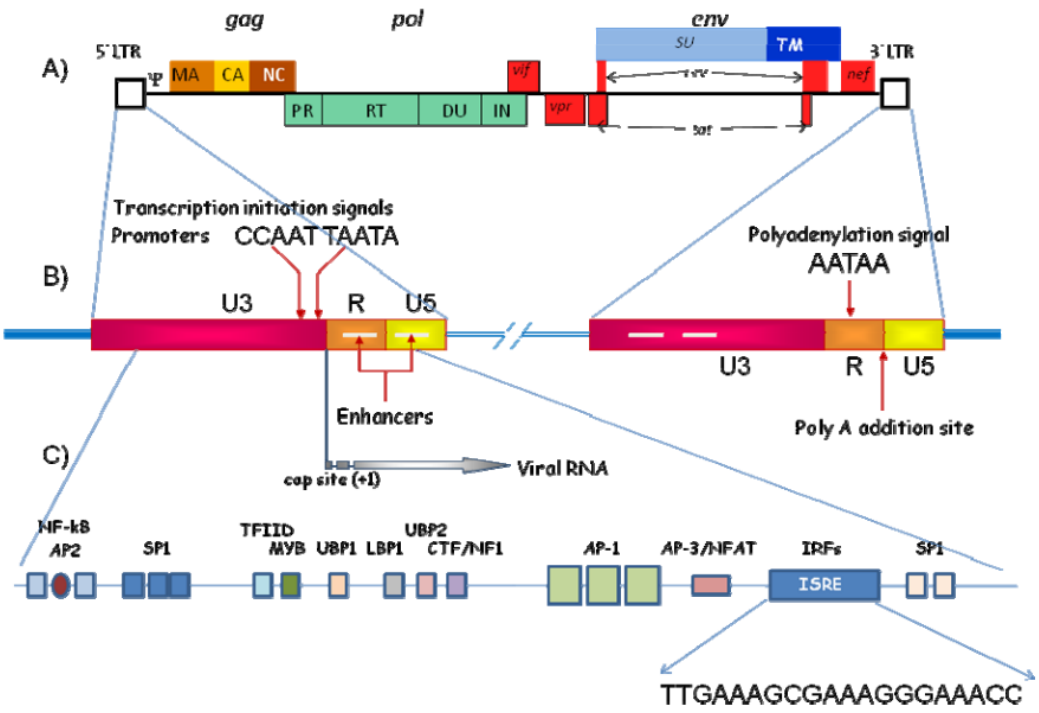


Figura 2. Representación esquemática de los genes del VIH-1 y RTL.

A) Estructura genómica del VIH-1. B) Ampliación de las dos Repeticiones Terminales Largas 5´ y 3´ (RTL) que flanquean la secuencia proviral. C) Algunos de los elementos reguladores en el 5' RTL. Fuente: (Gomez-Lucia, Collado, Miró, & Doménech, 2009).

1.6.1. Factores de transcripción

Existe diversidad en los sitios de unión de los factores de transcripción entre los subtipos del VIH del grupo M (Figura 3). En un estudio se demostró que el subtipo B presenta un sitio de unión exclusivo llamado USF, y que los sitios de unión de AP-1 están identificados en la mayoría de los RTL de subtipos, excepto para los subtipos B y D (Jeeninga et al., 2000). Por consiguiente, los RTL del VIH-1 promueven y modulan la transcripción proviral; incluso, la variabilidad genética de los RTL podría influir en la replicación viral y la progresión de la

enfermedad (Ramirez de Arellano, Martin, Soriano, Alcami, & Holguin, 2007). En esta investigación de tesis propongo estudiar in-silico si la capacidad de anclaje de los factores de transcripción a los sitios de unión depende de su diversidad genética en esa región.

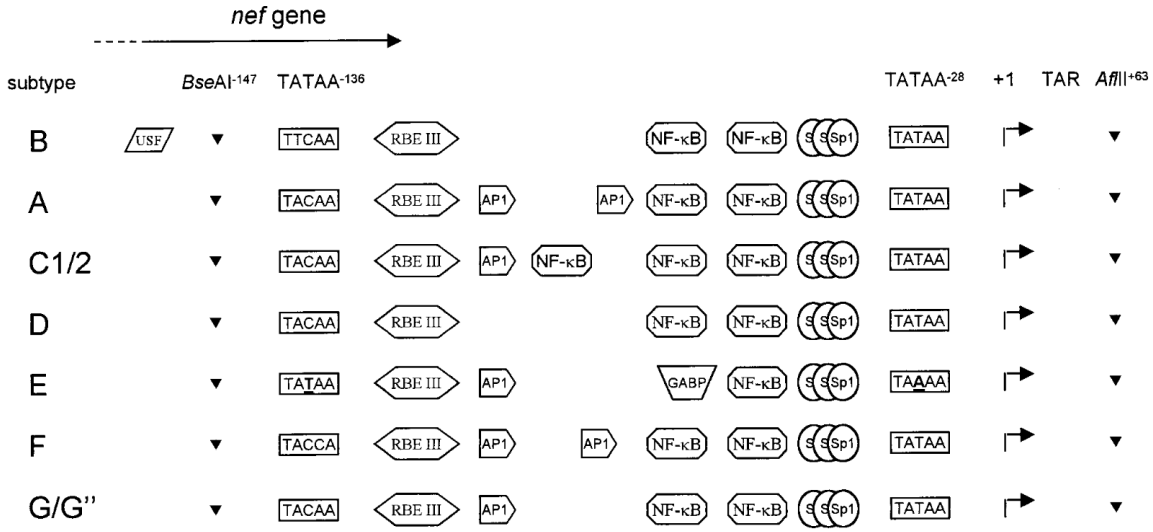


Figura 3. Organización del promotor RTL 3' de los subtipos A al G del VIH-1.

Fuente: (Jeeninga et al., 2000).

La transcripción del VIH da lugar en la región promotora, la cual dirige a la RNA-polimerasa celular en el proceso de la transcripción basal de los genes virales; el promotor del núcleo viral del subtipo B del VIH-1 comprende tres sitios de unión a Sp1 en tándem, una caja TATA y un elemento iniciador en el sitio de inicio de la transcripción (Figura 4) (Mbonye & Karn, 2017). Cerca de la región promotora, se encuentra la región de elementos aumentadores, la cual se encarga de la transcripción inducible de los genes virales. Esta región contiene una duplicación de 10 pares de bases [*GGGACTTTCC*], conocida como el sitio NF-kB y participa en la transcripción viral en respuesta a señales de activación, citocinas, mitógenos, y otros estímulos inmunológicos (Gómez-Román & Soler-Claudín, 2000). Por último, la región moduladora se encarga de la transcripción regulada. Esta región contiene múltiples

secuencias en cis que sirven como sitios de unión a diversos factores celulares tales como AP-1, NFAT-1, USF-1, entre otros factores de transcripción (Hokello, Lakhikumar Sharma, & Tyagi, 2021).

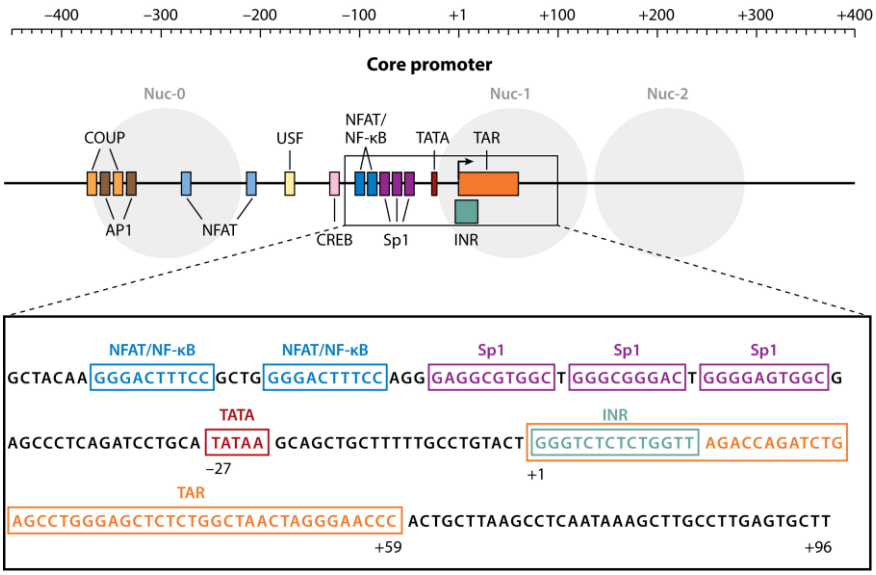


Figura 4. Estructura de la repetición terminal larga viral (arriba) y secuencia del promotor del núcleo viral de los sitios de unión de NFAT/NF-κB

CAPÍTULO II:

METODOLOGÍA

2.1. Población de estudio

Este es un estudio descriptivo, y retrospectivo, del 2011 al 2018, elaborado con secuencias nucleotídicas pertenecientes al proyecto de investigación "Determinantes Genéticos de la Incidencia de la Infección del VIH en Panamá, SINIP 9044.066" y el proyecto "Variabilidad Genotípica del Gen de la Integrasa del VIH, asociada a Resistencia a Drogas Antirretrovirales, FID16-IP-103" del Instituto Conmemorativo Gorgas (ICGES).

Las secuencias nucleotídicas, que constituyen nuestra población de estudio, consisten en muestras de sujetos naïve (sin tratamiento previo) y/o recién diagnosticados, pacientes con diferentes esquemas de tratamiento antirretroviral (TARV), procedentes de las clínicas de terapia antirretroviral del país. Todas las muestras del estudio pertenecen a individuos mayores de 18 años.

Para la construcción de la matriz Panamá se seleccionaron 55 secuencias que cumplieran con las siguientes características:

- ✓ Ser secuencias de pacientes panameños.
- ✓ Ser pertenecientes al VIH-1, grupo M, subtipo B.
- ✓ Contener el 80% de la RTL 3´ según la secuencia de referencia HXB2 (GenBank K03455.1).
- ✓ Ser secuencias nucleotídicas obtenidas por secuenciación de segunda generación con una resolución 10X.

Las 55 secuencias de la Matriz Panamá no fueron elegidas en base a características clínicas y/o epidemiológicas, sino únicamente por características estructurales a nivel nucleotídico.

La población de estudio se conformó principalmente por sujetos del sexo masculino (60%) y rango de edad entre 38-47 años (38%) (Tabla 1).

Tabla 1. Características epidemiológicas y clínicas de la población de estudio

Características	Total n=55 (%)
Género	
Masculino	33 (60)
Femenino	21 (38)
Sin información	1 (2)
Edad (años)	
18-27	8 (15)
28-37	17 (31)
38-47	21 (38)
48-61	8 (15)
NI	1 (2)
Periodo de diagnóstico de VIH (años)	
1998-2003	5 (2)
2004-2008	17 (31)
2009-2013	24 (44)
2014-2018	8 (15)
NI	1 (2)
Conteo celular de linfocitos T CD4+ (células/µL)	
< 200	24 (44)
200-499	19 (35)
≥ 500	6 (11)
NI	6 (11)
Carga Viral del VIH-1 (copias de ARN/mL plasma)	
< 1,000	3 (5)
1,000-10,000	6 (11)
>10,000	45 (82)
NI	1 (2)

Los datos son N (%)

Abreviaciones: VIH, virus de la inmunodeficiencia humana; NI, No Identificado

2.2. Construcción del set de datos de secuencias nucleotídicas de referencia y secuencias de estudio

La Matriz Los Álamos se elaboró con las secuencias de los subconjuntos más limpios de las alineaciones web, del VIH-1, grupo M de la página del Laboratorio de Los Álamos con genoma completo (https://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html) (Figura 5); la alineación utilizada fue la última actualización de datos del año 2020. En base a los objetivos del estudio se eliminó todas las secuencias que no pertenecían al VIH-1, grupo M, subtipo B y que no contuvo el 90% de la RTL 3' según la secuencia de referencia HXB2. Luego se eliminaron las secuencias que pertenecían a un mismo paciente/individuo, obteniendo un total de 65 secuencias aptas para el análisis. Por último, se igualo la cantidad de secuencias de la matriz de los Álamos a la cantidad de secuencias de la matriz de Panamá (n=55). Para ello se eliminaron las 10 primeras secuencias con mayor cantidad de lagunas en la región de estudio, RTL 3'.

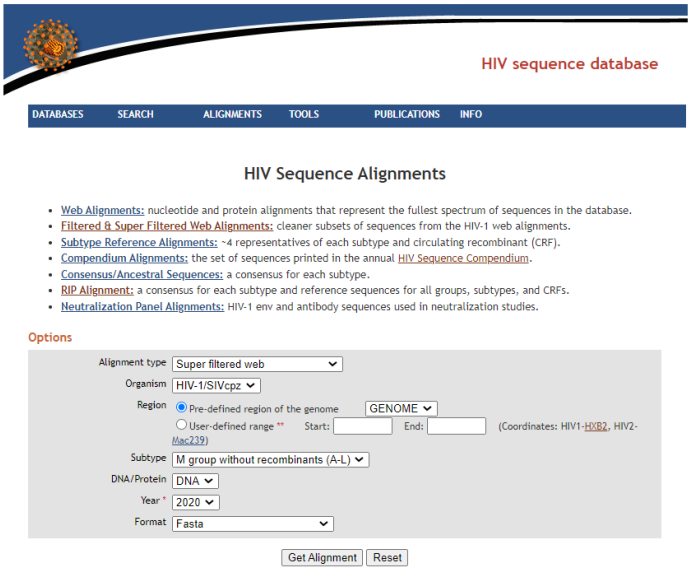


Figura 5. Base de datos de VIH del Laboratorio Nacional Los Alamos.

2.3. Análisis filogenético

Para la determinación de los subtipos que corresponden a las secuencias del Gorgas, se utilizó la herramienta de subtipificación en línea REGA versión 3.46 (https://www.genomedetective.com/app/typingtool/hiv). Para confirmar los resultados obtenidos en esta herramienta, se realizó una reconstrucción filogenética usando el programa MEGA (versión 7.0). Se descargó de la base de datos del Laboratorio de Los Álamos, secuencias de referencia de genoma completo de cada subtipo del VIH-1 grupo M y secuencias referencia del grupo N, O y P, para la confección de un árbol guía (Figura 6). El árbol guía se utilizó como referencia para determinar los subtipos de las secuencias no identificadas. El árbol guía involucró 43 secuencias de nucleótidos de referencia (REF), incluyendo 3 secuencias de outgroup, y. la Matriz Panamá (n=55); dando un total de 100 secuencias analizadas para la subtipificación.

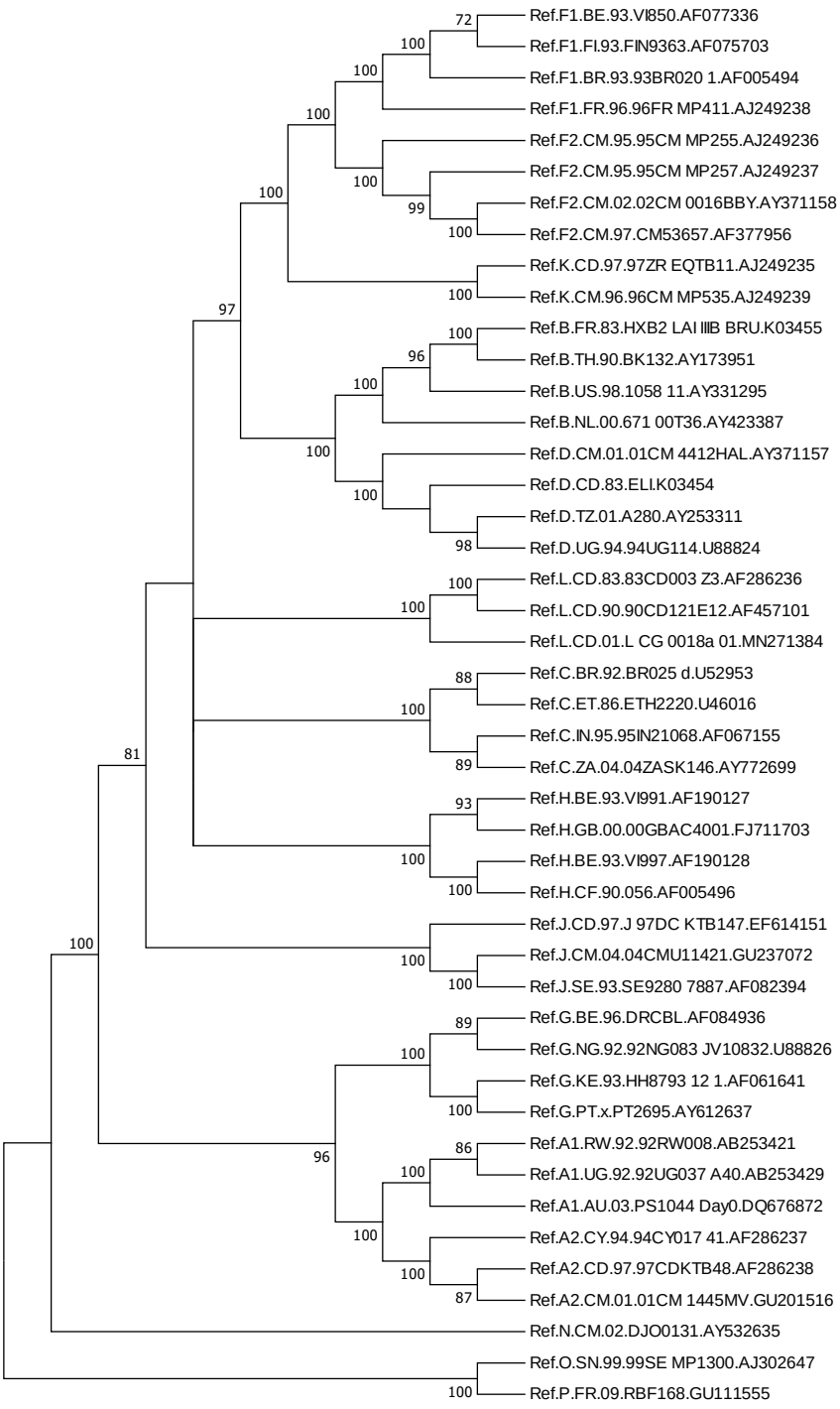


Figura 6. Árbol guía de las relaciones evolutivas de los taxones de matriz de secuencias panameñas.

Las secuencias de referencias de subtipos del VIH-1 se alinearon con las secuencias de Panamá utilizando el programa MUSCLE, junto con la base de datos de secuencias de

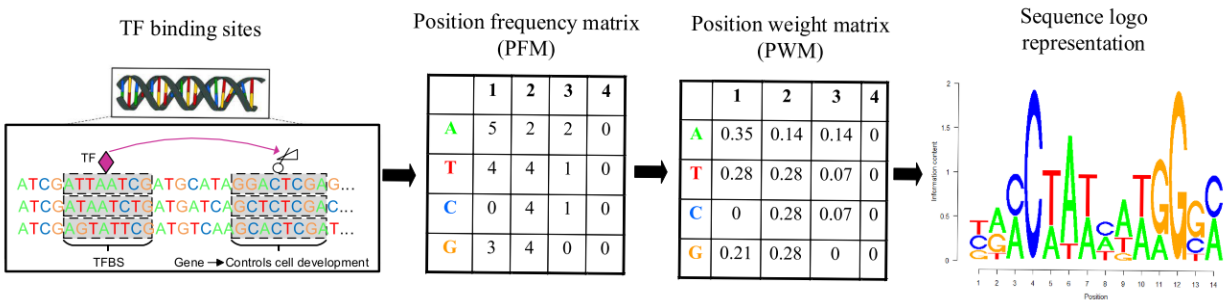
nucleótidos del Laboratorio de Los Álamos, utilizando la HXB2 como secuencia guía para el alineamiento (Tamura, Dudley, Nei, & Kumar, 2007). El programa MEGA7 eliminó todas las posiciones que contenían lagunas y datos faltantes (Kumar, Stecher, & Tamura, 2016).

La historia evolutiva se infirió utilizando el método Neighbor-Joining (Saitou & Nei, 1987). El porcentaje de árboles replicados en los que los taxones asociados se agruparon en la prueba de arranque de 500 repeticiones (Felsenstein, 1985). Las distancias evolutivas se calcularon usando el método Tamura-Nei (Tamura & Nei, 1993).

2.4. Selección de parámetros para el análisis de los sitios de unión del factor de transcripción

Para el desarrollo de los análisis bioinformáticos se estudiaron los siguientes parámetros:

A) Modelo de predicción de SFT. *S*e realizó una tabla comparativa con los métodos de predicción SFT que utilizan Aprendizaje Automático Tradicional (Traditional Machine Learning) y se eligió el mejor modelo en base a sus características, versatilidad y limitaciones según la pregunta de investigación. Según revisión bibliográfica se determinó que el mejor modelo es la matriz de peso de posición (MPP) (Figura 7) (Zeng, Gong, Lin, Gao, & Zhang, 2020).



	1	2	3	4
A	5	2	2	0
T	4	4	1	0
C	0	4	1	0
G	3	4	0	0

	1	2	3	4
A	0.35	0.14	0.14	0
T	0.28	0.28	0.07	0
C	0	0.28	0.07	0
G	0.21	0.28	0	0

Figura 7. El MPP es una matriz de N filas y cuatro columnas, en la que se describe la frecuencia de cada base en cada posición.

B) Base de datos. De igual forma, se realizó una tabla comparativa para la selección de una base de datos que incluyera perfiles de enlace de los factores de transcripción (FT) no redundantes, y formato de matrices de peso de posición (PPM) con acceso abierto a cualquier usuario; siendo JASPAR la mejor opción para la adquisición de las matrices y SFT previamente reportados en la literatura (Castro-Mondragon et al.).

C) Selección de factores de transcripción. Por último, se realizó una revisión bibliográfica de factores de transcripción que hayan sido previamente reportados en literatura por tener interacciones con el VIH. Los factores de transcripción escogidos cumplieron con las siguientes características: a) Tener un matriz de frecuencia de posición, b) tener un conjunto de sitios de unión de factores de transcripción previamente reportados (secuencias nucleotídicas de antecedentes background), c) haber sido reportados como FT que interactúan con el VIH-1, grupo M a nivel *in vivo*.

2.5. Identificación de los sitios de unión de los factores de transcripción de la RTL

Se identificaron los sitios de unión de los factores de transcripción por medio del programa en línea LASAGNA-Search que utiliza el algoritmo «LASAGNA» (Length-Aware Site Alignment Guided by Nucleotide Association) para la alineación SFT (Figura 8). Se introdujo cada matriz de los factores de transcripción en las secuencias alineadas en formato fasta y se eligió por secuencia la posición de cada sitio de unión con menor *valor de p*, siendo el rango entre *0,000475 y 0,0246* (Lee & Huang, 2013). El programa otorga información de relevancia para elanálisis del SFT como su posición (basada en 0), su valor de p, su valor de e y el puntaje que abarca los valores estadísticos.

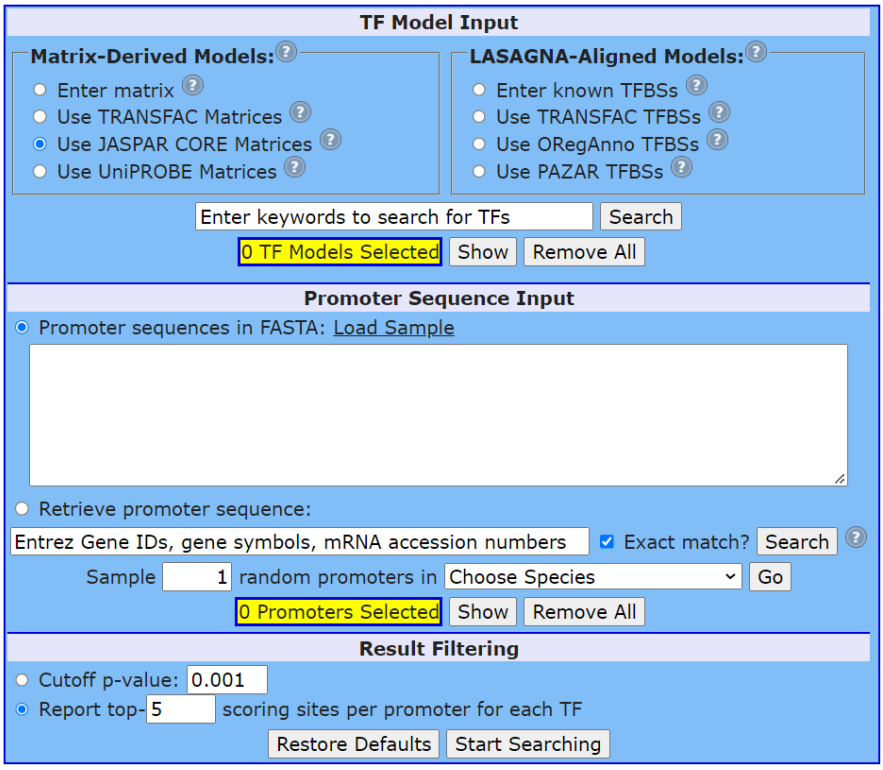


Figura 8. Programa en línea para la predicción de sitios de unión de los factores de transcripción.

2.6. Determinación de la diversidad genética de los sitios de unión de los factores de transcripción de las RTL del VIH-, grupo M

Análisis de frecuencia nucleotídica. Para visualizar la diversidad nucleotídica se hizo un análisis de patrón de nucleótidos que difieren entre sí «Signature Sequence Analysis» con la herramienta VESPA (Viral Epidemiology Signature Pattern Analysis) (https://www.hiv.lanl.gov/content/sequence/VESPA/vespa.html) (Korber & Myers, 1992). Esta herramienta, escrita en lenguaje C, está disponible en la base de datos de secuencias de VIH, Laboratorio Nacional de Los Alamos. Es un programa en lenguaje C corolario, SPCOUNT, que calcula la cantidad de aminoácidos o nucleótidos compartidos con una firma o nucleótidos característicos, dada para todas las secuencias incluidas en las matrices de nuestro estudio.

El programa VESPA (Análisis de patrones característicos de epidemiología viral) detecta residuos de aminoácidos o nucleótidos atípicos en un conjunto de secuencias de consulta en relación con un conjunto de secuencias de referencia. VESPA calcula la frecuencia de cada aminoácido (o nucleótido) en cada posición en una alineación para los dos conjuntos de secuencias que se comparan, seleccionando aquellos sitios para los cuales el aminoácido más común en el antecedentes (background) difiere del aminoácido más común en el conjunto de consulta (Korber & Myers, 1992). También, extrae las frecuencias de los aminoácidos distintivos. El patrón de firma para el conjunto de secuencias de consulta está definido por estos residuos característicos (Ou et al., 1992).

TABLE 1. FREQUENCIES OF THE DENTIST'S SIGNATURE PATTERN AMINO ACIDS

Signature:	A	I	A	G	A	E	E	V	I	H
Dentist:	1.00	1.00	1.00	1.00	1.00	0.67	1.00	1.00	0.83	1.00
Background:	0.06	0.13	0.06	0.16	0.25	0.25	0.06	0.25	0.19	0.16

Figura 9. Frecuencias de los aminoácidos del patrón característico del articulo (Korber & Myers, 1992).

En la figura 9 se presenta una ilustración de la metodología a emplear, utilizando como ejemplo el estudio de "Signature Pattern Analysis: A Method for Assessing Viral Sequence Relatedness" (Korber & Myers, 1992). La línea superior muestra el conjunto de diez aminoácidos que eran el aminoácido más común entre las seis secuencias disponibles del dentista, pero atípico en las Secuencias de antecedentes (background) (una firma mayoritaria). La segunda línea muestra las frecuencias de estos diez aminoácidos entre las secuencias virales del dentista y la tercera línea muestra sus frecuencias entre las 32 Secuencias de antecedentes (background) de la base de datos. Solo ocho de los diez sitios cumplieron con el criterio de una firma estricta, es decir, se conservaron perfectamente en todas las secuencias virales del dentista.

De esta forma se calculó la frecuencia de nucleótidos en cada posición de dos conjuntos de secuencias en una alineación (conjunto de consulta y conjunto de antecedentes (background)) (Ou et al., 1992; Singh et al., 2021). Las secuencias de consulta fue la base de datos de la región RTL 3' analizadas previamente de la matriz Los Alamos, y las secuenciadas por los investigadores del ICGES; mientras que, la Secuencia de antecedentes (background) fue la HXB2 considerada el genoma de referencia del virus de la inmunodeficiencia humana.

Se obtuvieron tres logos de referencia y las frecuencias de los PNC. Además, se hizo un análisis entre las dos bases de datos a estudiar, se colocó como secuencia de consulta la base de datos de secuencias panameñas y como Secuencia de antecedentes (background) la base de datos de secuencias del Laboratorio de los Alamos, para obtener el tercer logo y conjunto de datos de frecuencia nucleotídica. En total se realizaron tres experimentos de análisis de frecuencia nucleotídica por posición en VESPA.

- ✓ Experimento A: Como conjunto de consulta la Matriz Panamá y como Secuencia de antecedentes (background) HXB2.
- ✓ Experimento B: Como conjunto de consulta la Matriz Los Alamos y como Secuencia de antecedentes (background) HXB2.
- ✓ Experimento C: Como conjunto de consulta la Matriz Panamá y como Secuencia de antecedentes (background) la Matriz Los Alamos.

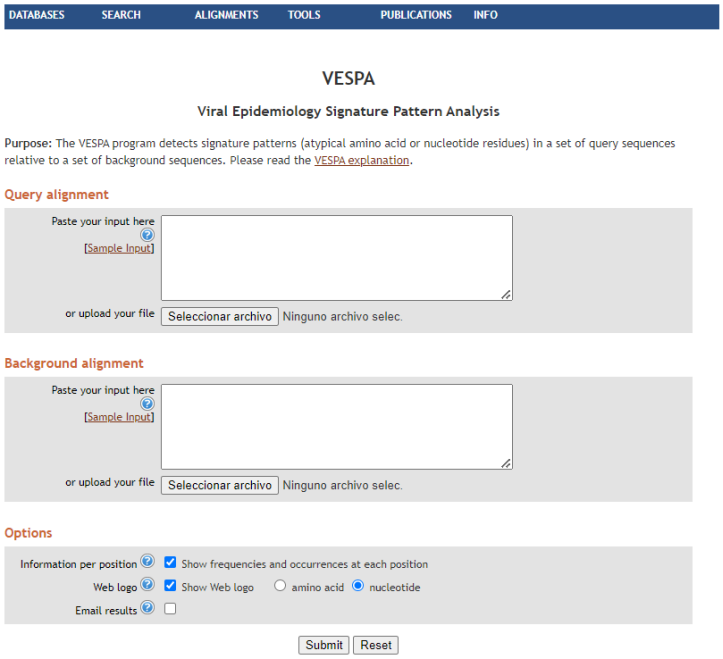


Figura 10. Programa en línea VESPA para análisis de diferencia nucleotídica.

Análisis de transversiones y transiciones en nucleótidos. Por medio de programa en línea de la base de datos del Laboratorio de Los Alamos (https://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html) se hizo un análisis que destaca las transiciones y transversiones nucleotídicas. Para el primer análisis se utilizó como secuencia de consulta la base de datos de secuencia panameñas y como Secuencia de antecedentes (background) la HXB2. Luego, para el segundo análisis se utilizó la secuencia de consulta, la base de datos del Laboratorio de los Alamos y como Secuencia de antecedentes (background) la HXB2.

2.7. Determinación de los sitios de unión de los factores de transcripción y su capacidad de anclaje a los factores de transcripción

Por medio del programa CiiiDER se realizó un análisis de enriquecimiento para identificar los SFT que están significativamente sobrerrepresentados o subrepresentados en comparación con un conjunto de secuencias previamente registradas (antecedente) (Gearing et al., 2019). Luego se introdujeron todos los SFT analizados de los conjuntos de secuencias de la región RTL 3' de la base de datos de secuencias de panameñas y de las bases de datos del Laboratorio de Los Alamos con las matrices de los cinco factores de transcripción a analizar. Para realizar el análisis de enriquecimiento se introdujeron las Secuencias de antecedentes (background) de los sitios de unión de los factores de transcripción previamente reportados en literatura.

Se generó una gráfica que muestra el límite de enriquecimiento (relación de proporción) y el límite de proporción logarítmica promedio que permite observar la proporción de regiones vinculadas para cada FT, y la puntuación de significancia; si el FT es mayor que cero (>0) está sobrerrepresentado y si es menor que cero (<0) está subrepresentado.

Para las gráficas de enriquecimiento, si un factor de transcripción dado tiene sitios de unión en n_S fuera de las regiones de búsqueda de N_S y n_B fuera de las regiones de antecedentes (background) de N_B, entonces:

$$Average.Log2.Proportion.Bound = \frac{1}{2}\log_2\left(\frac{n_S + 1/2}{N_S + 1/2}\right) + \frac{1}{2}\log_2\left(\frac{n_B + 1/2}{N_B + 1/2}\right)$$

$$Log2.Enrichment = \log_2\left(\frac{n_S + 1/2}{N_S + 1/2}\right) - \log_2\left(\frac{n_B + 1/2}{N_B + 1/2}\right)$$

$$Significance.Score = -\text{sign}(Log2.Enrichment) \times \log_{10}(P\text{−value})$$

Figura 11. Fórmula matemática para el cálculo del valor de significancia del anclaje (VSA)

CiiiDER utiliza el promedio logarítmico (log2) de la proporción de unión del FT al SFT con el enriquecimiento del SFT (log2), para obtener la puntuación de significancia de unión del FT al SFT (log10). Además, Ciiider genera una plot con un mapa de calor que representa el VSA. El VSA se utilizó como valor determinante de la calidad, capacidad y significancia de unión de los SFT al FT.

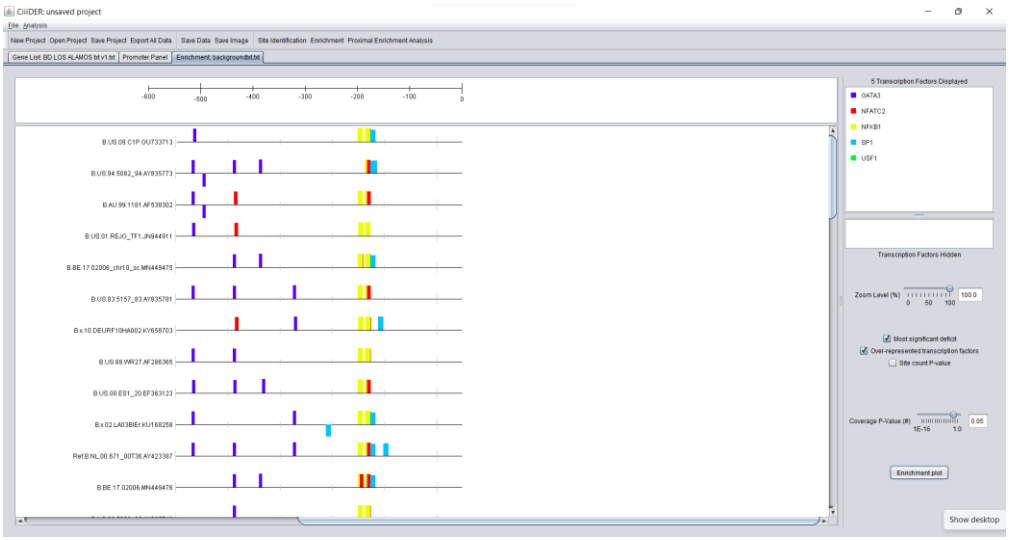


Figura 12. Programa bioinformático CiiiDER para análisis de identificación de SFT.

2.8. Análisis estadístico

Estadística descriptiva, distribución y normalidad. Para el desarrollo de análisis estadísticos descriptivos, se determinó el tipo de variable de los datos encontrados en la Tabla 6. Luego, se utilizó el programa Jamovi (Versión 1.8) obtenido de https://www.jamovi.org, para el análisis de datos. Se implementó el análisis de Skewness para conocer la distribución de los datos y el análisis de Shapiro-Wilk para conocer la normalidad de los datos (Altman & Bland, 1996; Fisher & Marshall, 2009; Love J, 2022).

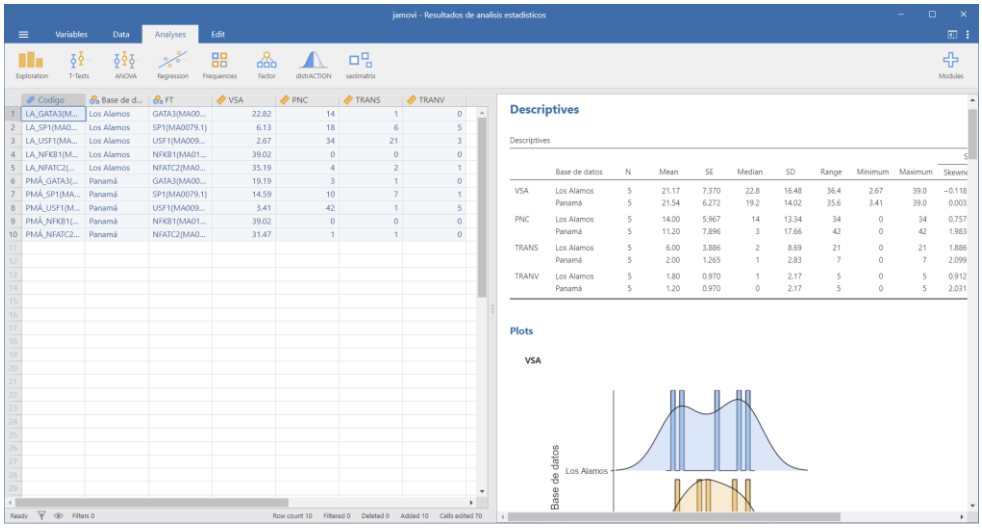


Figura 13. Programa para análisis estadísticos Jamovi

Análisis de matriz de correlación. Se instaló un conjunto de paquetes estadísticos en Jamovi para el análisis de correlación y creación de gráficos (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012; Revelle, 2019; Seol, 2022). Se realizó un análisis de matriz de correlación de Spearman, para datos no paramétricos, donde:

a) HI: La capacidad de anclaje de los factores de transcripción a los sitios de unión depende de su variabilidad genética en esa región.
b) Ha: La correlación entre la capacidad de anclaje (VSA) y el número de secuencias con PNC (diversidad polimórfica) es negativa.
c) H0: La correlación entre la capacidad de anclaje (VSA) y el número de secuencias con PNC (diversidad polimórfica) es positiva.

Luego, se generó una gráfica de la matriz de correlación con la densidad de las variables y su estadística. Además, se realizó una grafica de modelo Gausiano (EBIC). Se interpretaron los resultados en base a cada análisis y sus principios estadísticos (Bulmer, 1979).

CAPÍTULO III:

RESULTADOS

3.1. Búsqueda de secuencias nucleotídicas en Base de datos de Los Álamos

Para la construcción de la matriz de secuencias nucleotídicas de VIH-1 del laboratorio de los Álamos, se analizaron un total de 4,537 secuencias de genoma completo de alineaciones super filtradas. Se eligieron 51 secuencias del subtipo B con 100% de la RTL 3' según la HXB2 y se añadieron al análisis las 4 secuencias de referencia del subtipo B para obtener un total de 55 secuencias de la base de datos del laboratorio de Los Alamos.

La matriz del Laboratorio de los Álamos se constituyó por un 61.8% de secuencias de Estados Unidos de América entre los años 1983-2009, y 32.7% de secuencias pertenecientes a Francia (1983-2003), Holanda (2000), Tailandia (1990), Argentina (2007), Australia (1986-1999), Bélgica (2017), China (2007), España (1989), Japón (Sin información de los años de toma de muestra), Filipinas (2015) y Taiwán (1994) (Figura 14). El 5.5% restante pertenece a las secuencias sin información sobre el país de toma de muestra entre los años 2000-2010.

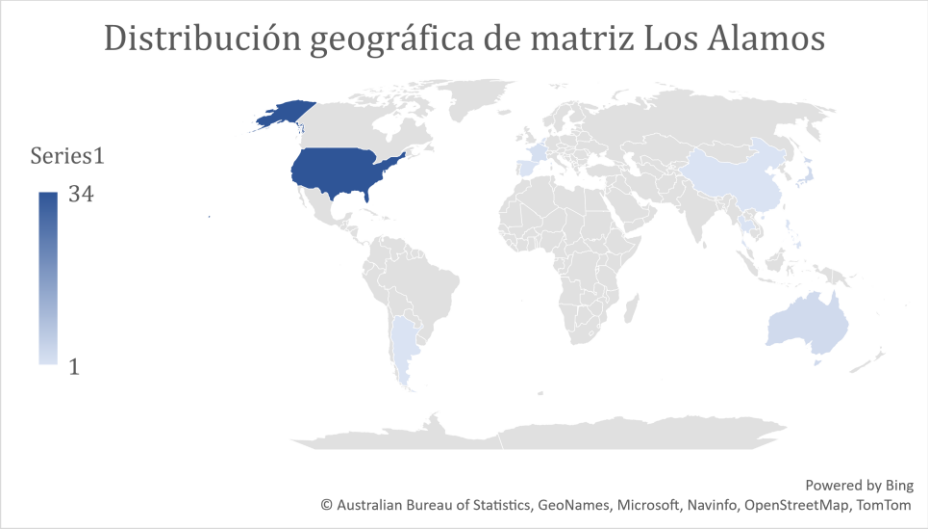


Figura 14. Distribución geográfica de matriz Los Álamos.

Tabla 2. Información de secuencias de la matriz Los Álamos

SUBTIPO	PAIS	AÑO	CÓDIGO DE MUESTRA	GENBANK
B	ARGENTINA	2007	DEURF07AR001	KY658686
B	AUSTRALIA	1999	1181	AF538302
B	AUSTRALIA	1987	MBC925	AF042101
B	AUSTRALIA	1986	MBC200	AF042100
B	BÉLGICA	2017	2006	MN449476
B	BÉLGICA	2017	02006 cen	MN449474
B	BÉLGICA	2017	02006 chr10 sc	MN449475
B	CHINA	1999	plwj	GU177863
B	ESPAÑA	1989	U61	DQ854716
B	ESTADOS UNIDOS DE AMERICA	2009	C1P	GU733713
B	ESTADOS UNIDOS DE AMERICA	2006	CH106 TF1	JN944897
B	ESTADOS UNIDOS DE AMERICA	2006	MDR 5a	KF990608
B	ESTADOS UNIDOS DE AMERICA	2005	MDR 1c	KF990605
B	ESTADOS UNIDOS DE AMERICA	2004	ES10 53	EF363127
B	ESTADOS UNIDOS DE AMERICA	2004	ES4 24	EF363124
B	ESTADOS UNIDOS DE AMERICA	2004	ES8 43	EF363126
B	ESTADOS UNIDOS DE AMERICA	2001	REJO TF1	JN944911
B	ESTADOS UNIDOS DE AMERICA	2000	ES1 20	EF363123
B	ESTADOS UNIDOS DE AMERICA	2000	RHPA TF1	JN944917
B	ESTADOS UNIDOS DE AMERICA	2000	THRO TF1	JN944930
B	ESTADOS UNIDOS DE AMERICA	1998	1058_11	AY331295
B	ESTADOS UNIDOS DE AMERICA	1996	5155 96	AY835753
B	ESTADOS UNIDOS DE AMERICA	1995	5073 95	AY835768
B	ESTADOS UNIDOS DE AMERICA	1994	5082 94	AY835773
B	ESTADOS UNIDOS DE AMERICA	1991	5048 91	AY835761
B	ESTADOS UNIDOS DE AMERICA	1991	DH12 3	AF069140
B	ESTADOS UNIDOS DE AMERICA	1990	WEAU160 GHOSH	U21135
B	ESTADOS UNIDOS DE AMERICA	1989	P896 89 6	U39362
B	ESTADOS UNIDOS DE AMERICA	1988	5160 88	AY835763
B	ESTADOS UNIDOS DE AMERICA	1988	WR27	AF286365
B	ESTADOS UNIDOS DE AMERICA	1987	5113 87	AY835758
B	ESTADOS UNIDOS DE AMERICA	1986	5084 86	AY835775
B	ESTADOS UNIDOS DE AMERICA	1986	5096 86	AY835749
B	ESTADOS UNIDOS DE AMERICA	1986	5127 86	AY835774
B	ESTADOS UNIDOS DE AMERICA	1986	AD87 ADA	AF004394
B	ESTADOS UNIDOS DE AMERICA	1986	YU 2	M93258
B	ESTADOS UNIDOS DE AMERICA	1985	5077 85	AY835769
B	ESTADOS UNIDOS DE AMERICA	1984	5019 84	AY835779
B	ESTADOS UNIDOS DE AMERICA	1984	MNCG MN	M17449
B	ESTADOS UNIDOS DE AMERICA	1984	SF33	AY352275
B	ESTADOS UNIDOS DE AMERICA	1983	5018 83	AY835777

B	ESTADOS UNIDOS DE AMERICA	1983	5157 83	AY835781
B	ESTADOS UNIDOS DE AMERICA	1983	SF2 LAV2 ARV2	K02007
B	FRANCIA	2003	LA06ToXa	KU168261
B	FRANCIA	1983	HXB2_LAI_IIIB_BRU	K03455
B	HOLANDA	2000	671_00T36	AY423387
B	JAPÓN	SIN INFO	JRC03B	AB565496
B	JAPÓN	SIN INFO	JRC05B	AB565497
B	JAPÓN	SIN INFO	JRC65B	AB565502
B	FILIPINOS	2015	DEMB15PH003	KY658690
B	TAILANDIA	1990	BK132	AY173951
B	TAIWAN	1994	TWCYS LM49	AF086817
B	SIN INFO	2000	LA02FolC	KU168257
B	SIN INFO	2002	LA03BlEr	KU168258
B	SIN INFO	2010	DEURF10HA002	KY658703

3.2. Análisis filogenético

Análisis filogenético con REGA. Análisis con el programa bioinformático para VIH, REGA, identifico 48 secuencias de 55 como subtipo B puro de VIH-1, 5 secuencias como subtipo semejante al B de VIH-1, y 2 secuencias como subtipo BD del VIH-1. La secuencia FID16-103-021 identificada como recombinante BD según REGA, comienza en la posición 335 y termina en la posición 9181 en relación con la secuencia de referencia NC_001802.1 para el virus de la inmunodeficiencia humana 1 (taxón: 11676) según el análisis de REGA (Figura 15, parte A). Además, el programa REGA realizo un Bootscan con un soporte de clúster: 0.873, tamaño de ventana 400 y tamaño de paso 50 (Figura 15, parte B). El análisis filogenético con el árbol guía, clasifico esta muestra como VIH-1 Subtipo B con un valor de Bootstrap de 99% (figura 17). En la figura 15 y 16 parte A, se observa en la parte superior el patrón de recombinación aproximado de cada secuencia, con un método de arranque mayor

a 70% y en la parte inferior el patrón de recombinación aproximado sin el método de arranque.

De igual forma, la secuencia FID16-103-023, identificada como recombinante según REGA, comienza en la posición 335 y termina en la posición 9180 en relación con la secuencia de referencia NC_001802.1 para el virus de la inmunodeficiencia humana 1 (taxón: 11676) según el análisis de REGA (Figura 16, parte A). Además, el programa REGA realizo un Bootscan con un soporte de clúster de 0.872, tamaño de ventana 400 y tamaño de paso 50 (Figura 16, parte B). El análisis filogenético con el árbol guía, clasifico esta muestra como VIH-1 Subtipo B con un valor de Bootstrap de 99%.

Análisis filogenéticos en MEGA7. Por medio de los análisis filogenéticos realizados en la matriz de las secuencias panameñas con el programa MEGA, se determinó que las 55 secuencias de la matriz de Panamá pertenecían al subtipo B. Ninguna de las secuencias resulto ser un recombinante. En la figura 17 se muestra el árbol óptimo que involucró 100 secuencias de nucleótidos pertenecientes a las 43 referencias de subtipos, 3 secuencias de outgroup y 55 secuencias de la matriz de Panamá. Hubo un total de 4,828 posiciones en el conjunto de datos final.

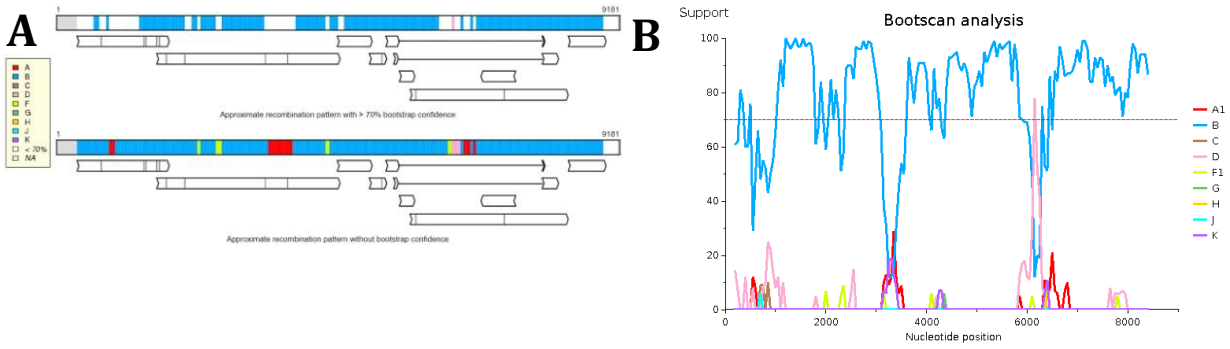


Figura 15. Mapa genético de la muestra FID106-103-021 (A) Mapa genético de la secuencia (B) Análisis de escáner de arranque de la secuencia.

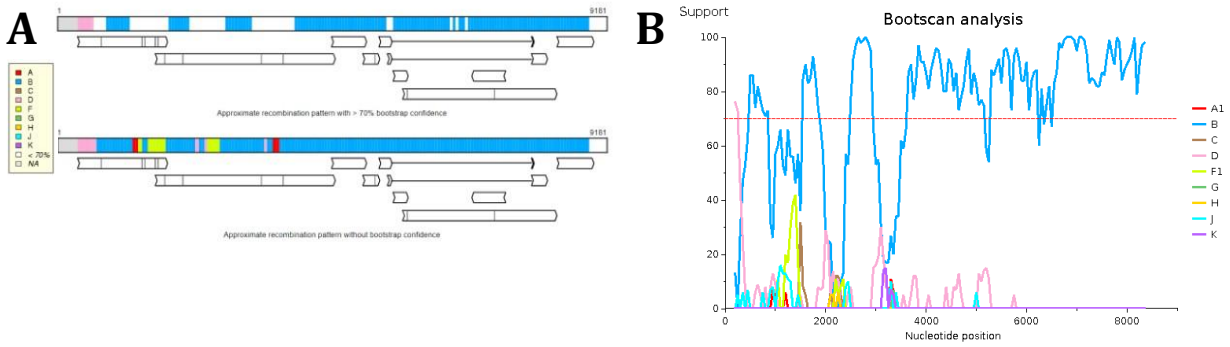


Figura 16. Mapa genético de la muestra FID106-103-023. (A) Mapa genético de la secuencia. (B) Análisis de escáner de arranque de la secuencia.

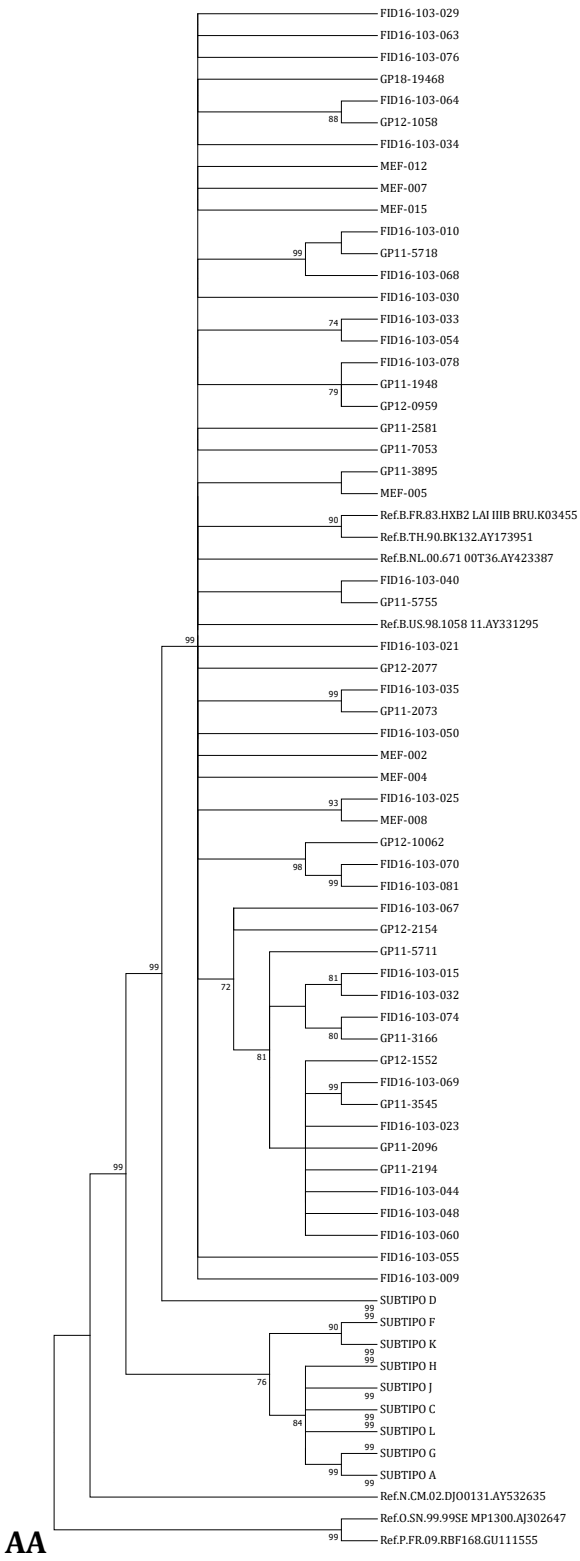


Figura 17. Árbol de las relaciones evolutivas de los taxones de matriz de secuencias panameñas.

3.3. Identificación de los sitios de unión de los factores de transcripción de la RTL

Se identificaron 5 sitios de unión (SFT) para cada factor de transcripción en cada secuencia (5 TF x 5 SFT x 110 secuencias nucleotídicas), obteniendo un total de 2,750 posiciones de sitios de unión. Se seleccionó la posición con mayor puntuación de cada SFT por secuencia, obteniendo un total de 550 posiciones para los 5 factores de transcripción del estudio. Todas las posiciones de la matriz tienen un *valor de p < 0.0246* (Tabla 3).

Se identificaron 9 posiciones diferentes para GATA3, 11 posiciones diferentes para SP1, 6 posiciones diferentes para USF1, 3 posiciones diferentes para NFKB1 y 10 posiciones diferentes del factor NFATC2 para la matriz de secuencias nucleotídicas de Panamá. Además, se identificaron 5 posiciones diferentes para el factor de transcripción GATA3, 16 posiciones diferentes para el factor SP1, 8 posiciones diferentes para el factor USF1, 3 posiciones diferentes para NFKB1, 10 posiciones diferentes para NFATC2 en la matriz de los Álamos (Tabla 3). Cada posición en su gran mayoría contaba con al menos dos variaciones del SFT por posición. Por ejemplo, el SFT GATA-3 para la posición 76 en ambas matrices, presento la secuencia nucleotídica *(T/A)GATTG* con un cambio nucleotídico en la posición 1 de T por A y viceversa.

Tabla 3. Identificación de los SFT con PNC obtenidos de LASAGNA y VESPA

Factor de transcripción	Sitios de unión de factores de transcripción	Posición del alineamiento en matriz de Panamá	Posición del alineamiento en matriz de Los Álamos	*Valor de p*	Valor de e
GATA3(MA0037.1)	**CGATAA**		210, (1)	*0,0005*	0,272
	AGATAG		160, (1)	*0,00175*	0,95
	AGATAG	165, (1)	165, (12)	*0,00175*	0,95
	(T/A)GATTG	76, (43)	76, (38)	*0,003725/0,004425*	2,03/2,71

	CGAGAG		296, (3)	*0,007975*	4,3
	AGATAT	31, (1)		*0,0131*	7,1
	AGATAC	110, (2)		*0,006*	3,3
	CGATCG	154, (1)		*0,0037*	2,01
	(C/A)GATAA	160, (2)		*0,0005/0,0082*	0,272/4,5
	AGATAA	187, (1)		*0,0082*	4,5
	TGATAG	265, (2)		*0,000975*	0,53
	CGAGAG	296, (2)		*0,007975*	4,3
SP1(MA0079.1)	**A(T/A)GGG(A/T)TGAT**	3, (2)		*0,00645/0,01095*	3,5/5,9
	AAG(G/A)CA(A/G)GAT	25, (2)	25 (2)	*0,003825/0,00605*	2,07/3,3
	AAGACATTCT	30, (1)		*0,006025*	3,3
	(T/C)TGGCAGAAT	79, (4)		*0,015/0,018*	8,1/9,7
	AAGCCACTGA	174, (1)		*0,007475*	4
	CC(T/A)GCATGGA	217, (2)	217 (8)	*0,017825*	9,6
	ATGGG(A/G)TGGA	222 (15)		*0,018125/0,005*	9,8/2,7
	GAGGCGCGGT	376, (1)		*0,00055*	0,297
	GAGGCGTAAT	376, (1)		*0,01365*	7,4
	GAGGCGTGGA	376, (18)	376, (3)	*0,0011*	0,59
	GAGGCGTGGT	376, (2)		*0,000425*	0,229
	GGGGCGTGAT	376, (1)		*0,00325*	1,75
	GAGCCTGGGA	478, (2)		*0,02435*	13,1
*	**GCCAGCCCTC**	406, (2)		*0.0264*	14.3
*	**TTCAAGCTAG**	137, (1)		*0.04375*	23.6
	AAGGG(T/A)TGAT		3, (3)	*0,00645/0,00335*	3,5/1,81
	AAGGCAAGAA		25 (1)	*0,005875*	3,2
	TTGGCAGT(A/G)T		79, (3)	*0,015/0,000525*	8,1/0,283
	AGGCCACTGT		174, (1)	*0,0145*	7,8
	A(T/C)GGGATGGA		222, (19)	*0,018125/0,009575*	9,8/5,2
	AAGACTCGGA		231, (1)	*0,01515*	8,2
	ACAGCAGGCT		267, (1)	*0,012525*	6,8
	TCAGCATTAT		283, (1)	*0,009925*	5,4
	AAGGCTGCAT		298, (1)	*0,0075*	4
	TGGGCTTTCT		335, (1)	*0,0012*	0,65
	GAGGCGTGAT		376, (1)	*0,000475*	0,257
	GAGGTGTGGT		376, (2)	*0,007975*	4,3
	TGGGCGGGGT		387, (1)	*0,00195*	1,05
	GAGCCTGGGA		478, (5)	*0,02435*	13,1
	TGGGCGCTCT		483, (1)	*0,00295*	1,59
USF1(MA0093.1)	**CACAAGG**	60, (4)	60, (1)	*0,0246*	13,4
	CACTTGA	150, (1)		*0,003125*	1,7
	CATGGGA	221, (8)	221, (6)	*0,0118*	6,4
	CATGTGG	253, (1)		*0,0008*	0,43
	ATGTGG	254, (1)	254, (1)	*0,01735*	9,4
	CACGTGG	287, (16)	287, (14)	*0,000325*	0,176

	CACAGGG	287, (1)		*0,01615*	8,8
	CACGTAG	287, (10)	287, (4)	*0,003975*	2,16
	CACTTGG	287, (3)	287, (1)	*0,003975*	2,16
	CACATGG	287, (7)	287, (20)	*0,001*	0,54
	CATGTGG	287, (3)	287, (3)	*0,0008*	0,43
	CAGGTGG		164, (1)	*0,005725*	3,11
	CACGGGA		221, (2)	*0,001275*	0,69
	CACCTGG		287, (1)	*0,003975*	2,16
	CAGGTGG		372, (1)	*0,005725*	3,11
NFATC2(MA0152.1)	**TATTCCA**	14, (1)		*0,00215*	1,17
	GGTTCCC	111, (1)	111, (1)	*0,001825*	0,99
	ATATCCA	112, (6)	112, (3)	*0,0038*	2,06
	ATTTCCA	112, (13)	112, (8)	*0,001325*	0,72
	GTATCCA	112, (2)	112, (2)	*0,00165*	0,9
	(G/T)GATCCA	154, (5)	154, (11)	*0,001575/0,001675*	0,86/0,91
	TGTTACA	201, (16)	201, (22)	*0,002175*	1,18
	TGATCCA	232, (2)	232, (1)	*0,001675*	0,91
	TGTTACA	249, (1)	249, (1)	*0,002175*	1,18
	CTTTCCA	279, (1)		*0,00435*	2,36
	CTTTCCA	353, (1)		*0,00435*	2,36
	CTTTCCA	367, (6)		*0,00435*	2,36
	GGATCCA		48, (1)	*0,001575*	0,86
	GTTTACA		50, (2)	*0,00215*	1,17
	TGTTACA		86, (1)	*0,002175*	1,18
	GTTTCCA		112, (1)	*0,000125*	0,068
	TTTTACA		314, (1)	*0,00385*	2,09
NFKB1(MA0105.1)	**GGGACTTTCCA**	349, (6)	349, (6)	*0,000475*	0,256
	GGGGACTTTCC	362, (2)	362, (1)	*0,00095*	0,51
	GGG(A/G)CTTTCC(A/G)	363, (47)		*0,000475/0,0008*	0,256/0,43
	GGGACTTTCC(A/G)		363, (48)	*0,000475/0,001275*	0,256/0,69

Posición, (Cantidad de secuencias con esa posición)
Morado; PNC existentes en el experimento A y B
Amarillo; PNC existentes en el experimento B y C
Verde; PNC existentes en el experimento A y C

3.4. Determinación de la diversidad genética de los sitios de unión de los factores de transcripción de las RTL del VIH-, grupo M.

Análisis de frecuencia nucleotídica. VESPA calculó la frecuencia de cada nucleótido en cada posición (columna) de la alineación para el conjunto de consulta y antecedentes (background). Se mostraron las posiciones en las que el carácter más común en el conjunto de consulta difiere del conjunto de antecedentes (background). Para el experimento A las posiciones en la alineación que difieren o los PNC son la posición 15, 23, 25, 51, 82, 108, 109, 139, 164, 168, 183, 196, 213, 220, 239, 244, 256, 262, 319, 324, 335, 343, 421, 501 (Figura 18, parte A). Para el experimento B se mostraron las siguientes posiciones con PNC; 15, 25, 51, 82, 108, 139, 164, 168, 183, 196, 198, 213, 220, 256, 262, 291, 319, 343, 421 (Figura 18, parte B). El ultimo análisis, experimento C, resulto que las posiciones 23, 109, 198, 239, 244, 291, 324, 335, 501 son PNC entre estos dos conjuntos de bases de datos (Figura 18, parte C).

Figura 18. Análisis de los PNC en VESPA con diferentes matrices. Logo de PNC del experimento A. Logo de PNC del experimento B. Logo de PNC del experimento C.

La tabla 4 muestra las secuencias nucleotídicas de los SFT (Columna 2), la posición de los SFT según el alineamiento de la RTL 3' por matriz (Columna 3 y 4), la posición del PNC (Columna 5), el análisis de transversión o transición (Columna 6) y las frecuencias polimórficas de los PNC (Columna 7 al 12). En las columnas de frecuencias nucleotídicas de PNC se colocaron dos tipos de frecuencias por experimento, la frecuencia del nucleótido más frecuente en la consulta y la frecuencia del nucleótido más común del antecedentes (background) en la consulta. En el experimento A, para GATA-3 se observa que adenina (A) es el nucleótido más frecuente en la matriz de Panamá para el PNC 213 y que guanina (G) que es el nucleótido más común para el antecedente (background) tiene una frecuencia de 0.073 en la matriz.

Tabla 4. PNC existentes en los SFT y su frecuencia nucleotídica según VESPA

						Experimento A		Experimento B		Experimento C	
Factor de transcripción	Sitios de unión de factores de transcripción	Posición del SFT en la matriz de Panamá basada en 0	Posición del SFT en matriz de Los Alamos basada en 0	Posición del alineamiento del PNC	PNC-Transición / PNC-Transversión segun la HXB2	Matriz Panamá (consulta)	secuencia HXB2 (fondo)	Matriz Los Alamos (consulta)	secuencia HXB2 (fondo)	Matriz Panamá (consulta)	Matriz Los Alamos (fondo)
GATA3(MA0037.1)	**CGATAA**		210, (1)	213	Transición			0.836A	0.073G		
	AGATAG		160, (1)	164				0.655G	0.200T		
	AGATAG	165, (1)	165, (12)	168		0.891G	0.018A	0.727G	0.255A		
	(C/A)GATAA	160, (2)		164		0.6G	0.182T				
SP1(MA0079.1)	**(T/C)TGGCAGAAT**	79, (4)		82	Transición	1.000G	0.000A				
	AAGCCACTGA	174, (1)		183	Transversión	0.782G	0.073A				
	CCTGCATGGA	217, (1)	217 (4)	220		0.509A	0.455T	0.6A	0.4T		
	CCAGCATGGA	217, (1)	217 (4)	220	Transversión	0.509A	0.455T	0.6A	0.4T		
	TTCAAGCTAG	137, (1)		139	Transversión	0.964T	0.018A				
	TTGGCAGT(A/G)T		79, (3)	82	Transición			0.982G	0.018A		
	AGGCCACTGT		174, (1)	183	Transición			0.818G	0.145A		
	AAGACTCGGA		231, (1)	239						0.455A	0.436G
	TCAGCATTAT		283, (1)	291	Transversión			0.473A	0.455G	0.564G	0.273A
	TGGGCTTTCT		335, (1)	343	Transversión			0.964T	0.036G		
USF1(MA0093.1)	**CATGTGG**	253, (1)		256	Transversión	0.782T	0.182A				
	ATGTGG	254, (1)	254, (1)	256	Transversión	0.782T	0.182A	0.673T	0.291A		
	CACGTGG	287, (16)	287, (14)	291	Transición			0.473A	0.455G	0.564G	0.273A

	CACAGGG	287, (1)		291						0.564G	0.273 A
	CACGTAG	287, (10)	287, (4)	291				0.473A	0.455G	0.564G	0.273 A
	CACTTGG	287, (3)	287, (1)	291	Transversión			0.473A	0.455G	0.564G	0.273 A
	CACATGG	287, (7)	287, (20)	291	Transición			0.473A	0.455G	0.564G	0.273 A
	CATGTGG	287, (3)	287, (3)	291				0.473A	0.455G	0.564G	0.273 A
	CAGGTGG		164, (1)	168	Transición			0.727G	0.255A		
NFATC2(MA0152.1)	**TATTCCA**	14, (1)		15	Transición	0.909T	0.091C				
	GGATCCA		48, (1)	51				0.709G	0.291A		
	GTTTACA		50, (2)	51	Transición			0.709G	0.291A		
	TTTTACA		314, (1)	319	Transversión			0.836A	0.164T		

Posición, (Cantidad de secuencias con esa posición)

Morado; PNC existentes en el experimento A y B

Amarillo; PNC existentes en el experimento B y C

* SFT con dos PNC en la misma secuencia

Análisis de transiciones y transversiones. La herramienta HIGHLIGHTER del Laboratorio de los Álamos comparó las secuencias de consulta, matriz de Panamá (Figura 19, parte A) y matriz de Los Álamos (Figura 19, parte B) con una única secuencia maestra (HXB2). Se destaco o resalto las transiciones y transversiones de ambas bases de datos. Se determinaron 8 transversiones y 32 transversiones en los SFT para la matriz Panamá y 5 transversiones y 42 transiciones en los SFT para la matriz Los Álamos. Las imágenes ilustran las transiciones y transversiones con el siguiente código de colores:

= Transiciones = Transversiones

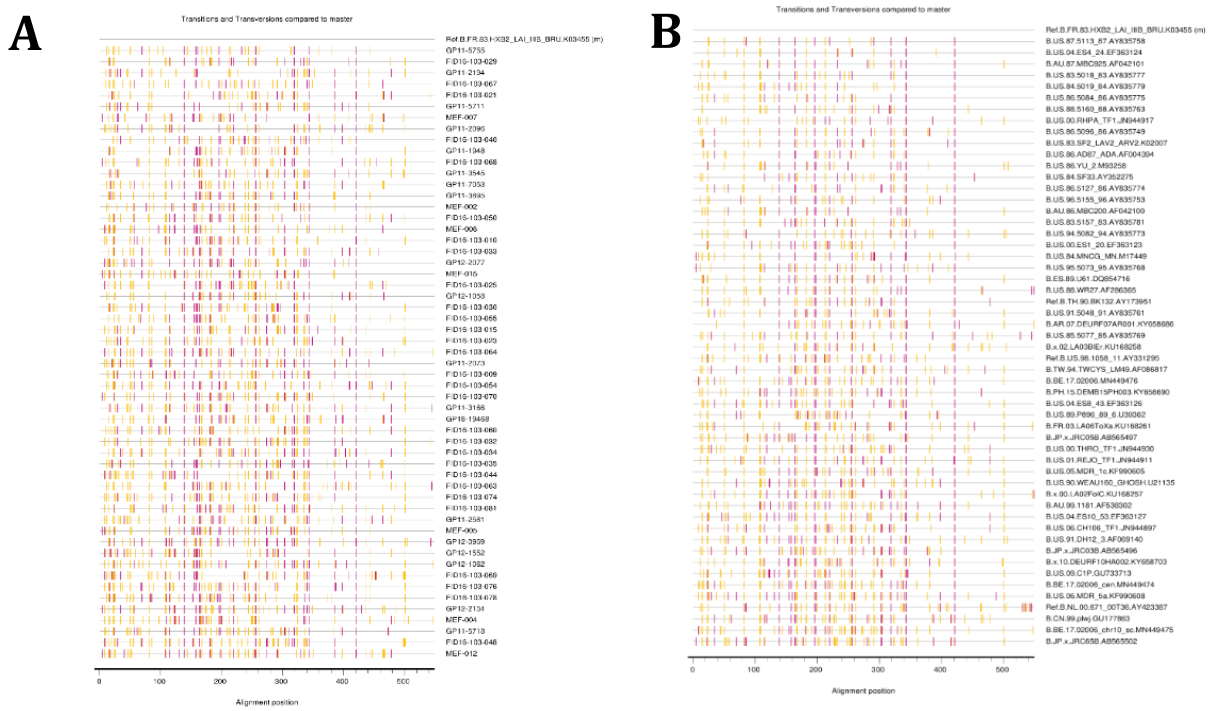


Figura 19. Transiciones y transversiones de la Matriz Panamá con la HXB2 (A) y Matriz Los Álamos con la HXB2 (B).

Tabla 5. Descripción de secuencias de la matriz de Panamá según resultados de VESPA y HIGHLIGHTER

Información de análisis bioinformaticos					Información del paciente					
Factor de transcripción	Sitios de unión de factores de transcripción	Posición del alineamiento del PNC	Secuencias de la base de datos de Panamá	PNC-Transición/ PNC-Transversión segun la HXB2	Sexo	Edad	Estado Clínico	Esquema de tratamiento resumido	Valor de carga viral	Valor de CD4+
GATA3(MA0037.1)	**AGATAG**	168	GP11-5711		M	58	Ninguno	AZT + 3TC + EFV	424138	358
	(C/A)GATAA	164	FID16-103-050		M	43	SIDA	DRV/r + RAL	41219	X
			GP11-5755		F	45	Asintomático	AZT + 3TC + EFV	142511	154
SP1(MA0079.1)	**(T/C)TGGCAGAAT**	82	FID16-103-040	Transición	M	27	No especificado	AZT + 3TC + LPV/r + RAL	4940	304
			FID16-103-070		M	46	No especificado	3TC + ABC + RAL	525700	82
			GP18-19468		M	52	Asintomático	EFV + LPV/r + RAL	18522	X
			MEF-015		F	45	Asintomático	Naive	49327	421
	AAGCCACTGA	183	FID16-103-063	Transversión	M	36	No especificado	3TC + LPV/r + RAL	122916	280
	CCTGCATGGA	220	FID16-103-054		M	50	No especificado	FTC + TDF + LPV/r	233006	111
	CCAGCATGGA		MEF-004	Transversión	M	33	Asintomático	Naive	145118	376
	TTCAAGCTAG	139	GP11-2073	Transversión	M	47	Asintomático	Sin Tratamiento	81504	493
USF1(MA0093.1)	**CATGTGG**	256	MEF-012	Transversión	M	35	Asintomático	Naive	345034	88
	ATGTGG	256	FID16-103-076	Transversión	F	26	Asintomático	FTC + TDF + RAL	71363	60
	CACGTGG	291	FID16-103-009	Transición	F	32	SIDA	FTC + TDF + DRV/r + RAL	46627	X
			FID16-103-010		X	X	X	X	X	X
			FID16-103-050		M	43	SIDA	DRV/r + RAL	41219	X
			FID16-103-063		M	36	No especificado	3TC + LPV/r + RAL	122916	280
			FID16-103-068		F	42	No especificado	FTC + TDF + RAL	90801	92
			FID16-103-023		F	41	No especificado	TDF + LPV/r + RAL	72722	126
			FID16-103-033		M	26	No especificado	3TC + LPV/r + RAL	324055	X
			FID16-103-054		M	50	No especificado	FTC + TDF + LPV/r	233006	111
			GP11-2194		F	38	Asintomático	3TC + ABC + EFV	20840	X
			GP11-2581		M	30	SIDA	AZT + 3TC + EFV	142666	332
			GP11-3895		F	41	Asintomático	TDF + FTC + EFV	38656	298
			GP11-5718		F	32	SIDA	AZT + 3TC + EFV	349800	50

			GP11-5755		F	45	Asintomático	AZT + 3TC + EFV	142511	154
			GP11-7053		F	49	Asintomático	AZT + 3TC + EFV	54057	718
			GP18-19468		M	52	Asintomático	EFV + LPV/r + RAL	18522	X
			MEF-004		M	33	Asintomático	Naive	145118	376
	CACAGGG	291	MEF-005		M	26	No especificado	Naive	36620	25
	CACGTAG	291	FID16-103-040		M	27	No especificado	AZT + 3TC + LPV/r + RAL	4940	304
			FID16-103-055		F	40	No especificado	AZT + 3TC + DRV/r + RAL	148538	329
			FID16-103-044		M	47	No especificado	TDF + LPV/r + RAL	6376	175
			GP11-1948		M	29	Asintomático	TDF + FTC + EFV	52288	15
			GP11-3166		M	47	Asintomático	AZT + 3TC + EFV	134636	34
			GP11-3545		F	38	Asintomático	AZT + 3TC + EFV	196273	169
			GP11-5711		M	58	Ninguno	AZT + 3TC + EFV	424138	358
			GP12-10062		M	30	Asintomático	Sin Tratamiento	3592037	673
			MEF-002		M	22	Asintomático	Naive	15751	732
			MEF-008		M	31	No especificado	Naive	232764	89
	CACTTGG	291	FID16-103-021	Transversión	M	24	No especificado	3TC + ABC + RAL	41114	510
			GP12-0959		F	32	Asintomático	AZT + 3TC + EFV	39082	14
			MEF-015		F	45	Asintomático	Naive	49327	421
	CACATGG	291	FID106-103-025	Transición	M	43	No especificado	3TC + ABC + LPV/r + RAL	3520	246
			FID106-103-034		M	45	SIDA	AZT + 3TC + ABC + RAL	19125	62
			FID106-103-081		M	46	No especificado	AZT + 3TC + RAL	947	36
			FID16-103-029		M	58	No especificado	LPV/r + RAL	53376	148
			FID16-103-067		M	32	No especificado		36353	280
			GP11-2073		M	47	Asintomático	Sin Tratamiento	81504	493
			GP12-2154		M	35	No especificado	AZT + 3TC + EFV	119213	13
	CATGTGG	291	FID106-103-069		F	18	No especificado	FTC + TDF + LPV/r + RAL	56	161
			FID106-103-078		F	43	No especificado	FTC + TDF + LPV/r + RAL	3502	269
			GP12-1058		F	50	Ninguno	AZT + 3TC + EFV	28775	508
NFATC2(MA0152.1)	**TATTCCA**	15	FID106-103-035	Transición	F	34	No especificado	TDF + LPV/r + RAL	890	368

3.5. Determinación de los polimorfismos de los sitios de unión de los factores de transcripción y su capacidad de anclaje a los factores de transcripción.

El análisis de enriquecimiento del programa CiiiDER comparó la distribución de los SFT predichos en un conjunto de regiones reguladoras de ADN con la distribución en un conjunto de Secuencias de antecedentes (background) (SFT previamente reportados en la matriz JASPAR de cada FT), para identificar con mayor precisión los verdaderos SFT y su significancia de anclaje.

La gráfica de enriquecimiento de la matriz de Panamá (figura 20, parte A) para el factor de transcripción GATA3, mostro un valor de unión de -1.08, un valor de enriquecimiento de 2.05 y un valor de significancia de 19.19. El factor SP1 mostro un valor de unión de -1.97, un enriquecimiento de 1.81 y una significancia de 14.59. El valor de unión de USF1 fue de -2.72, su valor de enriquecimiento de 1.58 y su significancia de 3.41. Para el factor NFKB1 se mostró un valor de unión de -1.81, un enriquecimiento de 3.62 y una significancia de 39.02. El factor NFATC2 mostro un valor de unión de -1.54, un enriquecimiento de 3.09 y una significancia de 31.47 (Tabla 6).

El resultado del análisis de enriquecimiento de la matriz del Laboratorio de los Álamos (Figura 20, parte B) mostro para el factor de transcripción GATA3, un valor de unión de -0.74, un valor de enriquecimiento de 1.48 y un valor de significancia de 22.82. El factor SP1 mostro un valor de unión de -1.68, un enriquecimiento de 2.39 y una significancia de 6.13. El valor de unión de USF1 fue de -2.63, su valor de enriquecimiento de 1.76 y su significancia de 2.67. Para el factor NFKB1 se mostró un valor de unión de -1.81, un enriquecimiento de

3.62 y una significancia de 39.02. El factor NFATC2 mostro un valor de unión de -1.32, un enriquecimiento de 2.64 y una significancia de 35.19 (tabla6).

En la tabla 6 se observa los valores obtenidos del análisis de enriquecimiento en CiiiDER (U, ENR, VSA), el cual se desarrolló por medio de la comparación de la frecuencia de los sitios de unión del factor de transcripción predichos en las secuencias de entrada con la frecuencia dentro de un conjunto de Secuencias de antecedentes (background). Las Secuencias de antecedentes (background) conformaban una matriz de 165 SFT reportados previamente en estudios in vitro en la base de datos JASPAR. CiiiDER genero un gráfico de enriquecimiento del factor de transcripción con déficit más significativo, a través de la fórmula de VSA (figura 11).

La matriz Los Álamos para el factor de transcripción GATA-3 obtuvo un VSA de 22.82, un total de 14 secuencias con SNP de las 55 analizadas; de las cuales las 14 fueron transversiones nucleotídicas (ver tabla 5 y 3). Para la matriz Panamá, el factor de transcripción tuvo un valor de significancia de anclaje de 19.19 puntos, un total de tres secuencias con PNC en su SFT de 55 secuencias estudiadas, siendo estas tres transversiones.

Tabla 6. Resumen de parámetros analizados en los distintos softwares utilizados según factor de transcripción y matriz de secuencias.

Matriz	FT	U*	ENR*	VSA*	PNC∘	TRANS■	TRANV■
Panamá	GATA3(MA0037.1)	-1.08	2.05	19.19	3	0	0
Panamá	SP1(MA0079.1)	-1.97	1.81	14.59	8	4	3
Panamá	USF1(MA0093.1)	-2.72	1.58	3.41	42	27	5
Panamá	**NFKB1(MA0105.1)**	**-1.81**	**3.62**	**39.02**	**0**	**0**	**0**
Panamá	NFATC2(MA0152.1)	-1.54	3.09	31.47	1	1	0
Los Alamos	GATA3(MA0037.1)	-0.74	1.48	22.82	14	1	0
Los Alamos	SP1(MA0079.1)	-1.68	2.39	6.13	15	4	2
Los Alamos	USF1(MA0093.1)	-2.63	1.76	2.67	44	35	2

Los Alamos	**NFKB1(MA0105.1)**	**-1.81**	**3.62**	**39.02**	**0**	**0**	**0**
Los Alamos	NFATC2(MA0152.1)	-1.32	2.64	35.19	4	2	1

U: valor de unión de FT al SFT

ENR: Enriquecimiento

VSA: Valor de significancia de anclaje del FT al SFT según el Software Ciiider

PNC: Polimorfismo de nucleotídeo característico

TRANS: Transiciones identificadas en HIGHLIGHTER

TRANV: Tranversiones identificadas en HIGHLIGHTER

* Información de Ciiider

◦ Información de VESPA

■ Información de HIGHLIGHTER

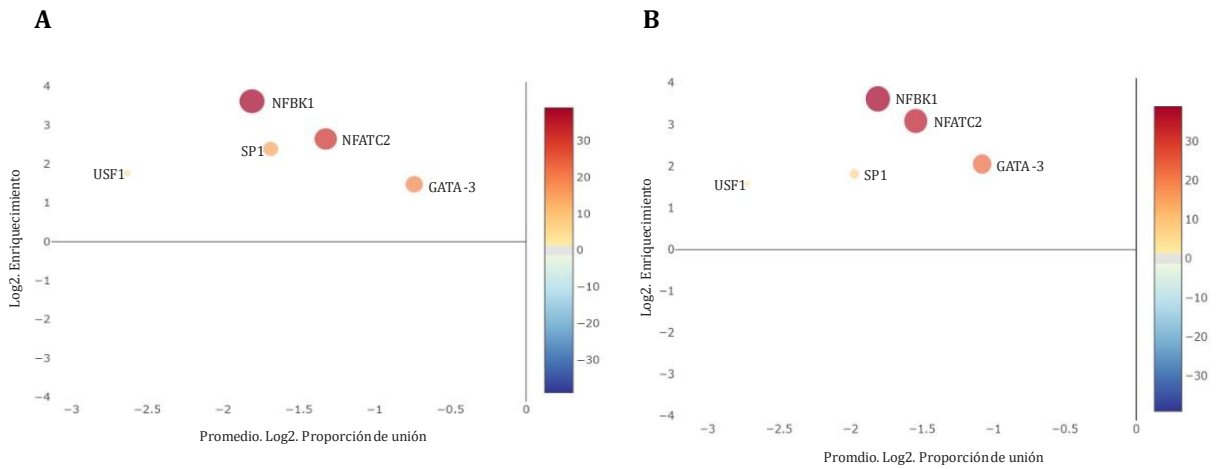


Figura 20. Análisis de enriquecimiento del factor de transcripción según Matriz Panamá (A) y Matriz de Los Álamos (B).

3.6. Análisis estadísticos

El resultado de VSA, fue analizado junto con la cantidad de secuencias con PNC, y la cantidad de secuencias con transiciones y transversiones en el PNC. Los resultados de distribución (Skewnes) y normalidad (Shapiro-Wilk) no son normales (Tabla 7 y figura 21

), por lo cual se trataron como datos no paramétricos.

La figura 22 muestra las dos matrices de correlación de ambas bases de datos, Panamá (A) y Los Álamos(B). Los asteriscos en el valor de rho de Spearman connotan el valor de p de la siguiente manera*: * p < .05, ** p < .01, *** p < .001*. El cuadro rojo corresponde a la correlación entre el valor de significancia de anclaje (VSA) y la cantidad de secuencias en esa matriz con polimorfismos de nucleótido característico (PNC), lo que responde a la pregunta de hipótesis planteada en metodología. El cuadro azul connota la correlación entre las transversiones y transiciones, y apunta a que existe una relación positiva entre estas dos variables (valor de rho de Spearman >0.9*). Por ultimo el cuadro rosado connota la correlación significativa (es decir con un valor de p > 0.05) encontradas en la matriz de los Álamos, pero no en la matriz de Panamá.

El *valor de p (>0.001)* para el análisis de correlación (Spearman) entre VSA y PNC es menor a 0.05 por lo cual la hipótesis nula se rechaza (ver metodología). El Rho de Spearman presento un valor de -1.000 para ambas bases de datos, lo que se considera una fuerte correlación negativa (Figura 22 y tabla 7). Por lo cual, se aceptó la Hi y se rechazó la Ho. Las demás correlaciones con un *valor de p <0.05*, no tenían relación teórica-científica.

Tabla 7. Estadística descriptiva de datos obtenidos de CiiiDER, VESPA y HIGHLIGHTER

							Skewness		**Shapiro-Wilk**	
	Media	**Mediana**	**DE**	**Rango**	**Mínimo**	**Máximo**	**Skewness**	**SE**	**W**	**p**
U*	-1.73	-1.745	0.621	1.98	-2.72	-0.74	-0.2282	0.687	0.958	0.767
ENR*	2.4	2.22	0.81	2.14	1.48	3.62	0.5384	0.687	0.895	0.192
VSA*	21.35	21.005	14.428	36.35	2.67	39.02	-0.0656	0.687	0.898	0.208
PNC∘	13.1	6	16.65	44	0	44	1.3681	0.687	0.761	0.005
TRANS■	7.4	1.5	12.668	35	0	35	1.8156	0.687	0.633	< .001
TRANV■	1.3	0.5	1.703	5	0	5	1.2723	0.687	0.8	0.015

U: valor de unión de FT al SFT

ENR: Enriquecimiento

VSA: Valor de significancia de anclaje del FT al SFT según el Software CiiiDER

TRANS: Transiciones identificadas en HIGHLIGHTER

TRANV: Tranversiones identificadas en HIGHLIGHTER

* Información de CiiiDER

◦ Información de VESPA

■ Información de HIGHLIGHTER

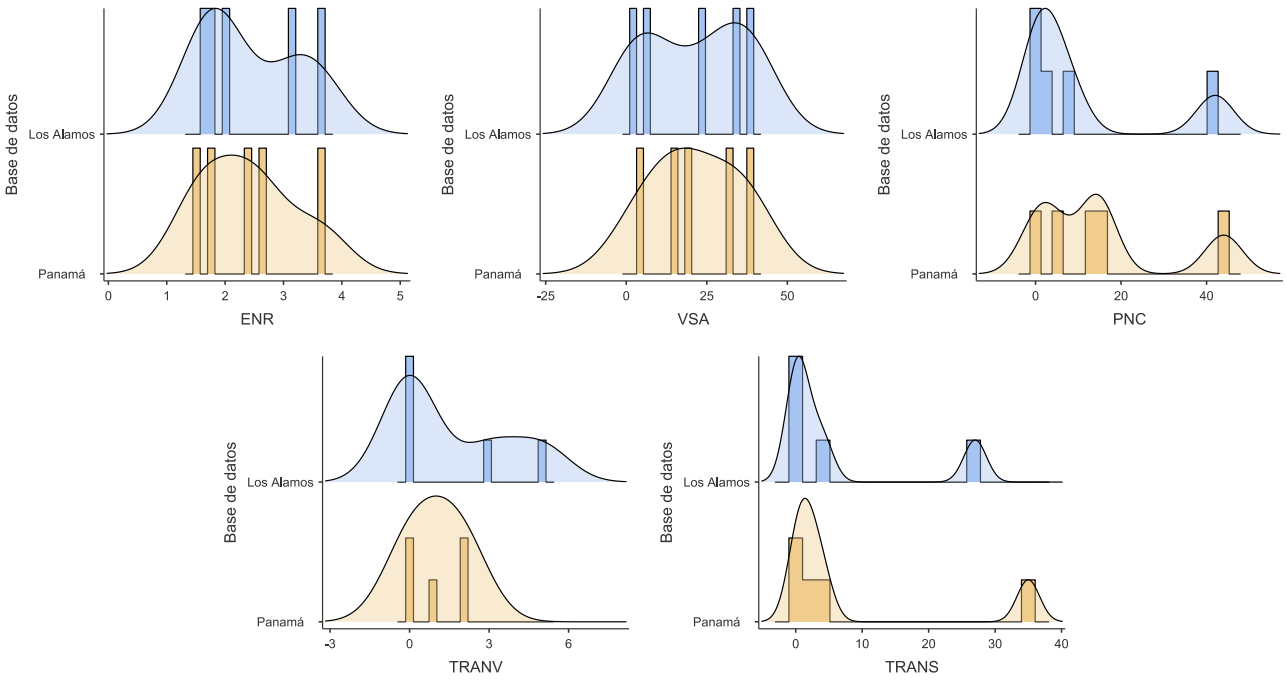


Figura 21. Histograma y densidad de distribución de los datos obtenidos de CiiiDER y HIGHLIGHTER

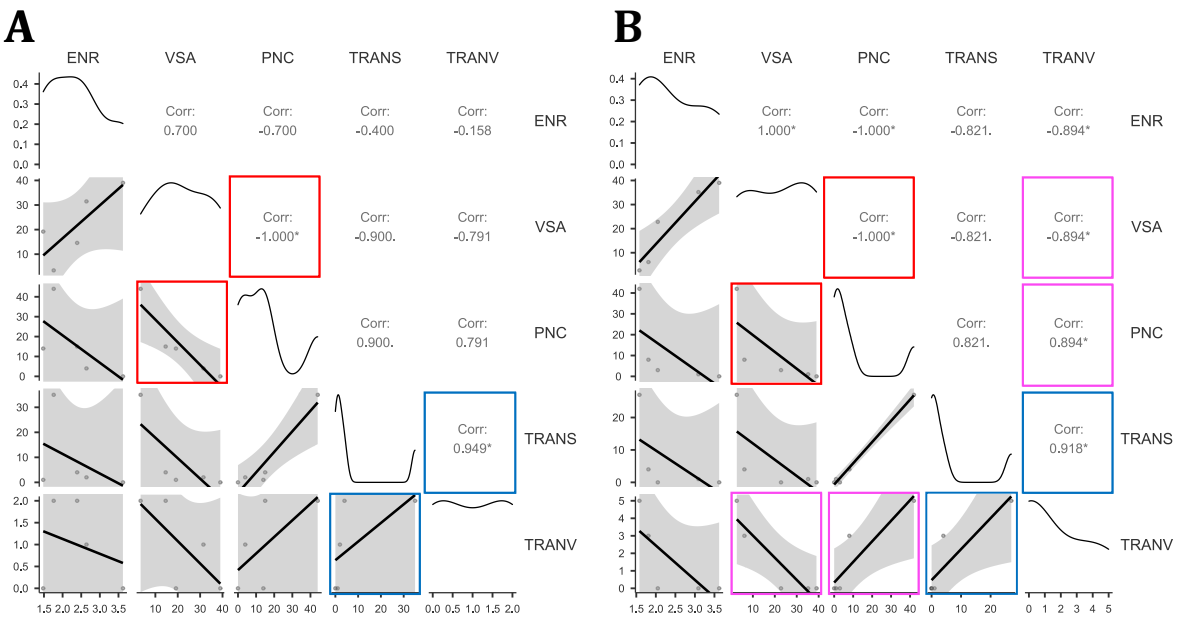


Figura 22. Matriz de correlación de resultados del estudio proveniente de diversos programas bioinformáticos. (A) Análisis de correlación de la Matriz de Panamá. (B) Análisis de correlación de la Matriz Los Álamos.

CAPÍTULO IV: DISCUSIÓN

Discusión

Se analizó el genoma completo de la matriz de secuencias nucleotídicas de Panamá. Los análisis filogenéticos revelaron que la matriz de secuencias nucleotídicas de Panamá pertenece al VIH-1 subtipo B. Nuestros resultados concuerdan con estudios previos de análisis filogenéticos y bootscaning realizados en el país en el año 2008, que establecen qué el subtipo B es la forma dominante (97%) en la epidemia de VIH-1 en Panamá (Ahumada-Ruiz et al., 2008); y estudios posteriores a este en el año 2014, determinaron que las infecciones por VIH-1 en Panamá siguen siendo dominadas por el subtipo B en un 98.9% (Mendoza, Bello, et al., 2014; Mendoza, Martinez, et al., 2014).

Estudios preliminares del factor de transcripción GATA-3 a nivel *in vitro*, identificaron las posiciones 78, 111, 161, 287 y 332 como sitios de unión para este factor (Yang & Engel, 1993). El programa LASAGNA-Search identificó 13 posiciones diferentes como posibles sitios de unión para GATA-3, con MPP optimizadas (Nandi & Ioshikhes, 2012), siendo la posición 76, la más frecuente en un 78% de los sitios de unión de la matriz de Panamá y un 69% en la matriz del Laboratorio de los Álamos. Sin embargo, los resultados por el método de MPP no concuerdan con la HXB2 que tiene la posición 26-41 como sitio de unión para GATA-3 (Singh et al., 2021). El análisis de VESPA no mostro ningún PNC para la posición 76 en ningún experimento realizado.

Estudios in vitro en células Jurkat demostraron que la mutación individual de los sitios de unión de GATA, da como resultado cambios cuantitativos muy pequeños en la transcripción mediada por HIV-1-LTR. Sin embargo, la mutación de todos los sitios GATA

dentro de la LTR si tienen una reducción significativa en la expresión génica mediada por la LTR del VIH-1, y ese efecto acumulativo es menor (de 2 a 3 veces) que la mutación de los dos sitios NF-xB (Yang & Engel, 1993). El SFT GATA-3 mostró 10 variantes diferentes a nivel nucleotídico (ver tabla 3), que eran estrechamente parecidas, pero no eran idénticas al *(A/T)GATA(A/G*) sitio de unión para factores GATA (C. A.-O. Mbondji-Wonje et al., 2018). Además, este sitio de unión presentó una VSA aceptable y moderadamente buena en comparación al resto de los sitios de unión estudiados (tabla 6).

Para el factor de transcripción SP1, se identificó la posición 222 en el 35% de los SFT en la matriz del laboratorio de los Álamos y el 28% para la matriz de Panamá. La posición 376 representó el 41% de secuencias de la matriz de Panamá, y concuerda con la posición del SFT para SP1 según la HXB2 (Jeeninga et al., 2000; Singh et al., 2021). Incluso un estudio *in vitro* reportó la posición 375 como SFT para el factor de transcripción (Qu et al., 2016). El factor SP1 presentó un total de 24 sitios de unión diferentes a nivel nucleotídico con el segundo VSA con menor anclaje del estudio.

USF1 obtuvo la mayor cantidad de SFT con PNC con un porcentaje de 80% para la matriz de Panamá y 76% para la matriz del Laboratorio de los Álamos. Siendo la posición 287 [*CACGTGG*] la más abundante entre todas las posiciones identificadas por LASAGNA-Search, con un porcentaje de frecuencia entre las secuencias de 72% para Panamá y 78% para Los Álamos. Resultados un estudio in vitro sugieren que USF es un regulador positivo de la activación transcripcional mediada por RTL y que la secuencia de SFT es *CACGTGG,* lo cual concuerda con la secuencia nucleotídica de la posición 287 en ambas matrices según análisis de LASAGNA-Search (Fagagna et al., 1995). Además, la posición 287 se encuentra

relativamente cerca de la posición del SFT USF1 según la HXB2 con tan solo 6nt de diferencia con respecto a la posición 281. En estudios in vitro fue reportada en la posición 273 de la RTL 3' [*TGGGGAGTGGC*] como sitio de unión para el factor de transcripción (Jeeninga et al., 2000).

LASAGNA-Search identifico para NFATC2 la posición 112 (G/A)TATCCA en un 25% para la matriz del Laboratorio de Los Alamos y un 38% para la matriz Panamá. Según el análisis de VESPA, NFATC2 obtuvo un PNC para la matriz de Panamá y 5 PNC para la matriz del Laboratorio de Los Alamos. Debido a la baja cantidad de SFT con PNC su valor de significancia fue de 31.47 para Panamá y 35.19 para Los Alamos, lo cual se establece como un excelente anclaje (ver figura 20). Estudios reportan la posición de este FT como 172 según la HXB2 (Singh et al., 2021). En ausencia de NF-KB, NFAT pueda regular la expresión del gen EGR2 o inhibir la activación de NF-KB de la transcripción del gen EGR2 a través de la unión competitiva a secuencias de unión superpuestas o similares (Hokello et al., 2021).

Por otro lado, el SFT NFKB1 no obtuvo ningún PNC según el análisis de VESPA. El programa LASAGNA-Search identificó un total del 90% de los sitios de unión para la matriz de Panamá y el 87% para la matriz de Los Álamos en la posición 363, que se encuentra relativamente cerca de la posición reportada en la HXB2 y de estudios in vitro en la posición 350 para este SFT (Qu et al., 2016). Al ser el único SFT sin polimorfismos significativos, obtuvo el valor de significancia de anclaje más alto, con un puntaje de 39.02 para Panamá y los Álamos. A pesar, de ser matriz con secuencias distintas, la ausencia de polimorfismos en el SFT permitió obtener el mismo valor de significancia, enriquecimiento y unión. Estudios previos in vitro demostraron que NF-κB juega un papel más importante en la transcripción

del VIH que NFAT; este FT se encarga de inducir y activar otros FT como AP-1 para el inicio de la transcripción de virus (Khan et al., 2012; C. A.-O. Mbondji-Wonje et al., 2018). Nuestros estudios concuerdan con la identificación de SFT a nivel in vitro que presentan la secuencia GGGACTTTCC en un 99% de coincidencia exacta con la secuencia de consenso indicada en este estudio (Elsheikh, Tang, Li, & Jiang, 2019). Esta misma secuencia nucleotídica fue la encontrada para la posición de 363 según el análisis de identificación de LASAGNA-Search.

Los resultados de transiciones y transversiones con el programa HIGHLIGHTER establecieron una notable diferencia entre la cantidad de transiciones por matriz; siendo las transiciones, la mutación más predominante. Probablemente, se debe a es más factible sustituir una estructura de un solo anillo (pirimidinas) por otra estructura de un solo anillo que sustituir un anillo simple por un anillo doble (purinas) (Stoltzfus & Norris, 2015; Wakeley, 1994). Nuestros resultados concuerdan con la literatura ya que encontramos mayor cantidad de transiciones que transversiones en los polimorfismos de nucleótido característico de ambas matrices (tabla 6).

Según los análisis estadísticos, existe una alta correlación negativa (Spearman rho -0.939 y *valor de p <0.001*) entre el valor de significancia de anclaje y el número de secuencias con polimorfismos de nucleótido característico (Figura 22). Esta correlación negativa indica que, a menor polimorfismo, existe una mayor capacidad de anclaje del factor de transcripción al sitio de unión. Se puede inferir según nuestros resultados que si existe una relación entre la diversidad genética de los sitios de unión (PNC) y la capacidad de anclaje de los sitios de unión de los factores de transcripción (VSA). Los SFT SP1 y USF1 fueron lo que mostraron mayor cantidad de SFT diferentes según el método MPP. Incluso, fueron los

SFT con mayor cantidad de PNC, lo que afectó su capacidad y significancia de anclaje. Por otro lado, NFKB1 y NFATC2 obtuvieron la menor cantidad de secuencias diferentes de SFT y el mayor VSA registrado. Entonces, la diversidad genética estudiada en los SFT (PNC) de ambas matrices mostraron una correlación negativa de igual magnitud con respecto al valor de significancia de anclaje de los factores de transcripción al sitio de unión.

Sin embargo, según el análisis de CiiiDER todos los VSA de los SFT estudiados presentaron un anclaje sobrerepresentado (ver figura 20), lo que significa que a pesar de que la diversidad nucleotídica afecto su capacidad de anclaje, a nivel *in silico* son considerados SFT aceptables para el anclaje del FT. Estudios in vitro han sugerido para los FT GATA, SP, USF las mutaciones puntuales a nivel nucleotídico no afectan de forma drástica el proceso de transcripción, pero mutaciones en los SFT NF-KB y NFAT pueden afectar la capacidad de los procesos de transcripción en el VIH-1 (Malcolm, Kam, Pour, & Sadowski, 2008; Yang & Engel, 1993).

Se sabe que la actividad diferencial de la RTL debido a la mutación o ausencia de un SFT puede compensarse mediante varios mecanismos, tales como, la ganancia de factores adicionales, la interacción entre SFT en estrecha proximidad o la deslocalización del SFT (C. Mbondji-Wonje et al., 2018). Por esta razón, es necesario la realización de estudios más completos con todo el complejo de proteínas involucradas en el proceso de transcripción del VIH-1 en la RTL. Estudios como los presentados en esta tesis, son útiles para contribuir a la producción de nuevas estrategias, como en el desarrollo de vacunas y tratamientos antirretrovirales; basado en la idea de la transactivación, como una forma de activar y

desactivar la transcripción y replicación, en respuesta a los cambios del entorno celular por medio de factores de transcripción (Elsheikh et al., 2019).

CAPÍTULO V: CONCLUSIONES & RECOMENDACIONES

CONCLUSIONES

- ✓ Las secuencias nucleotídicas estudiadas de genoma completo procedentes de sujetos panameños fueron dominadas por el VIH-1, subtipo B, según análisis filogenéticos.
- ✓ La diversidad genética de las regiones RTL 3' tiene una correlación negativa con respecto a el valor de significancia de anclaje del sitio de unión al factor de transcripción.
- ✓ El tipo de polimorfismo (PNC) mayoritario fueron las transiciones para ambas matrices estudiadas; como establece la teoría de selección molecular de transiciones y transversiones. Por lo que, se puede inferir que todos los SFT fueron sobre-reprepresentados ya que es menos probable que las transiciones den como resultado sustituciones aminoacídicas y que persistan como "sustituciones silenciosas" en polimorfismos de un solo nucleótido.

RECOMENDACIONES

- ✓ Se recomienda utilizar una matriz con un mayor número de secuencias según los factores de transcripción, para poder asociar una mutación especifica y establecer si a causa de ella existe una disminución en la capacidad de anclaje.
- ✓ Implementar al diseño experimental la región RTL 5' de secuencias panameñas y matriz del Laboratorio de los Álamos.
- ✓ Implementar el modelamiento de proteínas por homología para observar la interacción del factor de transcripción con el sitio de unión.

BIBLIOGRAFÍA

Agarwal-Jans, S. Timeline: HIV. (1097-4172 (Electronic)).

Ahumada-Ruiz, S., Casado, C., Toala-Gonzalez, I., Flores-Figueroa, D., Rodriguez-French, A., & Lopez-Galindez, C. (2008). High divergence within the major HIV type 1 subtype B epidemic in Panama. *AIDS Res Hum Retroviruses, 24*(11), 1461-1466. doi:10.1089/aid.2008.0153

Altman, D. G., & Bland, J. M. (1996). Statistics Notes: Detecting skewness from summary information. *BMJ, 313*(7066), 1200.

Berg, M. G., Yamaguchi, J., Alessandri-Gradt, E., Tell, R. W., Plantier, J. C., & Brennan, C. A. (2016). A Pan-HIV Strategy for Complete Genome Sequencing. *J Clin Microbiol, 54*(4), 868-882. doi:10.1128/JCM.02479-15

Bulmer, M. G. (1979). *Principles of statistics*: Courier Corporation.

Burdo, T. H., Nonnemacher, M., Irish, B. P., Choi, C. H., Krebs, F. C., Gartner, S., & Wigdahl, B. (2004). High-Affinity Interaction between HIV-1 Vpr and Specific Sequences That Span the C/EBP and Adjacent NF-κ B Sites within the HIV-1 LTR Correlate with HIV-1-Associated Dementia. *DNA and cell biology, 23*(4), 261-269.

Cabello, M., Mendoza, Y., & Bello, G. (2014). Spatiotemporal dynamics of dissemination of non-pandemic HIV-1 subtype B clades in the Caribbean region. *PLoS One, 9*(8), e106045. doi:10.1371/journal.pone.0106045

Castro-Mondragon, J. A.-O. X., Riudavets-Puig, R. A.-O., Rauluseviciute, I. A.-O., Lemma, R. A.-O., Turchi, L. A.-O., Blanc-Mathieu, R. A.-O., . . . Mathelier, A. A.-O. JASPAR 2022: the

9th release of the open-access database of transcription factor binding profiles. (1362-4962 (Electronic)).

Castro-Nallar, E., Perez-Losada, M., Burton, G. F., & Crandall, K. A. (2012). The evolution of HIV: inferences using phylogenetics. *Mol Phylogenet Evol, 62*(2), 777-792. doi:10.1016/j.ympev.2011.11.019

Cordeiro, N., Taroco, R., & Higiene, U. U. d. l. r. F. d. M. I. d. (2008). Retrovirus y VIH. *Temas de bacteriología y virología médica*.

Deeks, S. G., Overbaugh, J., Phillips, A., & Buchbinder, S. (2015). HIV infection. *Nature reviews Disease primers, 1*(1), 1-22.

Delatorre, E., & Bello, G. (2013). Phylodynamics of the HIV-1 epidemic in Cuba. *PLoS One, 8*(9), e72448. doi:10.1371/journal.pone.0072448

Delgado, R. (2011). Características virológicas del VIH. *Enferm Infecc Microbiol Clin, 29*(1), 58-65. doi:https://doi.org/10.1016/j.eimc.2010.10.001

Elsheikh, M. M., Tang, Y., Li, D., & Jiang, G. (2019). Deep latency: A new insight into a functional HIV cure. *EBioMedicine, 45*, 624-629. doi:https://doi.org/10.1016/j.ebiom.2019.06.020

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of statistical software, 48*, 1-18.

Fagagna, F. d. A. d., Marzio, G., Gutierrez, M. I., Kang, L. Y., Falaschi, A., & Giacca, M. (1995). Molecular and functional interactions of transcription factor USF with the long terminal repeat of human immunodeficiency virus type 1. *J Virol, 69*(5), 2765-2775. doi:doi:10.1128/jvi.69.5.2765-2775.1995

Felsenstein, J. (1985). CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. (1558-5646 (Electronic)).

Fisher, M. J., & Marshall, A. P. (2009). Understanding descriptive statistics. *Australian critical care, 22*(2), 93-97.

Gallo, R. C., & Montagnier, L. (2003). The discovery of HIV as the cause of AIDS. *N Engl J Med, 349*(24), 2283-2285.

Gearing, L. A.-O., Cumming, H. E., Chapman, R., Finkel, A. M., Woodhouse, I. B., Luu, K., . . . Hertzog, P. A.-O. (2019). CiiiDER: A tool for predicting and analysing transcription factor binding sites. (1932-6203 (Electronic)).

Gomez-Lucia, E., Collado, V., Miró, G., & Doménech, A. (2009). Effect of Type-I Interferon on Retroviruses. *Viruses, 1*, 545-573. doi:10.3390/v1030545

Gómez-Román, R., & Soler-Claudín, C. (2000). La importancia de la Secuencia Terminal Repetida Larga (LTR) en la patogenia del virus de la inmunodeficiencia humana. *Revista Biomédica, 11*(1), 61-71. Retrieved from https://www.revistabiomedica.mx/index.php/revbiomed/article/view/219

Goodsell, D. S. (2015). Illustrations of the HIV life cycle. *The Future of HIV-1 Therapeutics*, 243-252.

Hahn, B. H., Shaw, G. M., De, K. M., Cock, & Sharp, P. M. (2000). AIDS as a zoonosis: scientific and public health implications. *Science, 287*(5453), 607-614.

Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends Mol Med, 18*(3), 182-192. doi:10.1016/j.molmed.2011.12.001

Hokello, J., Lakhikumar Sharma, A., & Tyagi, M. (2021). AP-1 and NF-κB synergize to transcriptionally activate latent HIV upon T-cell receptor activation. *FEBS Lett, 595*(5), 577-594. doi:https://doi.org/10.1002/1873-3468.14033

Jeeninga, R. E., Hoogenkamp, M., Armand-Ugon, M., Baar, M. d., Verhoef, K., & Berkhout, B. (2000). Functional Differences between the Long Terminal Repeat Transcriptional Promoters of Human Immunodeficiency Virus Type 1 Subtypes A through G. *J Virol, 74*(8), 3740-3751. doi:doi:10.1128/JVI.74.8.3740-3751.2000

Junqueira, D. M., & Almeida, S. E. (2016). HIV-1 subtype B: Traces of a pandemic. *Virology, 495*, 173-184.

Junqueira, D. M., de Medeiros, R. M., Matte, M. C., Araujo, L. A., Chies, J. A., Ashton-Prolla, P., & Almeida, S. E. (2011). Reviewing the history of HIV-1: spread of subtype B in the Americas. *PLoS One, 6*(11), e27489. doi:10.1371/journal.pone.0027489

Khan, M. T., Mischiati C Fau - Ather, A., Ather A Fau - Ohyama, T., Ohyama T Fau - Dedachi, K., Dedachi K Fau - Borgatti, M., Borgatti M Fau - Kurita, N., . . . Gambari, R. (2012). Structure-based analysis of the molecular recognitions between HIV-1 TAR-RNA and transcription factor nuclear factor-kappaB (NFkB). (1873-4294 (Electronic)).

Kirchhoff, F. (2013). HIV life cycle: overview. *Encyclopedia of AIDS*, 1-9.

Kirchner, J. T. (2019). The origin, evolution, and epidemiology of HIV-1 and HIV-2. *Fundamentals of HIV Medicine 2019*, 14. Retrieved from https://books.google.es/books?hl=es&lr=&id=zS0zEAAAQBAJ&oi=fnd&pg=PA20&dq=The+origin,+evolution,+and+epidemiology+of+HIV-1+and+HIV-2&ots=ttOSzPW7P-&sig=7rWwVmCWB2O5eWGG4N3RotqDksA#v=onepage&q=The%20origin%2C%2

0evolution%2C%20and%20epidemiology%20of%20HIV-1%20and%20HIV-2&f=false

Korber, B., & Myers, G. (1992). Signature pattern analysis: a method for assessing viral sequence relatedness. (0889-2229 (Print)).

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol, 33*(7), 1870-1874.

Lee, C., & Huang, C.-H. (2013). LASAGNA-Search: an integrated web tool for transcription factor binding site search and visualization. *Biotechniques, 54*(3), 141-153. doi:10.2144/000113999

Liu, Y., Li, L., Bao, Z., Li, H., Zhuang, D., Liu, S., . . . Li, J. (2012). Identification of a novel HIV type 1 circulating recombinant form (CRF52_01B) in Southeast Asia. *AIDS Res Hum Retroviruses, 28*(10), 1357-1361. doi:10.1089/aid.2011.0376

Love J, D. D. R. S. (2022). The Jamovi project; version 1.8. Retrieved from https://www.jamovi.org

Maina, E. K., Adan, A. A., Mureithi, H., Muriuki, J., & Lwembe, R. M. (2021). A review of current strategies towards the elimination of latent hiv-1 and subsequent hiv-1 cure. *Current HIV Research, 19*(1), 14-26.

Malcolm, T., Kam, J., Pour, P. S., & Sadowski, I. (2008). Specific interaction of TFII-I with an upstream element on the HIV-1 LTR regulates induction of latent provirus. *FEBS Lett, 582*(28), 3903-3908. doi:https://doi.org/10.1016/j.febslet.2008.10.032

Mansky, L. M., & Temin, H. M. (1995). Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified

reverse transcriptase. *J Virol, 69*(8), 5087-5094. doi:10.1128/JVI.69.8.5087-5094.1995

Mbondji-Wonje, C., Dong, M., Wang, X., Zhao, J., Ragupathy, V., Sanchez, A. M., . . . Hewlett, I. (2018). Distinctive variation in the U3R region of the 5'Long Terminal Repeat from diverse HIV-1 strains. *PLoS One, 13*(4), e0195661. Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0195661

Mbondji-Wonje, C. A.-O., Dong, M., Wang, X., Zhao, J. A.-O., Ragupathy, V., Sanchez, A. M., . . . Hewlett, I. (2018). Distinctive variation in the U3R region of the 5' Long Terminal Repeat from diverse HIV-1 strains. (1932-6203 (Electronic)).

Mbonye, U., & Karn, J. (2017). The Molecular Basis for Human Immunodeficiency Virus Latency. *Annual Review of Virology, 4*(1), 261-285. doi:10.1146/annurev-virology-101416-041646

McLaren, P. J., & Fellay, J. (2021). HIV-1 and human genetic variation. *Nat Rev Genet, 22*(10), 645-657. doi:10.1038/s41576-021-00378-0

Mendoza, Y., Bello, G., Castillo Mewa, J., Martinez, A. A., Gonzalez, C., Garcia-Morales, C., . . . Pascale, J. M. (2014). Molecular epidemiology of HIV-1 in Panama: origin of non-B subtypes in samples collected from 2007 to 2013. *PLoS One, 9*(1), e85153. doi:10.1371/journal.pone.0085153

Mendoza, Y., Martinez, A. A., Castillo Mewa, J., Gonzalez, C., Garcia-Morales, C., Avila-Rios, S., . . . Bello, G. (2014). Human immunodeficiency virus type 1 (HIV-1) subtype B epidemic in Panama is mainly driven by dissemination of country-specific clades. *PLoS One, 9*(4), e95360. doi:10.1371/journal.pone.0095360

Mir, D., Cabello, M., Romero, H., & Bello, G. (2015). Phylodynamics of major HIV-1 subtype B pandemic clades circulating in Latin America. *AIDS, 29*(14), 1863-1869. doi:10.1097/QAD.0000000000000770

Nandi, S., & Ioshikhes, I. (2012). Optimizing the GATA-3 position weight matrix to improve the identification of novel binding sites. *BMC Genomics, 13*(1), 416. doi:10.1186/1471-2164-13-416

Nonnemacher, M. R., Irish, B. P., Liu, Y., Mauger, D., & Wigdahl, B. (2004). Specific sequence configurations of HIV-1 LTR G/C box array result in altered recruitment of Sp isoforms and correlate with disease progression. *J Neuroimmunol, 157*(1-2), 39-47.

Ou, C.-Y., Ciesielski, C. A., Myers, G., Bandea, C. I., Luo, C.-C., Korber, B. T. M., . . . Jaffe, H. W. (1992). Molecular Epidemiology of HIV Transmission in a Dental Practice. *Science, 256*(5060), 1165-1171. doi:doi:10.1126/science.256.5060.1165

Qu, D., Li, C., Sang, F., Li, Q., Jiang, Z. Q., Xu, L. R., . . . Wang, J. H. (2016). The variances of Sp1 and NF-kappaB elements correlate with the greater capacity of Chinese HIV-1 B'-LTR for driving gene expression. *Sci Rep, 6*(1), 34532. doi:10.1038/srep34532

Ramirez de Arellano, E., Martin, C., Soriano, V., Alcami, J., & Holguin, A. (2007). Genetic analysis of the long terminal repeat (LTR) promoter region in HIV-1-infected individuals with different rates of disease progression. *Virus Genes, 34*(2), 111-116. doi:10.1007/s11262-006-0054-z

Ramirez de Arellano, E., Soriano, V., & Holguin, A. (2005). [Regulation of transcription in different HIV-1 subtypes]. *Enferm Infecc Microbiol Clin, 23*(3), 156-162. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15757588

Rausch, J. W., & Le Grice, S. F. J. (2020). Characterizing the Latent HIV-1 Reservoir in Patients with Viremia Suppressed on cART: Progress, Challenges, and Opportunities. (1873-4251 (Electronic)).

Revelle, W. (2019). Psych: Procedures for psychological, psychometric, and personality research.[R package]. *Retrieved from [left angle bracket]* https://cran*. r-project. org/package= psych [right angle bracket]*.

Rodríguez, E. C. (2017). Revisión bibliográfica sobre VIH/sida. *Multimed, 17*(4).

Roebuck, K. A., & Saifuddin, M. (1999). Regulation of HIV-1 transcription. *Gene Expr, 8*(2), 67-84. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10551796

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. (0737-4038 (Print)). doi:D - NASA: 88189002 EDAT- 1987/07/01 00:00 MHDA- 1987/07/01 00:01 CRDT- 1987/07/01 00:00 PHST- 1987/07/01 00:00 [pubmed] PHST- 1987/07/01 00:01 [medline] PHST- 1987/07/01 00:00 [entrez] AID - 10.1093/oxfordjournals.molbev.a040454 [doi] PST - ppublish

Seol, H. (2022). Seolmatrix: Correlations suite for jamovi. [jamovi module]. . Retrieved from https://github.com/hyunsooseol/seolmatrix.

Sharp, P. M., & Hahn, B. H. (2011). Origins of HIV and the AIDS pandemic. *Cold Spring Harb Perspect Med, 1*(1), a006841. doi:10.1101/cshperspect.a006841

Singh, S., Kumar, A., Brijwal, M., Choudhary, A., Singh, K., Singh, R., . . . Dar, L. (2021). Intra-Clade C signature polymorphisms in HIV-1 LTR region: The Indian and African lookout. *Virus Res, 297*, 198370. doi:10.1016/j.virusres.2021.198370

Smyth, R. P., Davenport, M. P., & Mak, J. (2012). The origin of genetic diversity in HIV-1. *Virus Res, 169*(2), 415-429. doi:10.1016/j.virusres.2012.06.015

Stoltzfus, A., & Norris, R. W. (2015). On the Causes of Evolutionary Transition:Transversion Bias. *Mol Biol Evol, 33*(3), 595-602. doi:10.1093/molbev/msv274

Tamura, K., Dudley, J., Nei, M., & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol, 24*(8), 1596-1599. doi:10.1093/molbev/msm092

Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. (0737-4038 (Print)).

Velasco-de-Castro, C. A., Grinsztejn, B., Veloso, V. G., Bastos, F. I., Pilotto, J. H., Fernandes, N., & Morgado, M. G. (2014). HIV-1 diversity and drug resistance mutations among people seeking HIV diagnosis in voluntary counseling and testing sites in Rio de Janeiro, Brazil. *PLoS One, 9*(1), e87622. doi:10.1371/journal.pone.0087622

Wakeley, J. (1994). Substitution-rate variation among sites and the estimation of transition bias. *Mol Biol Evol, 11*(3), 436-442. doi:10.1093/oxfordjournals.molbev.a040124

White, S. N., Mousel, M. R., Herrmann-Hoesing, L. M., Reynolds, J. O., Leymaster, K. A., Neibergs, H. L., . . . Knowles, D. P. (2012). Genome-wide association identifies multiple genomic regions associated with susceptibility to and control of ovine lentivirus. *PLoS One, 7*(10), e47829. doi:10.1371/journal.pone.0047829

Yang, Z., & Engel, J. D. (1993). Human T cell transcription factor GATA-3 stimulates HIV-1 expression. (0305-1048 (Print)).

Zeng, Y., Gong, M., Lin, M., Gao, D., & Zhang, Y. (2020). A review about transcription factor binding sites prediction based on deep learning. *IEEE Access, 8*, 219256-219274.