UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

BIOLOGÍA EVOLUTIVA

LOCALIZACIÓN, CARACTERIZACIÓN Y DISTRIBUCIÓN DE SECUENCIAS SIMPLES EN LA PROTEÍNA gp120 DE HIV-1: IMPORTANCIA EN LA GENERACIÓN DE LA HIPERVARIABILIDAD RETROVIRAL

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTORA EN CIENCIAS

PRESENTA:

ANA MARÍA VELASCO VELASCO

TUTOR PRINCIPAL DE TESIS: DR. ANTONIO EUSEBIO LAZCANO ARAUJO REYES
FACULTAD DE CIENCIAS, UNAM
COMITÉ TUTOR: DR. LUIS FELIPE JIMÉNEZ GARCÍA
FACULTAD DE CIENCIAS, UNAM
DRA. MARÍA DEL CARMEN GÓMEZ EICHELMANN
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS, UNAM

MÉXICO, D.F. ENERO, 2014.

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

BIOLOGÍA EVOLUTIVA

LOCALIZACIÓN, CARACTERIZACIÓN Y DISTRIBUCIÓN DE SECUENCIAS SIMPLES EN LA PROTEÍNA gp120 DE HIV-1: IMPORTANCIA EN LA GENERACIÓN DE LA HIPERVARIABILIDAD RETROVIRAL

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTORA EN CIENCIAS

PRESENTA:

ANA MARÍA VELASCO VELASCO

TUTOR PRINCIPAL DE TESIS: DR. ANTONIO EUSEBIO LAZCANO ARAUJO REYES
FACULTAD DE CIENCIAS, UNAM
COMITÉ TUTOR: DR. LUIS FELIPE JIMÉNEZ GARCÍA
FACULTAD DE CIENCIAS, UNAM
DRA. MARÍA DEL CARMEN GÓMEZ EICHELMANN
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS, UNAM

MÉXICO, D.F. ENERO, 2014.

POSGRADO EN CIENCIAS BIOLÓGICAS
FACULTAD DE CIENCIAS
DIVISIÓN DE ESTUDIOS DE POSGRADO

OFICIO FCIE/DEP/923/13

ASUNTO: Oficio de Jurado

Dr. Isidro Ávila Martínez
Director General de Administración Escolar, UNAM
P r e s e n t e

Me permito informar a Usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día **27 de mayo de 2013**, se aprobó el siguiente jurado para el examen de grado de **DOCTORA EN CIENCIAS** del (de) alumno (a) **VELASCO VELASCO ANA MARÍA** con número de cuenta **91285167** con la tesis titulada: **"Localización, caracterización y distribución de secuencias simples en la proteína gp 120 de HIV-1: Importancia en la generación de la hipervariabilidad retroviral"**, realizada bajo la dirección del (de) **DR. ANTONIO EUSEBIO LAZCANO-ARAUJO REYES:**

Presidente:	DRA. YOLANDA LÓPEZ VIDAL
Vocal:	M. EN C. SAÚL ALFONSO DE LEÓN ROSALES
Secretario:	DR. LUIS FELIPE JIMÉNEZ GARCÍA
Suplente:	DRA. MARÍA EUGENIA JIMÉNEZ CORONA
Suplente:	DRA. MARÍA DEL CARMEN GÓMEZ EICHELMANN

De acuerdo con lo anterior, el (la) alumno(a) se apega a la nueva normatividad, particularmente en el artículo QUINTO TRANSITORIO en apego a lo establecido en el Artículo 31 del Reglamento General de Estudios de Posgrado (8 octubre de 2006).

Sin otro particular, me es grato enviarle un cordial saludo.

A t e n t a m e n t e
"POR MI RAZA HABLARA EL ESPÍRITU"
Cd. Universitaria, D.F. a 14 de noviembre de 2013.

Dra. María del Coro Arizmendi Arriaga
Coordinadora de Programa

FACULTAD DE CIENCIAS
DIVISION DE ESTUDIOS DE POSGRADO

MCAA/MJFM/ASR/grf*

A G R A D E C I M I E N T O S

- Agradezco al Programa de Posgrado en Ciencias Biológicas de la Universidad Nacional Autónoma de México por darme la oportunidad de continuar con mi formación académica y científica;

- el apoyo otorgado por el Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca otorgada para la realización de mis estudios de Doctorado;

- el apoyo del Proyecto CONACYT 50520-Q, al cual estuvo vinculado el desarrollo de este trabajo de tesis;

- a mi Tutor Principal, el Dr. Antonio Lazcano, por permitirme formar parte de su grupo, por su infinita paciencia, por todas sus valiosas enseñanzas durante la realización y término de este trabajo; pero sobre todo, por su apoyo académico, profesional, personal y calidad humana con las que siempre he contado;

- a los miembros de mi Comité Tutoral, Dra. Carmen Gómez Eichelmann y Dr. Luis Felipe Jiménez García por su interés, disposición y guía, así como por sus observaciones y aportaciones durante el desarrollo de este trabajo.

A G R A D E Z C O

- A los miembros de mi jurado, Dra. Yolanda López Vidal, Dr. Samuel Ponce de León, Dr. Luis Felipe Jiménez García, Dra. Carmen Gómez Eichelmann y Dra. María Eugenia Jiménez Corona, por la pronta revisión de este texto y sus acertados comentarios, los cuales enriquecieron su contenido;

- a Arturo Becerra, Luis Delaye y Ricardo Hernández por su ayuda con la elaboración de programas que se utilizaron en el proyecto;

- a Arturo Becerra y Luis Delaye por su apoyo y amistad en todo momento;

- a Sara Islas por su ayuda para revisar el texto y las referencias del artículo;

- a Héctor Miguel Cejudo Camacho por su invaluable ayuda para la recuperación datos y revivir computadoras dañadas o con virus, así como por la obtención e instalación de programas utilizados durante el desarrollo de este trabajo, por su gran conocimiento y habilidades en cómputo que me rescataron en momentos cruciales, gracias, gracias, gracias;

- a Alvaro García por realizar el análisis estadístico de los datos obtenidos;

- al Dr. Samuel Ponce de León y a la Dra. María Eugenia Jiménez por su apoyo y por las facilidades que me otorgaron para concluir este trabajo;

- a mis amigos del Laboratorio de Origen de la Vida, Sara y Erwin por todas la pláticas y discusiones tan “profundas” en las que nos enfrascamos, gracias por su apoyo y amistad;

- a Lourdes Agredano por su amistad, comprensión y apoyo;

- a David Amador por su amistad y apoyo;

- y de manera muy especial a la UNAM, por brindarme el privilegio de formar parte de su vida académica y científica durante todos estos años, gracias.

A mi mamá Anita Velasco y mi papá Bulmaro Velasco,

por todo, todo su apoyo y enseñanzas,

y por todo su amor

G R A C I A S

A Vero y Claudia Orozco Jiménez,

queridas amigas, gracias

por el tiempo que

compartimos,

LAS EXTRAÑO

ÍNDICE

ÍNDICE DE FIGURAS

R E S U M E N

Las regiones de baja complejidad (LCRs por sus siglas en inglés) son secuencias de ácidos nucleicos o de proteínas que presentan un sesgo en su composición. Su presencia ha sido confirmada en los tres grandes linajes celulares: Bacteria, Archaea y Eucarya. Sin embargo, a pesar de que en virus se han encontrado LCRs, tanto en genomas como en proteínas, es poco lo que se sabe de estas estructural moleculares en virus. En este trabajo se llevó a cabo la búsqueda , localización y caracterización de LCRs en la glicoproteína 120 (gp120) del virus de inmunodeficiencia humana tipo 1 (VIH-1), utilizando una base de 4117 genomas completamente secuenciados obtenidos de la base de datos de VIH de Los Álamos National Laboratory y del GenBank. La identificación de LCRs en las secuencias de las proteínas de gp120 analizadas se llevó a cabo utilizando el programa SEG. Se han encontrado LCRs en poco mas del 30% de las proteínas de nuestro banco de datos, las cuales se localizan distribuidas en cuatro de las cinco regiones hipervariables de gp120: V1, V2, V4 y V5. No se encontraron LCRs en V3. En las LCRs localizadas en V1, V2, V4 y V5 presentan un elevado contenido de asparagina, que probablemente se localiza en sitios de glicosilación, lo cual puede contribuir a la capacidad de los retrovirus de evadir el sistema inmune del hospedero. Nuestros resultados sugieren que las LCR representan una fuente no descrita hasta ahora de la variabilidad genómica en los lentivirus, las cuales pueden representar una fuente importante de variación antigénica en las poblaciones de VIH-1. Los resultados aquí presentados pueden ejemplificar los procesos evolutivos que pudieron aumentar el tamaño de genomas celulares primitivos de RNA y la función de las LCRs como una fuente de materia prima durante los procesos de adquisición evolutiva de nuevas funciones.

A B S T R A C T

Low complexity regions (LCRs) are sequences of nucleic acids or proteins defined by a compositional bias. Their occurrence has been confirmed in sequences of the three cellular lineages (Bacteria, Archaea and Eucarya), and has also been reported in viral genomes. We present here the results of a detailed computer analysis of the LCRs present in the HIV-1 glycoprotein 120 (gp120) encoded by the viral gene env. The analysis was performed using SEG progarm to analize a sample of 4117 completely sequenced and translated HIV-1 genomes available in public databases. We have identified LCRs in four different regions of the gp120 protein that correspond to four of the five regions that have been identified as hypervariable (V1, V2, V4 and V5). No LCR has been identified in the hypervariable region V3. The LCRs detected in the V1, V2, V4, and V5 hypervariable regions exhibit a high Asn content in their amino acid composition, which very likely correspond to glycosylation sites, which may contribute to the retroviral ability to avoid the immune system. The results presented here suggest that LCRs represent a hitherto undescribed source of genomic variability in lentivirus, and that these repeats may represent an important source of antigenic variation in HIV-1 populations. The results reported here may exemplify the evolutionary processes that may have increased the size of primitive cellular RNA genomes and the role of LCRs as a source of raw material during the processes of evolutionary acquisition of new functions.

I N T R O D U C C I Ó N

Existen diversos mecanismos que permiten explicar las variaciones en los tamaños de los genomas tanto de procariontes como de eucariontes, entre los que podemos mencionar la pérdida de material genético, duplicación completa del genoma, duplicación completa o parcial de un gen, o por adquisición de nuevo material genético por transferencia horizontal (Petrov, 2001; Kunin y Ouzounis, 2003). Además de los procesos mencionados anteriormente, las regiones de baja complejidad (LCRs por sus siglas en inglés), también llamadas secuencias simples (SS) representan uno de los mecanismos capaces de generar de variaciones en el tamaño del genoma, así como de diversidad genética (Hancock, 2002).

La mayor parte de los estudios de LCRs que se han realizados hasta ahora son en procariontes y eucariontes, y se han llevado a cabo muy pocas investigaciones en virus. Debido a su genoma de tamaño pequeño y compactado, se suele suponer que las mutaciones puntuales y los eventos de recombinación son los principales mecanismos que proporcionan diversidad en poblaciones virales sin que se produzcan modificaciones considerables en el tamaño del genoma (Keese y Gibbs, 2003).

CARACTERÍSTICAS DE LAS LCRs

Las LCRs o SS son regiones de ácidos nucleicos o proteínas que presentan un sesgo en la composición de sus residuos. Algunas LCRs están constituidas por homopolímeros o por motivos repetidos, mientras que en otras los residuos se presentan de una manera irregular (Wootton, 1994). Estas regiones exhiben una tasa evolutiva elevada, pues tienen una frecuencia muy alta de mutación (Huntley y Golding, 2000).

Las LCRs se pueden generar por eventos de recombinación, o por un error en el apareamiento de las bases durante la replicación del material genético. A este fenómeno se le denomina patinaje (slippage) durante la replicación (Schlötterer y Tautz, 1992) (Fig. 1). La generación de LCRs puede darse en regiones codificadoras y no codificadoras (Tautz *et al*., 1986). Cuando las SS se presentan en regiones codificadoras, esto lleva a un aumento de la variabilidad a nivel de proteínas (Sim y Creamer, 2002).

Tanto en procariontes como en eucariontes se han encontrado LCRs en una amplia gama de proteínas. Además de generar variabilidad, también se han reportado casos de LCRs conservadas que generalmente están asociadas a una función. Por ejemplo, forman parte importante de proteínas asociadas a membranas (Moxon, 1999), así como a sitios de unión a DNA como es el caso de los dedos de zinc (Heringa, 1998). También se han reportado regiones con un alto contenido de prolina (Pro) y glutamina (Gln) presentes en proteínas reguladoras de la transcripción (Gerber *et al*., 1994), las cuales intervienen en mecanismos de interacción entre proteínas (Kay *et al*., 2000). Además, se ha sugerido que la presencia de LCRs aumenta la variabilidad de las glicoproteínas de superficie de bacterias patógenas, en donde están involucradas en la producción de diversos patrones de glicosilación, lo cual lleva a la producción de proteínas con variación antigénica que le permiten al patógeno eludir al sistema inmune del hospedero (Hallet, 2001).

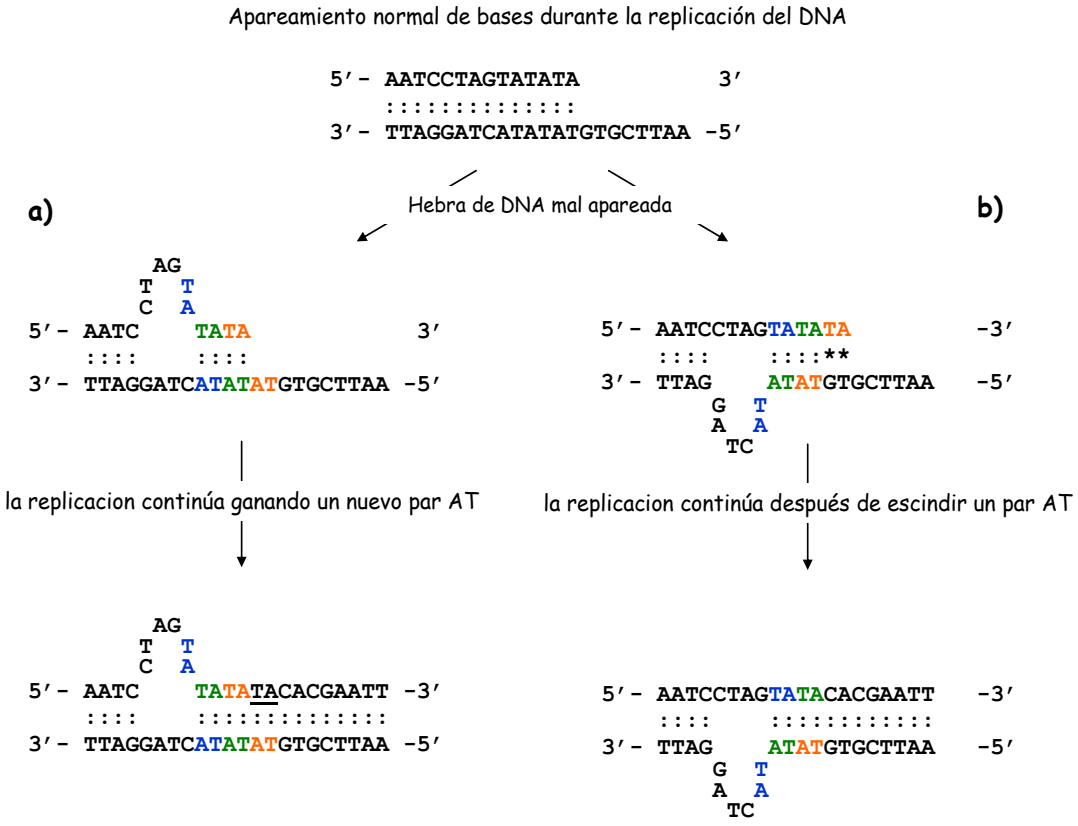


Fig.1 Generación de duplicaciones y deleciones por apareamiento erroneo entre tres repeticiones de AT contiguas (azul, verde, naranja) durante la replicación. a) en caso de que el mal apareamiento se de en la hebra que se está sintetizando, se obtiene la inserción de un par AT (subrayado); b) mientras que, cuando el mal apareamiento se da en la hebra molde se presenta la deleción de un par AT. Tomado y modificado de Li y Graur, 1991.

Por otro lado, la presencia de LCRs se ha asociado a varias patologías neurológicas en humanos, cada una de ellas ligada a un gen diferente. Los ejemplos incluyen la Atrofia Muscular Espinal y Bulbar ligada al cromosoma X, enfermedad de Huntington o Corea de Huntington, Ataxia Espinocerebelosa de tipo 1, atrofia dentatorubro-palidoluysiana y la enfermedad Machado-Joseph, también llamada Ataxia Espinocerebelosa de tipo 3 (Djian *et al*., 1996). Cada una de estas enfermedades está asociada a un número elevado de repeticiones en tandem del codón CAG que codifica para Gln. En el caso de la enfermedad de Huntington, el gen asociado a esta es *HTT*, el cual en condiciones normales codifica para 20 residuos de Gln, y cuando este número alcanza entre 35 a 40 residuos es cuando se desarrolla esta patología. Sin embargo, el número puede elevarse a más de 100 en generaciones sucesivas, lo cual provoca que estas enfermedades se manifiesten en etapas muy tempranas de la vida (Djian *et al*., 1996).

LCRs EN SISTEMAS VIRALES

Desde su descubrimiento, las LCRs han sido estudiadas sobre todo en sistemas celulares. Sin embargo, a principios de la década de los 80s se reportó por primera vez la presencia de LCRs en el genoma del virus de Epstein-Barr (Heller *et al*., 1982). Desde entonces se ha reportado la presencia de LCRs ya sea en genomas o en proteínas virales, como por ejemplo en el virus núcleo-citoplásmico grande de DNA (NCLDV) (Ogata y Claverie, 2007), alfavirus (Perera *et al*., 2001), HTLV-1 y Epstein-Barr (Cristilllo *et al*., 2001), y varios potivirus (Hancock *et al*., 1995). Se ha sugerido que el patinaje de replicación que origina LCRs en sistemas celulares también es el responsable de la generación de estas regiones en los genomas virales tanto de DNA como de RNA (Hancock *et al*., 1995).

LCRs EN EL VIRUS DE INMUNODEFICIENCIA HUMANA TIPO 1 (VIH-1)

El VIH-1 es un miembro de la Familia *Retroviridae* con genoma complejo de 9 a 11 Kb, constituido por 9 genes, de enorme importancia médica y de gran interés científico ya que es el agente causal del Síndrome de Inmunodeficiencia Adquirida (SIDA). Gabrielian y Bolshoy, (1999) fueron los primeros en reportar la presencia de LCRs en un genoma de HIV-1, aunque sin indicar en que región del genoma se encontraban. Este primer reporte fue seguido por el trabajo de Chen

et al. (2009), quienes también identificaron SS en el genoma del VIH-1 en repeticiones de di, tri, tetra, penta y hexanucleótidos, siendo las más comunes las formadas por el dinucleótido $(GA)_n$, donde n tiene valores de 4 a 13. La LCR mas larga que se encontró en todos las cepas de VIH-1 analizadas corresponde al hexanucleótido $(AAGAGG)_3$ repetido tres veces constituyendo una secuencia de 18 nucleótidos. Por otro lado, las proteínas accesorias Tat y Rev comparte la misma LCR: RKKRRQRRRPPQNS (Wootton, 1994) Esta región rica en arginina (Arg) es un motivo bifuncional que participa tanto en la unión a nucleótidos, así como en la localización nuclear (Cullen, 1998; Pollard y Malim, 1998). En el sitio http://www.bioafrica.net, también se reporta la presencia de LCRs en las poliproteínas Pol y Env, así como en las proteínas Vif, Vpu y Nef en el Virus VIH-1 cepa HXB2.

A pesar de su importancia como componentes de genomas y proteínas, de su amplia distribución y de las varias funciones con las que se les han asociado, aún queda mucho por conocer acerca del papel que juegan las LCRs en la estructura y funcionamiento de genomas y proteínas celulares, y aún es mucho menos lo que se sabe acerca del comportamiento de estas regiones en genomas y proteínas virales.

Tomando en cuenta lo anterior, este trabajo reporta los resultados del análisis de una base de datos de proteína de superficie gp120 del VIH-1, la cual presenta alta variabilidad (Freed y Martin, 1995), para identificar la presencia de LCRs en su estructura y tratar de entender que efecto tienen en esta proteína así como en el segmento del genoma que la codifica.

P U B L I C A C I O N E S

A R T Í C U L O

Journal of Theoretical Biology 338 (2013) 80–86

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Low complexity regions (LCRs) contribute to the hypervariability of the HIV-1 gp120 protein

Ana María Velasco [a,b], Arturo Becerra [a], Ricardo Hernández-Morales [a], Luis Delaye [a,1], María Eugenia Jiménez-Corona [b,2], Samuel Ponce-de-Leon [b], Antonio Lazcano [a,*]

[a] Facultad de Ciencias, UNAM, Ciudad Universitaria, Apdo. Postal 70-407, México D. F. 04510, Mexico
[b] Laboratorios de Biológicos y Reactivos de México, Amores 1240, Colonia Del Valle, México D. F. 03100, Mexico

HIGHLIGHTS

- We study the presence of low complexity regions (LCRs) in HIV-1 gp120 protein.
- LCRs were identified in the hypervarible region V1, V2, V4 and V5.
- A high number of glycosylation sites were found in LCRs.
- Our results suggest that LCRs are an important source of antigenic variation in HIV-1 gp120 protein.

ARTICLE INFO

Article history:
Received 10 January 2013
Received in revised form
1 August 2013
Accepted 31 August 2013
Available online 8 September 2013

Keywords:
Human immunodeficiency virus
Hypervariable regions
Glycosylation sites
Low complexity regions
LCRs

ABSTRACT

Low complexity regions (LCRs) are sequences of nucleic acids or proteins defined by a compositional bias. Their occurrence has been confirmed in sequences of the three cellular lineages (Bacteria, Archaea and Eucarya), and has also been reported in viral genomes. We present here the results of a detailed computer analysis of the LCRs present in the HIV-1 glycoprotein 120 (gp120) encoded by the viral gene *env*. The analysis was performed using a sample of 3637 Env polyprotein sequences derived from 4117 completely sequenced and translated HIV-1 genomes available in public databases as of December 2012. We have identified 1229 LCRs located in four different regions of the gp120 protein that correspond to four of the five regions that have been identified as hypervariable (V1, V2, V4 and V5). The remaining 29 LCRs are found in the signal peptide and in the conserved regions C2, C3, C4 and C5. No LCR has been identified in the hypervariable region V3. The LCRs detected in the V1, V2, V4, and V5 hypervariable regions exhibit a high Asn content in their amino acid composition, which very likely correspond to glycosylation sites, which may contribute to the retroviral ability to avoid the immune system. In sharp contrast with what is observed in gp120 proteins lacking LCRs, the glycosylation sites present in LCRs tend to be clustered towards the center of the region forming well-defined islands. The results presented here suggest that LCRs represent a hitherto undescribed source of genomic variability in lentivirus, and that these repeats may represent an important source of antigenic variation in HIV-1 populations. The results reported here may exemplify the evolutionary processes that may have increased the size of primitive cellular RNA genomes and the role of LCRs as a source of raw material during the processes of evolutionary acquisition of new functions.

1. Introduction

It has generally been assumed that due to the compactness and small size of viral genomes, the only mechanisms underlying viral diversity are point mutations and recombination events (Keese and Gibbs, 1993). In cells, however, other mechanisms exist that can generate diversity at different levels, including genome duplication, complete and internal gene duplication, as well as the acquisition of genetic material by horizontal transfer (Petrov, 2001). The generation of low complexity regions (LCRs) is recognized as a process that can enhance sequence variability, and also as a mechanism that can lead to variations in genome size (Hancock, 2002).

LCRs, also known as simple sequences, are regions of nucleic acids or proteins that present a bias in the composition of their residues. They often include repeats of one or more residues (Wootton, 1994), and are known to exhibit a high evolutionary

* Corresponding author. Tel.: +52 5/5622 4823; fax: +52 5/5622 4828.
E-mail address: alar@ciencias.unam.mx (A. Lazcano).
[1] Current address: Departamento de Ingeniería Genética, Cinvestav-Irapuato, Km. 9.6 Libramiento Norte, Carretera Irapuato-León, 36821 Irapuato, Guanajuato, Mexico.
[2] Current address: Dirección General de Epidemiología, Secretaria de Salud, Francisco de P. Miranda #177, 01480 México D.F., Mexico.

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.08.039

rate (Huntley and Golding, 2000). The mechanisms that have been proposed to explain their origin are (a) unequal crossover recombination events and (b) errors during DNA replication due to slippage (Hancock, 1995; DePristo et al., 2006), both of which lead to the appearance of repetitive patterns in sequences (Katti et al., 2001). LCRs have been found in nucleotide and protein sequences of organisms of the three major cellular lineages, i.e., Bacteria, Archaea and Eucarya (Hancock, 2002; Heringa, 1998; Andrade et al., 2001), where they can be present in different types of proteins with different functions. Although viral LCRs were first reported over thirty years ago in the Epstein–Barr virus (Heller et al., 1982), with few exceptions the analysis and characterization of these sequences have been studied mostly in cellular genomes. However, reports of LCRs in viral sequences, for example in potyviruses (Hancock et al., 1995), alphavirus (Perera et al., 2001), HTLV-1 (Cristillo et al., 2001) and HIV-1 (Gabrielian and Bolshoy, 1999; Chen et al., 2009; Bioafrica.net), support the possibility that LCRs also represent a source of variability in viral proteins. In order to test this possibility, we have studied in detail their presence in complete gp120 proteins identified in 3637 Env polyprotein sequences from 4117 HIV-1 genomes available in public databases as of December 2012. This analysis was undertaken as part of an ongoing study of completely sequenced HIV-1 genomes in order to assess the different rates of evolution of the viral components. While the relatively slow rate of evolution of the HIV-1 reverse transcriptase is well understood, the different mechanisms underlying the high rate of variability of HIV-1 surface protein deserve considerable attention.

The HIV-1 genome is an example of the streamlining processes that shape viral genomes. In spite of its small size, the HIV-1 genome is a complex structure that encodes structural and catalytic components, as well as an array of regulatory sequences typical of lentiviruses. The HIV-1 genome consists of nine genes: *gag*, *pol*, *env*, *vif*, *vpr*, *tat*, *rev*, *vpu* and *nef* (Fig. 1). The *env* gene encodes for the different envelope proteins. The signal peptide (SP) is located at the amino end of the Env polyprotein, and is followed by the surface (SU) and transmembranal (TM) protein sequences (Fig. 1). The availability in public databases of a significant high number of completely sequenced HIV-1 genomes has allowed us to undertake a detailed analysis of this virus in order to study the presence and distribution of LCRs in the gp120 protein.

Here we report the results of a detailed search analysis and characterization of LCRs in the HIV-1 gp120 protein, using a sample of 4117 completely sequenced genomes available in public databases as of December 2012. Our results show that LCRs are present in 31.4% of the HIV-1 gp120 proteins analyzed here. In some cases a single gp120 protein may be endowed with more than one LCR. A total of 1258 LCRs were found in the 1143 of the 3637 gp120 proteins analyzed. Our results indicate that in our sample three LCRs are present in the C4 domain, and one in each of the conserved domains C2, C3, and C5; and 23 in the signal peptide. The remaining 1229 LCRs are all located in four particular hypervariable regions of its sequence (V1, V2, V4 and V5). As argued here, the presence and location of LCRs in HIV-1 gp120 suggest that they play an important role in enhancing the antigenic variation in HIV-1 populations.

2. Materials and methods

2.1. Construction of HIV-1 protein databases from completely sequenced genomes

In the work reported here only those HIV-1 genomes with a complete set of genes were selected from the reported databases. A primary database was constructed with 4117 completely sequenced genomes of HIV-1 (Supplementary Table 1) from the Los Alamos National Laboratory and the NCBI databases (National Center for Biotechnology Information; ⟨www.lanl.gob⟩) available as of December 2012. Subsequently, a database of gp120 proteins for these HIV-1 genomes was constructed. Due to the presence of numerous end codons along the sequence of several genes, which do not produce proteins, complete gp120 could be obtained for 3637 of a total of 4117 viral genomes available. This represents 88.34% of the total sample.

2.2. Identification of LCR and estimate of their amino acid composition

In order to identify LCRs in our protein databases a search was performed using the SEG program (Wootton and Federhen, 1993), with the following parameters: window=12; locut=1.9; hicut=2.1; and −l option for protein analysis. The window value refers to a

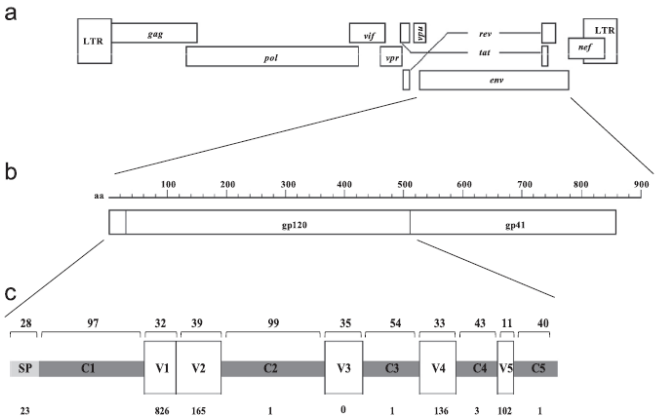


Fig. 1. (a) Schematic representation of the HIV genome; (b) structure of the *env* gene, which encodes for the Env polyprotein and gp120 protein. The Env polyprotein is composed by gp120 and gp41; (c) the gp120 protein sequence includes the signal peptide (SP) and the surface protein (SU) (Yamaguchi-Kabata and Gojobori, 2000). The hypervariable (V1, V2, V3, V4 and V5) regions are shown in white boxes, and the conserved regions are marked in gray boxes (C1, C2, C3, C4 and C5). The numbers above them indicate the length of each region (in number of amino acids). The numbers under the boxes indicate the number of sequences that exhibit LCRs. The genome and proteins of HIV-1 HXB2 (K03455) were used as reference system.

segment of 12 amino acids as the size of the window in which the search for LCRs is performed. Locut is the minimal detectable complexity value in the window analyzed, whereas hicut represents the upper limit of complexity value that the program can detect in the region defined by the window. These values allowed to the identification of LCRs and the measure of complexity at amino acid level in the sequences derived from the databases we have used. High complexity values correspond to LCRs with more variable amino acid composition in the window analyzed, while low complexity values indicate less variability in the amino acid composition. The amino acid contents of proteins and LCRs were estimated using the aacomp program (Pearson and Lipman, 1988).

2.3. Classification of LCR and function identification

A multiple sequence alignment was constructed for the HIV-1 gp120 proteins with LCRs, using the default set from muscle 3.8.31 (Edgar, 2004). In these alignments the LCRs are marked. Proteins from the 9719 bp HIV-1 HXB2 complete genome (K03455) were used as reference system in order to identify the gp120 protein sequences in the translated *env* genes, as well in the Env polyproteins we have analyzed. An inventory of the LCRs was carried out based on their size and position within protein primary structure. The gp120 protein domains were identified on the sequence of the gp120 reference protein, and in it we mapped the LCRs found in our study.

2.4. Tertiary structure localization

Both the tertiary structure and the amino acid sequences of HIV-1 gp120 protein were obtained from the PDB database (RCSB Protein Data Bank PDB): 1g9n. In order to locate the LCRs in the protein tertiary structure, the PDB amino acid sequence was aligned as mentioned above with the sequences of the database of gp120 proteins that we have analyzed. Tertiary structures were visualized with PyMOL v1.2r1 (The PyMOL Molecular Graphics System).

3. Results

3.1. LCRs in the HIV-1 gp120 protein

In this study we have analyzed the presence and amino acid composition of LCRs in HIV-1 gp120 proteins in a sample of 4117 completely sequenced genomes of HIV-1 available in the Los Alamos National Laboratory and the NCBI databases (National Center for Biotechnology Information; ⟨www.lanl.gob⟩) as of December 2012.

Of the 3637 gp120 proteins we have analyzed, 1143 (31.4%) exhibit LCRs. We have identified 1258 LCRs in this set of 1143 proteins. This discrepancy is due to the cases (107) in which some gp120 proteins exhibit more than one LCR (Supplementary Table 2).

All the 1143 gp120 proteins with LCRs were aligned as described in Section 2, and the location of each LCR was identified in the alignments (Fig. 1). In order to standardize our results, the positions of the LCRs in the HIV-1 gp120 proteins were determined using as a reference the gp120 sequence of the HIV-1 HXB2 strain (Korber et al., 1998). Completion of this mapping process led to the identification of four different sites in the gp120 sequence with LCRs, which are distributed along the protein (Fig. 1). It should be noted that none of these 1143 gp120 sequences contain simultaneously the four different LCRs (Supplementary Table 2).

A low complexity composition corresponds to an unusual content of amino acid or nucleotide sequences, where an over-representation of an amino acid or nucleotide is observed. This can be estimated using the SEG program (Wootton and Federhen, 1993). High complexity values thus correspond to LCRs with a more variable amino acid composition, while low values indicate less variability in the amino acid composition. For comparison, an estimate of the LCRs in a random sample of 374 Pol polyprotein sequences (which correspond to 10% of the retroviral genomes analyzed here) showed that only 34 of them exhibit LCRs with a length between 12 and 16 amino acids and with complexity values that range from 1.83 and 2.07. These LCRs are located in the reverse transcriptase protein. We have found no LCRs in the protease or integrase sequences. On the other hand, the complexity values of the 1258 LCRs identified in our sample, range from 0 to 2.6 (Supplementary Fig. 1). A complexity value of 0 corresponds, for instance, to the GlyGlyGlyGlyGly sequence in the V5 region of the FJ495937 gp120 protein (Supplementary Table 3).

Although there is considerable dissimilarity in the amino acid composition between the different LCRs reported here, in all of them the most abundant amino acid is Asn, followed by Ser, Thr and Gly. These abundances show variations in the different LCRs discussed here, as shown in Table 1. In the hypervariable sequences of gp120 proteins (both with or without LCRs), Asn is also the most amino acid, but is followed by Thr, Ser and Ile (Supplementary Table 4). There is substantial variation in the length of the different LCRs, with sizes ranging from 5 to 36 amino acids (Supplementary Fig. 1). The length of the different LCRs compared to that of the gp120 protein from HIV-1 HXB2 strain, represents a figure that ranges from 1.03% to 7.04% of the actual size of the protein.

The gp120 protein is involved in the virion binding to CD4, the primary membrane receptor that recognizes HIV-1 (Freed and Martin, 1995). This 483-amino acid protein presents five conserved regions (C1–C5), which are interspersed with five hypervariable regions (V1–V5) (Fig. 1). The hypervariable regions exhibit small deletions and insertions, and present a low degree of conservation (25% or less). These hypervariable regions are all found in loop-like structures at the surface of the protein (Modrow et al., 1987). The HIV-1 gp120 hypervariable regions are associated with the viral ability to evade the host immune system (Poignard et al., 2001). No LCR has been identified in the hypervariable region V3. The LCRs that we have identified in gp120 are all located in the hypervariable regions V1, V2, V4 and V5 (Fig. 1). In all of them, Asn is the most abundant amino acid, with values between 37.23% and 42.22% (Table 1).

Table 1
Amino acid composition of LCRs present in regions V1, V2, V4, and V5 of the gp120 protein (see Fig. 1). Amino acids are ordered by their overall abundance in the corresponding LCR. The total number (n) of amino acids in each LCR is indicated at the lower end of the corresponding column.

aa	V1	V2	V4	V5
ASN(N)	33.75	39.23	35.38	41.96
THR(T)	30.68	5.65	23.01	20.27
SER(S)	16.74	33.49	16.09	2.14
GLY(G)	3.14	2.80	5.71	22.14
ASP(D)	1.33	3.65	2.15	8.93
GLU(E)	0.27	2.85	3.56	2.86
ILE(I)	1.96	1.00	4.30	0.27
LYS(K)	1.47	4.27	1.10	0.63
TYR(Y)	0.36	3.23	0.94	0.00
ALA(A)	3.04	0.24	0.52	0.00
VAL(V)	2.44	0.38	0.21	0.00
TRP(W)	0.09	0.00	3.35	0.00
LEU(L)	1.15	0.09	1.47	0.00
ARG(R)	0.59	1.09	0.31	0.18
PRO(P)	1.10	0.14	0.16	0.00
GLN(Q)	0.15	0.81	0.31	0.27
CYS(C)	0.76	0.19	0.21	0.09
PHE(F)	0.16	0.00	0.79	0.00
MET(M)	0.40	0.57	0.05	0.00
HIS(H)	0.12	0.19	0.10	0.00
UNK(X)	0.29	0.14	0.26	0.27
	$n=12885$	$n=2108$	$n=1908$	$n=1120$

Table 2
Inventory of the potential glycosylation sites in the gp120 LCRs discussed in this work. The first and third columns indicate the amino acid composition of the two possible glycosylation sites using the single letter code. The second and fourth columns correspond to the frequency of possible glycosylation sites (X, any amino acid).

N–X–S	Frequency	N–X–T	Frequency
NAS	7	NAT	138
NCS	8	NCT	83
NDS	9	NDT	133
NES	4	NET	17
NFS	1	NFT	8
NGS	15	NGT	322
NHS	1	NHT	1
NIS	20	NIT	107
NKS	2	NKT	17
NLS	1	NLT	53
NNS	336	NMT	2
NPS	65	NNT	481
NQS	1	NQT	5
NRS	2	NRT	9
NSS	548	NST	365
NTS	105	NTT	586
NVS	25	NVT	86
NWS	2	NWT	8
NYS	13	NYT	25
NXS	1	NXT	8

The observed abundance of Asn is probably due to the location of these LCRs in hypervariable sites, where this amino acid is known to undergo a high level of glycosylation (Go et al., 2008; Marshall, 1974). The glycosylation process involves the addition of an oligosaccharide to a specific Asn residue, which is known to be part of the triplet sequence Asn–X–Thr/Ser, where X can be any other amino acid (Marshall, 1974). A search for possible N-linked glycosylation sites in the LCRs sequences allowed the identification of 3620 different potential sites in the data set reported here. As shown in Table 2, 2454 of these 3620 potential glycosylation sites correspond to the sequence Asn–X–Thr, and 1166 to Asn–X–Ser. A high number of glycosylation sites are present in the sequences of LCRs, where they vary from 9.1% to 100% of the total length of the LCRs. This provides the virus with a high variability in glycosylation patterns. There are major differences in the distribution of the potential glycosylation sites between the gp120 sequences lacking LCRs, and those that have them. In the former, some glycosylation sites are highly conserved, but in general tend to be distributed along the hypervariable regions. In contrast, in the LCRs reported here the glycosylation sites tend to be clustered in well-defined islands located towards the center of the LCRs (Supplementary Fig. 2).

Overlap of glycosylation sites is observed in the hypervariable regions of gp120 sequences with and without LCRs. There are 707 overlapping glycosylation sites in the hypervariable regions of the gp120 sequences with LCRs, and only 805 in those without LCRs, i.e., overlap is twice as abundant in the glycosylation sites present in LCRs. As summarized in Supplementary Table 3, only 70 of the 1258 LCRs reported here do not exhibit this kind of signaling sequence. The remaining 1188 LCRs exhibit one to eight potential glycosylation sites. The different numbers of glycosylation sites identified in the 1188 LCRs are shown in Table 2.

3.2. The localization of LCRs in protein tertiary structures

The availability of complete or partial tertiary structures of several HIV-1 proteins in the PDB database (RCSB Protein Data Bank PDB), has allowed the identification of the spatial location of the LCRs in the gp120 three-dimensional structure (Fig. 2). The gp120 protein is composed of five α-helix and 25 β-sheets (Kwong et al., 2000). The variable regions V1 and V2 are located between

Fig. 2. Positions of LCRs in the tertiary structure of HIV-1 gp120 protein (green color). PDB structure 1g9n was obtained from PDB (http://www.rcsb.org). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the β-2 and β-3 sheets; V3 is flanked by β-12 and β-13; V4 is present between β-18 and β-19; and finally, V5 is located between the β-23 and β-24 sheets (Kwong et al., 2000). The variable regions have a loop structure, and due to their high mutation rate they play a key role in escaping the host immune system (Yamaguchi-Kabata and Gojobori, 2000). As noted above, 1229 LCRs identified in the gp120 proteins discussed in this study fall in four of the well-defined hypervariable regions, i.e., the V1, V2, V4 and V5 regions. No LCRs are present in the V3 region. There are 826 LCRs in V1, 165 in V2, 136 in V4 and 102 in V5. The signal peptides of 23 gp120 proteins are endowed with one LCR each (Supplementary Table 3). One LCR is present in each of the conserved regions C2, C3 and C5, and three more are located in C4.

Although the complete tertiary structure of the gp120 protein is not yet available, we have identified the positions of LCRs in the V4 and V5 hypervariable regions in the partial structure. As expected, all the LCRs are located at the loops that constitute the variable regions (Fig. 2).

4. Discussion

Although viral LCRs were first reported over thirty years ago in the Epstein–Barr virus (Heller et al., 1982), with few exceptions analysis and characterization of these sequences have been studied mostly in cellular genomes. Currently, very little is known about LCRs and their effect in viral evolution, despite the reports that confirm their presence in the sequences of different RNA virus (Hancock et al., 1995; Perera et al., 2001; Cristillo et al., 2001; Gabrielian and Bolshoy 1999; Chen et al., 2009; Bioafrica.net). In order to analyze this issue, we have undertaken the study of the presence, distribution and characterization of LCRs in the HIV-1 gp120 protein. As a part of an ongoing comparative analysis of completely sequenced HIV-1 genomes available in the Los Alamos National Laboratory HIV and GenBank databases as of December 2012, we have analyzed gp120 sequences obtained from 4117 HIV-1 genomes. The results presented here include several HIV-1 subtypes (Supplementary Table 1). A total of 3637 gp120 sequences were analyzed, of which 1143 of them exhibit LCRs. In this sample we have detected a total of 1258 LCRs, which are distributed in four different positions of the gp120 protein sequence. The LCRs are in located in four of the five hypervariable regions (V1, V2, V4 and V5), which are known to play a key role in

the ability of the virus to evade the immune system of their human host.

A high level of glycosylation takes place in the hypervariable regions due to the presence of approximately 24 potential sites where this reaction can occur (Leonard et al., 1990). As summarized in Table 2, the most abundant residues in all the four LCRs reported here are Asn, Ser, Thr and Gly. This suggests the existence of many possible glycosylation sites. Indeed, we have detected a total of 3620 possible glycosylation sites in the LCRs reported here. No glycosylation sites are present in the LCRs located in the conserved regions (C2, C3, C4 and C5) or in the signal peptide. A multivariate analyses of variance (MANOVA) was performed in order to determine if there are significant differences between potential glycosylation sites found in LCRs in the hypervariable regions, and the remaining regions of the same sequences which are endowed with a higher level of complexity (hereinafter, HCR). A total of 5041 glycosylation sites were detected in the hypervariable regions V1, V2, V4 and V5 of the gp120 proteins with LCRs. Of these, a total of 3620 are located in the LCRs and 1421 in the HCRs of these same gp120 sequences (Supplementary Table 5). Significant differences were found ($p=0.0001$) between the number of glycosylation sites located in LCRs, and those in the HCR (Fig. 3). As noted above, in LCRs these glycosylation sites tend to be clustered in well-defined islands located towards the center of the low complexity regions, which most likely increases their chances of undergoing the glycosylation reaction. There is a total of 707 overlapping glycosylation sites in the 1143 gp120 sequences with LCRs, compared to 805 in 2494 gp120 proteins lacking low complexity regions.

The possible glycosylation sites detected in the low complexity regions can be divided in two different types of motifs, Asn–X–Thr and Asn–X–Ser. As noted above and as summarized in Supplementary Table 6, both types of glycosylation sites are present in comparable sites both in the variable (V) and in the conserved (C) regions of gp120 proteins with ($n=1143$) or without LCRs ($n=2494$).

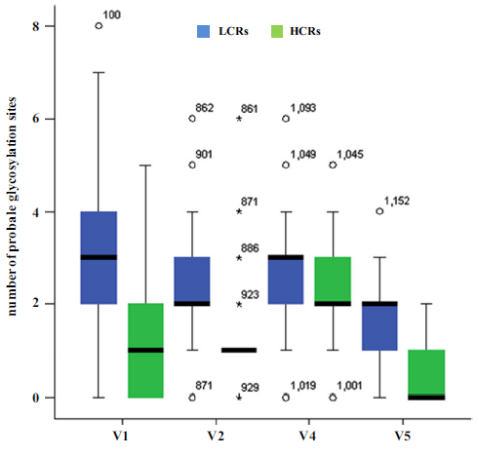


Fig. 3. Box plot diagram indicating the results of a multivariate analyses of variance (MANOVA) to determine the differences between potential glycosylation sites found in LCRs in the hypervariable regions (blue), and the remaining regions of the same sequences which are endowed with a higher level of complexity (HCR, in green). The absence of an overlap in each pair of regions analyzed demonstrates the difference in the number of glycosylation sites located in LCRs, and those in the HCR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Theoretically the high density of glycosylation sites present in the sequences of LCRs should provide the virus with a high variability in glycosylation patterns. However, the glycosylation efficiency is not equal in the two motifs. As shown by Kasturi et al. Kasturi et al. (1995), the Asn–X–Thr motif presents a glycosylation efficiency of 92%, while the corresponding value for Asn–X–Ser sequence is 52%. Glycosylation is blocked when X corresponds to Pro (Gavel and von Heijne, 1990). In the case of Asn–X–Ser motifs, the efficiency of glycosylation depends on the X amino acid. If X is Trp, Asp or Glu, an inefficient glycosylation take place (5, 19 and 24% respectively). An intermediate glycosylation level (43%) is observed when X is Leu, but it occurs at high levels if Phe or Ser are present (70% and 97% respectively) (Kwong et al., 2000).

Empirical data demonstrate that the most abundant glycosylation sites are the Asn–X–Thr motifs (Kwong et al., 2000). As shown in Table 2, our results indicate that in the motif Asn–X–Thr, Pro is always absent. In the case of the Asn–X–Ser sites identified in this work, the X corresponds to any amino acid, except Met (Table 2). These variants have been associated with glycosylation efficiencies that vary from 70% to 97% (Shakin-Eshleman et al., 1996). These motifs provide the gp120 protein with a high number of glycosylation sites, leading to sequences with a new immunogenicity that would allow the virus to avoid the response of the immune system (Shakin-Eshleman et al., 1996). On the other hand, it is notable that no LCRs were detected in V3, which is the predominant antigenic region of gpl20 (Simmonds et al., 1990). The significance of this hypervariable region as a potential target for structure-based drug design has been discussed extensively (Shakin-Eshleman et al., 1996; Wei et al., 2003; Sirois et al., 2005).

On the other hand, since no LCRs have been identified in the protease sequences of the random sample of 374 Pol polyprotein discussed here, we cannot apply the Chou distorted key theory (Sirois et al., 2005; Chou, 1996) to analyze the significance of our results in the design of therapies against HIV-1. The only possibility of application of the Chou's theory (Sirois et al., 2005; Chou, 1996) to assist in the design of drugs against HIV-1 infection based on the distribution of LCRs in the gp120 protein reported here, would be the design of unspecific antibodies exhibiting high affinity for the variable LCRs described here. Given the observed high rates of variability (Supplementary Table 3), this strategy appears to be of limited value for our case.

The relative abundance of amino acids in LCRs reflects in part their presence in proteins (Table 1), including those with a low biosynthetic cost, such as Ala, Gly, Pro, Ser and Glu (Radó-Trilla and Alba, 2012). It is likely that abundance of Gly in the LCRs reported here reflects not only the relatively high frequency of slippage of the codons encoding this amino acid (Chakraborty et al., 1997), but also its enhancement of chain flexibility, which would not affect negatively disordered coils were the LCRs reported here are located. The overrepresentation of Asn, Ser and Thr in Table 1 can be easily understood in terms of the glycosylation that most eukaryotic secreted- or membrane-bound proteins undergo (Kuriyan et al., 2013) and that plays a key role in molecular recognition. It is possible that their presence reflects, as noted by Marshall (Marshall, 1972), that single-base mutations can lead from the asparagine's AAU and AAC codons to those of serine (AGU and AGC) and threonine (ACU and ACC).

However, the number and abundance of glycosylation sites in V3 is lower than in the others variable region (Go et al., 2008; Simmonds et al., 1990; Wei et al., 2003). Although V3 is one of the largest variable regions, it is endowed with one or two glycosylation sites, while in V1, V2, V4 and V5 two to five glycosylation sites have been reported (Go et al., 2008; Wei et al., 2003). The low presence of glycosylation sites in V3 could be due to the fact that a higher level of glycosylation would interfere with the function of V3 binding to the receptor (Wei et al., 2003).

Previous studies in *Plasmodium falciparum* have shown that its genome A+T content correlates with the number and composition of LCRs, which may facilitate the generation of antigenic diversity (DePristo et al., 2006). In order to test this possibility, a Pearson analysis was performed with our dataset. Our analysis (not shown) shows a value of $p=0.083$, which indicates that no correlations exists between the A+T content of the *env* genes and the number of LCRs identified in the gp120 proteins they encode.

It is quite remarkable that a huge percentage of the LCRs identified in the gp120 sequences reported here are located in the hypervariable regions, and appears to play a central role as source for variability in glycosylation patterns. Overrepresentation of certain amino acids is a trait of the LCRs discussed here, and is also observed in tandem and cryptic amino acid repeats (Simon and Hancock, 2009). Glycosylation sites in LCRs are not randomly distributed along the region, but are clustered in well-defined islands located towards the central part of the low-complexity regions. This is in sharp contrast with the localization of glycosylation sites in the gp120 proteins without LCRs, where they distribution along the hypervariable regions

Like amino acid repeats, LCRs are located in disordered regions of proteins, which in the case of the gp120 protein discussed here, correspond to rapidly evolving regions. Our findings can be explained by a mechanism analogous to that observed in pathogenic bacteria (Moxon, 1999) unicellular eukaryotes such as *Plasmodium* and the Kinetoplastida (Depledge et al., 2007), were homopolymeric tracts and tandem arrays of short repeat motifs located in surface-associated proteins that interact with the host play a key role in immune evasion and influence pathogenicity and virulence (DePristo et al., 2006).

The existence of equivalent LCRs in the variable regions of the gp120 protein, which plays a key role in host recognition and infection, is a very strong indicator of their role in HIV-1 immune evasion. They are also consistent with a previous report of repetitive sequences present in HIV-1 involved in the high variability of glycosylation in V4 (Guglietta et al., 2010). As shown in Suppl. Fig. 1, the LCRs reported here are always found in segments of the gp120 protein that present a high degree of insertions, which show the significance of LCRs in producing variation in gene sizes.

LCRs have been associated with some human neuropathies such as X-linked spinal and bulbar muscular atrophy, Huntington disease, type 1 spinocerebellar ataxia, dentatorubral-pallidoluysian atrophy, and the Machado–Joseph disease, which develop when the proteins associated with these diseases exhibit more than 100 Gln residues in a homopolymeric sequence (Djian et al., 1996). Unfortunately, as of today there are no equivalent studies for viral LCRs. Analysis of other protein sequences in the HIV-1 viruses in our database led to the identification of LCRs in few sequences of the p6Gag that exhibit a duplicated late (L) domain (Pro–Thr–Ala–Pro) Colgrove, 2005. The presence of this duplication has been reported in several cases of maternal-infant HIV infections, and exhibits a trend toward association with lower maternal CD4 count [(Djian et al., 1996; Colgrove, 2005)]. The available databases do not allow us to corroborate an equivalent correlation of the presence of LCRs in both functional and nonfunctional sites in HIV-1 gp120 proteins.

There is evidence that protein regions that have originated in eukaryotes and viruses by replication slippage and unequal crossing-over or other mechanisms, can be co-opted for functional roles (Perera et al., 2001). The LCRs′ high number and variable composition suggest that they are an important source of variability in genomes, and it is tempting to conclude that the LCRs that we have detected play a role in the HIV-1 virus evolution. It could be argued that the presence of LCRs in only one third (31.4%) of the gp120 sequences included in the sample analyzed here speaks against their possible evolutionary role, and that since the generation of LCRs is a random process, these and other simple sequences may be the outcome of a neutral process. However, it is unlikely that this could explain the size, distribution and position of the LCRs in the gp120 proteins studied here, or the number and distribution of potential glycosylation sites located in the LCRs themselves. It is possible that the LCRs described were originated by a replication slippage mechanism and became fixed by a non-neutral process in viral populations, giving rise to variations that allow the viruses to escape from the host immune system. Purifying selection may prevent the incorporation LCRs in the highly conserved regions of the gp120 protein and of other viral proteins.

The LCRs reported in this work exhibit considerable variation in composition, length and distribution. This substantial intra-strain variation in the number, size and composition of LCRs in HIV-1 gp120 proteins demonstrates the stochastic origin of LCRs and can be interpreted as evidence of their polyphyletic origin. As shown by the LCRs present in a significant fraction of the gp120 sequences in our sample which are located in four of the five hypervariable regions (V1, V2, V4 and V5) (Fig. 1), simple sequences may produce antigenic variations as a result of the high rate of change in these regions, that would enable the virus to evade the host immune system without affecting its viability.

Together with previous reports of LCRs in the genomes of potyviruses (Hancock et al., 1995), alphavirus (Perera et al., 2001), HTLV-1 (Cristillo et al., 2001) and HIV-1 (Gabrielian and Bolshoy, 1999; Chen et al., 2009; Bioafrica.net), the results presented here demonstrate that in spite of the strong streamlining pressures acting over viral genomes, the appearance and maintenance of LCRs may play an important role as a source of genetic variation and possibly in genome evolution. Although the actual antiquity of the different viral groups remains an open issue, these reports exemplify the evolutionary processes that may have increased the size of primitive cellular RNA genomes, and provide support to the proposal that the random generation of homopolymeric tracts and tandem arrays of short repeat motifs may have been a source of raw material during the processes of evolutionary acquisition of new functions in early cellular evolution prior to the emergence of DNA genomes (Pool et al., 1998).

5. Conclusions

In spite of the strong streamlining pressures acting over viral genomes, the results presented here suggest that the appearance and maintenance of LCRs plays an important role as a source of genetic variation and possibly in genome evolution. Our results confirm the presence and the role of LCRs in population diversity and evolution in HIV-1. The LCRs reported here are all associated with hypervariable regions of the gp120 protein, and appear to enhance the ability of the HIV-1 to evade the immune system. Thus, LCRs can be a source of antigenic variations as a result of the high rate of change in these regions, as shown in the gp120 hypervariable regions. This mechanism, which is analogous to the variations shown in surface proteins of a number of prokaryotic pathogens (Moxon, 1999) and potyviruses (Hancock et al., 1995), suggest that in addition to the high rate of mutation and recombination events described for HIV-1, the random generation of LCRs is a source of antigenic variation that has not been considered as of today. In all cases, the LCRs located in gp120 protein and reported here are associated with hypervariable regions, which provide the HIV-1 virus with the capacity to avoid the host immune system.

Authors′ contributions

Author contributions: AL original idea; AMV collected the data, performed research, and wrote the manuscript, AMV, AL, RHM, LD, AB, MEJC, SPDL analyzed data and revised and discussed critically

the manuscript drafts. All authors have read and approved the final manuscript.

Acknowledgments

We are indebted to Luis Enrique Serrano Gutierrez, José Luis Torres Rodríguez and specially Hector Miguel Camacho and Alvaro García Pérez for their help with computer tools and data recovery, as well as to Sara Islas for help with the manuscript. Work reported here has been supported by a CONACYT Fellowship to AMV. The support of CONACYT (50520-Q) to AL and (100199) to AB is also gratefully acknowledged. Part of the work reported here was completed during a sabbatical leave of absence of AB, with support of DGAPA-UNAM, where he enjoyed the hospitality of Prof. Juli Peretó at the Instituto Cavanilles (Valencia, Spain). The support of the Posgrado en Ciencias Biológicas of the Universidad Nacional Autónoma de México to AMV is gratefully acknowledged.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2013.08.039.

References

Andrade, M.A., Perez-Iratxeta, C., Ponting, C.P, 2001. Protein repeats: structures, functions, and evolution. Journal of Structural Biology 134, 117–131.
Bioafrica.net (⟨http://www.bioafrica.net⟩).
Chakraborty, R., Kimmel, M., Stivers, D.N., Davison, L.J., Deka, R., 1997. Relative mutation rates of tri- and tetranucleotide microsatellite loci. Proceedings of the National Academy of Sciences 94, 1041–1046.
Chen, M., Tan, Z., Jiang, J., Li, M., Chen, H., Shen, G., Yu, R., 2009. Similar distribution of simple sequence repeats in diverse complete Human immunodeficiency virus type 1 genomes. FEBS Letters 583, 2959–2963.
Chou, K.C., 1996. Prediction of human immunodeficiency virus protease cleavage sites in proteins. Analytical Biochemistry 233, 1–14.
Colgrove, R.C., Millet, A., Bauer, G.R., Pitt, J., Welles, S.L., 2005. Gag-p6 Tsg101 binding site duplication in maternal-infant HIV infection. AIDS Research and Human Retroviruses 21, 191–199.
Cristillo, A.D., Mortimer, J.R., Barrette, I.H., Lillicrap, T.P., Forsdyke, D.R., 2001. Double-stranded RNA as a not-self alarm signal: to evade, most viruses purine-load their RNAs, but some (HTLV-1, Epstein–Barr) pyrimidine-load. Journal of the Theoretical Biology 208, 475–491.
DePristo, M.A., Zilversmit, M.M., Hartl, D.L., 2006. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. Gene 378, 19–30.
Depledge, D.P., Lower, R.P.J., Smith, D.F., 2007. A database of amino acid repeats present in lower eukaryotic pathogens. BMC Bioinformatics 8, 122, http://dx.doi.org/10.1186/1471-2105-8-122.
Djian, P., Hancock, J.M., Chana, H.S., 1996. Codon repeats in genes associated with human diseases: fewer repeats in the genes of nonhuman primates and nucleotide substitutions concentrated at the sites of reiteration. Proceedings of the National Academy of Sciences USA 93, 417–421.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792–1797.
Freed, E.O., Martin, M.A., 1995. The role of human immunodeficiency virus type 1 envelope glycoproteins in virus infection. Journal of Biological Chemistry 270, 23883–23886.
Gabrielian, A., Bolshoy, A., 1999. Sequence complexity and DNA curvature. Computer and Chemistry 23, 263–274.
Gavel, Y., von Heijne, G., 1990. Sequence differences between glycosylated and non-glycosylated Asn–X–Thr/Ser acceptor sites: implications for protein engineering. Protein Engineering 3, 433–442.
Go, E.P., Irungu, J., Zhang, Y., Dalpathado, D.S., Liao, H.X., Sutherland, L.L., Alam, S.M., Haynes, B.F., Desaire, H., 2008. Glycosylation site-specific analysis of HIV envelope proteins (JR-FL and CON-S) reveals major differences in glycosylation site occupancy, glycoform profiles, and antigenic epitopes' accessibility. Journal of Proteome Research 7, 1660–1674.
Guglietta, S., Pantaleo, G., Graziosi, C., 2010. Long sequence duplications, repeats, and palindromes in HIV-1 gp120: length variation in V4 as the product of misalignment mechanism. Virology 399, 167–175.
Hancock, J.M., 1995. The contribution of slippage-like processes to genome evolution. Journal of Molecular Evolution 41, 1038–1047.
Hancock, J.M, 2002. Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. Genetica 115, 93–103.
Hancock, J.M., Chaleeprom, W., Chaleeprom, W., Dale, J., Gibbs, A., 1995. Replication slippage in the evolution of potyviruses. Journal of General Virology 76, 3229–3232. ⟨http://www.lanl.gob⟩.
Heller, M., van Santen, V., Kieff, E., 1982. Simple repeat sequence in Epstein–Barr virus DNA is transcribed in latent and productive infections. Journal of Virology 44, 311–320.
Heringa, J., 1998. Detection of internal repeats: how common are they? Current Opinion in Structural Biology 8, 338–345.
Huntley, M., Golding, G.B., 2000. Evolution of simple sequences in proteins. Journal of Molecular Evolution 51, 131–140.
Kasturi, L., Eshleman, J.R., Wunner, W.H., Shakin-Eshleman, S.H., 1995. The hydroxy amino acid in an Asn–X–Ser/Thr sequon can influence N-linked core glycosylation efficiency and the level of expression of a cell surface glycoprotein. Journal of Biological Chemistry 270, 14756–14761.
Katti, M.V., Ranjekar, P.K., Gupta, V.S, 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Molecular Biology and Evolution 18, 1161–1167.
Keese., P., Gibbs, A., 1993. Plant viruses: master explorers of evolutionary space. Current Opinion in Genetics and Development 3, 873–877.
Korber, B., Foley, B.T., Kuiken, C., Pillai, S.K., Sodroski, J.G., 1998. Numbering Positions in HIV Relative to HXB2CG. In: Korber, B., Kuiken, C.L., Foley, B., Hahn, B., McCutchan, F., Mellors, J.W., Sodroski, J. (Eds.), Human Retroviruses and AIDS, Theoretical Biology and Biophysics Group. Los Alamos National Laboratory, Los Alamos, NM, pp. 102–111.
Kuriyan, J., Konforti, B., Wemmer, D., 2013. The molecules of life: physical and chemical reactions. Garland Science, New York, pp. 102–104.
Kwong, P.D., Wyatt, R., Majeed, S., Robinson, J., Sweet, R.W., Sodroski, J., Hendrickson, W.A., 2000. Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. Structure 8, 1329–1339.
Leonard, C.K., Spellman, M.W., Riddle, L., Harris, R.J., Thomas, J.N., Gregory, T.J., 1990. Assiggnment of intrachain disulfide bonds and characterization of potential glycosylation sites of the type 1 recombinant human immunodeficiency virus envelope glycoprotein (gp120) expressed in chines hamster ovary cells. Journal of Biological Chemistry 265, 10373–10382.
Marshall, R.D., 1972. Glycoproteins. Annual Review of Biochemistry 41, 673–702.
Marshall, R.D., 1974. The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins. Biochemical Society Symposium 40, 17–26.
Modrow, S., Hahn, B.H., Shaw, G.M., Gallo, R.C., Wong-Staal, F., Wolf, H., 1987. Computer-assisted analysis of envelope protein sequences of seven human immunodeficieny virus isolates: prediction of antigenic epitopes in conserved and variable regions. Journal of Virology 61, 570–578.
Moxon, E.R., 1999. Whole-genome analysis of pathogens. In: Stearns, S.C. (Ed.), Evolution in Health and Disease. Oxford University Press, New York, pp. 191–204.
National Center for Biotechnology Information (⟨http://www.ncbi.nlm.nih.gov⟩).
Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequences comparison. Proceedings of the National Academy of Sciences USA 85, 2444–2448.
Perera, R., Owen, K.E., Tellinghuisen, T.L., Gorbalenya, A.E., Kuhn, R.J., 2001. Alphavirus nucleocapsid protein contains a putative coiled coil α-helix important for core assembly. Journal of Virology 75, 1–10.
Petrov, D.A., 2001. Evolution of genome size: new approaches to an old problem. Trends in Genetics 17, 23–28.
Poignard, P., Saphire, E.O., Parren, P.W.H.I., Burton, D.R., 2001. GP120: biologic aspects of structural features. Annual Review of Immunology 19, 253–274.
Pool, A.M., Jeffares, D.C., Penny, D., 1998. The path from the RNA world. Journal of Molecular Evolution 46, 1–17.
RCSB Protein Data Bank PDB (⟨http://www.pdb.org/pdb/home/home.do⟩).
Radó-Trilla, N., Alba, M.M., 2012. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. BMC Evolutionary Biology 12, 155, http://dx.doi.org/10.1186/1471-2148-12-155.
Shakin-Eshleman, S.H., Spitalnik, S.L., Kasturi, L., 1996. The amino acid at the X position of an Asn-X-Ser sequon is an important determinant of N-linked core-glycosylation efficiency. Journal of Biological Chemistry 271, 6363–6366.
Simmonds, P., Balfe, P., Ludlam, C.A., Bishop, J.O., Brown, A.J., 1990. Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. Journal of Virology 64, 5840–5850.
Simon, M., Hancock, J.M., 2009. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. Genome Biology 10, R59.
Sirois, S., Sing, T., Chou, K.C., 2005. HIV-1 gp120 V3 loop for the structure-based drug design. Current Protein and Peptide Science 6, 413–422.
The PyMOL Molecular Graphics System, Version 1.0, L.L.C. Schrödinger,. ⟨http://www.expasy.org.spdbv⟩.
Wei, X., Decker, J.M., Wang, S., Hul, H., Kappes, J.C., Wu, X., Salazar-Gonzales, J.F., Salazar, M.G., Kilby, J.M., Saag, M.S., Komarova, N.L., Nowak, M.A., Hahn, B.H., Kwong, P.D., Shawn, G.M., 2003. Antibody neutralization and escape by HIV-1. Nature 422, 307–397.
Wootton, J.C., 1994. Sequences with 'unusual' amino acid compositions. Current Opinion in Structural Biology 4, 413–421.
Wootton, J.C., Federhen, S., 1993. Statistics of local complexity in amino acid sequences and sequence databases. Computational Chemistry 17, 149–163.
Yamaguchi-Kabata, Y., Gojobori, T., 2000. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. Journal of Virology 74, 4335–4350.

C A P Í T U L O


José Ángel Córdova Villalobos
Samuel Ponce de León Rosales
José Luis Valdespino
Editores
25 AÑOS de SIDA en MÉXICO
LOGROS, DESACIERTOS Y RETOS
Segunda edición
VIH-1M
VIH-10
VIH-2
SALUD
SECRETARÍA DE SALUD
Birmex
Instituto Nacional de Salud Pública
SALUD
CENSIDA

José Ángel Córdova Villalobos
Samuel Ponce de León Rosales
José Luis Valdespino
EDITORES

25 AÑOS de SIDA en MÉXICO

LOGROS, DESACIERTOS Y RETOS

SALUD
SECRETARÍA DE SALUD
Instituto Nacional de Salud Pública
SALUD
CENSIDA

25 años de SIDA en México. Logros, desaciertos y retos

Primera edición, 2008
Segunda edición, 2009

Portada: Laura Esponda
Ilustración: preparada por Antonio Lazcano y Ana María Velasco

Av. Universidad 655,
Santa María Ahuacatitlán
62100 Cuernavaca, Morelos, México

Impreso y hecho en México
Printed and made in Mexico

ISBN

Esta obra contó con el apoyo parcial de un *grant* educacional irrestricto de Bristol-Myers Squibb de México.

Agradecimientos

Los editores de la segunda edición del libro **25 años de SIDA en México. Logros, desaciertos y retos** deseamos manifestar nuestro más profundo agradecimiento a los doctores Mauricio Hernández Ávila y Mario Henry Rodríguez, quienes apoyaron en todo momento la publicación de esta obra; a la doctora María Eugenia Jiménez y al licenciado Gilberto Méndez, quienes participaron de manera importante en el proceso de integración de los diferentes capítulos, y a la Subdirección de Comunicación Científica y Publicaciones del Instituto Nacional de Salud Pública, encabezada por el licenciado Carlos Oropeza, por la coordinación editorial de la obra.

Contenido

capítulo
7

Origen y evolución del VIH

Ana María Velasco
Antonio Lazcano Araujo

Más allá de la tragedia colectiva que significa la epidemia de VIH/SIDA, que afecta todos los rincones del planeta, el surgimiento de un nuevo patógeno que se ha expandido con una rapidez aterradora se puede analizar no sólo como un problema médico y social sino también bajo la óptica de la evolución, la teoría central de las ciencias de la vida. Ello no ha sido fácil: en primer lugar, el estudio de los virus ha estado sujeto a una serie de prejuicios antropocéntricos que tienden a enfatizar su carácter patógeno y que han limitado la comprensión de su verdadera naturaleza y distribución biológica. De hecho, el mismo término, "virus", significa veneno en latín. En segundo lugar, la simplicidad estructural de los virus, que va unida a la posibilidad de ser cristalizados, ha llevado a discusiones interminables sobre si pueden ser considerados o no como sistemas propiamente vivientes. Finalmente, hay que tener presente que no es sino hasta en los últimos diez años cuando la conjunción afortunada de técnicas de secuenciación cada vez menos costosas, unidas a sistemas de cómputo cada vez más rápidos y baratos, han permitido la acumulación y análisis de una enorme cantidad de secuencias virales, sobre todo del VIH.

Se puede afirmar que el VIH es una de las entidades biológicas mejor estudiadas –lo que no implica, sin embargo, que todas las preguntas sobre su origen y evolución hayan sido resueltas en forma satisfactoria–. El VIH es un retrovirus, es decir, pertenece a la llamada familia *Retroviridae*, que se caracteriza por codificar una polimerasa llamada reverso transcriptasa o transcriptasa inversa. El hecho de que un gran número de vertebrados sean infectados por este tipo de virus abre de inmediato la posibilidad de buscar en ellos el origen del VIH, que estamos seguros brincó a nuestra especie a partir de otros primates. Aunque hasta hace relativamente poco tiempo un grupo de investigadores sostuvo que el VIH se propagó a la población humana debido a la aplicación de vacunas elaboradas con tejidos de riñones de simio que estaban contaminados, esta idea ha sido paulatinamente abandonada. De hecho, los datos disponibles indican que el origen de la epidemia de VIH/SIDA resultó de múltiples eventos de zoonosis en donde han jugado un papel protagonista los llamados virus de inmunodeficiencia de simios (VIS) que infectan a diversos primates y que brincaron hacia nuestra especie. Esta zoonosis debió haber sido provocada por la cercana convivencia entre el hombre y estos animales, incluyendo el contacto con sangre y otros tejidos contaminados durante la cacería, el comercio y el consumo de primates contaminados.[1]

La comparación de secuencias tanto de proteínas como de nucleótidos de un número creciente de muestras del VIH ha confirmado que, en realidad, existen dos grandes tipos de virus de la inmunodeficiencia humana, conocidos

como VIH-1 y VIH-2. Como lo indica el análisis evolutivo de las muestras de VIH-2, este virus tiene un origen independiente del VIH-1: los árboles filogenéticos lo ubican, por una parte, lejos del VIH-1 y, por otra, cercano a los llamados virus de la inmunodeficiencia de simios (VIS) provenientes de los llamados monos verdes (o *sooty mangabeys*, en inglés), los VISsm. El análisis evolutivo de las secuencias también ha demostrado que existen tres grandes ramas del VIH-1, denominadas M, N, y O, que parecen haber surgido como resultado de tres eventos independientes de infecciones zoonóticas con VIS (virus de inmunodeficiencia de chimpancés, VIScpz).[1,2] El VIH-2 se divide, a su vez, en ocho grupos (denominados con letras de la A a la H).[3] Es importante subrayar, sin embargo, que las fronteras que separan estos grupos y subgrupos están lejos de ser infranqueables: además de los grupos mencionados, tanto en el caso del VIH-1 como del VIH-2 se han encontrado virus recombinantes que presentan genomas en donde se reconocen secuencias génicas características de dos o más grupos. La existencia de estos virus recombinantes, unida a la elevada variabilidad genética tanto del VIH-1 como del VIH-2, complica de manera extraordinaria la posibilidad de reconstruir sus historias evolutivas. Ello explica, en buena medida, las filogenias y clasificaciones, a veces discordantes, de las que distintos grupos han informado, y que además utilizan distintas metodologías de análisis evolutivo. La tendencia actual es la de analizar secuencias de genomas virales completos para realizar descripciones que permitan caracterizar en forma adecuada la gran diversidad de tipos, subtipos y recombinantes de VIH. A pesar de estos problemas, se puede afirmar sin duda alguna el origen animal de la epidemia VIH/SIDA, que resultó del brinco evolutivo a nuestra especie a partir de otros primates.

Características generales

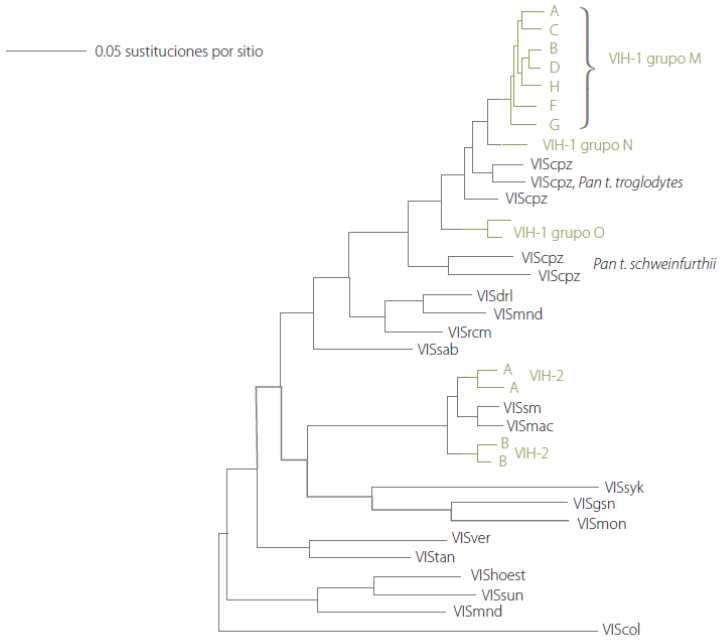
El VIH se ubica taxonómicamente en la subfamilia *Lentiviridae*, perteneciente a la Familia *Retroviridae*. Los miembros de esta familia se caracterizan por presentar un genoma compuesto por dos hebras sencillas de ARN^{+}, de aproximadamente 11kb contenidas dentro de un virión, por lo que se les considera pseudodiploides. Los retrovirus son característicos de los vertebrados, y su ciclo biológico depende de una polimerasa cuya capacidad para leer el genoma viral de ARN y sintetizar una hebra doble de ADN denominada provirus, la llevó a ser denominada reverso transcriptasa (RT) o transcriptasa inversa. El provirus, que se integra al genoma de la célula hospedera, está flanqueado por secuencias repetidas (o LTR, por sus iniciales en inglés) que son generadas por reverso transcripción.[4,5]

Aunque los retrovirus suelen infectar células somáticas, no todos son patógenos. El análisis de los genomas de diversos animales, incluyendo los primates y a los humanos mismos, ha mostrado cómo algunos retrovirus han infectado las células de las líneas germinales de sus hospederos. Ello se ha traducido en una colonización permanente del genoma del hospedero, lo que ha provocado que los descendientes del hospedero así infectado hereden en forma vertical los provirus. Los retrovirus que se integran al genoma de células de la línea germinal y se transmiten verticalmente son llamados endógenos, para diferenciarlos de aquellos que se transmiten de manera horizontal y que son denominados exógenos.[6]

En términos generales, se puede afirmar que los genomas retrovirales están constituídos por tres genes: gag, pol y env. El gen gag codifica para el conjunto de proteínas que forman la cápside, pol para la maquinaria catalítica del retrovirus, y env para las glicoproteínas de la envoltura, que están involucradas con la unión y penetración de la célula huésped y están asociadas a los demás componentes de la membrana del virus, que como es bien sabido tiene un origen celular y es adquirida por gemación. Cuando el genoma retroviral contiene únicamente a estos tres genes se considera de estructura sencilla, pero cuando el retrovirus presenta, además de los genes gag, pol y env, los llamados genes accesorios, entonces se le denomina complejo[6,7] (figura 1). Este es el caso del VIH, que presenta los genes accesorios vif, vpr, tat, rev y nef. Además, el VIH-1 posee el gen vpu, mientras que vpx es exclusivo del VIH-2.[8]

Como ya se mencionó, el VIH presenta una elevada variabilidad genética, la cual se debe a varios factores.[9] Esta variabilidad resulta, en primer término, de las propiedades de la polimerasa retroviral (RT), un ADN polimerasa que no posee propiedades de exonucleasa 3'-5', por lo que no tiene la capacidad de corregir los errores que ocurren durante la replicación. Ello se traduce en una elevadísima

FIGURA 1.
Filogenia de lentivirus de primates, relaciones evolutivas presentes entre el VIH-1 y el VIH-2 con diferentes grupos de VIS



Filogenia de lentivirus de primates donde se presentan las relaciones evolutivas presentes entre el VIH-1 y el VIH-2 con diferentes grupos de VIS, lo cual sugiere el origen independiente de estos dos grupos. De igual manera se observa de manera intercalada con varios VIS tanto a los grupos M, N, y O de VIH-1, como los grupos A y B de VIH-2. El árbol fue construido mediante el método de máxima verosimilitud utilizando para el análisis secuencias de nucleótidos de la polimerasa viral (pol), disponibles en bases de datos públicas. Las abreviaciones para los diferentes tipos de virus y sus hospederos son las siguientes: VIScol, colobus negro y blanco; VISdrl, drill; VISgsn, mono de nariz moteada; VISlhoest, mono L'Hoest; VISmac, macaco; VISmnd, mandril; VISmon, mono Campbell; VISrcm, mono testa roja; VISsab, mono Sabaeus; VISsun, mono cola dorada; VISsyk, mono Sykes; VIStan, mono tantalus. Modificado de Rambaut *et al.*, 2004

tasa de mutación *in vivo,* de aproximadamente $3.4x10^{-5}$ mutaciones por par de bases por ciclo de replicación.[10] De hecho, se ha calculado que la tasa de sustitución promedio es de alrededor de $10x10^{-3}$ sustituciones por sitio por año,[11] lo que implica de forma evidente que durante cada evento de replicación se produce un genoma de VIH genéticamente único. Es decir, en un individuo infectado existe una multitud de poblaciones genéticamente diversas pero relacionadas, que dan lugar a lo que se conoce como una cuasiespecie.[12] Por otra parte, cada célula infectada con VIH alberga entre 400 000 y 2 500 000 copias de ARN viral,[13] y diariamente se producen alrededor de 10^{10} viriones en el hospedero infectado, los cuales presentan una vida media en el plasma de aproximadamente 8 horas.[14] El rápido recambio viral, así como la elevada tasa de mutación, indican que en promedio se generan $3x10^{9}$ mutaciones cada día en la población viral de un paciente.[15]

Otro proceso que da lugar a la variabilidad genética en las poblaciones de VIH es la recombinación. Para que ésta tenga lugar, es necesario que dos virus divergentes

infecten a una misma célula, de suerte que los genomas de ambos tendrán la posibilidad de formar en su progenie partículas virales homodiploides y heterodiploides.[16] La infección *de novo* con virus heterodiploides puede resultar en la recombinación de las dos hebras que conforman su genoma, debido al cambio de molde por parte de la RT durante la síntesis de la hebra de ADN.[16] La posibilidad de llevar a cabo la recombinación puede estar influenciada por la procesividad de la polimerasa viral: se sabe que las polimerasas con procesividad limitada poseen una mayor oportunidad de brincar de una hebra de ácido nucleico a otra, lo que da lugar a progenies recombinantes.[12] Se ha calculado que en cada replicación viral ocurren de 7 a 30 eventos de recombinación.[17]

El origen del VIH

Poco tiempo después de haber descubierto la existencia de patógenos más pequeños que las bacterias, el propio Felix D'Herelle llegó a la conclusión que los virus representaban las formas de vida más antiguas.[18] Con ello se inició una tendencia, que persiste hasta nuestros días, que interpreta la simplicidad estructural de los virus como evidencia de un estadio primitivo. Esta idea subyace en algunas propuestas contemporáneas que pretenden explicar el origen de los retrovirus como resultado de la transición evolutiva de genomas primitivos basados en ARN hacia el ADN, lo que implica la aparición temprana de actividad de reverso transcriptasa.[19] Aunque el análisis de las estructuras cristalográficas indica que el llamado dominio *palm* de la RT es homólogo al del ADN polimerasa I de *Escherichia coli*,[20] y que esta región probablemente procede de una etapa ancestral anterior a la aparición de genomas celulares de ADN,[21,22] la distribución biológica de los retrovirus no permite, por el momento, ubicarlos en tiempos geológicos más antiguos que los que marcan el origen de los vertebrados.

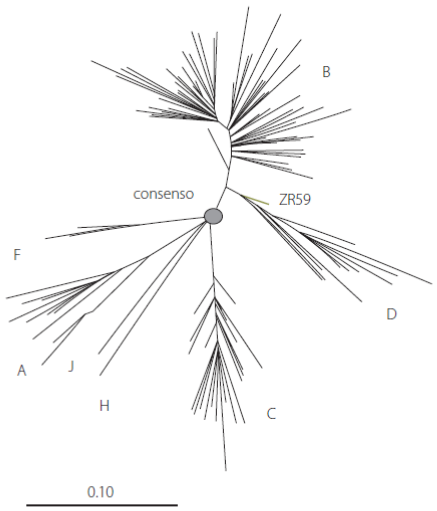
En realidad, la clave para acercarse al origen del VIH es saber con qué otros virus está relacionado. Se conocen más de 30 especies de primates africanos portadores de VIS, lo que sugiere que el VIH se originó en este grupo de animales. Más específicamente, los chimpancés son el único grupo de simios infectados por un virus relacionado cercanamente con el VIH-1.[23] Es importante subrayar que aunque a los retrovirus que infectan a todos estos animales se les conoce como virus de inmunodeficiencia de simios por presentar una estructura genómica similar a la del VIH, no hay evidencias que sugieran que provocan enfermedad alguna a sus hospederos naturales,[1] con la sola excepción de los macacos, en donde sí se han identificado como agentes causales de una inmunodeficiencia semejante al SIDA.[24]

La evidencia molecular de la cercanía evolutiva del VIH con otros retrovirus y la disponibilidad de secuencias genéticas muy pronto permitió la formulación de las primeras filogenias. Hasta la fecha, y debido al número siempre creciente de nuevas secuencias que se han venido acumulando en las bases de datos, la reconstrucción de la filogenia del VIH sigue siendo un punto central en su estudio, no solamente para conocer su origen, sino también la forma en que evoluciona, lo cual está íntimamente ligado a la capacidad del VIH para responder exitosamente a diferentes presiones de selección tales como la respuesta inmune de las personas infectadas y, en muchos casos, la evasión a la acción de los antirretrovirales.

El análisis filogenético ha confirmado que tanto el VIH-1 como el VIH-2 son retrovirus que brincaron a la especie humana a partir de otros primates, y a pesar de ello se ubican en ramas bastante separadas de los árboles evolutivos y están relacionados con diferentes grupos de VIS, lo cual implica que se originaron por procesos evolutivos independientes (figura 2). La variación de secuencias entre VIH-1 y VIH-2 puede ser hasta de 50%, mientras que las diferencias internas que separan a sus variantes son de aproximadamente 30 por ciento.[25] Como ya se mencionó, las poblaciones de VIH-1 se pueden dividir de manera natural en los grupos M, N y O, con una similitud de secuencias al seno de cada grupo que varía entre el 70 y el 90 por ciento. Las diferencias genéticas más grandes entre estos tres grupos se encuentran en el gen env (con diferencias de hasta 30% en la secuencia de nucleótidos), seguido por gag (con 20%) y, finalmente pol, con apenas 15 por ciento.[26]

Como se puede apreciar en la figura 1, cada uno de los tres subgrupos de VIH-1 está relacionado cercanamente con diferentes VIScpz de chimpancés. Aunque la cercanía filogenética que hay entre VIH-1 y VIScpz es evidente, durante un tiempo no se pudo afirmar con certeza que el chimpancé fuera el reservorio natural del VIH-1, ya que las

FIGURA 2.
Ubicación de la secuencia ZR59 en un árbol filogenético de los diferentes subgrupos del grupo M de HIV-1



Se aprecia que ZR59 se localiza muy cercanamente al ancestro de los subgrupos B y D. Para enraizar este árbol se utilizó una secuencia consenso que se obtuvo a partir de las secuencias consenso del gen env de cada subgrupo del grupo M de HIV-1. Modificado de Korber *et al.*, 2000.

tres muestras de VIScpz utilizadas para los análisis filogenéticos provenían de dos simios que aparentemente fueron infectados en su hábitat natural, mientras que se cree que un tercero adquirió la infección durante su cautiverio.[27] Sin embargo, el análisis filogenético de varias proteínas de VIScpz detectadas por anticuerpos y obtenidas a partir de muestras fecales de las cuatro subespecies de chimpancés, *Pan troglodytes troglodytes*, *P. t. schweinfurthii*, *P. t. verus* y *P. t. vellerosus* colectadas en el hábitat natural de cada una de las subespecies permitió confirmar una vez más que los chimpancés son el reservorio del VIH-1.[28]

Debido a que conocemos bien la evolución de los chimpancés, se puede afirmar que adquirieron el VIS después de la divergencia evolutiva de *P. t. verus* y *P. t. vellerosus*, pero muy probablemente antes de la divergencia de *Pan t.troglodytes*, *P. t. schweinfurthii*. Esta secuencia de eventos explicaría no sólo el que las dos primeras subespecies (*P. t. verus* y *P. t. vellerosus*) no estén infectadas con VIS, sino también la presencia de VIS en *Pan t. troglodytes*, *P. t. schweinfurthii* y la separación de VIScpz*Ptt* y VIScpz*Pts* observada en diferentes árboles evolutivos. Por otra parte, todas las muestras de VIH-1 utilizadas en este análisis filogenético se agruparon con VIScpzPtt, lo cual implica que *Pan t.troglodytes* es el reservorio natural del VIH-1. El hábitat natural de *Pan t.troglodytes* abarca Gabón, la República de Congo y el sur de Camerún. Esta última región es donde se originó el VIScpzPtt, que está directamente relacionado con el VIH-1, lo cual sugiere que es allí donde éste último virus pudo haber brincado a nuestra especie.

Como puede verse en la figura 2, el VIH-2 no se agrupa con el VIH-1, sino con ejemplares de VISsm de monos verdes (*Cercocebus torquatus atys*) y de VISmac de macacos. Las muestras de VIH-2 se encuentran distribuidas en ocho grupos. El análisis filogenético de las secuencias de los genes gag, pol y env de VIH-2, VISsm y VISmac disponibles en las bases de datos demostró que las muestras de VIH-2 se agrupan en ramas diferentes, en donde están mezcladas con los VISsm y VISmac.[29] Lo mismo ocurre con las dos muestras de VISmac obtenidas de *Macaca arctoides* y *M. nemestrina,* que se localizan en ramas diferentes. Ello indica que tanto en el caso de los humanos como en de los macacos las infecciones con VIH-2 y VISmac, respectivamente, resultaron de eventos independientes de transferencias virales entre especies a partir del contagio con VISsm de monos verdes. Es importante hacer notar que las muestras de VISmac analizadas fueron obtenidas de animales en cautiverio y que no hay evidencias de la presencia de VISmac en macacos salvajes, lo que indica que la infección ocurrió durante su confinamiento.[30]

¿Cómo brincó la infección de VIH a la especie humana?

Aunque las evidencias disponibles apoyan del todo la hipótesis del origen zoonótico del VIH, existen una serie de hipótesis alternativas, algunas de ellas plenas de fantasía. Algunos han sugerido que la aparición del virus está relacionada con la investigación con armamento biológico o

experimentos con nuevos patógenos.[91] Una de las hipótesis que alcanzó en el pasado una cierta notoriedad, proponía que el VIH-1 pasó a los humanos como resultado de la aplicación de vacunas orales contra el polio (OPV, por sus siglas en inglés) durante el programa de vacunación contra esta enfermedad que se llevó a cabo en África Central entre 1957 y 1960. Esta hipótesis se apoyaba en la creencia de que la cepa de OPV pudo haber crecido en células de riñón de chimpancé infectadas con VIScpz originarias de Stanleyville (ahora Kisangani, en la República Democrática del Congo). No hay nada que apoye esta idea. Por una parte, la búsqueda de VIS en las OPV restantes ha tenido resultados negativos. Por otra, tanto los cálculos de Korber,[92] que indican que el ancestro del grupo M del VIH-1 surgió décadas antes de que comenzaran los programas de vacunación, como los análisis filogenéticos de Worobey[93] de VIScpz recuperados de muestras fecales de simios recogidas en la República Democrática del Congo, en zonas que incluyen el hábitat natural de *Pan t.troglodytes*, indican que el origen de la epidemia de VIH/SIDA no provino de vacunas contaminadas.

El origen de la epidemia de VIH/SIDA se encuentra en otra parte. La posición de los distintos linajes de VIH-1 y VIH-2 en el árbol evolutivo de la figura 2 muestra como están intercalados con diferentes grupos de VIScpz, por una parte, y de VSsm y VISmac, respectivamente, lo que ha sido interpretado como evidencia del origen independiente de cada uno de los subgrupos de VIH-1 y 2.[1,94] Es una hipótesis audaz, que según Hahn[1] es resultado de una serie de zoonosis que desembocaron en el origen tanto del VIH-1 como del VIH-2. Las evidencias de la transmisión zoonótica de virus de primates al humano son:

a) la similitud en los genomas virales: la estructura de los genomas del VIH-1 y el VIH-2 es idéntica a la de los VIS con los que se agrupan. En el caso de VIH-2 y VISsm éstos presentan una proteína accesoria denominada Vpx que no ha sido encontrada en ningún otro lentivirus de primates;
b) las relaciones filogenéticas: en todos los estudios filogenéticos que se han realizado siempre se han encontrado asociados, por un lado, VIH-1 y VIScpz, y por otra parte, VIH-2 y VISsm (figura 2);
c) la prevalencia de VIH-1 y VIH-2 en hospederos naturales: existe un gran número de chimpancés y de monos verdes infectados en sus hábitats naturales. La frecuencia de infección en los monos verdes alcanza valores hasta del 22% en algunas manadas. En el caso de los chimpancés, puede ser de 29 a 35%;[28]
d) la distribución biogeográfica: en el caso del VIH-1 la evidencia más contundente la proporcionó un estudio filogenético en el cual una cepa de VIScpz obtenida de un chimpancé de Camerún quedó agrupada con una muestra de VIH-1 originaria de la misma zona. Es decir, la relación evolutiva que guardan entre sí es más cercana que la que tienen con cualquier otra cepa de VIS o de VIH. Por otra parte, las regiones que constituyen el hábitat natural de los monos verdes y aquéllas en donde se presenta el VIH-2 son las mismas, y abarcan regiones de Senegal, Costa de Marfil, Sierra Leona y Liberia en donde, por cierto, se han detectado las cepas más divergentes de VIH-2; y, finalmente,
e) las posibles rutas de transmisión: no es difícil imaginar la forma en la que los humanos pudieron haber sido infectados con VIS. A lo largo de la historia los pueblos africanos han estado cercanamente relacionados con una gran variedad de animales salvajes, entre ellos los simios. Algunos grupos los cazan y utilizan como alimento, y como ocurre en otras regiones del mundo, durante la preparación de la presa se está en contacto con la sangre y la carne del animal, que a veces puede ser consumida cruda. También se han utilizado por mucho tiempo como mascotas y se sabe que en estado silvestre son frecuentes los ataques de simios contra humanos, lo que puede provocar sangrado y mezcla de líquidos corporales de ambos, que pueden entrar en contacto con las heridas.

Evolución molecular del VIH-1

La clasificación de secuencias y la construcción de filogenias moleculares han confirmado, una y otra vez, que las poblaciones de VIH-1 se pueden dividir en tres grandes grupos, M, N y O,[95-97] cuyos clados se ubican entreverados con varios linajes de VIScpz (figura 1). Ello indica que el VIH-1 es resultado de tres eventos independientes de transmisión de VIS a humanos que pudieron haber ocurrido a principios del siglo XX.[1,90]

El análisis de muestras tomadas en las regiones en donde habita el *P. t. troglodytes*, el chimpancé identificado como el reservorio natural del VIH-1,[2,28] permitieron concluir que el grupo M se originó de un linaje de VIScpz*Ptt* aún presente entre poblaciones de *P. t. troglodytes* del sureste de Camerún, en donde probablemente se transmitió de manera local. Es probable que la infección haya migrado a lo largo de las márgenes del rio Sangha hasta el sur del río Congo, y de ahí a Kinshasa, en donde se diseminó rápidamente entre los habitantes de los cinturones de miseria de dicha ciudad. El grupo N, por su parte, que está limitado geográficamente a un número pequeño de pacientes en Camerún, está relacionado evolutivamente con un segundo linaje de VIScpz*Ptt*[28] del que se separó recientemente, y parece ser el resultado de un evento de recombinación entre un virus similar al VIH y otro de VIS.[2]

En cambio, el origen del grupo O del VIH-1 sigue siendo un problema abierto. Un estudio reciente obtuvo 591 muestras de primates de las zonas boscosas remotas del Camerún, de las cuales 378 pertenecían a *P. t. troglodytes* y *P. t. vellerosus,* y las restantes 213 a gorilas. Como era previsible, se encontró que los chimpances de la subespecie *troglodytes* estaban infectados con VIScpz*Ptt*, mientras que no se detectó ningún caso de infección en ejemplares de la subespecie *vellerosus*. Sorprendentemente, también se hallaron muestras en dos especies de gorilas infectados con VIS, a los que se denominó VISgor. Se obtuvieron fragmentos de los genes gag, pol, env y vif de todas las muestras infectadas. El análisis de las secuencias obtenidas demostró que los VIS presentes en los gorilas forman un conjunto estrechamente relacionado entre sí, que resultó ser mucho más cercano al grupo O de VIH-1 que a cualquier otro VIS. Así, se puede suponer que un VIS presente en chimpancés se dispersó hacia otras poblaciones de chimpancés y también a gorilas y humanos para dar origen, a partir de las comunidades de *P. t. troglodytes* del sur de Camerún, a cepas divergentes de VIScpz que al ser transmitidas a humanos dieron origen a los grupos M y N de VIH-1. Por otra parte, se pudieron haber transmitido virus del tipo del grupo O en forma independiente a gorilas y a humanos, o bien, primero a gorilas y de éstos a los humanos. Sin embargo, aún se desconoce la manera en que un VIS de chimpancé infectó a gorilas, ya que estos animales son herbívoros, y los encuentros entre gorilas y chimpancés parecen ser relativamente raros. Es posible, como sugiere Van Heuverswyn,[23] que en la cacería de gorilas, cuya carne se consume como alimento y que son utilizados en prácticas de medicina tradicional, pueda subyacer la zoonosis del grupo O de VIH-1.

El rumbo seguido por los distintos grupos (M,N,O) del VIH-1 después de la zoonosis ha sido completamente diferente. El grupo M es el que se adaptó con mayor éxito a su nuevo hospedero, se dispersó por todo el mundo y dio origen a múltiples subtipos: A-D, F-H, J y K (figura 2). En cambio, los grupos N y O muestran menor infectividad y parecen ser endémicos de Camerún y países vecinos en donde representan del 1 al 5% de las infecciones con VIH-1.[38] De hecho, el grupo N es extremadamente raro y a la fecha está conformado por los virus presentes en seis personas infectadas con esta variante. Por su parte, el grupo O está presente en un porcentaje que va del 2 al 5% de las muestras de pacientes seropositivos en Camerún. Los análisis filogenéticos no permiten, por el momento, dividir al grupo O en subgrupos.[39] Irónicamente, podemos reconocer al grupo O como el primero en ser descrito unos 12 años después del descubrimiento del VIH.[25] La primera infección provocada por este grupo de la que se tiene noticia se confirmó con el estudio de informes médicos y muestras de sangre y otros tejidos obtenidos de un marinero noruego, su esposa y su pequeña hija. El paciente parece haber viajado extensamente entre 1961 y 1965; visitó Nigeria y otros países de África occidental, así como Asia, Europa, el Caribe, Canadá y el Este de África. El paciente enfermó en 1966, fue seguido por su esposa en 1967 y su hija de dos años en 1969. Los tres murieron en 1976.[40]

Se cree que los nueve subtipos del grupo M se originaron en África Central.[41] Siempre aparecen como equidistantes en las filogenias moleculares, con una diferencia que va del 25 al 35% en el gen env, aunque al interior de cada subtipo se puede alcanzar hasta el 20%[34] (con excepción de los grupos B y D, que están mucho más cercanos entre sí pero se siguen considerando subgrupos en aras de la consistencia).[42] Las comparaciones de secuencias han permitido distinguir subconjuntos en A y F, a los que se designa como sub-subtipos A1 y A2, y F1 y F2, respectivamente. Debido a que para reconocer un subtipo como tal es indispensable demostrar que no ha habido recombinaciones, los subtipos I

y E, que originalmente fueron considerados junto con otros subtipos del VIH-1 M fueron anulados al demostrarse que eran producto de eventos de recombinación.[34]

La asociación de un porcentaje importante de virus recombinantes en nuevos casos de infecciones[25] vuelve casi indispensable una revisión detallada de lo que hasta ahora se han considerado como subtipos "puros"[43] para garantizar que no se trate de formas recombinantes. Por ejemplo, Abecasis *et al.*[44] demostraron que el subtipo G, que está reconocido como uno de los genomas parentales de CRF02_AG, es en realidad un recombinante compuesto por A/J y un genoma no identificado, en tanto que CRF02_AG resultó ser uno de los genomas parentales de este "subgrupo". Es importante subrayar que se han detectado eventos de recombinación en todos los niveles filogenéticos, lo mismo entre lentivirus de primates, que entre grupos, subtipos, e incluso intrasubtipos de VIH, que pueden afectar en forma importante no sólo las reconstrucciones filogenéticas, sino también otras inferencias sobre las características de la epidemia y la rapidez con la que varía.[45]

Nomenclatura y clasificación de las variedades del VIH-1: ¿un ejercicio taxonómico?

Los primeros intentos por clasificar y estudiar desde una perspectiva evolutiva al VIH se hicieron mediante secuencias parciales de gag y pol, pero la disponibilidad de genomas virales completamente secuenciados permitió la rápida identificación de subtipos y el descubrimiento de formas recombinantes, lo que hizo necesario definir los lineamientos de una clasificación más realista.[42] De acuerdo con este consenso, cualquier grupo nuevo que se pueda descubrir en el futuro deberá llamarse P, Q, R, etc.; si un nuevo subtipo es caracterizado, deberá nombrarse con la siguiente letra del alfabeto, en este caso N, y así sucesivamente; los sub-subtipos seguirán siendo aquellos linajes que no estén lo suficientemente alejados genéticamente como para considerarlos subtipos, mientras que los virus recombinantes que son cepas epidemiológicas se denominan Formas Recombinantes Circulantes (CRF, por sus siglas en inglés), se numeran en forma secuencial y se designan mediante las letras de los subtipos que los componen. Así, por ejemplo, CRF02_AG indica que se trata del segundo virus recombinante en ser definido como tal, y presenta componentes de los subtipos A y G. Cuando no sea posible determinar el origen de alguna parte del genoma, ésta se representará con una letra U, y cuando un genoma esté constituído por más de dos subtipos se le denominará "complejo", y después del número que le corresponde se agregarán las letras cpx, como ocurre en el caso CRF04_cpx. Para que un nuevo subtipo, sub-subtipo o recombinante sea asignado como tal, se debe comparar con tres genomas virales completamente secuenciados de individuos con infecciones no relacionadas epidemiológicamente entre sí,[9] o dos genomas completos y una o varias secuencias parciales que se agrupen con éstos. Las variantes desconocidas son designadas como U hasta que no se definan los criterios para su clasificación.

Las reglas establecidas para clasificar y nombrar las variantes del VIH-1 no deben ser vistas como un mero ejercicio de nomenclatura, sino como un instrumento valioso para facilitar la comprensión de la epidemia, por una parte y, por otra, para entender su evolución. De hecho, la extraordinaria diversidad de tipos, subtipos, sub-subtipos y formas recombinantes de VIH-1 nos obliga a reflexionar sobre el tiempo necesario para alcanzar semejante complejidad evolutiva, es decir, desde que surgió el VIH-1 y comenzó a expandirse entre los humanos. Ha habido varios intentos por dar respuesta a esta pregunta; por ejemplo, Gojobori *et al.*[46] han calculado que el último ancestro común del VIH-1 y VIH-2 existía hace 125 años, mientras que los trabajos de Smith *et al.*[47] lo ubican en el año 1951. Para Sharp y Li[48] la divergencia del ancestro de VIS, VIH-1 y VIH-2 se dio hace 150 años, en tanto que Mulder[49] propone que el último ancestro común de VIS infectó al ancestro de monos del Viejo Mundo hace 25 millones de años. Estas discrepancias son resultado no sólo de la diversidad de las muestras utilizadas, sino también de las distintas metodologías bioinformáticas empleadas. En este sentido, la ausencia de una muestra viral no contemporánea con la cual se pudieran comparar secuencias disponibles era una limitante extraordinaria. Este problema se resolvió cuando se analizaron 1 213 muestras de plasma sanguíneo que habían sido colectadas en África entre 1959 y 1982. Una de estas muestras, recolectada en 1959 y que correspondía a un paciente masculino de lo que ahora es Zaire resultó ser positiva. Se amplificaron cuatro

fragmentos del genoma viral de VIH-1 correspondientes al gen env, y el análisis filogenético de estas secuencias indicó que el virus, al que se denominó ZR59, era muy parecido al ancestro de los subgrupos B y D. Como se puede ver en la figura 2, ZR59 se ubica en la base de la rama del subgrupo D, cerca del subgrupo B. Ello indica que el último ancestro común de los subgrupos D y B debió haber existido pocos años antes de 1959. Más aún, la posición basal de ZR59 en la rama del subtipo D lo ubica cerca del nodo central del árbol, por lo que se puede deducir que el grupo M se originó en algún momento entre 1940 y 1950.[50]

Se han hecho varios intentos por determinar cuándo se originó el subtipo B.[28,32,51,52] Recientemente se utilizó una metodología parecida a la que permitió la identificación del ZR59, para saber dónde y cuándo surgió la epidemia causada por el subtipo B, para lo cual se analizaron muestras de cinco pacientes haitianos infectados con VIH-1, que migraron de Haití a Estados Unidos y que son reconocidos como parte de las primeras víctimas de SIDA. Las muestras fueron tomadas entre 1982 y 1983, y de ellas se obtuvieron secuencias parciales de env, que fueron comparadas con 117 secuencias de env provenientes de 19 países. Los resultados obtenidos muestran que las cepas más antiguas no africanas del subtipo B son las de origen haitiano, y presentan la mayor diversidad genética conocida en todo el mundo. Estos resultados se pueden explicar con la llegada a Haití de profesionistas infectados, que habían sido repatriados luego de trabajar en el Congo hasta el momento de su independencia. De acuerdo con este esquema, el virus pasó de Haití a los EUA, y de allí se dispersó por todo el mundo.[53]

La filogenia del VIH-2

Al igual que ocurre con el VIH-1, la filogenia del VIH-2 refleja su origen zoonótico, como lo muestra la dispersión de las secuencias correspondientes en medio de diferentes grupos de VIS en los árboles evolutivos (figura 2). El reservorio natural del VIH-2 son los monos verdes (*C. t. atys*), de donde provino el VISsm ancestral que evolucionó a la forma de VIH-2.[1] Actualmente se reconocen ocho grupos de VIH-2, que van de A-H, cada uno de los cuales se supone es resultado de eventos independientes de transferencia interespecífica.[54] Por otra parte, los datos que indican una elevada seroprevalencia de VIH-2 en Canchungo, una población de Guinea-Bissau, sugieren que este es el posible núcleo de origen de la epidemia, y que el brinco a los humanos ocurrió en la primera mitad del siglo XX.[29] Los niveles de infección parecen haberse mantenido estables, pero entre 1955 y 1970 ocurrió un crecimiento exponencial, que coincide con la guerra de independencia contra Portugal, durante la cual se dió un incremento notable de contagios por vía sexual y transfusiones durante este período bélico.[29]

Aunque los primeros casos de VIH-2 en Europa se dieron en Portugal y estaban relacionados con veteranos que sirvieron en el ejército de este país durante la guerra de independencia de Guinea-Bissau,[29] la distribución de VIH-2 sigue esencialmente restringida al África occidental, en donde se encuentran todos los grupos conocidos. Los dos tipos más comunes son el A y el B, cuyo último ancestro común pudo haber existido entre 1940 y 1945.[29] La mayoría de los genomas de VIH-2 que se han caracterizado hasta ahora pertenecen al grupo A y parecen estar distribuidos ampliamente en países de África Occidental, mientras que el grupo B parece estar restringido a Costa de Marfil y, en Europa, a Francia y Portugal. Los grupos C y D han sido identificados en Liberia, mientras que hay informes de E y F en Sierra Leona. Por último, la existencia del grupo G se determinó por el análisis del genoma completo obtenido de una cepa de un donador de sangre asintomático de Costa de Marfil, mientras que el grupo H fue caracterizado en un paciente masculino originario del mismo país.[54]

Conclusiones y perspectivas

A pesar de la falta de respuestas a muchas de las interrogantes que ha planteado el origen y la evolución del VIH, todos los datos disponibles indican que es el resultado de una serie de transferencias de virus que brincaron de otros primates a nuestra especie. Este no es un fenómeno raro: el análisis de genomas completamente secuenciados ha demostrado el papel que el transporte horizontal de genes o de grandes segmentos de ADN ha jugado en la evolución de todas las especies, desde los procariontes hasta los humanos mismos. Más allá de los factores estrictamente biológicos, es evidente que la transición de sociedades agrarias a un mundo globalizado ha modificado los patrones de transmisión de

enfermedades infecciosas, debido a procesos que van desde la producción industrial de alimentos, la mayor movilidad de individuos y grupos humanos, una urbanización desenfrenada y el desarrollo de cinturones de pobreza hasta el desarrollo de nuevas técnicas médicas como transfusiones y transplantes.[55] La rapidez con la que se ha expandido la epidemia de VIH/SIDA es resultado no sólo de las características biológicas del patógeno, sino también de una compleja red de factores que incluyen conductas individuales y los rasgos de la sociedad contemporánea.

¿Cuál es el futuro evolutivo del VIH? Una de las premisas fundamentales de los estudios evolutivos es la capacidad de explicar el pasado pero no la de predecir el futuro. Como han subrayado Heeney *et al.*,[56] la evidencia de un gran número de genomas de retrovirus endógenos que portamos en nuestros cromosomas indica que existen mecanismos de defensa y coexistencia entre retrovirus y humanos. No sabemos, sin embargo, si ese es el destino, a muy largo plazo, de la infección de VIH. Es también cierto, como lo muestran las filogenias moleculares, que los casos de transferencia interespecífica de los distintos tipos de VIH muestran la facilidad que tienen algunos sistemas virales para brincar las barreras de por sí tenues que separan a nuestra especie de otros primates. Como lo ha enfatizado recientemente Nathan Wolfe, de UCLA, la presencia de muchos otros VIS y la evidencia de cazadores africanos recientementre infectados con virus de otros primates como HTLV debería conducir a una mayor vigilancia epidemiológica, especialmente en los trópicos, donde la diversidad biológica (y, por lo tanto, de virus) es mayor.

Existen otros riesgos más inmediatos que se deberían atender. Más allá de las diferencias con las que algunos sectores sociales definen la lucha contra la epidemia de VIH/SIDA, la presencia de cepas resistentes a antirretrovirales (aún en pacientes que no han estado sometidos a terapias) habla, en términos evolutivos, de la respuesta de un sistema biológico que ante presiones de selección inéditas genera nuevas formas. Cada vez será más difícil enfrentar estas variantes. La conclusión es inevitable: si deseamos frenar la diversificación genética de la epidemia de VIH/SIDA, las medidas preventivas, entre las que destaca en forma notable el uso del condón y otras formas de sexo seguro, siguen siendo los mejores aliados para una sociedad que no está inerme del todo ante los mecanismos evolutivos de un patógeno.

Abstract

The discovery of HIV (Human Immunodeficiency Virus) as a causal agent for AIDS led to a limitless number of hypotheses about its origin and evolution, not all having the same scientific weight. Although the phylogenetic study of viruses is limited due to a series of factors that range from the lack of a fossil record to the polyphyletic origin of the genes that make up its genomes, the availability of a large quantity of sequences not only of HIV-1 and HIV-2 but also of the retroviruses that infect other primates has shown that the HIV/AIDS epidemic had its origins in interspecies contagion that led to the adaptation of the viruses known as simian immunodeficiency (SIV) to a new host, the human species. The available data indicate that the origin of the HIV/AIDS epidemic was the result of multiple zoonotic events where SIV that infect a variety of primates and jumped to our species have played a protagonist role. This zoonotics must have been caused by the close proximity between humans and these animals, including contact with blood and other contaminated tissue during hunting, selling, and consuming of contaminated primates. Actually, the key for discovering the origin of HIV is identifying the other viruses with which it is related. More than 30 known African primate species are SIV carriers, which suggests that HIV originated among this group of animals. More specifically, the chimpanzees are the only group of simians infected by a virus closely related with HIV-1. In spite of the rapid diversification of the different HIV strains, including the appearance of recombinant genomes, it is possible to reconstruct the structure and phylogeny of the large HIV-1 and HIV-2 groups and subgroups and, in most cases, the routes taken throughout its geographic expansion. In spite of the lack of answers to many of the

questions suggested by the origin and evolution of HIV, all of the available data indicate that it is the result of a series of transfers of the virus that jumped from other primates to our species. This phenomenon is not uncommon, the analysis of completely sequenced genomes has demonstrated the role that horizontal gene transfer or large DNA segments has played in the evolution of all of the species, from procarionts to humans. Today it can be said that HIV is one of the better-studied biological entities; that does not imply, however, that all of the questions about its origin and evolution have been satisfactorily resolved. Nevertheless, aside from the differences among how some social sectors define the fight against the HIV/AIDS epidemic, the presence of strains resistant to antiretrovirals suggest, in evolutive terms, that the response of a biological system to unknown selection pressures generates new forms, indicating that it will be more and more difficult to deal with these variants and that understanding its essential characteristics in detail will be necessary if we are to confront it successfully.

Referencias

1. Hahn BH, Shaw GM, De Cock KM, Sharp PM. AIDS as a zoonosis: scientific and public health implications. Science 2000; 287: 607-614.
2. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg M, Michael SF, *et al*. Origin of HIV-1 in chimpanzee *Pan troglodytes troglodytes*. Nature 1999; 397: 436-441.
3. Damond F, Worobey M, Campa P, Farfara I, Colin G, Mathernon S, *et al*. Identification of a highly divergent HIV type 2 and proposal for a change in HIV type 2 classification. AIDS Res Hum Retroviruses 2004; 20: 666-672.
4. Baltimore D. Viral RNA-dependent DNA polymerase. Nature 1970; 226:1209.
5. Temin HM, Mizutani S. RNA directed DNA polymerase in virions of Rous sarcoma virus. Nature 1970; 226:1211.
6. Gifford R, Tristem M. The evolution, distribution and diversity of endogenous retroviruses. Virus Genes 2003; 26:291.
7. Patience C. Our retroviral heritage. TIG 1997; 13:116.
8. Hirsch VM, Omsted RA, Murphey-Corb M, Purcell RH, Johnson PR. An african primate lentivirus (SIVsm) closely related to HIV-2. Nature 1989; 339: 389-392.
9. Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM. The challenge of HIV-1 subtype diversity. N Engl J Med 2008; 358: 1590-1602.
10. Mansky LM, Temin HM. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J Virol 1995; 69: 5087-5094.
11. Li W-H, Tanimura M, Sharp PM. Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol 1988; 5: 313-330.
12. Domingo E, Holland JJ. RNA virus mutations and fitness for survival. Annu Rev Microbiol 1997; 51:151-178.
13. Somasundaran M, Robinson HL. Unexpectedly high levels of HIV-1 RNA and protein synthesis in a cytocidal infection. Science 1988; 242:1554-1556.
14. Perelson AS, Neuman AU, Markowitz M, Leonard JM, Ho DD. HIV-1dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science 1996; 271:1582-1586.
15. Ramirez BC, Simon-Loriere E, Galetto R, Negroni M. Implications of recombination for HIV diversity. Virus Res 2008. doi: 10.1016/j.virusres.2008.01.007.
16. Hu W-E, Temin HM. Retroviral recombination and reverse transcription. Science 1990; 250:1227-1233.
17. Levy DN, Aldrovani GM, Kutsh O, Shaw GM. Dynamics of HIV-1 recombination in its natural target cells. Proc Natl Acad Sci USA 2004;101: 4204-4209.
18. Summers WC. Félix d'Herelle and the origins of molecular biology. Yale: University Press: 1999.
19. Lazcano A, Fox GE, Oró J. Life before DNA: the origin and early evolution of early Archean cells. En: RP Mortlock (ed.) The Evolution of Metabolic Function. Boca Raton, FL: CRC Press, 1992: 237-295.
20. Jeruzalmi D, Steitz TA. Structure of the T7 RNA polymerase complexed to the transcriptional inhibitor T7 lysozyme. Embo J 1998; 17: 4101-4113.
21. Delaye L, Vázquez H, Lazcano A. The cenancestor and its contemporary biological relics: the case of nucleic acid polymerases En: Chela-Flores J., Owen T., Raulin F. (eds.) First steps in the origin of life in the Universe: Proceedings of the Sixth Trieste Conference on Chemical Evolution. Dordrecht: Kluwer Academic Publisher, 2001: 223-230.
22. Becerra A, Delaye L, Islas S, Lazcano A. Very early stages of biological evolution related to the nature of the last common ancestor of the three major cell domains. Annu Rev Ecol Evol Sys 2007; 38: 361-379.

23. Van Heuverswyn F, Li Y, Neel C, Bailes E, Keele BF, Liu W, *et al.* Human immunodeficiency viruses: SIV infection in wild gorillas. Nature 2006; 444:164.
24. Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, Hahn BH. The origins of acquired immune deficiency syndrome viruses: where and when. Phil Trans R Soc Lond B 2001; 356: 867-876.
25. Tebit DM, Nankya I, Arts EJ, Gao Y. HIV diversity, recombination and disease progression: how does fitness "fit" into the puzzle? AIDS Rev 2007; 9:75-87.
26. Gao F, Robertson DL, Carruthers CD, Morrison SG, Jian B, Chen Y, *et al.* A comprehensive panel of near-full-length clones and reference sequences for non-sub-type B isolates of human immunodeficiency virus type 1. J Virol 1998; 72: 5680-5698.
27. Corbet S, Müller-Trutwin MC, Versmisse P, Delarue S, Ayouba A, Lewis J, *et al. env* sequences of simian immunodeficiency viruses from chimpanzees in Cameroon are strongly related to those of human immunodeficiency virus group N from the same geographic area. J Virol 2000; 74: 529-534.
28. Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, *et al.* Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science 2006; 313:523-526.
29. Lemey P, Pybus OG, Wang B, Saksena NK, Salemi M, Vandamme A-M. Tracing the origin and history of the HIV-2 epidemic. Proc Natl Acad Sci USA 2003; 100: 6588-6592.
30. Sharp PM, Bailes E, Gao F, Beer BE, Hirsch VM, Hahn BH. Origins and evolution of AIDS viruses: estimating the time-scale. Biochem Soc Trans 2000; 28 part 2: 275-282.
31. Hutchinson JF. The biology and evolution of HIV. Annu Rev Anthropol 2001; 30:85-108.
32. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, *et al.* Timing the ancestor of the HIV-1 pandemic strains. Science 2000; 288: 1789-1796.
33. Worobey M, Santiago ML, Keele BF, Ndjango J-BN, Joy JB, Labama BL, *et al.* Contaminated polio vaccine theory refuted. Nature 2004; 428:820.
34. Peeters M, Sharp MP. Genetic diversity of HIV-1: the moving target. AIDS 2000; 14 Suppl. 3: S31-S44.
35. Charneau P, Borman AM, Quillent C, Guétard D, Chamaret S, Cohen J, *et al.* Isolation and envelope sequence of a highly divergent HIV-1 isolate: definition of a new HIV-1 group. Virology 1994; 205: 247-253.
36. DeLeys R, Vanderborght B, Haesevelde MV, Heyndrickx L, van Geel A, Wauters C, *et al.* Isolation and partial characterization of an unusual human immunodeficiny retrovirus from two persons from west-central african origin. J Virol 1990; 64: 1207-1216.
37. Simon F, Mauclère P, Roques P, Loussert-Ajaka I, Müller-Trutwin MC, Saragosti S, *et al.* Identification of a new human immunodeficiency virus type 1 distinct from group M and O. Nature Med 1998; 4: 1032-1037.
38. Buonaguro L, Tornesello ML, Buonaguro FM. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenic and therapeutic implications. J Virol 2007; 81:10209-10219.
39. Apetrei C, Marx PA, Smith SM. The evolution of HIV and its consequences. Infec Dis Clin N Am 2004; 18: 369-394.
40. Jonassen TO, Stene-Johansen K, Berg ES, Hungnes O, Lindboe CF, Froland SS, *et al.* Sequence analysis of HIV-1 group O from norwwegian patients in the 1960's. Virology 1997; 231: 43-47.
41. Thomson MM, Pérez-Álvarez L, Nájera R. Molecular epidemiology of HIV-1 genetic forms and its significance for vaccine development and therapy. Lancet Infect Dis 2002; 2: 461-471.
42. Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, *et al.* HIV nomenclature proposal. Science 2000; 288: 55-56.
43. McCutchan FE. Global epidemiology of HIV. J Med Virol 2006; 78: S7-S12.
44. Abecasis AB, Lemey P, Vidal N, de Oliveira T, Peeters M, Camacho R, *et al.* Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. J Virol 2007; 81: 8543-8551.
45. Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. Nature Rev Gen 2004; 5:52- 61.
46. Gojobori T, Moriyama EN, Ina Y, Ikeo K, Miura T, Tsijimoto H, *et al.* Evolutionary origin of human and simian immunodeficiency viruses. Proc Natl Acad Sci USA 1990; 87: 4108-4111.
47. Smith TF, Srinivasan A, Schochetman G, Marcus M, Myers G. Nature 1988; 333: 573-575.
48. Sharp PM, Li W-H. Undestanding the origins of AIDS viruses. Nature 1988; 336: 315.
49. Mulder C. Human AIDS virus not from monkeys. Nature 1988; 333: 396.
50. Zhu T, Korber BT, Nahmias AJ, Hooper E, Sharp PM, Ho DD. An african HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature 1998; 391: 594-597.
51. Salemi M, Strimmer K, Hall WW, Duffy M, Delaporte E, Mboup S, *et al.* Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes ursing a new method to uncover clock-like molecular. FASEB J 2001; 15:276-278.
52. Robbins KE, Lemey P, Pybus OG, Jaffe HW, Youngpairoj AS, Brown TM, *et al.* U.S. human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. J Virol 2003; 77: 6359-6366.
53. Gilbert MTP, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. The emergence of HIV/AIDS in the Americas and beyond. Proc Natl Acad Sci USA 2007;104:18566-18570.
54. Kandathil AJ, Ramalingam S, Kannangai R, Divid S, Sridharan G, Molecular epidemiology of HIV. Indian J Med Res 2005; 121: 333-344.
55. Barnes E. Diseases and Human Evolution. Santa Fe: University of New Mexico Press, 2005.
56. Heeney JL, Dalgleish AG, Weiss RA. Origins of HIV and the evolution of resistance to AIDS. Science 2006; 313: 462-466.

DISCUSIÓN Y CONCLUSIONES

Los virus presentan un elevado nivel de traslapamiento en sus genes, lo cual minimiza su genoma a un grado máximo (Ilie y Popescu, 2006). Debido a esto, se podría suponer que cualquier proceso con el cual el genoma viral aumente o disminuya de tamaño tendría una fuerte presión de selección en contra. Por esto, se podría suponer que los virus no presentan LCRs, las cuales son capaces de aumentar o disminuir el tamaño del genoma (Hancock, 2002). Sin embargo, desde el primero reporte de LCRs en virus (Heller *et al*., 1982), y hasta la fecha continúan dándose a conocer mas virus que presentan SS tanto en su genoma como en sus proteínas.

El primer virus en el que se reportó la presencia de LCRs en el fue el de Epstein-Barr (Heller *et al*., 1982). Posteriormente, se encontró que las proteína de antígeno nuclear 1 (EBNA1) y 2 (EBNA2) codificadas respectivamente por los genes *ebna1* y *ebna2* de este virus, presentan una secuencia simple de 240 aminoácidos constituida por repeticiones de glicina-alanina (Gly-Ala) en el caso de la primera, mientras que en la segunda tiene una región de 42 Pro (Karlin et al., 1990). Debido al alto número de copias de Gly-Ala y Pro, así como al bajo nivel del diferencias en estas regiones, se ha sugerido que esta son de muy reciente adquisición. Esta sugerencia se opone a la idea de que en los virus no se puede dar este tipo de cambio en el genoma, ya que si tomamos en cuenta que la proteína EBNA1 reportada en el Genbank (ID:YP_401677), cuenta con un tamaño de 641 aminoácidos (1923 pb), y de este número corresponden a una SS 240 (720pb), entonces estaríamos hablando de que un tercio de la proteína está conformada por una LCR. Sin embargo, al comparar estos datos con los 171823 pb que constituyen el genoma del virus Epstein-Barr que codifica para esta proteína (GenBank ID:NC_007605), se observa que el tamaño total de la proteína corresponde al 1.1% del genoma, mientras que el tamaño de la SS representa apenas el 0.4%.

El tamaño de la LCR de EBNA1 contrasta enormemente con reportes previos (Wootton, 1994; http://www.bioafrica.net; Perera *et al*., 2001; Cristilllo *et al*., 2001; Hancock *et al*., 1995), así como los resultados presentados en esta investigación. La LCR mas grande que hemos encontramos es de 27 aminoácidos, y la mas pequeña es de nueve. Nuestros resultados también permiten comprender la manera en la que las SS localizadas en la proteína gp120

del VIH-1 representan una fuente de variabilidad, confiriéndole al virus la capacidad

de evadir al sistema inmune del hospedero gracias al elevado número de sitios de glicosilación que se presentan en las LCRs.

En base a los anterior, podemos suponer que aunque es probable que el efecto de las LCRs sobre el tamaño de los genomas virales no es relevante, la presencia de estas pequeñas SS a nivel de proteínas sí lo es, ya que representan una nueva fuente de variabilidad localizada en regiones clave de la proteína en donde son de gran importancia como lo demuestran nuestros resultados al encontrar LCRs únicamente en las regiones hipervariables de la proteína gp120. Nuestros resultados evidencian que este tipo de secuencias contribuyen a la diversidad genética y a la evolución en las poblaciones virales, y en este trabajo particularmente en el VIH-1.

R E F E R E N C I A S

1. Chen M, Tan Z, Jiang J, Li M, Chen H, Shen G, Yu R. 2009. Similar distribution of simple sequence repeats in diverse complete Human immunodeficiency virus type 1 genomes. *FEBS Lett* **583**: 2959-2963.

2. Cristillo AD, Mortimer JR, Barrette IH, Lillicrap TP, Forsdyke DR. 2001. Double-stranded RNA as a not-self alarm signal: to evade, most viruses purine-load their RNAs, but some (HTLV-1, Epstei-Barr) pyrimidine-load. *J Theor Biol* **208**: 475-491.

3. Cullen BR. 1998. Retroviruses as a model system for the study of nuclear RNA export pathways. *Virology* **249:** 203-210.

4. Djian P, Hancock JM, Chana H. 1996. Codon repeats in genes associates with human diseases:fewer repeats in the genes of nonhuman primates and nucleotide substitutions concentrated at the sites of reiteration. *Proc Natl Acad Sci USA* **93**: 417-421.

5. Freed EO, Martin MA. 1995. The role of human immunodeficiency virus type 1 envelope glycoproteins in virus infection. *J Biol Chem* **270**: 23883-23886.

6. Friedler A, Friedler D, Luedtker NW, Tor Y, Loyter A, Gilon C. 2000. Development of a functional backbone cyclic mimetic of the HIV-1 Tat arginine-rich motif. *J Biol Chem* **275**: 23783-23789.

7. Gerber HP, Seipel K, Georgiev O, Höfferer M, Hug M, Rusconi S, Schaffner W. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263**: 808-810.

8. Hallet B. 2001. Playing Dr Jekyll and Mr Hyde: combined mechanisms of phase variation in bacteria. *Curr Opin Microbiol* **4**: 570-581.

9. Hancock JM, Chaleeprom W, Chaleeprom W, Dale J, Gibbs A. 1995. Replication slippage in the evolution of potyviruses**.** *J Gen Virol* **76**:3229-3232.

10. Hancock JM. 2002. Genome size and the accumulation of simple sequence repeats: implications of new data from gnome sequencing projects. *Genetica* **115**: 93-103.

11. Heller M, van Santen V, Kieff E. 1982. Simple repeat sequence in Epstein-Barr virus DNA is transcribed in latent and productive infections. *J Virol* **44:** 311-320.

12. Huntley M, Golding GB. 2000. Evolution of simple sequences in proteins. *J Mol Evol* **51**: 131-140.

13. Ilie L, Popescu C. 2006. The shortest common superstring problem and viral genome compression. *Fundamenta Informatica* **73**: 153-164

14. Kay BK, Williamson MP, Sudol M. 2000. The importance of being proline: the interaction of prorine-rich motifs in signaling proteins with their cognate domains. *FASEB J* **14**: 231-241.
15. Kunin V, Ouzounis CA. 2003. The balance of driving forces during genome evolution in prokaryotes. *Genome Res* **13**: 1589-1594.

16. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. 1998. A consensus of protein repeats. *J Mol Biol* **293**: 151-160.

17. Moxon ER. 1999. Whole-genome analysis of pathogens**.** In *Evolution in Health & Disease*. Edited by Stearns SC. New York: Oxford University Press, 191-204.

18. Ogata H, Claverie JM. 2007. Unique genes in giant viruses: regularsubstitution pattern and anomalously short size. *Genome Res* **17**:1353-1361.

19. Perera R, Owen KE, Tellinghuisen TL, Gorbalenya AE, Kuhn RJ. 2001. Alphavirus nucleocapsid protein contains a putative coiled coil α-helix important for core assembly. *J Virol*, **75**: 1-10.

20. Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. *Trends in Genetics* **17**: 23-28.

21. Pollard VW, Malim MH. 1998. The HIV-1 Rev protein. *Annu Rev Microbiol* 1998, **52:** 491-532.

22. Sim KL, Creamer TP. 2002. Abundance and distributions of eukaryote protein simple sequences. *Mol Cell Prot* **1**:983-995.

23. Taitz D, Trick M, Dover GA. 1986. Criptic simlicity in DNA is a major source of genetic variation. *Nature* **322**: 652-656.

24. Wootton JC. 1994. Sequences with 'unusual' amino acid compositions. *Curr Opinion Struct Biol*, **4**: 413-421.