UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

Instituto de Investigaciones Biomédicas

“Evolución del virus de la inmunodeficiencia humana tipo 1 (HIV-1) a partir de una perspectiva de la teoría de las Cuasiespecies”

T E S I S

Que para obtener el grado académico de

MAESTRO EN CIENCIAS BIOLÓGICAS
(en Biología Experimental)

Presenta

BIOL. ALEJANDRO NABOR LOZADA CHÁVEZ

TUTOR DE TESIS: Dr. Marco Antonio José Valenzuela

COMITÉ TUTOR: Dr. Arturo Carlos II Becerra Bracho
Dr. Luis Padilla Noriega

MÉXICO, D.F. **MAYO, 2012**

UNAM
POSGRADO
Ciencias Biológicas

COORDINACIÓN

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 23 de abril de 2012, se aprobó el siguiente jurado para el examen de grado de **MAESTRO EN CIENCIAS BIOLÓGICAS (BIOLOGÍA EXPERIMENTAL)** del alumno **LOZADA CHAVEZ ALEJANDRO NABOR** con número de cuenta 509021225 con la tesis titulada **"EVOLUCIÓN DEL VIRUS DE LA INMUNODEFICIENCIA HUMANA TIPO 1 (HIV-1) A PARTIR DE UNA PERSPECTIVA DE LA TEORIA DE LAS CUASIESPECIES"**, realizada bajo la dirección del **DR. MARCO ANTONIO JOSÉ VALENZUELA:**

Presidente:	DR. ARTURO CARLOS II BECERRA BRACHO
Vocal:	DR. PABLO VINUESA FLEISCHMANN
Secretario:	DRA. BLANCA HAYDÉ RUIZ ORDAZ
Suplente:	DR. VÍCTOR MANUEL VALDÉS LÓPEZ
Suplente:	DR. LUIS PADILLA NORIEGA

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 ESPIRITU"
Cd. Universitaria, D.F., a 30 de Abril de 2012.

DRA. MARÍA DEL CORO ARIZMENDI ARRIAGA
COORDINADORA DEL PROGRAMA

Edif. de Posgrado P. B. (Costado Sur de la Torre II de Humanidades) Ciudad Universitaria C.P. 04510 México, D.F.
Tel. 5623-0173 Fax: 5623-0172 http://pcbiol.posgrado.unam.mx

“*AGRADECIMIENTOS OFICIALES*”

Mi total agradecimiento a:

El Posgrado en Ciencias Biológicas de la UNAM;

al Consejo Nacional de Ciencia y Tecnología (CONACYT)

por el apoyo económico (CVU: 293949. Becario: 225539) otorgado;

y a los miembros del *Comité tutor*:

Dr. Marco V. José,

Dr. Luis Padilla,

Dr. Arturo Becerra Bracho

“*AGRADECIMIENTOS Y RECONOCIMIENTOS*”

Mi especial agradecimiento y entero reconocimiento a

la M. en C. Irma Lozada Chávez

del Centro de interdisciplinario para la Bioinformática y Departamento de Ciencias Computacionales de Leipzig, Alemania, ***por fungir como tutora externa en la realización de este proyecto. El presente proyecto no hubiera logrado desarrollarse ni culminarse durante estos tres años sin su asesoría.***

Un amplio agradecimiento a los miembros del jurado de mi examen de obtención del grado:

Dr. Arturo Becerra Bracho,

Dr. Luis Padilla,

Dr. Pablo Vinuesa Fleischmann,

Dra. Blanca Haydé Ruiz Ordaz,

Dr. Víctor Valdés y

el Dr. Luis Delaye Arredondo,

por sus valiosas observaciones y comentarios sobre el proyecto y este escrito.

"…más agradecimientos"

>>>>>>>>>>>>>>>>>>>> °°° <<<<<<<<<<<<<<<<<<<<

Persistencia, honestidad, análisis, cansancio, sacrificio, coherencia y humildad son palabras, de muchas otras, que hemos ido aprendiendo a sentir y comprender a lo largo de nuestras vidas, y que en ocasiones difíciles las saboreamos hasta valorarlas en toda su esencia. El *agradecimiento*, una de esas palabras que empleamos y valoramos comúnmente, es algo que creemos se queda extremadamente pequeño cuando se sabe que se le debe tanto a una persona. Simplemente "gracias" no basta. Es así como en estas pequeñas líneas, estrechas y limitadas, agradezco a la única persona que siempre ha estado presente en mi vida, como hermana y como mentora, tanto en la vida académica como en la personal. Haciéndome siempre una mejor persona, a pesar de mi absurda resistencia, le agradeceré eternamente a **Irma** que cada historia de mi vida siempre mejorara por su causa, por su consejo, plática o conversación. Gracias, querida **Irma**, por sentir innumerables palabras conmigo, de las cuales muchas existen y, seguramente, otras tantas que aún no… Gracias por estar siempre en primera fila en mi vida. Gracias por hacerme comprender, a pesar de los consejos vacíos y evidencia ciega de otros, que lo *imposible* es *posible* cuando se trama bien una idea.

No habrá palabras para agradecerte todo tu apoyo académico que, desde mis inicios en la ciencia y en la divulgación científica, he tenido por parte tuya. Agradezco tu dedicación, dureza y pasión que reflejas en cada enseñanza, así como en cada evento de tu vida. ¡¡Todo digno de mencionarse!!

>>>>>>>>>>>>>>>>>>>> °°° <<<<<<<<<<<<<<<<<<<<

Gracias a **Marco José** por permitir integrarme a su grupo de trabajo, por todo el apoyo administrativo en el curso de esta maestría, por el espacio de trabajo y equipo de computo proporcionado, por su consejo y asesoría para este proyecto y, por permitirnos, a nosotros sus estudiantes, la libertad necesaria para crear y desarrollar nuestras propias ideas en nuestros proyectos. Gracias a **Juan Román Bobadilla** y a todos los estudiantes del laboratorio de Biología Teórica por sus comentarios y compañía: Isui, Alejandro, Víctor, Homero, Elisa y Miryam (galletita). Además, agradezco los útiles consejos de Tzipe y Gabriel.

Agradezco las continuas enseñanzas de mis padres, **Ana María** y **Nabor Félix**, para mostrarme que la vida es difícil en todo momento, para hacerme entender que 'la necesidad' se debe enfrentar sin vacilar y, principalmente, para revelarme que 'en las circunstancias más difíciles' sólo se cuenta con la verdadera familia. En verdad, gracias por esos momentos.

Quiero agradecer profundamente a mis amigos, los cuales son muy importantes en mi vida, ya que sin ellos, la vida tendría un sabor diferente y sería menos intensa. Gracias Isabel, Jorge, Fernanda, Alejandro C., Mariana, Stanley, Diana, Oscar y Carlos, Andrea, Alejandro O., Jimena, Josué, Irais, Mayitza, Cintia, Norma, Yolanda, Rafael, Laura, Cristina, Leopoldo, Beatriz y Mario. Y a todos los demás que ya no están conmigo, una lista corta que siempre la llevo en mí, mi total agradecimiento por sus enseñanzas para sobrevivir en esta vida y por recordarme que la amabilidad y la humildad es el lenguaje que los sordos pueden oír y los ciegos pueden ver.

Deseo agradecer especialmente a **Ivan Picaso**, a su padre y demás familia, el haberme brindado su amistad y la gran oportunidad de haber trabajado en su empresa durante varios años mientras desarrollaba mis estudios de licenciatura y de posgrado. Gracias por esos gratos y difíciles años.

Gracias a mi GRAN Universidad, la UNAM, por todo el conocimiento que hay ahí guardado y que es libre para todos. Gracias por enriquecer mi vida académica y personal, por todas las experiencias que tuve y la gente que conocí ahí y me brindó su apoyo incondicional: Laura, Azucena, Polo, Sergio, Maricela, Juanita, Pedro, Rafael, Teresa, Enrique, Israel y otros más.

Este proyecto fue realizado en el Laboratorio de Biología Teórica del Instituto de investigaciones Biomédicas de la UNAM, en colaboración con el Centro de interdisciplinario para la Bioinformática de la Universidad de Leipzig (Alemania), bajo la tutoría del Dr. Marco A. José Valenzuela y asesoría de la M. en C. Irma Lozada Chávez.

Asesorado institucionalmente (Comité Tutor) por:

Dr. Marco A. José Valenzuela	Instituto de Investigaciones Biomédicas, UNAM.
Dr. Arturo C. II Becerra Bracho	Facultad de Ciencias, UNAM.
Dr. Luis Padilla Noriega	Instituto de Investigaciones Biomédicas, UNAM.

Asesorado externamente a lo largo de todo el proyecto de esta tesis por:

M. en C. Irma Lozada Chávez	Centro Interdisciplinario para la Bioinformática, Leipzig, Alemania.

Esta tesis está dedicada a mi amada familia:

Irma Lozada Chávez

Dedicada a mis padres:

Ana María Chávez Morales

Nabor Félix Lozada Guerrero

Y ofrecida a mi universidad, la UNAM,

y a todo estudiante que busque el conocimiento dentro de esta, nuestra Ciudad Universitaria.

CONTENIDO

RESULTADOS Y DISCUSION

CONCLUSIONES

REFERENCIAS

MATERIAL SUPLEMENTARIO

RESUMEN

Un modelo que ha sido ampliamente usado y discutido para describir la evolución de los virus con genomas de RNA es la *Teoría de las Cuasiespecies*, propuesto por Manfred Eigen y Peter Schuster en 1977. En este modelo, se representa a los virus como nubes de genotipos replicantes que mantienen una distribución constante de mutantes con los fitness más altos a través de un equilibrio entre altas tasas de mutación y fuertes procesos de selección purificadora. Las propiedades más distintivas de las cuasiespecies son el umbral de error, la catástrofe de error y la extinción de error. El *umbral de error* es la tasa de mutación por debajo de la cual las poblaciones están equilibradas en un clásico balance de mutación-selección, en el cual, el genotipo favorecido puede mantener la replicación de la información a pesar de la alta tasa de mutación. El *error de catástrofe*, por otro lado, es la fase de transición ocurrida al traspasar el umbral de error, en la cual se pierde el genotipo favorecido, por ende también el comportamiento de cuasiespecies y posiblemente también la identidad viral. Finalmente, la *catástrofe de extinción* representa la perdida completa de la población viral a través de mutaciones letales.

Para comprobar si una población de virus con genomas de RNA puede evolucionar como *cuasiespecies*, se creó un programa computacional, con la estructura de un algoritmo genético, que incorpora realismo adicional el modelo original de cuasiespecies propuesto por Eigen-Schuster. Dada la complejidad de la caracterización de los parámetros biológicos de tal algoritmo, decidimos enfocar nuestras simulaciones en el Virus de la Inmunodeficiencia Adquirida serotipo 1 (HIV-1). Nuestro algoritmo incorpora parámetros biológicos del virus HIV-1 estimados con base en métodos experimentales (obtenidos de la literatura) y bioinformáticos para definir: la tasa de mutación, la frecuencia del tipo de mutaciones, el tiempo generacional, el tamaño efectivo de la población (*Ne*), así como el peso selectivo y la influencia reproductiva de las mutaciones en el genoma viral. Los resultados de nuestras simulaciones muestran la presencia un comportamiento de *cuasiespecies* en la dinámica evolutiva del virus HIV-1, logrando observar un *umbral de error* en la que los genotipos, en promedio, se mantienen con fitness altos cuando se encuentran sujetas a 1 y 2 mutaciones por genoma por ciclo de replicación. La presencia de una *catástrofe de error*, pero *no de extinción*, es posible observarse al incrementar la tasa de mutación a 3 y 4 mutaciones por genoma por ciclo de replicación, lo que conduce a un decaimiento drástico del fitness viral y a la pérdida del comportamiento de *cuasiespecies*. Los genotipos restantes en la dinámica evolutiva se mantienen replicando con un menor fitness, pero con gran robustez mutacional, debido a la persistencia de muchas mutaciones. El soporte estadístico de estos resultados, sin embargo, requieren de un mayor número de simulaciones (con más potencia en cómputo) para incluir generaciones >2,000 y valores de *Ne* más cercanos a los reales.

Nuestro *mapa del fitness mutacional* del genoma del virus HIV-1 (por genes, dominios funcionales y codones), nos permite observar que aunque la variación y la conservación nucleotídica se observan a lo largo del genoma del virus HIV-1, ambas también muestran una tendencia a estar limitadas o concentradas en regiones genómicas y estructurales específicas, y que predomina en particular el número de tripletes nucleotídicos seleccionados negativamente respecto a los positivos y neutros. De esta forma, la primera parte del genoma, conformada por los genes *5´LTR, gag, gag-pol, pol, vif* y *vpr*, se encuentra potencialmente sujeta a la influencia dominante de la *selección negativa*; mientras que la segunda mitad, conformada por los genes *tat, rev, vpu, env, nef* y *3'LTR,* se encuentra potencialmente sujeta a la acción principal de la *selección positiva.* Discutimos varios aspectos de la estructura genómica, evolutiva y funcional al nivel de las estructuras de RNA y las proteínas del virus HIV-1 que podrían soportar estos resultados. Todos estos resultados en conjunto sugerirían que la *presencia de una fuerte selección purificadora en genomas pequeños que se encuentran sujetos particularmente a altas tasas de mutación*, podría conducir al comportamiento de *cuasiespecies* de forma casi determinista, tal y como lo sugirieron Eigen-Schuster en su modelo original. Esperamos que este trabajo contribuya a tener una mejor perspectiva sobre la dinámica evolutiva de los virus de RNA.

Palabras clave: cuasiespecies, virus de RNA, umbral de error, error de catástrofe, extinción viral, robustez mutacional, epístasis, selección natural, deriva genética, balance mutación-selección, adecuación mutacional, evolución viral.

ABSTRACT

The quasispecies theory has been a model widely used and discussed to describe the evolution of viruses with RNA genomes, it was proposed by Manfred Eigen and Peter Schuster in 1977. In this model, the quasispecies are depicted as a cloud of viruses replicating genotypes that maintain a constant distribution of mutants with the highest fitness through a balance between high mutation rates and strong purifying selection. The most distinctive properties of quasispecies are the error threshold, the error catastrophe and the catastrophe of extinction. The *error threshold* is the mutation rate below which populations are in a classical *mutation-selection balance*, in which the genotype with the highest fitness can maintain the replication of the information despite the high mutation rates. The *error catastrophe*, on the other hand, is the phase transition above the mutation rate whereby the loss of the favored genotype, the behavior of quasispecies, and possibly also the viral identity is carry out through frequent deleterious mutations. Finally, the *catastrophe of extinction* represents the complete loss of the viral population through lethal mutations.

In order to demonstrate that a population of viruses with RNA genomes may evolve as quasispecies, we created a computer program, with the structure of a genetic algorithm, which incorporates additional realism to the original quasispecies model proposed by Eigen and Schuster. Given the biological complexity of the parameters needed to run this algorithm, we decided to focus our simulations exclusively on the Human Immunodeficiency Virus type 1 (HIV-1). Our algorithm incorporates biological parameters from HIV-1 virus based on experimental methods (obtained from the literature) and bioinformatics approaches to define: the mutation rate, frequency of mutation types, generation time, and effective population size (N_e), as well as the selective pressure and reproductive cost of mutations in the viral genome. The results of our simulations show the presence of the quasispecies behavior in the evolutionary dynamics of HIV-1 virus, thus, the presence of an *error threshold* during the evolutionary dynamic of HIV-1 virus can be observed. A consensus sequence (from the ancestral genome) is maintained with a high fitness at 1 and 2 mutations per genome per replication. However, the consensus sequence is lost when the initial population is submitted to replicate with 3 and 4 mutations every cycle. This transition phase is known as *error catastrophe*. Consequently, it is possible to note a drastic decrease of the viral fitness. Despite the loss of the quasispecies behavior, the remaining genotypes are still replicating with lower fitness, but with a strong mutational robustness due to the accumulation of negative mutations. The statistical support of our results, however, requires a greater number of simulations (which requires a higher computational power accordingly) to include a larger number of generations than 2,000 and Ne values closer to those ones found in natural populations.

In order to estimate a mutational fitness for the complete nucleotide sequence of HIV-1 virus, we create a *mutational fitness map* for the complete genome by classifying it in genes, functional domains and codons. Although this map shows both sequence variation and conservation throughout the genome of HIV-1 virus, it also surprisingly suggests the tendency to find the action of the positive *versus* negative selection in opposite genomic and structural regions of the viral genome. Thus, the first part of the genome, which include the genomic regions 5'LTR, gag, gag-pol, pol, vif and vpr, are potentially subject to the strong influence of negative selection; whilst the second half of the viral genome, which includes the genomic regions tat, rev, vpu, env, nef and 3'LTR, is potentially subject to the main action of positive selection. In general, the number of nucleotide triplets (i.e., codons in proteins) negatively selected with respect to the positive and neutral ones predominates on the viral genome. We discuss here several aspects and facts previously reported on the literature regarding the genomic structure, evolutionary and functional level of RNA structures and proteins from HIV-1 virus that can support our results and findings. All these results together suggest that the presence of a strong purifying selection in small genomes that are subject to replicate with high mutation rates can easily drive the behavior of quasispecies in RNA viruses, as suggested by Eigen and Schuster in their original model. We hope that this work can contribute to the better understanding of the evolution of RNA viruses and of the genetic material in highly constrained environments.

Keywords: quasispecies, RNA viruses, error threshold, error catastrophe, viral extinction, mutational robustness, epistasis, natural selection, genetic drift, mutation-selection balance, mutational fitness, viral evolution.

INTRODUCCIÓN

I. Los virus: una descripción general de su biología

Los *virus* son sistemas biológicos que necesitan de una célula hospedera para poder replicarse, y sólo así, generar las futuras progenies virales. Como consecuencia de esta dependencia, los sistemas virales han desarrollado dos funciones esenciales para los seres vivos: como agentes infecciosos y como vehículos de transferencia genética horizontal entre los organismos. Estas funciones virales han ocupado un sitio importante en el origen, la diversidad genética, la adaptación ecológica y la evolución de las especies en nuestro planeta (Filée J. et al., 2003, Forterre, 2006). No obstante, debido al controversial origen polifilético de los sistemas virales y a su posición paradójica entre la materia viva y la inerte, los virus han sido excluidos del árbol de la vida (Moreira and López-García, 2009).

Los virus están formados por una cápside de proteínas y por un genoma, y algunos virus además poseen una envoltura membranosa compuesta de lípidos, carbohidratos y proteínas. La *cápside* es una cubierta de proteínas que consiste de varias subunidades estructurales (capsómeros) que son idénticas. Existen diferentes tipos de cápside y éstas son clasificadas de acuerdo a su estructura. En general, los virus presentan 2 tipos de cápside, la helicoidal y la icosahédrica. El primero refleja la asociación de los capsómeros con la hélice del ácido nucleico; mientras que en el segundo, los capsómeros forman una sólida estructura que envuelve al genoma viral. A pesar de la similitud estructural que muestran los virus por dichas cápsides, ciertos virus son más complejos en su estructura [ver Figura I1]. De modo que algunas partículas virales pueden tomar diferentes estructuras dependiendo del tipo de célula a la que infecten, por lo que el arreglo de cada cápside alrededor del material genético es único para cada tipo de virus (Wagner et al., 2008). Asimismo, existe otra capa que cubre a la cápside, que no está necesariamente presente en todos los virus, y es conocida como *membrana* o *envoltura viral*, la cual está compuesta de biomoléculas pertenecientes a la membrana de la célula hospedera, como fosfolípidos y glicoproteínas (Carter and Saunder, 2007).

Mientras que un *virus* es la entidad biológica reconocida como un agente infeccioso que necesita de una célula hospedera para poder replicarse, un *virión* es la partícula viral completamente ensamblada y recientemente liberada de la célula hospedera (Cann, 2005). Los virus comparten características que hacen difícil una clasificación "natural" de los mismos. Algunas clasificaciones se basan en la forma de la cápside, el tipo de hospedero al que infectan o, en otros casos, se basan en el tipo y conformación del genoma que estos presentan.

Los genomas virales están compuestos de Ácido Desoxirribonucleico (DNA) o de Ácido Ribonucleico (RNA), y se encuentran estructurados en cadena sencilla (ssDNA o ssRNA) o en cadena doble (dsDNA o dsRNA). El tamaño del genoma de un virus va desde los 1,360 nucleótidos (nts) (e.g., *Virus de la descomposición de la hoja de cocotero* con un genoma de ssDNA) hasta los 1.2 millones pares de bases (pbs) (e.g., *Megavirus chilensis* con un genoma de dsDNA). Los mayores límites del tamaño del genoma se encuentran en los virus de DNA de cadena doble (Cann, 2005, Arslan et al., 2011). El contenido genómico es igualmente diverso, un genoma viral puede contener desde 4 genes (e.g., el *Fago MS2* con un genoma de ssRNA, (van Duin and Tsareva, 2006)) hasta los 911 genes (e.g., *Mimivirus* con un genoma de dsDNA, (Suzan-Monti et al., 2006)). Los genes virales se encuentran arreglados de forma compacta, mostrando una organización genómica sencilla o muy compleja (con genes anidados y corrimientos del marco de lectura de las proteínas para las que codifican). Asimismo, las regiones genómicas *no codificantes* (e.g., los LTRs –*Long Terminal Repeats*– en el *Virus de Inmunodeficiencia Humana,* con un genoma de ssRNA) presentan diversas funciones regulatorias y de reconocimiento con otras moléculas en los distintos procesos de la infección viral (Watts et al., 2009) [ver Figura I8].

Adicionalmente a estas características, la naturaleza polifilética de los virus se ve reflejada en el uso de diferentes polimerasas para llevar a cabo la replicación del material genético al interior de las células: DNA polimerasa dependiente de RNA (DdRp), RNA polimerasa dependiente de DNA (RdDp), RNA polimerasa dependiente de RNA (RdRp) y Reverso Transcriptasa (RTp) (Duffy et al., 2008). Los sistemas virales, en particular los basados en RNA, se caracterizan por poseer una gran variabilidad genética debido a las altas tasas de mutación generadas por las polimerasas durante la replicación de sus genomas. Por ejemplo, los virus de DNA presentan de $1x10^{-6}$ a $1x10^{-8}$ mutaciones por base por ciclo de generación;

mientras que los virus de RNA muestran de $1x10^{-3}$ a $1x10^{-5}$ mutaciones por base por ciclo de generación (Duffy et al., 2008, Gago et al., 2009). Finalmente, la evolución de los virus presenta procesos y características de gran importancia biológica, tales como la recombinación genética, epístasis, rearreglos, transferencia genética horizontal, segmentación de genomas, genes anidados, estructuración del RNA genómico, diferentes tasas de mutación y sustitución, gran variabilidad genética, ciclos generacionales cortos, tamaños poblacionales grandes, entre otros (Duffy et al., 2008). De esta forma, los virus han evolucionado hacia la especialización de la estructura genómica, lo que les ha permitido la invasión de casi todas las especies de los tres dominios de la vida: Bacteria, Arquea y Eucariota (La-Scola et al., 2003, Wagner et al., 2008).

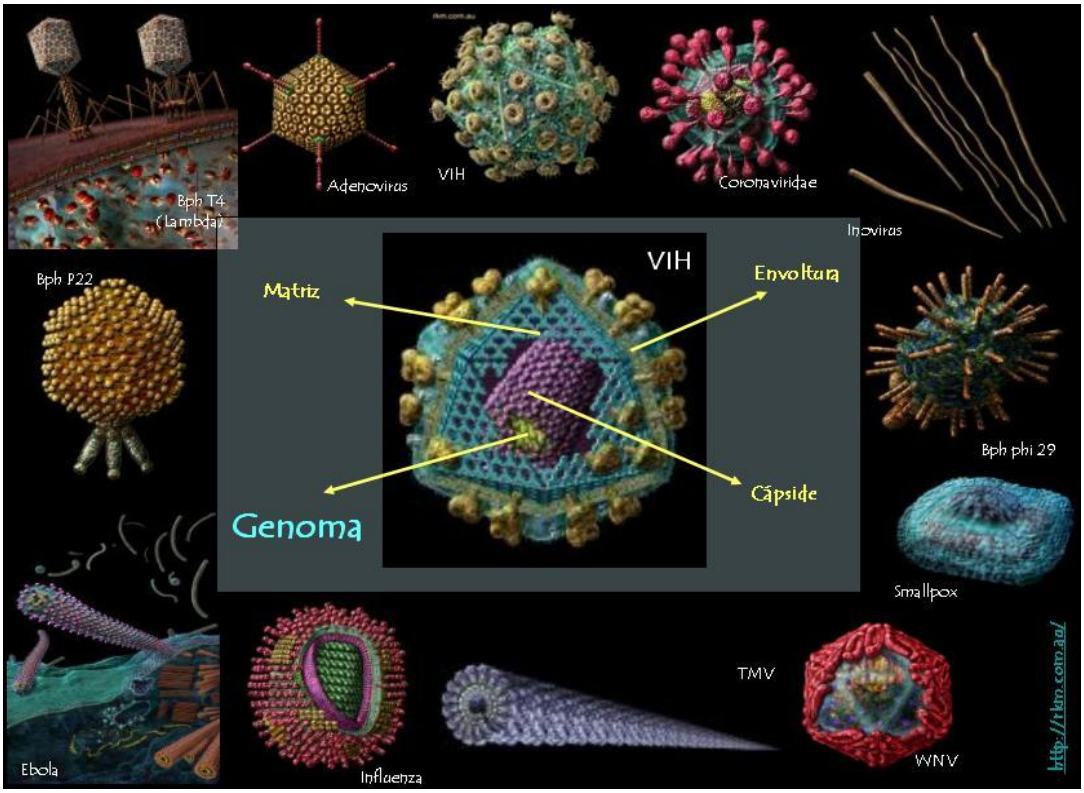


Figura I1. Diferentes tipos y simetrías en varias familias virales. En los virus, básicamente se muestran 4 tipos de estructuras morfológicas: *helicoidal*, en la que se observa al virus como un espiral, e.g. el virus de mosaico de tabaco (TMV); *icosahédrica*, en este tipo de simetría el virus presenta 20 caras con 12 ángulos, como el virus de inmunodeficiencia humana (HIV); *compuesta*, es aquella en la que el virus presenta las dos clasificaciones antes mencionadas, como el caso del bacteriófago T4 (BphT4). Finalmente, está la simetría *compleja*, en la que los virus con grandes genomas (dsDNA) presentan lípidos tanto en la envoltura como en las membranas externas, e.g. el virus de la viruela (smallpox). En el centro se muestran los componentes básicos de una partícula viral en el virus HIV: a) envoltura que cubre a la partícula, b) la matriz, c) la cápside y d) el genoma. Figura tomada y editada de (Lozada-Chávez, 2004).

II. Sobre el origen de los virus.

El origen de los virus ha sido un fenómeno complejo de evaluar debido a la paradójica y controversial naturaleza de los diversos grupos virales. Dentro de las teorías que tratan de explicar el origen de los virus se encuentran (Forterre, 2006, Wessner, 2010): 1) Teoría Progresiva, 2) Teoría Regresiva y 3) Teoría de la Coevolución. A continuación se describen brevemente.

Teoría Progresiva. Esta propuesta es también conocida como la *teoría del escape*, *origen celular* o como la *hipótesis del vagabundeo* o *nomadeo*. De acuerdo a esta teoría, los virus se originaron a través de un proceso progresista de la adquisición de sus componentes. Se ha sugerido, por ejemplo, que elementos genéticos móviles y/o piezas de material genético capaces de movilizarse dentro del genoma (e.g., transposones y retrotransposones) fueron capaces en un momento dado de salir de la célula hospedera cubiertos con parte de la membrana plasmática, de forma que ello les facilitó el reconocimiento y entrada a otra célula hospedera, replicar su material genético, liberarse e infectar sucesivamente a otras células.

Teoría Regresiva. Esta propuesta es conocida también como la *Teoría de la Degeneración***.** En contraste con la teoría anterior, los virus pudieron haberse originado por un proceso regresivo, a partir de la pérdida del material genético y sus componentes celulares. Esta propuesta está basada en el origen y estilo de vida observado de las bacterias endosimbiontes y parásitas. De acuerdo con los defensores de esta teoría, organismos autónomos desarrollaron inicialmente una relación simbiótica. Con el tiempo, esta relación se convirtió en parasítica, en la cual un organismo llegó a ser más y más dependiente del otro. Conforme avanzó la dependencia, el organismo fue perdiendo genes esenciales. Eventualmente fue incapaz de poder replicarse independientemente, llegando a ser un parasito intracelular obligado, es decir, un virus. Por ejemplo, se ha propuesto que las bacterias que viven como parásitos obligados intracelulares, e.g. las especies del género *Chlamydia* y *Rickettsia*, evolucionaron a partir de antepasados con estilos de vida libre (Andersson et al., 1998). Asimismo, se ha sugerido que las mitocondrias de las células eucariontes y *Rickettsia prowazekii* pudieron haber compartido un ancestro común con estilo de vida libre (Andersson et al., 1998). Los grandes virus de DNA nucleocitoplasmáticos (NCLDVs) también podrían ilustrar esta teoría; por ejemplo, los

Poxvirus tienen un gran número de enzimas y otros elementos que permiten producir por sí mismos RNA mensajeros (mRNA), a partir del genoma viral, que cumplen diversas funciones celulares dentro del citoplasma del hospedero (Wessner, 2010).

Teoría del virus primero (o teoría de la coevolución). En las teorías anteriores se asume que los virus y las células tuvieron orígenes independientes o que los sistemas celulares dieron origen a los virales. La teoría de la coevolución propone, por el contrario, que los virus pudieron haber existido primero en un mundo pre-celular como unidades auto-replicantes (Koonin and Martin, 2005). Con esta dinámica, los virus llegaron a ser más organizados y complejos. Eventualmente, surgieron enzimas para la síntesis de membranas y paredes celulares, evolucionando hasta la formación de una célula. Se ha sugerido ampliamente que la primera molécula auto-replicante fue de RNA y no de DNA. Una molécula de RNA que presenta características catalíticas y auto-replicantes es la ribozima, logrando proporcionar soporte a esta teoría. Una pregunta importante a responder es saber si los virus de RNA de cadena sencilla (ssRNA), son descendientes de antiguos sistemas de RNA pre-celular. Otros investigadores presentan esta propuesta con la participación de los NCLDVs, los cuales podrían haber jugado un papel importante en el surgimiento de células eucariontes a través de eventos parecidos a la endosimbiosis (Bell, 2001, Villarreal and DeFilippis, 2000).

III. La evolución viral y la teoría de las Cuasiespecies

A) Los virus de RNA: sus genomas y tasas de mutación

La versión de la base de datos genómicos actualizada hasta febrero del 2012 del Centro Nacional de Información Biotecnológica (del inglés *National Center for Biotechnology Information*, NCBI, http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html), tiene registrada la existencia de 893 genomas completos de virus de RNA de cadena sencilla, 151 de cadena doble y 116 retrovirus (i.e., aquellos que convierten su genoma de RNA a uno de DNA como parte de su ciclo de replicación). Dentro de los virus de RNA de cadena sencilla se encuentran aquellos denominados de “sentido positivo” (ssRNA+), lo que significa que el RNA genómico pasa directamente a ser traducido inmediatamente después de su entrada en la

célula hospedero, i.e., como si el genoma viral fuera un mRNA. También se encuentran descritos aquellos virus de RNA de cadena sencilla de "sentido negativo" (ssRNA-), en el cual el RNA genómico es copiado por una RNA polimerasa para formar un mRNA inmediatamente seguido de su entrada a la célula [ver ejemplos de virus ssRNA+ y ssRNA- en la Tabla I1].

Tabla I1. Tipos de genomas virales y ejemplos de algunos tipos de hospederos. La clasificación de los genomas está basada en el tipo de ácido nucleico y el RNA mensajero producido para su replicación (ver *). Las familias virales aquí listadas, y sus respectivos hospederos, sólo representan algunos ejemplos de los varios virus que se han identificado actualmente.

Genoma*	Familia♦	Virus♦	Hospedero♦	Enfermedad♦
dsDNA	*Herpesviridae*	Herpes simplex virus	Vertebrados (humano)	Cerebro, oral, genital, otros
ssDNA	*Geminiviridae*	Virus del estriado del maíz	Plantas	Enfermedad del estriado
dsRNA	*Reoviridae*	Rotavirus	Vertebrados (humano)	Gastrointestinal y respiratorio
ssRNA(+)	*Picornaviridae*	Fiebre aftosa	Vertebrados (ganado)	Alta fiebre, ampollas dentro en pies y boca y produce cojera
ssRNA(-)	*Orthomixoviridae*	Influenza virus	Vertebrados	Gripa
dsRNA-RT	*Hepadnaviridae*	Hepatitis B virus	Vertebrados	Cirrosis, carcinoma hepatocelular
ssRNA-RT	*Lentivirus*	Virus de inmunodeficiencia humana tipo 1	Vertebrados	Sindrome de inmunodeficiencia Adquirida (SIDA)

* Clasificación de diferentes tipos de genomas presentes en todos los grupos virales (Basados en la clasificación de Baltimore (1971) (Wagner et al., 2008). ss = cadena simple, ds = cadena doble, RT = retrovirus.

♦ Ejemplo de un grupo y tipo de virus, ejemplo de hospedero y tipo de enfermedad que representa un tipo de genoma viral que infectar a un organismo. Debe ser considerado que hay varios hospederos para una misma familia de virus y, asimismo, varios virus de diferentes familias que pueden infectar a similares hospederos.

A la fecha, los virus de RNA son los causantes de las enfermedades infecciosas de más relevancia médica para la población humana debido al número de muertes que han causado y a la rapidez de su propagación (Belshaw et al., 2008, Nichol et al., 2000). El segundo y el sexto lugar los ocupan el Síndrome de la Inmunodeficiencia Humana (SIDA) y el sarampión, respectivamente; mientras que otros tipos de virus de RNA contribuyen a la primera y tercera causas de muerte humana en el mundo, tales como las infecciones respiratorias (e.g., la neumonía viral y la influenza) y la diarrea, respectivamente (*The World Health Organization*, 2004: http://www.who.int/en/). Otras notables enfermedades humanas causadas por virus de

RNA incluyen: la poliomielitis, la hepatitis, la rubeola, la rabia, el ébola, la encefalitis, la leucemia, y el Síndrome Respiratorio Agudo Severo (SARS). Por el contrario, ninguno de los virus de DNA aparece en la lista de los primeros 30 lugares causantes de la mortandad humana, aún cuando se han descrito y secuenciado un número similar de genomas (Belshaw et al., 2008). Se ha considerado que gran parte del poder infectivo de los virus de RNA se debe a su gran variabilidad genética.

Una de las características más importante de los virus de RNA es su alta tasa de mutación. Las tasas de mutación de los virus de RNA estiman de entre 0.4 y 1.1 mutaciones por genoma por ciclo de replicación (Drake et al., 1998). Estas tasas de mutación afectan cada aspecto de la biología de los virus de RNA y son al menos un centenar de veces mayor que el estimado para los virus y bacterias con genomas de DNA (Duffy et al., 2008). Las diferencias entre las tasas de mutación de los virus de RNA y los virus de DNA parecen ser resultado de la ausencia de corrección de las polimerasas dependientes de RNA (Steinhauer et al., 1992). Las polimerasas dependientes de DNA de otros virus y organismos poseen tasas de mutación similares a las polimerasas dependientes de RNA, del orden de 10^{-4} a 10^{-5} por base por ciclo de replicación (pbs/r). Sin embargo, la tasa de error de las polimerasas dependientes de DNA es entonces reducida de 10^{-5} a 10^{-7} pbs/r a partir de la subsecuente corrección de mutaciones (Tippin et al., 2004). Robert Belshaw y colaboradores (2008) han sugerido diversas causas y consecuencias de las altas tasas de mutación en los virus de RNA, a continuación, se desglosa una breve reseña de sus propuestas.

- **Causas de una alta tasa de mutación**

***Historia de vida*.** Dado que muchos virus de RNA infectan a células que poseen un sistema inmune que se adapta constantemente (para reconocer y destruir a los patógenos), se ha propuesto que la presencia de una alta tasa de mutación refleja un mecanismo que permite a los virus adquirir rápidamente variabilidad genética y fenotípica para permanecer indetectables ante el sistema inmune del hospedero. Sin embargo, existen virus de RNA (con altas tasas de mutación) que infectan a bacterias, pero no hacen frente a una respuesta inmune, ya que la célula hospedera no la tiene. Así, las altas tasas de mutación de los virus de RNA no pueden

ser atribuidas fácilmente a su historia de vida.

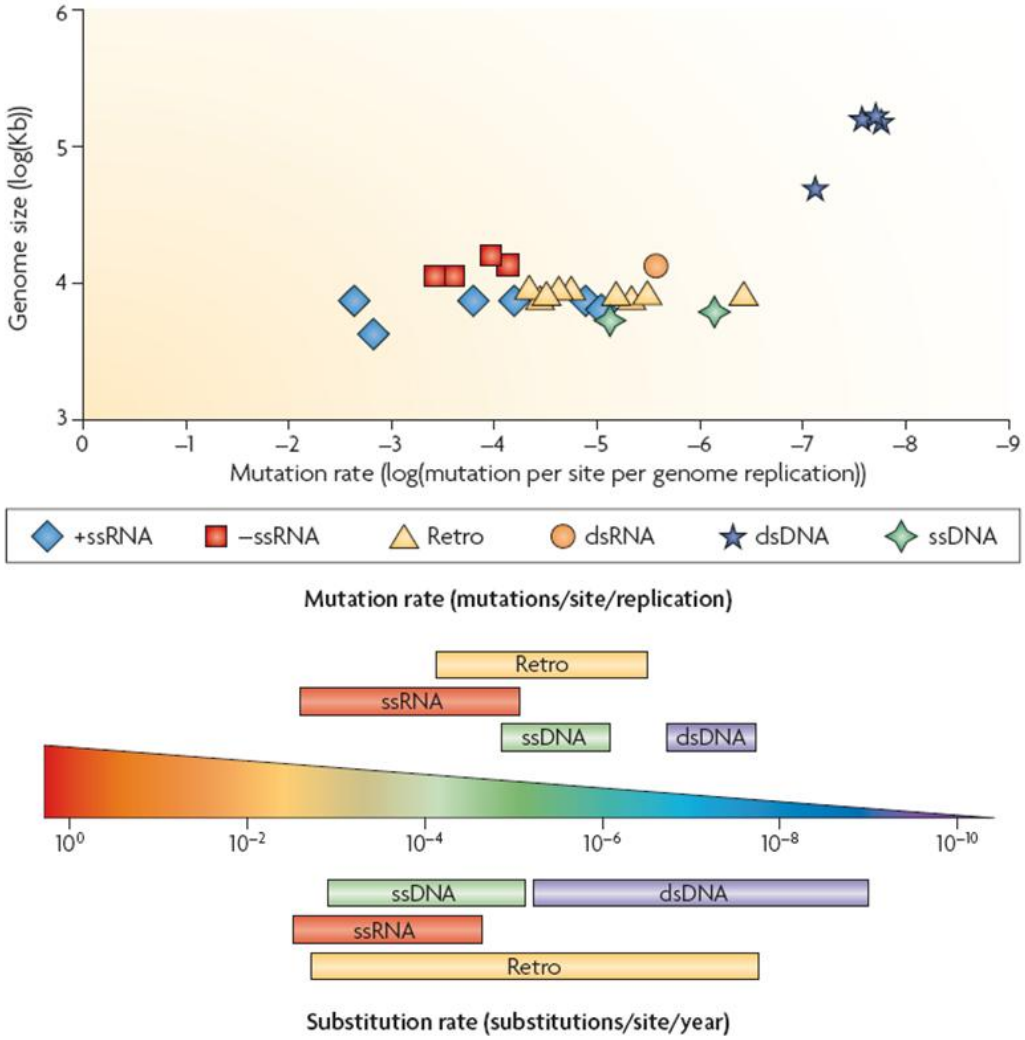


Figura I2. Tasas de mutación y sustitución en los diferentes grupos virales. *Panel superior:* Gráfica comparativa del tamaño del genoma contra la tasa de mutación de los virus de DNA y RNA por medio del promedio de la tasa de mutación espontánea en virus, ajustado a la tasa por genoma por replicación. *Panel inferior*: Comparación de la tasa de mutación contra la tasa de sustitución por año para los virus con genomas de dsDNA, ssDNA, ssRNA y retrovirus. Hay varias estimaciones para los virus que representan a genomas de DNA y RNA, para una revisión sobre el tema, ver Duffy et al. (2008). Figuras tomadas y editadas de Duffy et al. (2008).

Mecanismo de replicación. Es muy probable que el mecanismo de replicación viral tenga un efecto considerable sobre la tasa de mutación (Duffy et al., 2008). Existen dos mecanismos por los que se puede llevar a cabo la replicación del genoma viral, el primero, conocido como "replicación lineal reiterativa" (del inglés *stamping machine*), es aquel en la que el genoma de una sola partícula viral es usada como molde (o el complemento del molde único) para generar

toda la progenie de genomas virales que es producida en una célula infectada; dado que existe un solo templado en la célula hospedera, los genomas virales y las mutaciones se acumularían linealmente bajo este mecanismo de replicación. Por el contrario, en el segundo mecanismo, conocido como "replicación geométrica", algunos de los genomas virales producidos durante las primeras progenies pueden usarse como templados para generar los genomas de las progenies futuras (Chao et al., 2002, Duffy et al., 2008). Bajo este mecanismo de replicación, las mutaciones se acumularían geométricamente debido a que un templado mutado propaga su error a todas sus copias; y dado que existen diversos templados en la célula hospedera, los genomas virales pueden ser producidos a una tasa geométrica o exponencial (French y Stenger 2003, citado en: Duffy et al., 2008). A causa de la misma tasa intrínseca de error de la polimerasa, la "replicación lineal reiterativa" resulta en una tasa de mutación global más baja que en la "replicación geométrica", aunque es posible que algunos virus puedan usar ambos modos de replicación. Por ejemplo, el virus del mosaico del nabo (TuMV) (que posee un genoma de RNA de cadena sencilla) se replica principalmente, pero no exclusivamente, por medio del mecanismo de "replicación lineal reiterativa" (Martinez et al., 2011).

Compensación sobre la velocidad de la replicación. Parece ser la explicación más probable para entender la presencia de altas tasas de mutación en los virus de RNA. *Los virus pueden replicarse rápido o fidedignamente, pero no ambos*. De ahí que una replicación menos fidedigna permita una replicación más rápida, y en consecuencia, aumente la tasa de replicación y la adecuación del virus (Belshaw et al., 2008). Se han encontrado evidencias a favor y en contra de esta hipótesis. Investigaciones a favor han mostrado que los virus que tienen una tasa de mutación reducida, tal como el virus de la estomatitis vesicular, también presentan un *fitness* o adecuación reducida, este *fitness* es medido como el número de individuos que se generan en cada ciclo de replicación (Furió et al., 2005). Otros trabajos *in vitro* han observado que la reverso transcriptasa del virus HIV-1 presenta una relación negativa entre la tasa de polimerización y la tasa de mutación (Furió et al., 2007). Se ha reportado también que mutantes del fago T4 (cuyos cambios se localizaron en un par de nucleótidos en la región de su genoma que codifica para la DNA polimerasa) mostraron variación en su tasa de mutación con más de 4 órdenes de magnitud (Reha-Krantz, 1998). Dichas mutaciones incrementaron la fidelidad de la replicación, pero a la vez disminuyeron la tasa de replicación viral. Evidencia en contra de esta hipótesis se encontró en la reverso transcriptasa (RT) de

HIV-1, donde el cambio de un aminoácido por otro (M→V) en una posición específica reduce la tasa de mutación, mientras que el reemplazo del mismo aminoácido por otro tipo (M→A) incrementa la tasa de mutación; lo paradójico es que ambas tasas de mutación mostraron una alta tasa de polimerización (Pandey et al., 1996). Asimismo, una mutante de la replicasa de poliovirus que presentó un incremento en la fidelidad, no pareció mostrar una tasa de replicación reducida (Vignuzzi et al., 2006).

■ Consecuencias de una alta tasa de mutación

Viabilidad de la población. Dado que muchas mutaciones son dañinas, un incremento en la tasa de mutación en las poblaciones virales puede causar efectos sobre la población. Uno de los efectos más sencillos es suponer la generación de virus menos eficientes, tal que les impida infectar a sus próximos hospederos. La capacidad de infección, o incluso la extinción viral, puede ser entendida a partir de dos modelos teóricos propuestos. La pérdida de la capacidad infectiva puede ser entendida a partir de lo que se conoce como *umbral de error* (integrada en la teoría de las cuasiespecies que se describirá con más detalle en la próxima sección). En breve, el *umbral de error* propone un límite sobre la tolerancia de un genoma para soportar mutaciones que comprometan el fitness del virus, ya que bajo esta propuesta, todas las mutaciones son perjudiciales (Eigen, 1971, Bull et al., 2005). Esta aproximación ha sido ampliamente criticada y se piensa que el *umbral de error* puede ser un artefacto generado a partir de supuestos no realistas del modelo de las cuasiespecies (Bull et al., 2007, Belshaw et al., 2008). Como describiremos más adelante, el *umbral de error* sólo refleja un estado de transición del *fitness* de una población viral (por una posible pérdida de su capacidad infecciosa), pero no la extinción de la población. La *mutagénesis letal*, por otro lado, propone la extinción de la población viral a través de un incremento en la tasa de mutación en sitios que son letales para ésta. La propuesta de la mutagénesis letal ha sido usada como estrategia terapéutica, donde mutágenos químicos como la ribavirina pueden causar mutaciones letales para el virus, inhibiendo así el crecimiento de la población viral hasta su posible extinción (Eigen, 2002, Smith et al., 2005, Bull et al., 2007).

Robustez mutacional. La alta tasa de mutación en los virus de RNA puede ser asociada a genomas que son mutacionalmente robustos, es decir, que pueden tolerar altos niveles de mutación. Esta idea se ha introducido también en términos de la teoría de las cuasiespecies, y se describirá con mucho más detalle en la próxima sección.

Un genoma pequeño. El tamaño de los genomas de los virus de RNA es muy pequeño comparado con los de DNA; el tamaño típico promedio es de 10 kilo pares de bases (kb, 10,000 pbs), mientras que el genoma más grande se encuentra en coronavirus (30 kb), y el más pequeño se encuentra en lentivirus (4 kb). Más que el tamaño pequeño de los genomas de los virus de RNA en sí, lo que es de peculiar interés es el que ninguno de ellos haya evolucionado para poseer tamaños de genomas más grande. Una de las propuestas sobre el tamaño de los genomas de RNA viene de la teoría de las cuasiespecies y menciona que su tamaño es debido a sus altas tasas de mutación, específicamente argumenta que hay una relación inversa entre la longitud de la secuencia (contenido de información) y la tasa de mutación (la fidelidad de copiado). Consistente con tal tendencia, los genomas virales presentan un complejo nivel de compactación (i.e., regiones superpuestas o anidadas). No obstante, dicha relación puede ser contradictoria al realizar un análisis intuitivo: aquellos genomas que experimentan una tasa de mutación muy alta podrían no llegar a exceder ese tamaño ya que muchas mutaciones son perjudiciales. Muchos genomas virales y celulares de DNA, sin embargo, no parecen presentar una relación inversa entre la tasa de mutación y el tamaño del genoma, ya que dicha relación parece constante en estos organismos (Drake et al., 1998).

No es del todo claro si los virus de RNA con los genomas más largos, también poseen tasas de mutación más bajas. Evidencia indirecta muestra que las tasas de sustitución son más bajas en virus con genomas más grandes (Jenkins et al., 2002), aunque las tasas de sustitución pueden ser afectadas por la selección natural, la deriva genética y otros mecanismos tales como la recombinación y la transposición. Del mismo modo, hay genomas virales de RNA que tienen un genoma grande y una baja tasa de mutación, estos virus presentan polimerasas grandes que pueden sugerir mayor “capacidad de fidelidad” al copiar la información. Sin embargo, los virus con genomas pequeños de DNA (e.g., parvovirus) también presentan tasas de sustitución similares a los virus de RNA (Shackelton et al., 2005), de forma que esta hipótesis no puede ser sustentada del todo.

B) El concepto de Cuasiespecies y sus propiedades más significativas.

En este trabajo, nos enfocaremos a analizar y entender con más profundidad la teoría de las c*uasiespecies*. Esta teoría es de las más mencionadas en la comunidad científica para explicar la evolución de los virus de RNA. El concepto y modelo matemático fue inicialmente desarrollado por los químicos teóricos Peter Schuster y Manfred Eigen para describir el comportamiento y evolución de pequeñas moléculas de RNA (replicones) en el origen de la vida (Eigen and Schuster, 1977). En términos generales, el modelo de las cuasiespecies describe la evolución de ciertas entidades auto-replicantes dentro de un ambiente gobernado principalmente por leyes físico-químicas. De esta forma, el modelo se basa en cuatro supuestos:

(1) Las entidades auto-replicantes pueden ser representadas como secuencias compuestas de un pequeño número de monómeros, tales como los nucleótidos de adenina, timina, citosina, guanina (si consiste de DNA) o uracilo (si consiste de RNA).
(2) Nuevas secuencias entran al sistema solamente como resultado de un proceso de copiado de las secuencias existentes, ya sea un copiado correcto o erróneo.
(3) Los sustratos químicos (e.g., monómeros) que permiten la realización de tal copiado se encuentran siempre presentes en cantidades suficientes para llevar a cabo el proceso. El exceso de secuencias se depura en cada nueva generación.
(4) Las secuencias pueden descomponerse en sus monómeros iniciales. La probabilidad de la descomposición no depende de la "edad" de las secuencias, las secuencias con mayor o menor tiempo de vida son igualmente probables de descomponerse.

En el modelo de las cuasiespecies, las mutaciones se producen a través de los errores cometidos en el proceso de copiar las secuencias ya existentes, i.e., a partir de un proceso de replicación. Además, la selección surge porque los diferentes tipos de secuencias tienden a replicar a diferentes velocidades, lo que conduce a una disminución de las secuencias que replican más lentamente en favor de las secuencias que replican más rápido. Sin embargo, el modelo de las cuasiespecies no predice la extinción total de las secuencias que se replican más lentamente, sino que predice la extinción de las que se replican más rápido si es que se sobrepasa el balance respecto al número de mutaciones que se pueden acumular sin perder el

contenido informacional a expensas de mantener una alta replicación. A pesar de que las secuencias que se replican más lentamente no pueden mantener su nivel de abundancia por sí mismos, estas secuencias se renuevan constantemente tanto como las secuencias que se replican más rápido muten reversiblemente (en ellos) hasta que tengan la misma composición. Es importante recordar que para que esta reversión pueda llevarse a cabo, las secuencias deben tener tamaños cortos, de forma que el espacio composicional pueda ser amplia y rápidamente explorado. En el equilibrio, la eliminación de secuencias con una replicación más lenta (debido a la descomposición o la salida) es balanceada por la reposición; de modo que incluso las secuencias que replican relativamente de forma más lenta pueden seguir presentes en la abundancia de lo finito. De esta forma, una *cuasiespecie* es inicialmente definida como *una distribución de mutantes en equilibrio estacionario de tamaño infinito o finito que se encuentra centrada alrededor de una o varias secuencias consenso maestras* (Eigen, 1971, Eigen, 1993b). A través de la teoría de las cuasiespecies, Schuster y Eigen pretendieron explicar el origen de la vida (su dinámica metabólica y genética en base a RNA) a partir de la generación de cuasiespecies (Eigen and Schuster, 1977, Domingo et al., 1997).

No obstante, el concepto de las cuasiespecies ha sido utilizado para entender la evolución de los sistemas virales, debido a la alta similitud de las características compartidas entre los replicones prebióticos y los virus de RNA [ver Tabla I2]. Ambos sistemas son auto-replicantes, con un contenido informacional pequeño, replican y mutan a altas tasas, y sus componentes requeridos para replicar (e.g. nucleótidos en virus) son indefinidos mientras éstos sean provistos por la célula hospedera a la que infectan. De esta forma, la comunidad científica sugirió prontamente que los virus evolucionan como *cuasiespecies virales* (Eigen, 1993a). Esta propuesta fue inspirada inicialmente por los experimentos realizados por Spiegelman y colaboradores (1965) y, posteriormente, por los avances experimentales de Weissmann y colaboradores (1976, 1978) sobre la replicación *in vitro* del fago Qβ con genoma de RNA. Spiegelman, encontró que al introducir el genoma de RNA del bacteriófago Qβ en una solución con nucleótidos, más una enzima RNA replicasa y algunas sales, el RNA empezó a ser replicado inmediatamente *in vitro* por la enzima de RNA. Después de 74 generaciones, el genoma original (de 4,500 nts) terminó haciéndose muy pequeño (218 nts). Además, observó que la velocidad de la replicación aumentó en el proceso a pesar de las circunstancias (Spiegelman et al., 1965, Kacian et al., 1972). Esto atrajo a Eigen y temporalmente a Francis

Crick a formar una primera aproximación teórica sobre el fenómeno de la replicación auto-catalítica, empleando conceptos como 'copia maestra' y 'copias con errores' para describir a las cuasiespecies (Domingo et al., 1997). Posteriormente, surgieron los hallazgos de Wiessman y colaboradores al trabajar con el mismo bacteriófago Qβ, en el que hallaron una gran heterogeneidad y desventajas selectivas en la información genética del virus debido a una alta tasa de mutación por ronda de replicación. Haciendo énfasis en los cambios genéticos del virus (i.e., en la estructura genómica), ellos llegaron a la conclusión de que el bacteriófago Qβ no puede ser descrito como una única estructura definida, sino más bien como un peso promedio del gran número de diferentes secuencias individuales (Domingo et al., 1976, Domingo et al., 1978). Estos descubrimientos fueron expuestos a Eigen, quién incorporó la heterogeneidad poblacional y las tasas de mutación a las ideas previamente consultadas con Crick y, logró así consolidar, con la ayuda de Schuster, la *teoría de las cuasiespecies* (Domingo et al., 1997).

Tabla I2. Características similares entre las poblaciones teóricas de replicones y las virales con genomas de RNA. El signo ⊕ representa similitud, mientras que el ⊗ significa que no existe comparación o analogía parecida. (*) Interacción con moléculas del ambiente (proteínas, RNAs, etcétera que provienen de un sistema diferente, tal como el hospedero). Información recopilada de Eigen (1971, 1993b), Eigen y Schuster (1977), Domingo y Holland (1997).

Tipo de características	Viral RNA genomas	Cuasiespecies (Replicones)
Secuencia de información biológica de corta longitud	⊕	⊕
Molécula de información de RNA	⊕	⊕
Single stranded (ssRNA)	⊕	⊕
Altas tasas de mutación	⊕	⊕
Grandes poblaciones heterogéneas (N_e)	⊕	⊕
Tiempos generacionales cortos	⊕	⊕
Interacciones con otras moléculas (nivel de complejidad)*	⊕	⊗

Finalmente, el concepto de ‘cuasiespecies’ fue analizado experimentalmente en algunos virus de RNA por Esteban Domingo, John Holland, Manfred Eigen y Christof K. Biebicher, Howard M. Temin, entre otros (Domingo et al., 1997), encontrando que las poblaciones virales de RNA evolucionan en conjunto con altas tasas de mutación sin perder su información consenso y, por ende, tampoco pierden su capacidad infectiva. De esta forma, los hallazgos fueron extrapolados a la teoría de Eigen y Schuster por estos investigadores. Sobrevalorado por los experimentos iniciales, sin embargo, el concepto de cuasiespecie es entendido (y en muchas ocasiones erróneamente empleado) para describir solamente la persistencia de una cepa infectiva sometida a varias rondas de replicación (i.e., por medio de plaqueos o pasajes) a través de altas de tasas de mutación, sin tener en cuenta la dinámica general que describieron Eigen y Schuster para explicar la evolución de las moléculas de RNA. En una sección posterior se ahondará con más detalle en la problemática sobre las atribuciones biológicas y evolutivas del concepto de cuasiespecies.

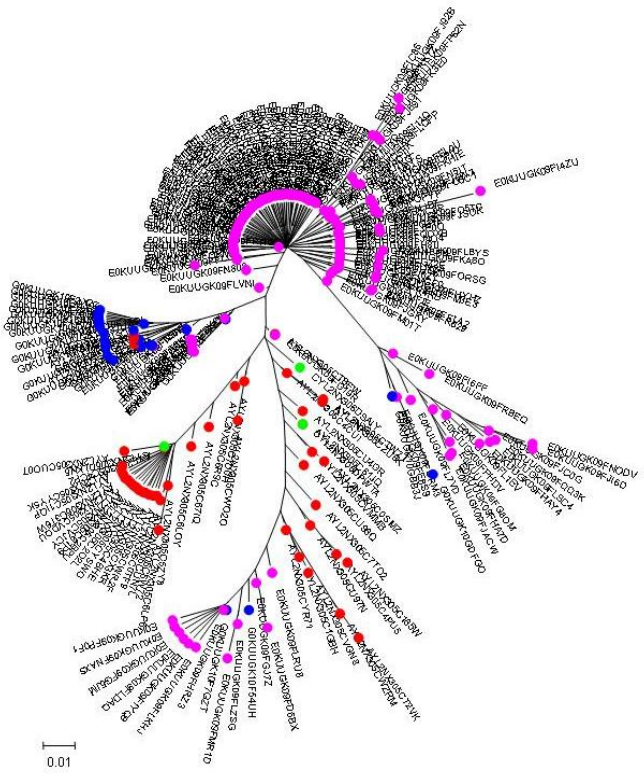


Figura I3. Árbol filogenético de la secuencia variable de aminoácidos de la región gp120-V3 del virus HIV-1. Distribución de secuencias a través del tiempo en pacientes infectados con HIV-1. El largo de las ramas refleja la “similitud” que presentan estas secuencias. Cuasiespecies provirales almacenados en monocitos (círculos rojos) y en linfocitos T (círculos verdes) extraídos de un paciente fueron analizados después de suprimir por un largo plazo la viremia. Un mes después se analizaron otras muestras del mismo paciente bajo terapia sin interrupción, se muestran los monocitos (círculos azules) y los linfocitos (círculos rosados-púrpuras) de nueva cuenta. Figura tomada de Rozera et al. (2009).

Las *cuasiespecies virales* se pueden definir, así, como *entidades auto-replicantes que conforman un grupo o “nube” de genotipos relacionados* (i.e., mutantes similares) *que coexisten en un ambiente de altas tasas de mutación* [ver Figura I4]. De esta forma, se espera

que una gran proporción de las secuencias descendientes en cada ciclo de replicación contengan una o más mutaciones respecto a la secuencia parental. Las mutaciones periódicas dentro de la población viral, si benefician la supervivencia y la reproducción de los individuos (relación conocida como adecuación, o *fitness* en inglés), pueden dar lugar a uno o más genotipos competentes (ver Figura I3). Estos genotipos son organizados alrededor de una o algunas secuencias que presentan los valores de *fitness* (adecuación) más altos dentro de la población. Las secuencias con los fitness más altos (o *fittest* en inglés) son denominadas *secuencias maestra* o *secuencias consenso.*

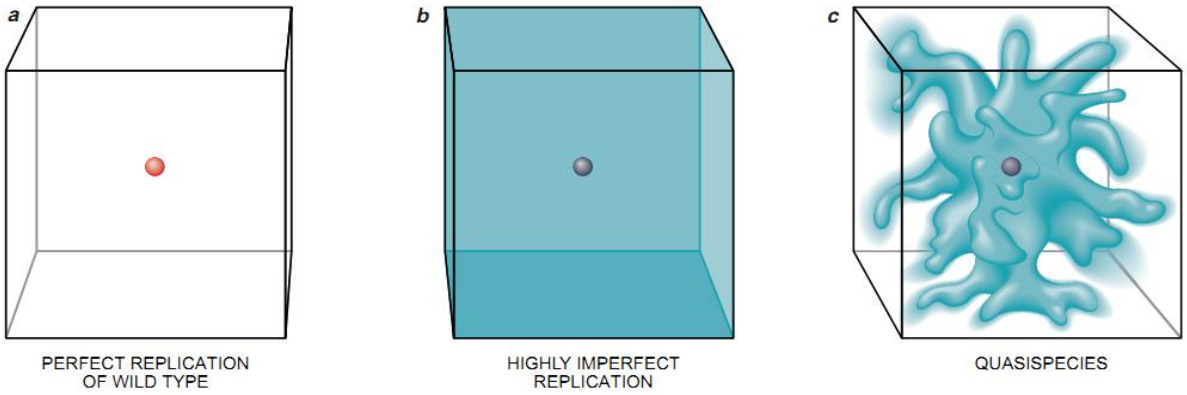


Figura I4. Dinámica poblacional de un virus dependiendo de su tasa de mutación y su proceso de replicación. Simplificación de la distribución de la secuencias en la secuencia espacio que contiene a la población viral. **(a)** Si el proceso de replicación es exacto, la descendencia viral podrá ocupar el mismo espacio siempre. **(b)** Si la replicación es altamente 'imperfecta', los virus generados ocuparían rápidamente cualquier posición de la secuencia espacio, y la población perdería su integridad informacional. **(c)** A una tasa intermedia de error, la población puede ser auto-sustentable, logrando parecer una nube de descendientes centradas sobre la secuencia consenso original. Esta nube representa una *cuasiespecie.* Figura tomada de Eigen (1993b).

El concepto de cuasiespecies es diferente al de *especie*, ya que los individuos de una especie poblacional poseen un genotipo cercanamente similar al parental. De forma que las *especies* descendientes de una población serán copias casi idénticas o fieles al individuo parental. Durante la evolución de las *especies* biológicas, el genotipo con el mayor *fitness* puede reemplazar al genotipo parental para dominar el espacio adaptativo de la *especie*. Es así como los genotipos individuales (o *especies*) pueden ser vistas como las unidades de selección de una población. Sin embargo, los genotipos de las *cuasiespecies* están constantemente cambiando debido a que están mutando rápidamente, de modo que el *fitness* de un genotipo particular no proporciona una relevancia evolutiva tan significativa como en la evolución de

las *especies* biológicas. La conexión (similitud) entre todos los genotipos que conforman la nube de *cuasiespecies* es lo que realmente se entendería como la unidad natural de selección de las poblaciones virales. Si muchos genotipos están cercanamente relacionados, e.g. cuando difieren en una mutación, entonces los genotipos de este grupo tienen una alta probabilidad de revertir mutacionalmente a cualquier otra secuencia de la nube de *cuasiespecies*. Es importante recordar nuevamente que esta dinámica evolutiva sólo es posible en secuencias cortas (genomas pequeños) y con altas tasas de mutación.

Debido a los razonamientos anteriores, se ha propuesto que las *cuasiespecies* aparecen debido a un balance entre la mutación y la selección. A este balance, en términos generales, se le conoce como el *balance de mutación-selección* (Bull et al., 2005, Wilke, 2005). El *balance de mutación-selección* se presenta cuando la selección natural incrementa la frecuencia de variantes *aptas,* mientras que la mutación introduce variantes *no aptas,* dando lugar a una distribución de equilibrio balanceado entre estos dos efectos. Este balance es definido como el equilibro que se da entre la tasa a la cual un gen aumenta el número de mutaciones recurrentes y su eliminación por selección natural. En general, la selección natural a nivel molecular puede ser divida en dos tipos: la *positiva* y la *negativa*. La selección negativa, también conocida como *purificadora* o *estabilizadora,* acelera la pérdida de aquellos alelos que son mutantes (o deletéreos) y que comprometen el *fitness* del genotipo o genotipos más óptimos, de forma que estos alelos se extinguen de la población o su frecuencia se mantiene a bajas tasas en las siguientes generaciones. Por otro lado, la selección positiva, también conocida como selección *direccional o diversificadora,* es un tipo de selección natural que favorece la fijación rápida de un alelo ventajoso en la población. La selección positiva es considerada como el mecanismo evolutivo mediante el cual se promueve la presencia de mutantes que poseen *fitness* mayores que el promedio de la población total, de modo que las frecuencias de dichos mutantes se incrementan en las siguientes generaciones. Así, la selección positiva promueve la diversidad genética, mientras que la selección negativa disminuye o purga esta diversidad eliminando a las variantes de la población (Castillo, 2007). Para Eigen y Schuster, la evolución de los genotipos del tipo *cuasiespecies* sólo tiene lugar a partir de una *selección purificadora* (Jenkins et al., 2001), donde las mutantes de la *secuencia consenso* disminuyen su *fitness* y son así seleccionadas en contra, para ser purgadas eventualmente de la población. De esta forma, las cuasiespecies evolucionan como una sola entidad, y no como genomas

individuales, y exclusivamente dentro de un 'umbral de error' específico que mantenga la replicación de la información genética a través de la generaciones virales.

A partir de esta dinámica evolutiva, los sistemas que evolucionan como cuasiespecies deberían exhibir alguna de sus tres propiedades o estados más distintivos (Bull et al., 2005). Estas propiedades o características se describen a continuación.

1. El ***umbral de error*** o ***límite de Eigen*** representa la tasa de mutación a través de la cual una población viral encuentra el equilibrio de un proceso tradicional de mutación-selección y, por ende, el genotipo favorecido (i.e., con el *fitness* más alto) puede mantener la replicación de la información genética a pesar de duplicarse bajo altas tasas de mutación (ver Figura I5C). Cuando la tasa de mutación cambia ligeramente (sin cruzar el umbral de error o límite de Eigen), la población se moverá a un nuevo "equilibrio" entre uno o más óptimos locales, estableciendo así una relación entre la mutación y la selección. Este fenómeno, que ya ha sido referido anteriormente como el "balance de mutación-selección", está siempre presente en las dinámicas evolutivas de las poblaciones virales, logrando alcanzar continuamente estados de optimización a diferentes presiones de selección El *umbral de error*, de hecho, es la zona de mutación estable que mantiene el comportamiento de cuasiespecies [ver Figura I5C].

2. La ***catástrofe de error***. Eigen y Schuster encontraron, sin embargo, que un aumento en la tasa de mutación por arriba del umbral de error, incluso uno casi insignificante, puede conducir a un cambio fundamental en la composición de los genotipos en la población. Este cambio es una transición de fase en términos físicos y se denomina *catástrofe de error* (o *Trinquete de Muller,* desde la genética de poblaciones), el cual conlleva a la pérdida del genotipo favorecido, y posiblemente también a la pérdida de la identidad viral a través de frecuentes mutaciones deletéreas (Eigen, 1993a, Eigen and Schuster, 1977, Eigen, 1992, Nowak, 2002). De esta forma, cuando el umbral de error se cruza, o las poblaciones no se pueden mantener en este "límite", se da lugar a la catástrofe de error con la consecuente pérdida del comportamiento de cuasiespecies [ver Figura I5B].

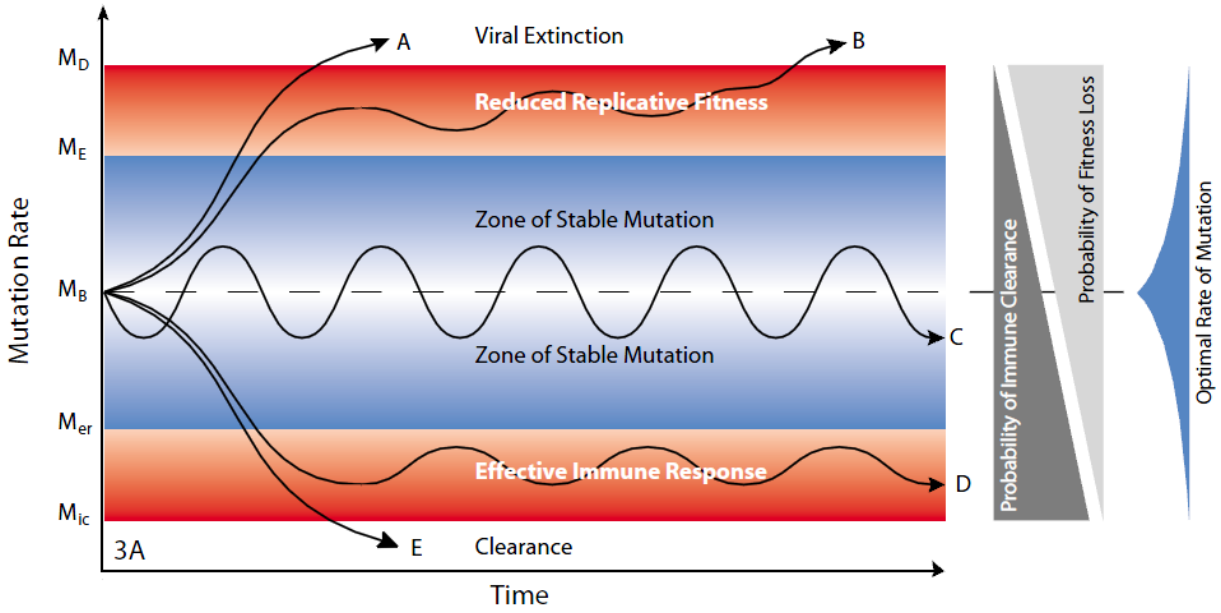


Figura I5. Propiedades de las Cuasiespecies. Se presenta una representación teórica de la distribución de las poblaciones virales en función de la tasa de mutación (Sallie, 2005). Esta distribución promueve la dinámica de las cuasiespecies: (C) balance de mutación-selección y el umbral de error; (B) representa la catástrofe de error y (A) la catástrofe de extinción. (D) y (E) representan el efecto de una tasa de mutación muy baja más la acción de un efectivo sistema inmune y el proceso de limpieza de partículas virales que se presenta en el hospedero, respectivamente.

No obstante, se ha reconocido tanto teórica como experimentalmente la presencia de otra población que se mantiene a pesar de la pérdida de las cuasiespecies. Este proceso, conocido como *el efecto cuasiespecies*, hace referencia a la redundancia de genotipos que pueden presentar las poblaciones virales ante el incremento en la tasa de mutación y altas presiones de selección. La redundancia se interpreta como el nivel de competitividad de un genotipo con una baja tasa de replicación pero con una gran *robustez mutacional* (Codoñer et al., 2006, Belshaw et al., 2008). La 'robustez mutacional', como se mencionó anteriormente, se define como la constancia de un fenotipo frente a mutaciones negativas. Este escenario es denominado también como "supervivencia de los más robustos o planos" (*flattest* en inglés), en contraposición al adjetivo de la "supervivencia del más apto" (*fittest en inglés*) propuesto por Charles Darwin (Darwin, 1869). Y refleja un *sistema de competencia-presencia* de dos poblaciones, en la que una población con una "frecuencia más baja" de crecimiento es capaz de competir contra aquellos que presentan "frecuencias más altas" de crecimiento cuando la tasa de mutación se incrementa considerablemente (Codoñer et al., 2006). La teoría de las cuasiespecies predice que "los replicadores más lentos" pueden ser favorecidos si incrementan la frecuencia de su progenie en la que estos, en promedio, son más aptos. Así, estas

poblaciones ocupan una pequeña región dentro del *fitness landscape*, no en lo más alto, sino en las regiones más planas de este valle adaptativo (Wilke, 2005, Lauring and Andino, 2010) (ver Figura I6). Se sabe que los sistemas robustos siguen siendo viables en el proceso infectivo a pesar de dichas presiones de selección (Aaskov et al., 2006).

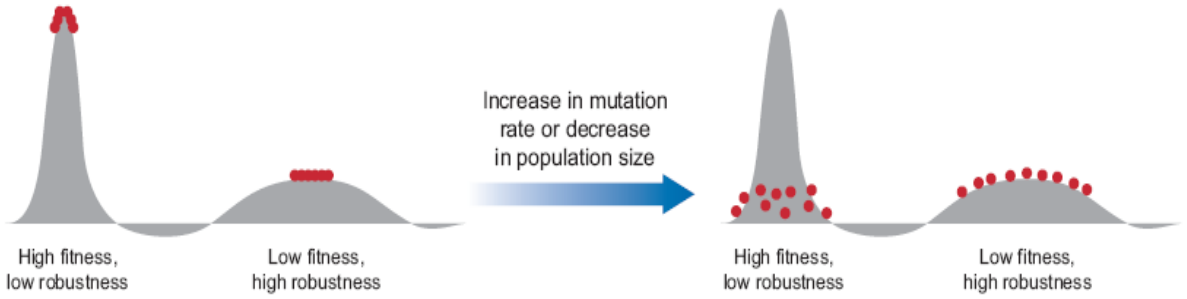


Figura I6. Sobrevivencia del más apto vs sobrevivencia del más robusto. Empleando la metáfora de Wright sobre el nivel de adaptación, se muestra un *paisaje adaptativo* caracterizado por un pico de 'alto fitness pero con baja robustez' (pico más alto), y otro con 'bajo fitness pero con alta neutralidad' (es decir, robustez, pico bajo) (izquierda). Fluctuaciones en el tamaño poblacional o en la composición genómica cambian esta dinámica: cuando la tasa de mutación incrementa o, el tamaño de la población disminuye, la población que se encuentra en el pico más alto, sufre un decaimiento considerable del fitness colocándose en regiones más planas, perdiéndose ese genotipo o, presentando un fitness más bajo que el de las poblaciones robustas (derecha). Las poblaciones 'robustas' y 'no robustas' presentan significantes efectos en su fitness. Figura tomada de Elena y Sanjuán (2007).

La robustez de las poblaciones virales se ha detectado recientemente de forma experimental en virus de RNA, *e.g.*, en el Virus de la Estomatitis Vesicular (VSV) (Torre and Holland, 1997), en el virus del dengue (Aaskov et al., 2006) y en "organismos digitales" *in silico* (Wilke et al., 2001). En estas investigaciones se muestran poblaciones 'defectuosas' (refiriéndose a las robustas) dada la alta tasa de mutación, que preservan su frecuencia y estabilidad poblacional. Este proceso, puede verse beneficiado en poblaciones pequeñas, en las que la deriva génica favorece la acumulación de mutaciones o, por el contrario, puede verse debilitado rápidamente por la disminución de la tasa de mutación o, teóricamente, por el incremento en el tamaño del genoma (Sanjuán, 2010, Sardanyés and Sanjuán, 2011).

Se ha considerado también la presencia de dos dinámicas poblacionales, esto es, la existencia de más secuencias maestras distribuidas en otros sitios del paisaje adaptativo. La concepción de un solo pico (y no un valle adaptativo) como espacio de exploración de las poblaciones virales, ha sido críticamente cuestionado por la Genética de Poblaciones. Aunque es sabido que un paisaje adaptativo puede ser más complejo, recientemente se ha corroborado que

pueden existir más picos adaptativos en las poblaciones de una misma especie viral, en la que al menos pueden coexistir dos secuencias maestras en espacios distantes del paisaje adaptativo (Schuster and Swentina, 1988). Schuster y colaboradores demuestran la presencia de dos secuencias maestras, las cuales pueden ser localizadas en regiones distantes separadas por zonas de *fitness* bajos, lo que indica que posiblemente no están conectadas entre sí. A este fenómeno se la ha llamado *degeneración de cuasiespecies.*

3. La ***catástrofe de extinción***. Se presenta con la pérdida completa de los genotipos que componen a la población viral, por ejemplo, a través de mutaciones letales. En otras palabras, esto ocurre cuando el mejor genotipo ya no puede reproducirse a sí mismo para mantener un mínimo tamaño poblacional [ver Figura I5A]. Así, altas tasas de mutación pueden causar la extinción, sólo si las mutaciones son letales. En términos prácticos, el ejemplo más trivial para distinguir entre la *catástrofe de error* y la *catástrofe de extinción* es el caso en el cual un solo genotipo existe y todas las mutantes son letales (Bull et al., 2005).

C) El problema biológico sobre el uso de la teoría de las Cuasiespecies

En años recientes, varios autores han cuestionado si la teoría de las cuasiespecies representa, en la naturaleza, la evolución de los virus de RNA o si ésta es relevante para el análisis de estos sistemas biológicos (Moya et al., 2000, Jenkins et al., 2001, Holmes and Moya, 2002, Comas et al., 2005). Existen al menos tres problemáticas sobre la modelación y utilización de la teoría de las cuasiespecies para explicar la evolución de los virus. En primer lugar, no es del todo claro o comprobado si el modelo evolutivo de las cuasiespecies es realmente diferente del que propone la genética de poblaciones para explicar la evolución viral (ver Tabla I3) (Wilke, 2005). En segundo lugar, no se han medido, estimado o comprobado del todo las propiedades de las cuasiespecies en los sistemas virales, sólo existen un par de ejemplos comprobados teórica y experimentalmente (Jenkins et al., 2001). En tercer lugar, no se han estimado ni modelado dinámicas evolutivas de cuasiespecies incorporando más realismo biológico a la propuesta original de Eigen y Schuster (Holmes and Moya, 2002). A consecuencia de estos factores, finalmente, la comunidad científica ha hecho un mal uso o empleo de los conceptos y propiedades de las cuasiespecies asumiendo, en la mayoría de los casos, que ésta es la

dinámica evolutiva a la que están sujetos todos los virus. A continuación vamos a detallar algunos ejemplos que ilustran esta problemática.

Tabla I3. Tabla comparativa de terminologías empleada en la "teoría de las cuasiespecies" y en la "Genética de Poblaciones". Es posible observar que en lo referente a las Cuasiespecies, se presenta un limitado número de elementos que participan en sus dinámicas evolutivas en comparación con la de la Genética de Poblaciones.

	Modelado de las poblaciones (modelos)	
Tipo de características	Teoría de las Cuasiespecies	Balance de Mutación-Selección
Locus/Loci *Descripción teórica*	Múltiples loci Determinístico: poblaciones infinitas	Uno-dos locus Efectos estocásticos y poblaciones finitas
Paisaje de Fitness (paisaje adaptativo)	Un único pico: No existen posiciones neutrales	Múltiples picos diferenciales: efecto de la selección positiva, negativa y la neutralidad
Predicciones	Umbral y catástrofe de error: extinción viral	Umbral de letalidad: No hay viral extinción, *no hay un umbral de error*
Fuerza evolutiva en acción	Sólo existe selección purificadora, no hay fluctuaciones en N_e	Deriva génica, hitchhiking, epístasis, pleiotropía, recombinación, otras fuerzas
Unidad de la Selección	Población	Individuos

El *umbral de error* se ha usado como una metáfora de las complicaciones sobre la presencia de altas tasas de mutación en un sistema que almacena poca información genética. Una problemática que se considera fue característica de la vida primigenia (Eigen, 1971, Schuster and Stadler, 2008), y se ha propuesto que desafía actualmente la evolución de los virus existentes con genomas de RNA (Anderson et al., 2004). De esta forma, el término "umbral de error" se ha usado como equivalencia del concepto de *cuasiespecies*, y éste último se ha usado como sinónimo de "alta tasa de mutación" y de "heterogeneidad poblacional". De la misma forma, se ha empleado erróneamente el concepto de 'error de catástrofe', ya que ha sido asociado frecuentemente a la "extinción viral", cuando formalmente está asociado a la "catástrofe de extinción", previamente descrito.

Sobre la validación experimental del concepto de las cuasiespecies, las evidencias empíricas argumentan que la alta tasa de mutación de los genomas de RNA debe conducir a un equilibrio

del proceso mutación-selección, el cual produce una gran cantidad de variación al mismo tiempo que favorece la rápida adaptación del virus a los cambios en el sistema inmune de sus hospederos (Bull et al., 2005, Wilke, 2005). De este modo, diversas observaciones experimentales sobre los virus de RNA han mostrado ciertas características que son compatibles con la teoría de las cuasiespecies, tales como: alta tasa de mutación/sustitución (Duffy et al., 2008, Mansky, 1998, Drake et al., 1998, Furió et al., 2005, Domingo and Holland, 1997), presencia de "secuencias consenso" (Domingo et al., 1978, Steinhauer et al., 1992, Rozera et al., 2009), presencia de mutaciones deletéreas (Domingo and Holland, 1997, Burch and Chao, 2000, Domingo et al., 2008, Duffy et al., 2008, Holmes, 2009) y selección purificadora (Holmes, et al., 2006, 2003a, Jerzak et al., 2005, en Holmes, 2009) (Muller y Bonhoeffer, 2005, en Santiago y Wagner, 2008) (Chare and Holmes, 2004, Elena and Sanjuán, 2005, Hughes and Hughes, 2007, Wertheim and Worobey, 2009). Sin embargo, a pesar de la frecuente referencia en la literatura al concepto de cuasiespecies para explicar la evolución de los virus, muy pocos estudios han demostrado que los virus (en particular los de RNA) realmente evolucionan como cuasiespecies. Actualmente, sólo se ha evaluado experimentalmente el comportamiento de cuasiespecies en muy pocos grupos virales: el fago Qβ (Domingo et al., 1978), el virus de la fiebre aftosa (FMDV) (Domingo et al., 2005, Perales et al., 2009, Pariente et al., 2003), el virus de la hepatitis tipo C (Forns et al., 1999, Martell et al., 1992), y el poliovirus (Vignuzzi et al., 2006), todos estos virus poseen un genoma de ssRNA (+).

Parte de la dificultad para evaluar el comportamiento de cuasiespecies en los virus es que esta teoría considera a la mutación solamente como un proceso degenerativo que reduce el *fitness* de las partículas virales, y no toma en cuenta a otras fuerzas evolutivas más allá de la selección purificadora. Eigen y Schuster no consideran a la mutación como un *proceso de exploración* para encontrar nuevos estados adaptativos bajo selección positiva o neutra. Una exploración adaptativa que definitivamente se lleva a cabo en las poblaciones virales. Si bien los autores no niegan la presencia de mutaciones neutrales, argumentan que ambos tipos de mutaciones (positivas y neutras) pueden ser casi totalmente ignoradas en sistemas biológicos con genomas pequeños, tamaños poblacionales grandes y con altas tasas de mutación (Eigen et al., 1988). Estas características permitirían que el espacio adaptativo, generado por un reducido número de posiciones neutras, pueda ser completamente y rápidamente explorado. Bajo esta

suposición, la población no se somete a la deriva génica o a otro tipo de selección que no sea la purificadora (Eigen, 1992). No obstante, se ha reportado ampliamente que los genomas virales de RNA muestran *selección positiva* en algunas de sus regiones genómicas codificantes, lo que permite a la población explorar otro espacio adaptativo (Price et al., 1997, Yoshida et al., 2011a).

Además de la selección natural, se han reportado otros mecanismos y fuerzas evolutivas que pueden participar en la fijación de mutaciones en las poblaciones de los virus, tales como la migración, la recombinación y los rearreglos, la transferencia horizontal, la deriva génica y el hitchhiking génico, entre otros (Domingo et al., 1997, Domingo and Holland, 1997, Furió et al., 2005, Duffy et al., 2008, Holmes, 2009). Dado que estos mecanismos son ajenos a la teoría de las cuasiespecies, la teoría de la *genética de poblaciones* propone que la participación de estos mecanismos son los que realmente controlan (y no sólo la tasa de mutación) la dinámica evolutiva de las poblaciones virales [ver Tabla I2]. Según la genética de poblaciones, la variación observada en las poblaciones virales no se debe a un equilibrio del proceso mutación-selección, sino que es derivada de la simple selección de diferentes genotipos en un ambiente altamente heterogéneo en el hospedero (Holmes and Moya, 2002). Asimismo, elementos usados en la teoría de las cuasiespecies, tales como la tasa de replicación y de mutación, el tamaño efectivo de la población y el efecto de las mutaciones sobre las nuevas generaciones, presentan variaciones considerables en las poblaciones naturales. Estimaciones exactas para cada uno de estos parámetros son difíciles de obtener y su variación podría representar problemas en el modelado de las cuasiespecies. Es así como la falta de cierto realismo biológico en el modelado de la dinámica evolutiva de los virus ha aumentando la incertidumbre sobre la posible existencia de las cuasiespecies en la naturaleza y sobre su participación en la evolución de los virus.

D. Reconciliación del problema biológico y motivación de este proyecto

La motivación del presente trabajo se basa en las observaciones experimentales que muestran evidencias de la presencia de cuasiespecies en algunos virus de RNA. Tales observaciones están basadas en: 1) la determinación de la tasa y frecuencia de las mutaciones, 2) la

comparación de secuencias consenso de ‘genomas’ que componen a una población y, finalmente, 3) la cuantificación de un *fitness relativo* basado en la persistencia de una población (consenso) a través del tiempo. De estos tres factores, los valores de *fitness* son los más difíciles de determinar de forma experimental, e incluso teóricamente, ya que es complicado estimar el número y tipo de mutaciones que pueden realmente afectar la tasa de sobrevivencia y reproducción de un virus; además que muchos otros factores, tales como eventos contingentes, pueden afectar el tamaño efectivo de la población (N_e), el impacto de las mutaciones en la población y, por ende, el *fitness* de un virus. Algunos de los factores que afectan el *fitness* de un organismo o virus se esquematizan en la Figura I7. Tanto para la teoría de las cuasiespecies como para la genética de poblaciones, el *fitness* es uno de los parámetros más importantes a estimar para entender la dinámica evolutiva de las poblaciones virales.

Es así como en el presente proyecto, tratamos de incorporar realismo biológico al modelado de la cuasiespecies. El objetivo principal de este proyecto es evaluar el posible comportamiento de cuasiespecies en el Virus de la Inmunodeficiencia Humana serotipo 1, HIV-1. Para ello, desarrollamos un algoritmo genético que reproduce bajos ciertos parámetros el balance mutación-selección del virus HIV-1. Enfocamos nuestros esfuerzos en incluir parámetros que son posibles de recuperar en la literatura científica a partir de los experimentos realizados en el virus HIV-1 y, en aquellos parámetros que son posibles de estimar a partir de los análisis genómicos realizados en este proyecto a través de los genomas completamente secuenciados del virus HIV-1 desde diversas partes del mundo. Además de incorporar en nuestro algoritmo genético la tasa de mutación y el tamaño efectivo de la población del virus HIV-1, también se incluyó la frecuencia del tipo de mutaciones que se han llevan a cabo durante el proceso de replicación de este virus, así como una medida del peso evolutivo que tienen estas mutaciones en la población del HIV-1. De esta forma, realizamos una estimación del *fitness mutacional*, también llamado *fitness relativo*, del virus HIV-1. Para este fin, estimamos el tipo de selección (positiva, negativa y neutra) al que están sujetos los 3,240 codones que componen el genoma completo del virus HIV-1. Se estimó la influencia de la selección a diferentes niveles del genoma codificante para proteínas (en genes completos, dominios-regiones funcionales y codones), y se compararon las predicciones con el tipo de selección que ha sido comprobado experimentalmente para ciertas posiciones y regiones del genoma del HIV-1.

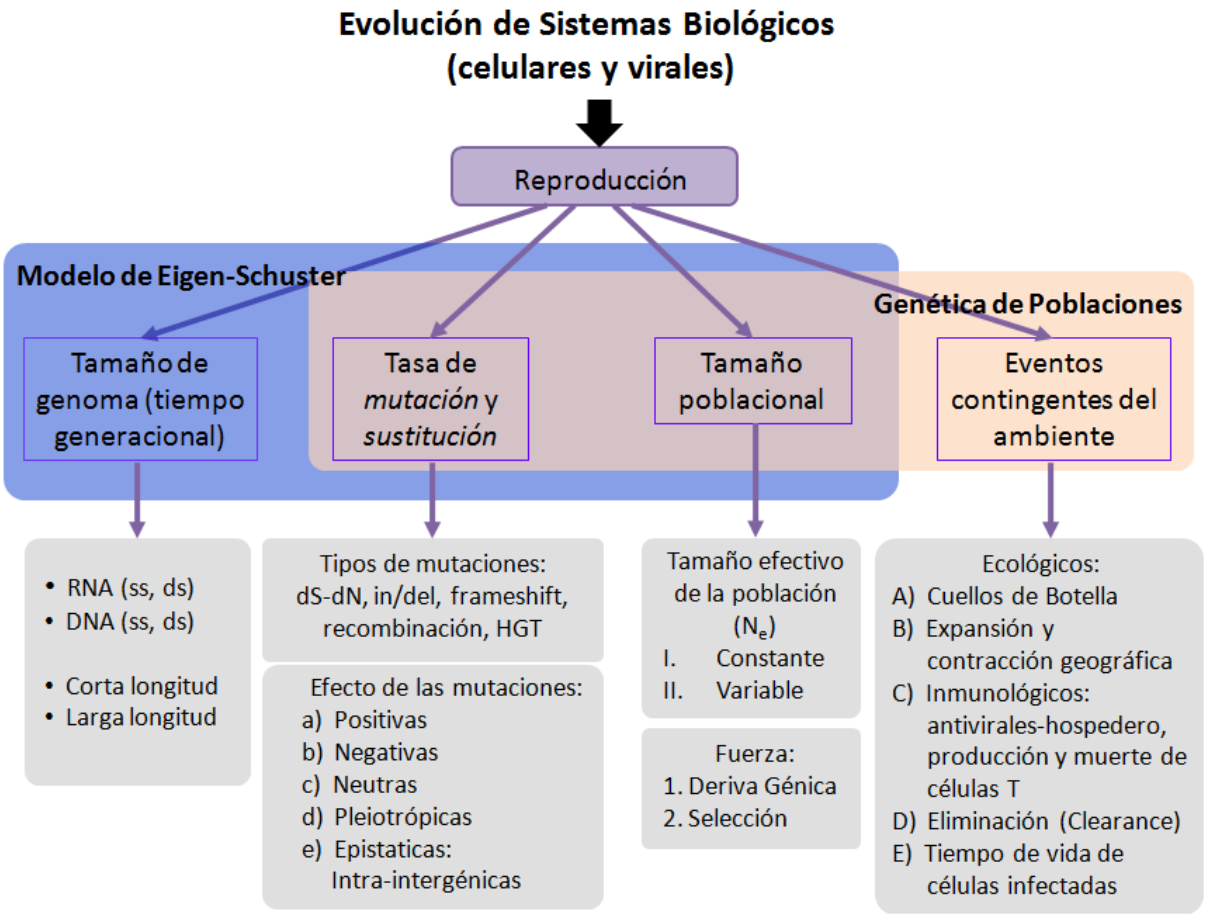


Figura I7. Factores que participan en la evolución de los sistemas biológicos. Basados en la *reproducción* (i.e., replicación) como la fase en la que se generan todas las poblaciones, y que éstas tienen diversas características que la distinguen, se puede realizar una generalización de estas características bajo las siguientes categorías: 1) tipo y tamaño del genoma, 2) tasa de mutación y sustitución, 3) tamaño poblacional, y, aquellos fenómenos que ejercen efecto sobre ellos conocidos como 4) eventos contingentes del ambiente. Así, para poder modelar la evolución de los sistemas biológicos, existen actualmente dos escuelas evolutivas que la explican empleando alguna de las categorías antes mencionadas. La primera, conocida como la teoría de las Cuasiespecies de Eigen-Schuster, emplea el *tamaño del genoma*, la *tasa mutación* y el *tamaño poblacional* como los únicos elementos participes de la evolución de los sistemas pre-celulares y virales (cuadro color azul). Se muestra una breve descripción de los elementos que conforman a estas categorías (los tres primeros cuadros color gris de izquierda a derecha). Esta corriente no toma en cuenta a los *eventos contingentes del ambiente*. El segundo, la Genética de Poblaciones, toma en cuenta la *tasa de mutación*, el *tamaño poblacional* y *los eventos contingentes del ambiente* (cuadro color almendra). Se muestra una breve descripción de los elementos que forman parte de estas categorías (los tres últimos cuadros color gris de derecha a izquierda). Esta escuela no toma en cuenta el *tipo y tamaño del genoma*. Dentro del tamaño poblacional, existe el término ***tamaño efectivo de la población*** (N_e), el cual se define como 'el número de individuos que se reproducen actualmente y que tendrán un aporte genético para la siguiente generación'. Esta cifra es difícil de calcular, pero resulta útil en el entendimiento de las poblaciones biológicas.

Nuestros resultados muestran que la mayor parte del genoma del virus HIV-1 se encuentra bajo selección purificadora. El *fitness mutacional*, la tasa de mutación y la frecuencia del tipo de mutaciones estimadas en este trabajo, nos permite obtener la dinámica de cuasiespecies (y sus características más distintivas) para el virus de HIV-1. Sin embargo, las limitaciones computacionales para incorporar el tamaño efectivo real (*Ne*) de la población en nuestro algoritmo genético restringen nuestras inferencias sobre la afirmación de que el virus HIV-1 puede estar evolucionando como cuasiespecies. Actualmente, nuestro algoritmo genético está siendo re-estructurado para realizar simulaciones en paralelo que soporten grandes tamaños poblaciones y fluctuaciones aleatorias en el tamaño de la población en diferentes tiempos generacionales de las poblaciones del virus HIV-1 simuladas computacionalmente.

No obstante, el presente trabajo muestra de forma robusta que, tal lo predicho por Eigen y Schuster (y también corroborado en el virus de Hepatitis C con genoma de RNA (Campo et al., 2008)), el genoma del virus HIV-1 está básicamente bajo selección purificadora. El número de codones neutrales es pequeño cuando se compara contra el número de codones que se encuentran bajo selección positiva y negativa. Mientras que la selección negativa domina la evolución de las regiones estructurales del virus (que se encuentran en la primera parte del genoma), la selección positiva actúa principalmente sobre la segunda parte de genoma y está restringida principalmente a aquellas regiones funcionales relacionadas con la interacción al hospedero. La presencia de la selección purificadora en genomas pequeños (con un “bajo” contenido informacional) y que están sujetos a altas tasas de mutación, podría así conducir al comportamiento de cuasiespecies de forma determinista. Esperamos que este trabajo contribuya a tener una mejor perspectiva sobre la dinámica evolutiva de los virus de RNA.

A continuación, describiremos brevemente la biología de nuestro modelo de estudio, para así proseguir con las hipótesis, su tratamiento metodológico, los resultados, su discusión y las conclusiones que hemos mencionado anteriormente.

E. El Virus de la Inmunodeficiencia Humana (HIV): un modelo biológico y médico

El *Virus de la Inmunodeficiencia Humana* tipo 1 (HIV-1) es un retrovirus que forma parte del genero Lentivirus y es causante de una de las enfermedades humanas más complejas, el Síndrome de la Inmunodeficiencia Humana (SIDA). HIV-1 se divide en 4 grupos: M, N, O, P. El grupo M es el más predominante y se divide en 11 subtipos, de la letra A a la K. El tipo 2 sólo se ha identificado en África. Se estima que 36.7 millones de personas en el mundo tienen SIDA (2010) (Organización Mundial de la Salud: http://www.who.int/topics/hiv_aids/), y se han destinado alrededor de 8,297 millones de dólares (2005) en investigaciones para encontrar una cura (UNAIDS: http://www.unaids.org/). Debido a ello, el virus del HIV-1 se ha convertido en uno de los modelos virales más ampliamente estudiado, por lo que se tiene una gran información en las bases de datos sobre su biología y patología. El primer genoma del virus del HIV-1 se secuenció en Estados Unidos (1985) (por diversas entidades educativas e institutos de investigación) bajo el nombre de *HXB2* y su identificador en el NCBI (del inglés, Centro Nacional para la Información sobre Biotecnología) es *K03455*.

El genoma del virus de HIV-1 es de RNA de cadena sencilla con sentido positivo (ssRNA+), y su replicación es bastante peculiar, dado que requiere de un intermediario de DNA que se copia a partir del RNA genómico de una secuencia molde. Este proceso se denomina *retrotranscripción* y se lleva a cabo a partir de una polimerasa denominada Reverso Transcriptasa. La retrotranscripasa es exclusivamente viral, lo que significa que no se encuentra en la célula hospedera y los virus tienen que empaquetarla en sus cápsides para asegurar la replicación viral en el próximo ciclo de infección. El genoma del virus de HIV-1 está dividido prácticamente en dos grandes regiones: la primera contiene tres genes principales o estructurales (*gag*, *pol* y *env*) y, la segunda región contiene seis genes auxiliares (dos reguladoras: *tat y rev,* y cuatro accesorios*: nef, vif, vpr* y *vpu*). Esta organización está rodeada en sus extremos por 2 regiones *no codificantes*: los LTRs 5’ y 3’ (repeticiones terminales largas, del inglés *Long Terminal Repeats*) [ver Figura I8]. Los ‘genes principales’ participan básicamente en la replicación y estructura del hospedero, mientras que los ‘auxiliares’ tienen múltiples funciones, pero la gran mayoría involucra la interacción con el hospedero (Frankel and Young, 1998, Carter and Saunder, 2007).

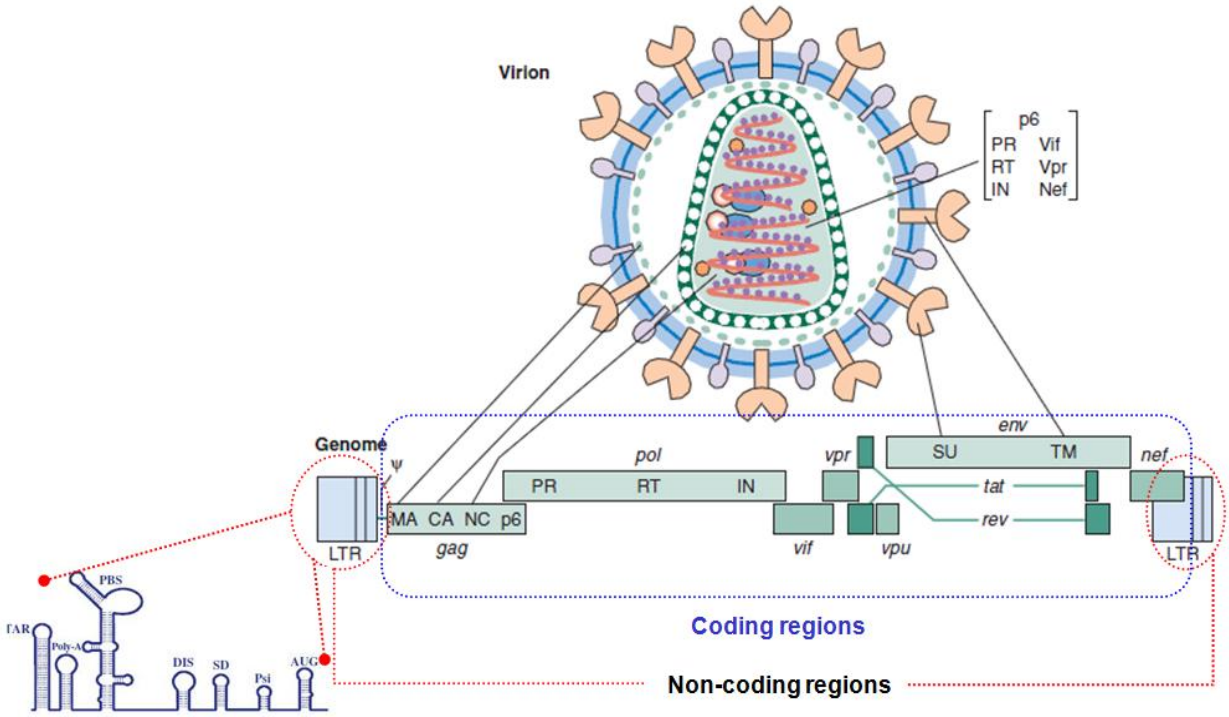


Figura I8. Organización genómica del virus HIV-1 cepa HXB2. En la parte superior se muestra una representación gráfica del virus HIV-1, desde el exterior al interior del virión: a) envoltura con compuesta de lípidos-proteínas-carbohidratos que es codificado por el gen Env, b) la cápside es conformada por proteínas, que son codificadas por el gen Gag (p24) , c) el genoma de RNA está estructurado por regiones codificantes (región delimitada por la línea azul punteada) y por regiones no codificantes (región delimitada por la línea roja punteada). Las regiones codificantes están conformadas por *genes estructurales (o principales)*: gag, pol y env; *genes auxiliares*: rev, tat (proteínas regulatorias) y vif, vpr, vpu y nef (proteínas accesorias). Las *regiones no codificantes* están conformadas por: R, U3 y U5. Los dominios presentes en 5' LTR: TAR, Poly-A, PBS, DIS, SD, Psi, AUG (adaptado de Wagner E. et al. 2008).

La replicación del genoma de HIV-1 es un proceso conformado por varios pasos [ver Figura 12], en el que intervienen diferentes elementos [ver Figura I9]. A continuación, se describen brevemente la función de cada uno de los elementos genómicos que participan en el proceso de replicación de este virus.

Repeticiones Terminales Largas (LTRs). Uno de los procesos clave es la expresión del genoma viral dentro del genoma del hospedero, el cual es regulado por los *LTRs*. Estos elementos son considerados como reguladores y sirven como puntos de convergencia de factores de transcripción y elementos de la maquinaria transcripcional del hospedero. Así, los LTRs interactúan con la maquinaria básica de transcripción de la célula hospedera. Para poder producir una infección progresiva, se requiere de una expresión suficiente de los genes de HIV-1. El principal impacto que tienen los LTRs se ve reflejado en el éxito completo de la

infección: producción, liberación o la latencia del virus (Frankel and Young, 1998).

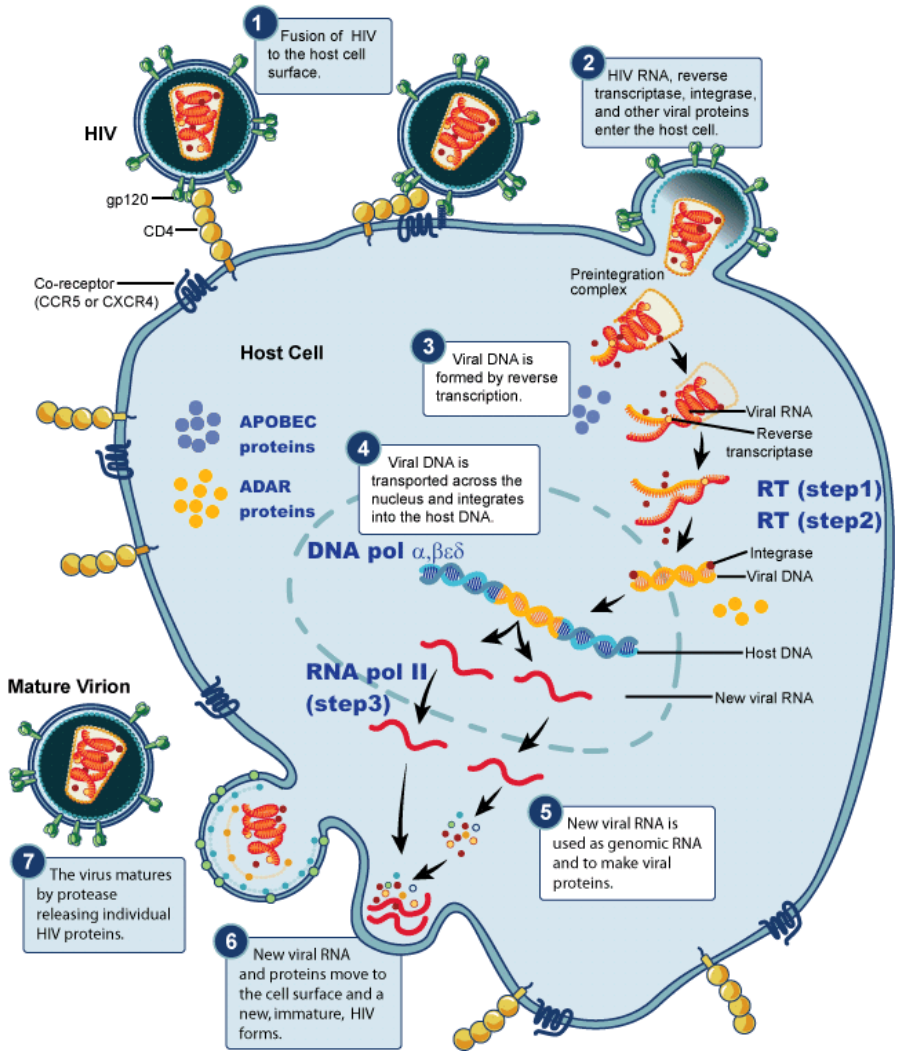


Figura I9. Ciclo de replicación y eventos de origen de la mutación en HIV-1. *La variación genética de las poblaciones de HIV-1* es el resultado de la incorporación de numerosas mutaciones en el proceso mismo de replicación del genoma viral. Dicha mutaciones se puede llevar a cabo durante 3 distintas etapas de polimerización del retrovirus dentro de la célula. Estas etapas se presentan a continuación a partir de la integración del virus a la célula hospedera: 1) en la síntesis de un ssDNA (-) a partir del genoma viral de RNA por medio de la RT, 2) en la síntesis de un ssDNA (+) a partir de la ssDNA (-) por medio de la RT. En este paso se considera la formación de una dsDNA (- y +), 3) cuando se replica el genoma de la célula infectada, y el genoma viral integrado en el genoma hospedero (provirus), a través de una DNA polimerasa dependiente de DNA (esta es opcional); y finalmente 4) cuando se realiza la síntesis del genoma viral de RNA (+) a partir del DNA integrado como cadena molde por medio de una RNA polimerasa tipo II. Asimismo, las moléculas de APOBEC y ADAR pueden participar en el proceso de mutación; mostradas en azul y amarillo, respectivamente. Figura tomada y modificada de: www.niaid.nih.gov/.

Genes *gag, pol y env*. El gen *gag* codifica para proteínas de la cápside. En general, la poliproteína Gag (p55) constituye un componente viral que es estrictamente requerido para el ensamble y la "gemación" por membrana de la partícula viral. Gag está conformada por otras proteínas que tienen varias funciones, por ejemplo, *MA* (matriz, p17) modula el blanco de la membrana plasmática, *CA* (cápside, p24) regula el ensamble de *gag* y forma el núcleo principal del virus, y *NC* (nucleocápside, p15) regula el empaquetamiento y condensación del genoma viral, así como también dirige la incorporación de la proteína reguladora Vpr en la 'gemación' de viriones y maneja la efectividad de la gemación (Frankel and Young, 1998, Wagner et al., 2008). El gen *pol* codifica para otras proteínas de gran importancia: *proteasa* (PR), *reverso transcriptasa* (RT) *e integrasa* (IN). La RT participa en la reverso-transcripción del genoma viral de RNA a DNA, mientras que PR e IN se encargan de integrar el nuevo genoma de DNA viral al genoma del hospedero. En conjunto, a estas tres proteínas se les conoce como el complejo de pre-integración (PIC).

El gen *env* codifica para 2 proteínas virales, gp120 y gp41 (SU, proteína de superficie, y TM, proteína transmembranal, respectivamente, en Figura I8), aunque éstas se sintetizan como una sola proteína inicialmente (gp160). La entrada viral a la célula hospedera es iniciada por la interacción entre HIV-1 y los receptores celulares de la superficie, *e.g.* CD4. Dicha interacción es mediada a través de dominios específicos de gp120 y, además, de la participación de otras proteínas receptoras transmembranales de la célula hospedera que funcionan como co-receptores virales, estos son CXCR4/fusina y CCR5; existen otros co-receptores, pero en menor cantidad (CCR3, CCR2b, Bonzo/STRL33 y BOB/GPR15). Por otro lado, la función de gp41 es mediar la fusión de membranas seguido de la unión de receptores (revisado en Frankel y Young, 1998).

Genes *vif* y *vpr*. Son catalogados como dos genes accesorios. El primero, conocido como *factor viral de infectividad*, promueve la infectividad, pero no la producción de partículas virales. Aunque el mecanismo exacto de cómo realiza este proceso no se conoce del todo, actualmente se sabe que cuando *vif* presenta mutaciones significativamente deletéreas (e.g., posición 144 de Serina), el fenotipo de estos mutantes genera una reducida síntesis de DNA y se producen intermediarios de la replicación altamente inestables. Estos resultados sugieren que *vif* participa antes o después de la síntesis de DNA (Mahy and Van Regenmortel, 2010, Frankel and Young, 1998). Actualmente, se sabe que esta región funcional tiene un papel

imprescindible en el éxito infectivo de HIV-1: el bloque de las proteínas APOBEC (del inglés, *apolipoproten B mRNA editing enzyme, catalytic polypeptide-like*). Dentro de la respuesta celular para combatir a HIV-1, se encuentra la participación de esta proteína, la cual es generada en la célula hospedera y es encapsidada en el proceso de ensamblaje de las partículas virales, de tal forma que su acción se activa cuando la reverso transcriptasa inicia su función. APOBEC es capaz de reconocer y editar nucleótidos de las moléculas de DNA y RNA, de cadena doble o sencilla. Las modificaciones más comunes son de Guanina a Adenina (G→A) o de Citocina a Uracilo (C→U), lo que provoca la alteración de la capacidad codificante de los mRNA que, en términos generales, se traduce en la producción de genomas mutantes virales con un número considerable de errores generados durante la reverso transcripción.

A la región *vpr* se le han adjudicado diversas funciones sin conocer específicamente a fondo su mecanismo de acción. Entre estas incluyen ataques contra la importación nuclear de complejos de integración, detección del crecimiento celular, la transactivación de genes celulares y la inducción de diferenciación celular. *Vpr* también participa en el transporte del completo de pre-interacción (PIC) y en la localización nuclear en células no divisibles (*i.e.* macrófagos). Aparentemente, el gen *vpr* se originó en HIV-2 y en SIV (Simian Inmunodeficiency Virus) a partir de un evento de una duplicación genética, posiblemente por recombinación (Peeters and Courgnaud, 2002). Además, se ha reportado que esta región viral puede influenciar la tasa de mutación (Mansky, 1996b, Mansky et al., 2000).

Regiones reguladores *Tat* y *Rev*. Estas regiones participan activamente en el proceso de replicación de HIV-1. La primera es responsable de incrementar la tasa de transcripción del genoma viral (~100 veces más) empleando a la RNA polimerasa II de su hospedero; ello debido a su capacidad de interactuar con factores transcripcionales celulares y de formar un complejo transcripcional. En ausencia de *tat*, las polimerasas no pueden transcribir más allá de unos cientos de nucleótidos. Sin embargo, en la actualidad no se sabe cómo *tat* puede hacer que las polimerasas lleguen a transcribir de forma progresiva todo el genoma de HIV-1 (~9.7kb) sin detenerse. Por su parte, *rev* actúa uniéndose a un elemento estructural de HIV-1 localizado en la proteína Env, el RRE (Elemento Sensible a Rev). Además *rev* es responsable de exportar, estabilizar y utilizar los mRNAs semi-procesados y sin procesar (i.e., secuencias

de trascritos cuyos nucleótidos conservan parcial o totalmente intrones) que contengan RRE (Frankel and Young, 1998).

Regiones accesorias *Vpu* y *Nef*. *vpu* promueve la degradación de moléculas CD4, las cuales atrapan a los productos de la glicoproteína Env (gp160), las proteínas gp120 y gp41, lo que facilita el transporte de estas proteínas hacia la superficie de la membrana para su ensamble en la partícula viral. El papel de *nef* radica en mantener altas cargas virales del virus reduciendo, al igual que *vpu*, la cantidad de moléculas de CD4 presentes en la célula. Además, modifica posiblemente la ruta de señalización de las células T $CD4^+$ y promueve la liberación de la partícula viral. La generación de virus con *nef* mutantes reflejan un decaimiento drástico en la síntesis de DNA después de la infección, lo que incita a pensar que esta región funcional participa en el ensamble, maduración y entrada del virus (68 en Frankel y Young, 1998).

HIPÓTESIS DE ESTE PROYECTO

Dado que el virus de la inmunodeficiencia humana tipo 1 (HIV-1) posee un genoma de RNA y que este virus posee aparentemente varias de las características descritas en el modelo original de Eigen y Schuster, se espera que este virus se comporte y evolucione como cuasiespecies. De esta forma, si HIV-1 se comporta como cuasiespecies, las características más distintivas de las cuasiespecies (tal como el umbral de error, la catástrofe de error, y la catástrofe de extinción) deberían ser observadas aún cuando se incorpore más realismo "biológico" en el modelado de su dinámica evolutiva.

OBJETIVOS DE ESTE PROYECTO

El principal objetivo de este proyecto se enfoca en incorporar *realismo biológico* al modelo original de las cuasiespecies de Eigen-Schuster a fin de evaluar el comportamiento de cuasiespecies en uno de los genomas virales modelo de RNA, el virus HIV-1. Para lograr tal meta, los objetivos específicos de esta tesis se describen a continuación:

1. Realizar un algoritmo genético computacional que permita simular la dinámica de las cuasiespecies según lo propuesto por Eigen y Schuster, pero que también incluya más parámetros y mecanismos que reflejen la biología de las poblaciones naturales de HIV-1.

2. Estimar, a partir de la literatura, la tasa de mutación por genoma por ciclo de replicación, la frecuencia del tipo de mutaciones, los tiempos generacionales y el tamaño efectivo de la población del HIV-1. Estos valores son indispensables para incorporar más realismo a las simulaciones realizadas con el algoritmo genético.

3. Calcular un "fitness mutacional" para todos los codones que conforman el genoma de HIV-1. Realizar estimaciones genómicas del tipo de selección (positiva, negativa o neutra) al que están sujetos los genes, los dominios funcionales del genoma de HIV-1 y sus respectivos codones. Comparar las estimaciones genómicas con las experimentales reportadas en la literatura. Este cálculo permite hacer una estimación más real del peso evolutivo que tienen las mutaciones en el genoma de HIV-1 e incorporarlo a las simulaciones computacionales.

4. Identificar con métodos estadísticos las propiedades de las cuasiespecies en los resultados obtenidos: secuencia consenso, el umbral de error, la catástrofe de error y la catástrofe de extinción.

MÉTODOS

I. Algoritmo genético que simula la dinámica evolutiva de las cuasiespecies

Se realizó un algoritmo genético en lenguaje de programación PERL (*quasispecies.pl)* que simula el modelo de cuasiespecies propuesto por Eigen y Schuster a través de los siguientes pasos [ver Figura M1 y ver Ecuación S1.1 en Texto S1].

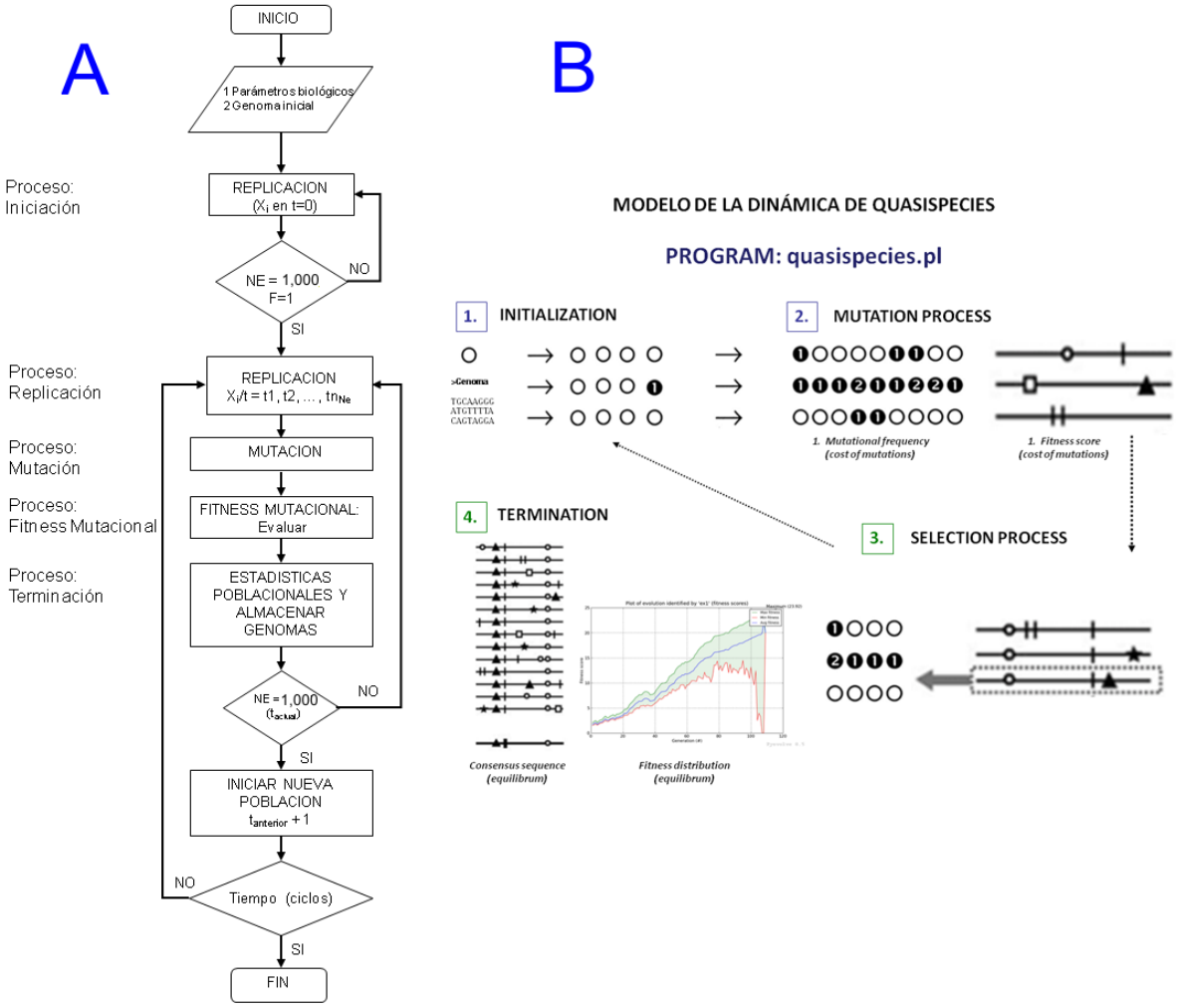


Figura M1. Diagrama de flujo y modelo de la dinámica del programa *quasispecies.pl*. A. El modelo de Eigen-Schuster está constituido por tres diferentes elementos: replicación, mutación y selección, mostrando los estados de la secuencias a través del tiempo. El programa incorpora esos elementos en un algoritmo genético y modela su dinámica en las poblaciones virales a través del tiempo a partir de los datos estimados para el virus HIV-1. En ***A***: se muestra el diagrama de flujo del programa *quasispecies.pl*. En ***B****:* representación del programa como un algoritmo genético que consiste de 4 pasos generales: 1) *iniciación*, 2) *mutación*, 3) el *proceso de selección* (evaluación) y 4) *terminación*. Cada uno de estos pasos se detalla en los métodos.

Quasispecies.pl: tasa de mutación y tamaño efectivo de la población para HIV-1

Para poder adicionar un realismo biológico a la teoría de las cuasiespecies e ilustrar el efecto de sus variables dentro de nuestras simulaciones, se realizó una búsqueda de datos con soporte experimental a través de la literatura para el virus de HIV-1. La búsqueda se enfocó en la estimación cuantitativa sobre la replicación del virus HIV-1, de los siguientes parámetros: el número de generaciones producidas al año, el tamaño efectivo de la población (N_e), el número de mutaciones por ciclo de replicación, la frecuencia y el tipo de mutaciones que se presentan en un ciclo de replicación. Los datos obtenidos y seleccionados se discuten ampliamente en la sección de resultados y se muestran en la Tabla R1. De este modo, nuestras simulaciones se realizaron con diferentes tasas de mutación (de 1 a 4 mutaciones por genoma por ciclo de replicación), con una N_e de 1,000 viriones, que corresponde entre el 1% y el 10% de la N_e real estimada para HIV-1. Las simulaciones se realizaron a lo largo de 1,000 y 2,000 generaciones, que correspondería aproximadamente a la mitad de los ciclos generacionales derivados desde la secuenciación del primer genoma completo del virus HIV-1 en 1981 a la fecha.

Quasispecies.pl: proceso de iniciación

Se parte de una población inicial (x_0) de 1,000 partículas virales de HIV-1 (N_e=1%), según los datos compilados en la literatura. La población inicial se conforma, así, del correspondiente número de copias de la secuencia modelo del genoma de HIV-1, la cepa HXB2, depositada con el número de acceso K03455 en el NCBI (http://www.ncbi.nlm.nih.gov/genomes/) y en Los Álamos (http://www.hiv.lanl.gov/). Cada uno de los genomas mantiene una longitud (*L)* constante de 9719 nucleótidos (nt), el tamaño de la secuencia parental inicial. Los tamaños de las poblaciones y la longitud de la secuencia permanecen constantes durante las simulaciones, aunque estos parámetros pueden ser cambiados fácilmente en el programa. En un ciclo de replicación, cada individuo en esta población da lugar a una nueva población *(x_i+1)* en una fracción correspondiente al *fitness* (respecto a la población total) heredado del ciclo anterior. Para iniciar la dinámica, todos los individuos de la población inicial (x_i) tienen un *fitness inicial* de *1* ($F_{xi/t0}=1$), por lo que cada individuo de la población inicial dará lugar a una progenie del mismo tamaño que la población inicial. Según la influencia de las mutaciones en el *fitness* del virus a partir del primer ciclo de replicación, el *fitness* cambiará en cada ciclo de replicación y para cada individuo de la población viral.

Quasispecies.pl: proceso de mutación

Cada individuo de la población y sus progenies replican a una tasa de mutación constante tal como y fue propuesto por Eigen y Schuster ($\mu = n$, donde n es el número de mutaciones). Se corrieron simulaciones con una sola mutación por genoma por ciclo de replicación, según los datos experimentales compilados para HIV-1 (ver Resultados). En simulaciones posteriores, se introdujeron dos, tres y hasta cuatro mutaciones por genoma por ciclo de replicación. Mientras que la posición donde se llevan a cabo las mutaciones en el genoma son completamente azarosas, la frecuencia del tipo de mutaciones no es azarosa, éstas fueron calculadas específicamente para purinas (*A* y *G*), pirimidinas (*C* y *T*), sustituciones sinónimas y no sinónimas, inserciones y eliminaciones, a partir de las frecuencias del tipo de mutaciones reportadas para HIV-1 en previos trabajos experimentales [ver Resultados]. El proceso de mutación sólo tomó en cuenta la simulación de mutaciones puntuales, excluyendo inserciones y eliminaciones.

A cada una de las mutaciones realizadas en el proceso anterior se les asigna un valor de *fitness mutacional o relativo*, es decir, un peso positivo, neutro o negativo sobre las futuras generaciones. Dicho *fitness relativo* fue previamente estimado (como se describirá más adelante) para todos los codones del genoma completo de HIV-1, e incorporado al programa en forma de una tabla que contiene las posiciones de los codones y la influencia del tipo de selección en forma de *scores,* el cual se explica a continuación. Cada codón del genoma tiene asignado un valor de selección calculado previamente (que se explicará con detalle más adelante), este valor puede ser negativo, positivo o $\cong 0$ (que entonces contará como un valor neutral). Según la frecuencia de tales valores, se estima un *fitness score* que determina qué valores influenciarán de forma negativa, positiva o neutra el fitness del individuo en el próximo ciclo de replicación (ver Tabla R6 y R7 en resultados). Dado que las mutaciones no son únicas en el genoma ni tampoco excluyentes de las que ocurren en cada ciclo de replicación, el *fitness relativo* previo y el *fitness score* nuevo se suman y promedian en cada ciclo de replicación para cada individuo, de forma que el *fitness mutacional o relativo* de cada individuo de una nueva población mantiene "una memoria" mutacional a lo largo de la evolución viral. Lo mismo aplica cuando se introducen más de una mutación por genoma por ciclo de replicación, se suman y promedian los *fitness score* (de cada una de las mutaciones)

del presente ciclo de replicación y se suman y promedian con el *fitness relativo* del ciclo anterior previo, a fin de contemplar posibles fenómenos de epístasis y mutaciones deletéreas acumulativas.

Quasispecies.pl: proceso de selección

Una población con *n* individuos y cada uno de ellos generando a su vez una progenie de *m* individuos, se organiza en una matriz bidimensional de tamaño *n* x *m* en cada generación. Por ejemplo, si las simulaciones empiezan con 1,000 individuos y cada individuo puede al menos generar el mismo número de individuos (o el doble si el *fitness relativo* es cercano a 2), nuestro programa genera cálculos para 1,000,000 individuos. De esta matriz bidimensional se obtienen aleatoriamente los individuos de la nueva población viral. La selección se realiza sin reemplazo y hasta alcanzar el número corresponde al tamaño efectivo de la población que estamos empleando. Cada uno de los individuos seleccionados aleatoriamente posee el genoma con las mutaciones almacenadas y los *fitness* correspondientes asignados en cada ciclo de replicación.

Es importante mencionar aquí, que se redujo la memoria RAM usada para almacenar dicha matriz dimensional al generar archivos de texto con las progenies de cada individuo de la población. Sin embargo, se requieren de más de 1.5 Giga bits (Gb) en memoria RAM y memoria física de al menos 380 Gb para poder realizar y almacenar cálculos durante 13 semanas con un tamaño inicial de población de 1,000 individuos a lo largo de 100 generaciones. De forma que el tiempo de cálculo del programa es dependiente del tamaño efectivo de la población, del número de mutaciones calculadas por ciclo de replicación por genoma y del número de generaciones que generan las simulaciones [ver sección *Quasispecies.pl: tiempo de cálculo y rendimiento relativo computacional* más adelante].

Quasispecies.pl: proceso de terminación

Después del proceso de selección y antes de empezar el próximo ciclo generacional, se realizan ciertas mediciones estadísticas, tales como la distribución del número y tipo de mutaciones a lo largo del genoma de HIV-1, la media, el máximo, el mínimo y el promedio del *fitness* mutacional para toda la población. La dinámica evolutiva se detiene cuando se alcanzan

los 1,000 ó 2,000 ciclos generacionales (ver *iniciación*). Consideramos que este número de generaciones es suficiente para que la población de HIV-1 alcance el equilibrio como cuasiespecies, en caso de que éste exista. Para lograr identificar este estado, se extrajeron cada 100 generaciones, los datos estadísticos descritos previamente y se calcularon las secuencias consenso de los genomas correspondientes a esas generaciones. Estos datos nos permiten identificar la presencia o ausencia del umbral de error, la catástrofe de error, y la catástrofe de extinción de la población del HIV-1.

Quasispecies.pl: tiempo de cálculo y rendimiento relativo computacional

El trabajo realizado por un programa construido mediante una serie de órdenes que requieren velocidad (procesador), espacio (memoria RAM) y tiempo (ciclos/segundo) para su ejecución, es diferente dependiendo del equipo en el que se trabaje. Una forma de estimar la potencia que requiere la implementación del programa es el *tiempo*; el tiempo es la medida de rendimiento de una máquina realizando un trabajo determinado. Aquí, nosotros calculamos el *tiempo de cálculo* y el *rendimiento relativo computacional* del programa *quasispecies.pl* basados en tres estimaciones sencillas, esto con la finalidad de obtener una implementación computacional óptima del programa en determinados equipos de computo al realizar varias ejecuciones al mismo tiempo. Así, el 'tiempo de cálculo' y el 'rendimiento relativo computacional' se estimaron de forma individual y comparativa, respectivamente, en dos ordenadores de diferentes características: un clúster y una PC portátil [ver Texto S2 en material suplementario]. El rendimiento relativo en el clúster fue 1.8 veces (tiempo promedio de CPU: 170,707 segundos, 99% y 3.8×10^{14} ciclos) que en la PC portátil (tiempo de CPU: 239,983 segundos, 114% y 4.8×10^{14} ciclos), es decir, el clúster cumple con un número doble de tareas que la PC portátil en menos tiempo, por lo que la implementación del programa puede llevarse a cabo más rápidamente con un equipo similar al clúster.

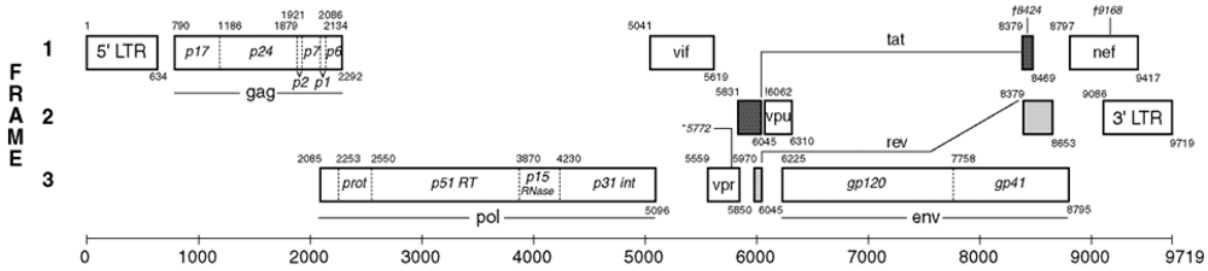
En perspectiva, el 'tiempo de CPU' y el 'rendimiento relativo' empleado por el programa *quasispecies.pl* es característico del tipo de ordenador que se emplee e, igualmente, del tamaño efectivo de la población, del número de mutaciones calculadas por ciclo de replicación por genoma y del número de generaciones que realizan las simulaciones [ver Tabla S2 en Texto S2 suplementario]. Actualmente, el rendimiento óptimo del programa *quasispecies.pl* requiere de

un nivel medio de computo (~1.5Gb RAM por simulación y con 2 procesadores como mínimo) y un espacio grande en el disco (~380Gb), esto debido a los datos numéricos biológicos usados para esta tesis (N_e = 1,000 viriones, generaciones = 1,000 ciclos y μ=1, 2, 3 y 4 mutaciones por ciclo de replicación por genoma). Considerando 4 simulaciones en sesión paralela con un diferente número de mutaciones de forma independiente, cada simulación consume aproximadamente 1.6Gb en RAM (~7Gb en RAM). Así, el programa no requiere mucha potencia del ordenador si se usa una N_e reducida ($N_e \ll 1000$, e.g. N_e=200), por lo que se puede emplear un equipo con un procesador de doble núcleo con 2Gb de RAM, ocupando entre 0.25—0.35Gb de RAM para cada simulación, y un espacio mínimo en disco de 10Gb para almacenar los cálculos. Emplear una máquina con características de bajo nivel (memoria RAM$\ll$4Gb y 1 procesador con un 1 núcleo, o modelos inferiores), con las cifras que nosotros empleamos en este estudio, representaría un gran problema en el tiempo de cálculo para finalizar el programa y la pronta obtención de datos.

II. Cálculo del fitness mutacional para el genoma de HIV-1

Para poder evaluar el efecto de las mutaciones producidas en el genoma de HIV-1 sobre el *fitness* de las nuevas poblaciones virales, se estimó para cada codón a lo largo del genoma viral de HIV-1 un *valor de selección* (*fitness*) positivo, negativo o neutro. Bajo esta perspectiva, las mutaciones neutrales no afectarán el fitness de los nuevos individuos, pero aquellas que sean positivas o negativas incrementarán o reducirán el *fitness* del individuo, respectivamente. Asimismo, y con la finalidad de detectar si varias regiones del genoma pueden estar sujetas a diferentes presiones de selección, se calculó el *fitness* de todo el genoma viral de HIV-1 en las *regiones codificantes* (11 genes y 37 regiones funcionales). Ver Figura M2 y R1 para ver la estructura y localización de las regiones funcionales dentro del genoma. Cada una de estas regiones funcionales presenta dominios, motivos o sitios de interacción con proteínas y con estructuras de RNAs funcionales, entre otras regiones más, las cuales pueden indicar los efectos que ejercen la *mutación* y la *selección* sobre la dinámica evolutiva viral. De esta forma, se realizó un cálculo *a priori* para dividir a cada gen por ventanas, $Win_{total} = x_{nt} / y_{sfr}$, con la finalidad de obtener sus regiones funcionales (ver Figura M2 y detalles en la serie

de Figuras S2 en material suplementario). El cálculo consiste en dividir el número total de nucleótidos de un gen (x_{nt}), entre el en número total de "ventanas mínimas" que realizan la cobertura de los dominios y motivos de proteínas de forma completa en tal gen (y_{sfr}).



	Sites	Location	Number of sub-functional regions
1)	*Region 5' LTR*	*(1 – 634)*	*0*
2)	Gag	(790 – 2292)	12 (gag & pol)
3)	Pol	(2085 – 5096)	
4)	Vif	(5041 – 5619)	9
5)	Vpr	(5559 – 5850)	1+
6)	Tat	(1^{*}= 5831 – 6045 & 2^{*}= 8379 – 8469)	7
7)	Rev	(1^{*}= 5970 – 6045 & 2^{*}= 8379 – 8653)	4
8)	Vpu	(6062 – 6310)	3
9)	Env	(6225 – 8795)	16
10)	Nef	(8797 – 9417)	12
11)	*Region 3' LTR*	*(9086 – 9719)*	*0*

Figura M2. Organización genómica del virus de HIV-1: estructura y localización. *Panel superior*: se muestra el genoma de HIV-1 en sus diferentes regiones funcionales y los 3 marcos de lectura en la que se transcriben sus genes (estructura genómica). *Panel inferior:* se muestra el número de regiones funcionales empleados en este estudio, su *localización* y el número de *sub-regiones funcionales* presentes en el genoma de HIV-1, basado en la estructura del genoma de referencia HXB2 (1983) y en la metodología empleada para su localización (ver métodos y Figuras S2 en material suplementario). Se ha considerando a la *región solapada gag-pol* como otra región funcional adicional no incluida en el cuadro inferior, por lo que serían 12 regiones funcionales (*sites*) empleadas en este estudio. Las posiciones de numeración para cada gen en relación con el genoma de referencia HIV-1 HXB2, se basaron en el *Compendio de bases de datos Los Alamos*: Human Retroviruses and AIDS, 1998 (Korber B, 1998). Figura tomada y adaptada de *http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html*.

A. Compilación de genomas completos del virus HIV-1

De las bases de datos NCBI (http://www.ncbi.nlm.nih.gov/) y Los Alamos (http://www.hiv.lanl.gov/), se colectaron 2,338 genomas completos del virus de inmunodeficiencia humana subtipo 1 (HIV-1) distribuidos en 83 países. La colección de los genomas se realizó a través de los siguientes filtros: *a)* genomas completos, *b)* sólo genomas

del tipo 1, *c)* genomas secuenciados desde el año 1982 hasta 2009, *d)* sin restricciones geográficas y *e)* sin presencia de recombinantes. Se ha estimado que la presencia de genomas recombinantes interfiere con la identificación de selección en las secuencias cuando los métodos de selección están basados en modelos de substitución de codones y filogenias, tal como los de maximum-likelihood que no toman en cuenta los efectos de la recombinación (Anisimova et al., 2003). De esta forma, eliminamos la presencia de genomas recombinantes (aunque ello implique una pérdida esencial del realismo biológico), a fin de evitar una estimación errónea de la selección. Se ejecutó una serie alineamientos para esta colección de genomas usado el programa *Muscle* v3.6 (Edgar, 2004). Posteriormente, se realizó un filtro de redundancia al 99% de identidad sobre la colección de genomas a través del programa *CD-HIT* v4.5.3 (Weizhoung and Godzik, 2006). Adicionalmente, se removieron los genomas completos que presentaron *gaps* dentro del genoma (considerando *1* gap como mínimo) y aquellos genomas en que las regiones funcionales estuvieron incompletas (80% ó 70% de la secuencia), como el caso de las regiones donde se encuentran los *LTRs* y genes truncados. Asimismo, se eliminaron aquellos genomas que presentaron errores de secuenciación (i.e., bases degeneradas) indicados a través de la presencia de la siguiente simbología en el genoma N, K, R, S, Q. Una lista oficial sobre estas *bases degeneradas* puede visualizarse en el sitio de Internet del Comité de Nomenclatura de la Unión Internacional de Bioquímica http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html#tab1 (NC-IUB por sus siglas en inglés). Como resultado de estos filtros, una colección final de 124 genomas completos y no redundantes de HIV-1 con una distribución en 35 países fue obtenida para este proyecto [ver Tabla M1]. Todos los genomas pertenecen al grupo M de HIV-1 (subtipos A, B, C, D y F). Se realizaron alineamientos de los 124 genomas completos a través del programa *clustalW* v2.0 (Thompson et al., 1994), así como la edición manual de los mismos a través del programa *BioEdit* v7.0.9.0 © (Hall, 1999). Las versión gráfica de los alineamientos se obtuvo con el programa *Geneious* v4.8.5 © (Drummund et al., 2011), los cuales se pueden observar en las Figuras suplementarias S1-A—S1-L del material suplementario.

Tabla M1. Colección de genomas del HIV-1. Se presentan 124 genomas obtenidos después de la aplicación de filtros descritos. Se muestran 5 continentes y sus países representantes. En el extremo derecho, se coloca el número total de genoma que representan a ese continente. Adicionalmente, se usó un genoma viral reportado por Watts y colaboradores (2005), en el cual se realizó un análisis de la estructura secundaria viral identificando varias *subregiones funcionales* a través de dos métodos.

Genomas Viral Analizados (1982-2009)

Continente	País	Numero of genomas	Total
América	Canadá	6	49
	Estados Unidos	23	
	Brasil	17	
	Venezuela	1	
	Argentina	2	
Europa	España	1	17
	Francia	3	
	Holanda	1	
	Grecia	7	
	Suecia	2	
	Estonia	1	
	Ucrania	1	
	Rusia	1	
Asia	China	5	15
	Macao	1	
	Japón	4	
	Vietnam	1	
	Corea del Sur	1	
	Afganistán	2	
	India	1	
África	Gabonesa (West C. África)	1	40
	Congo	4	
	Costa de Marfil (West África)	2	
	Camerún	1	
	Nigeria (East África)	1	
	Ghana (East África)	1	
	Botsuana (South África)	1	
	Kenia	12	
	Uganda	6	
	Tanzania	5	
	Senegal	1	
	Durban	4	
	Ruanda (Central África)	1	
Oceanía	Australia		
Basado en (*)	Cepa de HIV-1 *NL4-3*	1	1
Número total de genomas virales [♣]			*124*

* WATTS et al. 2009. *Nature,* 460**,** 711-716.

♣ Ver métodos.

B. Evaluación de la selección a nivel codificante por dos diferentes métodos predictivos

Para poder tener una efectiva y una más precisa valoración de *la selección* a lo largo del genoma viral, se realizó dicha evaluación usando dos métodos robustos e independientes: *la prueba de neutralidad de Tajima* y *la estimación de sustituciones sinónimas (dS) y no sinónimas (dN)*. La prueba de Tajima, se realizó en dos formas diferentes: (1) evaluación por regiones funcionales y (2) por codones. Con el segundo método sólo se evaluaron codones. En todas las evaluaciones, los alineamientos de los genomas completos fueron editados-cortados en sus diferentes genes y regiones funcionales con base en la estructura genómica del genoma de referencia HXB2 para HIV-1. Para el análisis por codones, los alineamientos se corrigieron manualmente para eliminar codones que presentan gaps (ya que los programas de cálculo de selección realizan corrimientos del marco de lectura cuando se encuentran con un gap) y para obtener el marco de lectura adecuado en aquellos genes anidados del genoma de HIV-1. Los valores de selección obtenidos para nucleótidos y codones superpuestos en aquellas regiones-genes sobrepuestas fueron promediados.

La diversidad genética puede ser descrita a partir de dos estimadores π (pi) y teta θ (teta). π se refiere a la *diversidad nucleotídica*, que es el número de nucleótidos diferentes por sitio entre dos secuencias tomadas al azar (Nei y Li, 1979) [ver *ecuación 1*]. θ, por otro lado, es calculado a partir de la expresión $\theta = 4Ne\mu$ (para secuencias de organismos diploides) y $\theta = 2Ne\mu$ (para secuencias de organismos haploides) (Kimura, 1968). Sin embargo, es difícil tener la determinación exacta de los parámetros *Ne* (tamaño efectivo poblacional) y μ (tasa de mutación), por lo que una manera indirecta de estimar θ es utilizando el número total de sitios segregativos (*S*) en un grupo de secuencias (un sitio segregativo es un sitio en donde las secuencias difieren). De esta forma, $\theta = K/a$, donde *K* es igual al número total de sitios segregativos en una muestra de secuencias dada y $a = 1 + 1/2 + ... + 1/n-1$, donde *n* es igual al número total de secuencias en la muestra (Watterson, 1975; Tajima, 1983). Se deduce de estas expresiones que π se ve afectada mayormente por los alelos que poseen mayor frecuencia y es independiente del tamaño de la muestra, mientras que θ se ve afectada por el tamaño de la muestra y por la deriva génica, es decir, por los alelos poco frecuentes (Tajima, 1983, 1989).

Así, la ***prueba de Tajima*** (su parámetro denominado *D*) calcula la diferencia con una confianza estadística entre los estimados *S* y π (haciendo que *D* pueda ser positivo o negativo) [ver ecuación 1]. De no existir diferencia entre *S* y π o que difieran sólo en lo esperado por el azar ($S \approx \pi$) (Fu and Li, 1993), *D*=0, entonces se aceptaría la hipótesis nula de neutralidad. Pero la selección, fluctuaciones demográficas y otras violaciones del modelo neutral cambiarán los valores esperados de *S* y π, haciéndolos significativamente diferentes como para obtener valores de *D*>0 (positivos), indicando selección positiva o direccional, o bien valores de *D*<0 (negativos), indicando selección purificadora o balanceadora (Rozas, 2009, Yang, 2006).

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{V(\hat{\theta}_\pi - \hat{\theta}_S)}} \qquad (1)$$

El número total de *S* está definido como $S=\hat{\theta}_S$, por lo que $\hat{\theta}_S=S/a_1$, en el que *S*, representa al número de sitios segregantes en una muestra de *n* secuencias y a_1, la longitud de la muestra dada por $a_1 = \sum_{i=1}^{n-1} 1/i$. Así, la diversidad nucleotídica π, representada como $\pi=\hat{\theta}_\pi$, se estima como $\pi = (\frac{n}{n-1})(\sum p_i p_j p_{ij})$, en el que π, representa la diversidad nucleotídica; *n*, el número de secuencias en la muestra; p_i o p_j, la proporción de la secuencia *i-ésima* o *j-ésima* en la muestra, respectivamente y, la π_{ij}, es la proporción de nucleótidos diferentes entre la secuencia *i-ésima* o *j-ésima*. La *V* representa a la varianza resultante de la diferencia entre los estimadores $\hat{\theta}_S$ y $\hat{\theta}_\pi$.

C. Estimación de selección a través de dS y dN por tres métodos: SLAC, FEL e IFEL

La prueba de neutralidad de Tajima se basa en supuestos sobre la historia demográfica de la población bajo estudio (i.e., compara estimados del tamaño efectivo de la población usando diferentes medidas de variación genética) (Li, 2011), y emplea la frecuencia de polimorfismo de secuencias del espectro de frecuencias sin tener en cuenta la información de *sustituciones sinónimas (dS)* y *no sinónimas (dN)* (Rozas, 2009). De este modo, se recurrió a la estimación de selección por otro método independiente. Estimados de sustituciones no sinónimas (*dN*)

significativamente diferentes de sustituciones sinónimas (*dS*) proveen evidencia de evolución no neutral (Ridley, 2004, Poon et al., 2009). Este segundo método usa probabilidad máxima (del inglés *máximum likelihood*) para la reconstrucción de secuencias ancestrales y calcula la sustitución de *dS* y *dN* en base a las relaciones filogenéticas.

Para estimar los sitios de aminoácidos bajo selección, se calcularon las tasas de *dS* y *dN* para todos los codones del genoma de HIV-1 a través de tres métodos: (1) el método de *probabilidad única de conteo ancestral* (SLAC), el cual calcula el número de cambios *dN* y *dS* que han ocurrido en cada codón a través de la historia evolutiva de la muestra (i.e., filogenia), identificando si *dN* es significativamente diferente de *dS*; (2) el modelo de *probabilidad de efectos fijos* (FEL), el cual ajusta tasas de sustitución entre *dN* y *dS* sobre una base de sitio-por-sitio; y (3) el modelo de *probabilidad de efectos fijos internos* (IFEL), el cual es esencialmente igual a FEL, excepto que la selección es probada a lo largo de las ramas internas de la filogenia (Kosakovsky-Pond and Frost, 2005, Kosakovsky-Pond et al., 2006). Estos modelos detectan codones sujetos a selección positiva (si $dN > dS$), negativa (si $dN < dS$) o evolución neutral (si $dN \sim dS$) bajo métodos de máxima verosimilitud. Se realizó dicho cálculo empleando el programa *HyPhy* v1.0beta (Kosakovsky-Pond et al., 2005).

Para realizar el análisis comparativo entre los tres métodos descritos previamente, se usaron 4 diferentes y sencillos parámetros para evaluar los resultados: *1)* en el análisis se usó solamente la región funcional *Vif*, dado que esta región es pequeña y de fácil análisis, *2)* se aplicaron dos modelos de sustitución de nucleótidos, *HKY85* y el modelo mejor obtenido a través de la *prueba de comparación de modelos* (*best fitting model*), integrado en el programa *HyPhy*. Asimismo, *3)* se usó un *umbral de significancia* de *0.05* y *4)* un árbol filogenético construido bajo el método de *Neighbor-Joining* mediante el uso de la distancia Tamura-Nei (Tamura and Nei, 1993). A pesar de que esta comparación ya se ha realizado de forma oficial por los autores del software (Kosakovsky-Pond and Frost, 2005), se decidió realizarla nuevamente para observar directamente si alguno de los tres modelos podría representar mejor o mostrar algún sesgo en nuestras predicciones. Se obtuvieron las mismas conclusiones de los autores: no son tan diferentes las predicciones de *dSs y dNs* obtenidas con los tres modelos, aunque la intensidad de la selección sí se ve diferenciada entre los tres modelos [ver Tabla R4 y Figura

R4 en resultados]. Se decidió usar el método SLAC para realizar la evaluación de todos los codones del genoma del virus HIV-1, el cual se ha determinado como el más sensible al número de muestras y el método más conservador en el análisis de secuencias (Kosakovsky-Pond and Frost, 2005, Poon et al., 2009, Delport et al., 2010).

D. Compilación de datos experimentales sobre selección de codones

A fin de robustecer la confianza de nuestras predicciones sobre el *fitness*, se realizó una búsqueda exhaustiva en la literatura, y en las bases de datos, de los codones que presentaran mutaciones con ventajas y desventajas selectivas para la eficiencia del virus HIIV-1 en la regiones codificantes para p6, proteasa y reverso transcriptasa (que se encuentran en los genes *pol* y *gag-pol*) [ver Figura R1, Figura S2-A y Tablas S1 en material suplementario]. Se decidió enfocarnos en esta región del genoma ya que es la que más datos experimentales y bioinformáticos poseen, lo que nos proporcionó una mayor cobertura de evaluación. En los datos experimentales, la fuerza de selección de una mutación en el *fitness* del virus se reporta con base en los efectos en su replicación, letalidad en la infección, evasión del sistema inmune, sobrevivencia e interacción con el hospedero. De los 439 codones que se encuentran en la región *pol* y *gag-pol* del genoma de HIV-1, se logró recuperar información para 400 codones con algún tipo de efecto selectivo sobre el *fitness* del virus. Los datos compilados de la literatura se muestran en la Tabla R5 en la sección de resultados y en la Tabla S1 del material suplementario disponible al final del manuscrito.

Las bases de datos consultadas para este estudio fueron *PFAM* (Finn et al., 2010), *Prosite* (Sigrist et al., 2010), *EMBL-EBI* (Cochrane et al., 2008), *Los Alamos* (http://resdb.lanl.gov/Resist_DB), la *base de datos de mutaciones positivas en HIV de UCLA* (UCLA-HIV database) (http://bioinfo.mbi.ucla.edu/HIV/) y *UniProt* (Jain et al., 2009). Mientras que la base de datos Los Alamos contiene exclusivamente información experimental, la base de datos HIV de UCLA contiene información experimental/bioinformática de selección de codones en las regiones proteasa y reverso transcriptasa en base al radio *Ka/Ks* (or ω, *dN/dS*). *Ka/Ks* es el radio del número de sustituciones no sinónimas por sitio no sinónimo (*Ka*) al número de sustituciones sinónimas por sitio sinónimo (*Ks*), el cual puede ser

usado como un indicador de presión selectiva actuando sobre una región codificante para proteína. Un radio *Ka/Ks* alto indica selección positiva, mientras que un radio *Ka/Ks* bajo indica selección purificadora.

E. Análisis estadístico de las predicciones

Posteriormente, y una vez obtenidos los resultados tanto de *Tajima* como de *SLAC*, se procedió a comparar ambos métodos para la parte de nuestras predicciones bajo una prueba estadística que evalúa la *sensibilidad* y *especificidad*. Los datos experimentales se compilaron de la literatura para la región *pol* y *gag-pol* (descritos anteriormente), y éstos fueron utilizados para realizar la comparación contra nuestras predicciones bioinformáticas [ver *cuadro* en Figura R4 y Tabla R5]. Para ello, se realizó un *script* en lenguaje Perl que considera las siguientes estimaciones: a) *sensibilidad*, la cual mide la proporción de datos *positivos reales* que son correctamente identificados, y b) la *especificidad*, la cual mide la proporción de datos *verdaderos negativos* que se han identificado correctamente. La sensibilidad está definida bajo la fórmula $S_{sen} = VP / VP + FN$, donde S_{sen} es la sensibilidad, *VP* representa a los *verdaderos positivos* y *FN* representa a los *falsos negativos*. Mientras que la especificidad se define como $S_{spe} = VN / VN + FP$, donde S_{spe} es la especificidad, donde *VN* son los *verdaderos negativos* y *FP* los *falsos positivos*.

De este modo, los *verdaderos positivos* (*VP*) corresponden al número de codones mutantes que, con nuestras predicciones y corroborados con los datos experimentales/bioinformáticos, presentan un determinado tipo de selección (en ambos positivos o en ambos negativos). Los falsos positivos (*FP*) corresponden al número de codones mutantes que presentan un tipo de selección con nuestras predicciones, pero que no son corroborados con los datos experimentales/bioinformáticos. Los verdaderos negativos (*VN*) corresponden al número de codones mutantes que no presentan ningún tipo de selección (i.e., son neutros) y que la ausencia de selección es corroborada con datos experimentales/bioinformáticos. Finalmente, los falsos negativos (*FN*) corresponden al número de codones mutantes que son erróneamente predichos como neutros en nuestras predicciones, cuando se ha encontrado evidencia de selección por medios experimentales u otros bioinformáticos.

RESULTADOS Y DISCUSIÓN

I. Parámetros biológicos para el programa quasispecies.pl: la tasa de mutación, la frecuencia del tipo de mutaciones, el tiempo generacional y el tamaño efectivo de la población.

La teoría de las cuasiespecies asocia los conceptos de fitness, crecimiento poblacional, mutación, heterogeneidad y distribución de las poblaciones a través de 4 parámetros generales y básicos que posee cualquier sistema viral: 1) un genoma (información genética), 2) replicación, 3) mutación y 4) selección en *poblaciones* [ver Texto S1: el modelo de Eigen-Schuster]. Estos parámetros son los que rigen la dinámica evolutiva de las cuasiespecies. No obstante, Eigen y Schuster usaron estimaciones teóricas para los diferentes parámetros que integran al modelo (Biebricher and Eigen, 2005). A continuación, describimos los datos experimentales que se han reportado para estos parámetros y que son indispensables para llevar a cabo nuestras simulaciones con el algoritmo genético.

A. Sobre la tasa de mutación de HIV-1.

La *tasa de mutación* en virus se refiere al número de errores genéticos (mutaciones puntuales, inserciones y eliminaciones) que son acumulados por unidad de tiempo (Ge et al., 2007), o por generación (por lisis para virus líticos obligados) (Chao et al., 2002), o por ronda de replicación genómica (Duffy et al., 2008). La más útil de estas medidas es frecuentemente la *tasa de mutación por ciclo de replicación*, aunque es difícil de determinar para HIV-1, en primer lugar, ya que requiere del conocimiento de los detalles del mecanismo de replicación viral (que se describieron en la introducción), si este se lleva a cabo a través de una "replicación lineal reiterativa" (del inglés *stamping machine*), o de una replicación "genómica geométrica", o por medio de una mezcla de ambos tipos de replicación (Drake, 2007).

En segundo lugar, la tasa de mutación para HIV-1 es difícil de estimar en la práctica experimental, según el tipo de cuantificaciones que se realicen, si estas se realizan *in vitro* o *in*

vivo, y el tipo de correcciones que se tomen en cuenta para las estimaciones finales. Se puede hablar, en función del tipo de estimación, de la tasa de mutación por par de bases por ciclo de replicación (μ_b), la tasa de mutación por genoma por ciclo de replicación (μ_g), o la tasa de mutación por genoma efectivo por ciclo de replicación (μ_{eg}) (Drake et al., 1998). Eigen y Schuster (1977) han sugerido que la tasa de mutación tolerable a eliminaciones no puede ser mayor a 1 mutación por genoma por ciclo de replicación, sin que ello no mitigue factores tales como una alta fecundidad, grandes poblaciones, y la recombinación. Se ha estimado que esta tasa de mutación corresponde a los virus de RNA líticos, mientras que se sugiere que los retrovirus deben tener tasas de mutación menores (Drake, 1993, Drake et al., 1998). La tasa de mutación es medida principalmente por dos vías, a partir de los *exámenes de fluctuación de Luria–Delbrück* o vía los *estudios de acumulación de mutaciones*.

Los *estudios de acumulación de mutaciones* se definen de dos maneras. Las medidas fenotípicas de la acumulación de mutaciones involucran someter a las poblaciones a cuellos de botella y seguir la frecuencia de la mutación a través de los cambios en el tamaño efectivo de la población (Drake, 1991 en: Duffy et al., 2008). Alternativamente, datos de mutaciones en secuencias pueden obtenerse a partir de genomas de poblaciones que han replicado después de un número conocido de generaciones (Baer et al., 2007 en: Duffy et al., 2008). Sin embargo, dado que este método no puede excluir todos los efectos de la selección natural, el método de acumulación de mutaciones puede reflejar en mayor medida la *tasa de sustitución de nucleótidos* que la de mutación. La *tasa de sustitución* se define como el número de cambios mutacionales fijados (por selección natural o por deriva genética) por sitio nucleotídico por unidad de tiempo (usualmente años). Las tasas de sustitución reflejan, de esta forma, el producto complejo de cuatro factores: la tasa de mutación subyacente, el tiempo generacional, el tamaño efectivo de la población (Ne) y el *fitness*, con mutaciones ventajosas fijadas en la población de forma más rápida que las mutaciones neutrales (Duffy et al., 2008).

El *método Luria–Delbrück*, también conocido como el *método de frecuencia de mutantes*, consiste en medir la frecuencia de mutaciones en un determinado fenotipo que surge al replicar poblaciones que se expanden clonalmente. Esta frecuencia se ajusta entonces para tener en cuenta el número de generaciones y el número de replicaciones del genoma dentro de

cada generación para obtener una tasa de mutación (por ciclo de replicación). Estas tasas se pueden calcular también a partir del número de poblaciones replicadas que no generan mutaciones (Drake, 1991; 1993 en: Duffy et al., 2008), suponiendo que el número de mutaciones sigue una distribución de Poisson. En el presente proyecto se usaron tasas de mutación para el virus de HIV-1 estimadas con el método de frecuencia de mutantes.

HIV-1 es genéticamente diverso tanto dentro como entre individuos infectados (Bonhoeffer et al., 1995, Butler et al., 2007, Shankarappa et al., 1999 en: Abram et al, 2010). La alta tasa de la replicación viral, el tamaño de la población viral en pacientes individuales, y la recombinación extensiva contribuyen a la variación genética de HIV-1. Sin embargo, la causa principal de la diversificación de HIV-1 son las mutaciones que surgen durante la replicación retroviral (Abram et al., 2010). En contraste a los virus de RNA líticos, un retrovirus replica precisamente tres o hasta cuatro veces por ciclo infectivo (ver Figura I9) (Smith et al., 2005, Menéndez-Arias, 2009, Abram et al., 2010): (i) cuando el RNA viral es transcrito a partir del provirus por la RNA polimerasa II dependiente de DNA (RNA Pol II) del hospedero, (ii) cuando el genoma viral de RNA de cadena sencilla (ssRNA) es convertido en DNA de cadena doble (dsDNA) a través de la reverso transcriptasa (RT), y (iii) cuando el provirus es copiado por la DNA polimerasa dependiente de DNA del hospedero cuando la célula infectada replica. La frecuencia del mutante resultante es entonces la suma de las tasas de mutación en los tres pasos, cuyas magnitudes no han sido factorizadas del todo (Drake et al., 1998). Debido a la notable fidelidad de la replicación del DNA celular (de 10^{-9} a 10^{-11} mutaciones por par de bases), es poco probable que las mutaciones que ocurren durante la replicación celular de los provirus contribuya significativamente al alto grado de diversidad del virus HIV-1 (O'Neil et al., 2002). De esta manera, muchas mutaciones son introducidas tanto durante la transcripción reversa llevada a cabo por la RT como en la transcripción de los provirus de DNA a genomas de RNA a través de la RNA Pol II.

De forma contraria a las DNA polimerasas que replican en la célula, la RT del virus HIV-1 carece de la actividad de corrección exonucleolítica de 3’ a 5’. Estudios *in vitro* e *in vivo* han sugerido que la RT es propensa a errores y que es probablemente la principal responsable de la diversidad genómica del virus HIV-1 (Roberts et al, 1988; Preston et al, 1988; Bebenek et al.,

1989; Nowak, 1990; Ji and Loeb, 1992; Mansky and Temin, 1995; Kim et al., 1996; O'Neil et al., 2002 en: Abram, et al., 2010). Estos estudios de fidelidad usando RT de HIV-1 en sistemas libres de células indican que las frecuencias de una mal incorporación de nucleótidos son bastante altas (desde $2.5x10^{-4}$ a $5.3x10^{-5}$ mutaciones por par de bases por ciclo de replicación) (Roberts et al., 1989, Roberts et al., 1988, Bebenek et al., 1989, Preston et al, 1988; Rezende y Prasad, 2004, Svarovskaia et al., 2003 en: Abram et al, 2010), aunque hay que destacar que estas estimaciones son típicamente mayores que la tasa global de mutación observada durante la replicación del virus HIV-1, lo cual indica que otros factores contribuyen a la fidelidad general *in vivo* (Preston et al., 1988, Mansky y Temin, 995, Ji y Loeb, 1992, Klarmann et al, 1997 en: O'Neil, 2002). Parte del problema es que la fidelidad de RT de HIV-1 *in vitro* depende de ambos, de la RT recombinante usada y de las condiciones de los experimentos. En los sistemas *in vivo* (cultivos celulares), por el contrario, existen varios componentes celulares que contribuyen a incrementar, así como también a disminuir, la fidelidad de la replicación de HIV-1. De esta forma, Abram y colaboradores (2010) han sugerido recientemente que no es posible discernir entre la contribución mutacional de la RT y la RNA Pol II en estudios *in vivo*.

Las estimaciones *in vivo* se realizan al copiar un templado de DNA o de RNA a través de la RT; este templado codifica para una región blanco en el que las mutaciones pueden ser fácilmente identificadas. Por ejemplo, el gen *lacZα* de *Escherichia coli* es uno de los más usados, aunque también han sido usados los genes *env* y las regiones LTRs de HIV-1. A pesar de que varios grupos de investigación han usado el mismo gen blanco (*lacZα*), diferentes tasas de error han sido reportadas, y también hay poco acuerdo sobre las proporciones y frecuencias de los tipos de mutaciones específicos que realiza la RT del HIV-1. Estimaciones experimentales *in vivo* basadas en frecuencias de mutantes para el gen *lacZα* sugieren que la μ_b de HIV-1 es de $3x10^{-5}$ a $4x10^{-5}$ mutaciones por par de bases por ciclo de replicación (Mansky and Temin, 1995, Mansky, 1996a). Al tomar en cuenta los filtros de mutaciones causadas por los métodos de secuenciación, análisis de polimorfismo y eficiencia de detección de mutaciones en estas estimaciones, Drake y colaboradores (1998) sugieren una μ_g de 0.1 a 0.2 en HIV-1 (cuando el RNA viral no está integrado en el genoma del hospedero). De forma que sólo 1 ó 2 de cada 10 genomas en un ciclo de replicación tendrían una mutación bajo estas

estimaciones. Usando el mismo gen blanco, *lacZα,* Abram y colaboradores (2010) han estimado recientemente una μ_b en HIV-1 *in vivo* de $2.4x10^{-5}$ muts/bp/replicación, lo cual sugiere 1 mutación por cada 10 ó 20 genomas en un ciclo de replicación.

Por otro lado, O'Neil y colaboradores (2002) identificaron *in vivo* mutaciones puntuales en los genes *env* y las regiones LTRs 3' y 5' en dos de las tres etapas de la replicación de dos cepas del virus HIV-1 (HXB2 y NL4-3), a fin de cuantificar las aportaciones de la RT y de la de RNA Pol II de forma independiente, para lo cual usaron ensayos de PCR, Polimorfismos en Conformación de Cadena Sencilla (SSCP del inglés *Single-stranded Conformation Polymorphism*) y secuenciación de DNA [ver Figura S4 en material suplementario]. Los autores encontraron que bajo esta estrategia metodológica, las tasas de mutación obtenidas para las dos cepas de HIV-1, HXB2 y NL4-3, fueron muy cercanas: 9.2×10^{-5} y 7.9×10^{-5} por par de bases por ciclo de replicación, respectivamente. Este trabajo argumenta una tasa de mutación en promedio de 8.5×10^{-5} por par de bases por ciclo de replicación, lo cual es al menos 2.5 veces más alta que las tasas de mutación reportadas con el gen *lacZα.* O'Neil y colaboradores (2002) discuten que las diferencias entre estas tasas de mutación se pueden deber a tres factores: (1) que la generación de mutaciones sea ligeramente más alta en las regiones terminales del genoma; (2) que la diferencia en la secuencia nucleotídica de los marcadores usados genere mutaciones diferencialmente; (3) el uso diferencial de las mutaciones por las metodologías usadas, mientras que el método SSCP evalúa ~80% de la mutaciones reales generadas, el conteo fenotípico a través del gen *lacZα* sólo utiliza ~40% de las mutaciones. Como ya se mencionó anteriormente, estudios previos *in vitro* han reportado que la μ_b de la RT de HIV-1 es de una mutación por cada 1700 ó 4000 bases nucleotídicas (i.e., 1/1700 ó 1/4000), y en regiones altamente mutables es de 1 por 70 nucleótidos polimerizados (Roberts et al., 1988, Preston et al., 1988). Estas estimaciones, en conjunto con la de O'Neil y colaboradores (2002), sugieren que el virus de HIV-1 podría tener una μ_g de 1 a 5 mutaciones por genoma por ciclo de replicación *in vivo*.

En el presente trabajo, se uso una tasa de mutación para las simulaciones de 1 hasta 4 mutaciones por genoma por ciclo de replicación. Este intervalo fue considerado suficiente para

ver el efecto de la mutación sobre la dinámica de cuasiespecies en nuestras simulaciones. Como se describirá más adelante, aún cuando la tasa de mutación fuese menor a 1 mutación por genoma por ciclo de replicación, nuestras simulaciones muestran que poblaciones de HIV-1 con *fitness* altos son capaces de soportar hasta 2 mutaciones por ciclo de replicación. Al incluir más de 2 mutaciones por genoma por ciclo de replicación, los genomas con fitness altos desaparecen, pero la población no se extingue, lo cual refleja que el virus HIV-1 posee una interesante robustez mutacional. Se discutirán estos hallazgos más adelante.

B. Sobre la frecuencia del tipo de mutaciones en HIV-1.

La variación genética de las poblaciones de HIV-1 es el resultado de la incorporación de numerosas mutaciones en el proceso mismo de replicación del genoma viral que, además de analizarse por medio de la tasa de mutación, también es posible cuantificarla por medio del cambio de cada nucleótido en cada ronda de replicación. Así, la *frecuencia de mutación* se refiere a la proporción de mutaciones específicas observadas en una población. Esta medida es empleada comúnmente para calcular y distinguir las frecuencias de las mutaciones, debido a que hay diferentes tipos de estas y su ocurrencia en los genomas no es la misma ni completamente al azar. En términos simples, "la frecuencia de mutaciones" es calculada como el número de mutantes recuperados entre la población total, ya sea para mutaciones de nucleótido únicas o múltiples, de tipo *frameshift*, inserciones, o eliminaciones (i.e., *deletions* o *indels*).

Actualmente hay un número considerable de estudios que han realizado esta estimación en diversos virus (revisado en Sanjuán et al., 2010), exponiendo una gran variedad en estos valores. Para HIV-1, existen al menos 4 reportes que han estimado los cambios de frecuencias de mutaciones para un ciclo de replicación [ver Tabla R1]. Las investigaciones realizadas por Mansky y Temin (1995), Mansky (1996) y Abram y colaboradores (2010) emplean el mismo sistema de detección, el cual consiste en el uso de un gen reportero conocido como *lacZ*α (que ya se describió previamente). Además de emplear el mismo sistema de detección, ellos también calculan la frecuencia de mutaciones en función del número de mutantes recobrados

en todos los experimentos. Todos los reportes han estimado, al menos, 4 clases de mutaciones: 1) *mutaciones de sustitución única*, 2) mutaciones tipo *frameshift de sustitución única*, 3) *sustituciones múltiples* e 4) *inserciones* y eliminaciones.

Los resultados de estas investigaciones muestran que hay una consistente relación en la sustitución de nucleótidos y otro tipo de mutaciones. Por ejemplo, las mutaciones más frecuentes en una ronda de replicación de HIV-1 son las de *clase 1*, seguidas de la de *clase 2* y, con una menor frecuencia, las de *clase 4*. Es decir, como se muestra en la Tabla R1, en promedio estas clases corresponden al 79%, 6% y 4%, respectivamente, del total de las mutaciones producidas en los experimentos de Abram y colaboradores (2010); el 48%, 33% y 11%, respectivamente, en Mansky (1996); el 66%, 27% y el 10%, respectivamente, en Mansky y Temin (1995); y finalmente, el 72%, 14%, 14%, respectivamente, en O'Neil y colaboradores (2002). Tomando en cuenta la clase 3 reportada únicamente en Mansky (1996) y en Abram y colaboradores, (2010), esta ocupa una fracción del 7% al 85% del total de las mutaciones generadas por cada ciclo de replicación en HIV-1.

Todas las investigaciones reportan a las hipermutaciones, transiciones y transversiones como las mutaciones más comunes entre nucleótidos; donde las transiciones son más abundantes que las transversiones. Por ejemplo, Mansky y Temin reportan que las mutaciones de G→A son mayores que las de C→T, y estas a su vez son más frecuentes que las del tipo T→C, T→G y T→A (para observar a detalle el número de mutaciones obtenidas, ver Tabla 2 y 4 en Mansky y Temin, 1995; Tabla 2 en Mansky, 1996; y Tabla 3 en Abram et al., 2010). Otra de las conclusiones en común de estos reportes es que las mutaciones de clase 2 (*frameshift* tipo (-) o (+)) son mucho más frecuentes en los nucleótidos T y A que en G y C [para observar la distribución de estas mutaciones, ver Figura S5 en material suplementario]. De forma interesante, si se toma en cuenta la frecuencia de las mutaciones múltiples (clase 3) de forma independiente a las otras clases, es notorio observar que estas presentan un gran porcentaje de mutaciones; siendo los cambios de G→A de los más abundantes, aunque ésto sólo es reportado por Abram et al., 2010 [ver Tabla R1].

Tabla R1. Tasa, frecuencia y fracción de mutaciones en HIV-1. Estimaciones calculadas en diferentes trabajos experimentales.

	Total global *		**Clase 1**⊙		**Clase 2**‡		**Clase 3**⋈		**Clase 4**♣	
lacZa: orientación y tipo mutaciones (s)	Frecuencia de mutación ♠	Tasa de mutación∞	Frecuencia de mutación	Fracción de mutación (%)♥	Frecuencia de mutación	Fracción de mutación (%)	Frecuencia de mutación	Fracción de mutación (%)	Frecuencia de mutación	Fracción de mutación (%)
ABRAM et al., 2010										
LZF (directo)										
Todas – clase 3[A]	361/172,562	$1.2x10^{-5}$	324/172,562	90	26/172,562	7			11/172,562	3
Todas + clase 3 dep. [B]	361/172,562	$1.4x10^{-5}$		75		6	71/172,562	16		3
Substituciones A→G [C]							64/172,562	90		
Otros [D]							7/172,562	10		
Todas + clase 3 indep.	669/172,562	$2.2x10^{-5}$		48		4	308/172,562	46		2
Substituciones A→G							289/172,562	94		
Otros							19/172,562	6		
LZR (reverso)										
Todas – clase 3	309/155,558	$1.1x10^{-5}$	280/155,558	91	16/155,558	5			13/155,558	4
Todas + clase 3 dep.	340/155,558	$1.3x10^{-5}$		83		5	31/155,558	9		4
Substituciones A→G							25/155,558	81		
Otros							6/155,558	19		
Todas + clase 3 indep	471/155,558	$1.7x10^{-5}$		60		3	162/155,558	34		3
Substituciones A→G							150/155,558	93		
Otros							12/155,558	7		
MANSKY (1996)										
3.12 directo										
Todas las clases	27/5,272	$4x10^{-5}$	13/5,272	48	9/5,272	33			3/5,272	11
Substituciones A→G							2/5,272	7		
MANSKY & TEMIN (1995)										
3.12 directo										
Todas las clases	38/8,678	$3.5x10^{-5}$	25/8,678	66	9/8,678	24			4/8,678	10
Substituciones A→G							nd∅	nd		
5.1 opuesto										
Todas las clases	28/6,746	$2.4x10^{-5}$	17/6,746	61	8/6,746	29			3/6,746	10
Substituciones A→G							nd	nd		
O'NEIL et al., (2002)[F]										
Todas las clases (LTRs)	21/215	$8.5x10^{5}$	15/215	72	3/215	14	nd	nd	3/215	14

Fila superior: *Total de frecuencias y tasas de mutaciones de todas las clases. ⊙ Mutaciones únicas de un solo nucleótido. ‡ Mutaciones tipo frameshift. ⋈Mutaciones múltiples. ♣ Mutaciones tipo *indel*. La frecuencia (♠) y tasa de mutación (∞) fueron calculadas como numero de mutantes recobradas por viriones totales analizados y la frecuencia dividida entre el reportero (nts). La fracción de mutación es el número de mutaciones de una particular clase expresada, como un porcentaje.

Columna izquierda inicial: [A] Todas las mutaciones, excepto las de las clase 3 debido a que las múltiples mutaciones podrían no ser hechas por la RNA pol II o la RT (clase 1, 2 y 4). [B] Todas las clases de mutaciones son contadas. [C] Mutaciones del tipo AG presentes. [D] Otro tipo de mutaciones no descritas aquí. [F] Los experimentos hechos por O'Neil et al., involucraron el análisis de ambos LTRs (5' y 3') en dos virus de referencia para HIV-1: HXB2 y NL4-3. Aquí se muestra el resumen del tipo, fracción y tasa de de mutaciones para este reporte. ∅nd, representa "no determinado".

Una gran similitud entre los resultados de Mansky y Temin (1995), Mansky (1995) y Abram y colaboradores (2010), es que la orientación en que se lleve a cabo la replicación del gen reportero (5'→3' o 3'→5') no afecta la tasa global de mutación, ni tampoco la frecuencia de estas significativamente. Mientras que los 4 trabajos obtuvieron resultados similares con respecto a las frecuencias del tipo de mutaciones que se llevan a cabo durante la replicación, existe una gran diferencia (de 2 ó 3 órdenes de magnitud) sobre las estimaciones de la tasa de mutación que fue calculada a partir de la frecuencia de sus mutantes (ver Tabla R1).

Las frecuencias de las mutaciones son una medida crucial para entender una gran parte de la evolución de HIV-1 a través de sus eventos de replicación. Los reportes de Mansky y Temin (1995), Mansky (1996), y el de Abram y colaboradores (2010), han sugerido que los mecanismos que participan en la variabilidad genética de HIV-1 van más allá de las polimerasas que participan en su replicación e, inclusive, más allá de la misma RT. Existen componentes virales y celulares que influyen en el tipo de mutaciones que son incorporadas en el genoma viral (ver Figura I9). Se pueden mencionar la participación activa de las proteínas virales no estructurales (*e.g. vpr, vif, tat*), y del tipo de conformación del genoma en cadena sencilla o doble (Mansky and Temin, 1995, Frankel and Young, 1998). Dentro de los componentes celulares se pueden mencionar la presencia de un ineficiente alineamiento de las cadenas complementarias durante la replicación del ácido nucleico (un proceso conocido como *slippage*), mecanismos celulares de reparación de mal apareamiento en las secuencias, así como la participación de proteínas del hospedero con efecto antiviral (APOBEC y ADAR). La acción de APOBEC y ADAR es considerada como un mecanismo de *hipermutación* provocado por la célula induciendo cambios de A→G, es decir, un mecanismo de *mutagénesis letal natural* (Frankel and Young, 1998, Yu et al., 2003). Las 'hipermutaciones' se definen como largos periodos de transición/sustitución de nucleótidos; aunque este término también se ha empleado para describir una elevada tasa de mutación de cualquier tipo, no necesariamente en una serie de nucleótidos. Este tipo de transiciones se ha encontrado en el virus HIV-1 (Caride et al., 2002, citado en: Duffy et al., 2008).

Al analizar en forma completa las estimaciones sobre la obtención y cálculo de la frecuencia de sustitución de nucleótidos en un ciclo de replicación para el genoma de HIV-1 y, a su vez,

conocer que hay una tendencia de cambio dependiendo del tipo de mutaciones y tipo de nucleótidos, se decidió estimar un consenso de las frecuencias reportadas en todas las investigaciones resumidas en la Tabla R1 con la finalidad de integrarlas en nuestro programa computacional y tener una aproximación más real del tipo de mutaciones que se realizan sobre las poblaciones virales. Nuestro algoritmo genético no toma en cuenta mutaciones del tipo *frameshift*, inserciones ni eliminaciones. Las frecuencias de mutaciones estimadas en porcentaje para cada nucleótidos y por tipo de mutación se muestran a continuación:

	A	T	G	C	TOTAL	Frequency
A	-	10	25	15	50	19
T	10	-	20	30	60	22
G	70	10	-	10	90	34
C	10	50	10	-	70	25
TOTAL	90	70	55	55	270	100

C. Sobre el tamaño efectivo de la población (N_e) de HIV-1.

La base intuitiva del concepto sobre el tamaño efectivo de la población (N_e), propuesto por S. Wright (1969), está asociado al tamaño poblacional que es relevante en términos evolutivos, i.e., el número de individuos reproductivos, ya que son éstos los que contribuyen a crear la generación siguiente en términos demográficos y sobre todo genéticos (Wrigth, 1969, Hedrick, 2000). Es por ello que el N_e se considera como el tamaño más pequeño teóricamente de la población a la que ésta puede evolucionar en la misma forma como la actual población bajo estudio. Dado que *Ne* está influenciado por varios factores (como aquellos que ocurren durante la transmisión de los virus entre los hospederos), N_e es usualmente mucho más pequeño que el tamaño total de la población. Por ejemplo, se ha notado una discrepancia de entre 100 a 100,000 órdenes entre el tamaño censado del total de la población y el N_e de HIV-1 (Liu and Mittler, 2008).

El producto del N_e y la tasa de mutación determinan el nivel de equilibrio de la variabilidad genética seleccionada neutralmente o débilmente en una población. Además, la efectividad de

selección para determinar si una mutación favorable se expande, o si una mutación deletérea es eliminada, es controlada por el producto de N_e y la intensidad de la selección. De esta forma, el N_e afecta de forma importante la variabilidad de los genomas y las tasas de evolución de sus componentes. Varias investigaciones han tratado de estimar N_e para HIV-1 (Brown, 1997; Nijhuis, et al.,1998; Rodrigo et al., 1999; Rouzine y Coffin 1999; Frost et al., 2000; Seo et al., 2002; Achaz et al.,2004; Shriner et al., 2004 en: Kouyos et al., 2006). Sin embargo, estimaciones divergentes y la complejidad inherente del concepto han creado una considerable confusión de cómo interpretar las estimaciones realizadas.

El N_e de las poblaciones naturales de HIV-1 puede estar influenciado por diversos procesos genéticos, ecológicos, evolutivos e infecciosos. Se mencionarán los tres primeros, mientras que el proceso infeccioso en el hospedero de discutirá más adelante (Kouyos et al., 2006):

1) *Una tasa desigual de reproducción viral:* frecuentemente unos pocos individuos producen la mayoría de la progenie viral, lo que incrementa efectos estocásticos a la vez de reducir el N_e. Se sabe, por ejemplo, que muchas células infectadas por HIV-1 (en estado latente o infectadas ineficazmente) producen pocos virus o ninguno (Nijhuis et al., 1998; Rodrigo et al., 1999; Rouzine y Coffin, 1991; Rouzine et al., 2001; Shriner et al., 2004, en: Kouyos et al., 2006).
2) *Una población estructurada:* la presencia de metapoblaciones (i.e., una colección de subpoblaciones) más que una gran población “bien mezclada” es lo que mejor describiría a la población de HIV-1 dentro del hospedero infectado (Frost et al, 2001, en: Kouyos, et al., 2006). Estas subpoblaciones experimentan frecuente extinción y recolonización, lo que generalmente reduce el N_e. Cada evento de recolonización representa un cuello de botella local, de modo que un pequeño número de virus son fundadores de una nueva subpoblación (Shriner et al., 2004; Haase et al., 1996; Frost et al., 2000; Wright, 1931; Bonhoeffer, et al., 1997 en Kouyos, et al., 2006).
3) *Las altas tasas de mutación y recombinación* también complican los intentos para realizar estimaciones sobre la N_e, ya que son muy frecuentes en las poblaciones virales (Liu and Mittler, 2008).

4) *Una recurrente selección sobre las poblaciones.* Es bien conocido, por ejemplo, que la selección puede reducir el N_e de HIV-1 (Brown, 1997; Achaz, et al., 2004; Seo et al., 2002; Shriner et al., 2004, en: Kouyos et al., 2006; Abram et al., 2010). Bajo los modelos neutrales, Brown (1997) estimó, por ejemplo, una N_e de ~1000 utilizando el gen *env*, mientras que Achaz et al. (2004), estimaron que N_e se encuentra entre 10^3 y 10^4 al usar la región *gag-pol*, investigaciones similares han convergido en estas últimas estimaciones (Drummond et al., 2002; Brown, 1997; Achaz, et al., 2004; Seo et al., 2002; Shriner et al., 2004, en: Liu y Mittler, 2008). Usando un modelo con selección, por el contrario, Rouzine y Coffin sugirieron que la N_e de HIV-1 se encuentra entre 10^5 y $5x10^5$.

Las fluctuaciones en las estimaciones del N_e de HIV-1 se deben, por un lado, a que es difícil tener un control estadístico de los factores que afectan la producción, distribución e infección viral; mientras que por el otro lado, los cálculos se realizan en diferentes momentos del ciclo reproductivo e infeccioso del virus. Las estimaciones se llevan a cabo, por ejemplo, en un momento específico en la infección y no a lo largo de toda la evolución de la enfermedad en diferentes individuos de una y entre varias poblaciones. Por lo que esta cifra, comúnmente, llega a ser sobrerrepresentada en la población.

Las cuantificaciones de las poblaciones producidas en ciclos de infección o por célula hospedera son las que se utilizan en la práctica para realizar las estimaciones de N_e en el virus de HIV-1. De manera que el término "tamaño de rotura" *(burst size* en inglés*)* es comúnmente usado para describir la cantidad total de virus producidos por una célula infectada (Nelson et al., 2004). Si se conoce el "tamaño de rotura", N, y el tiempo de vida promedio de una célula infectada productivamente, $1/\delta$ (que se ha estimado cerca de 2.2 días, con una vida media de 1.6 días (Perelson et al., 1996, Chen et al., 2007)) entonces, la tasa de producción viral, P, está dada por la siguiente relación $P=N\delta$, y denota la tasa promedio de producción viral a lo largo del tiempo de vida de una célula infectada reproductivamente. El "tamaño de rotura" se puede calcular *in vitro* a partir de la infección inducida de cultivos celulares de diferentes tejidos (i.e, a través de dilución de la carga viral generada e inoculación a lo largo de varios pasajes) con cepas de HIV-1 mantenidas en el laboratorio; o bien, se puede calcular *in vivo* al combinar análisis de imágenes con hibridación *in situ* en tejidos linfoides de pacientes infectados

crónicamente con HIV-1 [ver Figura S6 en material suplementario]. Se ha encontrado que las estimaciones del tamaño de rotura por ambos métodos son sumamente variables y, al igual que N_e, es un número difícil de calcular (Dimitrov et al., 1993, Chen et al., 2007, De Boer et al., 2010).

Se ha sugerido que la mayor producción viral de HIV-1 proviene de las células T $CD4^+$ en humano, ya que estas células son los principales blancos de la infección por el virus HIV-1 y el tejido linfoide posee más de 10^{11} células T $CD4^+$ (Perelson et al., 1996, Chen et al., 2007). Un ciclo completo de infección por HIV-1 en un una célula humana T $CD4^+$ se lleva a cabo en 3 ó 4 días (con un promedio de 2.6 días y un mínimo de 1.2 días (Perelson et al., 1996)) y procede en tres etapas (Dimitrov et al., 1993): (i) el período de iniciación de una infección durante el cual no se detecta ninguna producción viral, (ii) una fase en la que la producción inicial y liberación de la progenie viral se puede medir, esta producción se incrementa abruptamente hasta alcanzar un pico máximo, y (iii) un período final en el que la producción de viriones disminuye. Así, una célula humana que es infectada con HIV-1 puede producir en la segunda etapa de la infección entre 10,000 y más de 100,000 nuevas partículas virales durante su corto tiempo de vida (Drummond et al., 2002; Brown, 1997; Achaz, et al., 2004; Seo et al., 2002; Shriner et al., 2004, en: Liu y Mittler, 2008), o entre 10^5, $5x10^4$ (Chen et al., 2007) y $5x10^5$ (Rouzine and Coffin, 1999). Mientras que la estimación llega a ser de $5x10^{10}$ (Haase et al., 1996) si se asume que la cantidad máxima de RNA de HIV-1 en una célula infectada corresponde al tamaño de rotura, y de $10.3x10^9$ viriones si la estimación se realiza por día (Perelson et al., 1996). Sin embargo, las estimaciones entre 10^3 y 10^5 parecen contradictorias respecto al número de células humanas T $CD4^+$ infectadas productivamente, el cual se ha calculado entre 10^7 y 10^8 células (Chun et al., 1997, Haase, 1999). Así como también, las estimaciones de 10^{10} podrían entrar en conflicto con el número de células T $CD4^+$ disponibles en el tejido linfoide ($\sim10^{11}$).

Además, estimaciones *in vivo* para todo el cuerpo humano muestran que pacientes infectados crónicamente con HIV-1 (analizados por largos periodos de tiempo) poseen en promedio mucho menos que 10^3 (10,000) partículas virales libres potencialmente infecciosas y mucho menos que 10^4 (100,000) células productivamente infectadas (De Boer et al., 2010). En una

infección crónica de HIV-1, existe una *tasa constante de carga viral* (llamado *set-point* en inglés) debido a que hay una rápida rotación de células infectadas. Para mantener esta tasa constante de carga viral, se requiere así de un balance entre la producción y la eliminación viral, y entre la producción y la muerte de las células blanco. De modo que para formular una aproximación más realista del N_e de HIV-1, se deben considerar al menos dos procesos aparentemente contradictorios que se han registrado *in vivo* con diferentes metodologías y aproximaciones teóricas:

1) **La eliminación viral.** Debido a que la población de partículas víricas libres en sangre es mucho más baja de la población que se produce por células infectadas productivamente, se sugiere que la tasa a la cual las partículas virales son eliminadas es rápida e influye de forma importante en el N_e de HIV-1. La producción viral por cada célula infectada productivamente es alterada por la existencia de un proceso de *eliminación viral* (*viral clearance* en inglés) que es llevada a cabo naturalmente por el sistema inmune, mientras que la eliminación viral también puede ser llevada a cabo por diversas técnicas empleadas como tratamiento para remover la "contaminación viral" artificialmente. Algunas técnicas incluyen el uso de membranas especializadas, cromatografía o plasmaferesis; o bien, antivirales. De Boer y colaboradores elaboraron recientemente (2010) un modelo matemático que estima las tasas de limpieza del virus HIV-1 en varios órganos, incluyendo sangre y tejido linfático. Los autores encuentran que entre 10 y 100 partículas virales individuales son eliminadas rápidamente (en cuestión de minutos) del tejido linfático cuando la producción viral de una célula infectada productivamente excede 10^3 partículas, de modo que en menos de dos horas de puede reducir a la mitad la carga total de la población viral libre en sangre en un paciente infectado crónicamente con HIV-1.

2) **La presencia de un reservorio viral.** Gran parte de la población viral de HIV-1 que es producida en las etapas (i) y (ii) se encuentra atrapada en los tejidos linfoides, típicamente en las superficies membranales de las células dendríticas foliculares, donde los virus pueden sobrevivir por muchos meses, e incluso años en humanos (Keele et al., 2008). Este hecho ha sugerido que la mayor parte de la N_e de HIV-1 se encuentra resguardada en las células dendríticas foliculares, y que sólo una pequeña parte se encuentra libre en sangre. De

modo que la producción y la eliminación viral podría estar sesgada así por un intercambio equilibrado entre las partículas virales libres y aquellas resguardadas en el tejido linfoide (Hlavacek, et al., 1999, 2000, 2002; Müller et al., 2001 en: De Boer et al., 2010). Sin embargo, bajo el mismo estudio realizado por De Boer y colegas (2010) se encontró que la tasa de intercambio de las partículas virales del tejido linfoide hacia la sangre es baja, así como también que la eliminación viral tanto del tejido linfoide y de la sangre ocurre a una tasa mayor que el intercambio entre ambos órganos y que la muerte de las células infectadas productivamente. De modo que la rápida eliminación viral (que no afecta el tiempo de vida de las células infectadas productivamente) se debe principalmente a una infección no productiva, debido en parte a la reducida población de células T $CD4^+$ ($\sim 10^{11}$). Del mismo modo, los autores también apoyan la idea de que la perseverancia de un pequeño reservorio de HIV-1 en pacientes tratados con antivirales se puede explicar por la presencia de un mayor número de ligandos entre los virus restantes y los receptores de las células dendríticas foliculares (Hlavacek, et al., 1999, en: De Boer et al., 2010).

3) **Modo de infección.** Aún cuando la infectividad de HIV-1 durante la transmisión en cultivos celulares es bastante alta (1 unidad infecciosa por cada 10 partículas virales), es de hecho muy baja comparada con la infectividad de otros virus (e.g., se ha reportado el doble en el virus Semliki Forest) (Helenius et al., 1989). En un estudio de dinámica cinética realizado con datos de la acumulación de HIV-1 en los sobrenadantes del cultivo celular después de múltiples rondas de infección, Dimitrov y colaboradores (1993) encontraron que la *infectividad de HIV-1 durante la transmisión célula-célula* es entre 10^2 y 10^3 veces más grande que la *infectividad por las reservas de virus libres derivadas de las mismas células que inician la infección 'de novo'*. En consecuencia, el efecto de la interferencia viral sobre la cinética de la infección observada en la dispersión vía célula-célula es relativamente menor. En individuos infectados, la interferencia viral durante la transmisión célula-célula parece que no tiene un papel significativo en el tejido linfoide debido a la alta densidad y proximidad de los virus producidos por los linfocitos. Caso contrario ocurren en las células T $CD4^+$ susceptibles que se encuentran en circulación generalmente en el torrente sanguíneo, donde las partículas virales no infectivas y/o otros componentes de HIV-1 que no están asociados a las partículas virales, como la glicoproteína gp120, pueden bloquear la

infección. En un paciente infectado crónicamente con HIV-1, es importante tomar en cuenta además que la mayor parte de la población viral que se produce en el tejido linfoide se encuentra en el torrente sanguíneo, por lo que la interferencia viral es mayor para llevar a cabo una infectividad productiva (Dimitrov et al., 1993, De Boer et al., 2010).

Al revisar brevemente, pero de forma completa, los procesos genéticos, ecológicos, evolutivos e infecciosos que definen el tamaño efectivo de la población (N_e) de HIV-1, así como los procedimientos metodológicos que lo estiman, parece acertado suponer que la N_e de HIV-1 se encuentra entre 10^3 y 10^5 partículas infectivas en un paciente infectado crónicamente con este virus a lo largo de varios años. De esta forma, se decidió usar un N_e de 10^3 individuos por ciclo generacional (que permanece constante) para nuestro algoritmo genético dado que el tiempo de ejecución de nuestro algoritmo genético depende cuadráticamente del tamaño del genoma, del N_e y del número de generaciones utilizados. Por ejemplo, para ejecutar nuestro algoritmo genético con una N_e de 10,000 individuos y 10,000 generaciones (que se discutirá a continuación) se requerirían entre 12 y 18 meses de ejecución.

Las estimaciones hechas en el presente trabajo, sólo incorporan de un 1% al 10% de N_e del virus de HIV-1, si se toma en cuenta que $N_e = 10^4$, o si $N_e = 10^3$, respectivamente. Estos valores se consideran cercanos a representar la dinámica evolutiva de las poblaciones naturales de HIV-1. Pero en caso de que N_e sea mayor, $N_e = 10^5$, se estaría usando sólo el 0.1% de la población reproductiva eficientemente. Así, un N_e pequeño puede generar efectos evolutivos sesgados por un proceso de deriva genética. Por el contrario, si se considera que $N_e = 10^2$, se estaría implementando el 100% de *Ne* de HIV-1 en nuestras simulaciones (una estimación similar a lo calculado por Leigh, 1997); sin embargo, es poco probable que el tamaño efectivo de la población del virus HIV-1 sea tan pequeña en la naturaleza.

D. Sobre el tiempo generacional del virus HIV-1

El tiempo generacional de un virus se puede definir como el tiempo entre dos rondas de producción de la progenie viral, incluyendo el tiempo necesario para que los viriones

encuentren una célula huésped susceptible, seguido por adsorción y la infección de la célula hospedera, replicación viral y luego liberación. Además de poseer una gran variabilidad, las poblaciones virales con genomas de RNA se caracterizan por presentar tiempos generacionales cortos, de 2.6 días en promedio (Perelson et al., 1996), lo cual sugiere que estos virus tienen poblaciones muy grandes y tamaños reproductivos eficientes de la población. Dos reportes han realizado una estimación del número de ciclos de replicación por año para HIV-1, el primer reporte fue realizado por Coffin (1995), y el segundo lo realizaron Perelson et al. (1996).

Coffin observó los resultados de Daar et al. (1991) y de Clark et al. (1991) donde se muestra la producción de viriones en cuatro diferentes pacientes al inicio de una infección por HIV-1. En los reportes de Daar et al. y de Clark et al. se cuantifica la presencia de HIV-1 en DNA (por extracción de DNA y amplificación por PCR) en células infectadas por medio de la concentración de antígenos-anticuerpos durante la etapa de "seroconversión" (fase de una infección en la que los anticuerpos frente al agente infeccioso que la causa son detectados por primera vez, ya que hay mayor especificidad). Así, el número de virus producidos por células se reportó de alrededor ~10,000 copias las primeras dos semanas. Sin embargo, hubo un decaimiento espontáneo del número de copias de 100 veces su valor; así se mantuvo el resto del tiempo del experimento (55 días). Este efecto espontáneo ocurrió sin la presencia de antivirales y en las muestras de los cuatro pacientes. En ambos estudios, se argumenta que este alto nivel transitorio de HIV-1 posiblemente sea el reflejo de una etapa aguda de la infección (Daar et al., 1991, Clark et al., 1991). A partir de estos resultados, Coffin realiza una estimación teórica sobre el tiempo de vida de una célula infectada y el ciclo de replicación del virus basado en el hecho de que la reducción de la población viral fue casi de un 99%, de forma que el 1% restante se encargaría de producir a la progenie potencialmente infectiva. Coffin estima que el tiempo de vida de una célula infectada es de 1 a 2 días y, el ciclo de replicación debe ser menor a 1 día. Basándonos en esta estimación, el virus HIV-1 tendría 300 generaciones en un año.

Por otro lado, Perelson y colaboradores (1996) realizaron un modelo matemático basado en la aplicación de inhibidores de RT para calcular la carga viral producida por células en cinco pacientes. Ellos asumen en ese trabajo que HIV-1 infecta de forma constante (i.e., células

infectadas productivamente) y que el antiviral tiene un efecto del 100% (condiciones ideales). Basados en la obtención del número de viriones detectados en plasma, la tasa de eliminación del virus (*clearance*) y la tasa de pérdida de las células infectadas, los autores obtienen el lapso de vida de las células infectadas y el promedio del tiempo generacional *in vivo*. Ellos proponen que al sumar los inversos de la tasa de eliminación del virus (3.07 días) y de la tasa de pérdida de células infectadas (0.49 días), se obtienen los lapsos promedio de vida libre de un virión y de una célula infectada en días, los cuales reflejarían la duración del ciclo de vida del virus HIV-1. El valor para esta estimación es de 2.6 días, lo cual sugiere que habría ~140 generaciones del virus HIV-1 cada año.

Apoyados en estas propuestas y estimaciones, se puede realizar un cálculo de cuantos ciclos de replicación actualmente ha realizado el virus HIV-1 desde su descubrimiento. Un valor que también necesitamos estimar para nuestro algoritmo genético. Registros oficiales del Reporte Semanal de Morbilidad y Mortalidad (MMWR, 2001) han documentado que el primer caso de HIV-1 se registró en junio de 1981, por lo que han transcurrido aproximadamente 31 años hasta el presente año (2012). Asumiendo que el tiempo generacional de HIV-1 se da cada 2 días (promediando los estimados de Coffin y Daar), en un año se realizarían 180 generaciones. Esta estimación sugeriría que el virus de HIV-1 ha generado 5,580 generaciones desde 1981 hasta 2012. Por otro lado, si se deseara modelar la dinámica evolutiva del virus de HIV-1 desde su origen en África (1960) (Pickrell, 2006), o desde su arribo a América (1966) (Gilbert et al., 2007), entonces habrían transcurrido ya 52 años y el virus de HIV-1 tendría por lo menos 9,360 generaciones.

Debido a las limitaciones del tiempo de ejecución de nuestro algoritmo mencionadas anteriormente, se decidió dejar correr la dinámica por 2,000 generaciones, lo que corresponde a tan sólo al 31% del número de generaciones de HIV-1 desde hace 31 años que se identificó el primer caso infeccioso en humanos y 21% del número de generaciones desde hace 52 años en que se estima que el HIV-1 se originó en África.

II. Parámetros biológicos para el programa quasispecies.pl: el *fitness mutacional*

Uno de los factores clave sobre la teoría de las cuasispecies y de la evolución de los virus es el reconocimiento del tipo de mecanismos evolutivos que tienen a lugar en estos sistemas biológicos. Eigen y Schuster asumían inicialmente que cualquier nueva mutación en el genoma de los virus tendría una carga negativa que reduciría el *fitness* de las cuasispecies. Bajo la teoría de las cuasispecies, así, una selección purificadora sería la fuerza dominante en la evolución viral. Se pueden encontrar en la literatura científica diversos estudios que muestran la existencia de mutaciones positivas. Es por ello que uno de los objetivos más importantes de esta tesis fue la evaluación del tipo de selección en los codones y regiones funcionales de todo el genoma del virus HIV-1. Los resultados de tales observaciones se trataron con nuestro algoritmo genético para evaluar el peso evolutivo de las mutaciones que se realizaron en las progenies virales modeladas.

Usando como referencia el genoma completo del virus HIV-1 cepa HXB2, en la Figura R1 se puede observar la caracterización de 11 regiones funciones a lo largo de los 9 genes y las 2 regiones LTR del virus HIV-1 [ver Métodos]. Del mismo modo, se depuraron y editaron manualmente los alineamientos de 124 genomas no redundantes que fueron secuenciados en 35 países para obtener la colección completa de los tripletes de nucleótidos que codifican para proteínas. Se obtuvieron un total de 3,240 codones, y su distribución a lo largo de los genes del genoma del virus HIV-1 HXB2 (que tiene una longitud de 9,719 nucleótidos) se puede observar en la Tabla R2. Se realizaron dos análisis diferentes para evaluar la influencia de la selección en el genoma del virus HIV-1; mientras que el primero se enfocó en la secuencia nucleotídica de los genes y de las regiones funcionales al interior de cada gen, el segundo se enfocó exclusivamente en los codones. La presencia, ausencia y tipo de selección se analizaron así con dos metodologías independientes. La selección en los genes y las regiones funcionales sólo fue posible realizarla con la *prueba de neutralidad de Tajima*, mientras que el análisis de codones se realizó con la *prueba de neutralidad de Tajima* y la *hipótesis de selección por sustituciones sinónimas y no sinónimas empleando métodos filogenéticos de máxima verosimilitud*. La congruencia y robustez de las estimaciones realizadas por ambas pruebas fue analizada posteriormente [ver Métodos].

Tabla R2. Caracterización del genoma de HIV-1 en genes, LTRs, codones y regiones/dominios funcionales.

Gene	Codons (HXB2)[1]	Codons (alineamientos)[2]	Nivel de conservación[3]	Localización en el genoma[4]	No. Regiones functionales[5]	Descripción (Regiones Funcionales)[6]
5'LTR	268 (806 sitios)	232 (697 sitios)	368 sitios (53%)	1 – 634	6	**$W^{\blacklozenge}_1$:** U3, TAR, R, U5, PB y curva Psi
Gag	501	417	157 sitios (38%)	790 – 2292	3	**W_1:** Matriz (p17); **W_2:** cápside p24; **W_3:** cápside (p24) **W_4:** nucleocápside (p7, p6, p2, p1).
*Gag-Pol**	*1435*	18	23 sitios (42%)**	*nd*	2	**W_1:** p7 y proteasa (región sobrepuesta)
Pol	1003	895	1183 sitios (44%)	2085 – 5096	4	**W_1:** Proteasa, RT (p51); **W_2:** RT (p51); **W_3:** RNasa (p15), integrasa; **W_4:** integrasa
Vif	193	139	147 sitios (35%)	5041 – 5619	5	**W_1:** APOBEC, **W_2:** Unión a RNA, **W_3:** multimerización, **$W_{4,5}$:** Unión a la célula, asociación a membrana
Vpr	97	59	75 sitios (37%)	5559 – 5850	1	**W_1:** p10-15 **(**Homo-oligomerización**)** **W_2:** *nd*
Tat	101	85	75 sitios (29%)	CDS1: 5831-6045 CDS2: 8379 – 8469	2	**$W_{1,2,3}$:** p14 (empalme inicial a mRNA:exon1) **W_4:** p14 y p16 (empalme inicial mRNA:exon2)
Rev	116	103	78 sitios (25%)	CDS1: 5970-6045 CDS2: 8379 – 8653	3	**W_1:** Homo-multimerización, **W_2:** señal de exportación nuclear **W_3:** localización **$W_{3,4}$:** Poli-Arg
Vpu	83	56	36 sitios (21%)	6062 – 6310	2	**W_1:** Extracelular; **$W_{2,3}$:** región citoplasmática
Env	857	680	556 sitios (27%)	6225 – 8795	3	**$W_{1,2}$:** SU gp120 (extracelular), **W_3:** SU gp120 (extracelular), TM gp41 (región intracelular), y RRE **W_4:** TM gp41 (región intracelular)
Nef	124	106	104 sitios (33%)	8797 – 9417	3	**W_1:** MHC internalización; **W_2:** Dominio SH-3 CD4-internalization
3' LTR	100 (302 sitios)	76 (228 sitios)	68 sitios (30%)	9086 – 9719	6	**W_1:** Estructurado como 5'LTR
Total (Σ)* **	3526 (100%)	2848 (~81%)	35%	12 localizaciones	34	

[1] Numero de codones localizados en el genoma de referencia HXB2. [2] Numero de codones obtenidos de nuestro alineamiento final en HIV-1. [3] La conservación fue calculada considerando el número de nucleótidos invariables y el número total de nucleótidos en el alineamiento. [4] Puntos de referencia de genes dentro del genoma viral *HXB2*. [5] Número de regiones funcionales para cada gen. *nd*, no datos. ♦ Número de ventanas (Wn) localizadas en la Figura R2 y análisis por ventanas de regiones funcionales (ver serie de figuras en Figura S2); *n* significa la posición de la ventana. * Región sobrepuesta (secuencia y alineamiento original). ** Estimado basado en el número de sitios de nuestro alineamiento (54 sitios). *** Resumen total de todas las características.

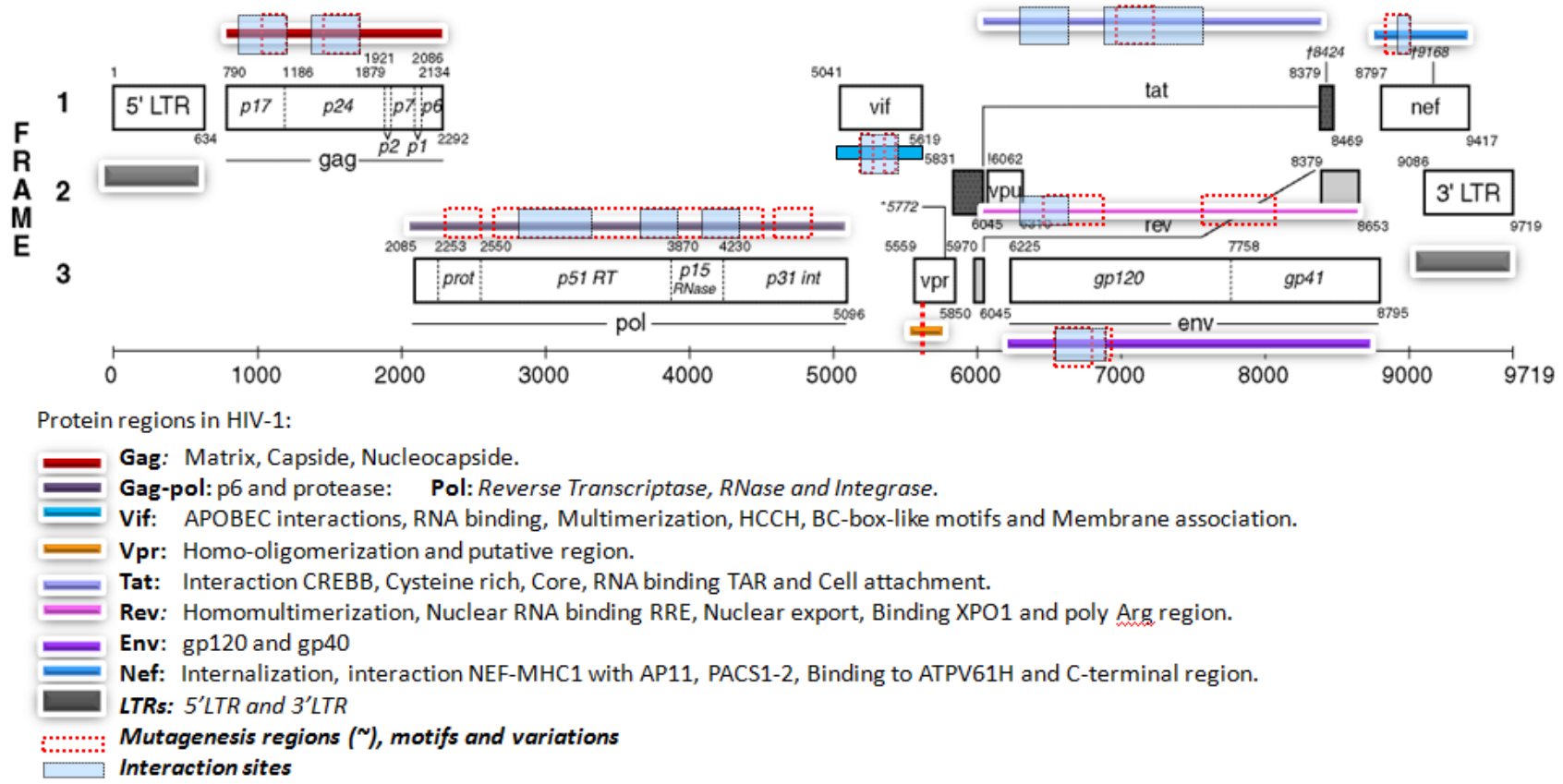


Figura R1. Mapa del genoma de HIV-1 HXB2. Los *genes y los LTRs* se muestran como rectángulos en blanco y negro (en FRAME 1, 2 y 3). El inicio del gen es indicado por el número en la esquina superior izquierda de cada rectángulo (ATG), mientras que el número en el registro inferior derecha representa la posición del codón de paro (STOP). Los rectángulos de color son la representación de la proteína de cada una de las regiones funcionales de HIV-1. Asimismo, se muestran los sitios de interacción unión con proteínas y moléculas de RNA (rectángulos pequeños azul claro)**.** Se presentan en recuadros de color rojo (línea punteada) todas las mutaciones experimentales reportadas que confieren resistencia a antivirales o que inducen un efecto letal sobre la replicación viral. Las posiciones de numeración para cada gen en relación con el genoma de referencia HIV-1 HXB2, se basaron en el *Compendio de bases de datos Los Alamos*: Human Retroviruses and AIDS, 1998 (Korber B, 1998). Para ver a detalle cada una de las posiciones dentro del genoma de HIV-1 ver Tabla S1.

A. Selección en los genes, regiones funcionales y codones del virus HIV-1

Selección en los genes. La *prueba de neutralidad de Tajima* se basa en el polimorfismo de las secuencias (θ y π), detectando valores positivos ($D > 0$), negativos ($D < 0$) o neutros ($D = 0$), que sugerirían la influencia de la selección positiva, negativa o neutra, respectivamente. Los valores del estimador D de Tajima para cada uno de los 9 genes y las dos regiones LTRs del virus de HIV-1 se pueden observar en la Tabla R3. La Figura R2 muestra (para nuestra sorpresa) una estructura genómica sujeta a dos presiones de selección: la primera parte del genoma, conformada por los genes ***5´LTR***, ***gag***, ***gag-pol***, ***pol***, ***vif*** y ***vpr***, se encuentra potencialmente sujeta a la influencia dominante de la **selección negativa**; mientras que la segunda mitad, conformada por los genes ***tat***, ***rev***, ***vpu***, ***env***, ***nef*** y ***3'LTR***, se encuentra potencialmente sujeta a la acción principal de la **selección positiva**. Sin embargo, los valores de la D de Tajima no fueron estadísticamente significativos ($p > 0.05$) sobre el análisis de los nucleótidos en los genes y las regiones de LTRs del virus de HIV-1. El análisis por codones, que se describirá más adelante, sí tuvo significancia estadística.

Tabla R3. Estadísticas del test de Tajima para los genes del virus HIV-1. Los valores estadísticos se obtuvieron en el análisis de cada región funcional del genoma de HIV-1. Los estimadores, θ y π, son inversamente proporcionales, si $\theta > \pi$, se presenta la selección negativa y, si $\theta < \pi$, hay selección positiva.

Tipo de selección	Reg. Func.	Pruebas Estadística			
		D de Tajima	θ	π	Valor de P*
Negativa	5'LTR	-0.1861	0.11731	0.1071	P > 0.10 *n.s.*
	Gag	-0.2638	0.10749	0.0989	P < 0.10 *n.s.*
	Gag-Pol	-0.5728	0.10411	0.0863	P < 0.10 *n.s.*
	Pol	-0.5702	0.10388	0.0861	P < 0.10 *n.s.*
	Vif	-0.0343	0.12045	0.1192	P > 0.10 *n.s.*
	Vpr	-0.1165	0.11769	0.1117	P > 0.10 *n.s.*
Positiva	Tat	0.7263	0.13088	0.1598	P > 0.10 *n.s.*
	Rev	0.5341	0.13846	0.1608	P > 0.10 *n.s.*
	Vpu	0.2138	0.14529	0.1548	P > 0.10 *n.s.*
	Env	0.0977	0.13512	0.139	P > 0.10 *n.s.*
	Nef	0.2195	0.12733	0.1357	P > 0.10 *n.s.*
	3'LTR	0.7126	0.13313	0.1618	P > 0.10 *n.s.*

* *n.s.*, no hay un valor significativo. Hay varios sitios con valor significativo de P (P < 0.01, 0.05) en los codones de nuestro alineamiento.

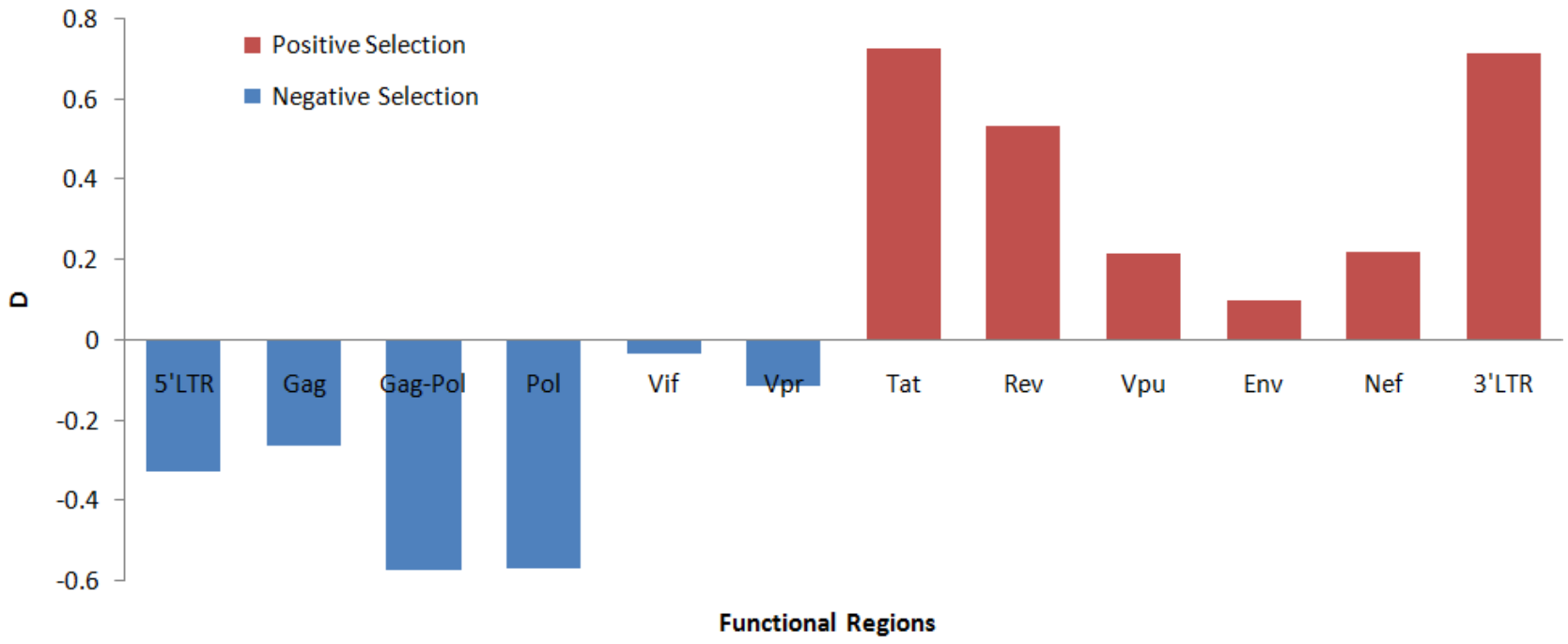


Figure R2. Detección de la selección en los genes y LTRs del genoma del virus HIV-1. En genoma del virus HIV-1 se divide en dos secciones según la influencia del tipo de selección que actúa sobre los genes y LTRs. Ambas secciones están conformadas por 6 regiones funcionales (11 + 1: Gag-Pol, región solapada). Cada gen y LTR presenta un sólo valor de la D de Tajima, representando todas aquellas subregiones que la componen [ver Figura R1]. Nuestras predicciones muestran, en general, que la primera mitad del genoma de HIV-1 está sujeta a selección negativa y que la otra mitad se encuentra bajo selección positiva ($p>0.05$). Los valores estadísticos para cada región funcional se encuentran en la Tabla R3. Una descripción más detallada sobre la selección de las subregiones para cada región funcional se muestra en la Figura R3, en la que se observa, el general, la misma distribución.

Selección en dominios funcionales. Dado que diferentes regiones de un mismo gen pueden estar sujetas a la acción de diferentes tipos de selección, e incluso a su ausencia, realizamos el análisis de la prueba de neutralidad de Tajima en las diferentes subregiones funcionales que conforman a cada gen del virus HIV-1. La caracterización de las subregiones funcionales (o dominios) se realizó con base a la identificación de dominios y motivos funcionales [ver métodos]. En la Figura R2 y R3 se puede observar así que se presenta en general la misma tendencia selectiva que la observada con los genes: la primera mitad del genoma, conformada por las regiones ***5´LTR***, ***gag***, ***gag-pol***, ***pol***, ***vif*** y ***vpr***, se encuentra sujeta potencialmente a la **selección negativa**; mientras que la segunda mitad del genoma, compuesta por ***tat***, ***rev***, ***vpu***, ***env***, ***nef*** y ***3'LTR***, se encuentra potencialmente sujeta a la **selección positiva**.

Sin embargo, la división por dominios funcionales permite ver que hay algunas regiones al interior de cada gen que siguen la tendencia selectiva contraria o la neutralidad. Dentro de la primera región del genoma de HIV-1 potencialmente sujeta a selección negativa se encuentran: (1) uno de los dos dominios *cápside* del gen ***gag*** (tercera ventana) muestra neutralidad; (2) dos de las cinco subregiones del gen ***vif***, como *la región de unión a RNA* (una de las 4 regiones de *interacción con el domino 3FF2 box de la proteína APOBEC)* y parte del *motivo HCCH* presentan selección positiva (tercera ventana); finalmente (3) *la región de asociación a la membrana celular* y *multimerización* del mismo gen (quinta ventana) se encuentra posiblemente sujeta a evolución neutral. Mientras que en la segunda mitad del genoma de HIV-1 sujeta potencialmente a la selección positiva, muestran una tendencia selectiva contraria: (4) los dominios *asociados al citoplasma* (tercera ventana) del gen ***vpu*** muestran selección negativa; (5) los dominios del gen ***env*** asociadas a la *región extracelular* y la *región del péptido fusión e inmunosupresión* (tercera ventana) se encuentran potencialmente sujetos a la selección negativa, así como también (6) los *motivos de interacción con el hospedero* que participan en procesos de unión e internalización (cuarta ventana) del gen ***nef***. En este mismo gen, se encuentra un dominio bajo evolución neutral potencialmente, el cual posee funciones de *internalización* y otros *sitios de interacción* al hospedero (tercera ventana).

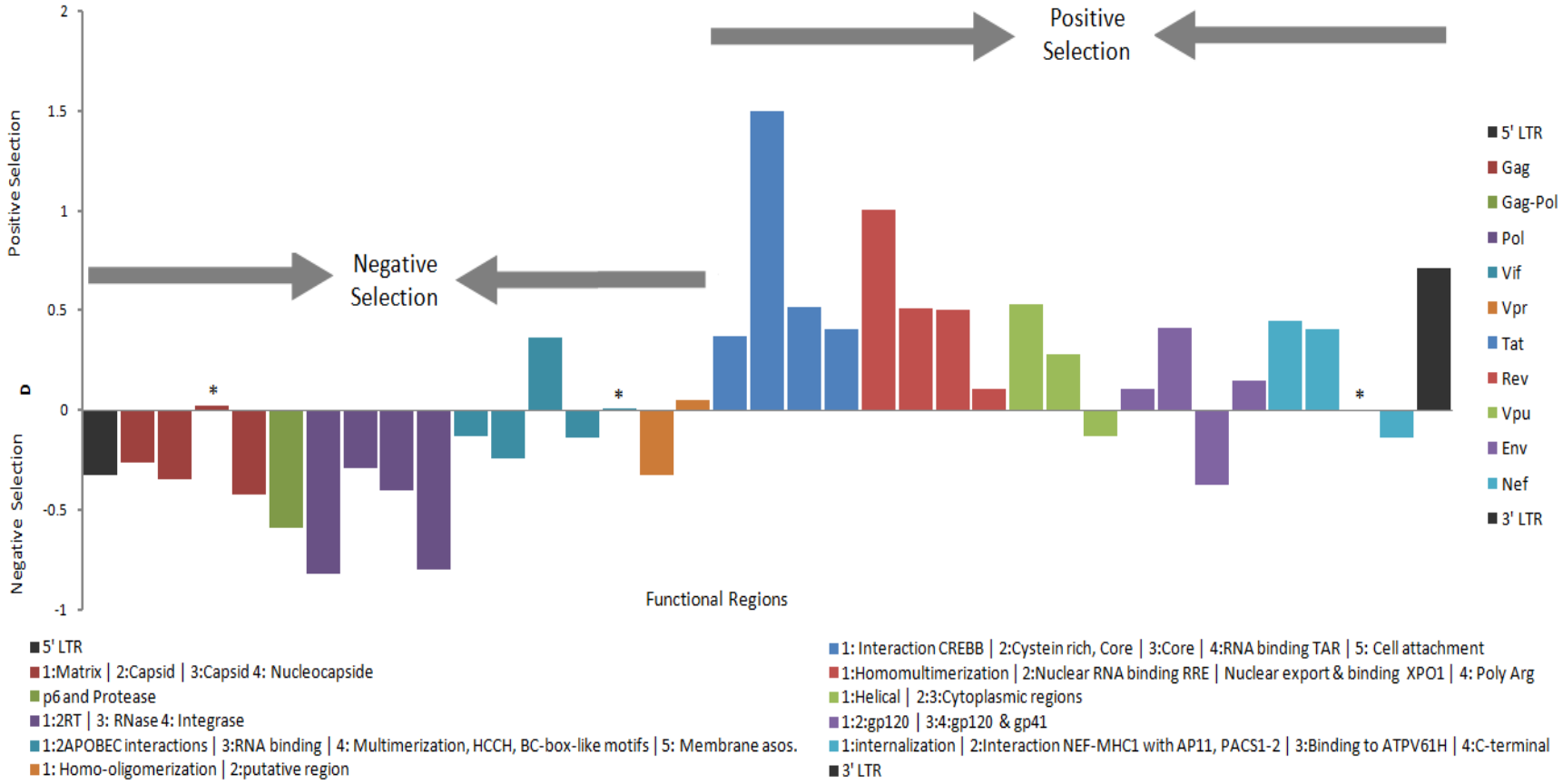


Figura R3. Evaluación de la selección por ventanas para cada región funcional en el genoma de HIV-1. *En la grafica:* se muestran las diferentes *regiones funcionales* de HIV-1 divididas en diferente número de ventanas (cada barra con el color correspondiente a la región funcional, ver el nombre en el lado superior derecho). Cada ventana representa diferente número y tipo de función para las *subregiones* que la componen (ver parte inferior, cuadros con su correspondiente color). Las primeras seis regiones funcionales (*5´LTR*, *Gag*, *Gag-Pol*, *Pol*, *Vif* y *Vpr*) presentan *selección positiva*. Mientras que los seis restantes (*Tat*, *Rev*, *Vpu*, *Env*, *Nef* y *3'LTR*) tienen selección negativa. (*) Tres subregiones presentan selección neutral. El número de ventanas correspondiente a la regiones funcionales son: *5'LTR: 1; Gag: 3; Gag-Pol: 2; Pol: 3; Vif: 6; Vpr: 2; Tat: 5; Rev: 5; Vpu: 2; Env: 3; Nef: 4 y 3'LTR:1.*

Selección por codones. Para poder identificar presiones de selección *a nivel de codón* a lo largo del genoma de referencia, se evaluó la selección a través de dos métodos independientes. El primero, conocido como la *prueba de neutralidad de Tajima* (usada previamente en el análisis de genes y por ventanas/regiones funcionales), se basa en el polimorfismo de las secuencias (θ y π) y, calcula la diferencia de sus estimadores para detectar a la selección positiva, negativa o neutral. Mientras que el segundo, usa probabilidad máxima (del inglés *máximum likelihood*) para la reconstrucción de secuencias ancestrales y calcula la sustitución de sitios sinónimos (dS) y no sinónimos (dN) en base a las relaciones filogenéticas. Este segundo método evalúa la acción de la selección positiva si dN > dS, y de la selección negativa si dN < dS, o evolución neutral si no hay variación estadística entre dN y dS, i.e., dS ~ dN. Para evaluar este segundo método se usaron 3 aproximaciones independientes que utilizan métodos de máxima verosimilitud: *SLAC, FEL* e *IFEL,* para obtener la aproximación que nos permitiera las predicciones más robustas [ver Material y Métodos]. En la Figura R5 y R6 se pueden observar los resultados de ambos métodos de predicción de selección a lo largo de los codones del genoma del virus HIV-1

La prueba de Tajima identificó 1,606 codones negativos, 654 neutrales y 980 positivos, los cuales representan el 50%, 20% y 30% del genoma, respectivamente. Sin embargo, los resultados no fueron estadísticamente significativos ($p>0.05$). Los valores de la D estadística de la prueba de Tajima muestran un valor máximo negativo de *-1.5*, un máximo positivo de *4.5* y la moda fue de *0*. Parte de nuestro problema para obtener predicciones de selección y neutralidad estadísticamente significativos posiblemente radica en que los estimadores π y θ no presentan suficiente varianza ($p>0.10$). Se sabe que π es sensible al número de mutaciones y que θ lo es para las frecuencias. Es posible que en la muestra, el número de mutaciones y/o la frecuencia de éstas no sean suficientes para poder rechazar estadísticamente la hipótesis nula. No obstante, sí obtuvimos un número pequeño de codones que presentan una significancia estadística ($p< 0.05$ y 0.01). Aunque el número es bajo, estos sitios son los que se emplearon para comparar, por medio de la prueba de sensibilidad y especificidad, los métodos de *dS* y *dN*, y la prueba de Tajima contra los datos experimentales [ver métodos].

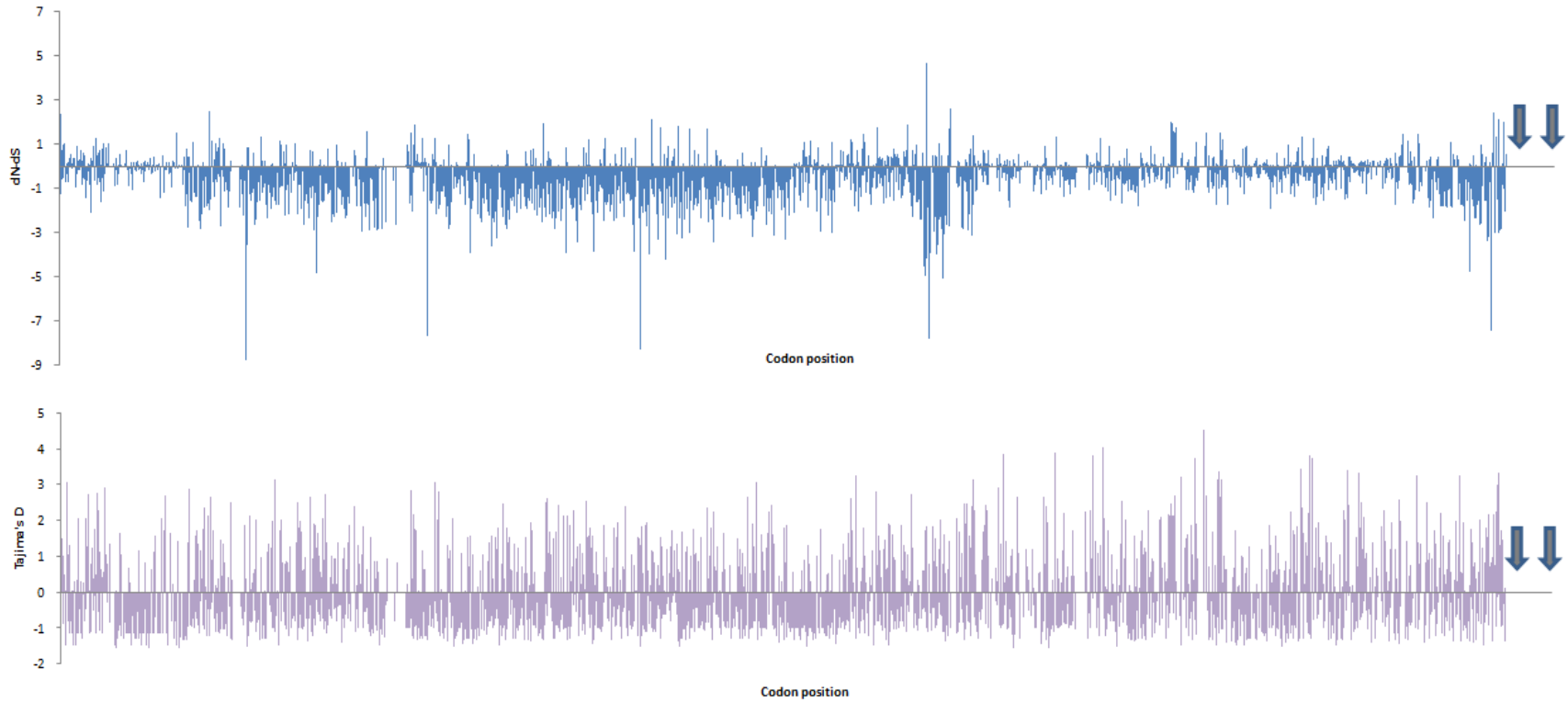


Figura R4. Representación grafica de la estimación de la selección por codones a lo largo del genoma de HIV-1. Para ambos métodos, se muestra la intensidad y distribución del tipo de selección que se estima para los 9,719 sitios que conforman al genoma de HIV-1. En ambas gráficas, se puede observar el tipo de selección a la que están sujetos los codones que constituyen al genoma viral, *i.e.* los valores positivos (por encima del cero), representan *selección positiva*, lo valores negativos (por debajo del cero) significan *selección negativa* y, los valores que son cero, significan *selección neutral*. En la parte superior, se presentan los valores obtenidos de la estimación de la estimación de dS y dN a través del programa *HyPhy v1.0beta*; todos los sitios presentan un alto grado de confianza ($p<0.05$). En la parte inferior, se muestran los valores obtenidos a través de la prueba de Tajima, basada en el análisis de polimorfismos de θ y π, obtenidos por medio del programa *DnaSP v5.10.01*; varios de los valores para estos sitios no son estadísticamente significativos ($p>0.05$). Para ambos métodos se muestra en la parte final derecha (flechas), el fragmento indefinido que corresponde al fragmentos final de la región funcional 3'LTR.

Al realizar las pruebas para detectar selección en el gen *vif* por medio de la estimación de *dS* y *dN*, se observó que los tres métodos, *SLAC*, *FEL* e *IFEL* [ver Material y Métodos], no presentaron diferencias significativas sobre la detección del número aproximado de codones positivos, negativos y neutrales [ver Tabla R4]. Sin embargo, la intensidad de la selección de los codones analizados sí varía significativamente entre los tres métodos, siendo el método SLAC el que muestra una mayor intensidad de la selección predicha [ver Figura R5]. Adicionalmente, se compararon los métodos contra sí mismos usando diferentes modelos de evolución de nucleótidos: *best fitting models* (012312) *versus* HKY85 (010010). Los resultados de estas comparaciones tampoco muestran una diferencia significativa entre los tres métodos [ver Figura S4-B, C y D en material suplementario]. De esta forma, se empleó el método SLAC para realizar la evaluación de todos los codones del genoma del virus HIV-1, ya que se ha determinado como el más sensible al número de muestras y el método más conservador en el análisis de secuencias (Kosakovsky-Pond and Frost, 2005, Poon et al., 2009, Delport et al., 2010). A diferencia de la baja significancia estadística obtenida con la prueba de Tajima, las predicciones de selección en los codones del genoma del virus HIV-1 estimados con el método SLAC son estadísticamente significativos ($p<0.05$).

Tabla R4. Comparación de los tres métodos que evalúan la presencia de la selección positiva, negativa o neutral para cada codón presente en el genoma de HIV-1 basados en *dS y dN*. La comparación se realizó bajo el uso de 2 modelos de sustitución diferente. Los codones detectados aquí, tienen un peso estadístico significativo.

	Pruebas de selección (by codons) en el gen VIF*		
MODEL	SLAC	FEL	IFEL
"012312"(a)	14 positivas 49 negativas 76 neutrales	14 positivas 55 negativas 69 neutrales	16 positivas 33 negativas 90 neutrales
"010010"(b)	12 positivas 49 negativas 78 neutrales	12 positivas 54 negativas 73 neutrales	16 positivas 33 negativas 90 neutrales

* Nivel de significancia por codón detectado: p-value < 0.05.

(a) Modelo de substitución a través de la "Prueba de comparación de modelos": Mejor modelo obtenido con in AIC de 22405.6

(b) Modelo de substitución: HKY85

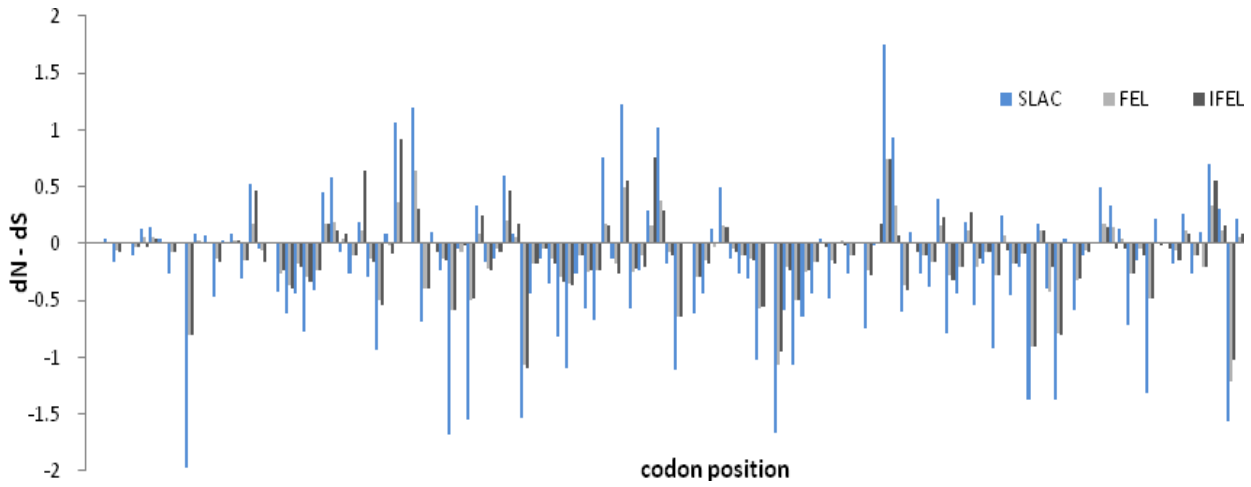


Figura R5. Gráficos comparativos de la prueba de selección (valores normalizados). Comparación de los tres métodos SLAC, FEL e IFEL. Para la comparación de cada uno de método independiente, hemos utilizado la herramienta de comparación de modelos (un modelo de codón) usando un paquete de *HyPhy*. Comparamos los mejores modelos de ajuste "012312" obtenidos con el modelo HKY85 (010.010).

Al igual que la prueba de Tajima, el método SLAC detectó la influencia de la selección en el genoma de HIV-1 con proporciones similares: el 58% de los codones se encuentra potencialmente bajo selección negativa, el 21% bajo la selección positiva, y un mismo porcentaje de los codones (21%) no presenta ningún tipo de selección, por lo que se sugiere que evolucionan neutralmente [ver Figura R6]. Sin embargo, es importante aclarar que las posiciones de nucleótidos que se encuentran perfectamente conservadas en los alineamientos de los genes del virus HIV-1 son identificadas, tal vez erróneamente, como posiciones nucleotídicas que se encuentran bajo evolución neutral por ambos métodos, ya que no hay diferencia estadística alguna entre los parámetros θ y π en Tajima, ni entre *dS* y *dN* en SLAC. Además, aquellas posiciones nucleotídicas que presentaron alguna eliminación o inserción fueron descartadas del análisis de selección por ambos métodos, ya que los programas utilizados descartan automáticamente aquellas columnas, o bien, realizan automáticamente corrimiento del marco de lectura de los codones. Este procedimiento metodológico se realizó particularmente en aquellas regiones con una alta proporción de *gaps* en los alineamientos, tales como las regiones solapadas y en algunas regiones de los LTRs. Por ejemplo, el 3'-LTR conformado por 101 posiciones en el genoma del virus HIV-1 HXB2 presenta diferentes longitudes en los genomas completamente secuenciados. De ahí que se decidió sólo analizar la parte inicial del 3'-LTR que está sobrepuesta con la región *nef*, mientras que la región

conflictiva del LTR fue considerada como neutral [flechas en Figura R4]. Así, existe un sesgo en la estimación de las posiciones que se encuentran evolucionando neutralmente al nivel de aminoácidos, el cual debe ser mucho menor de lo que nosotros hemos predicho aquí.

Tabla R5. Identificación de sitios con efecto selectivo sobre el fitness viral obtenidos experimentalmente en el virus HIV-1. Número de sitios totales (mutaciones) detectados en el genoma de HIV-1 que muestran un efecto sobre el fitness viral. El número reportado de sitios está basado en una identificación experimental.

		Numero de sitios detectados con 'efecto selectivo' de las mutaciones*			
Regiones Funcionales	Codons	Positivo	Negativo	Neutral	Detección Total
5' LTR	232	0	0	0	0
Gag & Pol	18	347	66	0	413
Vif	895	3	50	0	53
Vpr	139	0	13	0	13
Tat	85	0	14	0	14
Rev	103	0	6	0	6
Vpu	65	0	0	0	0
Env	680	0	6	0	6
Nef	106	0	32	0	32
3' LTR▪	76	0	3	0	3
Total de sitios vs. sitios del genoma completo					540[Δ]

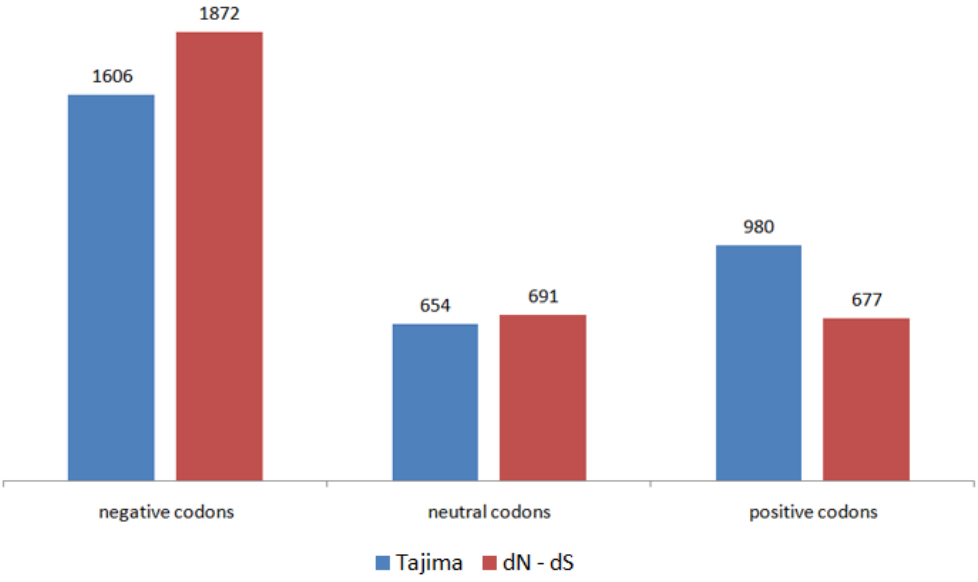
* Corresponde a los sitios identificados a través de reportes experimentales basado sobre esta base de datos (ver métodos). Además, se introdujo el número de sitios conservados detectados en el genoma de HIV-1 (ver tabla R2).

[Δ] Representa el 5.5% de los sitios del genoma completo de HIV-1, con 9719 sitios.

▪ 101 posiciones de la región final 3' LTR son nucleótidos indeterminados debido a que los datos están incompletos (ver Fig. R6).

En la Tabla R5 se muestra la influencia de la selección sobre ciertos codones que han sido obtenidos experimentalmente en los genes *gag*, *gag-pol* y *pol*. Así, se identificaron un total de 540 codones con *efectos ventajosos* (positivos) o *letales* (negativos), de los cuales, 413 se encuentran distribuidos en las regiones funcionales *gag* y *pol*, mientras que las otras 127 posiciones se distribuyen en otros genes [ver Tabla R5 y Tablas S1A-S1L en material suplementario]. Todas las mutaciones que se ha reportado en nuestra búsqueda, se han identificado como variaciones naturales o inducidas por mutagénesis. El número de sitios totales que se reporta aquí, basados en los criterios de búsqueda [ver Material y Métodos], representa alrededor de ~11% del número total de codones del genoma de HIV-1. Estos datos que se tratan como verdaderos positivos (*VP*), fueron usados para calcular la especificidad y sensibilidad de nuestras predicciones. En futuras simulaciones, los codones letales serán incluidos y tomados en cuenta en el *fitness mutacional* de nuestro algoritmo genético para

eliminar mutantes letales de las progenies generadas.



Selection Test*	Parameters*#				Sites*#		Result*	
	TP	FP	TN	FN	nd_t	nd_x	**Sensibility**	**Specificity**
*Tajima**	3	24	0	0	412	39	10%	0
*dS-dN**	228	160	0	0	26	39	60%	0

*Based on our aligned region of gag-pol and pol vs experimental support database (see methods and supplementary material). The results were obtained by the equations expressed in methods section.
TP: true positives; FP: false positives; TN: true negatives; FN: false negatives; nd_t: no detected sites by the selection test; nd_x: no detected sites by experimental reports.

Figura R6. Comparación de los métodos de *Tajima* y *dS y dN:* sensibilidad y especificidad. La mayoría de los codones del genoma de HIV-1, en ambos métodos, presentan selección negativa. En la tabla se muestran los datos numéricos que corresponden a los resultados obtenidos en las pruebas de sensibilidad y especificidad. Un total de 439 sitios conforman a las regiones gag y gag-pol, de las cuales 400 fueron evaluadas. La simbología de los *estimadores* (parámetros) es el siguiente: VP: verdaderos positivos; FP: falsos positivos; VN: verdaderos negativos; FN: los falsos negativos. Ver Métodos para conocer la forma en que se calcularon tales estimadores.

En la tabla de la Figura R6, se pueden observar los resultados de la significancia estadística entre los dos métodos de selección, a partir de los datos obtenidos en la Tabla R5 [ver Material y Métodos]. Se puede observar que la prueba de sustituciones sinónimas y no sinónimas fue más sensible en comparación a la de Tajima. Sin embargo, a ninguna prueba se le detectó especificidad, y esto se explica porque el cálculo de la especificidad requiere del número de codones predichos como *neutrales*, es decir, como "la no detección de la selección" (i.e., verdadero negativo, *VN*) dentro de la colección de datos experimentales compilados en la literatura. A la fecha, sin embargo, no tenemos conocimiento de que existan reportes de identificación de evolución neutral en los codones del genoma del virus HIV-1.

B. Las regiones genómicas de HIV-1 bajo selección purificadora *versus* direccional

En la Figura R7 se muestra un mapa donde se representan las estimaciones del tipo de selección y otras características de los genes del virus HIV-1, tales como la conservación nucleotídica, la estructuración del RNA (con datos obtenidos de Sanjuán y Bordería, 2011), las regiones de los genes solapados (anidados), así como los datos de selección obtenidos experimentalmente que fueron compilados a través de la literatura.

Aproximadamente, el 35% de los nucleótidos y el 67% de los aminoácidos del genoma del virus HIV-1 se encuentran relativamente (>90%) conservados [ver Figura R7 y ver alineamientos en Figura S1A-S1L en material suplementario]. Aunque la variación y la conservación nucleotídica se observan a lo largo del genoma del virus HIV-1, ambas también muestran una tendencia a estar limitadas o concentradas en regiones genómicas y estructurales específicas [ver Figuras R2 y R3]. No es difícil reconocer así que las regiones más conservadas son las que se encuentran sujetas a la selección negativa según nuestros resultados, mientras que las regiones más variables son las que se encuentran sujetas a la selección positiva [ver Figuras R7].

Se ha sugerido recientemente que la conservación y la variabilidad del genoma del virus HIV-1 están correlacionadas principalmente con las restricciones físico-químicas que son impuestas por la estructura secundaria del RNA viral, así como por la presencia de dominios de proteínas con plegamientos del tipo hélice α (Holmes, 2003, Watts et al., 2009, Woo et al., 2010, Snoeck et al., 2011, Sanjuán and Bordería, 2011). De este modo, los aminoácidos que son codificados por RNA genómico, que se encuentra formando estructuras funcionales altamente selectivas, son consistentemente menos variables a lo largo del genoma y evolucionan más lentamente que aquellos aminoácidos para los cuales tal selección al nivel del RNA se encuentra ausente (Sanjuán and Bordería, 2011). Por otro lado, mientras que más variación puede ser localizada en los plegamientos tipo β-plegadas, el plegamiento α-helix es el elemento más estable e importante en la estructura tridimensional de las proteínas (Richardson, 1981).

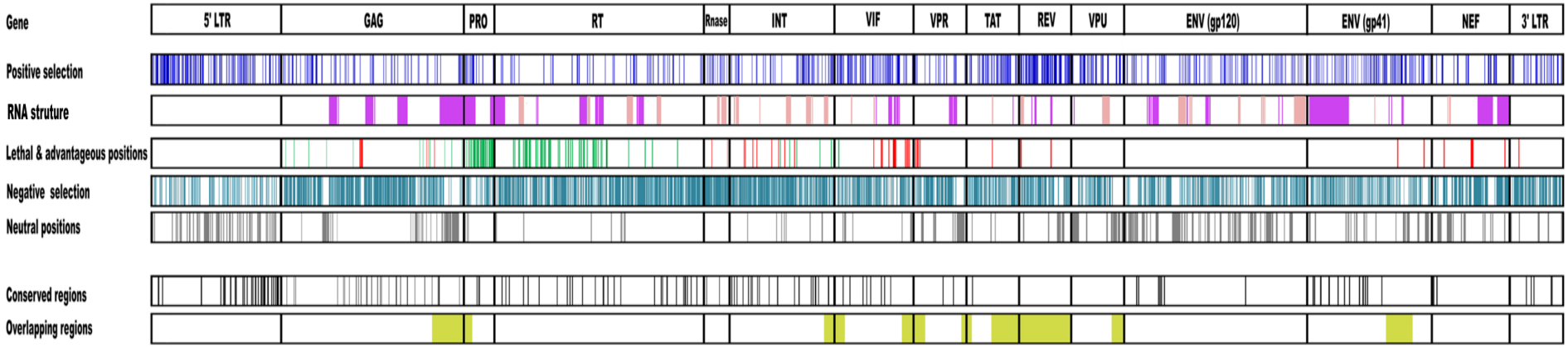


Figura R7. Representación grafica de la estimación de la selección por codones a lo largo del genoma de HIV-1. Mapa del genoma de HIV-1. Para claridad, el genoma es representado en forma lineal, con los genes y LTRs representados como una serie de "bloques concatenados" (barra superior). A continuación, las siguientes barras de datos son descritos: ***Selección positiva***: azul marino = representado por el 19.8% de los codones. ***Estructura de RNA***: rosa mexicano = RNA extensivamente estructurado (parámetro SHAPE < 0.25); rosa claro = RNA flexible (SHAPE > 0.5). Los datos de estructuración del RNA con la prueba SHAPE fueron obtenidos del trabajo realizado por Watts y colaboradores (2009) sobre la estructura secundaria completa del genoma de HIV-1; los datos se encuentran disponibles en: http://www.nature.com/nature/journal/v460/n7256/abs/nature08237.html. ***Posiciones letales y ventajosas*** (identificadas experimentalmente): rojo y verde, respectivamente = 1.7 % son letales y 3.5% son ventajosas. ***Selección negativa***: azul tenue = 54.1% de los codones tienen este tipo de selección. ***Posiciones neutrales***: gris = 13.9% de los codones. ***Codones conservados***: gris oscuro = 5.1% de los codones están conservados (calculado al 100% de conservación). Alrededor de un 85% de los nucleótidos están conservados, mientras que el 78% de los aminoácidos del genoma de HIV-1 se encuentran relativamente conservados (calculado considerando a la conservación >90%). ***Regiones sobrepuestas****:* verde limón = correspondientes a 7 zonas de los diferentes marcos de lectura en los genes de HIV-1: *gag-pol*, *integrasa-vif*, *vif-vpr*, *vpr-tat*, *tat-rev*, *vpu-rev*, y *env*. Los codones que presentan selección positiva, negativa reportados en este mapa, fueron identificados a través del programa *hyphy*, y cada detección de codón fue significativo (p = 0.05).

Los genes y regiones funcionales ***5´LTR***, ***gag***, ***gag-pol***, ***pol***, ***vif*** y ***vpr*** que se encuentra potencialmente bajo selección negativa, de acuerdo con nuestras predicciones, son aquellas regiones genómicas involucradas directamente en la eficiencia del sistema de replicación del genoma viral y su integración al genoma del hospedero. Aunque nosotros hemos encontrado también que ciertos dominios de los genes *gag*, *vif* y *vpr* se encuentran bajo selección positiva, tal como ha sido descrito en reportes previos (Soares et al., 2008, Snoeck et al., 2011). De hecho, se ha reportado que la selección purificadora "enmascara" la flexibilidad mutacional del la RT del virus HIV-1 (Smith et al., 2004). De este modo, la selección purificadora depuraría todas aquellas mutantes que comprometen la función de estos genes, garantizando con ello mantener la función más fundamental del virus HIV-1, esto es, su replicación. Los componentes principales que participan directamente en la replicación del virus HIV-1 (reverso transcripción de RNA→DNA) son los LTRs y los componentes integrados en el gen *pol*. Los *LTRs* tienen, en general, dos funciones, la primera es actuar como un *primer* o cebador para iniciar la retrotranscripción (con la ayuda de un tRNA) y, la segunda es la regulación del proceso de transcripción de proteínas virales. Los componentes que constituyen a *pol* (reverso transcriptasa: RT, proteasa: PR e integrasa: IN), conocidas como proteínas estructurales, son los encargados directos de la reverso transcripción del genoma viral y, a su vez, de la integración de este en el genoma del hospedero. Por otro lado, *vif* promueve la replicación viral *in vivo* al contrarrestar la acción de un factor celular anti-retroviral, APOBEC3G, el cual aumenta la tasa de mutación del virus HIV-1; mientras que *vpr* está involucrado en la importación nuclear de complejos de integración (tal como PIC: RT, PR e IN), la detección del crecimiento celular, la transactivación de genes celulares y la inducción de la diferenciación celular.

Por el otro lado, ha sido reconocida la presencia de presiones de selección positiva que constriñen la evolución del virus de HIV-1, como los generados por el sistema inmune o por los tratamientos antivirales, así como por eventos de cuellos de botella derivados de las transmisiones virales entre hospederos de la misma especie (Ryland et al., 2010, Liu et al., 2011). Estas presiones de selección han resultado en el desarrollo de estrategias adaptativas del virus para mantener una infección y reproducción exitosa. En términos generales, nuestros resultados muestran que las regiones ***tat***, ***rev***, ***vpu***, ***env***, ***nef*** y ***3'LTR*** muestran una fuerte

influencia de la selección positiva. Estas regiones están involucradas en las funciones que garantizan el reconocimiento de la célula hospedera y la interacción con diversos componentes moleculares para llevar a cabo parte de la replicación y la transcripción de los componentes virales y celulares. La acción de la selección positiva en varios de estos genes ya ha sido independientemente descrita en *tat, rev* y *vpu* (de Oliveira et al., 2004, Soares et al., 2008, McNatt et al., 2009), *env* (Wood et al., 2009, Woo et al., 2010, Yoshida et al., 2011b) y *nef* (Price et al., 1997, Zanotto et al., 1999). La función de *tat* es incrementar (más de 100 veces) la tasa de transcripción del genoma viral al interactuar con la RNA polimerasa II del hospedero, *rev* participa en la regulación del transporte de mRNAs virales para que éstos sean traducidos, *vpu*, al igual que *nef*, regulan la degradación de antígenos CD4 presentes en la célula, esto para incrementar el número de virus en la célula y promover la liberación celular. La proteína *env* media la unión y entrada del virión con la célula potencialmente infectable por medio de gp120, y es precursor de la formación de las proteínas de la superficie del virión.

Recientemente, la estructura secundaria del genoma entero del virus HIV-1 fue determinada experimentalmente (Watts et al., 2009) usando la prueba de reactividad SHAPE (del inglés *Selective 2'-hydroxyl Acylation analyzed by Primer Extensión reactivity*), la cual identifica nucleótidos que adoptan conformaciones flexibles o altamente estructuradas. Esta prueba evalúa la flexibilidad bioquímica nucleótido por nucleótido para formar pares de bases en una secuencia, la prueba se basa en la posición 2'hidroxilo de la ribosa, la cual es fuertemente asociada a la flexibilidad local del nucleótido. Los datos obtenidos del trabajo de Watts et al. (2009) muestran que RNA estructurales (con funciones importantes durante el ciclo de vida y evolución del virus) se forman en cualquier lugar del genoma viral de HIV-1.

Algunas de estas estructuras secundarias de RNA ya habían sido caracterizadas experimentalmente en diferentes partes del genoma de HIV-1. Por ejemplo, el elemento TAR (del inglés *trans-activating responsive element*) localizado en la región 5'LTR que forma una estructura blanco para la proteína Tat (Berkhout, 1992), el elemento RRE (del inglés *Rev response element*) en el gen *env* que exporta mRNAs del núcleo al citoplasma (Malim et al., 1989), la horquilla estructural localizada en la región *gag*/*pol* que mantiene el corrimiento del marco de lectura (Parkin et al., 1992), la estructura de RNA localizada en el dominio *gp120*

del gen *env* que facilita la recombinación (Moumen et al., 2001, Galetto et al., 2004), y el elemento GSL3 (del inglés *gag stem loop 3*) localizado en el gen *gag* que actúa en el empaquetamiento viral (Rong et al., 2003, Damgaard et al., 2004). Asimismo, ya se conocía desde hace tiempo que como una consecuencia de las presiones selectivas actuando al nivel del RNA, los nucleótidos involucrados en formar pares de bases intramoleculares en el RNA viral tienden a mostrar una menor variabilidad y tasas de evolución más bajas que aquellas bases no formando estructuras de RNA (Le et al., 1988, Le et al., 1989, Yoshida et al., 1997). Hasta la publicación del trabajo de Watts et al. (2009) no se sabía qué tan recurrente e importante es la estructuración del genoma de RNA del virus HIV-1.

Usando los datos de Watts et al. (2009), Sanjuán y Bordería (2011) demostraron además que la estructura del RNA y las proteínas no evolucionan de forma independiente en el genoma del virus de HIV-1. Una correlación negativa existe entre el grado del apareamiento de bases en el RNA genómico y la variabilidad de aminoácidos (en particular, en la segunda posición de sus codones). De modo que los aminoácidos codificados por estructuras de RNA funcionales son consistentemente menos variables a lo largo del genoma y evolucionan más lentamente que aquellos sitios de aminoácidos para los cuales la selección al nivel del RNA se encuentra ausente.

De esta forma, estructuras de RNA "relajadas o flexibles" pueden favorecer la acumulación de la variación genética en proteínas. Tal parece ser el caso de la acumulación de la variación genética en el gen *tat* y en las regiones hipervariables de la glicoproteína de envoltura de la superficie viral gp120 localizada en el gen *env*, las cuales se encuentran sujetas a selección positiva según nuestras predicciones (gen *tat* y dominios funcionales 1, 2 y 4 del gen *env* en la Figura R3), con excepción del dominio funcional 3 del gen *env* (el dominio de gp41) que se encuentra bajo selección negativa. El sistema inmune ejerce una fuerte presión de selección sobre la proteína de envoltura de la superficie gp120 localizada en el gen *env* a través de la neutralización de anticuerpos y la respuesta citotóxica de las células T (Wei et al., 2003, Kawashima et al., 2009). Mientras que la eficiencia del proceso de transcripción del genoma viral depende de la variabilidad genética de *tat*, debido a su capacidad de interactuar con factores transcripcionales celulares y de formar un complejo transcripcional con la RNA

polimerasa II del hospedero (Hauber et al., 1989, Ruben et al., 1989). En ambos casos, las presiones evolutivas han resultado en la presencia de muchos codones seleccionados positivamente. Las regiones de hipervariabilidad en la glicoproteína gp120 del gen *env* y en el gen *tat* se encuentran localizadas sobre la superficie de la estructura tridimensional de las proteínas para las que codifican, las cuales se encuentran expuestas al solvente y/o en sitios epítopos de células T CD8, T CD4 y anticuerpos (Woo et al., 2010, Snoeck et al., 2011). La selección positiva al nivel de los codones no parece ser el único factor que contribuye a la presencia de regiones hipervariables en la glicoproteína gp120 y en el gen *tat*, se ha comprobado recientemente que la flexibilidad de las estructuras de RNA localizadas en estas regiones también permiten la acumulación de mutaciones (Snoeck et al., 2011, Sanjuán and Bordería, 2011).

Contrariamente, cambios en la secuencia mantenidas por la selección positiva al nivel de proteínas pueden comprometer estructuras de RNA existente. Por ejemplo, el gen *pol* es el blanco de drogas antivirales dirigidas contra la proteasa (PR) y la reverso transcriptasa (RT). La extensión de la resistencia a la droga reportada entre individuos tratados es una consecuencia obvia de la fuerte presión de selección ejercida sobre las proteínas PR y RT (Lamei et al., 2004, Hué et al., 2009). Sin embargo, se ha reportado recientemente que la ruptura de pares de bases de las estructuras funcionales de RNA en los genes PR y RT, regiones altamente conservadas al nivel de aminoácidos, debe imponer un costo en el *fitness* del virus al mantener mutaciones en tales genes; de modo que existe un límite en la extensión de la resistencia a la droga *via* selección purificadora entre pacientes no tratados con antivirales (Hué et al., 2009, Snoeck et al., 2011, Sanjuán and Bordería, 2011). En individuos no tratados con drogas antivirales, de hecho, se ha encontrado que las regiones de la RT que codifican para plegamientos del tipo β-plegada se encuentran sujetas a una mayor selección purificadora que aquellas regiones de la RT que codifican para plegamientos tipo α-hélice (Leandro et al., 1999).

Es así como nuestras predicciones sobre las regiones y codones sujetos a un determinado tipo de selección corroboran las estimaciones realizadas parcialmente por otros grupos de trabajo

(de Oliveira et al., 2004, Soares et al., 2008, Snoeck et al., 2011). Así como también es posible observar en nuestros resultados (ver Figura R7) la interrelación entre un tipo de selección y la conservación al nivel de proteínas y de estructura de RNA que se ha sugerido recientemente (Sanjuán and Bordería, 2011), y que ya se discutió en los párrafos anteriores.

Finalmente, lo que nos parece de singular curiosidad traer a discusión es la predisposición espacial de una predominante selección purificadora en la primera mitad del genoma de HIV-1 (sólo 5'LTR, *gag* y *pol* cubren los primeros 5,100 nucleótidos de los ~10,000 nts que posee el genoma completo) con respecto a la selección direccional que predomina en la segunda mitad. Las diferentes cepas del virus HIV-1 difícilmente poseen un tamaño del genoma mayor a los 11,000 nts, y todas las regiones codificantes y de estructura de RNA adicionales a la cepa de referencia HXB2 (tal como el gen *rev*) se encuentran anidadas entre sí en los genes localizados en la segunda mitad del genoma. A diferencia de reportes previos, nosotros usamos una muestra bastante diversa de genomas completos del virus HIV-1, incluyendo diferentes subtipos virales de HIV-1, provenientes de diferentes países y años de secuenciación. De esta forma, se podría sugerir muy especulativamente que la predisposición del genoma de HIV-1 para promover persistentemente un tipo de selección en una u otra región del genoma refleja algún mecanismo de control de la maquinaria de la replicación o la presencia de una estrategia estructural genómica que promueva la acumulación de un determinado tipo de mutaciones en este genoma viral sin que comprometa las funciones más vitales del virus HIV-1.

De este modo, una primera especulación nos llevaría a sugerir que las estructuras de RNA no sólo modulan la localización y persistencia de la variabilidad codificante para proteínas, sino que también modularían la tasa de mutación y el *fitness mutacional* del virus, como ha sido sugerido previamente (Le et al., 1988, Le et al., 1989, Sanjuán and Bordería, 2011). Se ha reportado ampliamente que los ácidos nucleicos de cadena sencilla son más propensos al daño químico (Lindahl and Nyberg, 1974), a la ruptura de cadena (Parthasarathi et al., 1995), a la edición mediada por APOBEC3G (Yu et al., 2004), y se ha sugerido la interrelación entre la estructura de RNA y la fidelidad de la polimerasa (Ji et al., 1994, Yoshida et al., 1997, Contreras et al., 2002). El presente trabajo no logró incluir la influencia de la selección en las mutaciones que afectan las estructuras de RNA, debido a cuestiones de tiempo y a la falta de

una metodología adecuada para evaluar selección al nivel de RNA, aunque este objetivo sí estuvo considerado seriamente al inicio de este proyecto.

Por otro lado, la recombinación entre los genomas de HIV-1 (derivados de diversas partículas infectando a una misma célula hospedera), así como la presencia de la epístasis podrían generar en conjunto un mecanismo que sesgue la disposición espacial de un tipo de selección sobre otro en el genoma de HIV-1. La *epístasis* es definida como la dependencia funcional y evolutiva de la interacción generada entre dos posiciones de nucleótidos o aminoácidos dentro del mismo gen o locus (*epístasis intragénica*), como entre genes que segreguen de forma independiente o entre loci que estén ligados (*epístasis intergénica*). La presencia de la epístasis implica que las mutaciones no son independientes, sino que el fenotipo y *fitness* de una mutación depende de la presencia o ausencia de otra u otras mutaciones en el genoma (Elena et al., 2010). Cuando la interacción o dependencia entre dos mutaciones potencializa el resultado de tales mutaciones ventajosas o de tales mutaciones deletéreas, entonces se habla de que existe una *epístasis sinérgica positiva* o *negativa*, respectivamente. Si la interacción o dependencia de dos mutaciones cambia de forma opuesta el resultado de tales mutaciones ventajosas o de tales mutaciones deletéreas, entonces se habla de que existe una *epístasis antagónica negativa* o *positiva*, respectivamente (Michalakis and Roze, 2004, Elena and Sanjuán, 2007). En términos prácticos y dependiendo de su efecto sobre el *fitness*, la epístasis es positiva cuando este mejora el *fitness*; mientras que la epístasis es negativa cuando hace decaer el *fitness* (Jasnos and Korona, 2007).

Se ha reportado la existencia de epístasis positiva y negativa entre las mutaciones localizadas en las proteínas PR y RT del virus HIV-1. Sin embargo, no es posible establecer con claridad qué tipo de epístasis prevalece en diferentes regiones del genoma, si la intragénica, intergénica, positiva, negativa, sinérgica o antagónica (Bonhoeffer et al., 2004b, Wang et al., 2006, Parera et al., 2009). Se ha sugerido también ampliamente que la presencia de mutaciones epistáticas favorece la robustez mutacional de los genotipos virales que presentan *fitness* bajos, a fin de compensar la alta acumulación de mutaciones negativas o deletéreas en el genoma viral (Bonhoeffer et al., 2004a, Codoñer et al., 2006, de Visser et al., 2011, Jasnos and Korona, 2007, Michalakis and Roze, 2004, Sanjuán and Elena, 2006, Sanjuán et al., 2004,

Burch and Chao, 1999). Asimismo, se ha sugerido que el proceso de recombinación del genoma viral de HIV-1 podría favorecer la frecuencia de tales eventos epistáticos (Bonhoeffer et al., 2004a, de Visser et al., 2011). Estimados recientes sugieren que la RT intercambia los templados de RNA aproximadamente 10 veces por ciclo de replicación, lo cual produce en promedio un evento de recombinación por cada 1,000 pares de bases (Levy et al., 2004). Sin embargo, se debe tomar en cuenta que la epístasis no sólo depende de la recombinación, sino también de su interacción con el ambiente (condición geográfica y multiplicidad de los hospederos), el N_e, y la complejidad del genoma (Sanjuán and Elena, 2006). Así, se ha reconocido ampliamente que la recombinación y la epístasis están interrelacionados y son muy frecuentes en la evolución del virus HIV-1. Sin embargo, en el presente trabajo no incluimos la participación de tales procesos en las simulaciones realizadas con nuestro algoritmo genético, ya que no se cuentan con datos estimados ni precisos de su frecuencia, distribución y tipo de mutaciones que favorecen la recombinación y la epístasis a lo largo del genoma del virus HIV-1.

C. Definición y uso del fitness mutacional en el algoritmo genético quasiespecies.pl

Uno de los factores más importantes de nuestro algoritmo genético para simular la evolución de las poblaciones del virus HIV-1 es determinar el costo evolutivo que tienen las mutaciones en el genoma viral; es decir, definir si las mutaciones pueden aumentar, disminuir o mantener el *fitness* del virus. Para lo lograr tal objetivo, se realizaron tres pasos importantes, mientras que el primero permite identificar el peso selectivo de una mutación en una posición específica del genoma viral, los dos últimos pasos nos permiten asociar tal peso selectivo de una mutación o varias (i.e., *el fitness mutacional*) con el *fitness reproductivo* de la progenie viral. Este *fitness reproductivo* es lo que nosotros denominamos aquí como *fitness score*, y define cuántos nuevos genomas en la generación $t+1$ podrá generar un genoma parental que replica y muta en la generación presente t, según el peso selectivo o neutral de la mutación(es) a las que ha sido sujeto.

Para llevar a cabo el primer paso, se tuvo que calcular el *fitness mutacional* o *relativo* de cada posición del genoma con dos diferentes métodos de selección, Tajima y dS-dN. En nuestro caso se decidió que el análisis se realizara al nivel de tripletes nucleotídicos en los LTRs y codones (aquellos tripletes nucleotídicos que se encuentran en las regiones codificantes de los genes). Los valores obtenidos en este paso conforman un *mapa del fitness mutacional* (una mapa para Tajima y otro para dS-dN) del genoma del virus HIV-1, el cual se puede observar en la Figura R6. Este *mapa del fitness mutacional* se transformó en una *tabla de sitios* en formato texto plano que es necesaria para que el algoritmo genético localice la posición y obtenga (lea) el *fitness mutacional* de la mutación que se realiza en cada ciclo de replicación en un genoma viral en la generación t, a fin de estimar su *fitness score* con las matrices de categorías que se describen más adelante. La matriz de sitios está constituida por 9,719 sitios (que conforman el genoma del virus HIV-1 HXB2) divididos por tripletes, que en las regiones codificantes cada triplete representa un codón. Así, cada nucleótido del triplete del genoma tendrá el mismo valor que fue asignado para todo el triplete con el método correspondiente.

Dado que los valores de selección y neutralidad derivados en ambos métodos son continuos, se tuvo que categorizar tales valores en alguna de las tres categorías: *i)* selección positiva si los valores fueron > 0, *ii)* negativa si los valores fueron < 0, o *iii)* neutra si los valores fueron ≅ 0. Finalmente, dado que los valores de Tajima y dS-dN reflejan diferentes intensidades de la selección o la neutralidad (i.e., neutrales o cercanamente neutrales), se generó un *fitness score* (entre 0 y 2) compuesto por 21 grupos discretos que cubren todo el intervalo de la intensidad de la selección/neutralidad para todos los tripletes del genoma viral en cada método utilizado. El *fitness score* se realizó así a partir de la distribución de los codones entre el valor máximo (positivo) y el valor mínimo (negativo) de selección que predijo cada método a lo largo de 21 grupos de valores discretos de selección (ver Figura R8.A y R8.B).

El *fitness score* se introduce en el programa como una *matriz de categorías* que asigna un *fitness mutacional* a un *fitness reproductivo* (ver Tabla R6 y R7). De esta forma, si el *fitness score* ≅ 1 significa que la mutación o mutaciones no afectan el *fitness reproductivo* del genoma viral, de modo que su progenie generará en la siguiente generación ($t+1$) el mismo

número de nuevos genomas virales que este genoma parental produjo en la generación t. Bajo la misma lógica, si el *fitness score* se acerca a 0, significa que la mutación o mutaciones afectan negativamente el *fitness reproductivo* del genoma viral, de modo que su progenie disminuirá proporcionalmente (respecto a la población que genera el genoma parental en el tiempo t) el número de genomas nuevos en la próxima generación ($t+1$); se produciría una extinción viral si el *fitness score* = 0. Finalmente, si el fitness score se acerca a 2 significa que la mutación o mutaciones afectan positivamente el *fitness reproductivo* del genoma viral, de modo que su progenie se incrementará proporcionalmente (respecto a la población que genera el genoma parental en el tiempo t) el número de genomas nuevos en la próxima generación ($t+1$); se produciría exactamente el doble de la población viral si el *fitness score* = 2.

Dado que las mutaciones no son únicas en el genoma ni tampoco excluyentes de las que ocurren en las generaciones previas, el *fitness score* previo (*t-1*) y el *fitness score* nuevo (*t*) se suman y promedian para cada individuo en la generación presente, de forma que el *fitness score* de cada individuo de una nueva población (*t+1*) mantiene "una memoria" mutacional a lo largo de las generaciones virales. Lo mismo aplica cuando se introducen más de una mutación por genoma por ciclo de replicación, se suman y promedian los *fitness score* de las mutaciones generadas en el presente ciclo de replicación (*t*), el *fitness score* resultante se suma y promedia con el *fitness score* del genoma parental en la generación anterior (*t-1*), a fin de contemplar posibles fenómenos de epístasis y mutaciones deletéreas acumulativas.

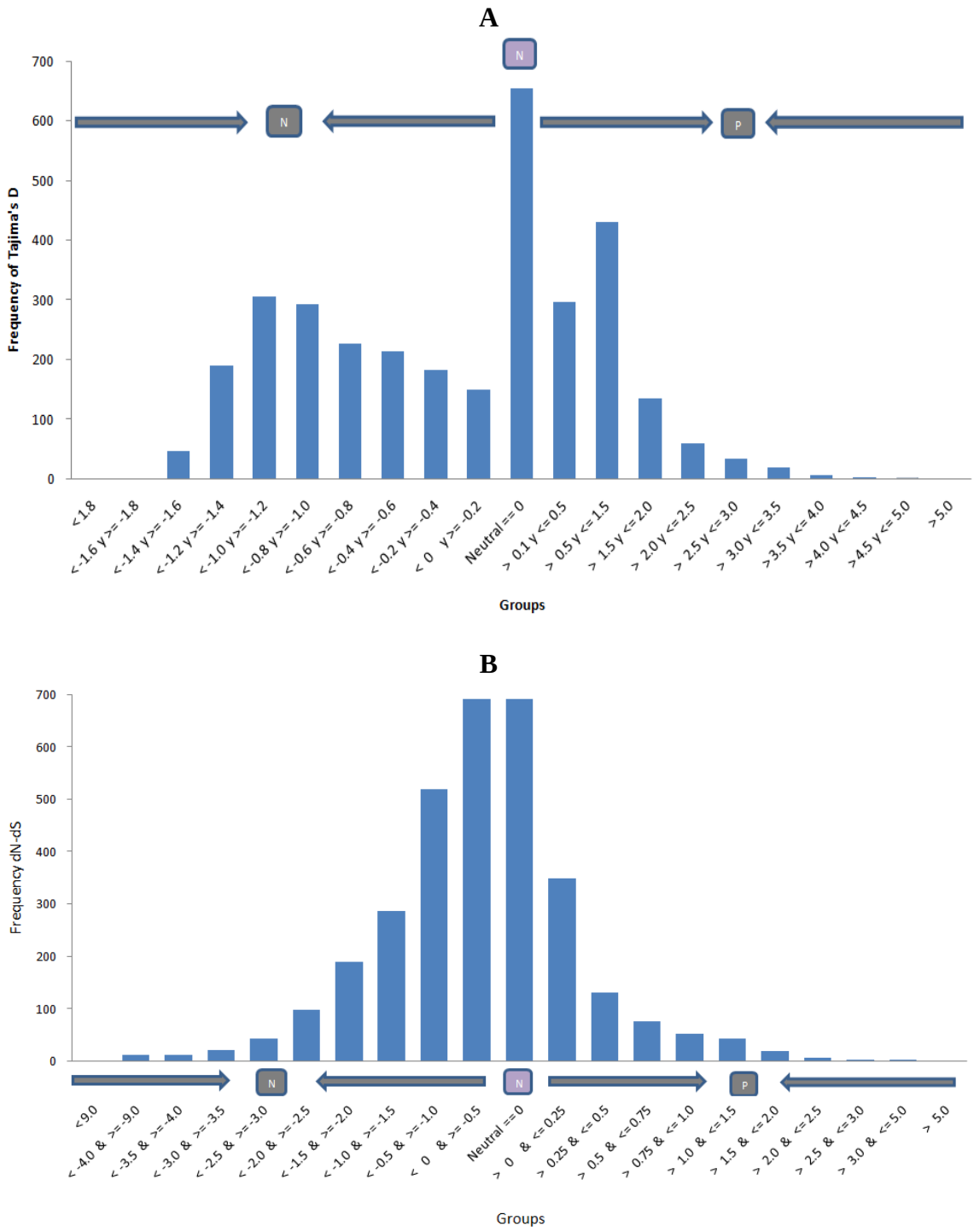


Figure R8. A y B: Histograma de la distribución de frecuencias en codones positivos, negativos y neutrales. A Basando en la prueba de neutralidad de Tajima. **B** estimado realizado por el numero de de dS – dN. Se clasificaron las frecuencias en tres categorías distintas: 1) *N* (*izquierda*), frecuencia de codones con selección negativa, 2) *N* (*centro*), codones que presentan neutralidad y 3) *P* (*derecha*), codones con selección positiva.

Tabla R6. Matriz de Fitness Mutacional (*score*). Obtención de los valores de *fitness mutacional* para todos los codones del genoma de referencia para HIV-1 a través del método *SLAC* [ver métodos]. Se presentan 3 categorías (negativo, neutral y positivo) en las cuales se ubicaron las frecuencias de los *valores de selección* de los codones en diferentes grupos. Asignación del fitness score por grupos. La categoría "negativa" y la "positiva" tienes 10 grupos, mientras que el "neutral" sólo 1. Esta matriz se empleo tanto para las simulaciones con dS y dN, como para las finales.

		Asignación de valores selectivos	
Tipo de Selección (Categoría)	Grupo de valores de dS – dN	Frecuencia de dS - dN (Σ)	Puntaje de Fitness
Negativa	*< -9.0*	*0*	*0*
	< -4.0 & >= -9.0	12	0.1
	< -3.5 & >= -4.0	12	0.2
	< -3.0 & >= -3.5	21	0.3
	< -2.5 & >= -3.0	42	0.4
	< -2.0 & >= -2.5	98	0.5
	< -1.5 & >= -2.0	189	0.6
	< -1.0 & >= -1.5	287	0.7
	< -0.5 & >= -1.0	519	0.8
	< 0 & >= -0.5	692	0.9
Neutral	*Neutral == 0*	*691**	*1*
Positiva	> 0 & <= 0.25	348	1.1
	> 0.25 & <= 0.5	131	1.2
	> 0.5 & <= 0.75	75	1.3
	> 0.75 & <= 1.0	52	1.4
	> 1.0 & <= 1.5	43	1.5
	> 1.5 & <= 2.0	19	1.6
	> 2.0 & <= 2.5	6	1.7
	> 2.5 & <= 3.0	2	1.8
	> 3.0 & <= 5.0	1	1.9
	> 5.0	*0*	*2*

* 101 posiciones de la región final 3' LTR son nucleótidos indeterminados debido a que los datos están incompletos (ver Fig. R6). Esta región es considerada como neutral basada en la prueba de dS-dN.

Tabla R7. Matriz de Fitness Mutacional con valores de Tajima (*score*). Obtención de los valores de *fitness mutacional* para todos los codones del genoma de referencia para HIV-1 a través del método de *Tajima* [ver métodos]. Se presentan, al igual que el anterior, 3 categorías (negativo, neutral y positivo) en las cuales se ubicaron las frecuencias de los *valores de selección* de los codones en diferentes grupos. Asignación del fitness score por grupos. La categoría "negativa" y la "positiva" tienes 10 grupos, mientras que el "neutral" sólo 1.

.

		Asignación de valores selectivos	
Tipo de Selección (Categoría)	Grupo para los Valores de Tajima	Frecuencia de la D de Tajima (Σ)	Puntaje Fitness
Negativa	< -1.8	0	0
	< -1.6 & >= -1.8	0	0.1
	< -1.4 & >= -1.6	47	0.2
	< -1.2 & >= -1.4	189	0.3
	< -1.0 & >= -1.2	306	0.4
	< -0.8 & >= -1.0	293	0.5
	< -0.6 & >= -0.8	226	0.6
	< -0.4 & >= -0.6	214	0.7
	< -0.2 & >= -0.4	182	0.8
	< 0 & >= -0.2	149	0.9
Neutral	*0*	*654**	*1*
Positiva	> 0.1 & <= 0.5	296	1.1
	> 0.5 & <= 1.5	430	1.2
	> 1.5 & <= 2.0	134	1.3
	> 2.0 & <= 2.5	60	1.4
	> 2.5 & <= 3.0	33	1.5
	> 3.0 & <= 3.5	18	1.6
	> 3.5 & <= 4.0	6	1.7
	> 4.0 & <= 4.5	2	1.8
	> 4.5 & <= 5.0	1	1.9
	> 5.0	0	2

* 101 posiciones de la región final 3' LTR son nucleótidos indeterminados debido a que los datos están incompletos (ver Fig. R6).

III. Simulación de la evolución de las poblaciones del virus HIV-1 con el programa *quasispecies.pl*

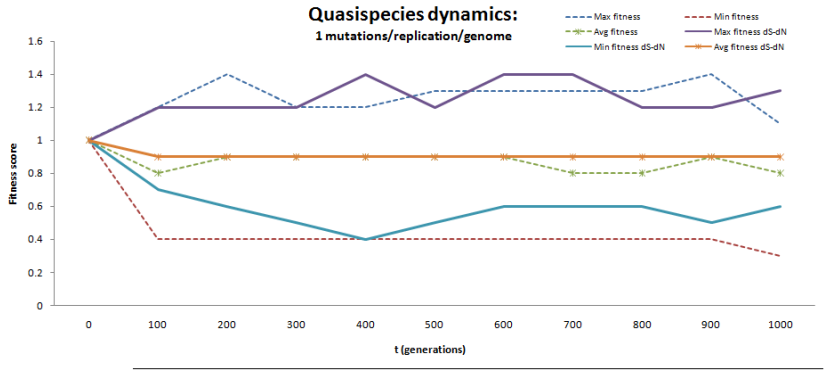
Para comprobar si una población viral con genomas de RNA puede evolucionar como *cuasiespecies*, se creó un programa computacional, con la estructura de un algoritmo genético, que incorpora realismo adicional el modelo original de *cuasiespecies* propuesto por Eigen-Schuster en 1977. Dada la complejidad de la caracterización de los parámetros biológicos de tal algoritmo, decidimos enfocar nuestras simulaciones en el Virus de la Inmunodeficiencia Adquirida serotipo 1, usando a la partícula viral HXB2 como cepa modelo de HIV-1. Nuestro algoritmo incorpora parámetros biológicos del virus HIV-1, estimados con base en métodos experimentales (obtenidos de la literatura) y bioinformáticos para definir: la tasa mutacional, la frecuencia del tipo de mutaciones, el tiempo generacional, el tamaño efectivo de la población, así como el peso selectivo e influencia reproductiva de las mutaciones en el genoma viral [ver Material y Métodos]. El programa *quasispecies.pl* genera poblaciones (de tamaño constante) de genotipos mutantes del virus HIV-1 bajo una serie de etapas definidas: en la *etapa de iniciación* se generan un N_e número de copias del genoma viral que poseen un *fitness inicial* de 1; en la *etapa de mutación*, cada genoma muta a un tasa específica, entonces, un *fitness mutacional* y un *fitness score* (reproductivo) es asignado a las mutaciones generadas en cada ciclo de replicación para las siguientes generaciones; en la *etapa de selección*, se genera una nueva población conformada por N_e genotipos seleccionados aleatoriamente a partir de toda la progenie viral generada en una generación. Este proceso se repite durante *n* generaciones [ver Material y Métodos].

Las simulaciones con el programa *quasispecies.pl* se realizaron con diversos tamaños de población, tasas de mutación y tiempos generacionales [ver Material y Métodos]. A continuación sólo mostramos dos *sets* de las simulaciones realizadas; otras simulaciones se pueden observar en el Material Suplementario de esta tesis. Para la primera simulación se emplearon los siguientes parámetros: un tamaño de población de 100 genomas, sometidos a 1, 2, 3 y 4 mutaciones por genoma por ciclo de replicación a lo largo de 1,000 generaciones. Estas simulaciones se duplicaron para usarse una *matriz de categorías de fitness score*s calculado por Tajima y otra por dS-dN. Los resultados de estas simulaciones pueden

observarse en la Figura R9, con la presencia de 1, 2, 3 y 4 mutaciones en las gráficas a, b, c y d, respectivamente. Para la segunda simulación se emplearon los siguientes parámetros: un tamaño de población de 1000 genomas, sometidos a 1, 2, 3 y 4 mutaciones por genoma por ciclo de replicación a lo largo de 1,000 generaciones. Debido a que el tiempo de cómputo para realizar este segundo *set* de simulaciones fue mayor a los 6 meses, sólo se muestran los resultados de las simulaciones que emplearon una *matriz de categorías de fitness score*s calculados a través del método dS-dN. Elegimos usar el *fitness score* calculado con este método de selección/neutralidad debido a que los resultados obtenidos con dS-dN presentan una mayor sensibilidad que los resultados obtenidos con el método de Tajima [ver Figura R5]. Los resultados de estas simulaciones pueden observarse en la Figura R10, con la presencia de 1, 2, 3 y 4 mutaciones en las gráficas a, b, c y d, respectivamente.

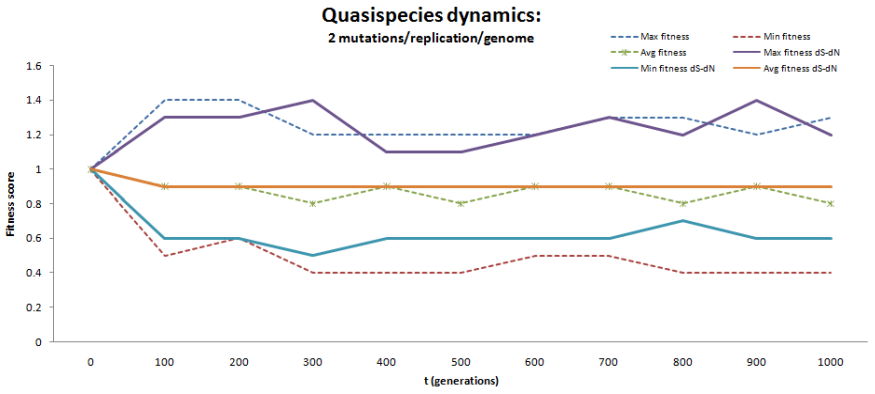
En ambos *sets* de simulaciones se observan patrones similares de la dinámica computacional realizada. Además, no se muestran diferencias significativas en las simulaciones obtenidas usando un *fitness mutacional* estimado con el método Tajima o con dS-dN. Si se observa la frecuencia de los genomas mutantes (sujetos a 1 y 2 mutaciones por genoma por ciclo de replicación) y sus respectivos *fitness score* (más alto, más bajo y el promedio) cada 100 generaciones en los incisos A y B de las Figuras R9 y R10, se podrá observar que la mayor parte de los genotipos virales se encuentran posicionados en los *fitness scores* más altos [ver Tablas bajo cada gráfico]. De hecho, la distribución de los genotipos mutantes no se centra sobre la categoría del *fitness score* más alto, sino que se distribuye con mayor proporción entre los *fitness score* que se encuentran más cercanos al *fitness score* más alto que se obtiene cada 100 generaciones (a modo de "nubes" de genotipos alrededor de una o varias *secuencias maestra* o *consenso* –genotipos *fittest*–). Esta tendencia se mantiene constante a lo largo de las 1,000 generaciones si se somete la replicación de los genomas virales a 1 y 2 mutaciones por genoma por ciclo de replicación. La dinámica de los genotipos virales observada aquí podría sugerir un comportamiento de *cuasiespecies* si se muestra que este comportamiento es dependiente de un *umbral de error*. El *umbral de error*, cabe recordar, es la tasa de mutación por debajo de la cual las poblaciones están equilibradas en un clásico balance de mutación-selección, donde el genotipo favorecido puede mantener la replicación de la información a pesar de la alta tasa de mutación.

A



dS - dN	1.1: 16	1.1: 10	1.1: 12	1.1: 11	1.1: 17	1.3: 1	1.1: 14	1.1: 12	1.1: 21	1.3: 1
Frequency of mutants	1.2: 1	1.2: 5	1.2: 1	0.8: 20	1.2: 2	1.1: 17	0.8: 18	1.2: 2	1.2: 1	1.1: 11
	0.8: 19	0.8: 15	0.8: 17	1.2: 5	0.8: 18	0.8: 18	0.7: 4	0.8: 11	0.8: 13	1.2: 1
	0.7: 3	0.7: 5	0.7: 7	0.7: 3	0.7: 8	1.2: 3	0.9: 38	0.7: 5	0.7: 7	0.8: 12
	0.9: 41	0.9: 54	0.9: 47	0.9: 35	0.9: 39	0.7: 6	0.6: 1	0.9: 49	0.9: 40	0.7: 1
	1.0: 21	0.6: 2	0.5: 2	1.5: 1	0.6: 2	0.9: 41	1.0: 25	0.6: 3	0.5: 1	0.9: 57
		1.0: 10	1.0: 15	0.6: 4	0.5: 1	0.6: 2	1.4: 1	1.0: 19	1.0: 18	0.6: 1
				0.4: 1	1.0: 14	1.0: 12				1.0: 17
				1.0: 20		1.4: 1				
Tajima	1.1: 10	1.1: 11	1.1: 23	1.1: 13	1.3: 1	1.3: 1	1.3: 1	1.3: 1	1.3: 1	1.1: 20
Frequency of mutants	1.2: 2	1.2: 2	1.2: 3	0.8: 20	1.1: 28	1.1: 17	1.1: 9	1.1: 13	1.1: 20	0.8: 16
	0.8: 22	0.8: 27	0.8: 22	1.2: 1	1.2: 4	0.8: 23	1.2: 1	1.2: 1	0.8: 32	0.7: 15
	0.7: 9	0.7: 8	0.7: 12	0.7: 12	0.8: 19	1.2: 6	0.8: 26	0.8: 24	0.7: 4	0.3: 1
	0.9: 28	0.9: 21	0.9: 15	0.9: 19	0.7: 5	0.7: 10	0.7: 13	0.7: 14	0.9: 20	0.9: 22
	0.6: 14	0.6: 7	0.6: 7	0.6: 14	0.9: 23	0.9: 17	0.9: 26	0.9: 23	0.6: 7	0.6: 13
	0.4: 1	0.4: 2	0.4: 2	0.4: 1	0.6: 9	0.6: 9	0.6: 8	0.6: 9	0.4: 3	0.4: 1
	0.5: 6	0.5: 2	0.5: 1	0.5: 1	0.4: 1	0.4: 3	0.4: 2	0.4: 3	0.5: 3	0.5: 2
	1.0: 9	1.0: 20	1.0: 16	1.0: 20	1.0: 11	1.0: 11	1.0: 12	0.5: 2	1.0: 10	1.0: 11
		1.4: 1				0.5: 4	0.5: 3	1.0: 11	1.4: 1	

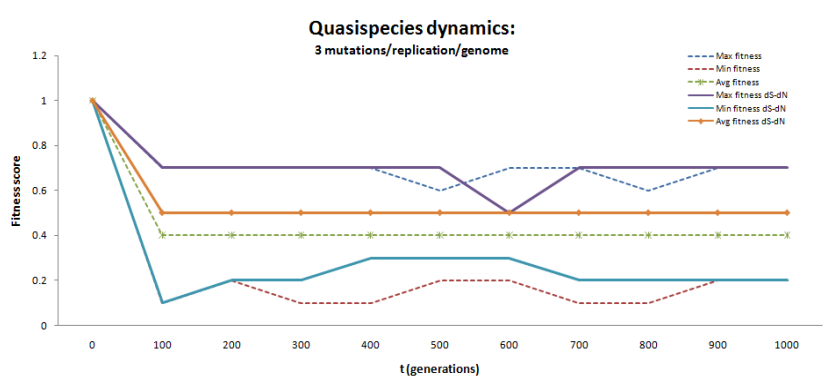
B



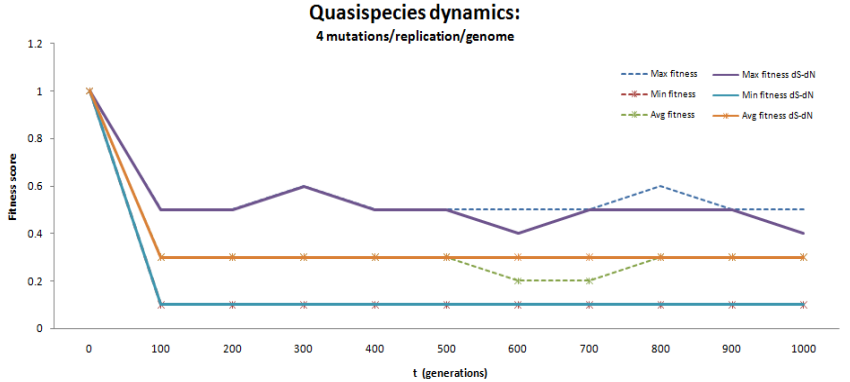
dS - dN	1.3: 1	1.3: 1	1.1: 11	1.1: 9	1.1: 21	1.1: 8	1.3: 2	1.1: 16	1.1: 12	1.1: 16
Frequency of mutants	1.1: 10	1.1: 12	1.2: 2	0.8: 28	0.8: 22	1.2: 4	1.1: 17	1.2: 3	1.2: 4	1.2: 2
	0.8: 18	1.2: 4	0.8: 17	0.7: 2	1.2: 3	0.8: 18	1.2: 3	0.8: 13	0.8: 18	0.8: 18
	1.2: 4	0.8: 20	0.7: 4	0.9: 45	0.7: 5	0.7: 4	0.8: 16	0.7: 4	0.7: 3	0.7: 3
	0.7: 4	0.7: 2	0.9: 46	0.6: 2	0.9: 35	0.9: 55	0.7: 1	0.9: 52	0.9: 44	0.9: 44
	0.9: 49	0.9: 38	0.5: 2	1.0: 15	0.6: 1	0.6: 1	0.9: 41	1.0: 13	0.6: 1	0.6: 3
	0.6: 1	0.6: 3	1.0: 18		1.0: 14	1.0: 11	0.6: 3		1.0: 18	1.0: 15
	1.0: 14	1.0: 21	1.4: 1				1.0: 18		1.4: 1	
Tajima	1.1: 20	1.1: 15	1.1: 12	1.1: 24	1.1: 15	1.1: 17	1.3: 2	1.3: 1	1.1: 28	1.3: 1
Frequency of mutants	0.8: 19	1.2: 4	1.2: 3	1.2: 3	1.2: 3	1.2: 6	1.1: 11	1.1: 18	1.2: 3	1.1: 18
	1.2: 1	0.8: 16	0.8: 22	0.8: 19	0.8: 12	0.8: 18	1.2: 4	0.8: 22	0.8: 22	1.2: 1
	0.7: 16	0.7: 16	0.7: 14	0.7: 9	0.7: 18	0.7: 10	0.8: 27	0.7: 17	0.7: 6	0.8: 16
	0.9: 18	0.9: 21	0.9: 21	0.9: 23	0.9: 19	0.9: 24	0.7: 8	0.9: 17	0.9: 16	0.7: 7
	0.6: 9	0.6: 15	0.6: 14	0.6: 12	0.6: 19	0.6: 4	0.9: 16	0.6: 8	0.6: 11	0.9: 22
	0.5: 6	1.0: 13	0.4: 2	0.4: 2	0.4: 1	0.5: 4	0.6: 12	0.4: 1	0.4: 2	0.6: 19
	1.0: 11	1.4: 1	0.5: 4	0.5: 1	0.5: 2	1.0: 18	0.5: 6	0.5: 5	1.0: 11	0.4: 5
	1.4: 1		1.0: 9	1.0: 8	1.0: 12		1.0: 15	1.0: 12	0.5: 2	1.0: 8
										0.5: 4

Figure R9 A y B. Dinámica de *cuasiespecies*: comparación del método de Tajima y el de dS-dN. Se realizaron las simulaciones con 100 genomas virales con 1000 ciclos de replicación con una tasa de mutación de 1, 2, 3 y 4 mutaciones por ciclo de replicación por genoma. En líneas punteadas representan la dinámica con el método de *Tajima*. Las líneas continuas representan la estimación por *dS-dN*. Se puede apreciar que ambos métodos muestran una dinámica similar. En la parte inferior de cada grafica se muestra la frecuencia de mutantes con su *fitness* correspondiente para cada ciclo de replicación. En la parte inferior de A y B, se muestra la frecuencia de mutantes obtenida en las diferentes dinámicas evolutivas para cada diferente método de detección de la selección: *dS-dN* y *Tajima*; presentado como ***a:b***, lo que significa 'valor de fitness' y 'frecuencia de mutantes', respectivamente.

C **D**



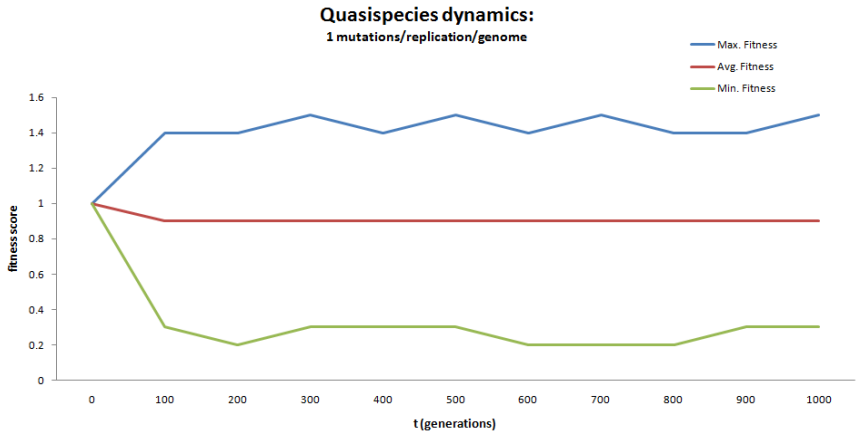
dS - dN	0.3: 7	0.3: 10	0.3: 8	0.3: 6	0.7: 2	0.3: 9	0.2: 1	0.2: 1	0.2: 1	0.2: 3
Frequency of mutants	0.1: 1	0.2: 1	0.2: 1	0.6: 7	0.3: 9	0.6: 6	0.7: 2	0.7: 2	0.7: 3	0.7: 2
	0.6: 5	0.6: 4	0.6: 5	0.4: 15	0.6: 9	0.4: 20	0.3: 9	0.3: 9	0.3: 4	0.3: 8
	0.4: 22	0.4: 14	0.4: 15	0.5: 71	0.4: 18	0.5: 66	0.6: 4	0.6: 1	0.6: 4	0.6: 5
	0.5: 65	0.5: 67	0.5: 70	0.7: 2	0.5: 63		0.4: 18	0.4: 20	0.4: 13	0.4: 22
	0.7: 1	0.7: 5	0.7: 2				0.5: 67	0.5: 68	0.5: 76	0.5: 61
Tajima	0.2: 4	0.2: 7	0.2: 9	0.2: 14	0.2: 7	0.2: 6	0.2: 6	0.2: 6	0.2: 6	0.2: 10
Frequency of mutants	0.1: 1	0.7: 3	0.1: 1	0.1: 1	0.3: 33	0.7: 1	0.1: 1	0.1: 1	0.7: 1	0.7: 2
	0.7: 2	0.3: 23	0.7: 1	0.7: 2	0.6: 6	0.3: 31	0.7: 4	0.3: 25	0.3: 25	0.3: 29
	0.3: 25	0.6: 6	0.3: 22	0.3: 24	0.4: 13	0.6: 2	0.3: 23	0.6: 3	0.6: 4	0.6: 5
	0.6: 8	0.4: 11	0.6: 9	0.6: 10	0.5: 42	0.4: 13	0.6: 7	0.4: 17	0.4: 22	0.4: 11
	0.4: 12	0.5: 51	0.4: 15	0.4: 11		0.5: 48	0.4: 8	0.5: 49	0.5: 43	0.5: 44
	0.5: 49		0.5: 44	0.5: 39			0.5: 52			



dS - dN	0.3: 7	0.3: 10	0.3: 8	0.3: 6	0.7: 2	0.3: 9	0.2: 1	0.2: 1	0.2: 1	0.2: 3
Frequency of mutants	0.1: 1	0.2: 1	0.2: 1	0.6: 7	0.3: 9	0.6: 6	0.7: 2	0.7: 2	0.7: 3	0.7: 2
	0.6: 5	0.6: 4	0.6: 5	0.4: 15	0.6: 9	0.4: 20	0.3: 9	0.3: 9	0.3: 4	0.3: 8
	0.4: 22	0.4: 14	0.4: 15	0.5: 71	0.4: 18	0.5: 66	0.6: 4	0.6: 1	0.6: 4	0.6: 5
	0.5: 65	0.5: 67	0.5: 70	0.7: 2	0.5: 63		0.4: 18	0.4: 20	0.4: 13	0.4: 22
	0.7: 1	0.7: 5	0.7: 2				0.5: 67	0.5: 68	0.5: 76	0.5: 61
Tajima	0.3: 51	0.3: 40	0.3: 48	0.3: 27	0.3: 40	0.3: 41	0.3: 38	0.3: 35	0.3: 42	0.3: 33
Frequency of mutants	0.2: 15	0.2: 21	0.2: 18	0.2: 28	0.2: 22	0.2: 18	0.2: 23	0.2: 20	0.2: 24	0.2: 29
	0.1: 22	0.1: 19	0.6: 1	0.1: 23	0.1: 19	0.1: 26	0.1: 24	0.1: 26	0.1: 21	0.1: 25
	0.4: 13	0.4: 19	0.1: 18	0.4: 22	0.4: 14	0.4: 15	0.4: 14	0.4: 17	0.4: 13	0.4: 14
		0.5: 2	0.4: 14	0.5: 1	0.5: 6	0.5: 1	0.5: 2	0.5: 3	0.5: 1	
			0.5: 2							

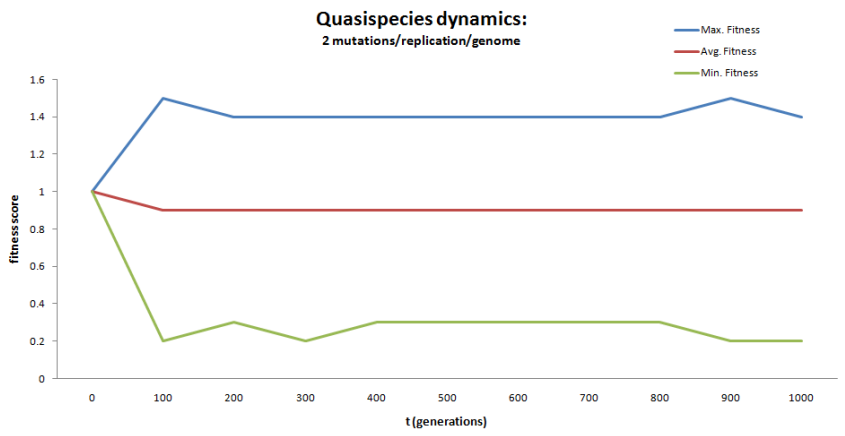
Figura R9 C y D (Continuación). Se realizaron las simulaciones con 100 genomas virales con 1000 ciclos de replicación con una tasa de mutación de 1, 2, 3 y 4 mutaciones por ciclo de replicación por genoma. Estos datos estadísticos son obtenidos del programa *quasispecies.pl* y fueron calculados en la etapa final de cada ciclo de replicación del programa *quasispecies.pl*. En esta etapa o *proceso de terminación* del programa, se realizan todas las evaluaciones estadísticas relacionadas al *fitness* de las poblaciones y sus correspondientes frecuencias (ver métodos). En la parte inferior de cada grafica se muestra la frecuencia de mutantes con su *fitness* correspondiente para cada ciclo de replicación; presentado como ***a:b***, lo que significa 'valor de fitness' y 'frecuencia de mutantes', respectivamente.

A



dS - dN									
Frequency of mutants									
1.3: 5	1.3: 6	1.3: 6	1.3: 2	1.3: 4	1.3: 4	1.3: 11	1.3: 5	1.3: 4	1.3: 2
1.1: 133	0.2: 1	1.1: 131	1.1: 145	1.1: 135	0.2: 1	0.2: 1	0.2: 1	1.1: 150	1.1: 126
1.2: 200	1.1: 142	1.2: 33	1.2: 22	1.2: 24	1.1: 117	1.1: 145	1.1: 151	1.2: 30	0.8: 164
0.8: 164	1.2: 31	0.8: 188	0.8: 190	0.8: 187	1.2: 23	1.2: 33	1.2: 30	0.8: 188	1.2: 38
0.7: 68	0.8: 172	0.7: 72	0.7: 66	0.7: 59	0.8: 184	0.8: 159	0.8: 183	0.7: 56	0.7: 69
0.3: 4	0.7: 76	1.6: 1	0.9: 336	0.3: 5	0.7: 79	0.7: 80	0.7: 85	0.3: 2	0.3: 6
0.9: 360	0.3: 7	0.3: 6	0.3: 10	0.9: 345	0.3: 2	0.3: 3	0.3: 8	0.9: 313	0.9: 352
0.6: 51	0.9: 344	0.9: 336	0.6: 40	1.5: 1	0.9: 360	0.9: 335	0.9: 305	0.6: 47	1.5: 1
0.4: 5	0.6: 44	1.5: 1	0.4: 2	0.6: 44	0.6: 50	1.5: 1	0.6: 48	0.4: 7	0.6: 44
1.0: 142	0.4: 7	0.6: 51	0.5: 34	0.4: 5	0.4: 4	0.6: 40	0.4: 9	0.5: 27	0.4: 8
0.5: 45	0.5: 26	0.4: 5	1.0: 146	0.5: 42	0.5: 30	0.4: 4	1.0: 132	1.0: 174	0.5: 41
1.4: 4	1.0: 144	0.5: 42	1.4: 8	1.0: 147	1.0: 141	0.5: 33	0.5: 37	1.4: 3	1.0: 145
	1.4: 1	1.0: 127		1.4: 3	1.4: 6	1.0: 156	1.4: 7		1.4: 5

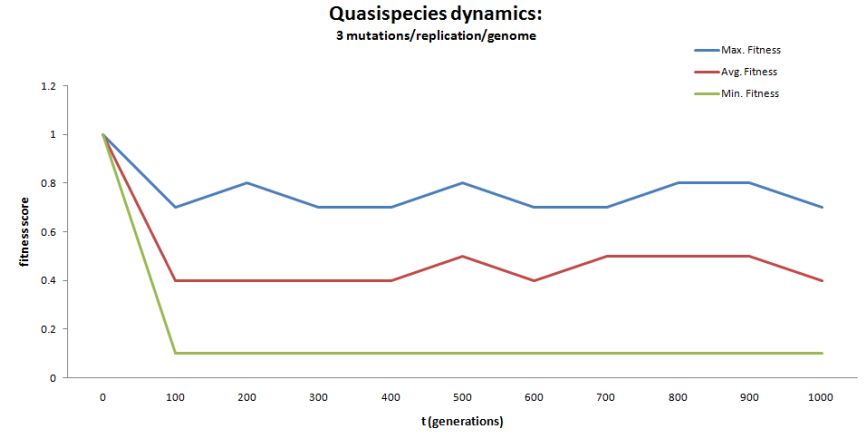
B



dS - dN									
Frequency of mutants									
1.3: 2	1.3: 5	1.3: 7	1.3: 5	1.3: 5	1.3: 2	1.3: 2	1.3: 7	1.3: 3	1.3: 4
0.2: 1	1.1: 157	0.2: 2	1.1: 120	1.1: 159	1.1: 148	1.1: 143	1.1: 154	0.2: 2	0.2: 1
1.1: 150	0.8: 176	1.1: 143	1.2: 25	1.2: 34	1.2: 25	0.8: 198	1.2: 30	1.1: 141	1.1: 146
1.2: 28	1.2: 25	0.8: 185	0.8: 176	0.8: 167	0.8: 192	1.2: 38	0.8: 181	0.8: 175	0.8: 187
0.8: 177	0.7: 51	1.2: 32	0.7: 80	0.7: 78	0.7: 60	0.7: 68	0.7: 58	1.2: 33	1.2: 22
0.7: 70	0.3: 5	0.7: 74	0.3: 5	0.3: 5	0.3: 3	0.3: 3	0.3: 2	0.7: 83	0.7: 84
0.3: 8	0.9: 335	0.3: 5	0.9: 360	0.9: 333	0.9: 340	0.9: 342	0.9: 354	0.3: 4	0.3: 6
0.9: 339	0.6: 52	0.9: 351	0.6: 56	0.6: 40	0.6: 46	0.6: 41	0.6: 34	0.9: 333	0.9: 332
1.5: 1	0.4: 6	0.6: 41	0.4: 10	0.4: 5	0.4: 4	0.4: 3	0.4: 3	1.5: 1	0.6: 40
0.6: 52	0.5: 36	0.4: 4	0.5: 33	0.5: 27	0.5: 26	0.5: 30	0.5: 26	0.6: 45	0.4: 7
0.4: 7	1.0: 148	0.5: 34	1.0: 128	1.0: 145	1.0: 153	1.0: 130	1.0: 146	0.4: 8	0.5: 30
0.5: 31	1.4: 5	1.0: 122	1.4: 3	1.4: 3	1.4: 2	1.4: 3	1.4: 6	1.0: 139	1.0: 141
1.0: 133		1.4: 1						0.5: 32	1.4: 1
1.4: 2								1.4: 2	

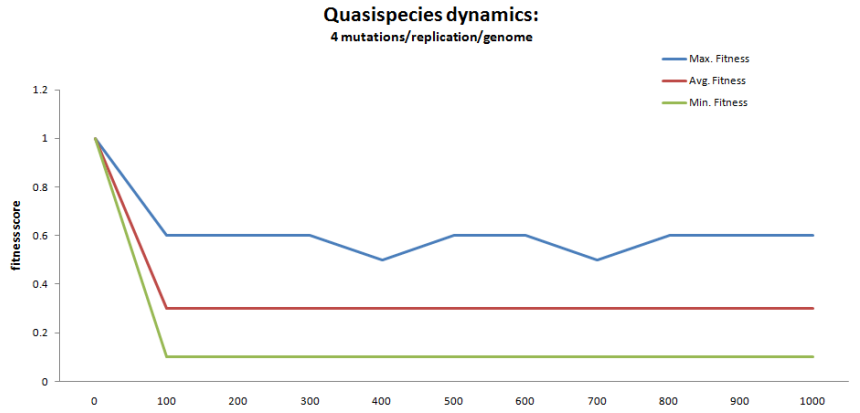
Figura R10 A y B. Se realizaron las simulaciones con 1000 genomas virales con 1000 ciclos de replicación con una tasa de mutación de 1, 2, 3 y 4 mutaciones por ciclo de replicación por genoma. En la parte inferior de cada grafica se muestra la frecuencia de mutantes con su *fitness* correspondiente para cada ciclo de replicación; presentado como ***a:b***, lo que significa 'valor de fitness' y 'frecuencia de mutantes', respectivamente.

C



dS - dN	0.3: 95	0.3: 89	0.3: 96	0.3: 98	0.3: 94	0.3: 83	0.3: 96	0.3: 76	0.3: 89	0.3: 108
Frequency of mutants	0.2: 45	0.2: 38	0.2: 55	0.2: 43	0.2: 39	0.2: 46	0.2: 54	0.2: 50	0.2: 39	0.2: 47
	0.1: 23	0.8: 1	0.1: 18	0.1: 25	0.8: 1	0.1: 25	0.1: 19	0.8: 3	0.8: 1	0.1: 24
	0.6: 67	0.6: 72	0.6: 62	0.6: 69	0.1: 18	0.6: 70	0.6: 82	0.1: 19	0.1: 19	0.6: 61
	0.4: 198	0.1: 29	0.4: 194	0.4: 198	0.6: 75	0.4: 203	0.4: 171	0.6: 80	0.6: 77	0.4: 193
	0.5: 554	0.4: 209	0.5: 554	0.5: 548	0.4: 195	0.5: 556	0.5: 553	0.4: 180	0.4: 175	0.5: 551
	0.7: 19	0.5: 546	0.7: 22	0.7: 20	0.5: 559	0.7: 18	0.7: 26	0.5: 573	0.5: 580	0.7: 17
		0.7: 17			0.7: 20			0.7: 20	0.7: 21	

D



dS - dN	0.3: 567	0.3: 525	0.3: 569	0.3: 551	0.3: 573	0.3: 575	0.3: 551	0.3: 564	0.3: 586	0.3: 574
Frequency of mutants	0.2: 201	0.2: 230	0.2: 213	0.2: 206	0.2: 190	0.2: 225	0.2: 209	0.2: 225	0.2: 209	0.2: 198
	0.6: 1	0.6: 4	0.6: 1	0.1: 102	0.6: 1	0.6: 1	0.1: 108	0.6: 1	0.6: 3	0.6: 2
	0.1: 76	0.1: 96	0.1: 85	0.4: 116	0.1: 99	0.1: 84	0.4: 98	0.1: 92	0.1: 82	0.1: 92
	0.4: 126	0.4: 116	0.4: 112	0.5: 26	0.4: 108	0.4: 92	0.5: 35	0.4: 98	0.4: 99	0.4: 115
	0.5: 30	0.5: 30	0.5: 21		0.5: 30	0.5: 24		0.5: 21	0.5: 22	0.5: 20

Figura R10 C y D. (Continua) Se realizaron las simulaciones con 1000 genomas virales con 1000 ciclos de replicación con una tasa de mutación de 1, 2, 3 y 4 mutaciones por ciclo de replicación por genoma. En la parte inferior de cada grafica se muestra la frecuencia de mutantes con su *fitness* correspondiente para cada ciclo de replicación; presentado como ***a:b***, lo que significa 'valor de fitness' y 'frecuencia de mutantes', respectivamente.

Consecutivamente, cuando los genotipos virales se sometieron a 3 y 4 mutaciones por genoma por ciclo de replicación (ver incisos C y D en Figuras R9 y R10), se observó una transición de fase en la dinámica evolutiva, en la que el *fitness score* de los genotipos disminuye drásticamente (por debajo de la mitad del *fitness* inicial, igual a 1) durante las primeras 100 generaciones, y se mantiene constante a lo largo de las 1,000 generaciones. De esta forma, la mayoría de estos genotipos mutantes se sitúan por debajo del *fitness promedio* (si se usa el *fitness score* estimado con el método de dS-dN) o sobre el *fitness más bajo* (si se usa el *fitness score* estimado con el método de Tajima) que se obtiene cada 100 generaciones. Este cambio de la dinámica evolutiva es lo que nosotros sugerimos como la presencia de un *error de catástrofe*, la cual se define como la transición ocurrida al traspasar el *umbral de error*. En esta fase, se pierde el genotipo favorecido, por ende, también se pierde el comportamiento de *cuasiespecies.*

En la medida en que nuestra "amortiguación o memoria mutacional" y la intensidad de la selección sobre las mutaciones (que se refleja más intensa y con una distribución más compleja con el método Tajima que con el método dS-dN, ver Figura R8 A y B) se acerquen más a lo que pasa en las poblaciones virales naturales, se podría sugerir que los genotipos virales mantienen un comportamiento de *cuasiespecies.* Así, las cuasiespecies del virus HIV-1 se mantendrían bajo un *umbral de error* entre 1 y 2 mutaciones por genoma por ciclo de replicación (incisos A y B en Figuras R9 y R10). Del mismo modo, los genotipos más aptos (quienes conforman a la cuasiespecie) pueden ser reemplazados por genotipos con *fitness* desventajosos, pero más robustos a la acumulación de mutaciones negativas (i.e, –genotipo *flattest*–), cuando un *error de catástrofe* ocurre a partir de 3 mutaciones por genoma por ciclo de replicación (incisos C y D en Figuras R9 y R10).

Nosotros no logramos observar la *catástrofe de extinción*, la cual representa la pérdida completa de la población viral a través de mutaciones letales, ya que no incluimos la influencia de las mutaciones letales en nuestro mapa del *fitness mutacional* ni en la matriz de categorización del *fitness score*. Además, el *fitness score* que implementamos en el presente estudio pondera el peso de las mutaciones que ocurren: i) entre las generaciones de un genoma parental y su progenie, y ii) entre las mutaciones en un mismo genoma viral, al promediar sus correspondientes *fitness score*; de modo que se amortigua el efectivo selectivo o neutral de las

mutaciones al costo de mantener una "memoria mutacional". Como se ha discutido en las secciones anteriores, se ha sugerido ampliamente que el efecto de las mutaciones en el *fitness* del virus HIV-1 no es del todo independiente de otras mutaciones anteriores o paralelas en el genoma (e.g., epístasis en aminoácidos y mutaciones compensatorias en estructuras de RNA). Desafortunadamente, no es posible evaluar cuánta "memoria mutacional" se encuentra en las poblaciones naturales del virus HIV-1, por lo que la ausencia o presencia de una *catástrofe de extinción* no pudo ser evaluada satisfactoriamente en este trabajo.

Para evaluar el contenido informacional de los genotipos mutantes obtenidos en nuestras simulaciones, se construyeron árboles de distancia usando el método de *Neighbour Joining* a través del programa *Clustalw versión 2.0*. En la Figura R11 se muestra el árbol de distancia obtenido con las simulaciones para 1,000 individuos a lo largo de 1,000 generaciones. El árbol de distancia se realizó con un alineamiento de 121 genomas virales obtenidos del *set* de simulaciones correspondiente, para ello se tomaron 10 genomas al azar en las generaciones número 100, 500 y 1,000 en cada una de nuestras simulaciones bajo 1, 2, 3 y 4 mutaciones. La última secuencia del alineamiento corresponde al genoma viral de referencia con la que se inició la simulación, etiquetado en el árbol como *HXB2*.

El árbol de distancia de los genomas mutantes nos permite corroborar primeramente que la pérdida del contenido informacional se realiza de manera proporcional respecto al número de mutaciones que se generan en el genoma viral, pero sobretodo con respecto al número de mutaciones que se acumulan a lo largo de las generaciones. Así por ejemplo, el contenido informacional de aquellos genomas mutantes que han divergido apenas unas 100 generaciones son más similares entre sí y al genoma de referencia (incluso si han acumulado 4 mutaciones por ciclo de replicación) que de aquellos genomas que han acumulado 1 mutación a lo largo de 500 y 1,000 generaciones.

Por otro lado, las gráficas (R9 y R11) obtenidas de las simulaciones nos sugieren que a partir de 3 y 4 mutaciones generadas por genoma por ciclo de replicación se reduce drásticamente el *fitness score* de los genotipos más aptos en la población viral. Pese a que el *fitness score* se puede mantener alto si la replicación de los genomas virales se realiza con no más de dos mutaciones en promedio por genoma por ciclo de replicación, la diversidad informacional

(i.e., variabilidad genética) de la población evolucionando con 2 mutaciones se incrementa significativamente, con respecto a las poblaciones evolucionando con 1 mutación, al paso de un mayor número de generaciones; de hecho, casi a la misma tasa que aquellas poblaciones evolucionando con 3 y 4 mutaciones.

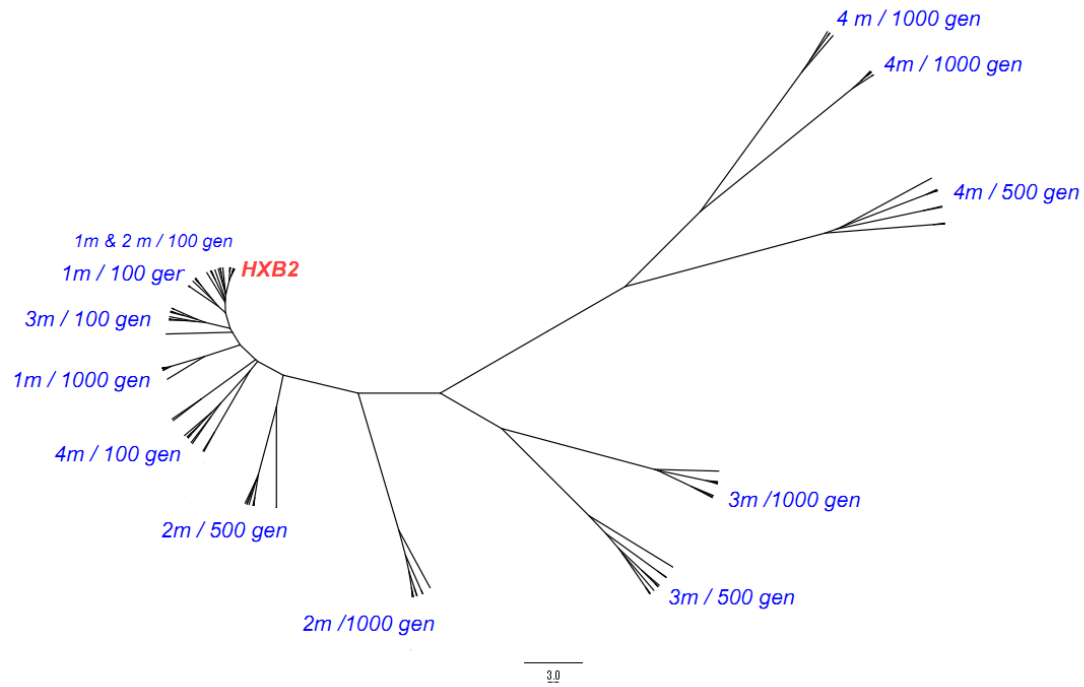


Figura R11. Árbol de distancias para las dinámicas de *cuasiespecies*. El árbol representa la distribución de 1000 genomas sujetos a 1, 2, 3 y 4 mutaciones por ciclo de replicación por genoma, durante 1000 generaciones en comparación con el genoma de referencia *HXB2* (en rojo). Dado que cada número de mutación es una simulación independiente, se extrajeron al azar 30 genomas a diferentes generaciones para cada simulación (*gen:* 100, 500 y 1000) y con su tasa de mutación correspondiente (en azul, representado como *N*m, donde *N* es el número de mutaciones por genoma por ciclo de replicación), obteniendo un total de 120 genomas. El árbol fue calculado empleando el porcentaje de similitud entre cada par de secuencias (PID). Este cálculo se realizó con la siguiente ecuación: PID= Numero de símbolos alineados *100 / numero más pequeño de posiciones en ambas secuencias. En esencia, esto representa el número de nucleótidos (bases o residuos) idénticos por 100 pares de bases.

A. *Cuasiespecies* en virus de RNA: nuestras predicciones en HIV-1 *versus* otros modelos

Al igual que la genética de las poblaciones, la teoría de las *cuasiespecies* es una de las teorías más novedosas en los últimos 40 años, ya que nos ha permitido observar nuevas dinámicas evolutivas de las poblaciones virales a lo largo del tiempo, y que se presume son diferentes de aquellas que conducen la evolución de los sistemas celulares.

Existen muchos trabajos experimentales que reportan la gran variabilidad genética de los virus de RNA. En algunos de aquellos reportes se caracteriza también la presencia de una secuencia mutante que se adapta a las presiones de selección que se le imponen después de varios ciclos de replicación, esta secuencia es llamada por los experimentalistas como la *secuencia consenso*, y extrapolan este concepto al de *cuasiespecies* en la mayoría de las ocasiones [ver Figura D1.A]. Sin embargo, el modelo original de la teoría de las *cuasiespecies* propuesto por Manfred Eigen y Peter Schuster en 1977, involucra más características que están siendo recientemente entendidas en la comunidad científica [ver Figura D1.B] (Bull et al., 2005, Wilke, 2005). Es así como algunos trabajos teóricos recientes han relacionado modelos matemáticos con ciertas estimaciones biológicas para definir el problema de las *cuasiespecies* con mayor claridad. Cuando se aplica el modelo de las *cuasiespecies* en la evolución de los virus de RNA, por ejemplo, las *cuasiespecies virales* se definen como nubes de genotipos replicantes que mantienen una distribución constante de mutantes con los *fitness* más altos a través de un equilibrio entre altas tasas de mutación y fuertes procesos de selección purificadora.

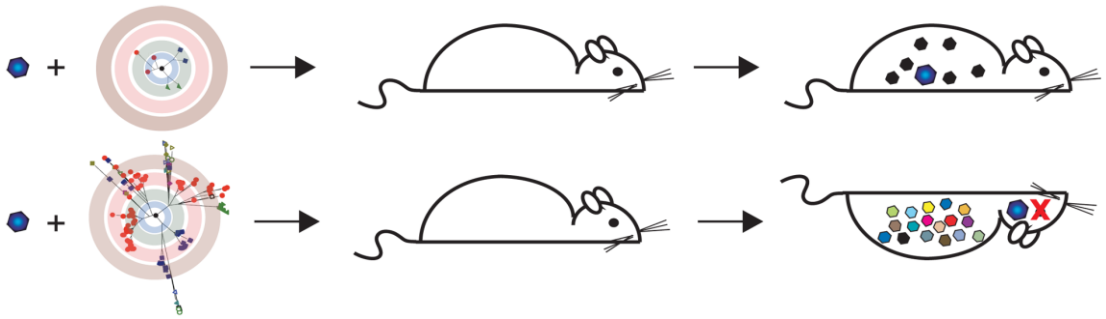


Figura D1.A. Resultados de los experimentos que se describen en Vignuzzi et al. (2006). Un clon neurovirulento de poliovirus se aisló de los cerebros de los ratones que habían sido infectados con una cepa de tipo silvestre. Los ratones no tratados fueron infectados con este clon derivado de una de una población genéticamente restringida (parte superior) o de una población genéticamente diversa (abajo). Aunque todos los ratones recibieron el clon neurovirulento, sólo aquellos infectados con una *cuasiespecies* diversa desarrollaron la enfermedad, ya que éstas facilitaron la entrada del clon neurovirulento en el sistema nervioso central. Figura tomada de Lauring y Andino (2010).

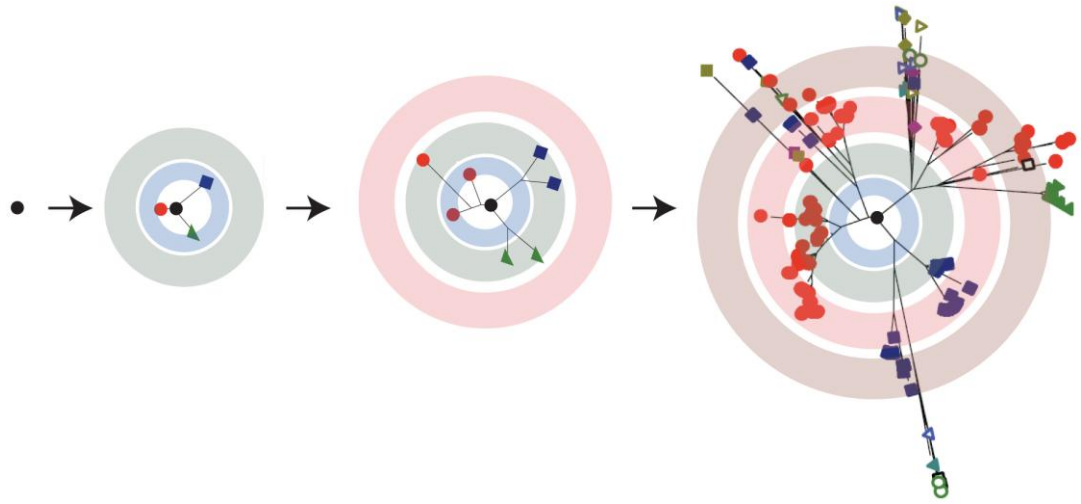


Figura D1.B. Representación de la dinámica de las *cuasiespecies*. Un virus replicando con una alta tasa de mutación generará un repertorio diverso de mutantes en el transcurso de unas pocas generaciones. En estos árboles mostrados, cada rama indica dos variantes unidas por un punto de mutación, y los círculos concéntricos representan los ciclos de replicación. La distribución resultante de estas mutantes se representa a menudo como una nube centrada en una *secuencia maestra* (punto color negro colocado en el centro del árbol). Este esquema de dos dimensiones es una simplificación excesiva de la conectividad intra-*cuasiespecies*. En las formulaciones matemáticas de la teoría de *cuasiespecies*, el espacio de secuencias es multidimensional, con numerosas ramas entre las variantes. Figura tomada de Lauring y Andino (2010).

Las propiedades más distintivas y cuantificables de la evolución de las *cuasiespecies* son el *umbral de error*, la *catástrofe de error* y la *extinción de error*. Estas propiedades son las que trataron de estimar en el presente trabajo con el algoritmo genético *quasiespecies.pl*. Una representación gráfica de estas propiedades puede ser observada en la Figura I5 (tomada de Sallie, 2005), y en la Figura D2 (tomada de Bull et al., 2005). El *umbral de error* es la tasa de mutación por debajo de la cual las poblaciones están equilibradas en un clásico balance de mutación-selección, en el cual, el genotipo favorecido puede mantener la replicación de la información a pesar de la alta tasa de mutación. El *error de catástrofe*, por otro lado, es la fase de transición ocurrida al traspasar el umbral de error, en la cual se pierde el genotipo favorecido, por ende también el comportamiento de *cuasiespecies* y posiblemente también la identidad viral. Finalmente, la *catástrofe de extinción* representa la perdida completa de la población viral a través de mutaciones letales.

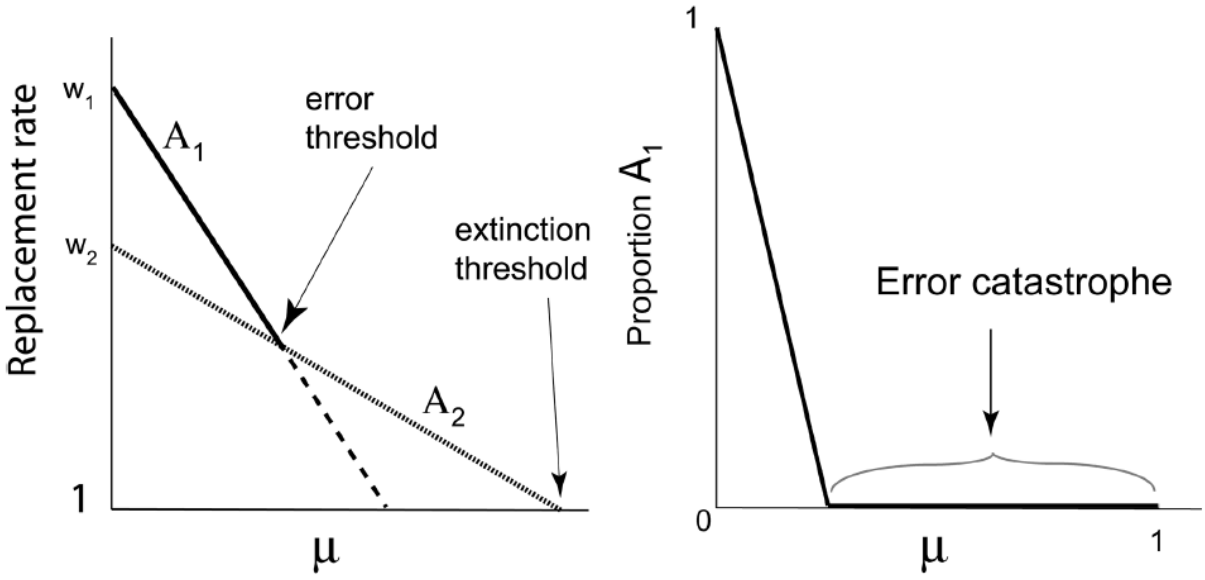


Figura D2. El umbral de error, la catástrofe de error y el error de extinción. Panel izquierdo: se muestra la tasa de sustitución de dos genotipos: A1 y A2. Cada uno de los genotipos presenta un *fitness* y una tasa de mutación diferente; para A1, esto se representa como w1(1-μ1). A1 es mostrado como líneas negras sólidas. Para el genotipo A2, su tasa de mutación y *fitness* corresponde a W2(1- μ2). A2 es mostrado como líneas grises. Bull et al, manejaron que la mutación sea μ1=kμ y a μ2=μ, con una k>1, por lo que las dos tasas de sustitución se pueden representar como líneas en el mismo plano. Puesto que el *fitness* del genotipo A1 (W1) es mucho mayor que el del genotipo A2 (W2), la restricción sobre k asegura una alta tasa de sustitución de A1 más que en A2 a una tasa baja de mutación. El punto en el cual las líneas se interceptan, es identificado como el *umbral de error*. Por lo tanto, para ilustrar la pérdida del genotipo A1 se usan guiones para mostrar la función de la tasa de sustitución. Si la tasa de sustitución cae por debajo de la unidad, la población se extingue, por lo que las funciones no se extienden por debajo de uno en el eje vertical. Panel derecho: la frecuencia de A1 decae cuando la tasa de mutación incrementa hasta que el *umbral de error* sea alcanzado. En las tasas de mutación más altas, sólo A2 está presente. La disminución en la frecuencia de A1 con una tasa de mutación hacia el *umbral de error* puede ser lineal (como se muestra aquí), cóncava o convexa, en función de los valores de los parámetros (lado de derecho ilustrado por w1=3.0, w2=2 y k=2).

Hemos seleccionado cuatro trabajos que simulan del modelo de *cuasiespecies* en diferentes sistemas para analizar comparativamente su realismo biológico y sus hallazgos contra los nuestros. Las características de estas aportaciones, incluyendo la nuestra, se muestran de forma comparativa en la Tabla D1. Aunque cada uno de estos reportes trata de modelar la dinámica de las *cuasiespecies* con gran apego al modelo original propuesto por Eigen y Schuster (1977), los parámetros "biológicos" empleados son aún limitados en todos los trabajos.

Tabla D1. Comparación del análisis realizado sobre el modelo de las *cuasiespecies* por diferentes trabajos.

	N_e/N[A]	Mutación (rango)[B]	Fitness (rango)[C]	Tiempo[D]	Longitud[E]	Fitness Mutacional[F]	Selección[G]	Genomas (información)[H]
Holmes et al., 2000 ♦	200,000	0.1-2.5	0.1 a 1.0	1,000 2,000	100	Biológico: codificante	neutral & negativa	A, T, G, C
Codoñer, et al., 2006♣	2,000	0.01-0.1	F(k)	40	64	Probabilidad estadistica[1]	Negativa	0, 1 (bits)
Bull et al., 2005♦	2*	0-1.0	W_1 y W_2**	20	¿?*	Probabilidad Arbitraria	neutral & negativa	0,1 (bits)
Sardanyés et al., 2008♣	500	0.01-0.072	1 a 0.003	1,000	32	Probabilidad estadistica[2]	Negativa	0,1 (bits)
Our research♦	100 1000	1-4	F <1> F	1,000 2,000	9719	Biológica: codificante & no codificante	Positiva & neutral negativa	A, T, G, C

[A] = Tamaño de la muestra o número de individuos; [B] = tasa de mutación; [C] = rango de *fitness* empleado para la secuencia maestra (F=1) y los mutantes (<1); [D] = tiempo en ciclos; [F] = longitud del genoma; [G] = tipo de selección tomada en cuenta para las simulaciones; [H] = tipo de información biológica en el genoma: cadena de bits o cadena de nucleótidos. [1, 2] Decaimiento de la información por acumulación de mutaciones. ♦ Se define a una sola cuasiespecie. ♣ Se definen a dos *cuasiespecies*.
* = No determinado en la publicación; ** *fitness arbitrario*: W_1 = máximo *fitness*, W_2 = *fitness* mutacional (población robusta).
♠ Población infinita o finita, altas tasas de mutación, cadena de información basadas en estimaciones estructurales y funciones (estructura de RNA) o perdida de la identidad por acumulación de mutaciones (perdida de la información).

Por ejemplo, Codoñer et al. (2006), Sardanyés et al. (2008) y Bull et al. (2005), emplean una cadena de bits (0-1) de *n* dígitos (en lugar de nucleótidos) para representar a ciertos genomas virales o viroides (digitales). La longitud de tales cadenas generalmente es más pequeña respecto a la que existen en los sistemas biológicos como las bacterias y los virus. En este tipo de investigaciones es común que la *secuencia maestra* esté compuesta así de una cadena uniforme de ceros de una longitud *L*. Cada cambio (mutación) que surge en el proceso de replicación (ciclos) es representado como 0→1. Así, cada genoma se va distanciando del original dependiendo del número de los cambios que haya acumulado. El número de ciclos o generaciones en estos reportes puede ir desde los 40 hasta los 2,000, mientras que el N_e usado va desde 2 hasta los 200,000 individuos o genotipos. El cambio/mutación es aleatorio en tales trabajos, sin embargo, cada posición (o celda, como se le conoce), viene acompañada de una probabilidad de mutación que la hace más o menos propensa a dichas mutaciones; cada evento de mutación se realiza a una tasa específica cada determinados ciclos de replicación. La mutaciones realizadas en estas simulaciones afectan, siempre negativamente, la capacidad de

replicación (*fitness*) de dichos genomas, e.g. si una secuencia progenie de $L = 20$ acumula 3 errores con respecto a la original, el efecto de la mutación será del 15% (aunque esto depende del efecto numérico sobre la posición mutada). En muy contados trabajos, se han introducido una fracción pequeña de mutaciones neutrales en las dinámicas de simulación de *cuasiespecies* (lo cual se discutirá más adelante). De este modo, en este tipo de simulaciones, es usual emplear a la mutación como un proceso degenerativo que reduce el *fitness*.

En la presente investigación, por el contrario, nosotros empleamos el genoma completo del virus modelo (en nucleótidos) para poder simular el proceso de replicación, mutación y selección. Asimismo, nosotros nos basamos en reportes experimentales y estimaciones bioinformáticas para obtener una aproximación biológica de dichos parámetros más que estimaciones teóricas basadas en la probabilidad. A continuación hacemos una breve discusión sobre las diferencias entre los parámetros usados y las estimaciones realizadas entre los cinco trabajos que modelan *cuasiespecies*.

Sobre la tasa de mutación. A diferencia de las investigaciones de Sardanyés et al. (2008), Codoñer et al. (2006) y Bull et al. (2005), nosotros usamos una tasa de mutación más alta, que va desde 1 a 4 mutaciones por genoma por ciclo de replicación. Como se ha discutido ampliamente en los resultados, si la tasa de mutación para el virus HIV-1 bien podría ser ligeramente menor o mayor en las poblaciones virales naturales, la tasa de mutación que nosotros usamos aquí bien cubre un amplio intervalo del espectro mutacional estimado por diversos reportes experimentales y teóricos para HIV-1. Sin embargo, en todas las simulaciones realizadas, incluyendo la nuestra, sólo se asume que ocurren mutaciones puntuales en el modelo. Está ampliamente reportado que mutaciones del tipo *frameshift* y recombinaciones también se encuentran en alta proporción en los genomas virales [ver los estimados para el virus HIV-1 en Tabla R1]. No obstante, el efecto que tales mutaciones tiene en la tasa de mutación o en el *fitness* del individuo es raramente conocido o cuantificado. La cuantificación de las mutaciones reversas constituye otro parámetro que permitiría simular con más realismo la frecuencia del tipo de mutaciones. En particular, la recombinación puede jugar dos papeles sobre el *umbral de error* y las poblaciones virales, 1) generar mutantes más vulnerables a una *catástrofe de error*, en caso de que la recombinación contribuya significativamente en la tasa de mutación, influenciando así directamente en el *umbral de*

error, y 2) generando mutantes que recombinan para regresar al genotipo original (secuencia maestra) (Bull et al., 2005).

Sobre el fitness mutacional. Un parte importante de la dinámica evolutiva de cualquier sistema es medir el efecto de las mutaciones en el *fitness reproductivo* de la siguiente generación, referido como *fitness mutacional* en la Tabla R8. Todos los trabajos sobre *cuasiespecies* a la fecha, excepto nosotros, estiman el *fitness mutacional* en base a estimados estadísticos, donde todas las mutaciones tiene un efecto negativo (Codoñer et al., 2006, Sardanyés et al., 2008) o neutro (Bull et al., 2005) sobre el *fitness* del individuo. En el presente trabajo una mutación puede tener un efecto negativo, neutral o positivo sobre el *fitness reproductivo* del virus. En nuestro modelo, las mutaciones no actúan de forma independiente (*modelo acumulativo dependiente*), a diferencia de otros modelos en los que los eventos de mutación son tratados como eventos *acumulativos independientes*. De modo que en nuestro modelo, los *fitness scores* del genoma parental y del genoma en la presente generación se promedia, así como también se promedian los *fitness score*s de todas las mutaciones que ocurren en un ciclo de replicación. Sin embargo, esta "memoria o dependencia mutacional" no se considera como un fenómeno de epístasis propiamente en nuestras simulaciones, ya que para ello requeriríamos conocer de antemano (a través de la experimentación por ejemplo) si la interacción de dos mutaciones presenta en la realidad un *fitness scores* por arriba o por abajo del promedio del genoma viral en la generación presente. Por ejemplo, si dos mutaciones negativas pueden derivar en un *fitness positivo*. Sin embargo, este proceso puede explicar la presencia y persistencia de una población robusta en nuestros resultados.

Otra limitación, tal vez una de las más importantes, sobre el *fitness mutacional* utilizado para modelar *cuasiespecies* en el presente trabajo, es que nuestra estimación es fija en el tiempo. Nosotros hemos calculado un *fitness* mutacional derivado de la historia mutacional depositada en los genomas completos del virus HIV-1. Este *fitness* mutacional representa la gran ventaja de evaluar la influencia de casi cada posición del genoma completo con base al resultado de la historia evolutiva de la población del virus HIV-1, la cual ha estado sujeta a diversos ambientes celulares y presiones de selección en diferentes localidades del planeta en los últimos 30 años. Sin embargo, sabemos que el efecto de una mutación en un tiempo t, puede cambiar drásticamente al tiempo t+1 en la naturaleza; de modo que el *fitness mutacional* de

una mutación en una posición específica del genoma viral no es constante a través del tiempo. Por ejemplo, una mutación neutral puede presentar *fitness* opuestos según el grupo de condiciones ambientales a los que esté sujeta en diferentes generaciones; de este modo, las mutaciones que pueden dirigir la variación antigénica o la resistencia antiviral pueden ser neutrales o deletéreas en la ausencia de anticuerpos o drogas antivirales o, en su caso, pueden ser ventajosas en la presencia de anticuerpos o drogas antivirales (Quer et al., 2001).

La presencia de mutaciones neutrales es de particular importancia en el modelo de las *cuasiespecies*, ya que se desestimó inicialmente su frecuencia en virus con genomas muy pequeños. Las *cuasiespecies* (conformada por "nubes o conjuntos" de genotipos) optimizan un mismo *fenotipo consenso* a través de caminos aleatorios (Schuster, 2003). Así, el conjunto de genotipos que mantienen un fenotipo particular se le conoce como *red neutral* [ver Figura D1.B] (Fontana and Schuster, 1987). De esta manera, la presencia de más de una mutación neutra puede alterar "el vecindario fenotípico" de un individuo, es decir, puede incrementar el grupo de distintos fenotipos al que puede acceder un genotipo a través de la adquisición de más mutaciones (Draghi et al., 2010), por lo que las características distintivas de las *cuasiespecies* (*umbral de error*, la *catástrofe de error* y la *extinción de error*) podrían tener a lugar sobre largos periodos generacionales (aún bajo altas tasas de mutación), o bien, nunca tener a lugar. Nuestras simulaciones muestran que las propiedades del comportamiento del tipo *cuasiespecies* se presenta durante la dinámica evolutiva de los genotipos virales de HIV-1 aún cuando el genoma viral posee potencialmente una considerable cantidad de mutaciones neutrales [ver Figura R5] (aún si menos de la mitad de estas mutaciones se encuentra sujeta a la selección positiva o negativa debido a una inadecuada o sesgada estimación metodológica). De este modo, nuestras simulaciones muestran el surgimiento de una población robusta (*survival of the flattest*) cuando los genotipos replican con 3 y 4 mutaciones por genoma por ciclo de replicación, lo cual refiere a la habilidad que tienen los fenotipos menos aptos –con grandes redes neutrales (*flattest*)– para desplazar a los fenotipos más aptos pero que poseen pequeñas redes neutrales (*fittest*) [ver Figura D3] (Wilke et al., 2001, Bull et al., 2005).

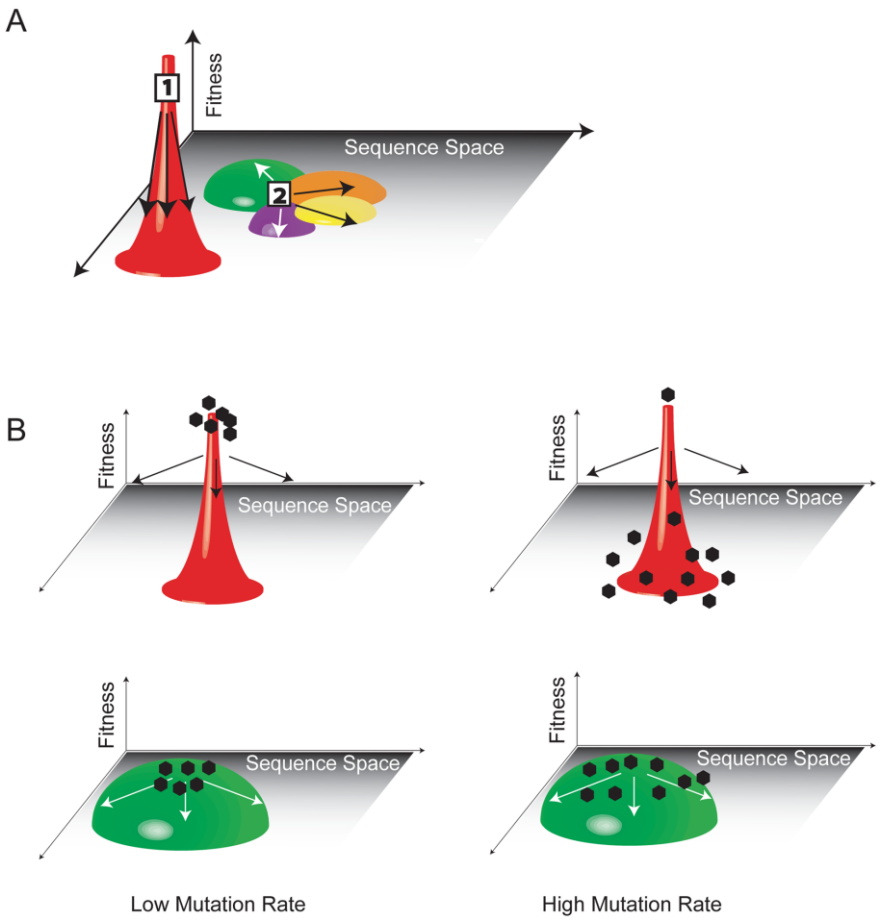


Figura D3. Representación gráfica del *efecto cuasiespecies*. (A) La población 1 tiene un *fitness* alto, pero se encuentra atrapado en el *espacio de secuencia* debido a que las mutaciones la conduce a una pérdida dramática del *fitness*. La población 2 es mutacionalmente más robusta debido a que la mutación conduce a una menor pérdida del *fitness*. La población más plana (*flattest*) está idealmente situada para moverse a través de la secuencia y acceder a otros picos locales del *fitness*, a través de las redes mutacionales de sus vecinos (indicados en diferentes colores). (B) A una tasa de mutación baja, las variantes serán genotípicamente estables y estarán agrupadas en la parte superior del pico de *fitness*. La variante con *fitness* alto será fácilmente superada por otros. A altas tasas de mutación, las variantes se extenderán hacia fuera del espacio de secuencia a través de los picos de *fitness* correspondientes. Las variantes sobre picos planos (en verde), se mantienen cerca de su *fitness* óptimo y tienen una mayor media del *fitness* que la población ubicada en el pico más pronunciado (ilustrado en rojo). La población más plana prevalecerá. Figura tomada de Lauring y Andino (2010).

Sobre la presencia de cuasiespecies en las dinámicas evolutivas. En una investigación hecha por Jenkins, Holmes y colaboradores (2001) se argumenta que no es posible concluir que las poblaciones virales del Virus de la Estomatitis Vesicular (VSV) y del *Foot-and-Mouth Disease Virus* (FMDV) se encuentren evolucionando como *cuasiespecies*. Este reporte es de particular importancia para nosotros porque es el único que incluye un similar realismo biológico en la modelación de las *cuasiespecies*. Holmes y colaboradores estiman un *fitness mutacional* con base la estructura del RNA, sitios sujetos a selección neutral y negativa a

través de los métodos dS-dN y *sesgo de codones*. Los autores realizan un programa que simula el modelo original de Eigen-Schuster empleando sólo un gen de dos diferentes virus: el gen G de VSV y el gen P1 de FMDV. Sus simulaciones se realizan con una N_e de 200,000 secuencias de 100 nucleótidos de longitud, de los cuales, sólo 80 sitios se encuentran sujetos a selección negativa (los sitios restantes son neutrales); las simulaciones se llevan a cabo a lo largo de 1,000 y 2,000 generaciones. Los resultados encontrados en este trabajo muestran que la *secuencia consenso* se pierde a una tasa de 0.9 (en FMDV) y 2.3 (en VSV) mutaciones por secuencia por ciclo de replicación. Cuando se incrementa el *fitness* de las mutaciones y el tamaño de la población, el *umbral de error* incrementó. Sin embargo, Holmes y colaboradores argumentan que la pérdida secuencia consenso no es suficiente para concluir que las poblaciones virales de estos virus se encuentra evolucionando como *cuasiespecies*. Por el contrario, los autores argumentan que los resultados de sus simulaciones sugieren que los genotipos virales del VSV y FMDV evolucionan bajo el *modelo del camino aleatorio* de la genética de poblaciones, el cual explica que la frecuencia de los nuevos genotipos mutantes que surgen en la población depende de la deriva génica; de modo que la frecuencia de cada alelo es un camino aleatorio entre la fijación y la perdida.

Contrario a Holmes y colaboradores (2001), nuestros resultados y los de otros grupos de trabajo muestran que las *cuasiespecies* se forman a partir de una alta tasa de mutación alcanzando un equilibrio en el tiempo, donde coexisten poblaciones heterogéneas con altos valores de *fitness*. En consecuencia, se logra identificar un *umbral de error*, bajo el cual se mantiene del comportamiento de *cuasiespecies* de los sistemas modelo (e.g., virus, secuencias digitales) que fueron analizados. En general, el *umbral de error* en otros reportes se localiza alrededor de 'la mitad' del espectro de tasas de mutación que ellos emplearon, *i.e.* ~0.05 (Codoñer et al., 2006), 0.5 (Bull et al., 2005), 0.078 (Sardanyés et al., 2008) y, en caso nuestro, este umbral se localiza en 3 mutaciones por ciclo de replicación para el virus HIV-1 [ver incisos A y B en Figuras R9 y R10]. En consecuencia, todos los reportes muestran la presencia de una *catástrofe de error* en las simulaciones cuando el umbral de error fue sobrepasado [ver incisos C y D en Figuras R9 y R10 en el presente trabajo, y Figuras 1, 8 y 18 en Bull et al., 2005; Figura 3 en Codoñer et al., 2006; Figura 4 en Sardanyés et al., 2008]. Así, se pueden perder las *cuasiespecies* de la evolución viral (de forma temporal o permanente), pero surgen en compensación genotipos redundantes y más robustos a las mutaciones (*el*

efecto cuasiespecies) [ver Figura D3]. La ausencia de una *extinción de error* (i.e., la extinción viral) también fue reportada en todos los trabajos, incluyendo el presente. La Figura D4 muestra un resumen de la dinámica de *cuasiespecies* identificada en nuestro trabajo, así como también se esquematiza sobre esta dinámica la participación de otros factores que no fueron analizados en el presente proyecto, tales como la epístasis, las mutaciones compensatorias y las redes neutrales.

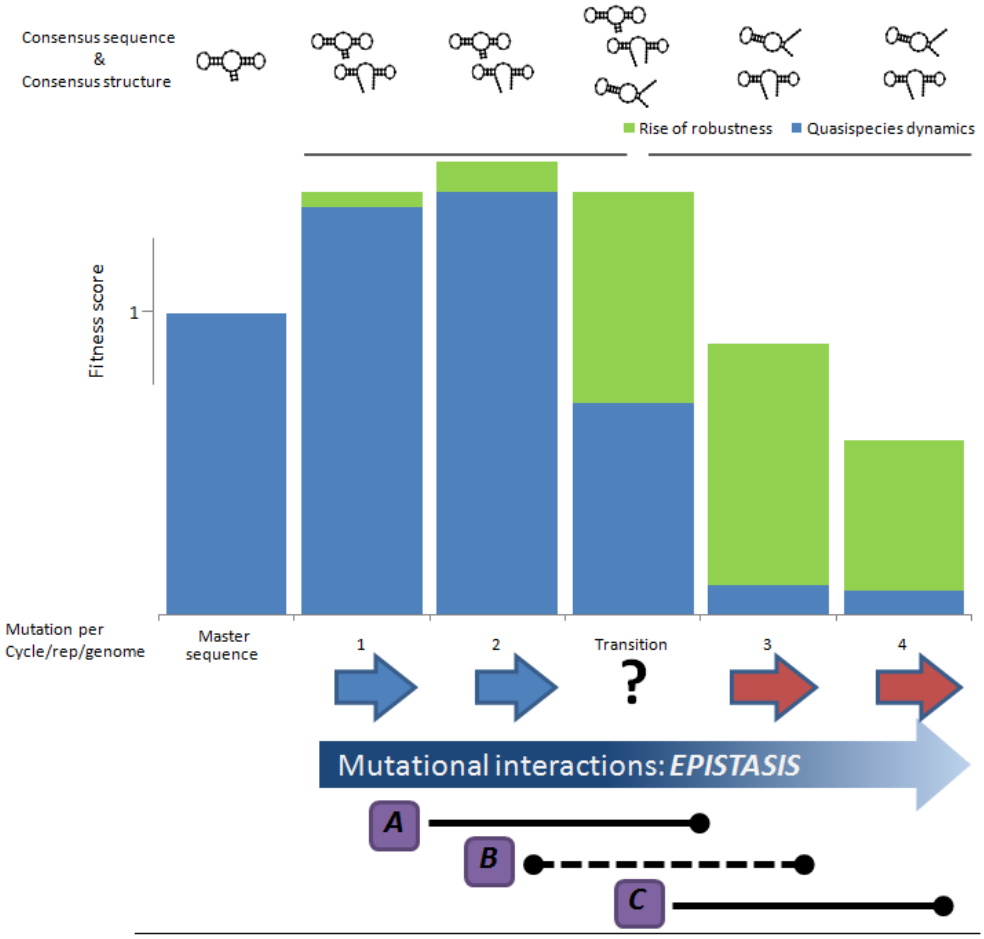


Figura D4. Dinámicas de las *Cuasiespecies* y el surgimiento de la Robustez. A partir de una población inicial (secuencia maestra, primera columna azul) y a través de una alta tasa de mutación, es posible obtener *cuasiespecies* con una sólida estructura del genotipo, tanto consenso como de la secuencia maestra (parte superior). Posteriormente esta es desplazada por otra población con más robustez mutacional, finalmente las *cuasiespecies* se pierden. Un aumento progresivo de la tasa de mutación induce la presencia de un umbral de error (2 mutaciones) y, consecutivamente, un error de catástrofe con 3 mutaciones (un ejemplo de ellos son las proteínas APOBEC que provocan hipermutaciones). En este proceso de transición (umbral → catástrofe) ocurren cambios estructurales drásticos en toda la población y, como consecuencia, se pierde la secuencia maestra y consenso (arriba). Con 2 o más mutaciones se incrementan las interacciones induciendo diferentes tipos epístasis (flecha de abajo). En **A.** *Epístasis sinérgica positiva* (aumenta el *fitness*), en **B.** *Epístasis sinérgica negativa* (disminuye el *fitness*), la cual se encuentra en el área de transición (**?**) y, en **C.** *Epístasis antagónica positiva* (amortigua e "incrementa el *fitness*").

CONCLUSIONES

Tomando en cuenta las interpretaciones teóricas y cuantificables más recientes sobre el modelo de las *cuasiespecies*, nosotros podemos concluir que las poblaciones del Virus de la Inmunodeficiencia Humana 1, un virus con genoma de RNA, se encuentran muy probablemente evolucionando como *cuasiespecies*. Nuestras simulaciones muestran la presencia de un *umbral de error* situado no más allá de 2 mutaciones por genoma por ciclo de replicación, logrando obtener una secuencia consenso sobre la que se encuentra la mayor frecuencia de mutantes con los *fitness mutacionales* más altos del sistema que analizamos. En consecuencia, también observamos una *catástrofe de error* cuando la población de genotipos virales se sometió a replicar con más de 3 mutaciones por ciclo de replicación. En esta fase evolutiva, la mayor frecuencia de los genotipos mutantes de localizó sobre los *fitness* localizados por abajo del promedio y en los más bajos de nuestras simulaciones. Pese a que estos genotipos mutantes poseen bajos *fitness*, se pueden mantener en la población a lo largo de las 1,000 generaciones simuladas debido a una robustez mutacional ante altas tasas de mutación. Nosotros no observamos una *catástrofe de extinción* probablemente debido a la ausencia de mutaciones letales en las estimaciones del *fitness mutacional* y *fitness score*. Sin embargo, nuestras predicciones carecen de significancia estadística rigurosa para validar la repetitividad de nuestros resultados, problema que se resolverá al incrementar la eficiencia computacional de nuestro algoritmo computacional *quasiespecies.pl* en un futuro cercano. Asimismo, una mayor eficiencia computacional de nuestro algoritmo permitirá simular poblaciones virales con estimaciones de N_e y un número de generaciones más grandes.

Como parte de la incorporación de más "realismo biológico" al modelo original de las *cuasiespecies*, creamos un mapa del *fitness mutacional* del genoma completo del virus HIV-1 (por genes, dominios funcionales y codones) a fin de estimar un *fitness mutacional* para cada posición del genoma y un *fitness reproductivo* de las mutaciones generadas al azar. Este mapa nos permite observar que aunque la variación y la conservación nucleotídica se observan a lo largo del genoma del virus HIV-1, ambas también muestran una tendencia a estar limitadas o concentradas en regiones genómicas y estructurales específicas; asimismo, es posible mostrar que predomina en particular el número de tripletes nucleotídicos seleccionados negativamente respecto a los positivos y neutros. De esta forma, la primera parte del genoma, conformada por

los genes *5´LTR, gag, gag-pol, pol, vif* y *vpr* (y encargada de la replicación del genoma viral y de su integración en el genoma del hospedero), se encuentra potencialmente sujeta a la influencia dominante de la *selección negativa*; mientras que la segunda mitad, conformada por los genes *tat, rev, vpu, env, nef* y *3'LTR* (y encargada de la interacción con la célula hospedera), se encuentra potencialmente sujeta a la acción principal de la *selección positiva.* Se analizaron diversos aspectos de la estructura genómica, evolutiva y funcional al nivel de las estructuras de RNA y las proteínas del virus HIV-1 que podrían soportar y explicar nuestras predicciones. Además, los resultados del presente trabajo son similares a los reportados por otros grupos de trabajo de forma independiente en diversas regiones del genoma de HIV-1.

Todos estos resultados en conjunto sugerirían que la *presencia de una fuerte selección purificadora en genomas pequeños que se encuentran sujetos particularmente a altas tasas de mutación*, podría conducir al comportamiento de *cuasiespecies* de forma casi determinista, tal y como lo sugirieron Eigen-Schuster en su modelo original. Sin embargo, las profundas consecuencias de la limitación del tamaño de los genomas de RNA, impuestos por las altas tasas de mutación y una fuerte selección purificadora, requieren reevaluar la naturaleza de aquellos sistemas biológicos que enmascaran la evolución dependiente entre estructuras de RNA y proteínas que son codificadas por la misma región genómica. Las altas tasas de mutación en los virus de RNA proporcionan, al menos teóricamente, el potencial de una adaptación molecular rápida (e.g., a las defensas inmunitarias adaptativas o la adquisición de nuevas enfermedades por transmisión zoonótica). Sin embargo, tal adaptación está, paradójicamente, muy limitada por la naturaleza de los genomas de los virus de RNA que, debido a su pequeño tamaño, presentan altos niveles de pleiotropía, mutaciones compensatorias, epístasis y evolución convergente. De modo que, en lugar de ser infinitamente flexibles, los virus de RNA deben verse como *bestias inquietas en una jaula pequeña*, ya que son capaces de adaptarse rápidamente al cambiar sólo entre una colección de fenotipos limitados. De esta forma, nosotros esperamos que la revisión teórica, experimental y los resultados mostrados en la presente tesis ayuden a complementar o sugerir nuevas ideas sobre el entendimiento de las *cuasiespecies* y la evolución de los sistemas virales.

REFERENCIAS

AASKOV, J., BUZACOTT, K., THU, H. M., LOWRY, K. & HOLMES, E. C. 2006. Long-term trnsmission of defective RNA viruses in humans and Aedes mosquites. *Science,* 311**,** 236-238.

ABRAM, M., FERRIS, A., SHAO, W., ALVORD, W. & HUGHES, S. 2010. Nature, position, and frequency mutation made in a single cycle of HIV-1 replication. *J Virol,* 84**,** 9864-9878.

ANDERSON, J., DAIFUKU, R. & LOEB, L. 2004. Viral error catastrophe by mutagenic nucleosides. *Annu Rev Microbiol,* 58**,** 183-205.

ANDERSSON, S. G. E., ZOMORODIPOUR, A., ANDERSSON J.O., SICHERITZ-PONTÉN, T., ALSMARK, U. C. M., PODOWSKI, R. M., NÄSLUND, A. K. N., ERIKSSON A.S., WINKLER, H. H. & KURLAND, C. G. 1998. The genome suequence of Rickettsia prowazenkii and the origin of mitochondria. *Nature,* 296**,** 133-143.

ANISIMOVA, M., NIELSEN, R. & YANG, Z. 2003. Effect recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics,* 164**,** 1229-1236.

ARSLAN, D., LEGENDRE, M., SELTZER, V., ABERGEL, C. & CLAVERIE, J. M. 2011. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc Nat Acad Sci,* 108**,** 17486-17491.

BELSHAW, R., GARDNER, A., RAMBAUT, A. & PYBUS, O. G. 2008. Pacing a small cage: Mutation and RNA viruses. *Trends in Ecol and Evol,* 23**,** 188-193.

BELL, P. J. L. 2001. Viral Eukaryogenesis: Was the ancestor of the nucleus a complex DNA virus? *J Mol Evol,* 53**,** 251-256.

BERKHOUT, B. 1992. Structural features in TAR RNA of human and simian immunodeficiency viruses: A phylogenetic analysis. *Nucl. Acids Res.,* 20**,** 27-31.

BIEBRICHER, C. K. & EIGEN, M. 2005. The error threshold. *Virus Res,* 107**,** 117-127.

BONHOEFFER, S., CHAPPEY, C., PARKIN, C. T., WHITCOMB, J. M. & PETROPOULOS, C. J. 2004a. Evidence for positive epistasis in HIV-1. *Science,* 306**,** 1547-1550.

BONHOEFFER, S., CHAPPEY, C., PARKIN, C. T., WITHCOMB, J. M. & CHRISTOS, J. P. 2004b. Evidence for Positive Epistasis in HIV-1. *Science,* 306**,** 1547-1550.

BULL, J., MEYERS, L. & LACHMANN, M. 2005. Quasispecies made simple. *PLoS Comput BIol,* 1**,** e61.

BULL, J., SANJUÁN, R. & WILKE, C. O. 2007. Theory of lethal mutagenesis for viruses. *J Virol,* 81**,** 2930-2939.

BURCH, C. L. & CHAO, L. 1999. Evolution by small steps and rugged landscape in the RNA virus phi6. *Genetics,* 151**,** 921-927.

BURCH, C. L. & CHAO, L. 2000. Evolvability of an RNA virus is determinated by its mutational neighbourhood. *Nature,* 406**,** 625-628.

CAMPO, D. S., Z. DIMITROVA, R.J. MITCHELL, J. LARA & KHUDYAKOV, Y. 2008. Coordinated evolution of the hepatitis C virus. *Proc Nat Acad Sci,* 105**,** 9685-9690.

CANN, A. J. 2005. *Principles of Molecular Virology,* San Diego, California.

CARTER, J. B. & SAUNDER, V. A. 2007. *Virology: Principles and Applications,* Oxford, UK, John Wiley & Sons, Ltd.

CASTILLO, A. C. 2007. La selección natural a nivel molecular. *In:* EGUIARTE, L. E., SOUZA, V. & AGUIRRE, X. (eds.) *Ecología Molecular.* Mexico, D.F.: CONABIO.

CLARK, S. J., SAAG, M., DECKER, W. D., CAMPBELL-HILL, S., ROBERSON, J. L., VELDKAMP, P. J., KAPPES, J. C. & SHAW, G. M. 1991. High titers of cytopathic virus in plasma of patients with symptomatic primary HIV-1 infection. *N Engl J Med,* 324**,** 954-960.

COCHRANE, G., AKHTAR, R., BONFIELD, J., BOWER, L., DEMIRALP, F., FARUQUE, N., GIBSON, R., HOAD, G., HUBBARD, T., HUNTER, C., JANG, M., JUHOS, S., LEINONEN, R., LEONARD, S., LIN, Q., LOPEZ, R., LORENC, D., MCWILLIAM, H., MUKHERJEE, G., PLAISTER, S., HOOPER, P., VAUGHAN, R., ZALUNIN, V. & BIRNEY, E. 2008. Petabyte-scale innovations at the European Nucleotide Archive. *Nucl Acids Res,* 37**,** D19-25.

CODOÑER, F. M., DARÒS, J. A., SOLÉ, R. V. & ELENA, S. F. 2006. The fittest verus the flattest: Experimental confirmation of the Quasispecies effect with subviral pathogens. *PLoS Pathog,* 2**,** e136.

COMAS, I., MOYA, A. & GONZALES-CANDELAS, F. 2005. Validating viral quasispecies with digital organisms: A re-examination of the critical mutation rate. *BMC Evol Biol,* 5**,** 5.

CONTRERAS, A. M., HIASA, Y., HE, W., TERELLA, A., SCHMIDT, E. V. & R.T., C. 2002. Viral RNA mutations are region specific and increased byribavirin in a full-length hepatitis C virus replication system. *J Virol,* 76**,** 8505-8517.

CHAO, L., RANG, C. U. & WONG, L. E. 2002. Distribution of spontaneous mutants and inferences about the replication mode of the RNA bacteriophage. *J Virol,* 76**,** 3276-3281.

CHARE, E. R. & HOLMES, E. C. 2004. Selection pressures in the capsid genes of plant RNA viruses reflect mode of transmission. *J Gen Virol,* 85**,** 3149-3157.

CHEN, H., DI-MASCIO, M., PERELSON, A., HO, D. & ZHANG, L. 2007. Determination of virus burst size in vivo using a single-cycle SIV in rhesus macaques. *Proc Nat Acad Sci,* 104**,** 19079-19084.

CHUN, T. W., CARRUTH, L., FINZI, D., SHEN, X., DIGUSEPPE, J. A., TAYLOR, H., HERMANKOVA, M., CHADWICK, K., MARGOLICK, J., QUINN, T. C., KUO, Y. H., BROOKMEYER, R., ZEIGER, M. A., BARDITCH-CROVO, P. & SILICIANO, R. F. 1997. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature,* 387**,** 183-188.

DAAR, E. S., MOUDGIL, T., MEYER, R. D. & HO, D. D. 1991. Transient high levels of viremia in patients with primary human inmunodeciency virus type 1 infection. *N Engl J Med,* 324**,** 961-964.

DAMGAARD, C. K., ANDERSEN, E. S., KNUDSEN, B., GORODKIN, J. & KJEMS, J. 2004. RNA interactions in the 5' region of the HIV-1 genome. *J Mol Evol,* 336**,** 369-379.

DARWIN, C. 1869. *On the origin of the species: By means of Natural Selection, or the preservation of favoured races in th struggle for life,* London, UK, John Murray.

DE BOER, R. J., R.M. RIBEIRO & PERELSON, A. S. 2010. Current estimantes for HIV-1 production imply rapid viral clearance in lymphoid tissue. *PLoS Comput BIol,* 6**,** e1000906.

DE OLIVEIRA, T., SALEMI, M., GORDON, M., VANDAMME, A. M., VAN RENSBURG, J., ENGELBRECHT, S., COOVADIA, H. M. & CASSOL, S. 2004. Mapping sites of

positive selection and amino acid diversification in the HIV genome: an alternative approach to vaccine design? *Genetics,* 167**,** 1047-1058.
DE VISSER, J. A. G. M., COOPER, T. F. & ELENA, S. F. 2011. The causes of epistasis. *Proc R Soc B,* doi:10.1098/rspb.2011.1537.
DELPORT, W., A. F. POON, S.D. FROST & KOSAKOVSKY, S. L. P. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics,* 26**,** 2455-2457.
DIMITROV, D., WILLEY, R., SATO, H., CHANG, L., BLUMENTHAL, R. & MARTIN, M. 1993. Quantitation of human inmunodeficiency virus type 1 infection kinetics. *J Virol,* 67**,** 2182-2190.
DOMINGO, E., ESCARMÍS, C., LÁZARO, E. & MANRUBIA, S. 2005. Quasispecies dynamics and RNA virus extinction. *Virus Res,* 107**,** 129-139.
DOMINGO, E., ESCARMÍS, C., MENÉNDEZ-ARIAS, L., PERALES, C., HERRERA, M., NOVELLA, I. S. & HOLLAND, J. J. 2008. Viral Quasispecies: Dynamics, interactions, and pathogenesis. *In:* DOMINGO, E., PARISH, C. R. & HOLLAND, J. J. (eds.) *Origin and evolution of viruses.* London, UK: Academic Press Elsevier.
DOMINGO, E., FLAVELL, R. A. & WEISSMANN, C. 1976. In vitro site-directed mutagenesis: generation and properties of an infecious extracistronic mutant of bacteriophage Qb. *Gene,* 1**,** 3-25.
DOMINGO, E., HOLLAND, J., BIEBRICHER, C. & EIGEN, M. 1997. Quasi-species: the concept and word. *In:* GIBBS, A. J., CALISHER, C. H. & GARCÍA-ARENAL, F. (eds.) *Molecular basis of virus evolution.* New York, USA: Cambridge University Press.
DOMINGO, E. & HOLLAND, J. J. 1997. RNA virus mutations and fitness for survival. *Annu Rev Microbiol,* 51**,** 151-178.
DOMINGO, E., SABO, D., TANIGUCHI, T. & WEISSMANN, C. 1978. Nucleotide sequence heterogeneity of an RNA phage population. *Cell,* 13**,** 735-744.
DRAGHI, J. A., PARSONS, T. L., WAGNER, G. P. & PLOTKIN, J. B. 2010. Mutational robustness can facilitate adaptation. *Nature,* 463**,** 353-355.
DRAKE, J., CHARLESWORTH, B., CHARLESWORTH, D. & CROW, J. 1998. Rates of spontaneous mutation. *Genetics,* 148**,** 1667-1686.
DRAKE, J. W. 1993. Rates of spontaneous mutation among RNA viruses. *Proc Nat Acad Sci,* 90**,** 4171-4175.
DRAKE, J. W. 2007. Too many mutants with multiple mutations. *Crit Rev Biochem Mol Biol,* 42**,** 247-258.
DRUMMUND, A. J., ASHTON, B., BUXTON, S., CHEUNG, M., COOPER, A., DURAN, C., FIELD, M., HELED, J., KEARSE, M., MARKOWITZ, S., MOIR, R., STONES-HAVAS, S., STURROCK, S., THIERER, T. & WILSON, A. 2011. Geneious v5.4. *http://www.geneious.com/*.
DUFFY, S., SHACKELTON, L. & HOLMES, E. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Gen* 9**,** 267-276.
EDGAR, R. 2004. MUSCLE: multiple sequence alignment with accuracy and high throughout. *Nucleic Acids Res,* 32**,** 1792-1797.
EIGEN, M. 1971. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften,* 58**,** 465-523.
EIGEN, M. 1992. *Steps towards life. A perspective on Evolution,* New York, US, Oxdord University Press.
EIGEN, M. 1993a. The origin of genetic information: Viruses as models. *Gene,* 135**,** 37-47.

EIGEN, M. 1993b. Viral quasispecies. *Sci Am,* 269**,** 42-49.
EIGEN, M. 2002. Error catastophe and antiviral strategy. *Proc Nat Acad Sci,* 99**,** 13374-13376.
EIGEN, M., MCCASKILL, J. & SCHUSTER, P. 1988. Molecular Quasi-Species. *J Phys Chem,* 92**,** 6881-6891.
EIGEN, M. & SCHUSTER, P. 1977. The Hypercycle. A principle of natural self-organization. Part A: Emergence of the Hypercycle. *Naturwissenschaften,* 64**,** 541-565.
ELENA, S. & SANJUÁN, R. 2007. Virus evolution: Insights from an experimental approach. *Annu Rev Ecol Evol Syst,* 38**,** 27-52.
ELENA, S. F. & SANJUÁN, R. 2005. Adaptative value of high mutation rates of RNA viruses: Separating causes from consequences. *J Virol,* 79**,** 11555-11558.
ELENA, S. F., SOLÉ, R. V. & SARDANYÉS, J. 2010. Simple genomes, complex interactions: epistasis in RNA virus. *Chaos,* 20**,** 026106.
FILÉE J., FORTERRE P. & J., L. 2003. The roles placed by viruses in the evolution of the their hosts: a view based on information protein phylogenies. *Res. Microbiol.,* 154**,** 237-243.
FINN, R., MISTRY, J., TATE, J., COGGILL, P., HEGER, A., POLLINGTON, J., GAVIN, O., GUNESEKARAN, P., CERIC, G., FORSLUND, K., HOLM, L., SONHAMMER, E., EDDY, S. & BATEMAN, A. 2010. The Pfam protein families database. *Nucl Acids Res,* 38**,** D211-222.
FONTANA, W. & SCHUSTER, P. 1987. A computer model of evolutionary optimization. *Biophys Chem,* 26**,** 123-147.
FORNS, X., PURCELL, R. H. & BUKH, J. 1999. Quasispecies in viral persistant and pathogenesis of Hepatisis C virus. *Trends in Microbiol,* 7**,** 402-410.
FORTERRE, P. 2006. Origin of the viruses an possible roles in major evolutionary transitions. *Virus Res,* 117**,** 5-16. .
FRANKEL, A. D. & YOUNG, J. A. 1998. HIV-1: Fifteen proteins and an RNA. *Annu Rev Biochem,* 67**,** 1-25.
FU, Y.-X. & LI, W.-H. 1993. Maximum Likelihood of population parameters. *Genetics,* 134**,** 1261-1270.
FURIÓ, V., MOYA, A. & SANJUÁN, R. 2005. The cost of replication fidelity in an RNA virus. *Proc Nat Acad Sci,* 102**,** 10233-10237.
FURIÓ, V., MOYA, A. & SANJUÁN, R. 2007. The cost of replication fidelity in human immunodeficiency virus type 1. *Proc Biol Sci,* 274**,** 225-230.
GAGO, S., ELENA, S., FLORES, R. & SANJUÁN, R. 2009. Extremely high mutation rate of a Hammerhead viroid. *Science,* 323**,** 1308.
GALETTO, R., MOUMEN, A., GIACOMONI, V., VERON, M., CHARNEAU, P. & NEGRONI, M. 2004. The structure of HIV-1 genomic RNA in the gp120 gene determines a recombination hot spot in vivo. *J Biol Chem,* 279**,** 36625-36632.
GE, L. M., ZHANG, J. T., ZHOU, X. P. & LI, H. Y. 2007. Genetic structure and population variability of tomato yellow leaf curl China virus. *J Virol,* 81**,** 5902-5907.
GILBERT, M. T. P., RAMBAUT, A., WLASIUK, G., SPIRA, T. J., PITCHENIK, A. E. & WOROBEY, M. 2007. The emergence of HIV/AIDS in the Americas and beyond. *Proc Nat Acad Sci,* 104**,** 18566-18570.
HAASE, A. T. 1999. Population biology of HIV-1 infection: viral and CD4+ T cell demographics and dynamics lymphatic tissue. *Annu Rev Immunol,* 17**,** 625-656.
HAASE, A. T., HENRY, K., ZUPANCIC, M., SEDGEWICK, G., FAUST, R. A., MELROE, H., CAVERT, W., GEBHARD, K., STASKUS, K., ZHANG, Z. Q., DAILEY, P. J.,

BALFOUR, H. H., ERICE, A. & PERELSON, A. S. 1996. Quantitation image analysis of HIV-1 infection in lymphoid tissue. *Science,* 274**,** 985-989.

HALL, T. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program of Windows 95/98/NT. *Nucl Acids Res,* 41**,** 95-98.

HAUBER, J., MALIM, M. H. & CULLEN, B. R. 1989. Mutational analysis of the conserved basic domain of human immunodeficiency virus tat protein. *J Virol,* 63**,** 1181-1187.

HEDRICK, P. W. 2000. *Genetics of populations,* Boston, US, Jones & Bartlett Publisher.

HOLMES, E. & MOYA, A. 2002. Is the quasispecies concept relevant to RNA viruses? *J Virol,* 76**,** 460-465.

HOLMES, E. C. 2003. Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol,* 11**,** 543-546.

HOLMES, E. C. 2009. The evolutionary genetics of emerging viruses. *Annu Rev Ecol Evol Syst,* 40**,** 353-372.

HUÉ, S., GIFFORD, R. J., DUNN, D., FERNHILL, E., PILLAY, D. & RESISTACE, B. O. T. U. C. G. O. H. D. 2009. Demonstration of sustained drug-resistant human immunodeficiency virus type 1 lineages circulating among treatment-naive individuals. *J Virol,* 83**,** 2645-2654.

HUGHES, A. L. & HUGHES, M. A. K. 2007. More Effective pirifying selection on RNA viruses than in DNA viruses. *Gene,* 404**,** 117-125.

JAIN, E., BAIROCH, A., DUVAUD, S., PHAN, I., REDSCHI, N., SUZEK, B., MARTIN, M., MCGARVEY, P. & GASTEIGER, E. 2009. Infrastructure for life sciences: design and implementation of the UniProt webasite. *BMC Bioinformatics,* 10**,** 136.

JASNOS, L. & KORONA, R. 2007. Epistatic buffering of fitness loss in yeast double deletion strains. *Nature Genet,* 39**,** 550-554.

JENKINS, G. M., RAMBAUT, A., PYBUS, O. G. & HOLMES, E. C. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol,* 54**,** 156-165.

JENKINS, G. M., WOROBEY, M., WOELK, C. H. & HOLMES, E. C. 2001. Evidence for non-quasispecies evolution of RNA viruses. *Mol Biol Evol,* 18**,** 987-994.

JI, J., HOFFMANN, J. S. & LOEB, L. A. 1994. Mutagenicity and pausing of HIVreverse transcriptase during HIV plus-strand DNA synthesis. *Nucl Acids Res,* 22**,** 47-52.

KACIAN, D., MILLS, D., KRAMER, F. & SPIEGELMAN, S. 1972. A replicating RNA molecule suitable for a detailed analysis of extracellular evolution and replication. *Proc Nat Acad Sci,* 69**,** 3038-3042.

KAWASHIMA, Y., PFAFFEROTT, K., FRATER, J., MATTHEWS, P., PAYNE, R., ADDO, M., GATANAGA, H., FUJIWARA, M., HACHIYA, A., KOIZUMI, H., KUSE, N., OKA, S., DUDA, A., PRENDERGAST, A., CRAWFORD, H., LESLIE, A., BRUMME, Z., BRUMME, C., ALLEN, T., BRANDER, C., KASLOW, R., TANG, J., HUNTER, E., ALLEN, S., MULEGA, J., BRACH, S., ROACH, T., JOHN, M., MALLAL, S., OGWU, A., SHAPIRO, R., PRADO, J. G., FINDLER, S., WEBER, J., PYBUS, O. G., KLENERMAN, P., NDUNG'U, T., PHILLIPS, R., HECKERMAN, D., HARRIGAN, P. R., WALKER, B. D., TAKIGUCHI, M. & GOULDER, P. 2009. Adaptation of HIV-1 to human leukocyte antigen class I. *Nature,* 458**,** 641-645.

KEELE, B. F., TAZI, L., GARTNER, S., LIU, Y., BURGON, T. B., ESTES, J. D., THACKER, T. C., CRANDALL, K. A., MCARTHUR, J. C. & BURTON, G. F. 2008. Characterization of the follicular dendritic cel reservoir of human inmunodeficiency virus type 1. *J Virol,* 82**,** 5548-5561.

KOONIN, E. V. & MARTIN, W. 2005. On the origin of genomes and cells within inorganic compartments. *Trends in Genet,* 21**,** 647-654.

KORBER B, K. C., FOLEY B, HAHN B, MCCUTCHAN F, MELLORS JW, AND SODROSKI J, EDS. 1998. A compilation and analysis of nucleic acid and amino acid equences. *Theoretical Biology and Biophysics Group.* Los Alamos National Laboratory, Los Alamos, NM.

KOSAKOVSKY-POND, S. L. & FROST, S. D. W. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol,* 22**,** 1208-1222.

KOSAKOVSKY-POND, S. L., FROST, S. D. W., GROSSMAN, Z., GRAVENOR, M. B., RICHMAN, D. D. & BROWN, A. J. L. 2006. Adaptation to different human population by HIV-1 revealed by codon-based analyses. *PLoS Comput BIol,* 2**,** e62.

KOSAKOVSKY-POND, S. L., FROST, S. D. W. & MUSE, S. V. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics,* 21**,** 676-679.

KOUYOS, R. D., ALTHAUS, C. L. & BONHOEFFER, S. 2006. Stochastic or deterministic: what is the effective population size of HIV-1? *Trends in Microbiol,* 14**,** 507-511.

LA-SCOLA, B., AUDIC, S., ROBERT, C., JUNGANG, L., LAMBALLERIE, X. D., DRANCOURT, M., BIRTLES, R., CLAVERIE, J. M. & RAOULT, D. 2003. A giant virus in amoebae. *Science,* 299**,** 2033.

LAMEI, C., PERLINA, A. & LEE, S. J. 2004. Positive Selection Detection in 40,000 Human Immunodeficiency Virus (HIV) Type 1 Sequences Automatically Identifies Drug Resistance and Positive Fitness Mutations in HIV Protease and Reverse Transcriptase. *J Virol,* 78**,** 3722-3732.

LAURING, A. S. & ANDINO, R. 2010. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog,* 6**,** e1001005.

LE, S. Y., CHEN, J. H., BRAUN, M. J., GONDA, M. A. & MAIZE, J. V. 1988. Stability of RNA stem-loop structure and distribution of non-random structure in the human immunodeficiency virus (HIV-I). *Nucl Acids Res,* 16**,** 5153-5168.

LE, S. Y., CHEN, J. H., CHATTERJEE, D. & MAIZE, J. V. 1989. Sequence divergence and open regions of RNA secondary structures in the envelope regions of the 17 human immunodeficiency virus solates. *Nucl Acids Res,* 17**,** 3275-3288.

LEANDRO, M., MENÉNDEZ-ARIAS, L., DE JORGE, R., VILLAHERMOSA, M. L., CONTRERAS, G., PÉREZ-AVILA, L., MOYA, A. & NÁJERA, R. 1999. Sequence analysis of the polymerase domain of HIV-1 reverse transcriptase in naive and zidovudine-treated individuals reveals a higher polymorphism in alpha-helices as compared with beta-strands. *Virus Genes,* 18**,** 203-210.

LEVY, D. N., ALDROVANDI, G. M., KUTSCH, O. & SHAW, G. M. 2004. Dynamics of HIV-1 recombination in its natural target cells. *Proc Nat Acad Sci,* 101**,** 4204-4209.

LI, H. 2011. A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Mol Biol Evol,* 28**,** 365-375.

LINDAHL, T. & NYBERG, B. 1974. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochem,* 13**,** 3405-3410.

LIU, Y., MACNEVIN, J. P., HOLTE, S., MCELRATH, M. J. & MULLINS, J. I. 2011. Dynamics of viral evolution and CTL responses in HIV-1 infection. *PLoS ONE,* 6**,** e15639.

LIU, Y. & MITTLER, E. 2008. Selection dramatically reduces effective population size in HIV-1 infection. *BMC Evol Biol,* 8**,** 133.

LOZADA-CHÁVEZ, I. 2004. *El papel de las secuencias simples en los proteomas virales.* Bióloga, Universidad Nacional Autónoma de México.

MAHY, B. W. L. & VAN REGENMORTEL, M. H. V. 2010. *Desk Encyclopedia of General Virology,* Oxford, UK, Academic Press Elsevier.

MALIM, M. H., HAUBER, J., LE, S. Y., MAIZE, J. V. & CULLEN, B. R. 1989. The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature,* 338**,** 254-257.

MANSKY, L. 1996a. Forward mutation rate of human inmunodeficiency virus type 1 in a T lymphoid cell line. *AIDS Res Hum Ret,* 12**,** 307-314.

MANSKY, L. M. 1996b. The mutation rate of human inmunodeficiency virus type 1 is influenced by the vpr gene. *Virology,* 222**,** 391-400.

MANSKY, L. M., PREVERAL, S., SELIG, L., BENAROUS, R. & BENICHOU, S. 2000. The interaction of Vpr with uracil DNA glycosylase modulates the human inmunodeficiency virus type 1 in vivo mutation rate. *J Virol,* 7415.

MANSKY, L. M. & TEMIN, H. D. 1995. Lower in vivo mutation rate of human inmunideficiency virus type 1 that predicted from the fidelity of purified reverse transcriptase. *J Virol,* 69**,** 5087-5094.

MARTELL, M., ESTEBAN, J., QUER, J., GENESCÀ, J., WEINER, A., ESTEBAN, R., GUARDIAN, J. & GÓMEZ, J. 1992. Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution *J Virol,* 66**,** 3225-3229.

MARTINEZ, F., SARDANYÉS, J., ELENA, S. F. & DARÒS, J. A. 2011. Dynamics of a plant RNA virus intracelullar accumulation: stamping machine vs. geometric replication. *Genetics,* 188**,** 637-646.

MCNATT, M. W., ZANG, T., HATZIIOANNOU, T., FOFANA, I. B., JOHNSON, W. E., NEIL, S. J. & BIENIASZ, P. D. 2009. Species-specific activity of HIV-1 Vpu and positive selection of tetherin transmembrane domain variants. *PLoS Pathog,* 5**,** e1000300.

MENÉNDEZ-ARIAS, L. 2009. Mutation rates and intrinsic fidelity of retroviral reverse transcriptase. *Viruses,* 1**,** 1137-1165.

MICHALAKIS, Y. & ROZE, D. 2004. Epistasis in RNA viruses. *Science,* 306**,** 1492.

MMWR 2001. Firs report of AIDS. *Morbidity and Mortality Weekly Report,* 50**,** 429-456.

MOREIRA, D. & LÓPEZ-GARCÍA, P. 2009. Ten reasons to exclude viruses from the tree of life. *Nat Rev Microbiol.,* 7**,** 306-311.

MOUMEN, A., POLOMACK, L., ROQUES, R., BUC, H. & NEGRONI, M. 2001. The HIV-1 repeated sequence R as a robust hot-spot for copy-choice recombination. *Nucl Acids Res,* 29**,** 3814-3821.

MOYA, A., ELENA, S. F., BRACHO, A. & MIRALLES, R. B. 2000. The evolution of RNA viruses: A population genetics view. *Proc Nat Acad Sci,* 97**,** 6967-6973.

NELSON, P. W., GILCHRIST, M. A., COOMBS, D., HYMAN, J. M. & PERELSON, A. S. 2004. An age-structured model on hiv infection that allows for variations in the production rate of viral particles and the death rate of productively infected cells. *Math Biosci Eng,* 1**,** 267-288.

NOWAK, M. 2002. From Quasispecies to unviersal grammar. *Z Phys Chem,* 216**,** 5-20.

O'NEIL, P. K., SUN, G., YU, H., RON, Y., DOUGHERTY, J. P. & PRESTON, B. D. 2002. Mutational analysis of HIV-1 Long Terminal Repeats to explore the relative contribution of reverse transcriptase and RNA polymerase II to viral mutagenesis. *J Biol Chem,* 277**,** 38053-38061.

PANDEY, V. N., KAUSHIK, N., SARAFIANOS, S. G., YADAV, P. N. & MODAK, M. J. 1996. Roles of methionine 184 of human inmnodeficiency virus type-1 reverse trscriptase in the polymerase function and fidelity of DNA synthesis. *Biochem,* 35**,** 2168-2179.

PARERA, M., PEREZ-ALVAREZ, N., BONAVENTURA, C. & MARTINEZ, M. A. 2009. Epistasis among Deleterious Mutations in the HIV-1 Protease. *J Mol. Biol.,* 392**,** 243-250.

PARIENTE, N., AIRAKSINES, A. & DOMINGO, E. 2003. Mutagenesis versus inhibition in the efficiency of extinction of foot-and-mouth disease virus. *J Virol,* 77**,** 7131-7138.

PARKIN, N. T., CHAMORRO, M. & VARMUS, H. E. 1992. Human immunodeficiency virus type 1 gag-pol frameshifting is dependent on downstream mRNA secondary structure: Demonstration by expression in vivo. *J Virol,* 66**,** 5147-5151.

PARTHASARATHI, S., VALERA-ECHEVERRIA, A., RON, Y., PRESTON, B. D. & DOUGHERTY, J. P. 1995. Genetic rearrangements occurring during a single cycle of murine leukemia virus vector replication: characterization and implications. *J Virol,* 69**,** 7991-8000.

PEETERS, M. & COURGNAUD, V. 2002. Overview of Primate Lentiviruses and their evolution in non-human primates in Africa. *Theoretical Biology and Biophysics Group, Los Alamos National Laboratory,* Los Alamos, NM.

PERALES, C., AGUDO, R. & DOMINGO, E. 2009. Counteracting Quasispecies Adaptability: Extinction of a ribavirin-resistant virus mutant by an alternative mutagenic treatment. *PLoS ONE,* 4**,** e5554.

PERELSON, A. S., NEUMANN, A. U., MARKOWITZ, M., LEONARD, J. W. & HO, D. D. 1996. HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. *Science,* 271**,** 1582-1586.

PICKRELL, J. 2006. Timeline: HIV & AIDS. *New Scientist.*

POON, A. F. Y., FROST, S. D. W. & KOSAKOVSKY-POND, S. L. 2009. Detecting signatures of selection from DNA sequences using Datamonkey. *In:* POSADA, D. (ed.) *Bioinformatics for DNA sequence analysis.* New York, USA: Human Press - Springer Science+Business Media.

PRESTON, B., POIESZ, B. & LOEB, L. 1988. Fidelity of HIV-1 reverse transcriptase. *Science,* 242**,** 1168-1171.

PRICE, D. A., GOULDER, P. J., LKLENERMAN, P., SEWELL, A. K., EASTERBROOK, P. J., BANGHAM, C. R. & PHILLIPS, R. E. 1997. Positive selection of HIV-1 citotoxic T lymphocyte escape varianst during primary infection. *Proc Nat Acad Sci,* 94**,** 1890-1895.

QUER, J., HERSHEY, C. L., DOMINGO, E., HOLLAND, J. J. & NOVELLA, I. S. 2001. Contingent neutrality in competing viral populations. *J Virol,* 75**,** 7315-7320.

REHA-KRANTZ, L. J. 1998. Regulation of DNA polymerase exonucleolytic proofreading activity: studies of bacteriophage T4 'atimutator' DNA polymerase. *Genetics,* 148**,** 1551-1557.

RICHARDSON, J. S. 1981. The anatomy and taxonomy of protein structure. *Adv Protein Chem,* 34**,** 167-339.

RIDLEY, M. 2004. *Evolution,* Oxford, UK, Blackwell Publishing.

ROBERTS, J. D., BEBENEK, K. & KUNKEL, T. A. 1988. The accuracy of reverse transcriptase from HIV-1. *Science,* 242**,** 1171-1173.

RONG, L., RUSSELL, R. S., LAUGHREA, M., WAINBERG, M. A. & LIANG, C. 2003. Deletion of stem-loop 3 is compensated by second-site mutations within the Gag protein of human immunodeficiency virus type 1. *Virology,* 324**,** 221-228.

ROUZINE, I. & COFFIN, J. 1999. Linkage disequilibrium test implies a large effective population number for HIV in vivo. *Proc Natl Acad Sci,* 96**,** 10758-10766.

ROZAS, J. 2009. DNA sequence polymorphism analysis using DnaSP. *In:* POSADA, D. (ed.) *Bioinformatics for DNA sequence analysis.* New York, USA: Human Press - Springer Science+Business Media.

ROZERA, G., ABBATE, I., BRUSELLES, A., VLASSI, C., D'OFFIZI, G., NARCISO, P., CHIMELI, G., PROSPERI, M., IPPOLITO, G. & CAPOBIANCHI, M. R. 2009. Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. *Retrovirol 6,* 15.

RUBEN, S., PERKINS, A., PURCELL, R. H., JUONG, K., SIA, R., BURGHOFF, R., HASELTINE, W. A. & ROSEN, C. A. 1989. Structural and functional characterization of human immunodeficiency virus tat protein. *J Virol,* 63**,** 1-8.

RYLAND, E. G., TANG, Y., CHRISTIE, C. D. & FEENEY, M. E. 2010. Sequence evolution of HIV-1 following mother-to-child transmission. . *J Virol,* 84**,** 12437-12444.

SALLIE, R. 2005. Replicative Homeostasis: A fundamental mechanism mediating selective viral replication and escape mutation. *Virol. J.,* 2**,** 10.

SANJUÁN, R. 2010. Quasispecies. *In:* MAHY, B. W. L. & VAN REGENMORTEL, M. H. V. (eds.) *Desk Encyclopedia of General Virology.* Oxford, UK: Academic Press Elsevier.

SANJUÁN, R. & BORDERÍA, A. V. 2011. Interplay between RNA structure and protein evolution in HIV-1. *Mol Biol Evol,* 28**,** 1333-1338.

SANJUÁN, R. & ELENA, S. F. 2006. Epistasis correlates to genomic complexity. *Proc Nat Acad Sci,* 103**,** 14402-14405.

SANJUÁN, R., MOYA, A. & ELENA, S. F. 2004. The contribution of epistasis to the architecture of fitness in an RNA virus. *Proc Nat Acad Sci,* 101**,** 15376-15379.

SARDANYÉS, J., ELENA, S. F. & SOLÉ, R. V. 2008. Simple quasispecies models for the survival-of-the-flattest effect: The role of space. *J Theor Biol,* 250**,** 560-568.

SARDANYÉS, J. & SANJUÁN, R. 2011. Quasispecies spatial models for RNA viruses with different replication modes and infection strategies. *PLoS ONE,* 6**,** e24884.

SCHUSTER, P. 2003. Molecular insights into evolution of phenotypes. *In:* CRUTCHFIELD, J. & SCHUSTER, P. (eds.) *Evolutionary dynamics. Exploring the interplay of selection, accident, neutrality, and function.* New York, USA: Oxford University Press.

SCHUSTER, P. & STADLER, P. F. 2008. Early replicons: Origin and evolution. *In:* DOMINGO, E., PARISH, C. R. & HOLLAND, J. J. (eds.) *Origin and evolution of viruses.* 2nd ed. London, UK: Academic Press Elsevier.

SCHUSTER, P. & SWENTINA, J. 1988. Stationary mutant distribution and evolutionary optimization. *Bull Math Biol,* 50**,** 636-660.

SHACKELTON, L. A., PARRISH, C. R., TRUYEN, U. & HOLMES, E. C. 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proc Natl Acad Sci,* 102**,** 379-384.

SIGRIST, C., CERUTTI, L., CASTRO, E. D., LANGENDIJK-GENEVAUX, P., BULLIARD, V., BAIROCH, A. & HULO, N. 2010. PROSITE, a protein domain database for functional characterization and annotation. *Nucl Acids Res,* 38**,** 161-166.

SMITH, R. A., ANDERSON, D. J. & PRESTON, B. D. 2004. Purifying selection masks the mutational flexibility of HIV-1 reverse transcriptase. *J Biol Chem,* 279**,** 26726-26734.

SMITH, R. A., LOEB, L. A. & PRESTON, B. D. 2005. Lethal mutagenesis of HIV. *Virus Res,* 107**,** 215-228.

SNOECK, J., FELLAY, J., BARTHA, I., DOUEK, D. & AMALIO, T. 2011. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. *Retrovirol,* 8**,** 87.

SOARES, A. E., SOARES, M. A. & SCHRAGO, C. G. 2008. Positive selection on HIV accessory proteins and the analysis of molecular adaptation after interspecies transmission. *J Mol Evol,* 66**,** 598-604.

SPIEGELMAN, S., HARUNA, I., BEAUDREAU, G. & MILLS, D. 1965. The synthesis of a self-propagating and infectious nucleic acid with a purified enzyme. *Proc Nat Acad Sci,* 54**,** 919-927.

STEINHAUER, D. A., DOMINGO, E. & HOLLAND, J. 1992. Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. *Gene,* 122**,** 281-288.

SUZAN-MONTI, M., SCOLA, B. L. & RAOULT, D. 2006. Genomic and evolutionary aspects of Mimivirus. *Virus Res,* 117**,** 145-155.

TAMURA, K. & NEI, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitocondrial-DNA in humans and chimpanzees. *Mol Biol Evol,* 10**,** 512-526.

THOMPSON, J., HIGGINS, D. & GIBSON, T. 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix code. *Nucleic Acids Res,* 22**,** 4673-4680.

TIPPIN, B., PHAM, P. & GOODMAN, M. F. 2004. Error-prone replication for better or worse. *Trends Microbiol,* 12**,** 288-295.

TORRE, J. C. D. L. & HOLLAND, J. J. 1997. RNA virus quasispecies populations can suppress vastly superior mutant progeny. *J Virol,* 64**,** 6278-281.

VAN DUIN, J. & TSAREVA, N. 2006. Single-Stranded RNA Phages. *In:* CALENDAR, R. L. (ed.) *The Bacteriophages* Second ed.: Oxford University Press.

VIGNUZZI, M., STONE, J. K., ARNOLD, J. J., CAMERON, C. E. & ANDINO, R. 2006. Quasispecies diversity determines pathogenesis trough cooperative interactions in a viral population. *Nature,* 439.

VILLARREAL, L. P. & DEFILIPPIS, V. R. 2000. A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. *J Virol,* 74**,** 7079-7084.

WAGNER, E. K., MARTINEZ, J. W., BLOOM, D. C. & CAMERINI, D. 2008. *Basic Virology,* Garsington Oxford, UK, Blackwell Publishing.

WANG, K., MITTLER, E. & SAMUDRALA, R. 2006. Comment on "Evidence for Positive Epistasis in HIV-1". *Science,* 312**,** 848.

WATTS, J. M., KRISTEN, K. D., GORELICK, R. J., LEONARD, C. W., JR-BESS, J. W., SWANSTROM, R., BURCH, C. L. & WEEKS, K. M. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature,* 460**,** 711-716.

WEI, X., DECKER, J. M., WANG, S., HUI, H., KAPPES, J. C., WU, X., SALAZAR-GONZALES, J. F., SALAZAR, M. G., KILBY, J. M., SAAG, M., KOMAROVA, N. L., NOWAK, M. A., HANN, B. H., KNOW, P. D. & SHAW, G. M. 2003. Antibody neutralization and escape by HIV-1. *Nature,* 422**,** 307-312.

WEIZHOUNG, L. & GODZIK, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 22**,** 1658-1659.

WERTHEIM, J. O. & WOROBEY, M. 2009. Relaxed selection and the evolution of RNA virus mucin-like pathogenicity factors. *J Virol,* 83**,** 4690-4694.
WESSNER, D. R. 2010. The origin of Viruses. *Nature Education,* 3**,** 37.
WILKE, C. O. 2005. Quasispecies theory in the context of populations genetics. *BMC Evol Biol,* 5**,** 44.
WILKE, C. O., WANG, J. L., OFRIA, C., LENSKI, R. E. & ADAMI, C. 2001. Evolution of ditgial organism at high mutation rates leads to survival of the flattest. *Nature,* 412**,** 331-333.
WOO, J., ROBERTSON, D. L. & LOVELL, S. C. 2010. Constraints on HIV-1 Diversity from Protein Structure. *J Virol,* 84**,** 12995-13003.
WOOD, N., BHATTACHARYA, T., KEELE, B. F., GIORGI, E., LIU, M., GASCHEN, B., DANIELS, M., FERRARI, G., HAYNES, B. F., MCMICHAEL, A., SHAW, G. M., B.H., H., B., K. & SEOIGHE, C. 2009. HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. *PLoS Pathog,* 5**,** e1000414.
WRIGTH, S. 1969. *Evolution and the genetics of populations,* Chicago, US, University of Chicago Press.
YANG, Z. 2006. *Computational Molecular Evolution,* New York, USA, Oxford University Press Inc.
YOSHIDA, I., SUGIURA, W., SHIBATA, J., REN, F., YANG, Z. & TANAKA, H. 2011a. Change of positive selection pressure on HIV-1 envelope gene inferred by early and recent samples. *PLoS ONE,* 6**,** e18630.
YOSHIDA, I., SUGIURA, W., SHIBATA, J., REN, F., ZIHENG, Y. & TANAKA, H. 2011b. Change of positive selection pressure on HIV-1 envelope gene inferred by early and recent samples. *PLoS ONE,* 6**,** e18630.
YOSHIDA, K., NAKAMURA, M. & OHNO, T. 1997. Mutations of the HIV type 1 V3 loop under selection pressure with neutralizingmonoclonal antibody NM-01. *AIDS Res Hum Ret,* 13**,** 1283-1290.
YU, Q., KONIG, R., PILLAI, S., CHILES, K., KEARNEY, M., S., P., RICHMAN, D. D., COFFIN, J. M. & LANDAU, N. R. 2004. Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. *Nat Struct Mol Biol,* 11**,** 435-442.
YU, Q., LANDAU, N. & KÖNIG, R. 2003. Vif and the role of antiviral cytidine deaminases in HIV-1 replication. *Theoretical Biology and Biophysic Group, Los Alamos National Laboratory,* LA-UR 04-7420**,** 12.
ZANOTTO, P. M. D. A., KALLAS, E. G., SOUZA, R. F. & HOLMES, E. C. 1999. Genealogical Evidence for Positive Selection in the nef Gene of HIV-1 *Genetics,* 153**,** 1077-1089.

MATERIAL SUPLEMENTARIO

Texto S1. El modelo de cuasiespecies. Una breve explicación.

Como se había mencionado previamente, la teoría de Eigen-Schuster hace referencia a dos aspectos muy importantes de la evolución de una población: 1) el cambio a través del tiempo, con efectos adaptativos, y 2) la pérdida de la información a través de una alta tasa de error, creando como consecuencia la pérdida de la identidad (catástrofe de error). En términos simples, la teoría está fundamentada en un modelo matemático (ver ecuación S1.1), en la que se constituye de tres parámetros que modelan a un sistema de evolución básico; estos son la *replicación*, la *mutación* y la *selección*. Cada parámetro está formada de un sistema de ecuaciones pequeño que modela su proceso (Luque, 2003). La dinámica de las poblaciones surge debido a la combinación de estos parámetros en las ecuaciones diferenciales (ver ecuación S1). Matemáticamente, la ecuación es un conjunto de ecuaciones diferenciales:

$$\frac{dx_i}{dt} = (A_i Q_{ii} - D_i)x_i(t) + \sum_{j \neq i} A_j Q_{ij} x_j(t) - \Phi(t)x_i(t), \qquad i = 1, \ldots, s \qquad \text{(S1.1)}$$

El primer término de la ecuación (en azul), representa el incremento en la población como resultado de la producción de replicas correctas. El segundo (en rojo) representa la aparición de nuevos copias mutantes como resultado de la replicación errónea llevada a cabo por otras secuencias, mientras que el tercero (en verde) está asociado a los parámetros de la población inicial y su cambio a través del tiempo. Donde *i* es el número de secuencias en la población de forma constante. Debido a que cada secuencia tiene un valor de productividad (reproducción), lo cual se traduce como el fitness en este modelo, la dinámica final se representa como una condición de adaptabilidad de las secuencias a altas tasas de error que afectan la productividad de la población, y por ende, la selección de la misma. La construcción detallada y representación matemática del modelo de Eigen-Schuster puede ser consultada en la revisión Realizada por Luque (2003).

Texto S2. *Quasispecies.pl:* tiempo de cálculo y rendimiento relativo computacional

Se calculó el *tiempo de cálculo* y el *rendimiento relativo computacional* del programa *quasispecies.pl* basados en tres estimaciones sencillas: 1) tiempo de CPU en segundos, 2) tiempo de CPU por ciclos/segundo y 3) el rendimiento relativo computacional. Para ello se comparó a una *PC portátil* con 2 procesadores de doble núcleo cada uno (4 núcleos en total) a 2.0 GHz y memoria RAM de 16Gb *versus* un *Cluster* de 16 procesadores con 4 núcleos cada uno (64 núcleos en total) a 2.4 GHz y con memoria RAM de 45Gb. Ambos ordenadores con sistema operativo Linux con arquitectura de 64 bits.

El *tiempo de cálculo* (CPU time, tiempo de la *Unidad Central de Procesamiento*) indica cuánto tiempo tarda en realizarse/ejecutarse un trabajo en el ordenador con un programa determinado. Se emplearon dos estimaciones individuales: a) una para medir el tiempo de CPU 'segundos' y otro para b) 'ciclos por segundo'. Así, con el número actual de individuos (N_e=1% en HIV-1), el 'tiempo de CPU en segundos' promedio en la PC portátil fue de 239983 segundos (114%), mientras que en el cluster fue de 150707 segundos (99%). El porcentaje indica el tiempo de transcurrido que se gasto esperando o ejecutando otros programas en el servidor (conexiones de procesadores, acceso a la memoria, dispositivos del disco duro y operaciones de entrada/salida, entre otros): por lo que el programa ocupa el 114% y 99% de tiempo de CPU para cumplir el programa, respectivamente. El tiempo de cálculo fue realizado con la siguiente ecuación:

$$CPUt_{total} = \frac{CPU_{user} + CPU_{sys}}{t_{real}} \qquad \text{(S2.1)}$$

Donde, $\boldsymbol{CPUt_{total}}$=porcentaje de tiempo transcurrido (tiempo de CPU) que se ocupa para realizar todos los procesos en el ordenador al ejecutar el *quasispecies.pl*, $\boldsymbol{CPU_{user}}$=tiempo de usuario, $\boldsymbol{CPU_{sys}}$=tiempo del ordenador, $\boldsymbol{t_{real}}$=tiempo transcurrido que tarda el programa. Los valores de los parámetros de la ecuación fueron obtenidos con el comando "*time*" de UNIX.

Asimismo, el cálculo del 'tiempo de CPU por ciclos/segundos' fue basado en las métricas de tiempo del hardware de los ordenadores (tasa de tiempo de reloj y tiempo real de cómputo). El tiempo de CPU representado como el número de ciclos que realiza el procesador para terminar

el programa (***cyc*$_{relojCPU}$**), refleja los mismos resultados que en el tiempo de CPU por segundos, en la que el cluster fue mejor ($3.8x10^{14}$ ciclos) que la PC portátil ($4.8x10^{14}$ ciclos), es decir, el clúster realiza mas ciclos que la PC portátil en menos tiempo. Esto se realizó con la siguiente fórmula:

$$Cyc_{relojCPU} = t_{real} * freq_{reloj} \qquad \text{(S2.2)}$$

Donde, ***Cyc*$_{relojCPU}$** es el rendimiento absoluto, representado como la frecuencia del reloj, *i.e.* el numero de ciclos (operaciones) por hertz (Hz) realizados por el procesador para el programa *quasispecies.pl*; ***freq*$_{reloj}$** es el número de ciclos del reloj del CPU que realiza el procesador del ordenador en Hz y ***t*$_{real}$** es el tiempo transcurrido real que tarda el programa *quasispecies.pl* en terminar.

Finalmente, el *rendimiento relativo* expresa el nivel de procesamiento (número de tareas completadas) de un programa dado en base a las características de un ordenador, comparado con otro de diferente nivel computacional. Así, el rendimiento relativo promedio fue 1.8 veces mejor en el cluster que en la PC portátil, esto significa que el número de tareas completadas en A es 1.8 el número de tareas que en la máquina B. Es decir, A fue más rápido que B. Esta cifra se calculó con los tiempos de cómputo para cada ordenador (ver ecuación 1S), y se dividieron como muestra la ecuación 3S.

$$CPUt_{relative} = \frac{ren_A}{ren_B} \qquad \text{(S2.3)}$$

Donde ***CPUt*$_{relative}$** es el rendimiento relativo, ***ren*$_A$** es el rendimiento del ordenador con más tiempo dado en segundos, y ***ren*$_B$** es el rendimiento del ordenador con menor tiempo dado en segundos. En perspectiva, no hubo una diferencia significativa, por lo que pueden emplearse ordenadores similares a los utilizados en este estudio para ejecutar el programa *quasispecies.pl*.

Tabla de Texto S2. **Tiempo de computo para el programa *quasispecies.pl*.** Se calculó cada característica del tiempo de CPU en distintas maquinas, A: cluster K53, y B: PC portátil. El programa se ejecuto con un Ne=200 y con el numero de generaciones T=2000. (1) Fue calculado con la ecuación S2.1. (2) Fue calculado con la ecuación S2.2. (3) Fue calculado con la ecuación S2.3

Tipo de Desempeño	**Cluster (A)**				**PC portátil (B)**				**Mejor desempeño**
	1m	**2m**	**3m**	**4m**	**1m**	**2m**	**3m**	**4m**	
CPU time (%)[1]	99	98	99	99	115	116	113	111	A
Rendimiento (ciclos)[2]	$4.0x10^{14}$	$4.8x10^{14}$	$3.2x10^{14}$	$2.3x10^{14}$	$4.5x10^{14}$	$6.0x10^{14}$	$4.5x10^{14}$	$3.9x10^{14}$	A≈B
Rendimiento relativo (segundos)[3]	166854	203274	134418	98285	227339	303279	229371	199942	A

Tabla de Texto S2. **Tiempo de computo para el programa *quasispecies.pl*.** Se calculó cada característica del tiempo de CPU en distintas maquinas, A: cluster K53, y B: PC portátil. El programa se ejecuto con un Ne=1,000 y con el numero de generaciones T=100. (1) Fue calculado con la ecuación S2.1. (2) Fue calculado con la ecuación S2.2. (3) Fue calculado con la ecuación S2.3

Tipo de Desempeño	**Cluster (A)**				**PC portátil (B)**				**Mejor desempeño**
	1m	**2m**	**3m**	**4m**	**1m**	**2m**	**3m**	**4m**	
CPU time (%)[1]	92	94	92	91	110	120	109	105	A
Rendimiento (ciclos)[2]	$4.0x10^{14}$	$5.1x10^{14}$	$3.7x10^{14}$	$3.2x10^{14}$	$6.9x10^{14}$	$7.9x10^{14}$	$5.7x10^{14}$	$4.8x10^{14}$	A
Rendimiento relativo (segundos)[3]	169178	214364	155648	135685	346198	399991	285872	244876	A

Figuras S1. Alineamientos ocupados para este estudio. Se muestran los alineamientos de las 11 regiones funcionales del genoma de HIV-1 (en orden de aparición): las regiones 5-LTR y 3-LTR, y los genes gag, gag-pol, pol, vif, vpr, tat, vpu, rev, env, nef. Inicialmente se alienaron los 2338 genomas, luego de los filtros correspondientes, se obtuvo una base de datos de 124 genomas finales, los cuales se desensamblaron en diferentes las regiones funcionales. En la ***parte superior*** de cada alineamiento, se muestra el porcentaje de identidad en forma de barras: *verde intenso* significa un alto porcentaje; mientras que el verde tenue (y colores más tenues) muestran un bajo porcentaje de identidad. En la ***parte inferior restante***, los colores sobrepuestos en cada secuencia a lo largo de todo el alineamiento, hace referencia a todas aquellas posiciones diferente (mutaciones) a la secuencia de referencia (el genoma *hxb2*). En otros alineamientos, se muestra esta misma categorización, pero debido a su longitud, las mutaciones son identificadas en color negro (i.e. gen *gag*). El número de secuencias en las regiones LTR del genoma viral es menor al número de secuencias obtenidas para las demás regiones, esto debido a la presencia de un número considerable de gaps y regiones incompletas a pesar de que se reportaron como genomas completos.

Figura S1-A. Región *5-LTR*

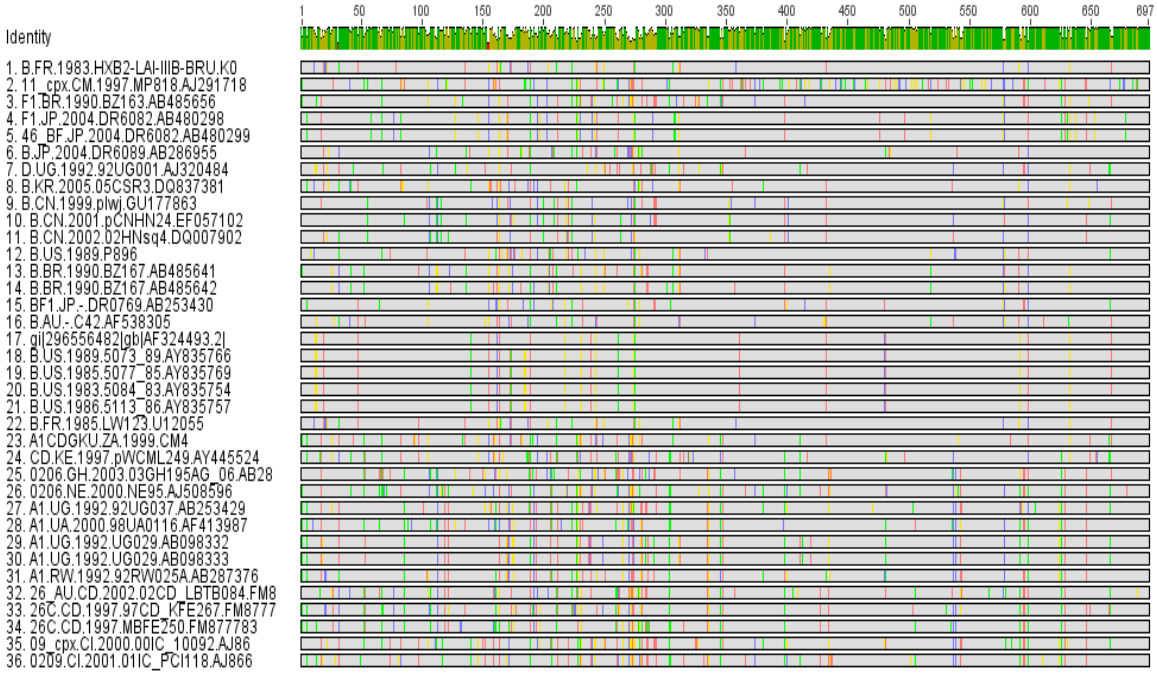


Figura S1-B. Región *3-LTR*

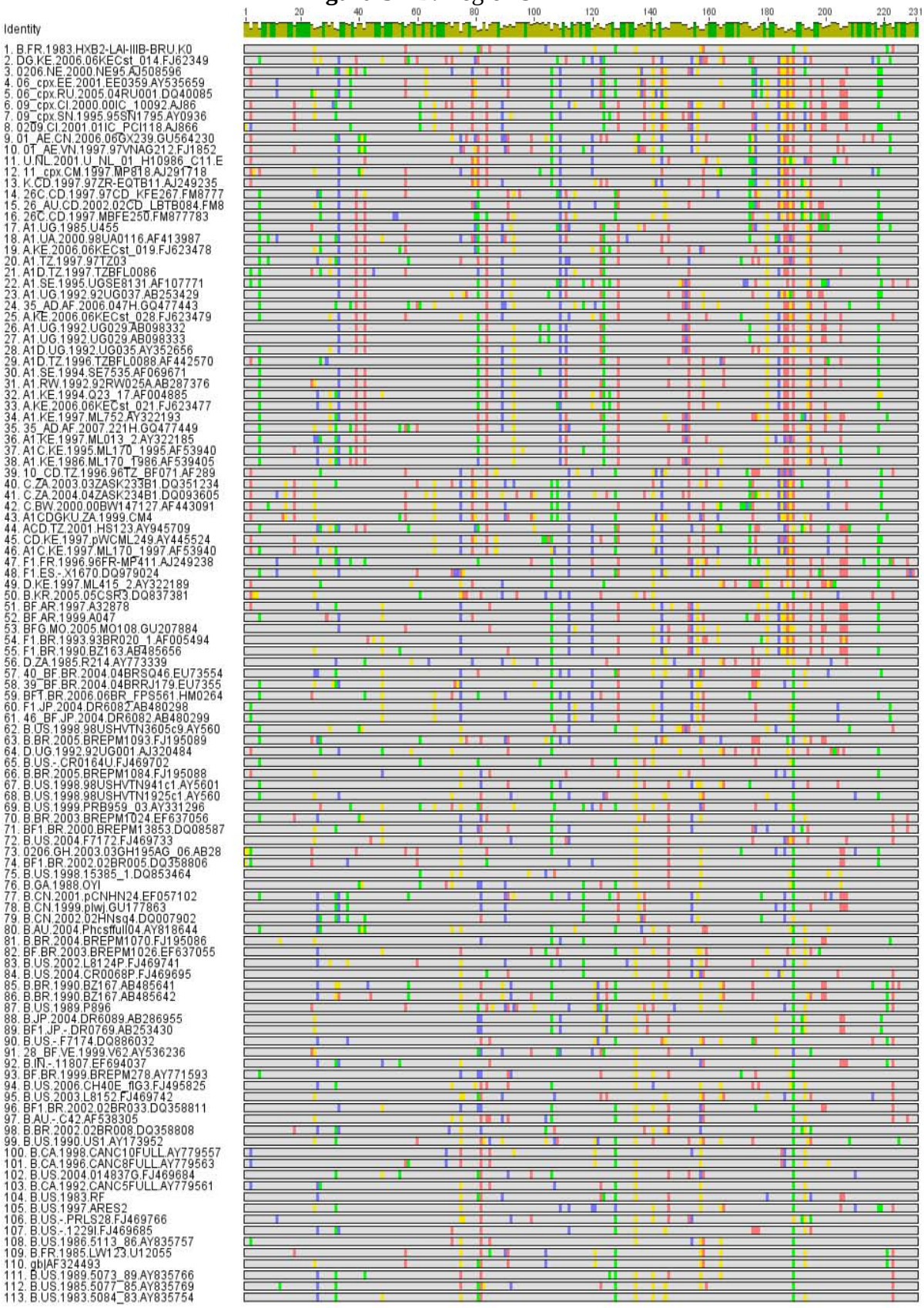


Figura S1-C. Gen *gag*

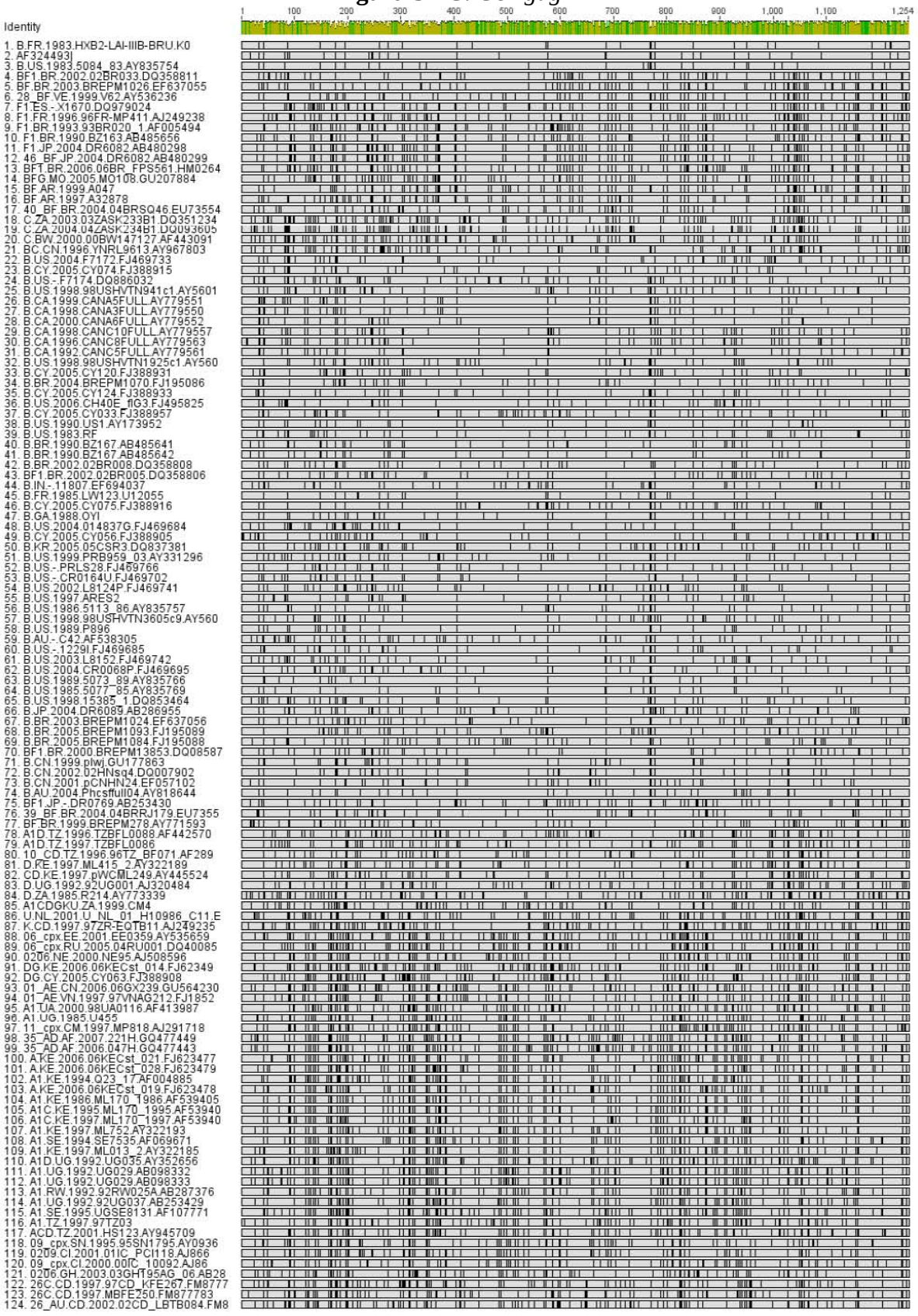


Figura S1-D. Región sobrepuesta *gag-pol*

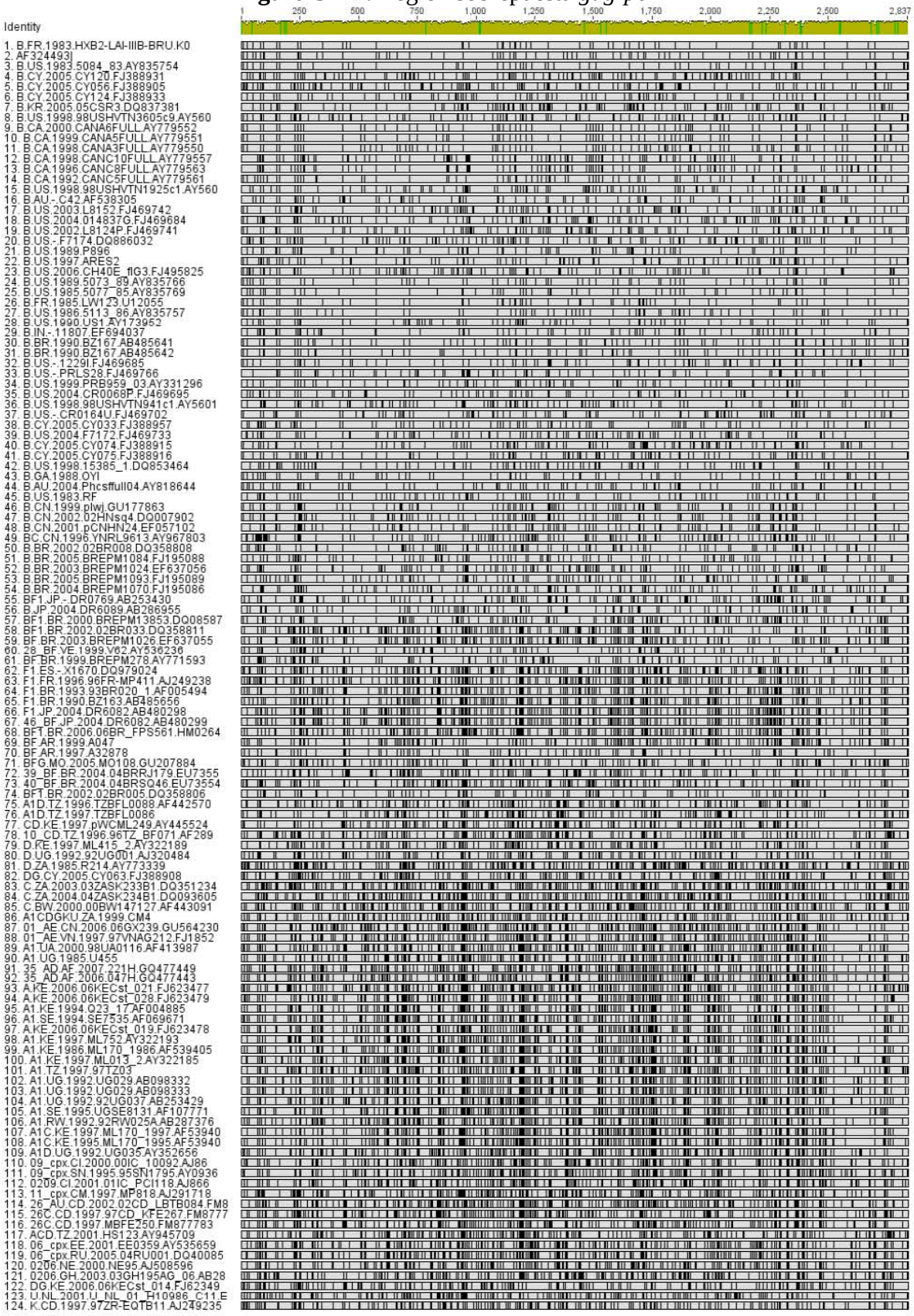


Figura S1-E. Gen *pol*

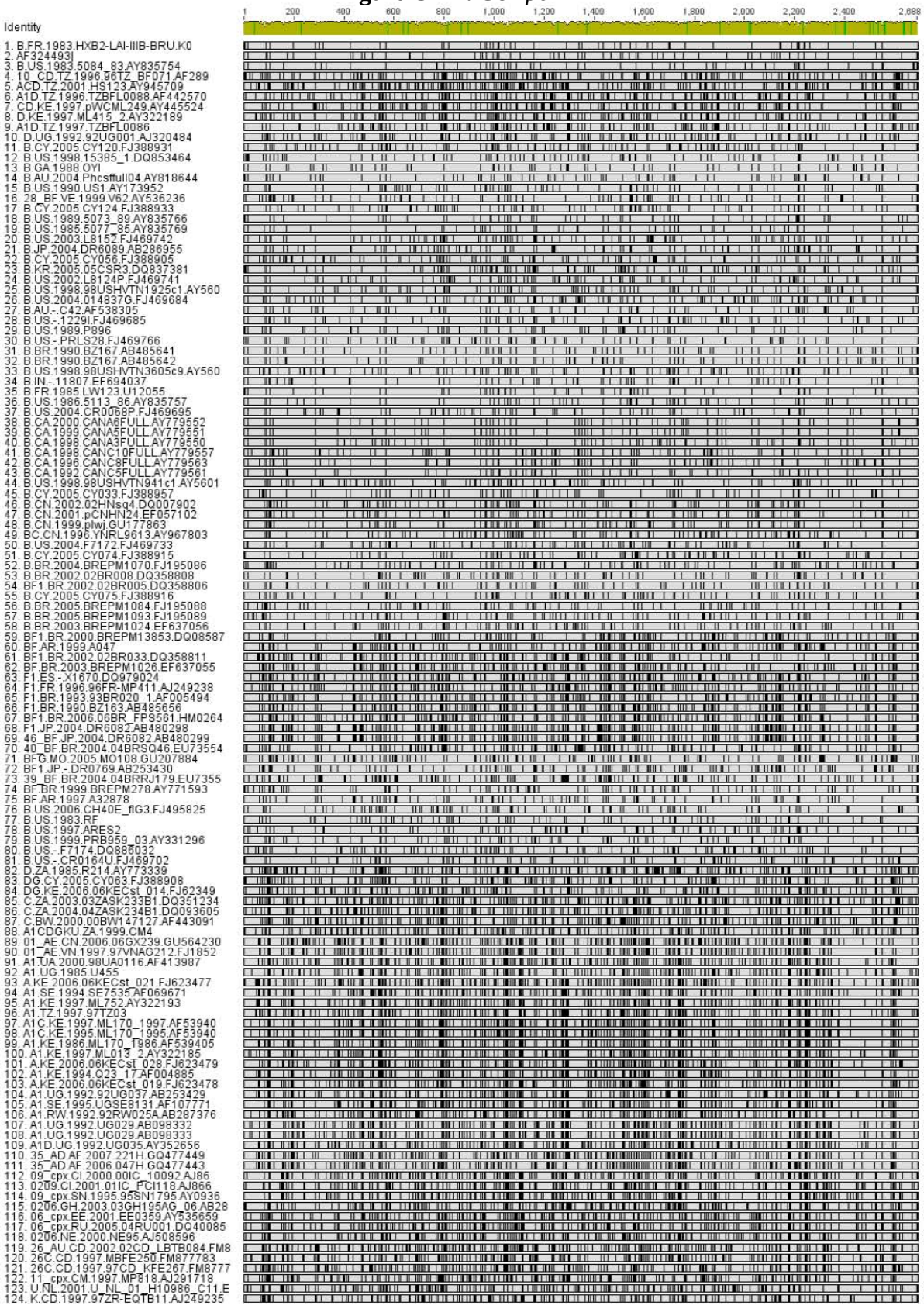


Figura S1-F. Gen *vif*

Figura S1-G. Gen *vpr*

Figure S1-H. Gen *tat*

Figura S1-I. Gen *vpu*

Figura S1-J. Gen *rev*

Figura S1-K. Gen *env*

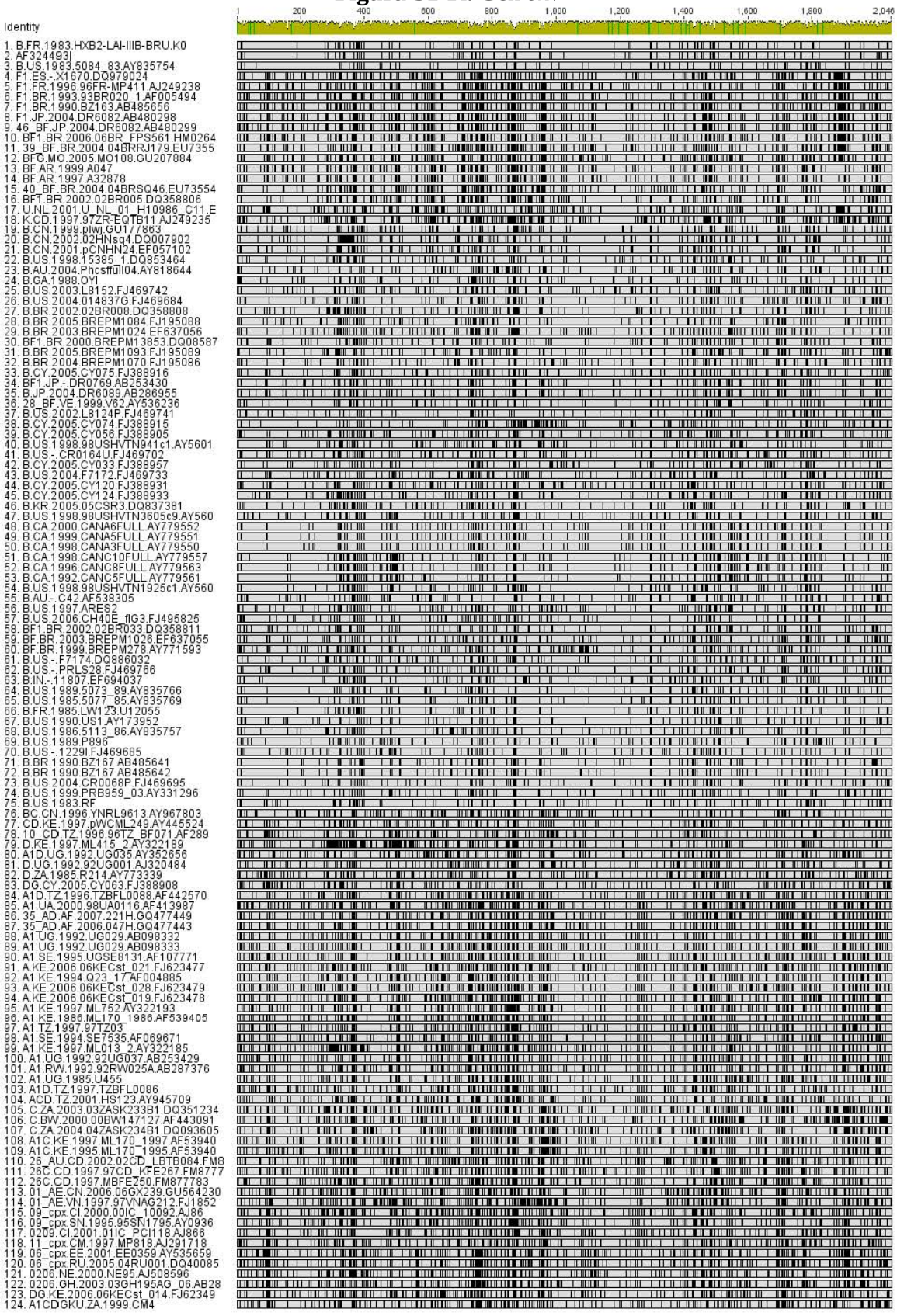
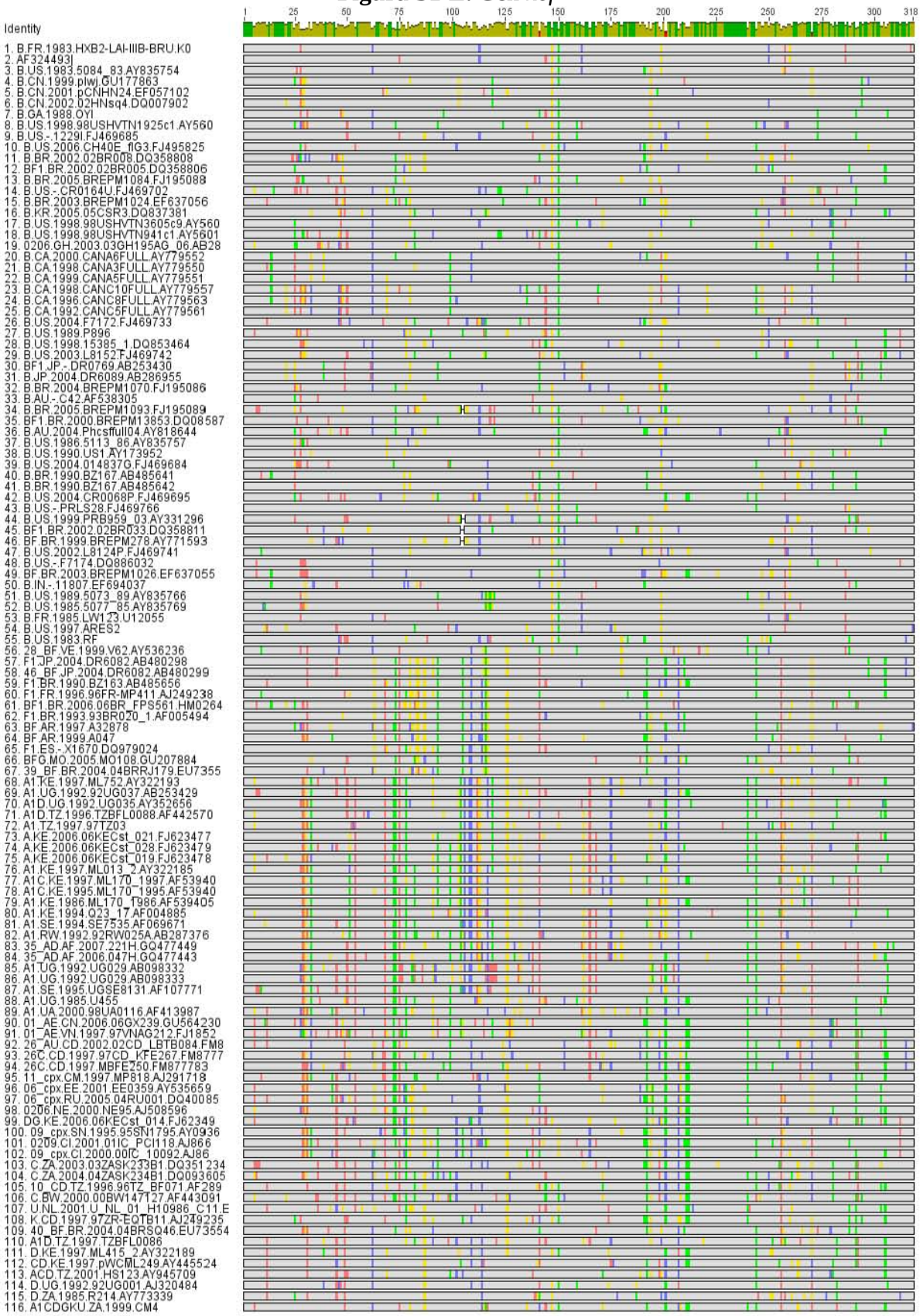
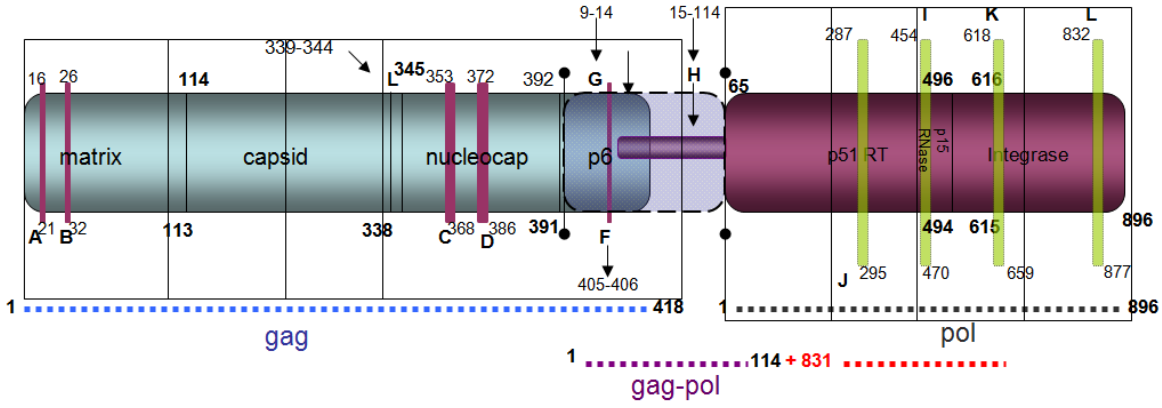


Figura S1-L. Gen *nef*



Figuras S2. Construcción de ventanas para 9 de las 11 regiones funcionales del genoma de HIV-1. (*A)* genes Gag, Gag-Pol (ovelapping region) y Pol: ventana de 4 ventanas de 315 nucleótidos, 1 de 342, y 4 de 672, respectivamente*; **(B)*** gen Vif: 5 de 111*; **(C)*** gen Vpr: 2 de 102; ***(D)*** *gen Tat*: 5 de 66; ***(E)*** *gen Rev:* 4 de 78, ***(F)*** *gen Vpu:* 3 de 57*; **(G)*** *gen Env:* 4 de 513; y ***(H)*** el *gen Nef:* 4 de 81. Cada una de estas ventanas representa un número de sub-regiones completas para poder estimar la selección. En las figuras, cada sub-región presenta su sitio de inicio (número superior izquierda) y fin (número inferior derecha) dentro de la secuencia, así como su diferenciación por letra.

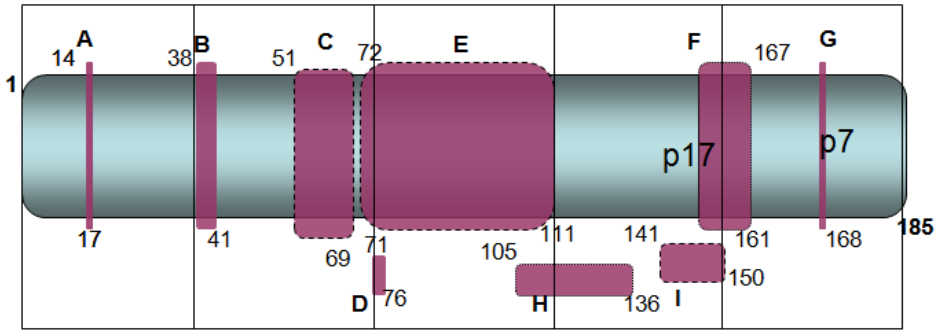
A



	Sites:	Lenght (aa)
A.	Nuclear export signal	5
B.	Nuclear localization signal	6
C.	CCHC-type 1	15
D.	CCHC-type 2	14
E.	PTAP/PSAP motif (*don´t match in alignment*)	-
F.	LYPX(n)L motif	2

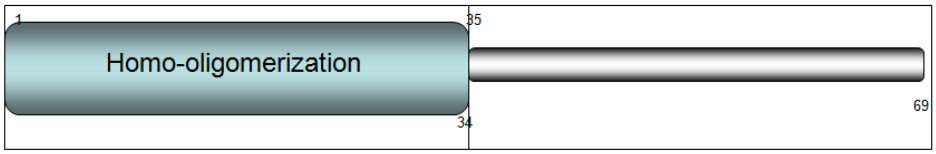
	Sites:	Lenght (aa)
G.	P6 (overlapping in Gag-Pol gen) and	5
H.	Protease (overlapping in Gag-Pol gen)	99
I.	Tryptophan repeat motif	16
J.	RT 'primer grip'	10
K.	Integrase type	41
L.	Integrase type	45

B



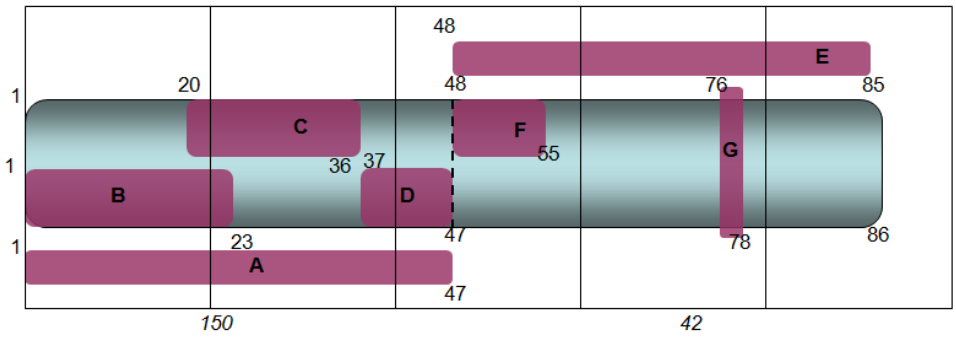
	Sites:	Length (aa)
	Total protein	185
A.	Interaction with host APOBEC3F; F1-box	4
B.	Interaction with host APOBEC3G; G.box	4
C.	Interaction with host APOBEC3F y 3G; FG-box	19
D.	Interaction with host APOBEC3F; F2-box	6
E.	RNA-binding	40
F.	Multimerization	14
G.	Membrane association	2
H.	HCCH motif	32
I.	BC-box-like motif	10

C



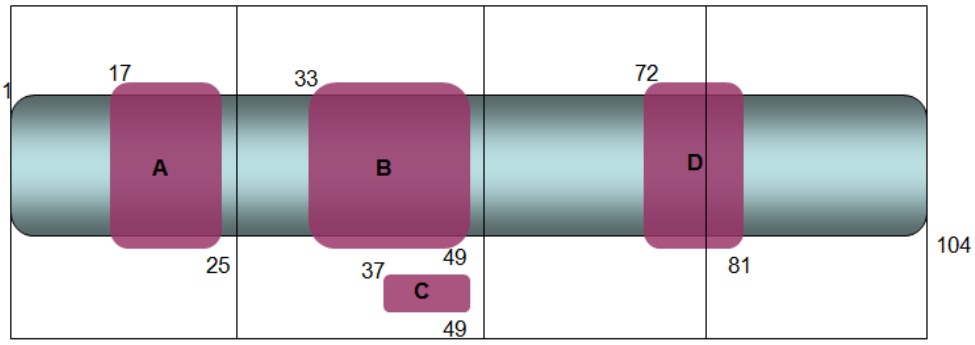
	Sites:	Lenght (aa)
	Protein total	69
A.	Homo-oligomerization	34

D



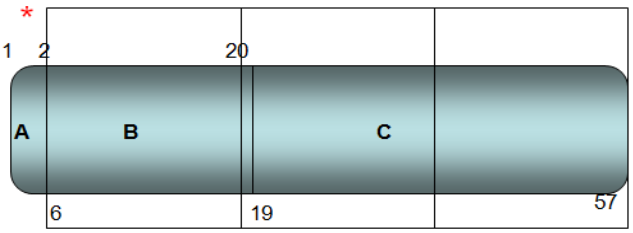
	Sites:	Length (aa)
	Protein total	85
A.	Trasactivation	48
B.	Interaction with human CREBBP	24
C.	Cysteine-rich	16
D.	Core	11
E.	Interaction with the host capping enzyme RNGTT	38
F.	Nuclear localization signal, RNA-binding (TAR), and protein transduction	9
G.	Cell attachment site	3

E



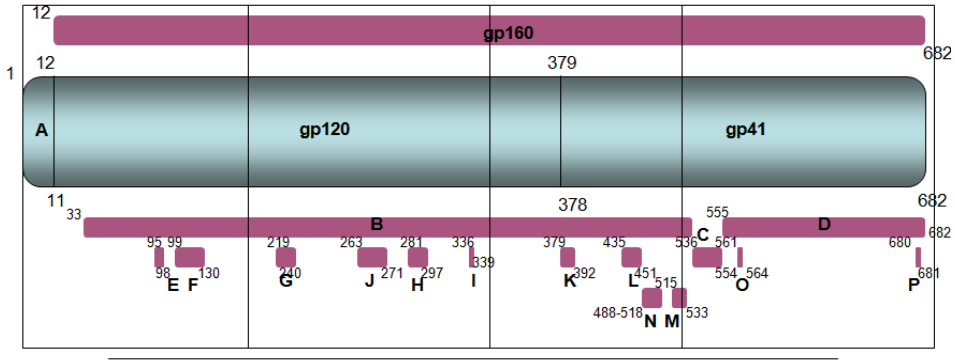
	Sites:	Length (aa)
	Protein total	104
A.	Homomultimerization	9
B.	Nuclear localization signal and RNA-binding (RRE)	17
C.	Nuclear export signal and binding to XPO1	12
D.	Poly-Arg	13

F



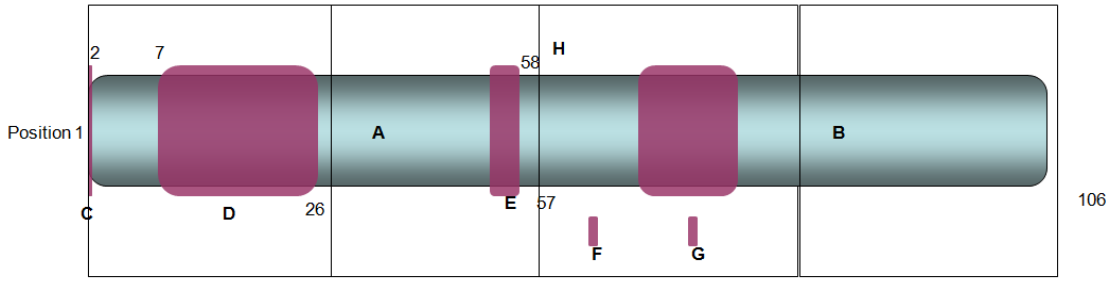
Sites:		Length (aa)
	Protein total	81
A.	Extracellular	*Not match in alignment*
B.	Helical	18
C.	Cytoplasmic	38

G



Sites:		Length (aa)
A.	Signal peptide	32
B.	Extracellular	652
C.	Helical	21
D.	Cytoplasmic	151
E. F.G. H. I.	are: V1, V2 , V3, V4 and V5	26, 40, 35, 34, 11 (respectively)
J.	CD4-binding loop	11
K.	Fusion peptide	21
L.	Inmunosuppression	17
M.	MPER, binding to GalCer	22
N.	Putative region unknown	35
O.	YXXL motif; contains endocytosis signal	4
P.	Di-leucine internalization motif	2

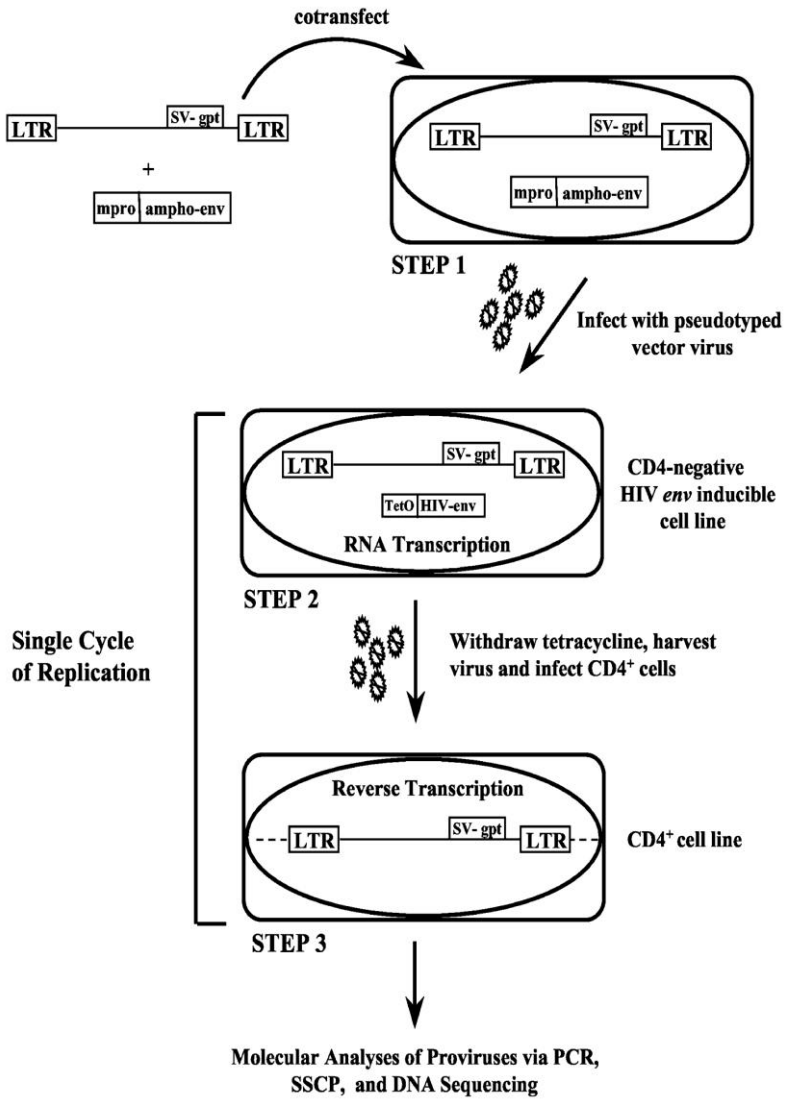
H

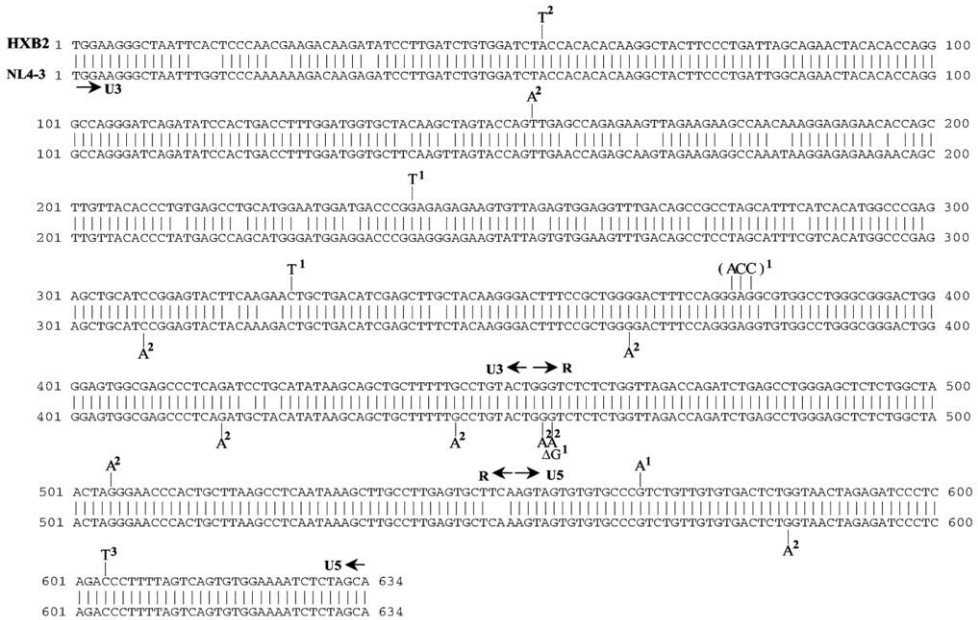


Sites:		Length (aa)
A.	N terminal; associated with host plasma membrane	56
B.	C terminal core protein	149
C.	Removed; by host	1
D.	Necessary for MHC-I internalization	20
E.	Acidic, stabilize the interaction of NEF/MHC-I whith host AP11; necessary for MHC-I internaliz. and interaction with host PACS1 and PACS2	4
F.	Poly-Glu	4

Sites:		Length (aa)
G.	PxxP; stabilize the interaction of NEF/MHC-I with host AP1M1, for internalization	4
H.	SH3-binding; interaction with Src family tyrosine kinase	10
I.	Mediates dimerization, NEF-PTE1 interaction	17
J.	Binding ATP6V11	33
K.	Di-leucine internalization motif; necessary for CD4 internalization	2
L.	Diacidic; necesary fot internalization	2

Figuras S4. Construcción de protocolos para identificar las mutaciones empleando vectores de expresión celular.

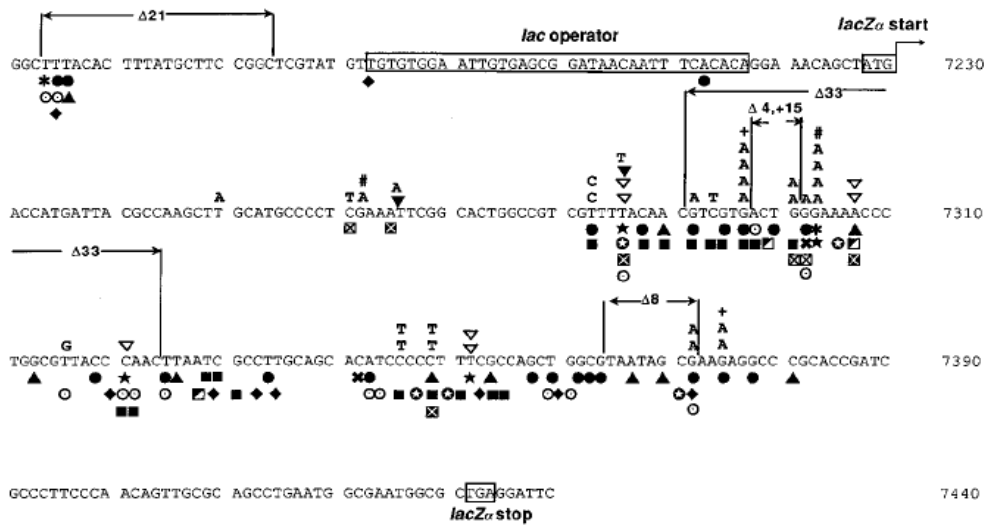


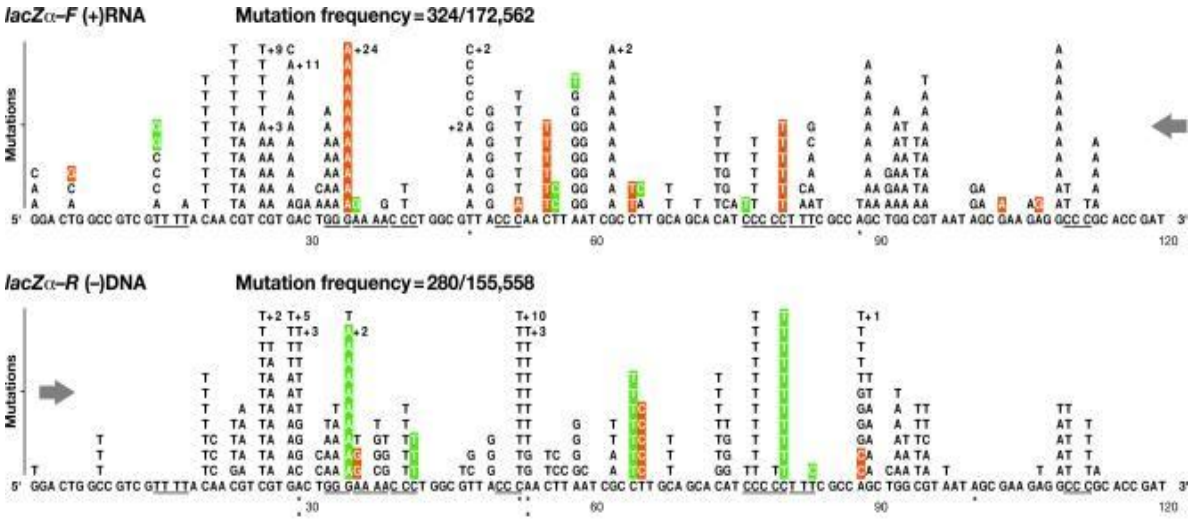


Panel superior. Alineación de secuencias LTR de las cepas HXB2 y NL4 3 de VIH-1 en las que se muestran las localizaciones de mutaciones puntuales identificados por SSCP y secuenciación del ADN. Las mutaciones de las secuencias marcadas *por encima* de las secuencias son las identificadas en el provirus VIH-gpt, mientras que los que están *debajo* se encuentra en el provirus NL4-3gpt. Todas las mutaciones mostraron que son sustituciones d base única, excepto para una deleción en el nucleótido 456 y tres sustituciones consecutivas de bases (en paréntesis) de los nucleótidos 377 al 379. Las mutaciones marcados con *superíndice 1* ocurrieron sólo en el 5'-LTR. Las mutaciones marcados con *superíndice 2* se produjeron en ambos LTRs, mientras que las mutaciones marcadas con *superíndice 3* se encuentra sólo en el 3'-LTR.

Panel izquierdo. Protocolo para estudiar la tasa de mutación en VIH-1. El vector viral fue cotransfectado con un plásmido, el cual a env anfotrópico MLV en células T293, una línea celular de embrión de riñón de humano (etapa 1 en las células). El vector del virus serotipado se utilizó para infectar una linea celular CD4-negativo del VIH-1 inducible de *env* (etapa 2). Tras la inducción, el virus se recolectado y usado para infectar a células diana HeLa T4 CD4-positivas, seguido por la selección y aislamiento de clonas de células (etapa 3). Los provirus en el paso 3 se analizaron adicionalmente. La replicación celular de las etapa 2 y 3 es confinada a un solo ciclo de replicación. mpro, el promotor MLV U3; ampho-env, el gen anfotrópico de envoltura MLV; SV-gpt, promotor del gen temprano de SV40 y las secuencias codificantes para el gen gpt; tetO, operador de tetraciclina heptamerizado fusionado a un promotor mínimo de citomegalovirus.

Figura S5. Sistema de detección de mutaciones para los diferentes tipos de sustitución de nucleótidos en HIV-1.

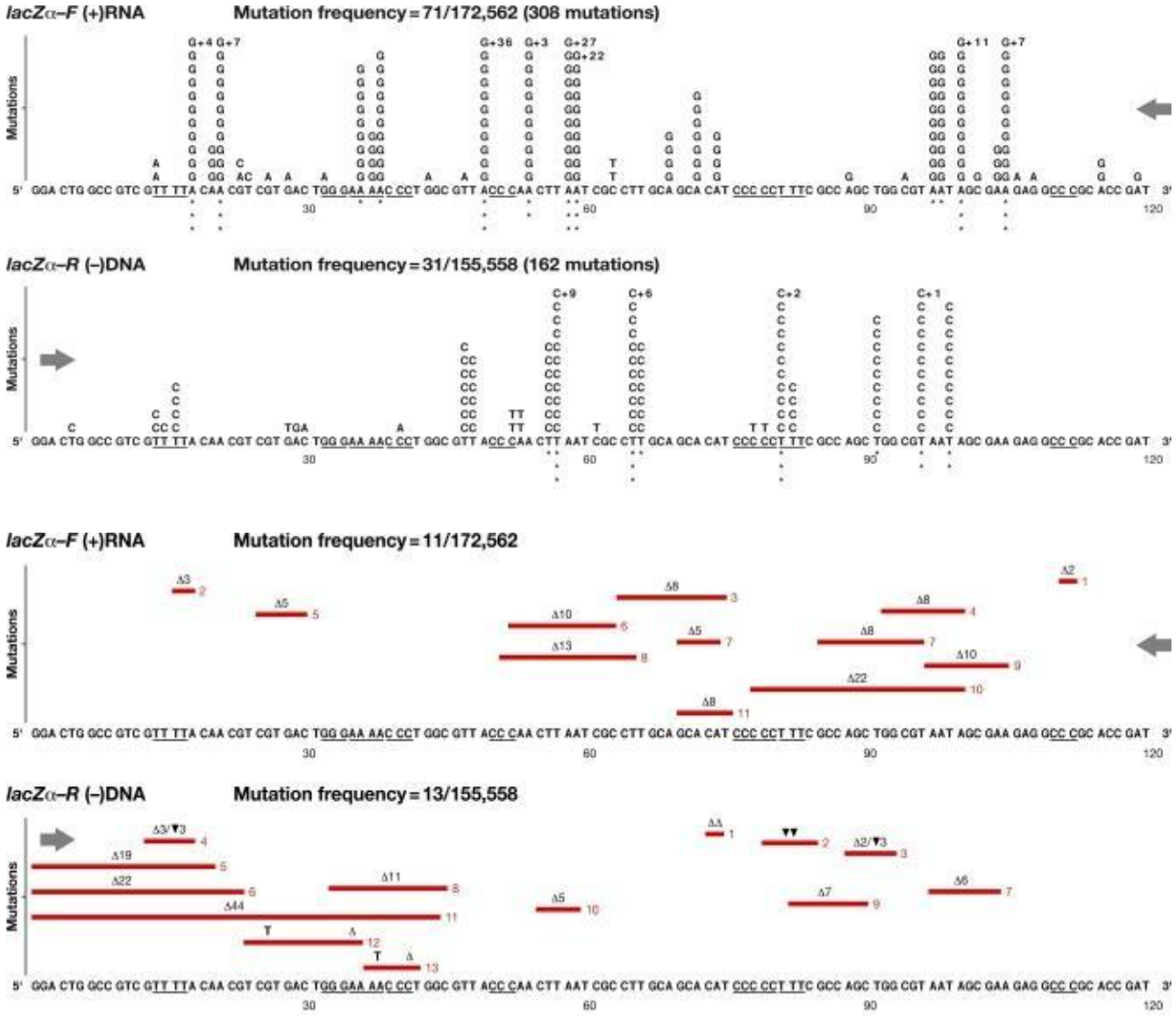




Left panel (top). Plus-strand nucleotide sequence of the lacZa gene region in HIV shuttle 3.12. The start for nucleotide numbering is the beginning of the 59 long terminal repeat. The start and stop codons of the lacZa open reading frame (small boxed sequences) and the lac operator sequence (large boxed sequence) are shown. Nucleotide positions of base pair substitutions (letters above the sequence), +1 frameshifts (letters with ∇ (black color) above the sequence), -1 frameshifts (∇ (white color) above the sequence), deletions and deletions with insertions (solid black lines with arrows above the sequence and adjacent to the deletion names) are indicated above the sequence. Base substitution mutations occurring in the lacZa gene region of the same mutant are designated with a + or #, respectively. Symbols below nucleotides in the nucleotide sequence indicate the locations of base substitution and frameshift mutations that have been characterized in cell-free studies with purified HIV-1 reversetranscriptase, using the sense-strand of the lacZa gene region as an RNA template or a DNA template. Target sites for base substitution mutations in experiments using an RNA template are indicated by ⊙ for the study by Boyer et al., and ■ and **X** (hot-spot site for base substitutions) for the study by Ji and Loeb; target sites for frameshift mutations are indicated as ♦ and 'start inside of a box' (hot-spot site for frameshift mutations) for the Boyer et al. study and i for the Ji and Loeb study. Target sites for base substitution mutations in experiments using a DNA template (6) are indicated by ● and ⋆ (hot-spot site for base substitutions), and frameshift mutations are indicated by Δ (black color) and ⋆ (hot-spot sites for frameshift mutations). Additional symbols below nucleotides in the nucleotide sequence indicate the locations of base substitution and frameshift mutations that have been characterized in the BLV shuttle vector BLV-SVNEO/ACYCLacZ (34). Target sites for base substitution and frameshift mutations are indicated below the nucleotide sequence by ☒ and ☒ (black color) respectively.

Right panel (down): Class 1 mutations: single nucleotide substitutions detected in lacZ (4-vector system). The numbers, types, and locations of the independent substitution errors are shown for both the forward lacZ orientation (plus-strand nucleotide sequence) and the reverse lacZ orientation (negative-strand nucleotide sequence). Opposing directional arrows indicate the actual sequence context and direction of minus-strandDNA synthesis during reverse transcription. The total length of the lacZ target sequence was defined as 174 nt, representing codons 6 to 63 from GGA to the first TAA termination codon. Single nucleotide substitution errors are shown as letters above the original wild-type template sequence, limited to 11 per position, with additional errors indicated by n. Runs of 3 or more identical nucleotides are underlined. Misalignment/slippage of the primer or template strand that could result in a substitution error is highlighted in orange or green, respectively. Mutational hot spots for which there are significant differences in the forward and reverse lacZ orientations are indicated by asterisks: *, P<0.01; **, P< 0.001; ***, P< 0.0001 (see Materials and Methods). The mutation frequency is the number of lacZ DNAs with a mutation(s) divided by the total number of adjusted recovered clones. **Class 2 mutations:** single nucleotide frameshifts detected in lacZ (4-vector system). Single nucleotide frameshift errors are indicated as triangles above the wild-type sequence: single nucleotide deletions (-1) are shown as upright triangles and single nucleotide additions (+1) as inverted triangles. Runs of 3 or more identical nucleotides are underlined. If a frameshift occurred within a run, its exact position within the run cannot be determined, and the deletions and additions are marked over the last nucleotide in the run in the direction of minus-strand DNA synthesis. Frameshift errors in nucleotide runs that could have arisen by a misalignment/slippage of the primer or template strand are highlighted in orange or green, respectively. Mutational hot spots for which there are significant differences in the forward and reverse lacZ orientations are indicated by asterisks: *.

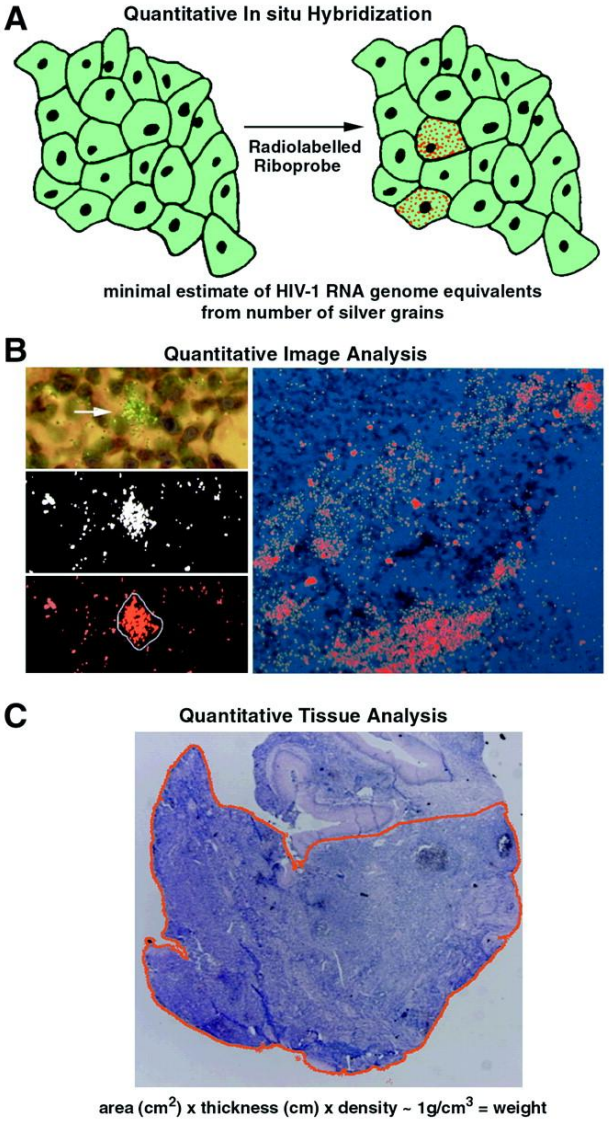
Figura S5. (Continua)



Class 3 mutations: multiple nucleotide substitutions detected in *lacZα* (4-vector system). Multiple mutations are shown as letters above the original wild-type template sequence, limited to 11 per position, with additional errors indicated by +*n*. Runs of 3 or more identical nucleotides are underlined. Mutational hot spots for which there are significant differences in the forward and reverse *lacZα* orientations are indicated by asterisks: *, $P < 0.01$; **, $P < 0.001$; ***, $P < 0.0001$ (see Materials and Methods). The mutation frequency is the number of *lacZα* DNAs with a mutation(s) divided by the total number of adjusted recovered clones.

Class 4 mutations: insertions, deletions, and/or deletions with insertions (indels) detected in *lacZα* (4-vector system). Indel mutations (deletions, deletions with insertions, and duplications) are shown as red lines extending over the positions of the mutations, with each *lacZα* DNA sequence numbered accordingly. The lines approximate the sizes of the indels, with the actual number of nucleotides inserted or deleted preceded by an inverted or upright triangle, respectively. Additional substitution errors within indels are shown as letters above the original wild-type template sequence. Runs of 3 or more identical nucleotides are underlined. The mutation frequency is the number of *lacZα* DNAs with a mutation(s) divided by the total number of adjusted recovered clones.

Figura S6. Métodos de estimación de la producción y eliminación de la progenie viral.



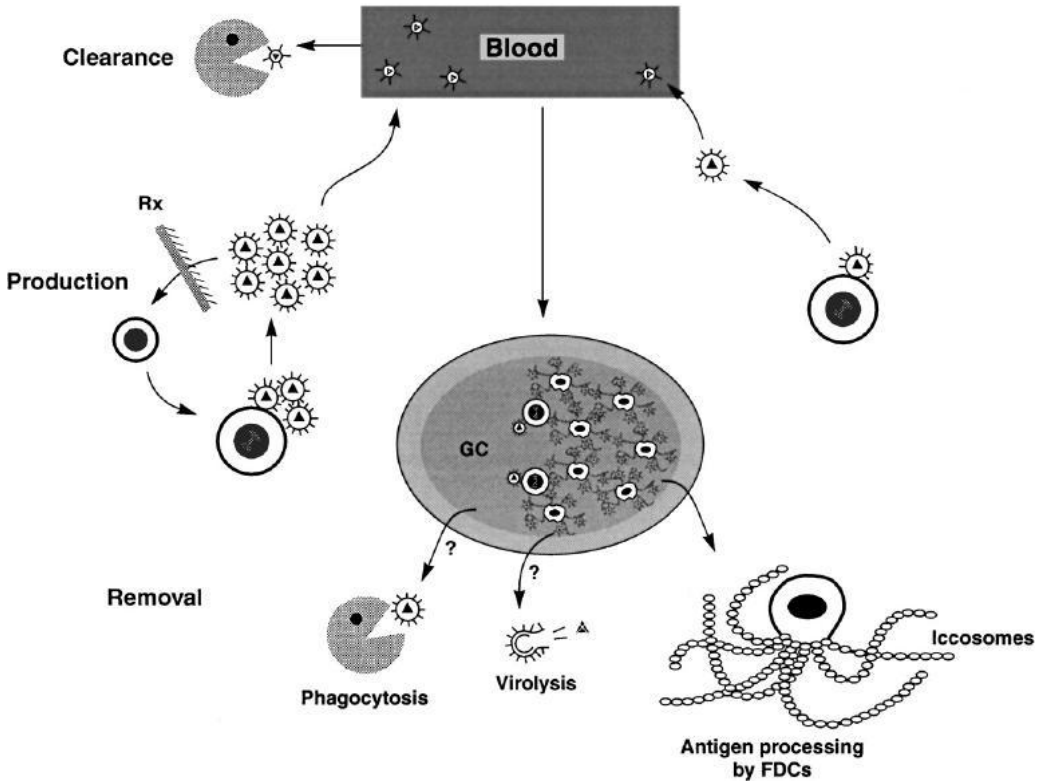


Figure 4 Hypothetical mechanisms of virus production and clearance. Most HIV-1 is produced in continuing de novo cycles of infection and reinfection of CD4+ T cells with short life spans, but smaller amounts of virus are produced by cells with longer life spans. Most of the virus in LT is in immune complexes stored on FDCs cleared by slow and possibly fast processes such as virolysis and phagocytosis. Treatment (Rx) interrupts the cycles of infection, levels of production and virus levels decline in the blood and FDC pool with biphasic kinetics related to the different life spans.

Figure 3 Quantitative image analysis of viral populations in LT. (*A*) The number of silver grains (*red dots*) overlaying a cell with viral RNA is proportional to the number of copies of viral RNA that bound the radiolabeled probe. The copy number can therefore be back-calculated from the grain count. (*B*) This task has been facilitated by quantitative image analysis as described in the text. In *C*, the traced area of the section is used to determine its weight. Population sizes are expressed per gram of tissue, which can be extrapolated to total body estimates. (Modified and reproduced with permission from Haase AT et al 1996. *Science* 274:985–89.).

Figura S7. Determinación del tiempo generacional en HIV-1. Resumen esquemático de la infección de HIV-1 *in vivo*. En el centro se muestran una población de viriones libres de células, los cuales son muestreados cuando la carga viral en el plasma es medido. En el lado derecho, se muestra la dinámica viral infectando continuamente a células blanco en una etapa inicial de la infección. La vida media de una célula productiva es de 1.6 días ($t_{1/2}$); mientras que el tiempo generacional de HIV-1 es de 2.6 días. En el lado izquierdo, se muestra la etapa posterior a la infección de HIV-1, en la que células del hospedero sirven como alojo para los virus que permanecerán de forma latente o, como células infectadas con virus defectuosos. Figura tomada de Perelson et al. (1996).

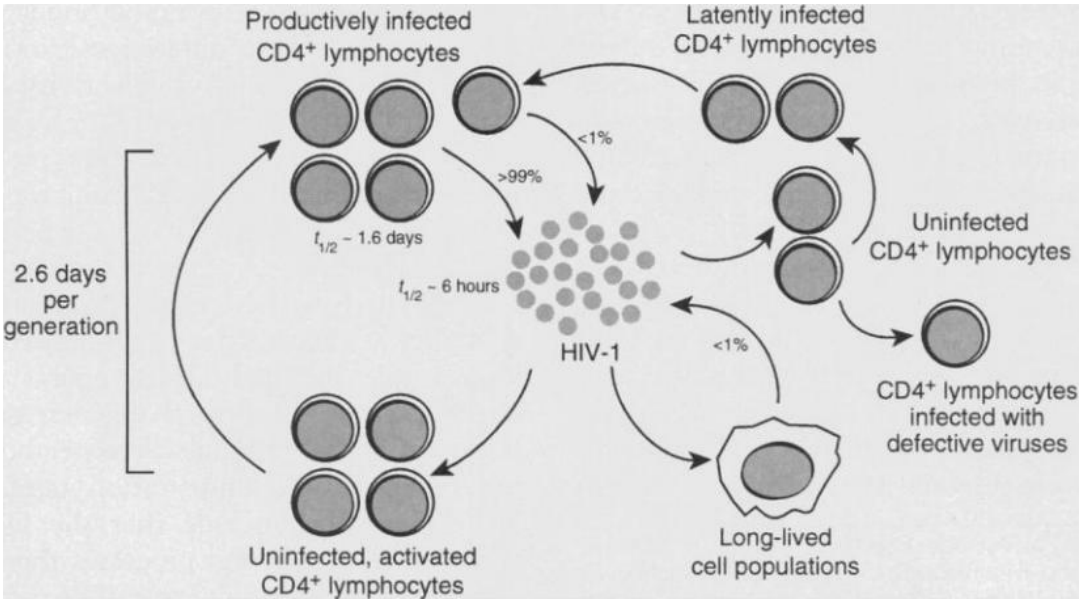


Figure S8. Gráficas comparativas de la prueba de selección en dS y dN realizada en el programa HyPhy (valores normalizados). *(A)* Comparación de SLAC con diferentes modelos de sustitución. *(B)* Comparación de FEL con diferentes modelos de sustitución. *(C)* Comparación de IFEL con diferentes modelos de sustitución. Para cada comparación independiente, nosotros usamos el modelo la herramienta de comparación de nucleótidos (codon modelo compare) implementado en el programa HyPhy. Nosotros comparamos, por tanto, dos modelos, uno obtenido por el programa (*best fitting models*) y otro elegido por nosotros, siendo "012312" *versus* HKY85 (010010), respectivamente.

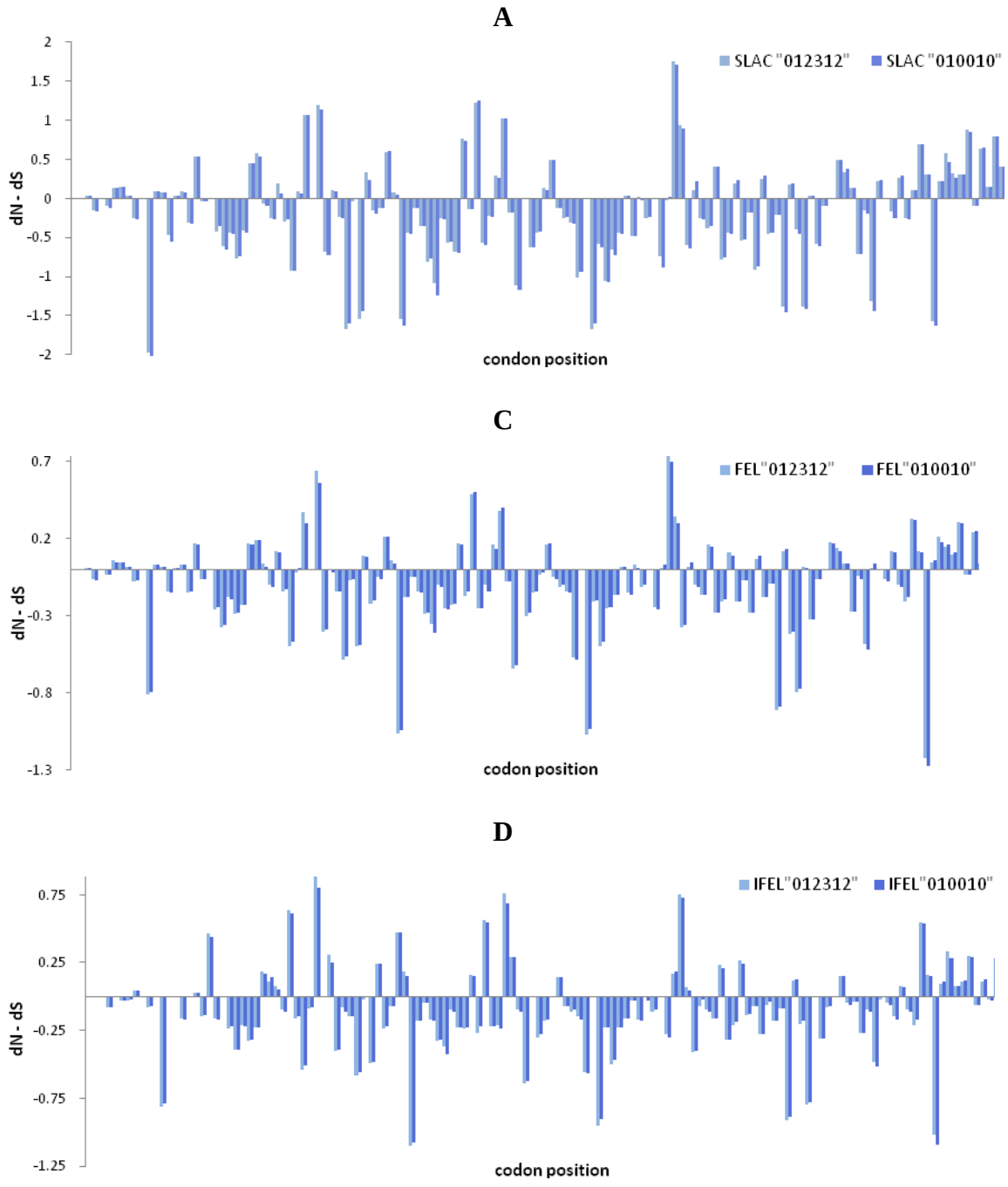


Figura S9. Dinámica de Cuasiespecies empleando el Fitness Mutacional usando Tajima. Simulaciones con 200 N_e y con 200 ciclos. Se empleó una tasa de mutación de 1 a 4 por ciclo de replicación por genoma.

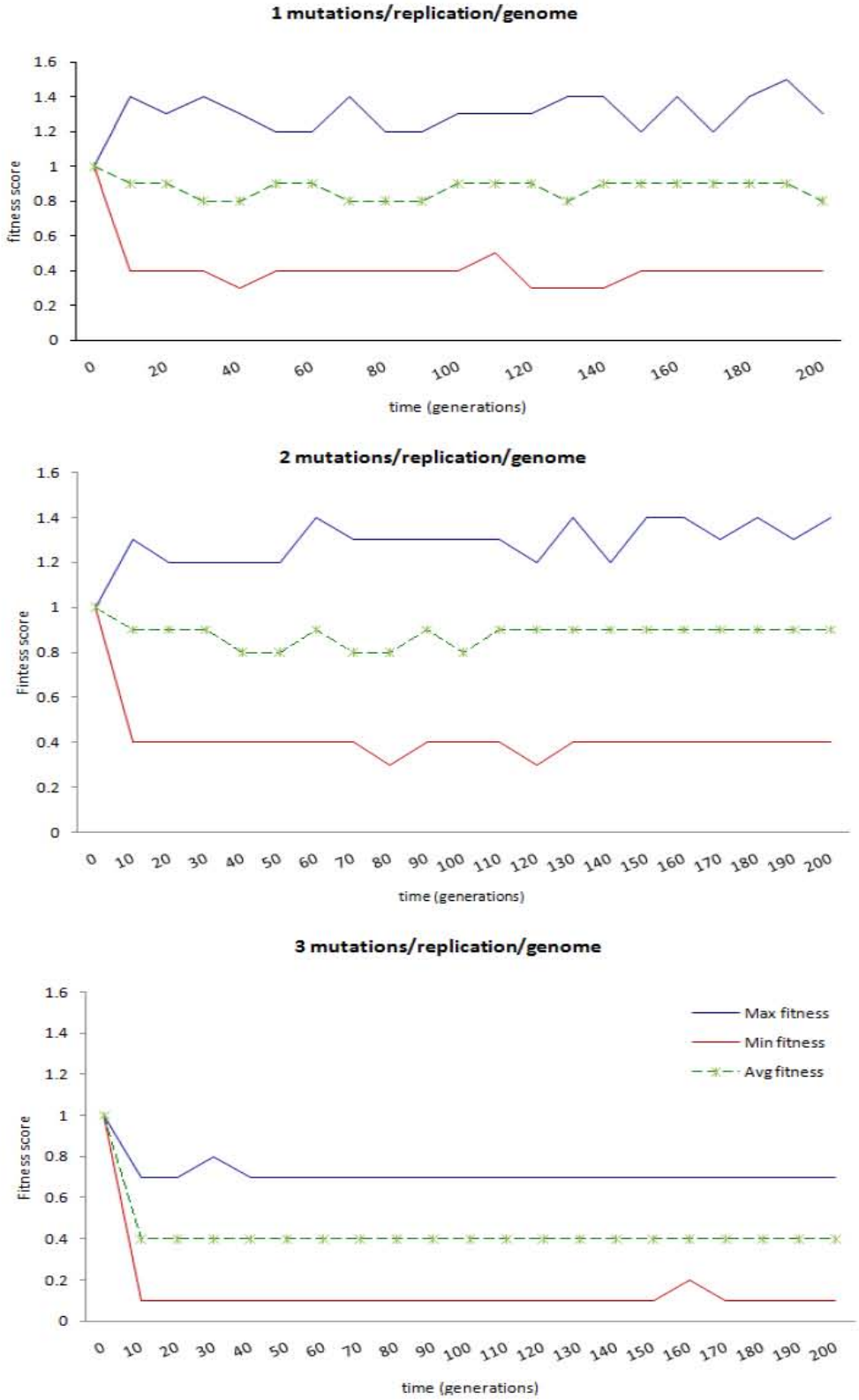
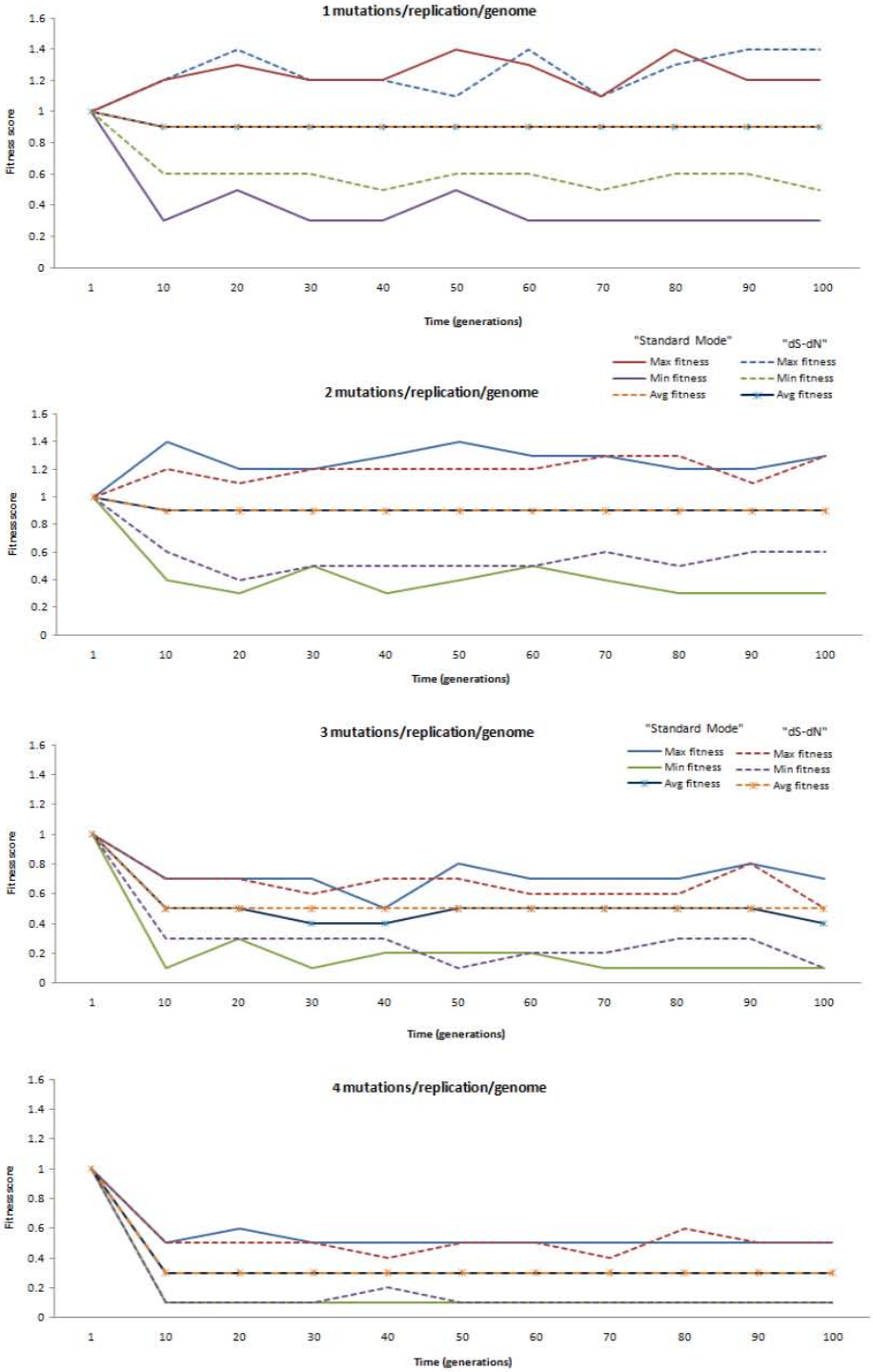


Figura S10. Dinámica de Cuasiespecies: comparación de los métodos dS-dN y el "modo estándar". Simulaciones con 100 genomas virales, 100 ciclos de replicación y un espectro mutacional (μ=1, 2, 3, 4 ciclo/rep/genoma). En esta sección de comparación de métodos, se obtuvieron los mismos resultados del *fitness mutacional* de *dS-dN* y del *modo estándar*. Sin embargo, existe un ligero efecto del *modo estándar*: expande la distribución del fitness en la población.



Tablas S1. Mapeo de sitios predichos con mutaciones que tienen un efecto positivo o negativo en el fitness del virus HIV-1 y que han sido reportados experimentalmente. HIV-1 genome HXB2:

A. Proteína GAG (poliproteína gag).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Initiator Methionine	1	1	Removed; by host By similarity
Molecule	2 – 500	499	Gag polyprotein
processing	2 – 132	131	Matrix protein p17 By similarity
	133 – 363	231	Capsid protein p24 By similarity|
	364 – 377	14	Spacer peptide p2 By similarity
	378 – 432	55	Nucleocapsid protein p7 By similarity
	433 – 448	16	Spacer peptide p1 By similarity
	449 – 500	52	p6-gag By similarity
Regions: Zinc finger	390 – 407	18	CCHC-type 1
	411 – 428	18	CCHC-type 2
Motif	16 – 22	7	Nuclear export signal
	26 – 32	7	Nuclear localization signal
	455 – 458	4	PTAP/PSAP motif
	483 – 492	10	LYPX(n)L motif
Sites	132 – 133	2	Cleavage; by viral protease By similarity
	363 – 364	2	Cleavage; by viral protease By similarity
	377 – 378	2	Cleavage; by viral protease By similarity
	432 – 433	2	Cleavage; by viral protease By similarity
	448 – 449	r2	Cleavage; by viral protease By similarity
Amino acid modification	387	1	Asymmetric dimethylarginine; in Nucleocapsid protein p7; by host PRMT6 By similarity
	409	1	Asymmetric dimethylarginine; in Nucleocapsid protein p7; by host PRMT6 By similarity
	2	1	N-myristoyl glycine; by host By similarity
Experimental Info: Mutagenesis	18	1	K → A: Replication-defective, induces nuclear mislocalization of matrix protein; when associated with G-
	22	1	R → G: Replication-defective, induces nuclear mislocalization of matrix protein; when associated with A-
	27	1	K → A: No effect on subcellular localization of matrix protein; when associated with A-18 and G-22. *[Ref.1]*

B. Proteína GAG-POL.

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Initiator Met	1	1	Removed; by host By similarity
Molecule	2 – 1435	1434	Gag-Pol polyprotein
Processing	2 – 132	131	Matrix protein p17 By similarity
	133 – 363	231	Capsid protein p24 By similarity
Peptide	364 – 377	14	Spacer peptide p2 By similarity
	378 – 432	55	Nucleocapsid protein p7 By similarity
Peptide	433 – 440	8	Transframe peptide Potential
	441 – 488	48	p6-pol Potential
	489 – 587	99	Protease
	588 – 1147	560	Reverse transcriptase/ribonuclease H
	588 – 1027	440	p51 RT
	1028 – 1147	120	p15
	1148 – 1435	288	Integrase By similarity
Regions:	508 – 577	70	Peptidase A2
Domain	631 – 821	191	Reverse transcriptase
	1021 – 1144	124	RNase H
	1201 – 1351	151	Integrase catalytic
Zinc finger	390 – 407	18	CCHC-type 1
Zinc finger	411 – 428	18	CCHC-type 2
Zinc finger	1150 – 1191	42	Integrase-type
DNA binding	1370 – 1417	48	Integrase-type
Region	217 – 225	9	PPIA/CYPA-binding loop
	814 – 822	9	RT 'primer grip' By similarity
Motif	16 – 22	7	Nuclear export signal By similarity
	26 – 32	7	Nuclear localization signal By similarity
	985 – 1001	17	Tryptophan repeat motif By similarity

B. (continua).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Sites:			
Active	513	1	For protease activity; shared with dimeric partner By similarity
Metal binding	697	1	Magnesium; catalytic; for reverse transcriptase activity By similarity
	772	1	Magnesium; catalytic; for reverse transcriptase activity By similarity
	773	1	Magnesium; catalytic; for reverse transcriptase activity By similarity
	1030	1	Magnesium; catalytic; for RNase H activity
	1065	1	Magnesium; catalytic; for RNase H activity
	1085	1	Magnesium; catalytic; for RNase H activity
	1136	1	Magnesium; catalytic; for RNase H activity
Sites	132 – 133	2	Cleavage; by viral protease By similarity
	221 – 222	2	Cis/trans isomerization of proline peptide bond; by human PPIA/CYPA
	363 – 364	2	Cleavage; by viral protease By similarity
	377 – 378	2	Cleavage; by viral protease By similarity
	393 – 394	2	Cleavage; by viral protease Potential
	426 – 427	2	Cleavage; by viral protease Potential
	432 – 433	2	Cleavage; by viral protease Potential
	440 – 441	2	Cleavage; by viral protease By similarity
	488 – 489	2	Cleavage; by viral protease By similarity
	587 – 588	2	Cleavage; by viral protease By similarity
	988	1	Essential for RT p66/p51 heterodimerization By similarity
	1001	1	Essential for RT p66/p51 heterodimerization By similarity
	1027 – 1028	2	Cleavage; by viral protease; partial By similarity
	1147 – 1148	2	Cleavage; by viral protease By similarity
Amino acid mod.	132	1	Phosphotyrosine; by host By similarity
Lipidation	2	1	N-myristoyl glycine; by host By similarity
Natural variations	496	1	R → K Confers to resistance to A-77003; when associated with other amino acid changes.

B. (continua).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Natural Variations	496	1	R → Q Confers to resistance to A-77003.
	498	1	L → F Confers resistance to amprenavir, atazanavir, lopinavir; when associated with other amino acid
	498	1	L → I / L → R / L → V / L → Y. Confers resistance to indinavir, lopinavir, ritonavir and
	503	1	I → V Confers resistance to tipranavir.
	504	1	G → E Confers resistance to lopinavir, ritonavir and saquinavir; when associated with other amino acid
	508	1	K → I / K → M / K → R. Confers resistance to lopinavir, to indinavir and nelfinavir
	511	1	L → I Confers resistance to BILA 2185 BS.
	512	1	L → I Confers resistance to amprenavir, indinavir, lopinavir, ritonavir and saquinavir; when associated
	518	1	D → N Confers resistance to nelfinavir; when associated with other amino acid changes.
	520	1	V → I Confers resistance to A-77003, amprenavir, atazanavir, indinavir, kynostatin, lopinavir, ritonavir
	521	1	L → F Confers resistance to atazanavir nelfinavir and ritonavir; when associated with other amino acid
	522	1	E → Q Confers resistance to lopinavir; when associated with other amino acid changes.
	523	1	E → D Confers resistance to tipranavir.
	524	1	M → I / M → L. Confers resistance to nelfinavir and ritonavir; when associated with other amino acid
	525	1	S → D Confers resistance to indinavir and tipranavir; when associated with other amino acid changes.
	529	1	R → K Confers resistance to tipranavir. |
	533	1	K → I Confers resistance to DMD-323; when associated with other amino acid changes.
	534	1	M → F / M → I / M → L . Confers resistance a antivirals*
	535	1	I → V Confers resistance to amprenavir, lopinavir, kynostatin, ritonavir and saquinavir; when associated
	536	1	G → V Confers resistance to A-77003, amprenavir, indinavir, ritonavir, saquinavir and telinavir; when
	538	1	I → L / I → V. Confers resistance to antivirals when associated with other amino acid changes*.
	541	1	F → L / F → Y. Confers resistance to antivirals when associated with other amino acid changes*.
	542	1	I → A / I → L / I → M / I → S / I → T. Confers resistance to amprenavir and lopinavir
	542	1	I → V Confers resistance to indinavir, lopinavir, ritonavir and saquinavir; when associated with other
	543	1	K → R Confers resistance to nelfinavir.
	545	1	R → K Confers resistance to nelfinavir.
	546	1	Q → E Confers resistance to lopinavir and ritonavir; when associated with other amino acid changes.
	548	1	D → E Confers resistance to tripanavir.

B. (continua).

Sequences annotation (features)			
Feature Key	*Position*	*Lenght*	*Description*
Natural variations	546	1	Q → E Confers resistance to lopinavir and ritonavir; when associated with other amino acid changes.
	548	1	D → E Confers resistance to tripanavir.
	549	1	Q → H Confers resistance to lopinavir; when associated with other amino acid changes.
	551	1	L → P / L → T. Confers resistance to atazanavir, indinavir, lopinavir, ritonavir and saquinavir; when associated with
	553	1	E → Q Confers resistance to lopinavir; when associated with other amino acid changes.
	554	1	I → F Confers resistance to indinavir, ritonavir and saquinavir; when associated with other amino acid changes.
	557	1	H → Y Confers resistance to lopinavir; when associated with other amino acid changes.
	559	1	A → I / A → L / A → T /A → V. Confers resistance to antivirals; when associated with other amino acid changes*.
	561	1	G → S Confers resistance to indinavir, nelfinavir, ritonavir and saquinavir; when associated with other amino acid
	565	1	V → I Confers resistance to indinavir, nelfinavir, ritonavir and saquinavir; when associated with other amino acid
	570	1	V → A / V → F / V → I / V → S / V → T . Confers resistance to antivirals; when associated with other amino acid
	572	1	I → A / I → V. Confers resistance to antivirals; when associated with other amino acid changes*.
	576	1	N → D / N → S Confers resistance to nelfinavir, atazanavir, indinavir and nelfinavir
	577	1	L → M Confers resistance to atazanavir; when associated with other amino acid changes.
	578	1	L → M Confers resistance to indinavir, lopinavir, nelfinavir, ritonavir and saquinavir; when associated with other amino
	579	1	T → S Confers resistance to lopinavir, ritonavir and saquinavir; when associated with other amino acid changes.
	581	1	I → L Confers resistance to antivirals; when associated with other amino acid changes*.
	628	1	M → L Confers resistance to zidovudine; when associated with other amino acid changes.
	631	1	E → A Confers resistance to lamivudine.
	631	1	E → D Confers resistance to zidovudine; when associated with other amino acid changes.
	639	1	P → R Confers resistance to stavudine.
	641	1	N → D Confers resistance to stavudine.
	649	1	A → V Confers multi-NRTI resistance; when associated with other amino acid changes.
	652	1	K → R Confers resistance to abacavir, adefovir, didenosine, lamivudine, stavudine, tenofir and zidovuline; when
	654	1	D → A / D → E / D → G D / z→ N / D → S. Confers resistance to zidovudine and multi-NRTI resistance.
	655	1	S → G / S → N / S → Y.Confers multi-NRTI resistance; when associated with other amino acid changes.
	656	1	T → A / T → D/T → G/ T → N. Confers resistance to antivirals when associated with other amino acid changes*.
	657	1	K → E / K → R/ K → S.Confers resistance to antivirals when associated with other amino acid changes*.
	661	1	L → I / L → V. Confers resistance to antivirals when associated with other amino acid changes*.
	662	1	V → I /V → L/V → M/V → T. Confers resistance to antivirals when associated with other amino acid changes*.
	664	1	F → L Confers multi-NRTI resistance; when associated with other amino acid changes.

B. (continua).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Natural variations	744	1	P → S Confers resistance to lamivudine.
	748	1	Q → L Confers resistance to pyrophosphate analog PFA.
	766	1	V → D Confers resistance to efavirenz, tivirapine and trovirdine; when associated with other amino acid changes.
	768	1	Y → C Confers multi-NNRTI resistance.
	771	1	M → I / M → T / M → V. Confers resistance to antivirals when associated with other amino acid changes*.
	775	1	Y → C / Y → H / Y → L. Confers resistance to antivirals when associated with other amino acid changes*.
	776	1	V → I Confers resistance to HBY 097.
	777	1	G → A / G → C / G → E/ G → Q / G → S/ G → T/ G → V. Confers resistance to antivirals when associated with
	795	1	H → Y Confers resistance to lamivudine, pyrophosphate analog PFA and zidovudine.
	797	1	L → W Confers resistance to zidovudine.
	798	1	R → K Confers resistance to lamivudine and zidovudine.
	801	1	L → F Confers resistance to ph-AZT and zidovudine.
	802	1	T → F / T → Y. Confers resistance to zidovudine; when associated with other amino acid changes.
	806	1	K → Q / K → R. Confers resistance to antivirals when associated with other amino acid changes*.
	812	1	P → H Confers resistance to efavirenz, emivirine, HBY 097 and quinoxaline; when associated with A-17.
	823	1	P → L Confers resistance to atevirdine and delavirdine.
	825	1	K → T Confers resistance to atevirdine and zidovudine; when associated with other amino acid changes.
	870	1	L → I Confers resistance to delavirdine, efavirenz and nevirapine.
	905	1	Y → F Confers resistance to delavirdine and nevirapine.
	920	1	G → D / G → E. Confers resistance to abacavir, lamivudine and zidovudine.
	973	1	T → I Confers resistance to abacavir, lamivudine and zidovudine.
Mutagenesis	217	1	P → A: 3-fold decrease of PPIA-binding affinity. *[Ref.1]*
	218	1	V → A: 2.7-fold decrease of PPIA-binding affinity. *[Ref.1]*
	219	1	H → A or Q: 8-fold decrease of PPIA-binding affinity. *[Ref.1]*
	220	1	A → G / A → V: 44 & 31-fold decrease of PPIA-binding affinity, respectively. *[Ref.1]*
	221	1	G → A / G → V: 31 & 154-fold decrease of PPIA-binding affinity, respectively. *[Ref.1]*
	222	1	P → A P → V: 36 6 150-fold decrease of PPIA-binding affinity, respectively. *[Ref.1]*
	223	1	I → A / I → V: ~1.2-fold decrease of PPIA-binding affinity. *[Ref.1]*
	224	1	A → G / A → V: ~2.3-fold decrease of PPIA-binding affinity. *[Ref.1]*
	225	1	P → A: 1.6-fold decrease of PPIA-binding affinity. *[Ref.1]*
	394	1	N → F or G: Decreases infectivity and replication. *[Ref.2]*
	400	1	H → C: Complete loss of infectivity and in vitro chaperone activity. *[Ref.3]*
	405	1	C → H: Complete loss of infectivity and DNA synthesis. *[Ref.3]*

B. (continua).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Mutagenesis	421	1	H → C: Partial loss of infectivity. Complete loss of in vitro chaperone activity. *[Ref.3]*
	426	1	C → H: Partial loss of infectivity. *[Ref.3]*
	1065	1	E → Q: Complete loss of RNase H activity. *[Ref.4]*
	1136	1	D → N: Complete loss of RNase H activity. *[Ref.4]*
	1159	1	H → C: No effect on integrase activity in vitro. *[Ref.5]*
	1163	1	H → C or V: 75% increase of integrase activity in vitro. *[Ref.5]*
	1187	1	C → A: Complete loss of integrase activity in vivo. *[Ref.6]*
	1190	1	C → A: Complete loss of integrase activity in vivo. *[Ref.6]*
	1200	1	Q → C: 75% increase of integrase activity in vitro. *[Ref.5]*
	1208	1	W → A: Complete loss of integrase activity in vivo. *[Ref.6]*
	1211	1	D → A or V: Complete loss of integrase activity in vivo and in vitro. *[Ref.5] [Ref.6] [Ref.7]*
	1213	1	T → A: No effect on infectivity. *[Ref.6]*
	1222	1	V → P: Complete loss of integrase activity. *[Ref.6]*
	1228	1	S → A / S → R: Complete loss of integrase activity in vivo & no effect on integrase activity in vitro, respectively.
	1262	1	T → A: No effect infectivity. *[Ref.6]*
	1263	1	D → A or I: Complete loss of integrase activity in vivo and in vitro. *[Ref.5] [Ref.6] [Ref.7]*
	1270	1	G → A: No effect on infectivity. *[Ref.6]*
	1282	1	I → P: Complete loss of integrase activity in vivo. *[Ref.6]*
	1298	1	V → A: No effect on infectivity. *[Ref.6]*
	1299	1	E → G or P: Complete loss of integrase activity in vitro. *[Ref.5] [Ref.6] [Ref.7]*
	1306	1	K → P: Slow down virus replication. *[Ref.6]*
	1326	1	A → P: Complete loss of integrase activity in vivo. *[Ref.6]*
	1346	1	R → C: 75% increase of integrase activity in vitro. *[Ref.5]*
	1382	1	W → A: Complete loss of infectivity. No effect on integrase activity in vitro. *[Ref.5] [Ref.6]*

Referencias GAG:

[1] Dupont S., Sharova N., DeHoratius C., Vibasius C.M., Zhu X., Bukrinskaya A.G., Stevenson M. and Green M.R. 1999. A novel nuclear export activity in HIV-1 matrix protein required for viral replication. Nature 402:681-685.

Referencias GAG-POL:

[1] Yoo S., Myszka D.G, Yeh C., McMurray M., Hill C.P. and Sundquist W.I. 1995. Molecular recognition in the human immunodeficiency in the HIV-1 capsid/cyclophilin A complex. J. Mol. Biol. 269:780-795.
[2] Thomas J.A., Shulenin S., Coren L.V., Bosche W.J., Gagliardi T.D., Gorelick R.J. and Oroszlan. 2006. Characterization of human immunodeficiency virus type 1 (HIV-1) containing mutations in the nucleocapsid protein at a putative HIV-1 proteae cleavage site. 2006. Virology 354:261-270.
[3] Guo J., Wu T., Kane B.F., Johnson D.G., Henderson L.E., Gorelick R.J. and Levin J.G. 2002. Subtle alterations of the native zinc finger structure have dramatic effects on the nucleic acid chaperone activity of human immunodeficiency virus type 1 nucleocapsid protein. J. Virol. 76:4370-4378.
[4] Cristofaro J.V., Rausch J.W., Le Grice S.F. and DeStefano J.J. 2002. Mutations in the ribonuclease H active site of HIV-RT reveal a role for this site in stabilizing enzyme-primer-template binding. Biochem. 41:10968-10975.
[5] Leavitt A.D., Shiue L. and Varmus H.E. 1993. Site-directed mutagenesis of HIV-1 integrase demonstrates differential effects on integrase functions in vitro. J. Biol. Chem. 268:2113-2119.
[6] Cannon P.M., Wilson W., Byles E., Kingsman S.M. and Kingsman A.J. 1994. human immunodeficiency virus type 1 integrase: effect on viral replication of mutations at highly conserved residues. J.Virol. 68:4768-4775.
7[10] Gaur M. and Leavitt A.D. 1998. Mutation in the human immunodeficiency virus type 1 integrase D,D(35)E motif do not eliminate provirus formation. J. Virol. 72:4678-4685.
[*] See website uniprot for details in viral resistance: http://www.uniprot.org/

Referencias VIF:

[1] Yang X. and Gabuzda D. 1998. Mitogen-activated protein kinase phosphorylates and regulates the HIV-1 Vif protein. J. Biol. Chem. 273:29789-29887.
[2] Mehle A., Strack B., Ancuta P., Zhang C., McPike M. and Gabuzda D. 2004. Vif overcomes the innate antiviral activity of APOBEC3G by promoting its degradation in the ubiquitin-proteasome pathways. J. Biol. Chem. 279:7792-7798.
[3] Yang X., Goncalves J. and Gabuzda D. 1996. Phosphorylation of Vif and its role in HIV-1 replication. J. Biol.Chem. 271:10121-10129.
[4] Goncalves J., Shi B., Yang X. and Gabuzda D. 1995. Biological activity of human immunodeficiency virus type 1 requires membrane targeting by C-terminal basic domains. J. Virol. 69:7196-7204.

C. Proteína VIF (factor de infectividad del virión).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Molecule processing	1 – 192	192	Virion infectivity factor By similarity
	1 – 150	150	p17 By similarity
	151 – 192	42	p7 By similarity
Regions	14 – 17	4	Interaction with host APOBEC3F; F1-box
	40 – 44	5	Interaction with host APOBEC3G; G-box
	54 – 72	19	Interaction with host APOBEC3F and APOBEC3G; FG-box
	74 – 79	6	Interaction with host APOBEC3F; F2-box
	75 – 114	40	RNA-binding Potential
	151 – 164	14	Multimerization By similarity
	171 – 172	2	Membrane association By similarity
Motif	108 – 139	32	HCCH motif By similarity
	144 – 153	10	BC-box-like motif
Sites	150 – 151	2	Cleavage in virion (by viral protease) By similarity
Amino acid modification	96	1	Phosphothreonine; by host MAP4K1 Probable
	144	1	Phosphoserine; by host
	155	1	Phosphothreonine; by host
	165	1	Phosphoserine; by host MAP4K1 Probable
	188	1	Phosphothreonine; by host
Mutagenesis	96	1	T → A: 90% loss of reverse transcriptase activity in virions; no effect on the ability to decrease
	96	1	T → E: Complete loss of viral infectivity in non permissive cells; no effect on the ability to decrease APOBEC3G level. *[Ref.1]*
	114	1	C → S: Reduces the ability to decrease APOBEC3G level; when associated with S-133. *[Ref.2]*
	133	1	C → S: Reduces the ability to decrease APOBEC3G level; when associated with S-114. *[Ref.2]*
	144	1	S → A: 90% loss of viral infectivity in non permissive cells; no effect on the ability to decrease APOBEC3G level. *[Ref.3]*
	157 – 160	4	KKIK → AAIA: 75% loss of membrane binding; decrease Pr55Gag binding. *[Ref.4]*
	173 – 179	7	RWNKPQK → AWNAPQA: 40% loss of membrane binding; decrease Pr55Gag binding.
	179 – 184	6	KTKGHR → ATAGHA: 25% loss of membrane binding; decrease Pr55Gag binding.
	188	1	T → A: No effect on the ability to decrease APOBEC3G level. *[Ref.2]*

D. Proteína VPR (proteína viral R).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Molecule processing	1 - 78	78	Protein vpr
Region	1 - 42	42	Homo-oligomerization

E. Proteína TAT (proteína de transactivación regulatoría).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Molecule processing	1 – 86	86	TAT protein
	1 – 48	24	Interaction eith human CREBBP
	22 – 37	16	Cysteine-rich
	38 – 48	11	Core
	49 – 86	38	Interaction with the host capping enzyme RNGTT
	49 – 57	9	Nuclear localization signal, RNA-binding (TAR), and protein transduction
	78 – 80	3	Cell attachment site
Site	11	1	Essential for Tat's translocation through the endosomal membrane
Modified residue	28	1	N6-acetillysine; by host PCAF
	50	1	N6-acetillysine; by host EP300 and GCN5L2
	51	1	N6-acetillysine; by host EP300 and GCN5L2
	52	1	Asymmetric dimethylarginine; by hosto PRMT6
	53	1	Asymmetric dimethylarginine; by hosto PRMT6
	71	1	Glycyl lisien isopeptide (Lys-Gly)
Natural variants:	73 – 86	14	Missing in isoform Short

F. Proteína REV (regulador de la expresión de proteínas virales/transactivator de Anti-represión).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Molecule processing	1 – 116	116	Protein Rev
Regions	18 – 26	9	Homomultimerization By similarity
Motif	34 – 50	17	Nuclear localization signal and RNA-binding (RRE) By similarity
	73 – 84	12	Nuclear export signal and binding to XPO1 By similarity
compositional bias	38 – 50	13	Poly-Arg
Amino acid Modifications	5	1	Phosphoserine; by host CK2 *[Ref.1]*
	8	1	Phosphoserine; by host CK2 *[Ref.1]*
	92	1	Phosphoserine; by host By similarity
	99	1	Phosphoserine; by host By similarity

References REV:

[1] Ratner L., Fisher A., Jagodzinski L.L., Mitsuya H., Liou R.S., Gallo R.C. and Wong-Staal F. 1987. Complete nucleotide sequences of functional clones of the AIDS virus. AIDS Res. Hum. Retroviruses 3:57-69.

G. Proteína ENV (glicoproteína de envoltura gp160).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Signal	1- 32	32	
Molecule	33 – 856	824	Envelope glycoprotein gp160
processing	33 – 511	479	Surface protein gp120 By similarity
	512 –	345	Transmembrane protein gp41 By similarity
Regions: topological domains	33 – 684	652	Extracellular Potential
Transmemb.	685 –	21	Potential
topological domains	706 –	151	Cytoplasmic Potential
Regions:			
gp120 / gp41	131 –	26	V1
	157 -	40	V2
	296 –	35	V3
	364 –	11	CD4-binding loop
	385 –	34	V4
	461 –	11	V5
	512 –	21	Fusion peptide Potential
	576 –	17	Immunosuppression
	662 –	22	MPER; binding to GalCer
Coiled coil	633 –	35	Potential
Motif	712 –	4	YXXL motif; contains endocytosis signal
	855 –	2	Di-leucine internalization motif
Site	511 –	2	Cleavage; by host furin By similarity

G. (continua).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Amino acid modifications:			
Lipidation	764	1	S-palmitoyl cysteine; by host *[Ref.1] [°]*
	837	1	S-palmitoyl cysteine; by host *[Ref.1] [°]*
Glycosylation	88	1	N-linked (GlcNAc...); by host Potential
	136	1	N-linked (GlcNAc...); by host Potential
	141	1	N-linked (GlcNAc...); by host Potential
	156	1	N-linked (GlcNAc...); by host Potential
	160	1	N-linked (GlcNAc...); by host Potential
	186	1	N-linked (GlcNAc...); by host Potential
	197	1	N-linked (GlcNAc...); by host Potential
	230	1	N-linked (GlcNAc...); by host Potential
	234	1	N-linked (GlcNAc...); by host Potential
	241	1	N-linked (GlcNAc...); by host Potential
	262	1	N-linked (GlcNAc...); by host Potential
	276	1	N-linked (GlcNAc...); by host Potential
	289	1	N-linked (GlcNAc...); by host Potential
	295	1	N-linked (GlcNAc...); by host Potential
	301	1	N-linked (GlcNAc...); by host Potential
	332	1	N-linked (GlcNAc...); by host Potential
	339	1	N-linked (GlcNAc...); by host Potential
	356	1	N-linked (GlcNAc...); by host Potential
	386	1	N-linked (GlcNAc...); by host Potential
	392	1	N-linked (GlcNAc...); by host Potential
	397	1	N-linked (GlcNAc...); by host Potential
	406	1	N-linked (GlcNAc...); by host Potential

G. (continua).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Amino acid Modifications:			
Glycosylation	448	1	N-linked (GlcNAc...); by host Potential
	463	1	N-linked (GlcNAc...); by host Potential
	611	1	N-linked (GlcNAc...); by host Potential
	616	1	N-linked (GlcNAc...); by host Potential
	624	1	N-linked (GlcNAc...); by host Potential
	637	1	N-linked (GlcNAc...); by host Potential
	674	1	N-linked (GlcNAc...); by host Potential
Disulfide bond	54 ↔ 74		By similarity
	119 ↔ 205		By similarity
	126 ↔ 196		By similarity
	131 ↔ 157		By similarity
	218 ↔ 247		By similarity
	228 ↔ 239		By similarity
	296 ↔ 331		By similarity
	378 ↔ 445		By similarity
	385 ↔ 418		By similarity
	598 ↔ 604		By similarity
Mutagenesis	764	1	C → S: Complete loss of palmitoylation, decreased association with host cell membrane lipid rafts,
	837	1	C → S: Complete loss of palmitoylation, decreased association with host cell membrane lipid rafts, decreased incorporation onto virions and severe reduction of infectivity; when associated with S-

Referencias ENV:

[*] Miyauchi K., Kim Y., Latinovic O., Morozov V. and Melikyan G.B. 2009. HIV enters cells via endocytosis and dynamin-dependent fusion wiht endosomes. Cell 137:433-444.

[°] Yang C., Spies C.P. and Compans R.W. 1995. The human and simian inmunodeficiency virus envelope glycoprotein transmembrane subunits are palmitoylated. PNAS 92:9871-9875.

[1] Yang C., Spies C.P. and Compans R.W. 1995. The human and simian immunodeficiency virus envelope glycoprotein transmembrane subunits are palmitoylated. PNAS 92:9871-9875.

H. Proteína NEF (factor negativo).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Initiator Met.	1	1	Removed by host
Molecule	2 – 206	205	Nef protein
Processing	58 – 206	149	C-terminal
	2 – 57	56	N-terminal, associates with the host plasma membrane
	7 – 26	20	Necessary for MHC-I internalitation
Regions	62 – 65	4	Acidic; stabilizes the interaction of NEF/MHC-I with host AP1M1; necessary for MHC-I
	69 – 78	10	SH3-binding, interaction with Src family tyrosine kinases
	108 – 124	17	Mediates dimerization, Nef-PTE1 interaction
Motif	148 - 180	33	Binding to ATP6V1H
	72 – 75	4	PxxP; stabilizes the interaction of NEF/MHC-I with host AP1M1; necessary for MHC-I internalization
	164 – 165	2	Di-leucine internalization motif; necessary for CD4 internalization
	174 – 175	2	Diacidic; necessary for CD4 internalization
Comp. bias	62 – 65	4	Poly_Glu
Site	20	1	Might play a role in AP-1 recruitment to the Nef-MHC-I complex
	57 - 58	2	Cleavage; by protease
Modif. aa:	2	1	N-myristoyl glycine; by host
Experimental	20	1	M → A: Complete loss of Nef-induced MHC-I down-regulation, MHC-I is internalized but not
mutagenesis	62 – 65	4	EEEE → AAAA: Complete loss of Nef-induced MHC-I down-regulation, MHC-I is not internalized.
	62 – 65	4	EEEE → AAEA: About 50% loss of Nef-induced MHC-I down-regulation. *[Ref. 2 and 3].*
	62 – 65	4	EEEE → About 50% loss of Nef-induced MHC-I down-regulation. *[Ref. 2 and 3].*
	62 – 65	4	EEEE → DDDD: No effect on Nef-induced MHC-I down-regulation. *[Ref. 2 and 3].*

H. (continua).

Sequences annotation (features)			
Feature Key	*Position (s)*	*Length*	*Description*
Mutagenesis	62 – 64	3	EEE → AAA: About 50% loss of Nef-induced MHC-I down-regulation *[Ref. 2 and 3].*
	62 – 63	2	EE → AA: Almost no effect on Nef-induced MHC-I down-regulation.
	63 – 65	3	EEE → AAA: About 50% loss of Nef-induced MHC-I down-regulation. *[Ref. 1]*
	63 – 64	2	EE → AA: Almost no effect on Nef-induced MHC-I down-regulation.
	64 – 65	2	EE → AA: Almost no effect on Nef-induced MHC-I down-regulation.
	72	1	P → A: Complete loss of Nef-induced MHC-I down-regulation, MHC-I is not internalized; when
	75	1	P → A: Complete loss of Nef-induced MHC-I down-regulation, MHC-I is not internalized; when
	123	1	D → E: Complete loss of Nef-induced MHC-I down-regulation, CD4 down-regulation, and
Sequence conflict	29	1	R → G in CAA26946. *[Ref. 4]*

Referencias NEF:

[1] Blagoveshchenskaya A.D., Thomas L., Feliciangeli S.F., Hung C.H., Thomas G. 2002. HIV-1 Nef downregulates MHC-I by PACS-1 and PI3k-regulated ARF6 endocytic pathway.Cell 111:853-866.

[2] Baugh L.L., Garcia J.V. and Foster J.L. 2008. Functional characterization of the human immunodeficiency virus type 1 Nef acidic domain. J. Virol. 82:9657-9667.

[3] Atkins K.M., Thomas L., Youker R.T., Harriff M.J., Pissani F., You H. and Thomas G. HIV-1 Nef binds PACS-2 to assemble a multikinase cascade that triggers major histocompatibility complex class I (MHC-I) down-regulation: analysis using short interfering RNA and knock-out mice. J. Biol. Chem. 283:11772-11784.

[4] Ratner L., Starcich B.R., Josephs S.F., Hahn B.H. Reddy E.P., Livak K.J., Petteway S.R. Jr., Pearson M.L., Haseltine W.A. and Wong-Staal F. 1985. Polymorphysm of the 3' open reading frame of the virus associated with acquired immune deficiency syndrome, human T-lymphotropic virus type III. Nucleic Acids Res. 13:8219-8229.

Tabla S2. Lista de los genomas de HIV-1 compilados en el presente trabajo. 124 genomas se empelaron para este estudio. Se muestran el país de procedencia de cada genoma y el número de secuencias empleadas. La etiqueta de acceso está en el siguiente orden separado por un punto: subtipo, país, año de obtención, nombre de la secuencia y número de acceso al Genebank.

País	Numero de GeneBank	Numero	País	Numero de GeneBank	Numero	País	Numero de GeneBank	Numero
Canada (CA)	B.CA.2000.CANA6FULL.AY779552	6	**Brazil (BR)**	B.BR.1990.BZ167.AB485641	17	**China (CN)**	B.CN.1999.plwj.GU177863	5
	B.CA.1999.CANA5FULL.AY779551			B.BR.1990.BZ167.AB485642			B.CN.2002.02HNsq4.DQ007902	
	B.CA.1998.CANA3FULL.AY779550			B.BR.2002.02BR008.DQ358808			B.CN.2001.pCNHN24.EF057102	
	B.CA.1998.CANC10FULL.AY779557			B.BR.2005.BREPM1084.FJ195088			BC.CN.1996.YNRL9613.AY967803	
	B.CA.1996.CANC8FULL.AY779563			B.BR.2003.BREPM1024.EF637056			01_AE.CN.2006.06GX239.GU564230	
	B.CA.1992.CANC5FULL.AY779561			B.BR.2005.BREPM1093.FJ195089		**Macao (China) (MO)**	BFG.MO.2005.MO108.GU207884	1
Gabonese (West C. Africa) (GA)	B.GA.1988.OYI	1		B.BR.2004.BREPM1070.FJ195086				
				BF.BR.1999.BREPM278.AY771593		**Holanda (NL)**	U.NL.2001.U_NL_01_H10986_C11.E	1
United States (US)	B.US.1998.98USHVTN3605c9.AY560	23		BF1.BR.2000.BREPM13853.DQ08587				
	B.US.1998.98USHVTN1925c1.AY560			BF1.BR.2002.02BR033.DQ358811		**Japón (JP)**	BF1.JP.-.DR0769.AB253430	4
	B.US.2003.L8152.FJ469742			BF.BR.2003.BREPM1026.EF637055			B.JP.2004.DR6089.AB286955	
	B.US.2004.014837G.FJ469684			F1.BR.1993.93BR020_1.AF005494			F1.JP.2004.DR6082.AB480298	
	B.US.2002.L8124P.FJ469741			F1.BR.1990.BZ163.AB485656			46_BF.JP.2004.DR6082.AB480299	
	B.US.-.F7174.DQ886032			BF1.BR.2006.06BR_FPS561.HM0264				
	B.US.1989.P896			39_BF.BR.2004.04BRRJ179.EU7355		**Vietnam (Asia) (VN)**	01_AE.VN.1997.97VNAG212.FJ1852	1
	B.US.1997.ARES2			BF1.BR.2002.02BR005.DQ358806				
	B.US.2006.CH40E_flG3.FJ495825			40_BF.BR.2004.04BRSQ46.EU73554		**Ucrania (Europa) (UA)**	A1.UA.2000.98UA0116.AF413987	1
	B.US.1989.5073_89.AY835766							
	B.US.1985.5077_85.AY835769		**Venezuela (VE)**	28_BF.VE.1999.V62.AY536236	1	**Corea de Sur (KR)**	B.KR.2005.05CSR3.DQ837381	1
	B.US.-.PRLS28.FJ469766							
	B.US.1999.PRB959_03.AY331296		**Argentina (AR)**	BF.AR.1999.A047	2	**Australia (AU**	B.AU.2004.Phcsffull04.AY818644	2
	B.US.2004.CR0068P.FJ469695			BF.AR.1997.A32878			B.AU.-.C42.AF538305	
	B.US.1998.98USHVTN941c1.AY5601					**Rep. Cyprus (Grecia) (CY)**	B.CY.2005.CY120.FJ388931	7
	B.US.-.CR0164U.FJ469702		**Vector Structure**	AF324493|	1		B.CY.2005.CY056.FJ388905	
	B.US.1983.5084_83.AY835754						B.CY.2005.CY033.FJ388957	
	B.US.2004.F7172.FJ469733		**Spain (ES)**	F1.ES.-.X1670.DQ979024	1		B.CY.2005.CY074.FJ388915	
	B.US.1998.15385_1.DQ853464						B.CY.2005.CY075.FJ388916	
	B.US.1983.RF						B.CY.2005.CY124.FJ388933	
	B.US.1990.US1.AY173952		**France (FR)**	*B.FR.1983.HXB2-LAI-IIIB-BRU.K0*	3		DG.CY.2005.CY063.FJ388908	
	B.US.1986.5113_86.AY835757			B.FR.1985.LW123.U12055				
	B.US.-.1229I.FJ469685			F1.FR.1996.96FR-MP411.AJ249238		**Russia (RU)**	06_cpx.RU.2005.04RU001.DQ40085	1
Afghanistan (AF)	35_AD.AF.2007.221H.GQ477449	2	**Kenya (KE)**	A.KE.2006.06KECst_021.FJ623477	12	**Tanzania (Africa) (TZ)**	A1D.TZ.1996.TZBFL0088.AF442570	5
	35_AD.AF.2006.047H.GQ477443			A.KE.2006.06KECst_028.FJ623479			A1D.TZ.1997.TZBFL0086	
Democ. Rep. Congo	26_AU.CD.2002.02CD_LBTB084.FM8	4		A1.KE.1997.ML752.AY322193			10_CD.TZ.1996.96TZ_BF071.AF289	
(Africa) (CD)	26C.CD.1997.97CD_KFE267.FM8777			A1.KE.1986.ML170_1986.AF539405			A1.TZ.1997.97TZ03	
	26C.CD.1997.MBFE250.FM877783			A1.KE.1997.ML013_2.AY322185			ACD.TZ.2001.HS123.AY945709	
	K.CD.1997.97ZR-EQTB11.AJ249235			A1C.KE.1997.ML170_1997.AF53940				
Côte d' Ivoire (CI)	09_cpx.CI.2000.00IC_10092.AJ86	2		A1C.KE.1995.ML170_1995.AF53940		**Rwanda (Central Africa) (RW)**	A1.RW.1992.92RW025A.AB287376	1
(West Africa)	0209.CI.2001.01IC_PCI118.AJ866			A1.KE.1994.Q23_17.AF004885				
				A.KE.2006.06KECst_019.FJ623478		**India (IN)**	B.IN.-.11807.	1
Cameroon (CM)	11_cpx.CM.1997.MP818.AJ291718	1		CD.KE.1997.pWCML249.AY445524				
				D.KE.1997.ML415_2.AY322189		**Uganda (Africa) (UG)**	A1.UG.1992.UG029.AB098332	6
Sweden (SW)	A1.SE.1995.UGSE8131.AF107771	2		DG.KE.2006.06KECst_014.FJ62349			A1.UG.1992.UG029.AB098333	
	A1.SE.1994.SE7535.AF069671							
				DG.KE.2006.06KECst_014.FJ62349			A1.UG.1992.92UG037.AB253429	
Estonia (North Europe) (EE)	06_cpx.EE.2001.EE0359.AY535659	1	**Ghana (East Africa) (GH)**	0206.GH.2003.03GH195AG_06.AB28	1		A1.UG.1985.U455	
							D.UG.1992.92UG001.AJ320484	
Niger (East Africa (NE)	0206.NE.2000.NE95.AJ508596	1	**Botswana (South Africa) (BW)**	C.BW.2000.00BW147127.AF443091	1		A1D.UG.1992.UG035.AY352656	

Tabla S3. Calculo de la conservación de los codones, nucleótidos y aminoácidos en el genoma de HIV-1. La conservación de codones se estimo basado en el número total de codones que no presentaron cambios al 100% (idénticos) en cada uno de los alineamientos (*1*). La conservación de nucleótidos se calculó considerando todos aquellos sitios invariables (idénticos al 100%) en el alineamiento y en aquellos que presentaron pocos cambios con una significancia razonable (score >80%), por lo que el nivel de conservación que presentamos es de 100% y >80% (*2*). La conservación de aminoácidos se computó con el programa *score_conservation.phy* implementado por Capra y Singh (2007) (*3*). Bajo la hipótesis de que los sitios conservados están bajo una significante presión evolutiva, y que en las posiciones bajo presión se espera que tengan una distribución de aminoácidos muy diferente a aquellas columnas que no la tengan, Capra y Singh calculan esta diferencia empleando el análisis de divergencia de Jensen-Shannon (JSD).

Gen	Codones[1]	Nucleótidos (%)[2]	Aminoácidos (%)[3]
5'LTR	36	88.63	79.2
Gag	31	88.42	79.4
*Gag-Pol**	-	*90.3*	78.9
Pol	58	87.3	77.5
Vif	10	87.19	75.3
Vpr	2	86.41	76.6
Tat	1	83.3	78.5
Rev	0	81.86	77.4
Vpu	0	82.1	75.5
Env	21	79.6	76.8
Nef	5	89.89	78.3
3' LTR	6	81.51	77.1
Total (Σ)* **	**170 (~1.8%)**	**84.78**	**77.6**

Referencias suplementarias:

- Livingstone C.D. and Barton G.J. 1993. Protein Sequence Alignments: A Strategy for the Hierarchical Analysis of Residue Conservation. *CABIOS* 9(6):745-756.
- Luque, B. 2003. An introduction to physical theory of molecular evolution. Cen. Eur. J. Phys. 3:516-555.
- Capra, J.A. and Singh, M. 2007. Predicting functionally important residues from sequence conservation. Bioinformatics, 23(15):1875-82.

Extractos de los programas empleados: *quasispecies.pl*.

```
#!/usr/bin/perl

# quasispecies.pl
# May 2010
# Questions to Irma Lozada-Chavez and Alejandro N. Lozada-Chavez:
# ilozada@bioinf.uni-leipzig.de

# Perl modules -----
use strict;
use diagnostics;
use FileHandle;                 # To avoid $| = 1;
use DirHandle;                  #
use Cwd;                # Determine the process's current working directory
use Getopt::Long;       # Check if it's necessary to switch to Getopt::Long::order
use List::Util qw(shuffle);
srand (time ^ $$ ^ unpack "%L*", `ps axww | gzip -f`);
#-------------------------------------------------------------


# Usage message -----
#-------------------------------------------------------------
my $usage= << "USAGE";
$0: is a simple computer simulation to test quasispecies theory on RNA virus evolution

 usage: perl $0 -i <infile> -c <cycles> -p <population> -m <mutation> -f <fitness>
                                -s <sampling> -d <default[y/n]>

 arguments:
   -i <infile>                  File containing the nucleotide genome-master sequence in FASTA format.
                                (mandatory)

   -o <outfile>         Subdirectory path where the outfile results will be written.
                                (optional)

   -t <table>                   File containing the mutational-map nucleotide position across the genome:
                                A. neutral, positive and negative mutations.
                                B. synonymous and non-synonymous positions.
                                (mandatory)

   -c <cycles>                  Number of rounds of production of viral progeny (replication cycles)
                                to mantain the recursive mutation-selection process.
                                (1000 by default)

   -m <mutation>        Mutation rate of the viral population (mutations/site/replication)
                                = mutations per genome per round of replication.
                                (1 by default for retroviruses)

   -p <population>      Effective population size = the smallest theoretical population size .
                                This size must be kept constant.
                                (10000 by default)

   -s <sampling>        The sampling range to report the sequences and statistics of $0.
                                (100 cycles by default)

   -d <default>         Run $0 with all default parameters: yes [YyTt] / no [NnFf]
                                ('n' by default)
                                NOTE:  You can use this option to run $0 in a parallel way (cluster)
                                changing the DEFAULT parameters in the script.

Questions to Irma Lozada-Chavez and Alejandro N. Lozada-Chavez: ilozada at bioinf.uni-leipzig.de
USAGE
#-------------------------------------------------------------
```

Extractos de los programas empleados: *sesibilidad_especificidad.pl*

```
#!usr/bin/perl

##############################################
#
# Este programa evalua la sensibilidad y especificidad
# de dos métodos Tajima vs dN/dS.
#
##############################################

use strict;
#use warnings;

#--------------------------
#  VARIABLES
#--------------------------

my $in;
my $infile;
my @table2;
my %hash;
my $hash;
my $foreach;
my @table;
my $todos;
my @array_hash;
my $i;
my $pos_codon = 0;
my $pos_tajima = "";
my $pos_hyphy = "";
my $pos_exp = "";

my $correction;

#--------------------------
#               USAGE
#--------------------------

&usage()
   if scalar @ARGV !=1;

$in = $ARGV[0];

#--------------------------
# OPEN & READ FILE
#--------------------------

open(INFILE, "<".$in) or die "\n Can not open file: $in:
                                                -> Don't exist this file!
                                                -> Please re-check your DIR to open the infile.\n
Don't forget USAGE: perl evaluating_two_methods.pl infile > outfile\n\n";

$infile = <INFILE>;
chomp $infile;

############################
# MAIN                                  #
############################


@table= <INFILE>;
#chomp @table;
#print  "@table";

foreach $foreach (@table){
               $foreach =~ s/^\s+$//gis;
               #$foreach =~ s/\t/:/;
               }
```