Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

ESCOLA NACIONAL DE SAÚDE PÚBLICA
SERGIO AROUCA
ENSP

"Elementos de transposición en el genoma de Anopheles gambiae"

por

Rita Daniela Fernández Medina

Tese apresentada com vistas à obtenção do título de Doutor em Ciências na área de Saúde Pública.

Orientador principal: Prof. Dr. Claudio José Struchiner
Segunda orientadora: Prof.ª Dr.ª Claudia Marcia Aparecida Carareto
Terceiro orientador: Prof. Dr. José Marcos Chaves Ribeiro

Rio de Janeiro, setembro de 2009.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

ESCOLA NACIONAL DE SAÚDE PÚBLICA
SERGIO AROUCA
ENSP

Esta tese, intitulada

"Elementos de transposición en el genoma de Anopheles gambiae"

apresentada por

Rita Daniela Fernández Medina

foi avaliada pela Banca Examinadora composta pelos seguintes membros:

Prof. Dr. Elgion Lucio da Silva Loreto

Prof. Dr. Alexandre Afranio Peixoto

Prof. Dr. Gonzalo José Bello Bentancor

Prof. Dr. Adauto José Gonçalves de Araújo

Prof. Dr. Claudio José Struchiner – Orientador

Tese defendida e aprovada em 21 de setembro de 2009.

Este trabajo fue realizado en el Programa de Computación Científica (PROCC) de la Fundación Osvaldo Cruz y fue financiado con recursos de la Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) y la Fundação de Amparo à Pesquisa de Estado de Rio de Janeiro (FAPERJ).

A Antonia y Felipe

...si un río há de llegar al mar lo hace a base de girar a derecha e izquierda, cuando sin lugar a dudas sería más rápido, más *práctico*, ir directamente hasta su meta en lugar de complicarse la vida con todas esas curvas, logrando tan sólo alargar el camino tres veces –tres coma catorce veces, para ser exactos- tal y como han verificado los científicos con una precisión científica y bella.

-Es como si se vieran *obligados* a girar, ¿comprende?, parece algo absurdo, si lo piensa no puede evitar tomarlo como algo absurdo, pero el hecho es que *deben* avanzar de esa manera, trazando una curva tras otras, y no es una manera absurda ni lógica, no es correcto o erróneo, es su manera, simplemente, su manera, y ya está.

Alessandro Baricco
City, 1999

AGRADECIMIENTOS

Comecei o doutorado em Saúde Pública na ENSP, em março de 2004. Nos mais de cinco anos transcorridos foram muitas as pessoas que estiveram por perto, que me apoiaram e acompanharam em diferentes momentos, compartilhando momentos de alegria e desesperação. A elas dedico este agradecimento.

Ao Dr. Claudio Struchiner, pelo apoio institucional que me brindou no PROCC e por ter me convidado a trabalhar com elementos de transposição -que não era o meu projeto de tese inicial- e que acabou sendo um caminho muito rico e interessante.

Ao Dr. José Marcos Ribeiro, sem quem esta tese não teria sido possível. Agradeço as discussões, a disponibilidade para responder minhas duvidas e questões, para programar em função de minhas necessidades, pela leitura crítica de manuscritos e tese e por todo o trabalho envolvido no desenvolvimento de AnoTExcel.

À Dra Claudia Carareto, quem aceitou a co-orientação da tese praticamente no final, mas sem quem não teria sido possível a finalização da mesma. Agradeço a disponibilidade, a ajuda desinteressada, as leituras feitas do material em tempos recordes, as discussões e o recebimento no seu laboratório no IBILCE (UNESP). Muito obrigada pela compreensão e carinho.

Aos colegas e amigos do PROCC com muito carinho, em especial a Lu y Ronaldo, a Ernesto, Aline, Franklin, Camila & Felipe. A Felipe Figueiredo pelos programinhas em Pearl! a Carlos y Ana Lúcia, muito obrigada pela ajuda administrativa!

Aos amigos “cariocas”. A Andreita por tantas y tantas cosas, a Elvirita & Rafa, a Caro -a la distancia-, a Horacio, a Maria José, amiga desde o começo dos meus dias no Rio, a Kaori & Georges, a Gonzalo & ZeZé, a Federico & Natacha, a Bete, Dora, Claudio & Tania.

A los amigos de otras bandas, a Flor por esta amistad de años, que desafia las distancias, a Pau, a Pablo con mucho amor, a Marce por el incentivo y el cariño de siempre.

Un gracias especial y afectuoso a Ana Lila, quien realmente me ayudo a entrar (y salir!) de la fase final de esta tesis.

À “familia carioca”, os Santos-Peixoto com muito carinho.

A “la familia porteña”, mis queridos Rabossis con todo el amor.

A mis padres queridos Rosa & Daniel, por haberme apoyado incondicionalmente en todos mis proyectos de vida, por el amor y en especial, a mamá por la ayuda enorme de los últimos tiempos.

Por ultimo, a Antonia y Felipe, mis preciosuras que le dan sentido a todas las cosas. Y a Fer, por el amor de tantos años, y por la paciencia infinita de los últimos tiempos.

INDICE

Contenido del CD

-Carpeta conteniendo

Anexo 1
Cuadro 1_Relaciones entre taxa utilizados para clasificar elementos de transposición
Tabla 1_ Criterios de Clasificación de elementos de transposición

Anexo 2
-Matriz de distancias-p de elementos COPIA en AnoTExcel.xls
-Matriz de distancias-p de elementos GYPSY en AnoTExcel.xls
-Matriz de distancias-p de elementos GYPSY en Repbase.xls
-Matriz de distancias-p de elementos Pao-Belen AnoTExcel.xls

Anexo 3
Filogenias de elementos NLTRs analizados en la sección de Resultados
Filogenias de elementos Clase II analizados en la sección de Resultados

-Base de datos AnoTExcel.xls
-LINKS archivos de los *hiperlinks* de la base de datos

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE CUADROS

LISTA DE ABREVIATURAS

- Aa - Aminoácidos
- cDNA - DNA complementar
- dN - Sustituciones en posiciones no sinónimas
- DNA - Ácido desoxirribonucléico
- dS - Sustituciones en posiciones sinónimas
- DS- Desvío Standard
- EN - Envoltura
- EST - Expressed sequence tags
- Gag – gen específico de antígeno
- Gb - Gigabases
- Indels – inserciones y deleciones
- INT - Integrasa
- Kb - Kilobases
- LARD - *Large Retrotransposon Derivatives*
- LINE - *Long Interspersed Elements*
- LTR - Long Terminal Repeat (Terminaciones Repetitivas Largas)
- Mb - Megabases
- MITE - *Miniature Inverted Repeat Transposable Elements*
- mRNA - Acido ribonucleico mensajero
- NLTR - Non-LTR
- Nts - nucleótidos
- OMS - Organización Mundial de la Salud
- ORF - Open Reading Frame
- pb - Pares de base
- PepA17 - Peptidasa A 17
- Pol – Gene de la polimerasa

- RB - Base de datos Repbase
- RM – Programa *Repaeat masker*
- RNAm - Ácido ribonucléico mensageiro
- RT – Transcriptasa Reversa
- RT - Transcriptase reversa
- SINE - *Short Interspersed Elements*
- SNAC - *Small Non-Autonomous CACTA elements*
- TE - Elementos de Transposición
- Tf - Base de datos Tefam
- TIR - Terminal Inverted Repeat
- TRIM -*Terminal Repeat Retroelements in Miniature)*

RESUMEN

Hasta la fecha, no existen mecanismos eficientes de control de la malaria, una de las enfermedades infecciosas más importantes del mundo. En las últimas décadas, la transformación genética de los mosquitos transmisores de la malaria ha sido propuesta como una alternativa para su control. Para ello, además del desarrollo de técnicas para la introducción de DNA foráneo en células germinales de mosquitos y la identificación de genes capaces de bloquear o reducir la transmisión de los parásitos a humanos, es necesario el desarrollo de métodos eficientes para la introducción y fijación de los genes refractarios a la malaria en las poblaciones naturales de mosquitos. Se ha sugerido para tales fines, el uso de elementos de transposición debido a las características biológicas de invasión y propagación en genomas eucariotas que presentan.

En esta tesis se propuso analizar los elementos de transposición presentes en el genoma del mosquito *Anopheles gambiae*, uno de los principales vectores de la malaria en el mundo.

Los resultados presentados se encuentran divididos en tres partes, primeramente la descripción de AnoTExcel, una base de datos con información detallada de los elementos de transposición presentes en el genoma de *Anopheles gambiae*. Por otra parte, a partir de esta base se identificaron y caracterizaron elementos nuevos, no detectados previamente en ningún genoma. Por último, se analizaron elementos representativos de las diferentes clases de elementos de transposición desde una perspectiva evolutiva. También se propuso el análisis de redes (*Network analysis*) para inferir las interrelaciones entre elementos de transposición pertenecientes a la misma familia.

Palabras clave: Elementos de transposición, *Anopheles gambiae*, malaria, evolución

RESUMO

Até hoje não existem mecanismos eficientes de controle da malaria, uma das doenças infecciosas mais importantes do mundo. Nas últimas décadas, a transformação genética dos mosquitos transmissores tem sido proposta como uma alternativa para o controle de esta doença. Para isso, além do desenvolvimento de técnicas para a introdução de DNA em células germinais dos mosquitos e da identificação de genes capazes de bloquear ou reduzir a transmissão de parasitas aos humanos, é necessário o desenvolvimento de métodos eficientes para a introdução e fixação de genes refratários à malaria nas populações naturais de mosquitos. O uso de elementos de transposição tem sido sugerido para tais fins pelas características biológicas de invasão e propagação em genomas eucariotas que presentão estes elementos.

Na presente tese, analisaram-se os elementos de transposição presentes no genoma do mosquito *Anopheles gambiae*, um dos principais vetores da malaria no mundo.

Os resultados apresentados se encontram divididos em três partes, primeiramente a descrição de AnoTExcel, uma base de dados com informação detalhada dos elementos de transposição presentes neste genoma. Por outra, a partir desta base, foram identificados e caracterizados elementos novos, não detectados previamente em nenhum genoma. Por último, foram analisados elementos representativos pertencentes às diferentes classes de elementos de transposição desde uma perspectiva evolutiva e foi proposta uma análise de redes (Network analysis) para inferir as inter-relações entre elementos de transposição pertencentes à mesma família.

Palavras chave: Elementos de transposição, *Anopheles gambiae*, malaria, evolução

ABSTRACT

Up today, there are no efficient mechanisms for the control of Malaria, one of the most important diseases in the world. In the last decades, the genetic transformation of the malaria vectors have been proposed as an alternative. For achieving so, besides the development of techniques for the introduction of foreign DNA into the mosquitoes germinal cells and the identification of genes able to block or reduced the parasites transmission to human, the development of efficient methods for the introduction and fixation of the refractory genes into the mosquitoes natural populations is needed. In this regard, the use of transposable elements has been suggested as a driver system due to their biological characteristics of eukaryotic genomes invasion and propagations.

In this thesis the analysis of the transposable elements in the genome of *Anopheles gambiae*, one of the most important vectors of malaria, has been proposed.

The results are divided in three parts, first of all a description of AnoTExcel, a database with detailed information of the transposable elements present in the mosquito genome. The characterization of Novel elements that have not been described before is also described. Last, an evolutionary analysis of representative elements from different classes and families of elements has been performed. Finally, we have also proposed a network analysis for inferring the relationships between elements within the same family.

Key words: Transposabel elements, *Anopheles gambiae*, malaria, evolution

INTRODUCCIÓN

LA MALARIA

La Malaria es una de las enfermedades infecciosas más importantes del mundo, En 2006 se registraron según las estimaciones de la Organización Mundial de la Salud (OMS) aproximadamente 247 millones de casos de malaria entre 3300 millones de personas en riesgo [**Guerra et al., 2006**] produciendo como resultado casi un millón de muertes, principalmente de menores de cinco años. En 2008 había 109 países con malaria endémica [**Who, 2008**]. Factores epidemiológicos y ecológicos juegan un papel importante en la gravedad de los casos clínicos de malaria así como también en la intensidad de su transmisión.

Las manifestaciones de la malaria, así como la carga de la enfermedad en diferentes regiones del mundo dependen de diversos determinantes. Los mismos se encuentran relacionados con la dinámica de los tres agentes principales de la malaria: el hospedero intermediario (humanos), el agente etiológico (parásitos pertenecientes al genero *Plasmodia*) y el vector (mosquitos pertenecientes al genero *Anopheles*) y pueden ser clasificados en factores intrínsecos y extrínsecos [**Breman, 2001**]. Los componentes intrínsicos relacionados a las poblaciones humanas dependen de factores genéticos[1] así

[1] Una serie de polimorfismos genéticos en las poblaciones humanas tienen impacto en el curso de la infección por malaria. La anemia falciforme y las talasemias son dos condiciones que cuando están presentes en heterocigosis presentan distinto grado de protección contra la infección por plasmodios mientras que en homocigosis pueden llegar a ser fatales. Estos polimorfismos humanos se encuentran presentes en proporciones más elevadas en regiones africanas endémicas para malaria. La anemia falciforme provoca una deformación de los glóbulos rojos que impide la penetración y desarrollo posterior de los parásitos de la malaria en estas células. La "Hipótesis de la malaria" se refiere a la hipótesis levantada por J.B. Haldane en 1948 en el congreso de Genética de Estocolmo, que sugiere que la distribución geográfica de la malaria por *falciparum* y la Talasemia en regiones mediterráneas sugiere que los individuos heterocigotas para talasemia deben tener una mayor resistencia a la malaria. Para esto debemos pensar que la infección por malaria debe haber tenido un fuerte impacto en las poblaciones humanas ancestrales de manera tal de ejercer una presión selectiva para la fijación de alelos relacionados a la protección**.**

como también del grado de inmunidad a la malaria presente en los individuos de la población[2]. La especie de parasito prevalente en una determinada región es un factor importante relacionado con las características clínicas de la malaria. Por último, la longevidad de los mosquitos y su antropofília son los factores más importantes relacionados con la población de vectores.

Los factores extrínsecos incluyen factores ambientales como temperatura, periodos de lluvia y seca y humedad así como condiciones socioculturales como la pobreza, las migraciones, y el aumento poblacional, entre otros [**Breman, 2001**]. De todos los factores relacionados con la malaria, la densidad y longevidad de los mosquitos son los que juegan el papel más importante en la transmisión de la enfermedad [**Breman, 2001**]. La transmisión es proporcional a la densidad de los mosquitos, al número de picadas por persona por día así como a la probabilidad de supervivencia del mosquito. Esto último, como ya fue mencionado, es fundamental para el desarrollo de los parásitos ya que es necesario que el mosquito sobreviva el tiempo suficiente de manera de permitir que el ciclo de vida del *Plasmodium* se complete y llegue a las glándulas salivales del mosquito en su forma infecciosa.

[2] El grado de inmunidad de los individuos o de las poblaciones susceptibles a la malaria es un determinante fundamental en la respuesta clínica a la infección así como también en la transmisión de la malaria. La inmunidad contra malaria se obtiene paulatinamente y se dice que es edad-dependiente ya que los individuos obtienen mayor grado de protección a medida que se encuentran expuestos a sucesivas infecciones [**Carter & Mendis, 2002**]. El número de inoculaciones experimentadas por un individuo, así como los intervalos entre diferentes infecciones son determinantes para el status inmunológico de un individuo [**Gupta, 1999**]. Se ha propuesto el termino “inmunidad dependiente de la duración de la exposición a malaria”, en vez de ‘inmunidad-dependiente de la edad’ ya que no es la edad del individuo en si, sino la constante exposición a la enfermedad la que se relaciona con inmunidad protectora a malaria [**Carter & Mendis, 2002**].

Por su parte, la malaria puede clasificarse según su endemicidad en tres categorías diferentes. Malaria endémica estable que se produce en regiones donde los individuos se encuentran continuamente expuestos a tasas de inoculación parasitarias relativamente constantes, como es el caso de África Subsahariana. Malaria endémica inestable que se da en poblaciones expuestas casi constantemente a la malaria, pero donde existen fluctuaciones en las tasas de inoculación a humanos que pueden variar entre uno o más años. Este es el caso de algunos países del Mediterráneo, varios países del Pacífico y América Central y del Sur y el Caribe. Es particularmente significativa en términos de la gravedad de la enfermedad cuando los períodos entre las inoculaciones se hacen mayores a un año ya que la inmunidad contra la malaria se pierde rápidamente en el transcurso de ese periodo lo que provoca una morbilidad mayor. Por ultimo, Malaria epidémica que ocurre cuando se da un aumento significativo del número de casos en una población o grupo de individuos. En estos casos, cuando el agente patogénico es el *Plasmodium falciparum* puede ocasionar los casos más graves de la enfermedad. Esto pueden ocurrir en algunas regiones de África, así como en ciertas parte de Asia y América Latina [**Carter & Mendis, 2002**].

Las tasas de transmisión por su parte, pueden ser elevadas o bajas independientemente del tipo de endemicidad presente en una región determinada. En regiones de transmisión intensa los individuos adultos normalmente contraen inmunidad contra la enfermedad y sufren formas leves después de una exposición continua a lo largo del tiempo. En estas condiciones

los niños menores de cinco años y las mujeres embarazadas resultan más vulnerables a desarrollar formas severas de la enfermedad. Así, la malaria es una de las principales causas de morbilidad y mortalidad infantil y la infección de mujeres embarazadas es responsable por casos de muerte materna, de anemia severa, así como también de bajo peso al nacer y mortalidad neonatal. Por su parte, en regiones donde la transmisión es inestable a lo largo del tiempo, todos los individuos de la población se encuentran igualmente vulnerables a la infección y pueden ocurrir epidemias afectando a un gran número de individuos de diferentes edades.

Los factores mencionados anteriormente determinan en diferente grado la evolución clínica de la malaria, que puede presentarse como enfermedad asintomática (con parasitémia pero sin síntomas), como malaria clínica (caracterizada por episodios febriles de diferente grado de intensidad y frecuencia) o como malaria grave (que incluye anemia severa, síndrome neurológico y eventualmente coma y muerte) [**Breman, 2001**].

La distribución geográfica de la malaria incluye todas las regiones tropicales y algunas subtropicales del mundo. África es el continente más afectado, con más del 90% de los casos anuales. En América Latina, Brasil es el país con mayor número de casos registrados anualmente, la mayor parte de los cuales ocurren en la región de la Amazonia Legal. En ciertas regiones de África, principalmente la región subsahariana, los individuos pueden presentar varios episodios de malaria en un mismo año.

Se cree que la malaria ha acompañado a las poblaciones humanas por más de 50.000 años [**Joy et al., 2003**]. Existen evidencias que indican que existió una expansión dramática de la población parasitaria africana unos 6000 años atrás en conjunción con una serie de cambios en las poblaciones humanas que incluyen la emergencia de la agricultura y el aumento de la transmisión de malaria a humanos.

El agente etiológico de la malaria

El agente etiológico de la malaria es un protozoario perteneciente al genero *Plasmodium* del cual existen cuatro especies diferentes (*Plasmodium vivax, Plasmodium. ovale, Plasmodium malariae* y *Plasmodium falciparum*) y cuya transmisión a humanos se produce mediante la picada de un mosquito infectado del genero *Anopheles*. *Plasmodium falciparum* y en bastante menor grado *Plasmodium vivax* son las especies que tienen el mayor impacto en la salud pública mundial, provocando un grado mayor de morbilidad y mortalidad. *Plasmodium falciparum* es responsable por la mayor parte de las infecciones en el continente africano y es, a su vez, responsable por un mayor número de casos de malaria severa y mortalidad [**Gupta et al., 1994**]. *Plasmodium vivax*, por su parte es el más prevalente a nivel mundial pero si bien causa una morbilidad importante, incluso con recaídas provocadas meses después de contraída la infección inicial [**Greenwood et al., 2005**], raramente se encuentra asociado a casos graves de malaria [**Nosten, 1999**]. *Plasmodium malariae* se asocia a complicaciones renales y pacientes no tratados pueden mantener parasitémia durante años [**Elsheikha & Sheashaa, 2007**]. *Plasmodium ovale* es

el menos prevalente de los cuatro parásitos a nivel mundial, es más común en África y produce también una enfermedad caracterizada por recaídas a lo largo del tiempo.

Ciclo de vida del *Plasmodium*

El ciclo de vida de estos parásitos es sumamente complejo [**Miller et al., 2002**]. Una parte se lleva a cabo en diferentes órganos del hospedero intermediario (humanos) y otra, en el sistema digestivo de mosquitos hembras que son los hospederos definitivos (Figura 1). Existen también varias formas parasitarias (sexuadas y asexuadas) que componen el ciclo de vida, que puede ser resumido de la siguiente manera.

La picadura de un mosquito infectado inyecta en sangre *esporozoítos* presentes en las glándulas salivales del mosquito. Los *esporozoítos* migran rápidamente al hígado donde infectan hepatocitos y posteriormente, maduran a una forma parasitaria denominada *merozoíto*. Estos son liberados de las células hepáticas y pasan a circular en sangre infectando glóbulos rojos. Dentro de estas células los *merozoítos* se multiplican y lisan los glóbulos rojos liberando gran cantidad de *merozoítos* en sangre que, a su vez, infectan nuevos glóbulos rojos. Esta fase se denomina eritrocitária, ya que el parasito se encuentra dentro de los eritrocitos. Este proceso sincronizado de los *merozoítos* es responsable por la sintomatología de la malaria, caracterizada por episodios febriles y escalofríos intensos intercalados por periodos asintomáticos. La duración de los periodos asintomáticos caracteriza a la malaria provocada por diferentes especies de

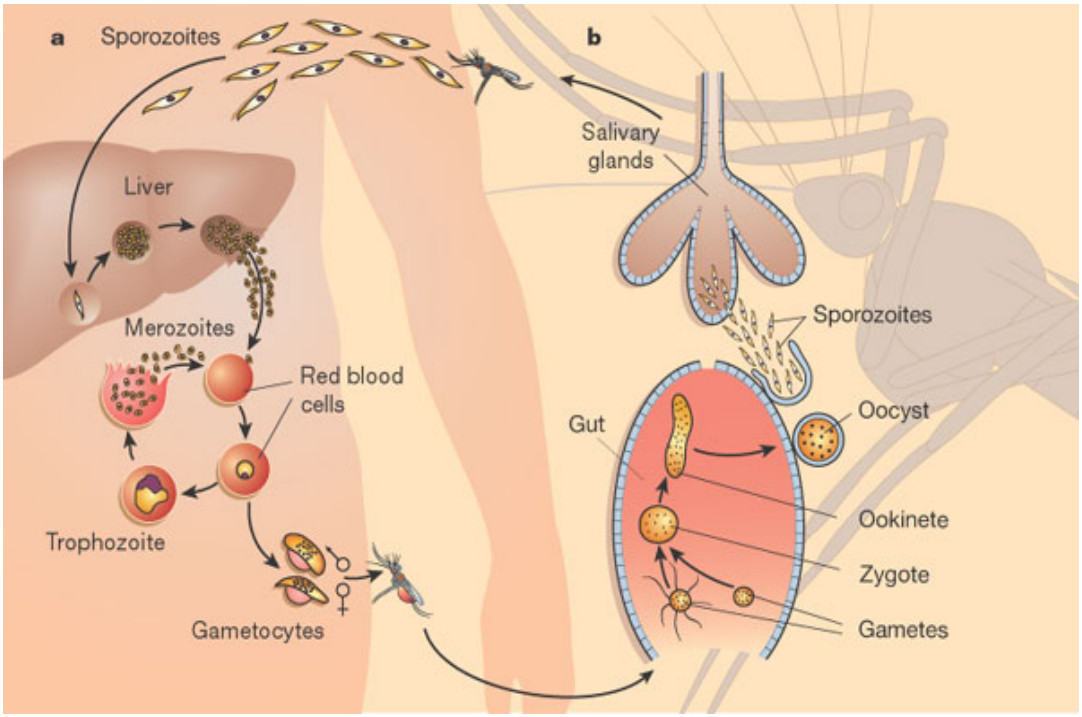


Figura 1. Ciclo de Vida de *Plasmodium falciparum*.
Esporozoitos son inyectados con la saliva del mosquito. Los parásitos migran al hígado donde se multiplican hasta romper los hepatocitos, apareciendo un nuevo estadio del parasito conocido como **merozoitos**. Estos pueden reinfectar hepatocitos o ser liberados en el torrente sanguíneo donde infectan eritrocitos. Esta es conocida como la fase eritrocitaria donde los merozoitos se convierten en **trofozoitos**. La mayoría de los merozoitos continúan en este ciclo, mientras que algunos se desarrollan en **gametocitos** masculinos y femeninos. Al ser picado por un nuevo mosquito, el individuo infectado transmite los gametocitos al mosquito. En el interior de éste los gametocitos se diferencian en **gametas** que al fusionarse producen **zigotos**. Los zigotos, a su vez, se convierten en **oocinetos** que invaden la pared intestinal del mosquito, donde se desarrollan en **ooquistes**. Los ooquistes crecen, se rompen y liberan una nueva generación de **esporozoitos**, que hacen su camino a la glándulas salivares del mosquito donde se encuentran prontos para la infección de un nuevo individuo. Adaptado de Dyann F. Wirth, 2002

plasmodios. Varía de 48 horas en las infecciones por *Plasmodium vivax* y *ovale*, casi 48 horas para *Plasmodium falciparum* y aproximadamente de 72 horas para *Plasmodium malariae*. El periodo de incubación de la malaria varía también en relación al parasito infectante pero comúnmente es de 7 a 30 días. Los

periodos más cortos suelen ser para *Plasmodium falciparum*, y los mayores para *Plasmodium malariae*.

Posteriormente, algunos *merozoítos* se diferencian en formas sexuadas o *gametocitos*. Un nuevo mosquito al alimentarse de un individuo infectado ingiere *gametocitos* (masculinos y femeninos) que después de un proceso de gametogénesis en el lumen del intestino del mosquito se convierten en *gametas*. Éstas, a su vez, al fecundarse, forman un *cigoto* que posteriormente se transforma en un *oocineto* que es la forma móvil que atraviesa la pared intestinal y se aloja en la membrana basal para diferenciarse en un *oocisto*. Algún tiempo después estos *oocistos* por medio de sucesivas meiosis se convierten en millares de *esporozoitos* que migran hacia las glándulas salivales donde se encuentran listos para infectar a un nuevo individuo.

El vector de la malaria

Mosquitos del genero *Anopheles* son responsables por la transmisión de la malaria a humanos. Existen más de 400 especies de *Anophelinos* entre las cuales entre 30 y 40 transmiten alguna de las 4 especies de plasmodios a humanos. *Anopheles gambiae* y *Anopheles funestus* son las especies más importantes en relación a la transmisión de uno de los parásitos mas agresivos de la malaria, el *Plasmodium falciparum*.

La capacidad vectorial, que es una medida que representa el potencial de transmisión de un patógeno por parte de una población de mosquitos, está relacionada con (a) la densidad de los mosquitos en relación a los hospederos

vertebrados, (b) la frecuencia con que el vector se alimenta de sangre humana, (c) la competencia del vector para permitir que el parasito finalice su ciclo de vida, (d) la duración del periodo de latencia en el vector y (e) la expectativa de vida del mosquito. La misma varía en las diferentes especies de mosquitos. *Anopheles gambiae* es un excelente vector de la malaria por tres características principales relacionadas con su capacidad vectorial, que son su alta antropofília, su competencia y su elevada longevidad, lo que permite al parasito completar el ciclo de vida dentro del mosquito y a su vez transmitirlo a un nuevo individuo.

Medidas para el control de la malaria

Las diversas formas parasitarias de los plasmodios así como la presencia de dos hospederos en el ciclo de vida hacen de la malaria una enfermedad compleja y de difícil control. Además, la emergencia y posterior propagación de resistencia de los parásitos a drogas antimaláricas y de bajo costo como la cloroquina, así como la resistencia de los mosquitos a los insecticidas utilizados en su control, son los dos factores principales que imponen un gran desafío a los esfuerzos de control de la malaria en vastas regiones del mundo. Otros factores relacionados con el agravamiento de la malaria incluyen cambios climáticos, deforestación, crecimiento poblacional relacionado a procesos de urbanización no planificados e incluso el desplazamiento rápido de los mosquitos vectores en regiones geográficas distantes por el aumento de los viajes en avión. La situación de la malaria a nivel mundial ha empeorado significativamente a partir de los años 80s

y hoy en día se presenta como uno de los problemas de salud más importante a nivel mundial [**WHO, 2008**].

Control de mosquitos vectores

A fines del siglo XIX, Ronald Ross demostró que cierto tipo de mosquitos se infectaban con el parasito de la malaria a partir de pacientes infectados. A partir de este descubrimiento se realizaron posteriormente campañas de control de vectores basadas en la protección individual contra las picaduras de mosquitos, la inspección de casas para prevenir contacto con mosquitos adultos y el tratamiento de aguas para atacar los sitios de desarrollo de larvas.

El control de mosquitos puede darse mediante el ataque a las formas larvarias o a los mosquitos adultos. Esto ultimo tiene efectos directos e importantes en la transmisión de la malaria, ya que, para que un mosquito sea capaz de transmitir malaria debe ingerir gametocítos de una persona infectada y a su vez, sobrevivir el tiempo suficiente para que los mismos se desarrollen en el mosquito y migren a las glándulas salivales [**Killeen et al., 2002**] para posteriormente infectar a un nuevo individuo. Este proceso dura un mínimo de 10 días en el caso de *Plasmodium falciparum*, por lo tanto cualquier medida que aumente la mortalidad de mosquitos adultos puede producir una reducción substancial de la transmisión de malaria. Por su parte, los larvicidas tienen un impacto menos directo en la transmisión de la malaria y al mismo tiempo los habitats larvarios son más difíciles de identificar. Aún así, programas de control centrados en el ataque a larvas se han mostrado efectivos, como por ejemplo, el programa de

erradicación de *Anopheles gambie* en la región norte del Brasil luego de su introducción accidental en 1930 y 1940 [**Killeen et al., 2002b**].

Las primeras campañas regionales de erradicación de la malaria, basadas en el control de mosquitos y en el uso de drogas antimaláricas, fueron llevadas adelante a fines de los años 40s. Posteriormente, se realizó el Programa Global de Erradicación de la Malaria de la Organización Mundial de la Salud en 1955 basado en el uso de *Dicloro-Difenil-Tricloroetano* (DDT), insecticida de acción residual altamente efectivo, así como tratamientos masivos con cloroquina (antimalárico de baja toxicidad). Gracias a estas intervenciones fue posible la eliminación de la malaria de diversas regiones del mundo, incluida Europa, América del Norte, el Caribe y partes de Asia. Lamentablemente, vastas regiones del mundo, incluyendo gran parte de los países africanos, no lograron la erradicación. Además, el uso exagerado de la cloroquina por parte de la población así como el uso limitado del DDT en estos países provocaron un agravamiento de la situación. Tanto los mosquitos como los plasmodios han generado resistencia al DDT y a la cloroquina respectivamente, dejando así, pocos medios capaces de controlar la malaria. Este panorama se ha agravado en las últimas décadas. En el año 1998, después de cuarenta años del inicio del programa global de erradicación de la malaria, la malaria vuelve a ser parte de la agenda mundial de salud con una iniciativa denominada *Roll Back Malaria* creada conjuntamente por la Organización Mundial de la Salud (OMS), la Fundación de Naciones Unidas para los niños (UNICEF) y el Programa de Desarrollo de Naciones Unidas (UNDP). El objetivo principal de esta iniciativa

era disminuir la carga de la enfermedad a nivel mundial en un 50% para el año 2010 a partir de intervenciones adaptadas a los niveles locales y de fortalecimiento del sector de salud.

Los métodos de control utilizados hoy en día incluyen el uso de insecticidas en el interior de las viviendas, el uso de mosquiteros embebidos en insecticidas, el drenaje de aguas, el control biológico basado en el uso de especies predadoras de larvas de mosquito, como larvicidas bacterianos (*Bacillus thuringiensis)* y hongos entomopatogénicos capaces de parasitar o predar mosquitos, así como también el tratamiento de pacientes con drogas antimaláricas como la cloroquina.

No obstante, estos métodos se muestran poco efectivos para solucionar el problema global de la malaria, lo que incentiva el uso de nuevas metodologías de control.

MOSQUITOS GENÉTICAMENTE MODIFICADOS COMO MECANISMO DE CONTROL DE LA MALARIA

Métodos clásicos

Métodos de control genético de insectos han sido utilizados previamente, principalmente para combatir pestes agrícolas o ganaderas. La generación y posterior liberación de insectos estériles (conocida como Técnica de Insectos Estériles -SIT por sus siglas en inglés-) producidos mediante la exposición de los mismos a radiación o agentes químicos mutagénicos es considerada la forma más simple y tradicional de control genético de insectos [**Gould &**

Schliekelman, 2004]. Esta técnica clásica de manipulación genética fue inicialmente propuesta en la década del 50 por EF Knipling [**Knipling 1965, 1968**]. El objetivo principal es causar mutaciones en las gametas de los machos de forma de imposibilitar el desarrollo de embriones viables o de producir progenie estéril una vez cruzados esos machos con hembras salvajes. Las teorías de liberación de individuos estériles se fundamentan en la liberación de machos estériles ya que en una población existe un excedente del potencial de cruzamiento de los machos [**Gould & Shliekelman, 2004**]. El principal desafío en este tipo de método es encontrar la dosis de mutágeno tal que produzca disrupción cromosómica sin afectar la capacidad reproductiva de los insectos o la viabilidad del esperma de los mismos. El paradigma de uso de esta técnica fue la total erradicación del gusano barrenador (*Cochliomyia hominivorax*) de América del Norte y Central en la década del 50 y la prevención de nuevas invasiones desde Sudamérica, mediante la liberación en periodos regulares de nuevos insectos estériles en una zona geográfica en Panamá, que se prolonga hasta los días de hoy. Esta técnica ha sido testada también para el control de otros insectos, entre ellos la mosca *Tse-tse* (vector de tripanosomiasis a humanos y ganado en algunas regiones de África) y la mosca de la fruta *Ceratitis capitata* [**Robinson, 2002a, 2002b**]. Sin embargo, el uso de este tipo de metodologías se ha mostrado ineficiente para el control de mosquitos, particularmente para *Anopheles gambiae*. A pesar de haber mostrado algunos resultados preliminares promisorios en laboratorio y a pequeña escala con *Anophelinos* [**Andreasen & Curtis, 2005**] la liberación de mosquitos estériles no

se mostró eficiente en experimentos de campo [**Benedict & Robinson, 2003**]. Problemas relacionados con los métodos físicos para la separación de los insectos machos [**Alphey & Adreasen, 2002**] además de una baja competitividad de los individuos estériles en comparación con los mosquitos salvajes [**Benedict & Robinson, 2003**], dificultan la utilización de esta metodología como método de control de la malaria. Por otra parte, los mosquitos son insectos más frágiles que las moscas u otros insectos para tolerar la radiación a la que son sometidos, ya sea como adultos o pupas. En el contexto de SIT, se han realizado curvas dosis-respuesta de irradiación de mosquitos *Anopheles*. Curtis que trabajó con *Anopheles gambiae* [**Curtis, 1978**] y Patterson con *Culex quinquefasciatus* [**Patterson, 1970**] encontraron una mortalidad elevada horas después del tratamiento de las pupas o una expectativa de vida menor en el caso de adultos de *Anopheles stephensi* irradiados [**Sharma et al., 1978**].

Otros métodos genéticos de control incluyen el uso de insectos con translocaciones cromosómicas y métodos que interfieren exclusivamente con la viabilidad de las hembras [ver **Curtis, 2006**].

Nuevas Metodologías

Mecanismos de control genético de mosquitos basados en transgénesis (técnica de introducción de un gen exógeno en un organismo vivo que produce la aparición de una nueva propiedad biológica además de la transmisión de dicha característica a la progenie) han surgido recientemente como una alternativa

complementaria a las metodologías existentes para el control de la malaria y otras enfermedades transmitidas por vectores como el dengue o la filariasis.

El avance en las técnicas de biología molecular así como el desarrollo de técnicas que permiten la rápida secuenciación y ensamblado de genomas completos han dado impulso a nuevas líneas de investigaciones en el área de control genético de insectos vectores. Entre ellas, el uso de insectos portadores de genes letales dominantes (RIDL según sus siglas en inglés que significan *Release of insects carrying a dominant lethal* gene) como alternativa a las estrategias SIT fue desarrollado en *Drosophila* [**Thomas et al., 2000**] y también ha sido aplicado a *Aedes* [**Touré, et al., 2004**]. Esta estrategia se basa en la liberación de insectos machos que portan un gen letal dominante reprimible. El gen represor puede estar relacionado a alguna sustancia ambiental no presente en la naturaleza, que puede ser utilizada durante la producción masiva de los insectos en el laboratorio -como por ejemplo, un aditivo en la dieta de los insectos o un antibiótico. Los mosquitos RIDL liberados en la naturaleza al cruzarse con hembras salvajes producen una progenie heterocigota que al no encontrarse en presencia de la sustancia que reprime la expresión del gen letal produce la muerte de la población heterocigota [**Thomas et al., 2000*; Alphey & Andreasen, 2002***].

Por otra parte, el descubrimiento de bacterias simbiontes del intestino de artrópodos con herencia materna como *Wolbachia* [**Yen & Barr, 1971**] asociadas con el fenómeno de incompatibilidad citoplasmática (producido cuando mosquitos portadores de la bacteria se cruzan con hembras no portadoras

produciendo elevada mortalidad de la progenie), atrajo la atención por su posible uso como un agente biológico de control de vectores. Estas bacterias pueden ser utilizadas como un mecanismo de control poblacional de mosquitos debido al potencial de producir una disminución poblacional, o como sistema para la diseminación de un genotipo deseado entre la población dada la desventaja de las hembras no infectadas que se crucen con machos infectados, es decir como un *driver* asociado a un transgen [**Collins & Paskewitz, 1995**].

En 1991 se llevo a cabo una reunión científica organizada por la OMS, llamada "Perspectivas del control de la malaria mediante la manipulación genética de sus vectores" ("*Prospects for malaria control by genetic manipulation of its vectors", Tucson, Arizona, EEUU)*. En ella se planteó el desarrollo de un programa de generación de mosquitos genéticamente modificados como metodología de control de la malaria [**Alphey et al., 2002**]. Este proyecto se concentró principalmente en tres áreas de investigación: (i) el desarrollo de herramientas de ingeniería genética que puedan ser utilizadas con los vectores de la malaria para la introducción de los transgenes en la población de mosquitos; (ii) la identificación de genes capaces de bloquear la transmisión de parásitos y (iii) el desarrollo de métodos eficientes para la introducción y fijación de los genes refractarios en las poblaciones naturales de mosquitos [**Alphey et al., 2002**].

Si bien, en principio, una gran parte de los mosquitos *Anopheles* son capaces de transmitir los parásitos de la malaria, se han identificado alelos de genes del sistema inmune de mosquitos que inhiben el desarrollo de los *Plasmodios* [**Osta et al., 2004**] impidiendo así su transmisión. Existen también moléculas exógenas

con acción antiparasitária en mosquitos (por ejemplo, el péptido Gomesina aislado de una araña que tiene efectos antimicrobianos [**Silva, et al., 2000**]. La introducción de este tipo de genes en la población de mosquitos mediante la ligación de los mismos a un mecanismo que permita la introducción y fijación de estos genes en la población sería una alternativa de control de la malaria.

Es poco probable que un gen de resistencia a la infección por plasmodios en mosquitos consiga fijarse en una población de vectores por si solo porque la prevalencia de infección por plasmodios en las poblaciones de mosquitos es relativamente baja [**Beier et al., 1999**] y además, porque la ventaja selectiva de un gen de resistencia claramente disminuye cuando la enfermedad (en este caso la infección por plasmodios de los moquitos) se hace menos prevalente [**Boete & Koella, 2003**]. Por lo tanto, es indispensable la utilización de un sistema de introducción y transmisión de el/los gen/es de resistencia dentro de las poblaciones naturales de mosquitos. Por su naturaleza, los TEs pueden inserirse en el genoma de gametas y así aumentar su representación en las siguientes generaciones, de esta manera genes ligados a estos elementos podrían transmitirse en la población con una frecuencia mayor que la Mendelliana [**Kidwell & Ribeiro, 1992**] convirtiéndose de esta manera en un herramienta para la introducción y fijación de un gen deseado en una población. La figura 2 muestra esquemáticamente el proceso de transmisión de elementos de transposición en comparación a alelos convencionales entre generaciones.

Uno de los sistemas de *drivers* genéticos utilizados en la generación de mosquitos transgénicos han sido los elementos de transposición. Varios

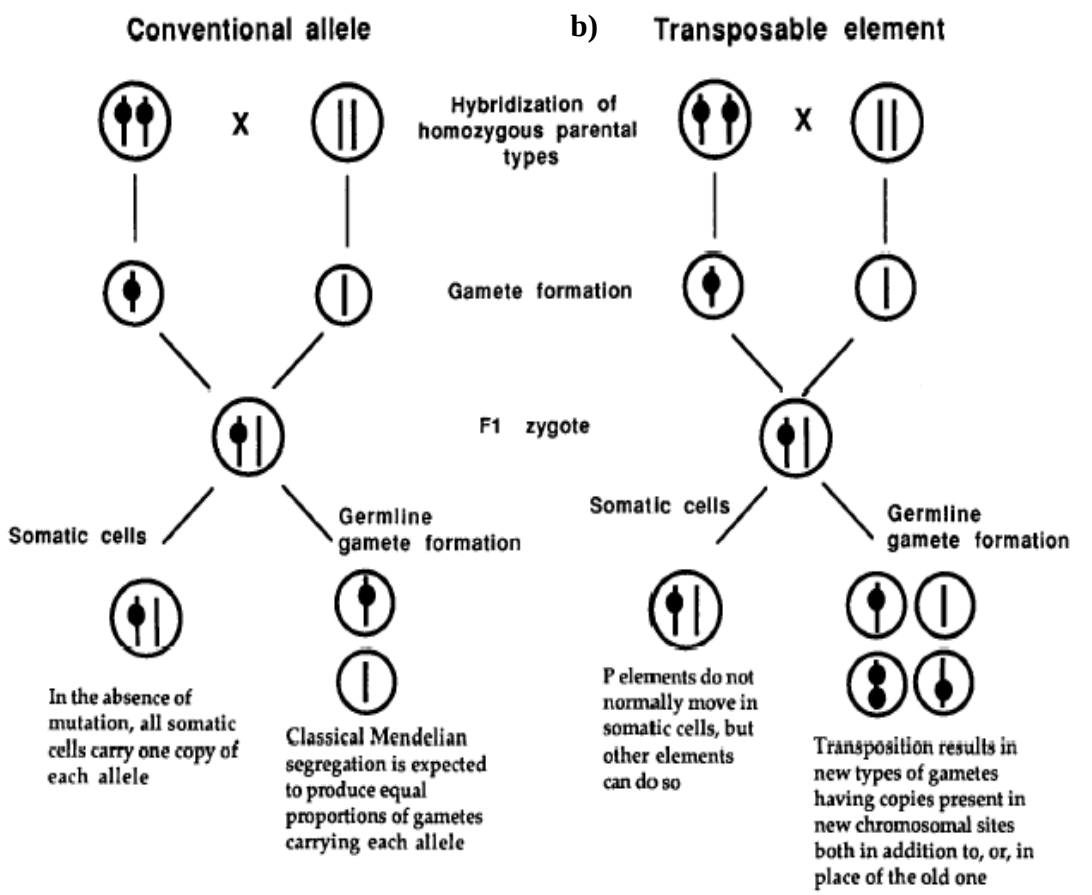


Figura 2. **Replicación de un alelo con segregación Mendeliana en comparación a un elemento de transposición.**
a) Segregación clásica de un alelo convencional. Individuos homocigotas para la presencia (+/+) o ausencia (-/-) de un determinado alelo producen gametas 100% idénticas (+) o (-) respectivamente. EL cruzamiento de estos individuos genera una progenie de individuos 100% heterocigotos cargando una copia del alelo y cuyas gametas tendrán una representación del alelo del 50%. b) Para un elemento de transposición, los individuos heterocigotos de la progenie generaran gametas en donde dicho elemento se encuentre representado en más del 50% de las mismas.
Kidwell, MG & Ribeiro JMC. 1992.Parasitology Today 8(10) 325-329

elementos han sido identificados en una variedad de insectos y utilizados en la introducción de genes en mosquitos: Hermes, Mariner, Minos y PiggyBac [**O'Brochta et al., 1996; Atkinson & Michel, 2002**; **Moreira et al., 2002**; **O'Brochta, 2003**]. El primer ejemplo de transformación de insectos con un elemento de transposición tuvo lugar en el año 1982. Rubin y Spradling [**1982**]

utilizaron como vector de transformación al elemento P[3] de *Drosophila melanogaster*.

Posteriormente, se vio que este elemento es específico de *Drosophila*, lo que limitó su uso como un vector de transformación de otros insectos. En los años 80 el elemento Hobo [**Gelbart, 1989**], también identificado en varias especies de *Drosophila*, fue utilizado como vector de transformación [**Blackman *et al*., 1989**], pero presentó poca eficiencia en la transformación de insectos fuera del genero *Drosophila*.

Posteriormente, el elemento Minos, aislado de *Drosophila melanogaster* [**Franz & Savakis, 1991**] fue utilizado en experimentos de transformación de lepidopteros [**Klinakis, 2000a**], del mosquito *Anopheles stephensi* [**Catteruccia, 2000**] y de células de mamíferos [**Klinakis, 2000b**].

La transformación genética de mosquitos de manera estable comienza en 1998 [**Coates et al., 1998**] con la transformación de *Aedes aegypti* con el elemento de transposición Mos1 de la superfamlia Tc1-Mariner. También, Jasinskiene *et al*. [**1998**] utilizaron al elemento Hermes para transformar al mismo mosquito. PiggyBac fue utilizado como vector para la transformación de varias especies de insectos [**Handler & Harrell, 1999**, **2001**; **Handler & McCombs, 2000**; **Grossman *et al*., 2000**; **Peloquin et al., 2000**; **Lobo et al., 1999; Tamura et al., 2000, Kokoza et al., 2001, Ito et al., 2002, Nolan et al., 2002, Moreira et al., 2002**]. Recientemente, la generación de mosquitos *Anopheles stephensi* transgénicos que presentaron un mayor *fitness* que insectos salvajes al ser alimentados con sangre infectada por plasmodios (*Plasmodium berghei*)

[3] Elemento de transposición de clase II descubierto en el genoma de *Drosophila.*

[**Marrelli et al., 2007**], renovó el interés por el uso de mosquitos modificados genéticamente como un medio para el control de la malaria.

ELEMENTOS DE TRANSPOSICIÓN

Definición

Los elementos de transposición (TEs), también llamados transposones, son elementos genéticos presentes en los genomas de prácticamente todos los organismos vivos. Se caracterizan por su capacidad de replicar independientemente del genoma que los alberga, por ser capaces de movilizarse dentro del genoma a diferentes localizaciones cromosómicas y por presentarse de forma repetitiva en el genoma.

Los TEs fueron descubiertos en el genoma del maíz por B. McClintock en los años 50 [**McClintock, 1950**]. En las décadas siguientes, nuevos elementos de transposición fueron descubiertos en otros organismos, incluyendo los elementos Tn en bacterias, los elementos Ty en levaduras [**Fulton et al., 1987**] y elementos P en *Drosophila* [**O'Hare & Rubin, 1983**]. En los últimos 30 años, elementos de transposición han sido identificados prácticamente en todos los genomas analizados y en la medida que nuevos genomas son secuenciados y analizados nuevos elementos continúan siendo descritos.

El interés científico que generan estos elementos ha venido en aumento desde su descubrimiento. Sea desde la perspectiva de su uso como herramientas genéticas para la transformación de diversas especies como para su utilización como agentes mutagénicos que permitan la identificación de nuevos genes.

Para los estudios que buscan comprender la dinámica de los genomas, la existencia de elementos, que no siendo funcionales para los genomas en los que habitan, se mantienen y reproducen durante largos periodos evolutivos es un tema de discusión. También, la relación de estos elementos con retrovirus, despierta gran interés en términos científicos. Las relaciones filogenéticas entre estos dos tipos de elementos es obvia, sin embargo el origen de unos y otros y su relación permanece desconocida.

La perspectiva bajo la cual los TEs han sido estudiados ha mudado desde su descubrimiento hasta el presente así como también la perspectiva bajo la cual estos elementos han sido concebidos por diferentes autores. Originalmente fueron llamados de genes controladores, ya que parecían regular la expresión de genes relacionados con la variegación en los granos del maíz. Para algunos autores, los TEs fueron considerados parásitos intragenómicos debido a su habilidad de invasión, proliferación y supervivencia de nuevos genomas durante largos periodos de tiempo evolutivo de forma independiente del genoma en el cual replican [**Doolitle & Sapienza, 1980**; **Orgel & Crick, 1980**]. Bajo esta perspectiva, los TEs son capaces de persistir en un dado genoma siempre y cuando su propia replicación no interfiera o no induzca efectos deletéreos en el genoma hospedero. Así, no son considerados como una parte constitutiva de los genomas sino elementos externos que establecen con el genoma una relación de tipo parasitaria. Se ha propuesto también un modo de vida semi-parasítico [**Hickey, 1982**] para los TEs que predice que si un TE es transmitido al 100% de la progenie podrá fijarse en la población, incluso provocando hasta un 50% de

reducción del *fitness*. Otro tipo de perspectiva para entender la naturaleza de estos elementos es que su replicación no necesariamente es perjudicial para los genomas hospederos, sino que podría tener efectos beneficiosos para el propio organismo, otorgando variabilidad genética y plasticidad además de, eventualmente, permitir la aparición de nuevas funciones en el organismo. Por ejemplo, [**McDonald, 1993**; **McFadden & Knowles, 1997**; **Kidwell & Lisch, 2000**] han detectado secuencias de elementos transponibles asociadas a genes del hospedero o a regiones regulatorias [**Makalowski, 2000**; **Nekrutenko & Li, 2001**; **Jordan et al., 2003**; **Silva et al., 2003**; **Feschotte, 2008**]. Algunos de los roles propuestos han sido el mantenimiento de la regiones cromosómica teloméricas en *Drosophila* [**Levis et al., 1993**] o su relación con el sistema de recombinación V(D)J de inmunoglobulinas en mamíferos [**Craig, 1996**; **Jones & Gellert, 2004**].

Clasificación de los elementos de transposición

El termino “elementos de transposición” se ha utilizado históricamente para referirse a dos clases bien diferenciadas de elementos basándose en la molécula intermediaria de su replicación (RNA o DNA). Los elementos de clase I, también conocidos como retrotransposones, son aquellos que tienen RNA como intermediario, y que dependen de un paso de transcripción reversa mediado por la enzima transcriptasa reversa durante su transposición. En cuanto a los mecanismos de transposición, se dice que los elementos clase I utilizan un mecanismo de “copiado y pegado” (*copy and paste*) o transposición replicativa,

ya que en el evento de transposición permanece una copia en el locus de orígen y una nueva copia aparece en otro locus. Estos elementos se clasifican a su vez en dos grupos según su mecanismo de transposición. A saber, LTRs (Long Terminal Repeats) y NLTRs (Non-LTRs). Ambos son primeramente transcriptos a mRNA, los LTRs son posteriormente retrotranscriptos a DNA e inseridos en el genoma, mientras que los NLTRs son retrotranscriptos de forma simultánea a su inserción en una nueva posición cromosómica (Figura 3a).

Los clase II se dice que tienen un mecanismo de “corte y pegado” (*cut and paste*) o transposición no replicativa, pudiendo la misma ser conservativa o no conservativa. En este tipo de mecanismo, el propio elemento que se escinde de su posición cromosómica original es inserido en una nueva localización cromosómica. Este evento puede implicar la simple movilización de la copia en cuestión o una duplicación, caso los mecanismos correctores del DNA completen el DNA faltante utilizando la hebra de DNA de la cromatida hermana (Figura 3a).

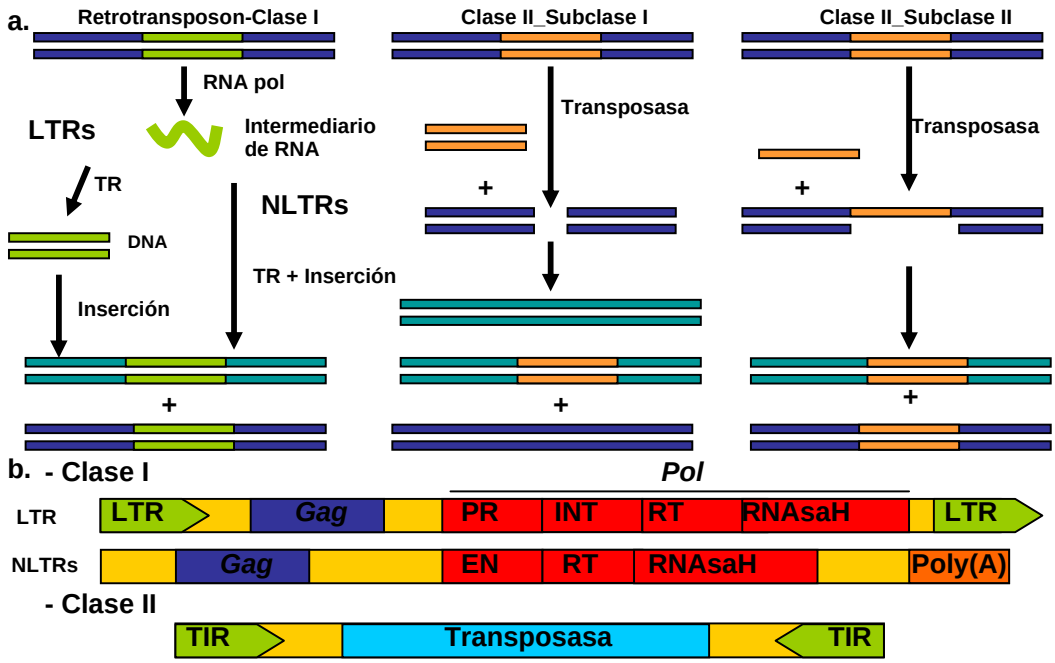


Figura 3. (a) Mecanismos de Transposición de elementos clase I y clase II y (b) Estructura Genética esquemática de los elementos de transposición. LTR=*Long Terminal Repeats*; NLTR= *Non-LTRs* Gag=Gen de nucleocápside; Pol=Gen de la polimerasa, que se encuentra conformado por los dominios: PR=Proteasa; INT=Integrasa; RT=Transcriptasa Reversa; EN=Endonucleasa; RNAsaH. TIR=*Terminal Inverted Repeats*

La diferencia entre las dos clases de elementos no solo se basa en su mecanismo de transposición sino también en su estructura genética (Figura 3b). Los elementos de clase II son genéticamente muy simples, poseen un gen que codifica para la enzima transposasa (responsable por la transposición de estos elementos) y dos regiones invertidas repetitivas que se encuentran a ambos extremos del gen de la transposasa. Los elementos de clase I son genéticamente bastante más complejos, poseen varios genes relacionados a su mecanismo de transposición, incluyendo un gen gag y un gen que codifica para la enzima transcriptasa reversa y, por lo menos, un gen de endonucleasa.

La clasificación dicotómica de los TEs ha resultado limitada para clasificar nuevos elementos de transposición que no se encuadran completamente dentro de las categorías mencionadas. Ejemplo de esto son los llamados MITE (*Miniature Inverted Repeat Transposable Elements*), que son elementos presentes en algunos genomas eucariotas cuya estructura genética consta de dos TIRs (*Terminal Inverted Repeats*) sin capacidad codificante. Presentan una altísima representación y dependen para su replicación de elementos autónomos.

A fines de 2007 surgió una nueva propuesta de clasificación de estos elementos [**Wicker et al., 2007**] que aplica criterios mecanísticos y enzimáticos para llegar a un sistema jerárquico de clasificación que incluye los niveles de clase, subclase, orden, superfamilia, familia y subfamilia (ver Tabla 1 en Anexo 1).

La nueva clasificación mantiene las dos clases (retrotransposones o clase I y transposones propiamente dicho o clase II) pero subdivide a los clase II en dos subclases. Los de clase I fueron divididos en cinco órdenes, de acuerdo con su estructura, organización y filogenia de la transcriptasa reversa (RT).

El orden LTR (*Long Terminal Repeat*), que presenta como característica principal una repetición terminal larga, sitios de duplicación y dominios codificantes de proteinas gag y pol, fue subdividido en cinco superfamilias: Copia, Gypsys, *Bel-Pao*, *Retrovirus* y *ERV*. El orden *DIRS*, cuyos componentes no generan sitios de duplicación en el local de inserción, se encuentra representado por las superfamílias *DIRS*, *Ngaro* y *VIPER*. El orden *PLE* (*Penelope*-like elements) contiene elementos que codifican una RT, contenidos

solamente en una superfamília: *Penelope*. Los ordenes *LINE* (Long Interspersed Nuclear Elements) o también llamados NLTRs (Non-LTR) y *SINE* (Short Interspersed Nuclear Elements), no presentan *LTR*s. Los *NLTRs* se encuentran agrupados en cinco superfamílias (*R2*, *RTE*, *Jockey*, *L1* e *I*) y los *SINE*s en tres (*tRNA*, *7SL RNA* y *5sRNA*). Las dos subclases de la clase II difieren en el número de hebras de DNA cortadas en el momento de la excisión Tabla 1 en Anexo 1.

En la subclase 1, las dos hebras son cortadas y en la 2 solamente una lo es. La subclase 1 está compuesta por el orden *TIR* (Terminal Inverted Repeat), que presenta sitio de duplicación en el lugar de inserción y está compuesta por nueve superfamílias (*Tc1-Mariner, hAT, Mutator, Merlin, Transib, P, PiggyBac, PIF – Harbinger, CACTA*) y, por el orden *Crypton*, con una superfamília (*Crypton*), cuyos transposones no poseen *TIR*s. La subclase 2 está representada por dos ordenes, con una superfamília cada uno: Helitron (*Helitron*) y Maverick (*Maverick*). Los *Helitrones* no generan sitio de duplicación y presentan un *hairpin* en la extremidad 3', y los *Maverick*, poseen *TIRs* largos.

Abundancia de los elementos de transposición

Existe una enorme diferencia en la proporción del genoma "ocupado" por TEs en diferentes especies. En un extremo tenemos el ejemplo del maíz cuyo genoma está compuesto en más de un 70% por TEs [**Meyers et al., 2001**], en el otro *Sacharomyces cerevisiae* o *Drosophila melanogaster* con aproximadamente 3% del genoma ocupado por TEs [**Kim et al., 1998, Kaminker et al., 2002**]. Por su

parte, existe también variabilidad en cuanto a la representación relativa de cada clase de elemento en diversos genomas. Así, por ejemplo, el genoma humano contiene aproximadamente 3 millones de copias de elementos de transposición, equivalente a 46% del mismo [**Lander et al, 2001**] de los cuales 3/4 partes corresponden a elementos NLTRs. Ya, el genoma del arroz se encuentra altamente poblado por elementos de clase II [**Jiang et al., 2004**]. En el caso del mosquito *Anopheles gambiae,* que posee 16% del genoma ocupado por TEs [**Holt et al., 2002**], la proporción de las diferentes clases de elementos presentes en el genoma es similar. Las diferencias en términos de abundancia de elementos transponibles así como de las proporciones de las diferentes clases de elementos en algunos genomas continua siendo tema de debate (para una discusión más completa sobre estos temas ver **Brookfield, 2005**; **Feschotte, 2004**; **Kidwell, 2002**; **Zhang & Wessler, 2004**). Por un lado, puede pensarse que algunas familias de elementos podrían tener efectos de tipo fundador en un determinado genoma, es decir que ciertos elementos podrían "colonizar" un genoma y de alguna manera inhibir la reproducción de otras clases de elementos. Alternativamente, puede pensarse que ciertos genomas podrían ser más o menos permisivos a la amplificación de transposones. Por ejemplo, puede constatarse que la diferencia entre el tamaño del genoma de *Aedes aegypti* (1,3Gb) y *Anopheles gambiae* (245Mb) se debe a una enorme diferencia en la abundancia de elementos de transposición.

En el genoma de *Anopheles gambiae* se han identificado previamente cuarenta tipos diferentes de elementos de transposición, siendo los clase I los

más abundantes, aunque también se han identificado elementos pertenecientes a las principales familias de la clase II [**Holt et al., 2002**].

Antecedentes del uso de TEs en la transformación genética de mosquitos

La utilidad de los elementos de transposición como *drivers* genéticos está determinada, entre otras características, por el potencial de invasión del elemento en cuestión en la población de organismos a ser transformada así como por el mantenimiento del gen refractario como parte de la estructura del elemento de transposición. Esto es, si en pocos ciclos de vida o en pocas generaciones un transposón utilizado en un experimento de transformación pierde su carga (gen refractario de interés), el éxito de un programa de control genético sería escaso. Se requiere, entonces, de un fuerte vínculo entre el elemento de transposición y el gen refractario. En este respecto, cabe señalar que un mecanismo que podría llevar a la perdida de vinculo entre el transposón y su carga son los procesos de recombinación [**Riehle et al, 2003, Scott et al., 2002**]. Para el caso particular de los elementos de transposición, una preocupación adicional es la aparición de deleciones internas en los elementos, lo que eventualmente podría acabar con la integridad del gen de resistencia, inviabilizando también el éxito de un programa de control genético de mosquitos.

El presente trabajo se inició con la idea de detectar y caracterizar elementos de transposición nuevos en el genoma de *Anopheles gambiae* bajo la perspectiva de su uso como potenciales *drivers* genéticos para la generación de mosquitos transgénicos. Inicialmente se creó una base de datos que pretendía presentar

los elementos de transposición encontrados en este genoma e incorporar una caracterización de los mismos. Se identificaron y caracterizaron también elementos nuevos, no descritos previamente. En el proceso de análisis de los elementos identificados en el genoma observamos la existencia de una variedad de “formas” deterioradas de estos elementos, lo que llevó a la pregunta que se pretende explorar aquí y que está relacionada con el proceso de deterioración de los diferentes elementos en el genoma de uno de los principales vectores de la malaria. ¿De qué manera se relacionan los TE completos, potencialmente activos, con elementos deteriorados pertenecientes a la misma familia? ¿Cuáles son las dinámicas de los TEs dentro de las diferentes clases y familias? Estas son algunas de las preguntas que esta tesis pretende abordar.

METODOLOGÍA

GENERACIÓN DE ANOTEXCEL

- Detección de elementos Repetitivos en el genoma de *Anopheles gambiae*.

El programa PILER (del inglés *Parsimonious Inference of a Library of Elementary Repeats*) [**Edgar & Myers, 2005**] fue utilizado para la identificación de secuencias repetitivas dentro del genoma de *Anopheles gambiae* (versión 3.7, Febrero/2006). Este programa busca todas las secuencias repetitivas presentes en un genoma determinado. Para ello, se basa en auto-alineamientos generados por el algoritmo PALS (del inglés *Pairwise Aligner for Long Sequences*) que se utiliza para hallar alineamientos locales de un genoma sobre sí mismo. Se utilizó el método de búsqueda llamado DF que busca familias de elementos repetitivos dispersos en el genoma, lo que mayoritariamente corresponde a elementos de transposición, aunque también puede detectar otro tipo de familias repetitivas en el genoma como exones parálogos y microsatélites, por ejemplo. Este programa, bajo la configuración utilizada, busca familias de 3 o más secuencias repetitivas globalmente alineables, de más de 400 nucleótidos. Una vez detectadas las secuencias repetitivas en el genoma, el programa asigna estas secuencias a familias de secuencias similares que sean globalmente alineables entre ellas. Posteriormente genera una biblioteca conteniendo las secuencias consenso correspondientes a cada familia. En el caso que dos familias sean localmente alineables en una determinada región de su secuencia, ambas familias serán asignadas a la misma Superfamilia.

Con el objetivo de agrupar todas aquellas familias que potencialmente correspondiesen a la misma familia de elementos de transposición, a partir del archivo con todas las secuencias individuales identificadas por PILER, se realizó un agrupamiento de las mismas considerando en un mismo cluster todas las secuencias con más de 90% de identidad en más de 90% del largo de la secuencia. Una vez generados estos clusters fueron denominados numéricamente en forma decreciente (menor número asignado al cluster con mayor número de secuencias).

- Alineamientos Múltiples

Todas las secuencias dentro del mismo cluster fueron subsecuentemente alineadas utilizando un programa de alineamiento denominado MUSCLER (desarrollado por el Dr. J.M. Ribeiro) que alinea secuencias progresivamente según su tamaño utilizando el programa MUSCLE [**Edgar, 2004**]. Se utilizó este algoritmo debido a la dificultad de alinear secuencias de tamaños muy diferentes y, en algunos casos, pertenecientes a regiones diferentes del mismo elemento y a la dificultad para generar alineamientos óptimos de programas como CLUSTAL [**Higgins & Sharp, 1988**] o el propio MUSCLE. El programa MUSCLER utiliza al programa MUSCLE para ejecutar un alineamiento de múltiples secuencias utilizando los siguientes parámetros y penalidades: center= -1; gapopen= -500 y gapextend= -50. Un análisis visual de todos los alineamientos fue realizado con la ayuda del programa MEGA1.4 [**Kumar et al., 2004**].

- Secuencias Consenso y Centroide

Se obtuvieron secuencias consenso a partir de los alineamientos múltiples obtenidos para cada cluster. En la generación de la secuencia consenso se consideró, para cada posición del alineamiento, el nucleótido que apareciera en más de 50% de las secuencias independientemente del número de secuencias totales que estuviesen representadas en esa posición nucleotídica. Aquellas posiciones en donde menos del 50% de las secuencias se encuentran representadas se indican mediante la utilización de letras pequeñas en la secuencia consenso. De esta manera nos aseguramos de obtener la mayor secuencia consenso posible.

Con el objeto de detectar clusters que pudieran pertenecer a la misma familia de TE pero con diferente grado de deterioración, las secuencias consenso de cada cluster fueron a su vez agrupadas en nuevos clusters con diferente grado de identidad (35, 50,75 y 90%) sobre más del 50% del largo de las secuencias.

También se calculó la secuencias centroide para cada cluster y se comparó con la secuencia consenso con el objetivo de determinar cuán representativa de un determinado cluster era la secuencia consenso obtenida. La secuencia centroide de cada cluster fue escogida como aquella que obtuvo el mayor valor al sumar todos los valores obtenidos después de realizar un *blast* de todas las secuencias contra todas las secuencias de cada cluster. Como una forma de corroborar la representatividad de la secuencia consenso de cada cluster (utilizada posteriormente en varios análisis realizados), se comparó la misma con la

secuencia centroide por medio de un *blast* de una contra la otra. El tamaño de ambas secuencias fue también comparado.

- Detección de Marcos de Lectura Abiertos (ORFs = Open Reading Frames)

La presencia de ORFs fue deducida a partir de los seis marcos de lectura posibles correspondiente a las secuencias consenso utilizando el programa de dominio público "*ORF Finder*" en su versión online (http://www.ncbi.nlm.nih.gov/gorf/gorf.html).

- Detección de Dominios Proteicos Conservados

La presencia de dominios proteicos conservados en los ORFs detectados fue deducida utilizando el programa de dominio público "*Conserved Domains*" en su versión online (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi).

- Identificación de características particulares de TE: Repeticiones Invertidas (TIR), Repeticiones Directas (LTR) y secuencias palindrómicas

Un *blast* de cada secuencia dentro de cada cluster contra (a) si misma y (b) su secuencia reversa y complementaria, se realizó mediante el programa "*Blast-2-Sequences*" [**Tatusova and Madden, 1999**] para la detección de TIRs, LTRs y secuencias palindrómicas. Las secuencias invertidas repetidas fueron clasificadas como terminales si se encontraban presentes en los primeros 10% y en los últimos 90% nucleótidos de cada secuencia, caso contrario, fueron consideradas como repeticiones sub-terminales.

El programa LTR-FINDER [**McCarthy & McDonald, 2003**] fue también utilizado para ayudar con la identificación de elementos LTRs nuevos en los datos analizados.

***- Blast* (*Basic Local Alignment Search Tool*)**

El programa *Blast*n fue utilizado para comparar cada una de las secuencias consenso con las siguientes bases de datos: (1) Tefam (Tf) (http://tefam.biochem.vt.edu/tefam/); (2) REPBASE (RB) (http://www.girinst.org/repbase/update/index.html) [**Jurka et al., 2000 y 2005**]; (3) la colección de elementos repetitivos del genoma de *Anopheles gambiae* descrita por Smith et al, 2007; (4) la base de datos del *Gene Ontology* (http://www.geneontology.org/) [**Ashburner et al., 2000**]; (5) la base de datos no repetitiva de proteínas del *Genbank*; (6) la base de datos no repetitiva de nucleótidos del NCBI (National Center for Biotechnology Information); (7) un set de 180.000 ESTs (Expressed Sequence Tags) de *Anopheles gambiae*; (8) una base de datos de ESTs individuales; (9) la base de datos *Pfam* [**Finn et al., 2006**]; (10) una base de datos generada con secuencias proteícas del gen *gag*; (11) una base de datos generada con secuencias proteícas de genes de transposasas y (12) una base de datos generadas con fragmentos genómicos de 50Kb con una sobreposición de 10Kb en cada extremo.

La herramienta t*blast*x fue también utilizada con las bases específicas de elementos de transposición Tefam y Repbase con el objetivo de facilitar la identificación de elementos altamente divergentes o deteriorados. Todos los hits

con valores-E menores a 10^{-15} fueron considerados positivos para el propósito de anotación.

- Distribución cromosómica

Se analizó también la proporción de elementos repetitivos presentes en cada cromosoma. Para ello, todas las secuencias obtenidas con el programa PILER fueron *blast*eadas contra el genoma de *Anopheles* usando un tamaño de palabra de 100 nucleótidos en el programa *Blast*N (por lo tanto, solo fueron considerados aquellos *matches* que tuviesen por los menos 100 nucleótidos consecutivos con un *match* perfecto). Las coordenadas cromosómicas fueron identificadas.

- Anotación de secuencias

Los clusters fueron manualmente anotados y organizados según correspondiesen a elementos de transposición en CLASE I (LTRs y NLTRs), CLASE II y NESTED (elementos ANIDADOS, caracterizados como conjuntos de secuencias con alta identidad que se encuentran conformadas por diferentes elementos o fragmentos de elementos de transposición pertenecientes a diferentes familias o clases), o a repeticiones sin características de TEs. Para ello se consideró la similitud de las secuencias consenso contra TEs depositados en las bases específicas de TEs (Tf y RB) así como la presencia de características particulares de elementos de transposición como la presencia de LTRs, TIRs, o similitud con proteínas específicas de origen retroviral o

transposasas. El programa "*Repeat Masker*" (RM) en su versión online [**Kohany et al., 2006**] fue ejecutado para aquellos clusters que fueron clasificados como TEs (http://www.girinst.org/censor/index.php).

La clasificación de los elementos de transposición se basó en los criterios sugeridos por Wicker et al, 2007 para elementos de transposición en Clase, Subclase, Orden, Superfamilia y Familia. Además de considerar también si se trataba de elementos completos o fragmentos de elementos.

- Construcción de la base de datos

La base de datos denominada AnoTExcel (ideada y generada por el Dr. J.M. Ribeiro) contiene toda la información colectada para cada familia de elemento repetitivo, se presenta en la forma de una planilla Excel conteniendo células con *hyperlinks* a los archivos con los diferentes resultados obtenidos (alineamientos nucleotídicos, resultados de *Blast*, secuencias consenso, etc). Las diferentes líneas de esta planilla (totalizando 245) corresponden a cada uno de los clusters que está compuesto por un número diferente de secuencias (los menores con 3 secuencias y los más numerosos con 250). La información fue organizada en tres bloques principales conteniendo información relacionada con (1) la identidad de los elementos repetitivos (títulos presentados en azul); (2) las características estructurales de los mismos (títulos en verde) y (3) la homología de las secuencias con otras secuencias depositadas en diferentes bases de datos (títulos en rojo).

El análisis estructural de las secuencias consenso incluyó el análisis de la presencia de ORFs –el tamaño del ORF mayor, la secuencia aminoacídica del mismo y el marco de lectura en el que se encuentran están incluidas como columnas en AnoTEXcel. También, cuatro columnas de la planilla incluyen el tamaño de las secuencias mayor, menor y consenso así como el cociente entre el tamaño de la secuencia mayor y menor en cada cluster. Se realizó un análisis de las deleciones presentes en los extremos de las secuencias de cada cluster. Las deleciones presentes en los extremos 3' y 5' fueron calculadas y fue creada una descripción gráfica de las mismas. Los resultados se presentan en una columna denominada "*Mean average length of sequence with gaps and link to truncation analysis*" que muestra una representación del alineamiento con gaps como un archivo .txt.

También se presentan como columnas de AnoTExcel: (i) la distribución cromosómica de las repeticiones (con columnas presentando la proporción de elementos repetitivos en cada cromosoma); (ii) la fracción de secuencias dentro de cada cluster con secuencias invertidas terminales, sub-terminales y palindrómicas; (iii) los alineamientos múltiples de las secuencias de cada cluster. Por otra parte, los resultados de los *blast* contra las diferentes bases de datos se presentan como columnas diferentes. Varias informaciones relacionadas con los *blast*s se presentan en diferentes columnas, como por ejemplo, (1) el valor-E, (2) la extensión del match, (3) % de identidad, etc.

ANÁLISIS EVOLUTIVOS

- Análisis Filogenético

Se utilizó el método de *Neighbor-Joining* [**Saitou & Nei, 1987**] para inferir la historia evolutiva de las diferentes secuencias analizadas, sea considerando las secuencias aminoacídicas o nucleotídicas (según el caso), implementado en el programa MEGA4 [**Tamura et al, 2007**]. Las distancias evolutivas que fueron utilizadas para inferir la filogenia fueron calculadas utilizando el método de Corrección de Poisson para los alineamientos aminoacídicos y de Máxima Verosimilitud Compuesto (*Maximum Composite Likelihood*) en el caso de los alineamientos nucleotídicos. Todas las posiciones conteniendo *gaps* fueron eliminadas del análisis (método de deleción completa). El porcentaje de los árboles en los cuales los diferentes taxa se presentan juntos en el test de *bootstraping* utilizando 1000 remuestreos, se presenta en las ramas de los árboles. Las distancias-p fueron también calculadas utilizando del programa MEGA4.

Detección de la Presencia de Selección

Con el objetivo de detectar la presencia de selección actuando en las secuencias que conforman cada cluster se realizó el test de Tajima y la relación entre sustituciones sinónimas (dS) y no sinónimas (dN).

- Test de Tajima

Para establecer si las secuencias dentro de cada cluster evolucionan bajo un proceso neutro o no fue utilizado el test de Tajima que fue ejecutado en los alineamientos nucleotídicos correspondientes a las secuencias completas para cada cluster analizado. El mismo fue implementado en el programa DNAsp4.50.3 [**Rozas et al., 2003**].

- dN/dS

También se analizaron la relación entre las mutaciones en posiciones sinónimas y no sinónimas para aquellos clusters conteniendo secuencias completas, como un indicador de la existencia de presiones de selección operando a nivel de las regiones codificantes halladas. La relación entre el número de substituciones sinónimas por sitio sinónimo y de substituciones no sinónimas por sitio no sinónimo fue calculada utilizando el método de Kumar [**Nei & Kumar, 2000**], según se encuentra implementado en MEGA4.

- Análisis de redes (Network Analysis)

Se utilizó un análisis de redes basado en *Median-Joining Networks* [**Bandelt et al., 1999**] para inferir las relaciones y los padrones de expansión de algunas familias de elementos de transposición presentes en AnoTExcel. Se utilizó el programa NETWORK versión 4.5.1.0 para realizar los análisis de *Median Joining*.

- Deterioración Estructural

Para analizar si existen patrones diferentes de deleciones en los clusters pertenecientes a diferentes clases y familias se utilizó un programa (diseñado en lenguaje *Pearl* por Felipe Figueredo). El programa analiza las posiciones de las deleciones presentes en las diferentes secuencias de cada cluster a partir del alineamiento global de las secuencias. Para cada posición del alineamiento contabiliza la proporción de secuencias que tienen un nucleótido (cualquiera) o un gap, para después graficar la posición del alineamiento versus la proporción de secuencias presentando deleciones en esa posición.

RESULTADOS & DISCUSIÓN

ANOTEXCEL: BASE DE DATOS DE ELEMENTOS DE TRANSPOSICÓN EN EL GENOMA DE *Anopheles gambiae.*

Un aspecto importante en el estudio de los elementos de transposición (TEs) en un genoma es el conocimiento de la diversidad de los mismos en el genoma, así como la clasificación y organización de los elementos ya identificados en una determinada especie. A partir de la década de los 90 y en un esfuerzo conjunto para organizar y presentar elementos de transposición identificados en diversos genomas eucariotas, se materializó la base de datos de elementos repetitivos más completa existente hasta la fecha, considerada una colección de referencia de TEs [**Feschotte, 2008**], denominada Repbase (RB) [**Jurka, 2000**] y disponible *online* en la siguiente dirección electrónica: (http://www.girinst.org/repbase/update/browse.php). Esta base consta de secuencias prototípicas de DNA repetitivo generadas a partir de secuencias obtenidas en diferentes estudios y presenta además información sobre nomenclatura, clasificación y detalles biológicos de los elementos allí presentes. Esta base crece constantemente mediante la incorporación de nuevos elementos y es considerada una colección de referencia tanto para objetivos de anotación, como de enmascaramiento de elementos en proyectos genómicos. Allí es donde se encuentran depositados la mayoría de los TEs identificados en organismos eucariotas.

Una base más específica de TEs en los genomas de los mosquitos *Anopheles gambiae* y *Aedes aegypti* es la base denominada TEfam (Tf). Esta base,

también disponible *online* (http://tefam.biochem.vt.edu/tefam/), presenta secuencias consenso de algunos elementos identificados en los genomas de los mosquitos mencionados. La misma, si bien no es completa, tiene algunos elementos que no han sido depositados en RB, principalmente pertenecientes a la clase II y NLTRs. El principal objetivo de Tf es proveer información para la sistematización y caracterización, a nivel genómico, de elementos de transposición en los genomas de los dos mosquitos mencionados.

Ambas bases presentan secuencias consensos y no disponibilizan las secuencias a partir de las cuales los consensos fueron generados. Así, el material allí presente sirve fundamentalmente de base para la obtención de secuencias de referencia para estudios comparativos u otro tipo de estudios con TEs pero no permiten analizar las diferentes secuencias de elementos presentes en una misma familia.

En el presente trabajo, se utilizó una estrategia combinada de identificación de TEs presentes en el genoma del mosquito (Figura 4), que incluyó el uso del programa PILER-DF (*Parsimonious Inference of a Library of Elementary Repeats*) como primer método de *screening* y posteriormente una caracterización detallada de cada una de las familias de elementos identificadas por el programa. Además, se analizaron características estructurales de los TEs identificados así como también el grado de identidad de las secuencias identificadas con secuencias previamente depositadas en una serie de bases de datos biológicas, incluyendo bases específicas de elementos de transposición. Toda la información colectada se presenta como una base de datos, denominada AnoTExcel y descripta detalladamente a continuación.

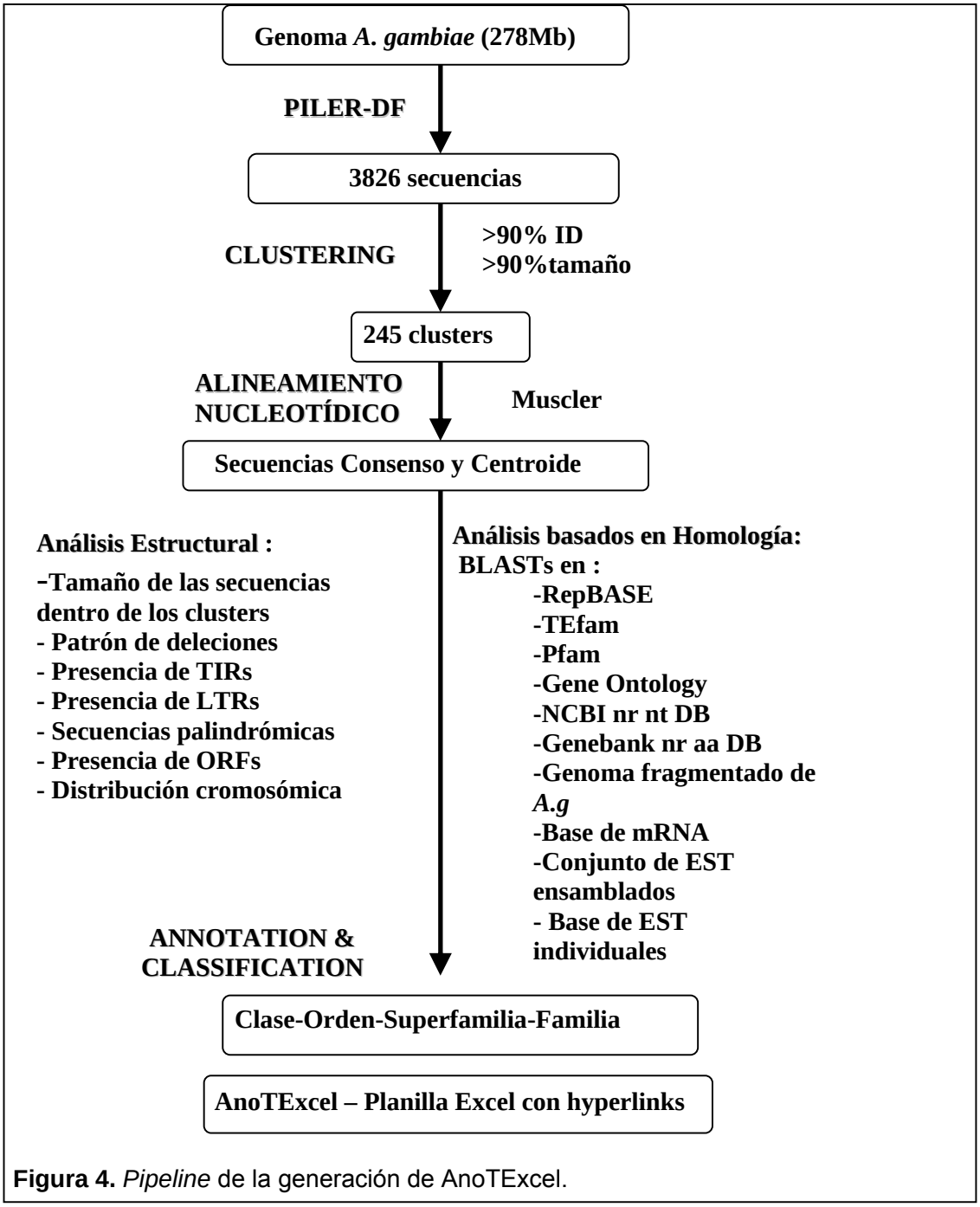


Figura 4. *Pipeline* de la generación de AnoTExcel.

Presentación de AnoTExcel

AnoTExcel es una base de datos de elementos repetitivos presentes en el genoma de *Anopheles gambiae,* producto de la búsqueda y organización de secuencias de TEs en este genoma. AnoTExcel se presenta en forma de planilla

Excel, donde cada célula contiene, en formato de *hyperlink,* archivos con los resultados obtenidos en cada uno de los análisis realizados para cada familia de elementos caracterizada (Ver CD en anexo). Esta forma de presentación facilita enormemente la visualización de los análisis realizados en la generación de la propia base, así como la utilización de este material por otros investigadores.

Cada línea de la planilla corresponde a un cluster que contiene un conjunto de secuencias con alto grado de identidad (>90%), correspondiendo en muchos casos a una familia de elementos de transposición (siguiendo los criterios de clasificación sugeridos por **Wicker et al, 2007** y presentados resumidamente en la Tabla 1 en el Anexo 1). Cada columna de AnoTExcel presenta información obtenida a partir de los diferentes análisis realizados, la mayoría de las veces utilizando a las secuencias consenso (*hyperlink* en la Columna A de AnoTExcel) generadas para cada cluster como base de análisis y de búsqueda. AnoTExcel puede ser copiada y mantenida en una computadora personal. Ocupa aproximadamente 178 Mb de espacio físico (planilla más links) y puede modificarse a voluntad, agregando resultados de nuevos estudios, o retirando columnas o links que no sean de interés del investigador.

En AnoTExcel se encuentran depositadas 3826 secuencias repetitivas dispersas que fueron subsecuentemente agrupados en 245 clusters. Estas secuencias son consideradas intactas –por ser globalmente alineables entre ellas- y aisladas porque se encuentran delimitadas por secuencias únicas, o sea, corresponden a eventos de inserción de secuencias repetitivas diferentes. El programa PILER clasifica a las secuencias repetidas obtenidas en grupos de *familias* y *superfamilias* de secuencias. Las familias son consideradas conjuntos de

instancias del mismo elemento y todas las secuencias dentro de cada grupo tienen el mismo tamaño. Por otra parte, las superfamilias, según PILER, incluyen diversas familias que comparten homología en diferentes regiones y cuyas secuencias no tienen necesariamente el mismo tamaño, ni el mismo grado de identidad, ni corresponden a la misma región. Las diferentes secuencias dentro de una determinada familia según Piler son numeradas siguiendo el siguiente criterio: TRS_Número de Superfamilias_Número de familia. De esta manera secuencias que pertenecen a una misma Superfamilia, pero a familias diferentes compartirán el mismo primer número y diferirán en el segundo. Estas numeraciones pueden observarse en las secuencias de los alineamientos presentados en AnoTExcel en la columna denominada *Muscler Output* (columna AO). Los clusters en AnoTExcel se encuentran identificados con números consecutivos del 1 al 245, ordenados de forma decreciente, según el número de secuencias que contienen.

La estrategia utilizada en la generación de AnoTExcel permitió agrupar, dentro de un mismo cluster, a grupos de secuencias que componen familias y superfamilias de elementos de transposición según la nomenclatura del programa PILER-DF, es decir, secuencias pertenecientes a la misma familia de TE pero de tamaños diferentes o pertenecientes a regiones diferentes del mismo elemento. Y, por lo tanto, con distinto grado y/o tipo de deterioración[1].

[1] Cabe señalar que la nomenclatura que utiliza el programa PILER de *familias* y *superfamilias* es igual a la nomenclatura sugerida para clasificar elementos de Transposición por Wicker et al, 2007 y seguida en esta tesis. Las llamadas Superfamilias en el programa PILER, pueden corresponder a Familias de elementos de transposición, y las denominadas Familias según PILER pueden representar sub-familias en la nomenclatura de transposones. Las nominaciones familias y superfamilias tal cual utilizadas por Piler no son utilizadas en ningún momento en el

AnoTExcel no se presenta como una base de datos exhaustiva de los elementos de transposición presentes en el genoma de *Anopheles gambiae* ya que no contiene toda la diversidad de elementos previamente descripta en este genoma. De cualquier manera, fueron identificados elementos pertenecientes a las dos clases y a las principales superfamilias y familias de TEs descriptas previamente en el mosquito (Figura 5). Presenta como principal ventaja y como diferencial en relación a las otras bases, el hecho de mostrar la totalidad de las secuencias pertenecientes a cada familia de elementos repetitivos identificada. Además, presenta los alineamientos de las secuencias de cada familia en formato fastA, las secuencias consenso y centroides para cada familia, detalles de sus características estructurales (secuencias y alineamientos de TIRs, LTRs) y resultados de Blasts realizados en diferentes bases de datos. Todo esto compone un material rico y novedoso que puede utilizarse para diversos fines según el objetivo que se persiga.

Los elementos de transposición identificados en AnoTExcel fueron clasificados en: clase, subclase, orden, superfamilia, familia (siguiendo los criterios de **Wicker et al, 2007**) y tipo (que no es una categoría de clasificación taxonómica sino que considera las características estructurales, es decir, si las familias corresponden a elementos completos o fragmentados o, en casos más específicos si corresponden a Solo-LTRs, MITEs, ClaseII-NA, etc).

presente texto y son únicamente utilizadas para referirse a la clasificación de elementos de transposición.

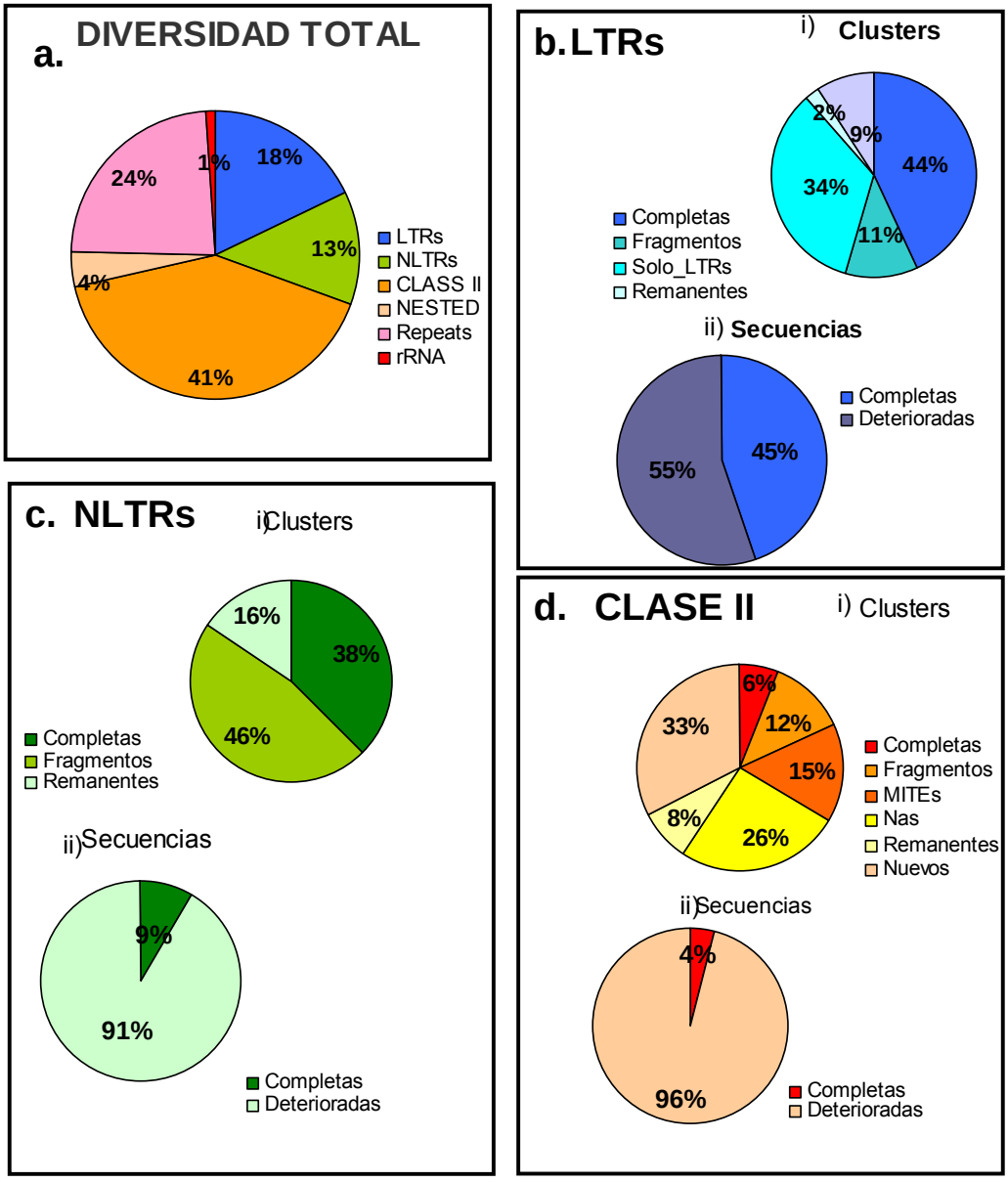


Figura 5. Diversidad de los elementos repetitivos en AnoTExcel a. Diversidad (representada como el número de clusters) de los elementos repetitivos identificados en AnoTExcel. **b.** Para elementos **LTRs (i)** Proporción de **clusters** cuyas secuencias consenso son completas o que presentan diferente grado de deterioración clasificadas como **Completas:** aquellas que contienen el 100% del tamaño de los elementos canónicos, incluyendo región interna y LTRs; **Fragmentos:** corresponden a secuencias que representan una región de un elemento, **SOLO-LTR:** secuencias correspondientes a uno o ambos LTRs sin región interna, **Remanentes:** fragmentos de elementos completos con identidad muy baja (menos de 70%) y **Nuevos:** corresponden a elementos descriptos por primera vez en AnoTExcel. **(ii)** Proporción de **secuencias** completas o con algún grado de deterioración. **c.** idem b. para elementos **NLTRs d.** idem b. para elementos **clase II**. MITEs: Miniature Inverted Terminal Repeats, Nas: elementos no autónomos ya caracterizados en RB

Un valor-E menor a 10^{-15} fue considerado significativo para fines de clasificación. Todo el proceso de anotación y clasificación de los clusters en AnoTExcel se llevó a cabo manualmente, considerando aquellas características singulares de TEs (como la presencia de LTRs, TIRs, o *matches* positivos con enzimas específicas de elementos de transposición) además de los *matches* obtenidos en los *BLASTs* realizados en las diversas bases de datos. El primer criterio considerado en todos los casos fue un *match* positivo y significativo con elementos de transposición ya caracterizados y depositados en Tf o RB mediante *blastN* (detección a nivel de la familia de TE) o *TblastX* (detección a nivel de la superfamilia o el orden del TE).

La metodología utilizada para la detección de TEs utilizada en el presente trabajo permitió la detección de un importante número de familias (clusters en AnoTExcel) (Figura 5a) que incluyen secuencias de elementos completos principalmente para los elementos LTRs y NLTRs (Figura 5b y 5c). Ya los elementos de clase II se encuentran notablemente deletados (Figura 5d).

Resumiendo, AnoTExcel se presenta como una plataforma para el análisis de elementos repetitivos en *Anopheles gambiae,* que contiene familias de TEs con diferentes grados de deterioración, e incluye los resultados de una serie de análisis permitiendo su uso por investigadores interesados en diferentes aspectos de los TEs. El estudio de las familias presentes aquí puede ayudar a una mejor comprensión de algunas características importantes de los TEs presentes en este genoma, tales como la variabilidad en las regiones terminales (TIRs o LTRs), los procesos de deterioración o la distribución cromosómica de las diferentes familias o clases de elementos. Es decir, reúne información

compilada que puede servir como punto de partida para futuras investigaciones que tengan como objetivo realizar estudios específicos sobre características específicas de las familias de TEs en *Anopheles* o de estudios comparativos con otras especies.

La plataforma utilizada para la generación de AnoTExcel puede servir para generar una base análoga basada en el genoma secuenciado de otros organismos, lo que puede ser de gran utilidad en estudios de las dinámicas de familias de elementos de transposición específicos y también en estudios comparativos. Esto permitiría estudiar las dinámicas de ciertos TEs en ambientes genómicos y selectivos diferentes, además de poder asociar dinámicas de replicación y deterioración a los genomas en las cuales estos elementos se reproducen, permitiendo analizar los genomas en términos de su permisividad a la amplificación de TEs en general o de algunos elementos en particular. Estudios comparativos pueden también evidenciar el proceso de transmisión horizontal de algunos elementos así como de procesos de deterioración propios de algunos elementos. Este tipo de material puede también aportar información para estudios de utilización de elementos de transposición como marcadores moleculares entre diferentes poblaciones o formas moleculares de *Anopheles*, en caso que nuevos genomas de este complejo sean secuenciados.

La información que surge a partir de proyectos de secuenciación genómicos ofrece la posibilidad de un cambio cualitativo en el estudio de los procesos evolutivos que operan sobre los genomas. En el caso de los elementos de

transposición en particular permite estudiar la dinámica de los diferentes elementos presentes en uno o varios genomas, trascendiendo a limitaciones impuestas por el uso exclusivo de técnicas laboratoriales para tales fines.

CARACTERIZACIÓN DE LOS ELEMENTOS DE TRANSPOSICIÓN EN ANOTEXCEL

Cuando se pretende realizar una caracterización de cualquier conjunto de elementos u objetos con ciertas características comunes, es esencial comprender cuáles son esas características y cuáles los puntos en que difieren. Caso ya exista una clasificación, comprender claramente los criterios utilizados en la misma es esencial. Existen varias formas "razonables" de observar similitudes o diferencias en cualquier conjunto de elementos y una clasificación puede basarse en cualquiera de las innumeras características que emergen de un dado conjunto de datos. En particular, la clasificación de los elementos de transposición se ha tornado en los últimos tiempos en un tema controvertido [**Wicker et al., 2007**, **2009**; **Kapitonov & Jurka, 2008; Seberg & Petersen, 2009**]. La clasificación de estos elementos por debajo del taxón "Superfamilia" (esto es, los taxa correspondientes a familia, subfamilia, etc) no es clara, lo que queda en evidencia en la propia bibliografía específica de TEs donde diferentes autores se refieren a familias, linajes, clados, etc para referirse a secuencias que comparten el mismo grado de semejanzas en diferentes organismos o en el mismo organismo para referirse a diferentes TEs. De esta manera, no queda claro a qué nos referimos cuando citamos a un conjunto de elementos pertenecientes a una misma familia o linaje. En Repbase se encuentran

depositadas “secuencias consenso prototípicas de grandes familias y subfamilias de elementos repetitivos” que se encuentran depositadas en la siguiente dirección (http://www.girinst.org/repbase/update/index.html). Sin embargo, no está explícito a qué se refiere el término “familia” a nivel de similitudes o grado de identidad entre las secuencias que la componen. Ya, en Tefam encontramos el término “Elementos” para referirse a las “múltiples copias generadas por eventos de transposición”.

En esta tesis, solo fueron considerados secuencias identificadas en el genoma del mosquito *Anopheles gambiae* y no fue el objetivo aportar a la controversia existente en relación a la anotación y clasificación de TEs. No obstante, en la tentativa de caracterizar a los elementos de transposición hallados en AnoTExcel, ha sido inevitable notar problemas de clasificación presentes en las diversas familias de elementos descritos y depositados previamente en Repbase y Tefam. Esto incluye la denominación de “familias” diferentes para conjuntos de secuencias con altos grados de identidad, que deberían ser consideradas como miembros de una misma familia, así como también problemas de denominación de elementos prácticamente idénticos como pertenecientes a familias diferentes. Estos problemas entre otras cosas, llevan a una sobrestimación del número de familias que componen a algunas superfamilias, además de dificultar una comprensión de la diversidad y abundancia real de las familias de TEs en los genomas analizados.

En AnoTExcel se identificaron elementos de transposición pertenecientes a las Clases *I* y II, además de siete clusters de elementos anidados (Nested) (correspondiendo, en total, al 74,6% del total de secuencias identificadas por

PILER) (Figura 5a, pagina 51). El resto de los clusters fueron clasificados como “Repeticiones” sin características distintivas de TEs (donde se incluyen algunas secuencias de tipo satélites) y clusters conteniendo secuencias pertenecientes a rRNAs.

A continuación se presenta una caracterización detallada de los elementos de transposición presentes en AnoTExcel: 36 clusters con características de TEs corresponden a elementos que no habían sido caracterizados previamente en el genoma de *Anopheles gambiae*, es decir, constituyen familias nuevas de TEs. Estos corresponden a cuatro elementos del orden LTR y 32 pertenecientes a la clase II (incluyendo a diversas superfamilias).

ELEMENTOS CLASE I- ORDEN LTR

Los elementos LTRs, de modo general, se caracterizan por la presencia de dos regiones estructurales: las repeticiones terminales largas (LTRs propiamente dichos) en ambos extremos y la región interna (*I*) compuesta por genes estructurales. Estos genes son el gen *gag* que codifica para proteínas estructurales (matriz, cápside y nucleocápside), el gen *pol* que codifica para proteínas relacionadas con la transcripción de estos elementos: AP (Proteinasa Aspartica), RT-PR (Transcriptasa Reversa-Proteasa) RNaseH e INT (Integrasa), y, en algunos casos, un gen de envoltura *env* relacionado con retrovirus pero que en el caso de retrotransposones o bien se encuentra ausente o no es funcional [**Lerat & Capy, 1999**]. El orden relativo de estos dominios puede variar entre una superfamilia y otra. Estos elementos se encuentran relacionados con retrovirus, y han sido propuestos varios escenarios evolutivos para explicar la

relación existente entre ellos [**Xiong & Eickbush, 1990, McClure, 1991, Capy et al., 1996**]. Los LTRs propiamente dichos, presentes en las regiones 3' y 5' de cada elemento son considerados idénticos en el momento de inserción, pero posteriormente estos evolucionan de manera neutral y cada LTR experimenta mutaciones independientemente [**Bergman & Bensanson., 2007**].

En AnoTExcel los elementos del orden LTR corresponden al 18% del total de clusters identificados (Figura 5a) y están representados por miembros de las tres principales superfamilias de LTRs presentes en el genoma *Anopheles gambiae*: *Ty1-Copia*, *Ty3-Gypsy* y *Pao_Bel*. Los elementos pertenecientes a la superfamilia *Pao-Bel* son los más abundantes en términos del número de secuencias totales identificadas seguidos por las familias *Gypsy*s (Tabla 1). La superfamilia *Copia* es la menos representada con tan solo dos clusters y siete secuencias identificadas.

Tabla 1. Número (N) y porcentaje (%) de clusters y de secuencias identificados en cada Superfamilia de elementos del orden LTR. Obs= no se encuentran contemplados los elementos Nuevos encontrados en esta clase.

	Clusters		**Secuencias**	
LTRs	**N**	**%**	**N**	**%**
Gypsy	22	55,0	93	43,7
Copia	2	5,0	7	3,3
Pao-Bel	16	40,0	113	53,1
Total	**40**		**213**	

A continuación se presenta un análisis filogenético de los miembros de las diferentes superfamilias de LTRs presentes en AnoTExcel, junto a secuencias de referencia obtenidas en RB (Figura 6). El análisis se basa en las regiones comunes identificadas en el dominio de la Transcriptasa Reversa de las

secuencias consenso de todos los clusters clasificados como LTR Full y LTR Novel en AnoTExcel.

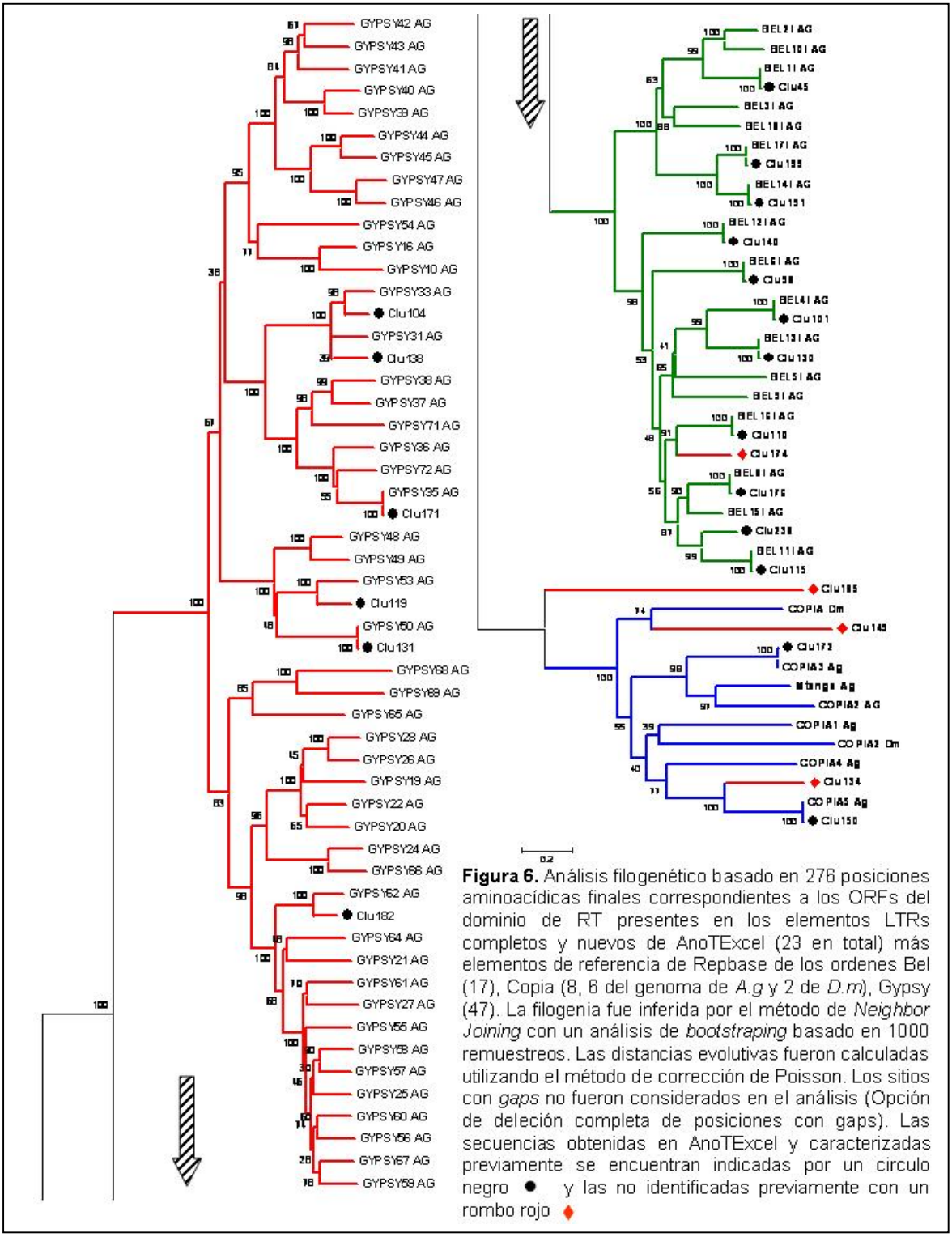


Figura 6. Análisis filogenético basado en 276 posiciones aminoacídicas finales correspondientes a los ORFs del dominio de RT presentes en los elementos LTRs completos y nuevos de AnoTExcel (23 en total) más elementos de referencia de Repbase de los ordenes Bel (17), Copia (8, 6 del genoma de *A.g* y 2 de *D.m*), Gypsy (47). La filogenia fue inferida por el método de *Neighbor Joining* con un análisis de *bootstraping* basado en 1000 remuestreos. Las distancias evolutivas fueron calculadas utilizando el método de corrección de Poisson. Los sitios con *gaps* no fueron considerados en el análisis (Opción de deleción completa de posiciones con gaps). Las secuencias obtenidas en AnoTExcel y caracterizadas previamente se encuentran indicadas por un circulo negro ● y las no identificadas previamente con un rombo rojo ♦

El objetivo de este análisis fue entender cuán representativas son las familias de elementos identificadas en AnoTExcel en relación a las familias de elementos identificadas previamente en el genoma de *Anopheles*. Además de estudiar la diversidad de elementos presentes en dicho genoma y la estructura de estas familias en términos de subfamilias. Todas las secuencias analizadas se agruparon con secuencias de las superfamilias a las cuales fueron asignados en AnoTExcel (sin análisis filogenético) con valores de *bootstraping* de 100. La única excepción corresponde al cluster 185 (caracterizado en AnoTExcel como un elemento Nuevo perteneciente a la superfamilia *Pao-Bel*) que en la filogenia ocupa una posición intermedia entre los elementos pertenecientes a las superfamilias *Copia* y *Pao-Bel*, aunque sin peso estadístico.

Superfamilia *Copia*

La superfamilia *Copia* está representada en el genoma de *Anopheles gambiae* por 5 familias diferentes: *Copia*1-5_AG [**Kapitonov et al., 2003; Pavlicek et al., 2003**] y la familia *Mtanga* [**Rohr et al., 2002**]. La distancia-p media (ver Matriz de distancias-p para *Copia* en Anexo 2) calculada a partir de las secuencias consensos presentes en RB es 0,6608 (mínima=0,5141; Máxima=0,7486) lo que confirma la clasificación como familias diferentes pertenecientes a la Superfamilia *Copia*. De los elementos detectados en AnoTExcel, cuatro clusters fueron clasificados como *Copia* (Figura 6). Uno de ellos corresponde a la familia *Copia* 3_AG y otro a la familia *Copia*-5_AG (clusters 172 y 150, respectivamente) (Figura 7). Los otros dos corresponden a elementos que

pertenecen a esta superfamilia (clusters 149 7 134) pero que presentan

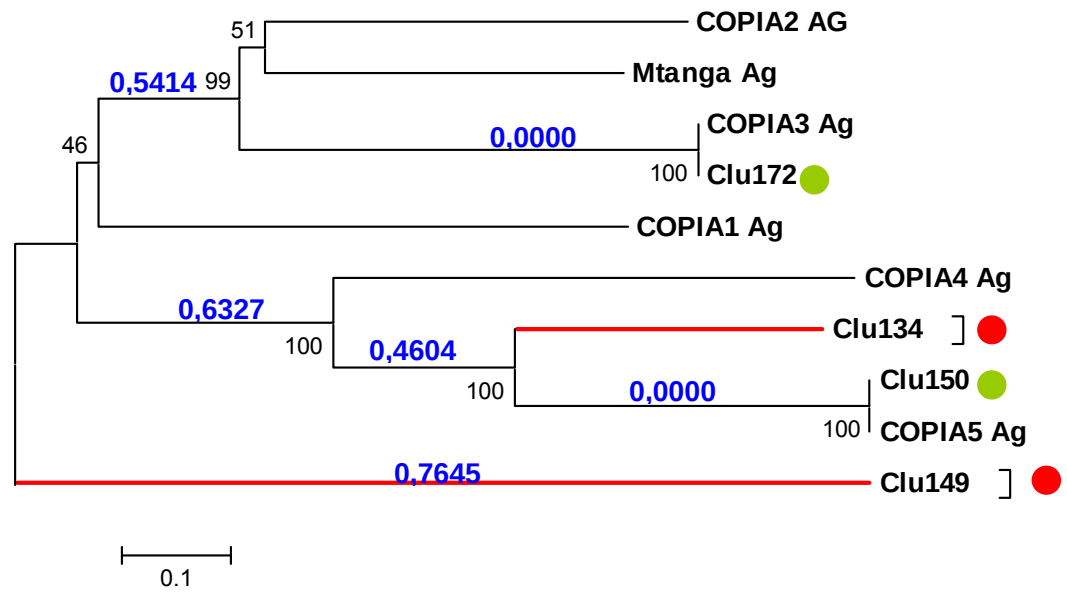


Figura 7. Filogenia realizada con el método de *Neighbor-Joining* basada en el alineamiento aminoacídico (354 posiciones) de los ORFs presentes en las secuencias consenso de AnoTExcel y en el conjunto de secuencias consenso pertenecientes a la Superfamilia *Copia* en *Anopheles gambiae* según Repbase. Los valores de *bootstraping* obtenidos después de 1000 muestreos se presentan en las ramas en negro. Para los nodos con valores de *boostraping* significativos (*I*.e mayores a 70) se presentan los valores de distancias-p obtenidos para las secuencias que se encuentran a la derecha del nodo respectivo (en azul). Con un círculo rojo se indican dos familias de elementos *Copia* no identificados previamente y en verde familias de elementos identificados en AnoTExcel ya identificadas previamente.

distancias aminacídicas muy grandes con otras secuencias (0,4604 y 0,7645 respectivamente) por lo tanto parece tratarse de familias nuevas, que sugerimos identificar como *Copia*6_AG y *Copia*7_AG siguiendo la nomenclatura utilizada en Repbase (estos clusters serán descriptos detalladamente en la sección 4.2).

Los clusters 150 y 172 de AnoTExcel, contienen 4 y 3 secuencias respectivamente, siendo que las familias a las cuales corresponden tienen, según la información depositada en RB, 5 y 14 secuencias, respectivamente que son 99 y 98% idénticas. El programa PILER no detectó la presencia de estas

otras secuencias en el genoma de *Anopheles* ni de miembros pertenecientes a las otras familias *Copia*.

Superfamilia *Pao-Bel*

Los elementos pertenecientes a la superfamilia *Pao-Bel* descriptos previamente y depositados en Repbase suman 18 familias en total que tienen entre ellas una distancia-p media a nivel aminoacídico de 0,7023 (min=0,3646; Max=0,8389). En AnoTExcel se encuentran representantes de once de estas familias. Las distancias-p entre algunas de las mismas (ver Matriz de distancias-p para elementos Pao_*Bel* en el anexo 2 y figura 8) permite concluir que por lo menos cuatro familias *Pao-Bel* previamente caracterizadas componen en verdad dos familias con subfamilias asociadas. Es el caso de los elementos *Bel*-2_AG, *Bel*-7_AG que tienen 41% de distancia aminoacídica en la región utilizada para la filogenia (región conservada entre todas las secuencias de la transcriptasa reversa). El elemento *Bel*-7 es un elemento considerado no autónomo por la falta del dominio para la transcriptasa reversa (RT) y la integrasa (INT) que presenta un ORF de 720 aa, en cuanto *Bel*-2 presenta un ORF de 1750 aa que incluye los dominios de RT e INT. También los elementos *Bel*-14 y 17_AG –dos elementos no autónomos- presentan una distancia en la región analizada (0,3646) compatible con la clasificación de las mismas como miembros de subfamilias. El programa PILER con las configuraciones utilizadas no detectó la variabilidad ya reportada previamente en esta superfamilia, pero por otra parte, se identificaron dos familias de elementos no descritos previamente. Los clusters identificados con los números 185 y 174, según la clasificación realizada a partir de lo datos presentes en RB corresponden a

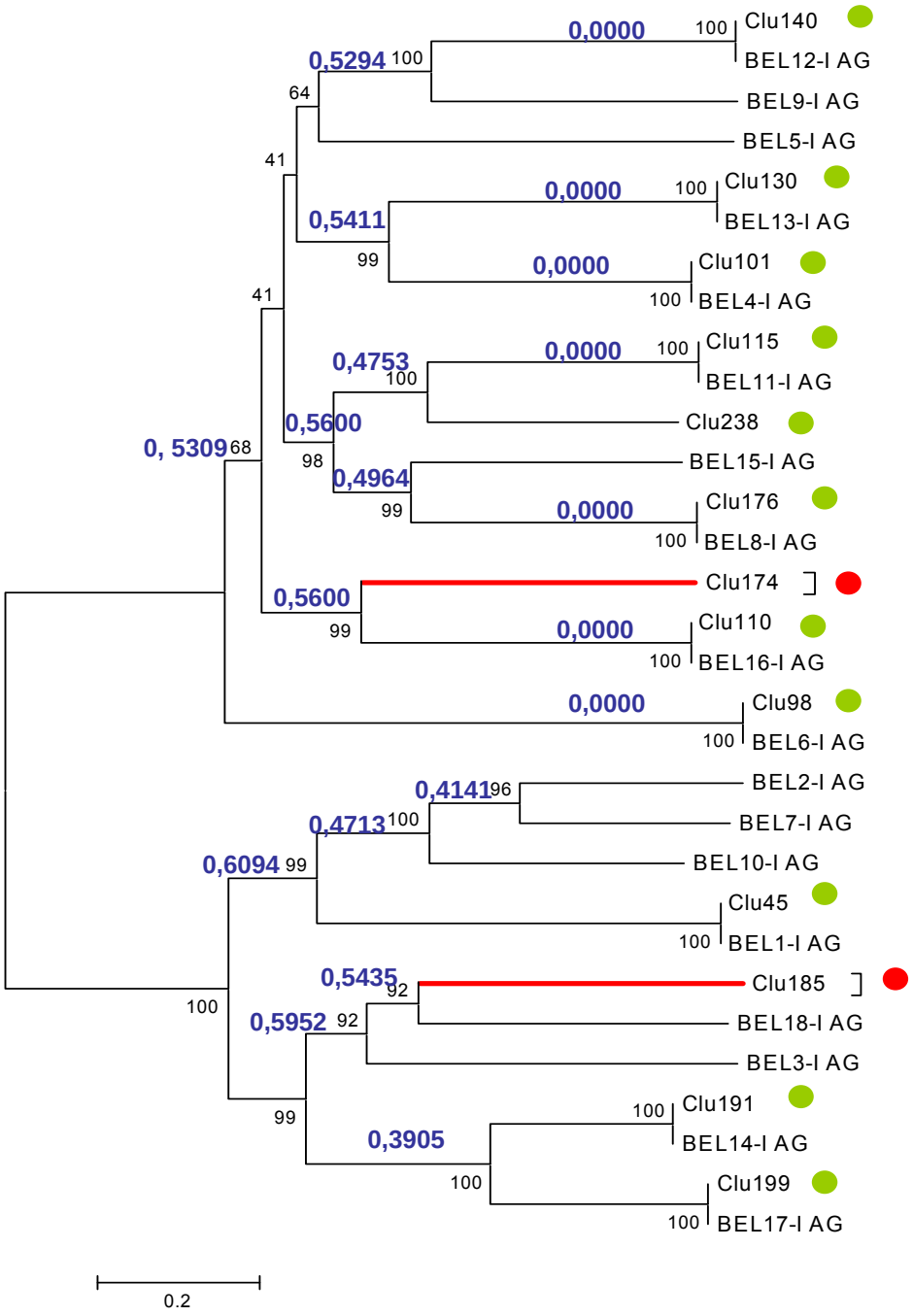


Fig 8 Filogenia realizada con el método de *Neighbor-Joining* basada en el alineamiento aminoacídico (425 posiciones) de los ORFs presentes en las secuencias consenso de la Superfamilia *Pao-Bel* de AnoTExcel más secuencias referencia de Repbase de esta superfamilia. Ver la leyenda de la figura 7 para mayores especificaciones. Los valores de bootstraping obtenidos después de 1000 muestreos se presentan en las ramas en negro. Para los nodos con valores de boostraping significativos (I.e mayores a 70) se presentan los valores de distancias-p obtenidos para las secuencias que se encuentran a la derecha del nodo respectivo (en azul). Las secuencias obtenidas en AnoTExcel y caracterizadas previamente se encuentran indicadas por un círculo verde y las no identificadas previamente con un círculo rojo

familias *Pao-Bel* no detectadas previamente, o eventualmente a miembros bastante distantes. Estos clusters fueron denominados *Bel*-20_AG y *Bel*-21_AG y serán descriptos con mayor detalle en la sección 4.2.

Superfamilia *Gypsy*s

Los elementos *Gypsy* en Repbase componen la superfamilia de elementos más diversa en términos del número de familias que la integran [**Tubio et al., 2004, 2005**]. Sin embargo, a partir de las distancias-p calculadas entre las secuencias consenso de las diferentes familias presentes en Repbase (ver matriz de distancias para elementos *Gypsy* en anexo 2) podemos concluir que una parte de esas secuencias constituyen subfamilias por presentar distancias-p muy bajas entre si (menores a 20%). La superfamilia *Gypsy* ha sido tradicionalmente clasificada en nueve linajes diferentes, según análisis filogenéticos de los dominios de RT, RNAseH, e INT [**Malik & Eickbush, 1999**] utilizados también en el presente análisis. Seis de estos linajes han sido reportados en insectos. En *Anopheles gambiae*, cinco de ellos han sido descriptos (*Gyspy, Mag, Mdg1, CsRn1* y *Mdg3*). Tubio et al., [**2004**] reportaron la identificación de una enorme variedad de familias diferentes dentro de cada uno de estos linajes basándose en filogenias de las regiones conservadas de los dominios mencionados anteriormente y en la agrupación de las secuencias confirmada por altos valores de *bootstraping*. Sin embargo, una gran parte de esas secuencias tienen distancias aminoacídicas entre sí menores a 20%, constituyendo, o bien subfamilias, o bien el producto de eventos de transposición de un mismo elemento o de elementos muy próximos dentro de la misma subfamilia.

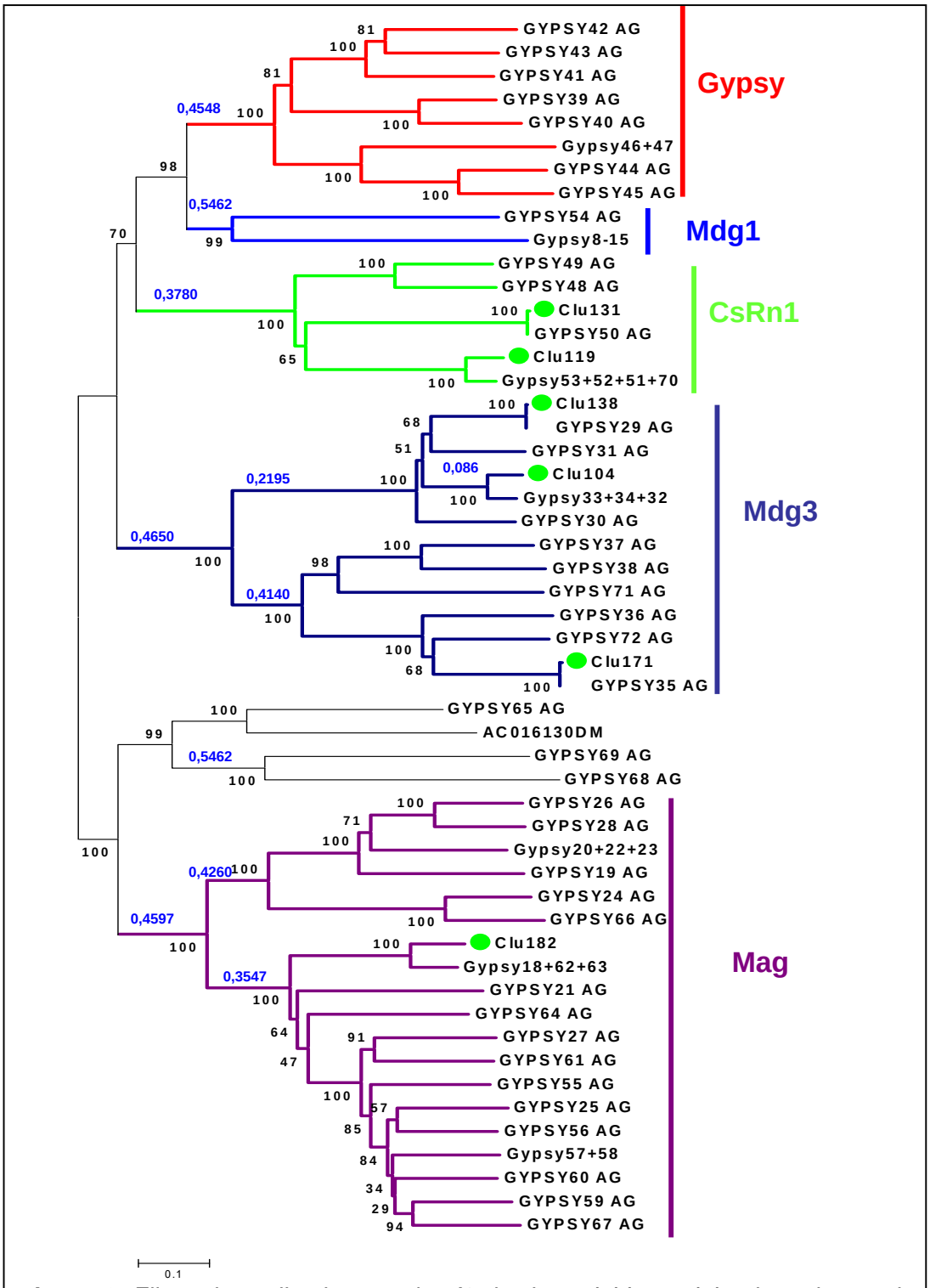


Figura 9. Filogenia realizada con el método de *Neighbor-Joining* basada en el alineamiento aminoacídico (465 posiciones) de los ORFs presentes en las secuencias consenso de la Superfamilia *Gypsy* de AnoTExcel más secuencias referencia de Repbase de esta superfamilia. Ver la leyenda de la figura 7 para mayores especificaciones. Con un circulo verde se indican las secuencias identificadas en AnoTExcel. Los considerados Linajes diferentes se indican a la derecha del árbol en diferentes colores.

En el anexo 2 se muestra la matriz de distancias para los elementos Gyspsy presentes en Repbase (Matriz_dist-p_GYPSY_REPBASE), indicando las distancias-p aminoacídicas entre las diferentes secuencias de *Gypsy*s reportados. Es evidente que muchas de las secuencias consideradas y denominadas como familias de elementos diferentes componen secuencias que son el producto de eventos de transposición dentro de una misma familia. Así, por ejemplo, todos los elementos dentro del denominado linaje *Mdg1* tienen entre si distancias que oscilan entre 0,0235 y 0,2431, con una media de 0,2353 (ver matriz de distancias para elementos *Gypsy* en Repbase en el Anexo 2) lo que, además teniendo en cuenta sus características genéticas estructurales (ver **Tubio, 2004**) las define como secuencias pertenecientes a una misma familia. En cuanto al linaje denominado *CsRn1* parece contener secuencias que conforman una única familia. El linaje *Mag* estaría compuesto por dos familias que tienen secuencias con 0,3547 y 0,4260 de distancia entre si. Considerando estos datos, se realizó una filogenia de los elementos *Gypsy*s completos identificados en AnoTExcel (seis clusters) junto a secuencias consenso para las familias *Gypsy*s mencionadas anteriormente (figura 9). Los elementos *Gypsy* identificados en AnoTExcel corresponden a los linajes *CsRn1*, *Mdg3* y *Mag*. No fueron identificados elementos pertenecientes a los linajes *Gypsy* y Mdg1.

ELEMENTOS CLASE I- ORDEN NLTRS

Los elementos NLTR presentan entre 3 y 8 kb de largo y han sido descritos en prácticamente todos los organismos eucariotas analizados. Contienen, en general, dos genes: el gen de RT indispensable para su replicación, y el gen gag

presente también en los retrotransposones LTRs y en retrovirus. Otras estructuras típicamente presentes en los elementos NLTRs son promotores para la polimerasa II y señales de polyadenilación AATAAA en el extremo 3'.

En total, en AnoTExcel, fueron identficados 27 clusters (634 secuencias) como elementos NLTR (Ver Tabla 2), de los cuales 5 fueron considerados remanentes de NLTRs por su baja identidad a nivel aminoacídico con NLTRs ya descritos.

Tabla 2, Número (N) y porcentaje (%) de clusters y secuencias identificados en cada Superfamilia de elementos del orden NLTR.

	Clusters		**Secuencias**	
NLTR	**N**	**%**	**N**	**%**
RTE	3	11,1	242	38,2
Outcast	2	7,4	6	0,9
Jockey	4	14,8	47	7,4
I	1	3,7	4	0,6
CR1	17	63,0	335	52,8
Total	**27**		**634**	

Elementos pertenecientes a las superfamilias *RTE*, Outcast, *Jockey*, *I* y *CR1* fueron identificados siendo la inmensa mayoría pertenecientes a la superfamilia *CR1*. Los elementos NLTRs fueron clasificados como *Full* o *Fragment* (completos o fragmentados) en base al tamaño de la secuencia consenso, es decir, que dentro de un cluster clasificado como Completos pueden existir secuencias individuales fragmentadas (Ver figura 5, página 51). Es por ello que la proporción de clusters con secuencias completas es mucho mayor a la proporción de elementos completos cuando se analizan las secuencias individualmente. El análisis de la deterioración estructural de estos elementos será analizado posteriormente. No fueron identificados en AnoTExcel secuencias pertenecientes a las superfamilias *Loner, L1, R1, R4* y *CRE.*

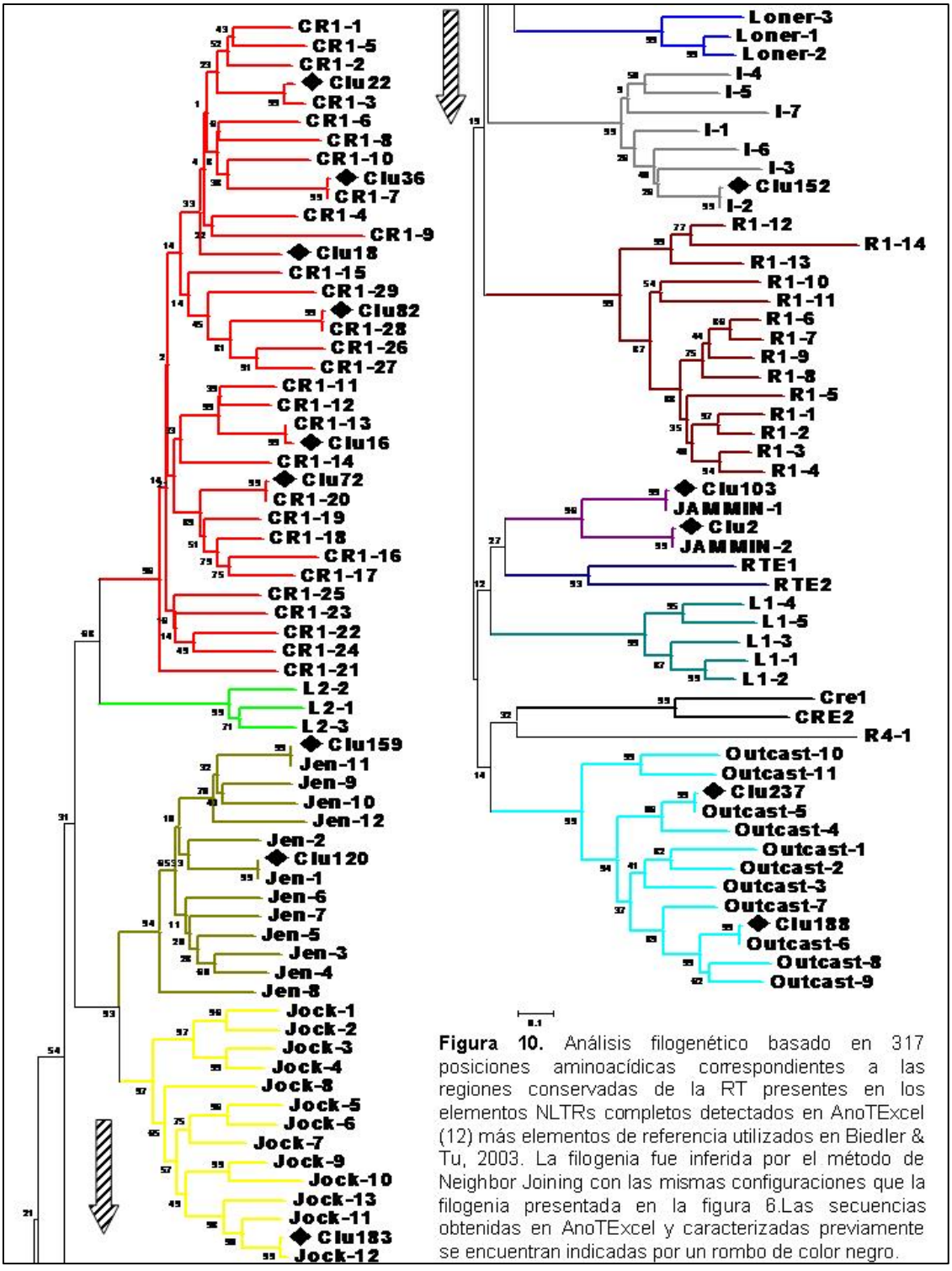


Figura 10. Análisis filogenético basado en 317 posiciones aminoacídicas correspondientes a las regiones conservadas de la RT presentes en los elementos NLTRs completos detectados en AnoTExcel (12) más elementos de referencia utilizados en Biedler & Tu, 2003. La filogenia fue inferida por el método de Neighbor Joining con las mismas configuraciones que la filogenia presentada en la figura 6.Las secuencias obtenidas en AnoTExcel y caracterizadas previamente se encuentran indicadas por un rombo de color negro.

La totalidad de los elementos NLTR que se encuentran en AnoTExcel a partir de los cuales se obtuvieron ORFs (completos o parciales) fueron analizados filogenéticamente junto al conjunto de secuencias de referencia de RB y TF

(Figura 10). La filogenia confirma la clasificación sugerida siguiendo los criterios de clasificación a partir de la información presente en AnoTExcel. Para una mejor visualización y análisis de estos elementos, se realizaron filogenias parciales para cada familia de elementos: *CR1*-1; *Jockey*; *I*; *RTE* y Outcast (Figura 11 a, b, c, d y e).

Superfamilia *CR1*

Se analizaron seis elementos de la familia CR-1 presentes en AnoTExcel, cuatro de ellos correspondiendo a elementos completos mientras que los otros dos (Clusters 22 y 36) fueron clasificados como fragmentos, pero fue posible incluirlos en este análisis por presentar regiones conservadas del ORF utilizado en la filogenia. Ésta confirma la clasificación realizada en AnoTExcel para todos los clusters con excpeción de dos (números 72 y 22) que aparecen con denominaciones diferentes en Tf y RB. Estas discrepancias de deben a que en RB y Tf existen elementos prácticamente idénticos depositados con denominaciones diferentes. Es el caso de los elementos de RB *CR1*-2, *CR1*-3, *CR1*-4, *CR1*-5 y *CR1*-6 que corresponden a los elementos de la base Tf *CR1*-Ele3, *CR1*-Ele23, *CR1*-Ele27, *CR1*-Ele19, *CR1*-Ele20 y *CR1*-Ele13, respectivamente. Todos estos casos muestran menos de 1,5% de diferencias nucleotídicas a lo largo de la secuencia completa (datos no mostrados). Los elementos *CR1*_Ele 23, Ele27 y Ele13 parecerían corresponder a fragmentos de los elementos completos depositados en RB. Esto demuestra una vez más, los problemas de anotación y criterios clasificatorios que existen en la denominación de TEs, lo que dificulta el análisis de estas familias y genera incertidumbre en

relación a la abundancia y diversidad de las diferentes familias de TEs en *Anophelesgambiae*

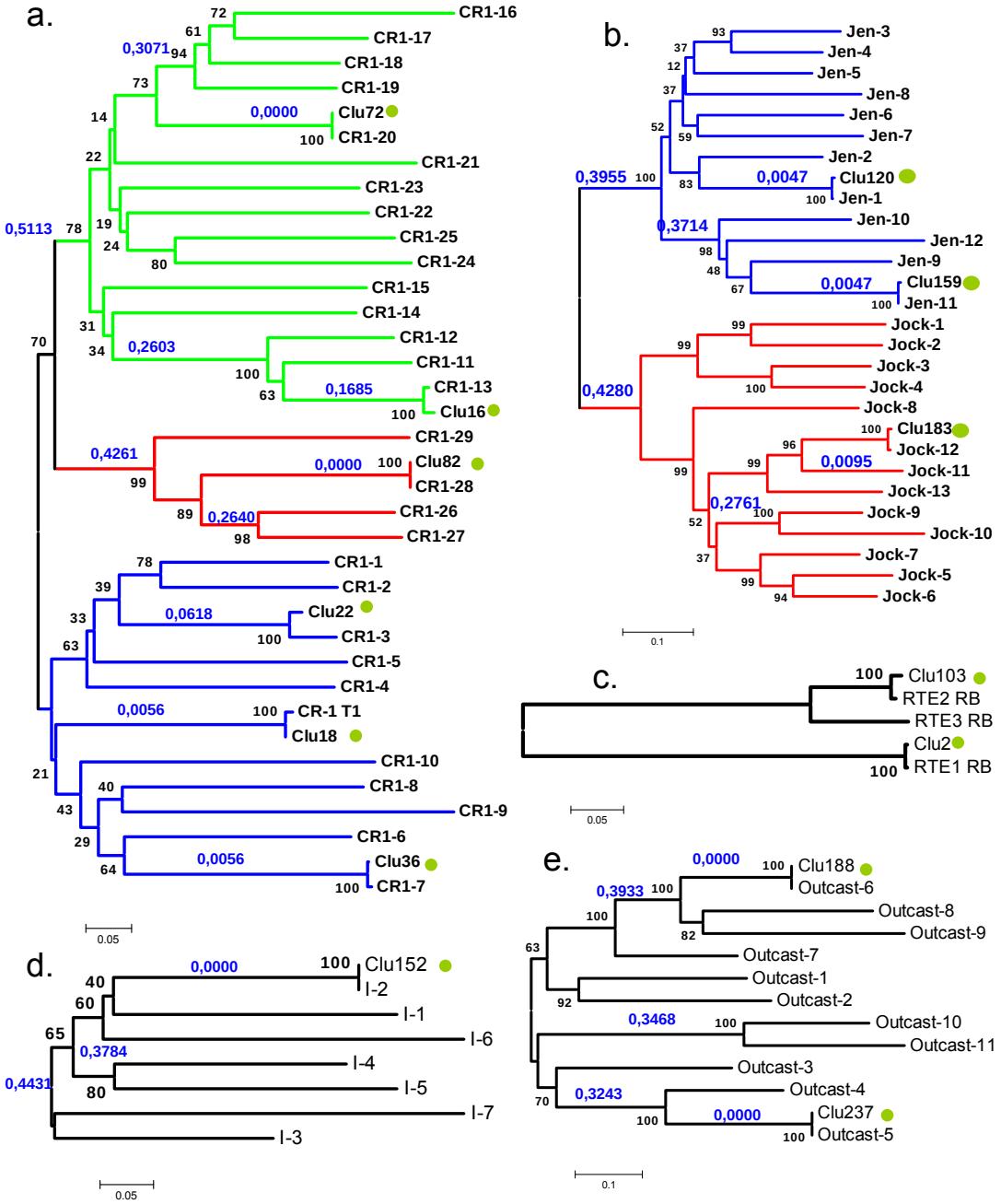


Figura 11. Análisis filogenéticos de los elementos de las diferentes familias de NLTRs presentes en AnoTExcel y secuencias consenso de RB y Tf. Los análisis fueron realizados con las mismas configuraciones detalladas anteriormente en la figura 7. (a) elementos *CR1* (178 posiciones aminoacídicas); (b) elementos *Jockey* (210 posiciones); (c) elementos *RTE* (472 posiciones); (d) elementos *I* (259 posiciones) y (e) elementos Outcast (222 posiciones).

Basándonos en la filogenia presentada en la figura 11a y en las distancias-p entre las secuencias de esta filogenia puede decirse que la superfamilia de elementos *CR1* se encuentra compuesta por tres familias (indicadas en diferentes colores en la Figura 11a). En AnoTExcel se identificaron elementos pertenecientes a estas tres familias, pero no fueron detectadas elementos pertenecientes a otras familias.

Superfamilia *Jockey*

Los elementos *Jockey* en *Anopheles* (Figura 11b) parecen también estar subdivididos en dos subfamilias de acuerdo a las distancias-p entre sus secuencias. Todos los elementos pertenecientes a esta superfamilia, identificados previamente en *Anopheles gambiae,* se encuentran depositados únicamente en la base Tefam. Algunos de los elementos reportados como *Jockey* tienen grados de identidad elevados, mayores al 75%, lo que lleva a la conclusión que también en este caso se trata de elementos que no deberían ser clasificados como diferentes familias sino como diferentes inserciones de la misma familia. Estos elementos *Jockey* han sido denominados *Jockey*_Ele1 a *Jockey*_Ele25 y doce de ellos (elementos denominados 14 a 25) tienen también un nombre sinónimo (Jen 1 a 12). Basándonos en las distancias aminoacídicas en la región analizada en la filogenia, se observa que estas diferentes denominaciones representan en verdad dos familias de elementos diferentes, tal como se observa en la figura 11b. En AnoTExcel se identificaron cuatro clusters pertenecientes a la superfamilia *Jockey*, pero tan solo tres de ellos con elementos completos.

Superfamilia *RTE*

Según la base de datos RB esta superfamilia está compuesta por tres familias, denominadas *RTE*-1,2 y 3_AG. Siendo RTE3 una versión truncada en la región 5'. En Tefam se encuentran depositados dos elementos denominados *RTE*_Ele1 y Ele2. La secuencia depositada como RTE1_AG en RB y la secuencia *RTE*_Ele2 en TF tienen una distancia nucleotídica a lo largo de sus 3351 nucleótidos de 1,21% y la familia denominadas RTE2_AG en RB y *RTE*_Ele1 en TF tienen una distancia de 1,46%, lo que indica que se trata de secuencias que corresponden a las mismas familias mostrando más una vez problemas de identificación. En AnoTExcel se identificaron dos clusters (2 y 103, respectivamente) correspondientes a dos de estas familias (RTE2_AG y RTE1_AG respectivamente) (Figura 11d). EL cluster 2 posee 232 secuencias con diferente grado de deterioración y será analizado con detalle en la sección 4.3.

Superfamilia *I*

Tan solo una familia de elementos *I* fue identificada en AnoTExcel y clasificada como *I*_ele2 (Figura 11c). Según la base TF, esta superfamilia está compuesta por siete familias (cuatro de ellas con elementos truncados), cuyas secuencias se encuentran depositadas exclusivamente en esa base de datos.

Superfamilia *Outcast*

Esta superfamilia está compuesta por 11 familias de elementos depositados en TF. En AnoTExcel fueron identificados sólo dos clusters correspondientes a las familias Outcast 5 y 6 (Figura 11e)

CLASE II

Los elementos de clase II están caracterizados por la presencia de TIRs y también por TSD (*Target Site Duplications*). Esta clase de elementos ha sido recientemente subdividida en dos subclases (1 y 2) dependiendo del número de hebras de DNA que son cortadas durante la transposición [**Wicker et al., 2007**]. La gran mayoría de los elementos clase II en el genoma de este mosquito están inactivos presentando diversos grados y patrones de deterioración. En RB se encuentran depositadas tan solo cuatro familias de elementos de clase II completos (*Mariner1_AG, Mariner2_AG, Tc1-1_AG y TsessebeII*), todos pertenecientes a la superfamilia *Tc1-Mariner*. En Tf son seis, siendo que dos de ellos corresponden a los mismos elementos *TsessebeII y Tc1-1* de RB. Los elementos clase II en AnoTExcel fueron clasificados en base a la presencia de TIRs en las secuencias así como en la existencia de *matches* positivos con elementos clase II previamente caracterizados. Estos corresponden a las superfamilias *Tc1-Mariner, Harbinger, hAT, PiggyBac, P e Ikirara* en la subclase *I* y *Helitron* de la subclase II, totalizando71 clusters diferentes con 1324 secuencias (Tabla 3).

La mayoría de los mismos corresponde a elementos deletados, no activos, presentando diferentes grados de deterioración, que llevaron a la clasificación de los mismos en: “Fragmentos”, “Mites” o “NA” (no autónomos según denominaci Repbase). Esta variedad de elementos deteriorados será analizada en mayor detalle en la sección 4.3.

Tabla 3. Número (N) y porcentaje (%) de clusters y secuencias identificados en cada Superfamilia de elementos de la Clase II. Obs= no se encuentran contemplados los elementos Nuevos encontrados en esta clase.

	Clusters		**Secuencias**	
CLASS II	**N**	**%**	**N**	**%**
C1_Mar	37	52,1	765	57,8
Helitron	5	7,0	20	1,5
Harbinger	3	4,2	91	6,9
hAT	8	11,3	70	5,3
PiggyBac	5	7,0	106	8,0
P	11	15,5	256	19,3
Ikirara	2	2,8	16	1,2
Total	**71**		**1324**	

Fueron considerados como familias de elementos completos seis clusters, tres perteneciente a la superfamilia Tc1_Mariner, uno a la familia P y dos corresponden a Ikirara. La figura 12 muestra la filogenia de los tres elementos Tc1-Mariner con referencias obtenidas de Tf. A partir de las distancias entre estos elementos es evidente que el conjunto de elementos correspondientes a mariners constituyen diferentes elementos pertenecientes a una única familia.

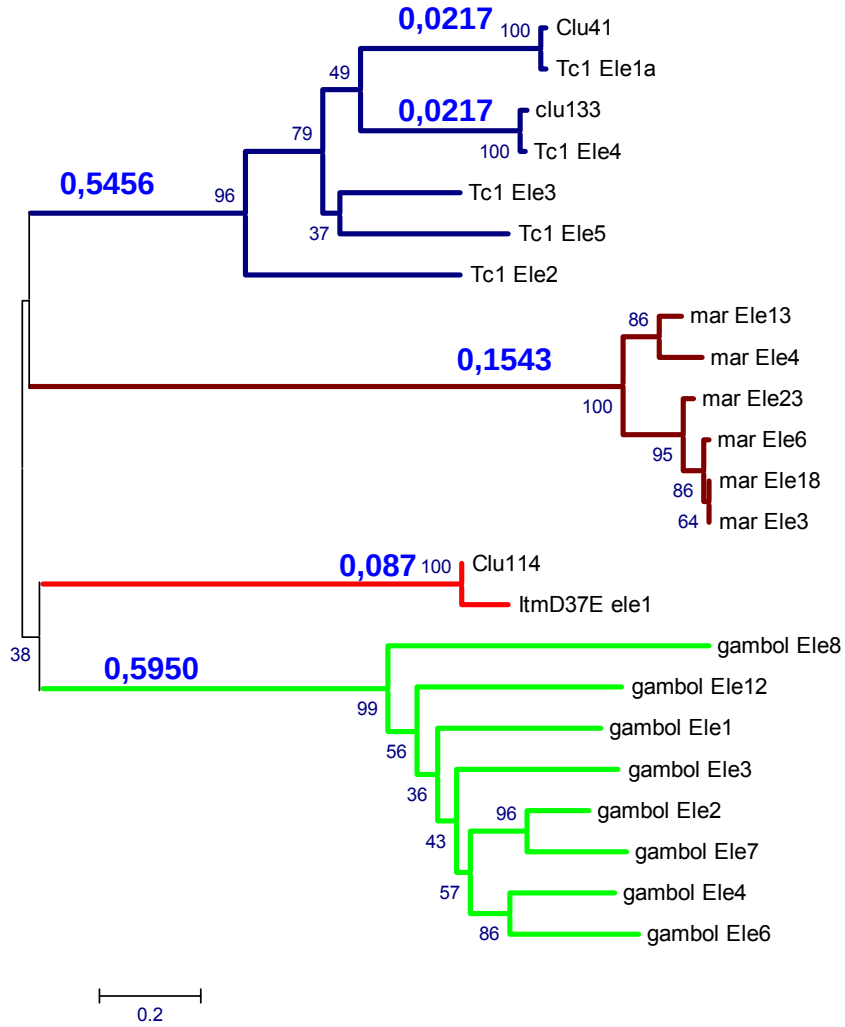


Figura 12- Análisis filogenético de elementos de la superfamilia Tc1-Mariner presentes en AnoTExcel. Esta filogenia está basada en el alineamiento aminoacidico correspondiente al ORF de estos elementos que incluye 362 posiciones. Ver Figura 7 para mayores detalles

COMPARACIÓN DE ANOTEXCEL CON REPBASE Y TEFAM

La base de datos presentada aquí, AnoTExcel (AT), no pretendió ser una base exhaustiva de TEs en el genoma de *Anopheles*, y como fue mencionado anteriormente no presenta la diversidad de familias que ha sido caracterizada previamente en el genoma del mosquito (por ejemplo, los elementos clase II: Herves, PiggyBac completos, algunas familias de hATN, algunas familias de

Mariners, y algunas familias *Gypsy*s pertenecientes a la clase *I*, entre otros no fueron identificados). Las configuraciones utilizadas al ejecutar el programa PILER pudieron interferir con la detección de algunas de estas familias.

AnoTExcel, no pretende ser una alternativa a Repbase para los transposones del mosquito *Anopheles*, a diferencia de ésta y de Tefam, presenta para las familias consideradas todas las instancias de TEs encontradas, alineadas entre sí y con secuencias consenso para cada una de esas familias. AnoTExcel puede cargarse en un computador personal junto a todos los links asociados a ella, lo que permite su uso a criterio del investigador. También, por tratarse de una planilla de Excel con *hyperlinks* de tipo .txt no necesita de la instalación de ningún programa específico y también permite una visualización y tratamiento de los datos de manera sencilla y accesible.

La Figura 13 presenta una comparación en términos del número de clusters presentes en las tres bases de datos. Repbase presenta elementos pertenecientes a las tres clases de TEs totalizando 190 familias diferentes. Tefam presenta solamente elementos NLTR y Clase II, totalizando 156 familias de elementos.

Como fue mencionado anteriormente, algunos de los elementos presentes en RB se encuentran depositados en Tf con nombres diferentes. Por otra parte, algunas secuencias se encuentran descritas como pertenecientes a diferentes familias cuando en realidad tienen distancias nucleotídicas muy pequeñas entre sí, lo que indica que se trata de diferentes subfamilias o hasta de instancias de transposición diferentes de un mismo elemento.

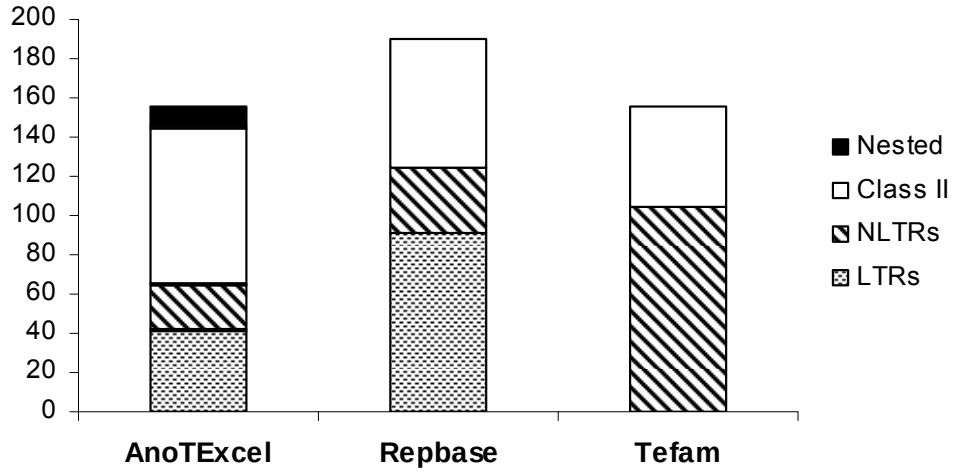


Figura 13. Comparación de bases de datos AnoTExcel, Repbase yTefam
Proporción de familias de elementos de transposición representadas en cada una de las bases de datos analizadas.

En cuanto a la clasificación de los TEs, no se propone aquí un nuevo sistema de clasificación. La reciente propuesta de Wicker et al., 2008 (Ver Cuadro 1 en Anexo 1) parece una clasificación jerárquica útil de los TEs. Parece justificado remitirse a ella y analizar a las familias de TEs considerando la divergencia que presentan sus secuencias entre si, para poder incluirlas dentro de familias comunes o como nuevas familias. Es probable que los TEs caracterizados en otros genomas sufran de los mismos problemas de anotación y clasificación. Esto lleva a una sobreestimación de la diversidad de algunas superfamilias (como es claramente el caso de los elementos *Gypsy*s en *Anopheles*) y también a dificultades a la hora de caracterizar un elemento nuevo encontrado en un determnado genoma.

El número de elementos de transposición diferentes presentes en Repbase, Tefam y AnoTexcel es de 190, 156 y 158, respectivamente. Por tanto, en AnoTExcel se encuentran presentes y analizados en detalle, 83% del total de

elementos identificados previamente en el genoma de *Anopheles gambiae* y presentes es RB lo que incluye una serie de elementos descritos aquí por primera vez. AnoTExcel tiene, por su parte, propiedades únicas descritas anteriormente (como la facilidad en la recuperación de secuencias individuales o de alineamientos de secuencias de una misma familia) permitiendo los análisis propuestos en los objetivos de esta tesis.

CARACTERIZACIÓN DE ELEMENTOS NUEVOS EN EL GENOMA DE *Anopheles gambiae*

La identificación de elementos nuevos en un genoma es importante para conocer la diversidad de los elementos de transposición que conviven en ese genoma así como para entender la dinámica de los mismos en el genoma. Por otra parte, bajo la perspectiva del uso de elementos de transposición como *drivers* genéticos, la identificación de nuevos elementos que potencialmente puedan ser utilizados en este tipo de abordajes permite abrir líneas de investigación relacionadas con este objetivo.

La detección de nuevos TEs se realiza a partir de una vasta serie de herramientas que pueden basarse tanto en técnicas laboratoriales como en el uso de herramientas de bioinformática. El estudio de nuevos TEs comienza con la identificación de copias repetitivas presentes en el genoma, seguido del alineamiento de las secuencias, la clasificación en familias o subfamilias, y la reconstrucción de una secuencia consenso que se utiliza normalmente para la posterior caracterización de esa familia de elementos [**Jurka, 1998**]. Considerando que los elementos de transposición en una familia surjan a partir de uno o unos pocos elementos activos, las secuencias consenso obtenidas a partir de algunos miembros de una determinada familia son considerados una aproximación razonable al elemento original activo que dio lugar a la familia en estudio [**Jurka & Milosavljevic, 1991; Ivics et al., 1997; Lampe *et al*., 1996; Miskey *et al*., 2003].**

Las herramientas de bioinformática para la detección de TEs pueden dividirse básicamente en tres: (*I*) métodos basados en la detección de secuencias repetidas en diferentes localizaciones cromosómicas sin tener en cuenta la homología de las secuencias con elementos ya caracterizados; (ii) métodos basados en la identidad de secuencias nucleotídicas o aminoacídicas con secuencias pertenecientes a elementos de transposición conocidos y (iii) métodos basados en la identificación de características estructurales compartidas por cierto conjunto de elementos, como por ejemplo la presencia de terminaciones repetidas invertidas (*Terminal Inverted Repeats* -TIRs) para los TEs de clase II, o las terminaciones largas repetidas (*Long Terminal Repeats* -LTRs) para los elementos de clase *I* pertenecientes al orden LTR [Ver por ejemplo, **Tu & Coates, 2004; Edgar & Myers, 2005; Quesneville 2003, 2005 y Bergmann & Quesneveille, 2007** como excelentes revisiones de los métodos de detección e identificación de TEs en genomas]. Otras tentativas de identificación de elementos de transposición en un genoma se han basado en las diferencias de composición de nucleótidos de los TEs en comparación con genes del hospedero utilizando modelos probabilísticos como *Hiden Markov Models* (HMM) [**Andrieu et al, 2004**]. Los autores de este trabajo muestran que esta metodología puede resultar complementaria a los métodos basados en la homología de secuencias a TEs conocidos. Existen también otros trabajos donde se detectan elementos LTR nuevos sin depender de una biblioteca de TEs conocidos para analizar la homología de las secuencias [**Rho et al., 2007**]. El programa LTR-STRUC detecta automáticamente elementos LTRs basándose en características estructurales distintivas de este tipo de elementos [**McCarthy**

& McDonald, 2003]. El programa LTR-Finder, por su parte, predice la presencia de elementos LTR completos en una dada secuencia de DNA considerando características estructurales de este tipo de elementos.

La metodología de identificación de nuevos transposones en un genoma basándose en sus características repetitivas ha sido considerada ineficiente [**Quesneville et al., 2003**] debido, por una parte, a que las copias de TEs suelen encontrarse altamente deterioradas en los genomas y por lo tanto conservan pocas regiones de homología y, por otra, por las duplicaciones de segmentos de DNA que son relativamente comunes y difíciles de distinguir de verdaderos TEs [**Quesneville, 2003**]. Sin embargo, mostramos aquí que esta metodología, junto al análisis de características particulares de los elementos de transposición, permite no solo la identificación de una vasta mayoría de elementos identificados previamente, sino y más importante, la detección de familias no descriptas anteriormente, a pesar de los varios estudios ya realizados con el objetivo de identificar TEs en el genoma de *Anopheles gambiae.*

La estrategia de identificación de TEs en el genoma de *Anopheles gambiae* utilizada aquí permitió la identificación de una serie de elementos no identificados previamente, algunos de los cuales muestran claras evidencias de actividad actual o reciente.

ELEMENTOS LTR NUEVOS

Cuatro clusters (identificados con los números 134, 149, 174 y 185) fueron considerados como posibles elementos nuevos, *i*.e no descriptos previamente en el genoma de *Anopheles gambiae* de acuerdo a la información presente en

AnoTExcel y a la filogenia realizada con todos los elementos LTRs presentada anteriormente (Figura 6 pagina 58).

Los cuatro clusters poseen secuencias con LTRs en los extremos, lo que llevó a definirlos como elementos pertenecientes al orden LTR. No poseen *match* positivo por *BlastN* con ningún elemento depositado en RB o Tf, pero si por *BlastX* a proteínas de las superfamilias *Bel* y *Copia* en RB, lo que, junto a la localización de estas secuencias en la filogenia de los LTRs, indica que estos elementos pertenecen a estas superfamilias. En la tabla 4 se presentan las principales características de estos elementos que son descritos a continuación en mayor detalle.

Tabla 4 Características de Elementos LTRs Nuevos

ID	#	N	D-p	N LTRs @	D-p LTR $	Dist-p LTRs promedio *	T	PBS	PPT	ORF	D.C	Tajima´s D	dN/dS
Clu134	4	4399	0.0009	149	0,0000 0,0000 0,0067 0,0067	0.006	192308	Pro (AGG)	11/15 4234-4248	4050	RVE RVT_2 RNAseH	-0,065 NS	dN=0,0004 dS=0,0012 w=0,3333
Clu149	4	2245	0.0082	168	0,0241 0,0000 0,0060 0,0060	0.0168	538462	Val (AAC)+ Met (CAT)	11/15 246-260	1845	N	-0,6545 NS	dN=0,0060 dS=0,0090 w=0,6666
Clu174	3	5769	0.0068	213	0,0000 0,0093 -	0.014	448718	-	-	5274	RT_PepA 17 PepA17 RVE	ND	dN=0,0027 dS=0,0129 w=0,2093
Clu185	3	3803	0.0033	227	0,0087 0,0218 0,0087	0.0143	458333	-	-	2781	N	ND	dN=0,0009 dS=0,0007 w=1,2

ID Número identificador del cluster
Número de secuencias presentes en el cluster
N Tamaño Total de las secuencias en Nts
D-p Distancia-p entre las secuencias completas del cluster
@ Tamaño de los LTRs en nucleótidos
$ Distancia-p entre los LTRs 5-3' de cada secuencia individual en el cluster
***** Distancia-p promedio entre todos los LTRs de las secuencias en el cluster
T Tiempo (en años) desde la Transposición según formula T=k/2( ) k=valor en columna G y u=1,56x10^{-8}
PBS Primer Binding Site (Sitios de Ligación de primer)
PPT PolyPurine tract (señal de polypurinas)
ORF Open Reading Frame (marco de lectura abierto)
D.C Dominios Proteicos Conservados. N=Ninguno detectado
dN/dS Relación entre distancia en posiciones no-silentes y silentes

Superfamilia *Copia*

Cluster 134

Este cluster está compuesto por cuatro secuencias de 4399 nts de largo, que poseen una distancia-p a nivel nucleotídico -considerando el largo de toda su secuencia- de 0,0009 (ds= 0,0003). Tiene LTRs de 149 nts de largo que presentan en conjunto una distancia-p total de 0,0060 (desvío standard, ds=0,004), siendo que en dos de las secuencias los LTRs -5' y 3'- son idénticos (distancia-p=0,000) y en las otras dos existe un único nucleótido de diferencia entre ellos. Es decir, estos elementos tienen una identidad muy elevada, tanto en su región codificante como en la región que corresponde a los LTRs, lo que indica que probablemente se encuentran activos o que recientemente lo estuvieron. Basado en las distancias medias entre los LTRs de todas las secuencias en el cluster y utilizando la tasa de substituciones neutrales del genoma de *Drosophila melanogaster* calculamos que esta familia tiene menos de 200.000 años de evolución en el genoma del mosquito.

Posee un ORF de 4050 nucleótidos (1350 aa) que contiene regiones conservadas para integrasa (RVE), Transcriptasa Reversa (RVT-2) y RNAsaH (Figura14), indicando también que esta familia se mantiene activa hasta el presente o hasta poco tiempo atrás.

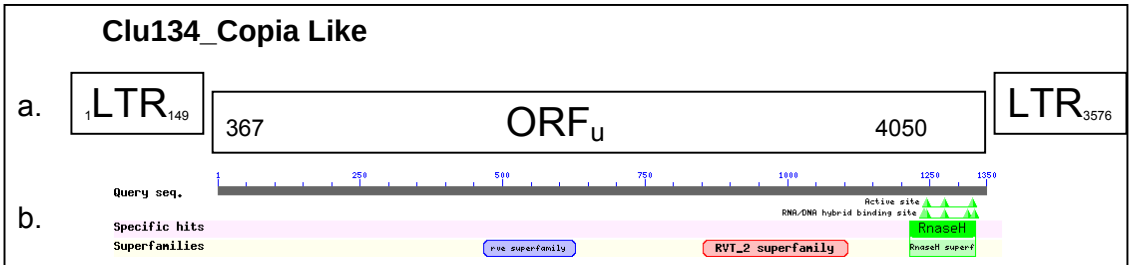


Figura 14. Estructura génica y Dominios Conservados Presentes en la secuencia consenso del cluster 134. a. Los rectángulos corresponden a los dos LTRs que flanquean una región interna equivalente a un marco de lectura abierto único. Los números indicados en los rectángulos corresponden a posiciones nucleotídicas. **b.** Análisis de Dominios Conservados presentes en la región correspondiente al ORF en aminoácidos. Los números en azul corresponden a posiciones aminoacídicas dentro del ORF. El rectángulo azul muestra el dominio correspondiente a la integrasa, en rojo dominio de la transcriptasa reversa y en verde de la RNAsaH (datos obtenidos con el programa *Conserved Domains:* http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml).

Las inserciones de esta familia se encuentran presentes en los cromosomas 3L y 2L. El programa LTR-Finder detectó la presencia de un Sitio de Ligamiento de Primer (PBS según sus siglas en inglés) para Prolina (AGG) y una señal de polipurinas en la posición 4234 en las cuatro secuencias, datos que también apuntan a la existencia de actividad en estos elementos.

Los resultados obtenidos al realizar *BlastN* contra las bases Tf y RB indican que este elemento no fue caracterizado previamente y los datos de *tBastX* en RB enfatizan el hecho de tratarse de un elemento perteneciente a la superfamilia *Copia*, no identificado previamente.

Según el programa *Repeat Masker* [**Kohany et al., 2006**] la secuencia consenso posee dos regiones a lo largo de su secuencia con identidades a nivel nucleotídico que varían entre 65 y 69% con *Copia*5_AG.

La Figura 7 (página 60) que presenta la filogenia de elementos *Copia* ya detectados en el genoma de *Anopheles* previamente muestra que este elemento tiene una distancia aminoacídica aproximadamente de 46% con *Copia*5_AG, es

decir, se encuentra en el límite de identidad que permitiría incluirlo como miembro de esta familia. Sin embargo, sus LTRs (149 nts de largo) no muestran identidad con los LTRs de la Familia *Copia*5_AG (de 108 nts de largo), lo que sugiere que se trata de familias diferentes.

Cuando analizadas las secuencias pertenecientes a los elementos *Copia* previamente caracterizados en *Anopheles gambiae*, las distancias-p aminoacídicas entre las diferentes familias *Copia* poseen una media de 0,6696 (ds 0,0575) que no difiere de las distancias de la secuencia consenso del cluster 134 con cada una de estas familias (media 0,6299; ds 0,1228).

El análisis de las substituciones en posiciones no sinónimas (dN) en relación a aquellas en posiciones sinónimas (dS) del ORF en la secuencia consenso de este cluster, mostró que estas secuencias presentan seis veces más substituciones en posiciones sinónimas que en posiciones no-sinónimas (ω=0,1667) indicando que existe un presión negativa o purificadora sobre ellas. Ya el test de Tajima para el análisis de evolución neutral de las secuencias en este cluster no fue significativo (Tabla 4) es decir, no podemos descartar la hipótesis nula que lo elementos se encuentren evolucionando de forma neutral. Esto indicaría que estas secuencias no se encuentran activas, sin embargo, el hecho de tratarse de un cluster con solo cuatro secuencias con variaciones muy pequeñas entre si (la distancia-p=0,0009 y existen tan solo cuatro sitios segregantes a lo largo de todo el alineamiento) puede influir en el poder del test para detectar la existencia de presiones selectivas actuando sobre estas secuencias.

Por su parte, no se detectaron indicios de expresión, ya que se obtuvieron *matches* negativos tanto en la biblioteca de EST como en la de mRNA (Ver resultados presentados en la base de datos AnoTExcel). Siguiendo la numeración y nomenclatura utilizadas en RepBase se propone la denominación de este elemento como ***Copia*-6_AG**. Se concluye que este cluster de secuencias corresponde a una familia joven, de pocos elementos aparentemente activos en el genoma de *Anopheles gambiae.*

Cluster 149

Este cluster está compuesto por cuatro secuencias de aproximadamente 2245 nucleótidos de largo presentando LTRs de 168 nucleótidos de largo y un marco de lectura abierto de 1845 nucleótidos que no presenta ningún dominio conservado (Figura 15).

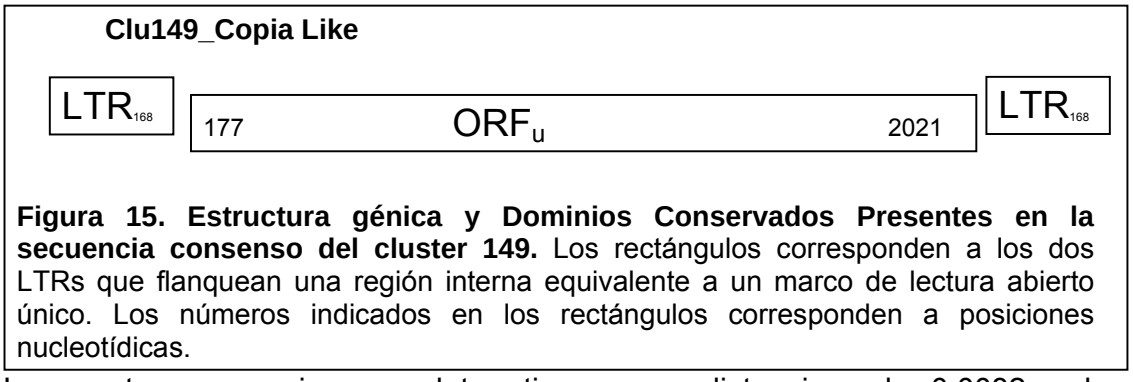


Figura 15. Estructura génica y Dominios Conservados Presentes en la secuencia consenso del cluster 149. Los rectángulos corresponden a los dos LTRs que flanquean una región interna equivalente a un marco de lectura abierto único. Los números indicados en los rectángulos corresponden a posiciones nucleotídicas.

Las cuatro secuencias completas tienen una distancia-p de 0,0082 y los respectivos LTRs 3' y 5' (de 168 Nts de largo) de cada secuencia presentan una distancia-p media 0,0168. El calculo basado en la distancias medias entre los ORFs permite inferir que este elemento fue activo más de 500.00 años atrás.

Las inserciones de estos cuatro elementos se presentan en los cromosomas X, 2L y 3R. El programa LTR-finder detectó la presencia de dos PBS diferentes en

las secuencias, tres de ellas presentan sitios para Valina (AAC) mientras el cuarto presenta para Metionina (CAT), estas diferencias en los miembros de una misma familia es indicativo de una posible ausencia de actividad de la misma. Los resultados obtenidos al realizar *BlastN* contra las bases Tf y RB indican que este elemento no fue caracterizado previamente y los datos de *tBastX* en RB indican que este elemento puede tratarse de un elemento perteneciente a la superfamilia *Copia*. Según el programa Repeat Masker, la secuencia consenso no posee regiones de identidad con ningún elemento depositado previamente en Repbase. Tanto en la filogenia basada en todos los elementos LTRs (Figura 6) como aquella basada únicamente en elementos *Copia* presentes en el genoma de *Anopheles*, (Figura 7) el elemento correspondiente al cluster 149 se encuentra considerablemente distante del resto de los elementos *Copia* presentes en el genoma de *Anopheles* (las distancias p con los elementos *Copia* 1 a 5 son respectivamente de 0,8126; 0,8169; 0,7712; 0,8456 y 0,8376). Estas distancias aminoacídicas son significativamente mayores que el conjunto de todas las distancias entre las cinco familias de elemento *Copia*. Sin embargo, comparte un valor de *bootstraping* de 100 con el resto de los elementos de esta superfamilia (Figura 6). El análisis de dN/dS mostró un valor de 0,6644, es decir que la presión es relajada, por su parte el valor de Tajima si bien fue negativo, no fue significativo para este conjunto de secuencias. En este caso, como con el cluster 149 contamos también con pocas secuencias lo que puede interferir con la significancia de este test estadístico.

Las características generales de esta familia indican que se trata de una familia no activa y que, como el caso de todos los otros miembros de la superfamilia

Copia, está conformada por pocos elementos. Por otra parte, en el caso de esta familia obtuvimos *match* positivo con la biblioteca de EST y de mRNA (Ver datos en AnoTExcel y tabla 4) indicando que el elemento es transcripto y también traducido a mRNA. La secuencia de este elemento tiene 99% de identidad en más de 1200 nucleótidos con un cDNA obtenido de *Anopheles gambiae*.

Siguiendo la numeración y nomenclatura utilizadas en RepBase se propone la denominación de este elemento como ***Copia*-7_AG**.

Superfamilia *Pao-Bel*

Cluster 174

Este cluster esta compuesto por tres secuencias de 5769 nucleótidos de largo, con LTRs de 213 nts que poseen una distancia-p global de 0,0068, dos de ellas presentan ambos LTRs mientras que la tercera presenta únicamente el LTR de la región 3'.

La distancia calculada a partir de los LTRs 5 y 3' de las dos secuencias completas fue de 0,014. El calculo de tiempo desde la transposición basado en la identidad entre los LTRs de estas secuencias indica un tiempo de aproximadamente 450.000 años. La secuencia consenso presenta un único marco de lectura abierto de 5274 nucleótidos (1758 aa) con dominios conservados para RT-PepA17 y RVE (Integrasa) lo que indica que puede tratarse de secuencias activas (Figura 16). Sin embargo, cabe señalar que el programa *LTR-Finder* no detectó ninguna característica compatible con un elemento LTR. Las inserciones de estos cuatro elementos se presentan en los cromosomas 2L y 3R. Los resultados obtenidos al realizar *BlastN* contra las

bases Tf y RB indican que este elemento no fue caracterizado previamente sin embargo los resultados de *tBastX* en RB indican que este elemento pertenece a

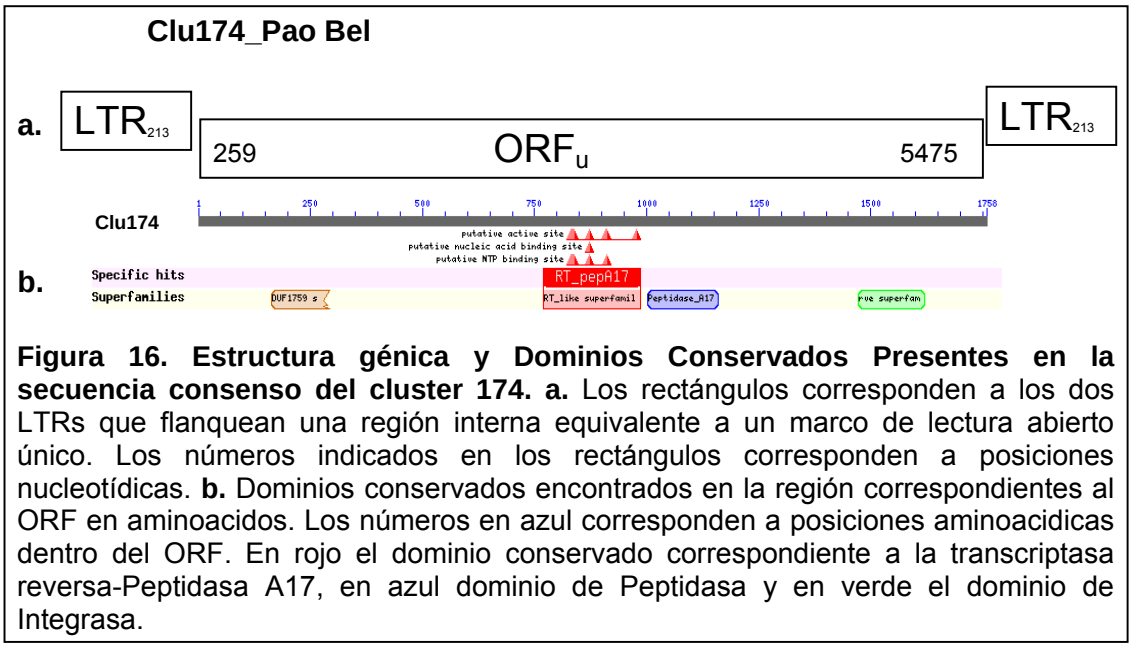


Figura 16. Estructura génica y Dominios Conservados Presentes en la secuencia consenso del cluster 174. a. Los rectángulos corresponden a los dos LTRs que flanquean una región interna equivalente a un marco de lectura abierto único. Los números indicados en los rectángulos corresponden a posiciones nucleotídicas. **b.** Dominios conservados encontrados en la región correspondientes al ORF en aminoacidos. Los números en azul corresponden a posiciones aminoacidicas dentro del ORF. En rojo el dominio conservado correspondiente a la transcriptasa reversa-Peptidasa A17, en azul dominio de Peptidasa y en verde el dominio de Integrasa.

la superfamilia *Pao-Bel* ya que presenta un poco más de 50% de identidad a nivel aminoacídico con elementos *Bel* conocidos en el genoma de *Anopheles gambiae* en diferentes regiones a lo largo de su secuencia. Según el programa Repeat Masker, la secuencia consenso posee regiones de identidad (entre 63-65%) dispersas a lo largo de su secuencia con *Bel*16_AG. En la filogenia basada en todos los elementos LTRs (Figura 6, pagina 58) el elemento correspondiente al cluster 174 se agrupa con los elementos *Bel*, precisamente con *Bel*16_AG con un alto valor de bootstraping y la distancia que presenta con esa familia a nivel aminoacidico es del 56% (Figura 8, página 62). Por su parte, los LTRs de *Bel*16 y de las secuencias del cluster 174 tienen diferentes tamaños y baja identidad, indicando que no se trata de secuencias de la misma familia. El análisis de dN/dS mostró un valor de 0,2093 indicando una presión de selección negativa

sobre el ORF en esta secuencia. El test de Tajima no pudo ser ejecutado debido a que este cluster tiene tan solo tres secuencias.

Las características generales de esta familia indican que se trata de una familia activa de elementos *Bel* con pocos elementos. Siguiendo la numeración y nomenclatura utilizadas en RepBase se propone la denominación de este elemento como ***Bel*20_AG**.

Cluster 185

Este cluster se encuentra compuesto por tres secuencias de 3803 nucleótidos de largo, con una distancia-p de 0,0033 y con LTRs de 227 nucleótidos que presentan una distancia-p global de 0,0143. Esta distancia indica un tiempo desde la transposición de aproximadamente 500.000 años en el genoma de *Anopheles*.

La secuencia consenso tiene un marco de lectura abierto de 1845 nucleótidos que no presenta dominios conservados para proteínas conocidas (Figura 17).

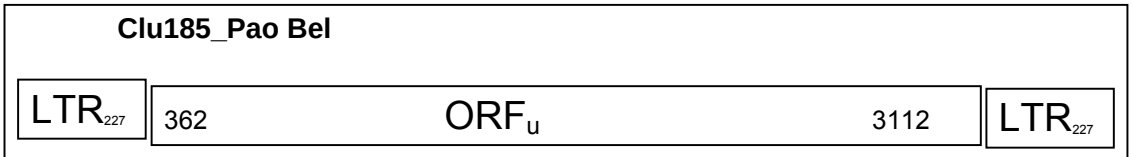


Figura 17. Estructura génica y Dominios Conservados Presentes en la secuencia consenso del cluster 185. Los rectángulos corresponden a los dos LTRs que flanquean una región interna equivalente a un marco de lectura abierto único. Los números indicados en los rectángulos corresponden a posiciones nucleotídicas.

Las inserciones de estos cuatro elementos se presentan en los cromosomas X, 2L y 3R. Como en el caso del cluster 174 descrito anteriormente, el programa LTR-Finder no detectó ninguna característica particular de elementos LTR en la secuencia consenso. Los resultados obtenidos al realizar *BlastN* contra las

bases Tf y RB indican que este elemento no fue caracterizado previamente sin embargo los resultados de *tBastX* en RB indican que pertenece a la superfamilia *Pao-Bel* ya que presenta alrededor de 50% de identidad con elementos *Bel* conocidos en el genoma de *Anopheles gambiae* en diferentes regiones a lo largo de su secuencia. El programa *Repeat Masker* (RM) indica que la secuencia consenso posee dos regiones de aproximadamente 500 nucleótidos con identidades de 62 y 63% con *Bel*-18_AG. En la filogenia de la figura 6 (pagina 58) la secuencia consenso de esta familia se agrupa con los elementos *Copia*, pero sin valor de bootstraping que lo valide. Teniendo en cuenta que según el programa RM y según el BlastX realizado en RB esta secuencia es semejante a un elemento de la superfamilia *Bel*, analizamos esta secuencia en una filogenia junto a otros miembros de esta familia (Figura 8, pagina 62). En esta figura se agrupa con 99% de *bootstraping* con los elementos *Bel* 18, 3, 14 y 17_AG y con 92% con la familia *Bel*18_AG. La distancia aminoacídica con esta familia es de 0,5435. Esta distancia es compatible con las distancias existentes entre miembros pertenecientes a diferentes familias dentro de esta superfamilia (media = 0,7023 con un rango de 0,3650 a 0,8389). Las substituciones en posiciones sinónimas y no sinónimas están presentes en la misma proporción, dando como resultado un ω=1, lo que indica que estas secuencias se encuentran evolucionando de forma neutra.

Las características generales de esta familia indican que se trata de una familia no activa de pocos elementos. Siguiendo la numeración y nomenclatura utilizadas en RepBase se propone la denominación de este elemento como ***Bel*-21_AG**.

ELEMENTOS CLASE II NUEVOS

Secuencias tipo MITEs de TEs identificados previamente

Un total de 16 clusters de AnoTExcel fueron clasificados como MITEs. Una parte de ellos (siete clusters) consiste en elementos MITEs ya descritos previamente y actualmente depositados en RB[2] que corresponden en su totalidad a MITEs asociados a elementos P [**Quesneville et al., 2006**].

De los elementos restantes clasificados como MITEs en AnoTExcel, cuatro (clusters 49; 189; 85 y 57) tienen TIRs que poseen alta identidad con parte de los TIRs de elementos de la familia *Gambol* (elementos *Gambol_6 –Parker, Gambol_2 –Ozzie, Gambol_3 -Whistler y Gambol_10 –Pi-Pi*, respectivamente descritos en Tefam) (Figura 18). Estas familias pertenecen a la superfamilia Tc1-Mariner [**Coy & Tu, 2005**]. Las secuencias identificadas en AnoTExcel y relacionadas a estos elementos tienen estructuras genéticas particulares (Figura 18) correspondientes a fragmentos de la región de TIRs de los elementos completos. El cluster 189 tiene alta identidad con la región 5' del TIR del elemento *Gambol_Ele2*, pero la totalidad de su secuencia, de 675 nucleótidos, aparenta ser un TIR ya que la versión reversa y complementaria produce la misma secuencia (Figura 18a).

Los clusters 85 y 49 corresponden a elementos que presentan TIRs, con alta identidad con una parte de los TIRs de elementos *Gambol* (Figuras 18b y c). Esto lleva a pensar que estos elementos pueden replicar como MITEs utilizando

[2] Los clusters 91, 123, 226, 7, 28, 154 y 37, todos ellos clasificados como MITEs de elementos P fueron descritos por **Quesneville et al., 2006**. Los *BlastN* y *BlastX* en Anotexcel no resultan en *matches* positivos porque estos elementos fueron depositados en esta base posteriormente a la generación de AnoTExcel. De cualquier manera, todos ellos, salvo el cluster 154 fueron identificados por estar depositados en la base de datos no redundantes del NCBI.

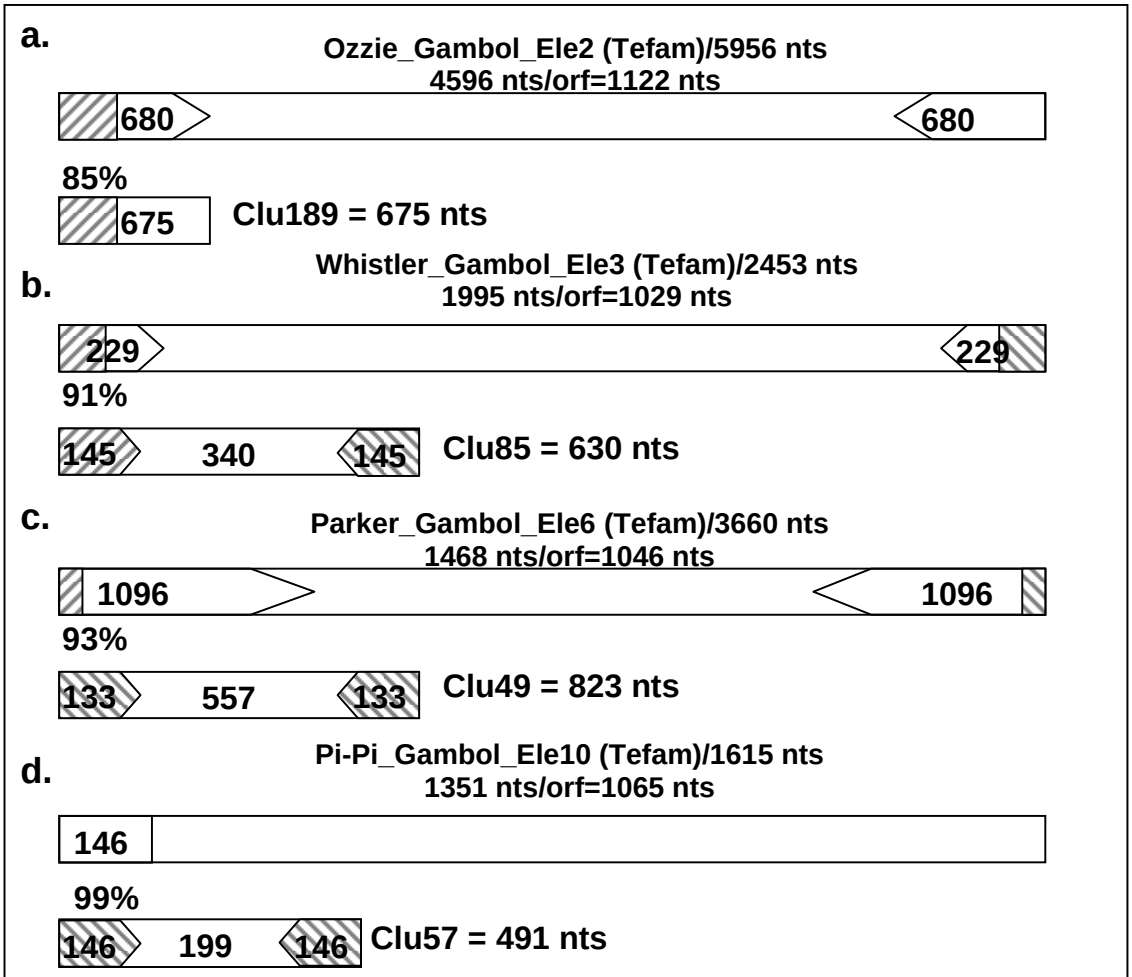


Figura 18. Estructura génica de cuatro elementos de tipo MITE relacionados a elementos de la familia *Gambol*. La estructura genética de los elementos completos depositados en Tefam se encuentra en la parte superior. Los TIRs se esquematizaron con flechas y las regiones de identidad con rectángulos rallados. (a.) Cluster 189 (b.) Cluster 85 (c.) Cluster 49 (d.) Cluster 57.

la transposasa de los elementos activos, pero tan solo con una pequeña región del TIR del elemento original para ser reconocida. Estos clusters no son altamente abundantes, presentan 9 y 19 secuencias respectivamente, que tienen entre si distancias-p nucleotídicas de 0,053 (ds=0,0049) para el cluster 85 y de 0,0373 (ds=0,0033) para el cluster 49. Presentando distancias aun menores en las regiones correspondientes a los TIRs (0,0315 –ds=0,0055- para el cluster 85 y 0,0271 –ds=0,0051- para el cluster 49).

Por ultimo, el cluster 57 clasificado originalmente como MITE debido a su pequeño tamaño y a la presencia de TIRs conservados presenta 99% de identidad con la región 5' del elemento *Gambol_10* (Figura 18d). Sin embargo, esta región de la secuencia de este elemento Gambol no parece corresponder con un TIR. En este caso un escenario posible para explicar esta estructura génica podría ser una recombinación ectópica entre dos elementos *Gambol* con diferentes sentidos (*sense* y *antisense*) seguido de una deleción interna.

Esto muestra la existencia de una diversidad de elementos deteriorados de tipo MITE (o productos derivados de MITEs) que eventualmente podrían formar parte de la dinámica de esta familia así como de otras familias de elementos de clase II.

Por otra parte, el cluster 21 (Figura 19) según RB, corresponde al elemento Marina, descrito en el orden de insectos *Endopterygota*. Sin embargo, según Tf

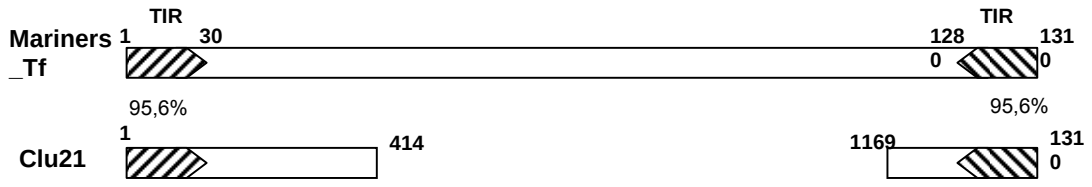


Figura 19. Estructura génica de la secuencia consenso correspondiente al cluster 21 en AnoTExcel y de los elementos Mariner _ele 5, 6, 8, 12, 15, 16, 17 y 18 depositados en TF. Los números corresponden a las posiciones nucleotídicas y los porcentajes indican el grado de identidad entre las regiones indicadas con rallas.

tiene una alta identidad con un conjunto de elementos *Mariners* denominados *Mariner_ele 5, 6, 8, 12, 15, 16, 17 y 18* que tienen una distancia-p media entre si de 0,0370. Las familias *Mariner_ele 5* y 6 son versiones truncadas, que mantienen solo 412 nucleótidos de la región 3'. Estas secuencias *Mariner* contienen, además, TIRs con altas identidades (97,33%), lo que lleva a pensar

que se trata de secuencias que corresponden a una misma familia. Los elementos *Mariners* mencionados tienen un tamaño promedio de 1300 nts, poseen capacidad codificante y son elementos activos según la información presentada en Tf. La secuencia consenso del cluster 21 de AnoTExcel -con 39 secuencias- tiene un tamaño de 560 nts y mantiene conservadas las regiones 3 y 5' (con una identidad de 94,3% entre si) con excepción de algunas secuencias que solo presentan la región 5' o 3'. La distancia-p entres los TIRs de los elementos *Mariners* y los TIRs en las secuencias del cluster 21 es de 0,0446 (ds=0,0082) y son perfectamente alineables y del mismo tamaño. Todas estas características indican que las secuencias del cluster 21 corresponden a un elemento de tipo MITE relacionado a los elementos *Mariners* antes mencionados y presentes en Tf.

El Cluster 9 esta compuesto por 72 secuencias, tres de las cuales tienen una inserción de 205 nts con alta identidad con el elemento MITE *AgaP12MITE205B* (distancia – p = 0,0275 sd=0,0076) [Quesenville, et al 2006] (Figura 20).

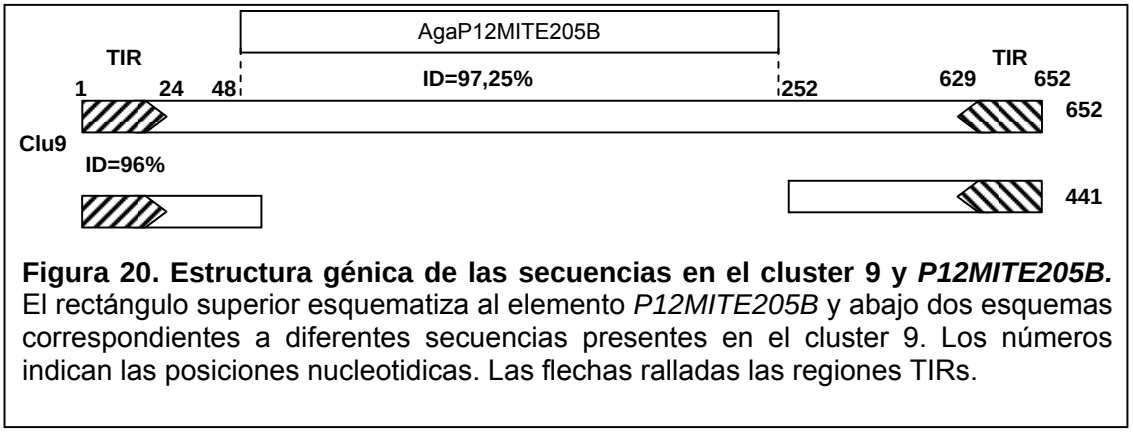


Figura 20. Estructura génica de las secuencias en el cluster 9 y *P12MITE205B.* El rectángulo superior esquematiza al elemento *P12MITE205B* y abajo dos esquemas correspondientes a diferentes secuencias presentes en el cluster 9. Los números indican las posiciones nucleotidicas. Las flechas ralladas las regiones TIRs.

El resto de las secuencias de 441 nucleótidos de largo presentan solamente las regiones 3 y 5' con alta identidad con las primeras, donde se encuentran presentes TIRs de 24 nucleótidos que no pertenecen a ningún elemento de

transposición conocido. Aparentemente, se trata de un elemento MITE de la familia *P* que adquirió TIRs de 24 nucleótidos que no se encuentran relacionados a TIRs de ningún elemento clase II conocido.

Por último, el cluster 35 consta de 24 secuencias de 447 nts de largo y con TIRs de 40 nucleótidos que tienen alta identidad (dist p=0,0708) con los TIRs presentes en el elemento *Harbinger1_AG* (Figura 21). Se han identificado MIITEs correspondientes a este elemento en otros genomas [Grzebelus, et al 2006, Yang & Hall, 2003] aunque no en el genoma de *Anopheles*.

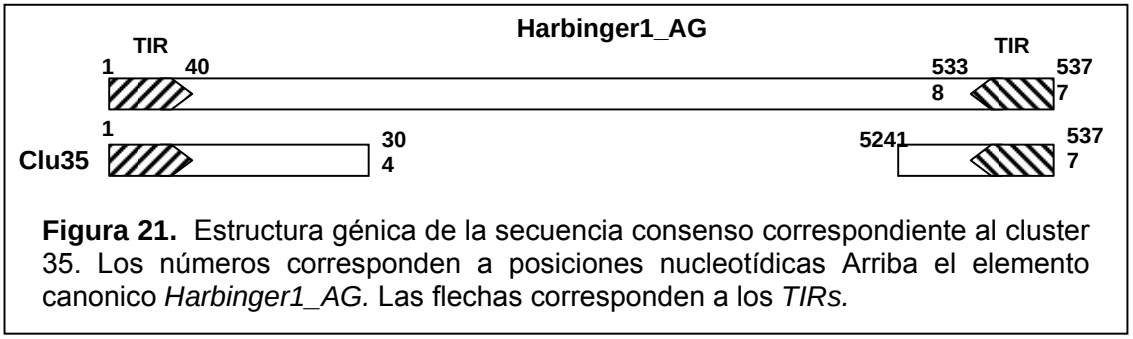


Figura 21. Estructura génica de la secuencia consenso correspondiente al cluster 35. Los números corresponden a posiciones nucleotídicas Arriba el elemento canonico *Harbinger1_AG.* Las flechas corresponden a los *TIRs.*

Elementos clase II no identificados previamente

Un total de 32 clusters cuyas secuencias presentan TIRs variando entre 18 y 204 nucleótidos de largo y tamaños de secuencias entre 430 y 1678 nucleótidos fueron identificados en AnoTExcel. Algunos de estos clusters presentan un número considerable de secuencias (Ver información en la Tabla 5 a continuación). Estas secuencias no poseen identidad con otros elementos clase II descriptos previamente ni en las regiones internas ni en los TIRs, lo que llevó a clasificarlos como elementos nuevos.

Los TIRs de estos clusters pueden clasificarse en terminales y subterminales y en algunos casos se trata de secuencias palindrómicas.

Siete de estos clusters (clusters número 166; 144; 240; 168; 137; 197 Y 39) contienen secuencias con más de 1000 nucleótidos pero ninguno de ellos presentan ORFs de tamaños compatibles con la producción de proteínas, con excepción del cluster 240 que presenta un ORF de 603 nts. El análisis de la secuencia aminoacídica de este ORF mostró que no presenta dominios conservados. El tamaño de estas secuencias llama la atención ya que aparentemente son elementos mayores a los clásicos MITEs que tienen en promedio 500 nts de largo y un máximo de 800 nts.

El mecanismo por el cual los elementos de tipo MITE son generados es desconocido. Se desconoce si existe una perdida gradual de la región codificante o si estos elementos son generados en su forma final por una única deleción interna. Podemos imaginar que estos elementos de más de 1000 nts se encuentren en un proceso de degradación que eventualmente lleva a la formación de MITEs menores.

ID	N	TIR	Sec.	TajimaD	SigD	mRNA	EST
Clu38	432	32	23	-1,451	N.S		
Clu12	441	27	64	**-2,5025**	**	+	+
Clu164	441	46	3	n.d.	n.d.		
Clu19	449	18	40	**-1,6763**	#		+
Clu15	472	137	51	**-1,866**	*		
Clu50	509	27	19	**-1,6422**	#		+
Clu52	555	69	18	**-1,9485**	*		+
Clu44	597	18	21	**-1,7042**	#		
Clu31	645	18	26	**-1,7116**	#		+
Clu43	663	29	21	**-2,0299**	*		+
Clu151	674	22	4	-0,6977	N.S		
Clu97	674	171	8	-0,6289	N.S		
Clu27	685	18	30	-1,3225	N.S		+
Clu162	699	20	3	n.d.	n.d.		+
Clu40	716	60	22	**-2,1982**	**		+
Clu66	729	28	14	**-2,0028**	*		
Clu186	751	28	3	n.d.	n.d.		
Clu87	798	31	9	**-1,6083**	#		
Clu111	817	204	6	-1,0545	N.S		+
Clu48	833	34	19	-1,5696	N.S		
Clu121	877	33	5	-1,119	N.S		
Clu11	923	26	65	**-2,267**	**		+
Clu75	947	36	11	**-1,6656**	#		
Clu13	970	30	60	**-1,7505**	#		
Clu25	988	34	32	**-2,3989**	**	+	+
Clu39	**1086**	27	22	**-2,0878**	*	+	+
Clu197	**1087**	33	**3**	n.d.	n.d.		
Clu137	**1176**	68	**4**	-0,5774	N.S		+
Clu168	**1346**	23	**3**	n.d.	n.d.		
Clu240	**1636**	23	**3**	n.d.	n.d.		+
Clu144	**1672**	70	**4**	-0,5558	N.S	+	+
Clu166	**1675**	161	**3**	n.d.	n.d.		+

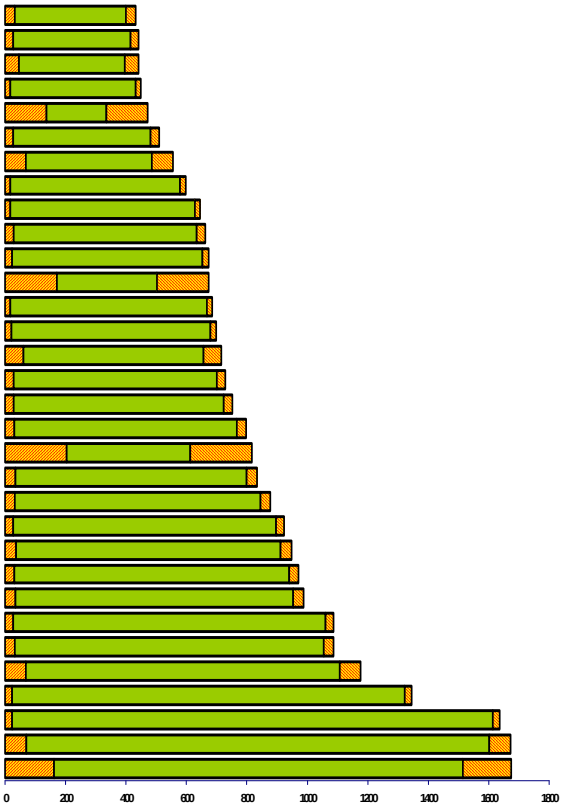


Tabla 5. Características de los elementos de tipo MITE no descritos previamente. Se encuentran resaltados los elementos con más de 1000 nts, y los resultados significativos del test de Tajima. ID: Número de cluster; N: número de nts, TIR: tamaño del TIR; SigD: significancia del test de Tajima. n.d: no determinado; #, *P* , < 0.10; * *P*, < 0.05; * * *P*, < 0.01; *** *P*, < 0.001 A la derecha de la tabla se muestra una representación esquemática de los elementos identificados. En verde, las regiones internas que no poseen identidad con ningún elemento conocido y a rallas rojas y amarillas se representan los TIRs.

Del total de los 32 clusters clasificados como Nuevos, una parte sustantiva (17 en total) muestra *matches* positivos con la base de datos de ESTs (ver AnoTExcel y Tabla 5). El hecho de tratarse de familias de elementos que presentan alta identidad, algunos de ellas bastante numerosos, y también, por ser expresados, indica que estos elementos pueden cumplir un papel dentro de la dinámica del genoma o dentro de la familia de elementos a la cual se encuentran emparentados. Se realizó el test de neutralidad de Tajima para

corroborar si alguna de estas familias se encuentra bajo presión de selección. La tabla 5 muestra los valores del test obtenidos para los 32 clusters. Dieciséis clusters presentan valores significativos, indicando que las secuencias que los componen se encuentran conservadas por selección purificadora y dado que ninguna de estas secuencias tiene capacidad codificante podemos inferir que pueden estar relacionados a alguna función reguladora ya sea de genes del hospedero como de los propios elementos de las familias a la cuales pertenecen. Por su parte, diez de los clusters con valores significativos para el Test de Tajima presentan evidencias de estar expresados.

Resumiendo, la metodología aplicada aquí permitió identificar elementos LTRs y de Clase II no descritos previamente. Los LTR corresponden a elementos completos, y, en algunos casos, con características de actividad y bajo número de copias, lo que sumado a la alta identidad que presentan las secuencias entre si, puede indicar que se trata de familias jóvenes, con reciente introducción en el genoma del mosquito. Por otro lado se identificaron una serie de elementos de clase II no descritos previamente. Estas familias fueron clasificadas como tales debido a la presencia de secuencias repetitivas invertidas con alta identidad entre sí. Algunos de estos elementos parecen ser versiones altamente deterioradas de elementos de clase II previamente descritos (*Gambol*, *Harbinger* o *Mariner*), otros corresponden a elementos que no han sido identificados o que se han perdido del genoma de *Anopheles*. Estos últimos elementos poseen TIRs que no presentan identidad aparente con ningún TIR previamente descrito, lo

que lleva a cuestionar la forma en que los mismos replican dentro del genoma ya que no podemos identificar una transposasa capaz de reconocerlos.

Por otra parte, se identificaron elementos tipo MITE que solo mantienen identidad con una pequeña región de los TIRs de elementos identificados previamente, como es el caso de los posibles MITEs de elementos *Gambol* reportados aquí. Esto lleva a pensar en la posibilidad de que estos elementos sean capaces de replicar de forma parasítica en el genoma con tan solo pequeñas regiones de los TIRs presentes en los elementos completos.

Los elementos de clase II no poseen capacidad codificante ni características de actividad. Sin embargo, muchos de estos elementos tienen secuencias nucleotídicas que indican que se encuentran bajo presión de selección positiva, y corresponden, además, a secuencias expresadas.

La visión de los elementos de transposición como parásitos genómicos viene siendo cuestionada hace mucho tiempo y varios trabajos dan cuenta de la contribución de estos elementos en la evolución de mecanismos de regulación génica así como también en la creación de estructuras génicas novedosas (para revisión ver **Feschotte, 2008**).

La existencia de TEs, sin capacidad codificante, que mantienen TIRs y secuencias conservadas y que además se expresan, levanta la posibilidad de que los mismos contribuyan con algún rol en el genoma o como elementos regulatórios dentro de la red de elementos de una misma familia.

Cabe señalar, que todos los elementos identificados como clase II nuevos contienen TIRs pero ninguno de ellos presenta homología con elementos clase II

conocidos (resultados no mostrados aquí) por lo tanto, de jugar un papel en la regulación de una familia de transposones, puede pensarse que o bien son reconocidos por enzimas transposasas de otros elementos, o de elementos aún no identificados en este genoma.

Es evidente que existe una gran diversidad de elementos de transposición con diferentes grados de degradación. Algunos presentan características de MITEs, es decir pequeño tamaño y alta identidad y conservación de los TIRs relacionados a elementos clase II completos. Otras secuencias identificadas aquí, sin embargo, contienen solamente parte de los TIRs identificados en elementos de transposición previamente caracterizados. Esto puede indicar que, por lo menos para el caso de estas familias, tan solo una parte del TIR es esencial para el reconocimiento de la transposasa y la transposición del elemento.

EVOLUCIÓN DE TES EN EL GENOMA DE *Anopheles gambiae*

Los elementos de transposición (TEs) se encuentran presentes en prácticamente todos los genomas de organismos eucariotas. Varios proyectos de secuenciación de genomas completos en diversas especies que han sido finalizados en los últimos tiempos han confirmado la naturaleza ubicua de los TEs en los genomas eucariotas así como también su presencia en algunos procariotas [algunos artículos de revisión al respecto que pueden ser consultados: **Brookfield 2005; Feschotte *et al*., 2002 y Kidwell, 2002**]. Estos elementos no solo se encuentran presentes en estos genomas sino que representan, en algunos casos, un alto porcentaje de los mismos. Este dato fue sorprendente ya que evidenció que la cantidad de DNA presente en un determinado organismo no es directamente proporcional a la complejidad del mismo, así como tampoco lo es el número de genes. Este DNA aparentemente "excedente" en algunas especies se ha relacionado a DNA sin funciones aparentes, como el DNA repetitivo. Por ejemplo, los genomas de *Anopheles gambiae* [**Holt et al., 2002**] y *Aedes aegypti* [**Nene et al., 2007**] tienen, respectivamente, 286 Mb y 1.3 Gb en tamaño, lo que no redunda en una diferencia de complejidad ni en diferencias significativas en el número de genes. Por su parte, el genoma de *Aedes aegypti* contiene mayor contenido de elementos de transposición que el genoma de *Anopheles gambiae* (50 y 16 %

de estos genomas están ocupados por elementos de transposición respectivamente) [**Nene et al., 2007; Holt et al., 2002**].

Si bien los TEs determinan en muchos casos diferencias significativas en el contenido de DNA genómico de especies muy próximas evolutivamente, la mayor parte de los TEs identificados en los genomas actuales corresponde a secuencias degeneradas, o sea remanentes de elementos de transposición activos[3] [**Feschotte *et al*., 2002**]. En las palabras de Feschotte y Pritham [**2007**] "…una porción importante de la mayor parte de los genomas eucariotas multicelulares e incluso de algunos genomas unicelulares es un enorme cementerio de elementos de transposición". En un análisis realizado con los elementos de transposición presentes en los genomas de *Drosophila melanogaster* y *Anopheles gambiae* [**Quesneville et al., 2003**] se encontró que la media del tamaño de las secuencias de TE alineadas a los elementos completos era de 12 y 24%, respectivamente. Es decir, los elementos presentes actualmente en estos genomas, en media, corresponden a pequeños fragmentos de los elementos originales, evidenciando una alta deterioración de los mismos, principalmente por medio de deleciones.

La historia evolutiva de un dado TE puede resumirse como un proceso que consiste principalmente en tres etapas: invasión, amplificación y senescencia. Si bien estas etapas fueron propuestas como un modelo para explicar el ciclo de

[3] En muchos casos los ancestrales activos de una familia de TEs no se encuentran presentes en los genomas actuales y las secuencias consensos han sido inferidas a partir de secuencias remanentes de TEs hallados en los genomas actuales. Dos casos de particular interés en este aspecto son los elementos Sleeping Beauty (SB) [**Ivics, et al., 1997**] y Prince Frog (PR) [**Miskey et al, 2003**]. La secuencia de los ancestros de SB y PR fueron deducidas a partir de secuencias de elementos no activos (corregidas para actividad) obtenidas de los genomas de peces salmonídeos y de *Rana pipiens*, respectivamente. Las secuencias consenso obtenidas a partir de los fragmentos hallados y su generación en el laboratorio permitió la "resurrección" de los mismos como elementos activos.

vida de elementos clase II [**Hartl et al., 1997**] también pueden pensarse como las etapas correspondientes al ciclo de vida de elementos de clase *I* e inclusive de retrovirus. Esta comparación es válida si consideramos a las tres etapas del ciclo en sí, pero cada una de ella difiere significativamente entre ambas clases. Es posible afirmar que aquellos elementos presentes en los genomas actuales, ya sea como formas completas y activas o como secuencias remanentes o deterioradas, son aquellos elementos que fueron capaces de establecer una estrategia de invasión y permanencia exitosa en algún momento de su camino evolutivo. La invasión de un genoma (considerada como la aparición de un elemento de transposición en el genoma de una especie donde anteriormente no estaba presente pudiendo darse por transmisión horizontal o por recombinación o reactivación de un elemento degradado) es un paso crítico en la evolución de un TE [**LeRouzic & Capy, 2005**]. En esta primera fase operan las presiones selectivas más importantes, posteriormente la evolución de un TE progresará de forma neutra. El proceso de invasión de un elemento es así visto como un equilibrio entre la capacidad de amplificación del elemento y la tolerancia del genoma hospedero a la presencia y replicación de estos elementos. En la etapa de amplificación puede ocurrir la transmisión horizontal a otro genoma, lo que asegura la subsistencia de esta familia. Si bien existen evidencias de transmisión horizontal entre especies diferentes, las mismas son indirectas y el mecanismo por el cual ocurre es desconocido. Finalmente, el proceso de deterioración de las secuencias de estos elementos lleva a la pérdida de identidad de las mismas (esencialmente debido a la falta de una presión purificadora sobre la integridad y funcionalidad de estos elementos).

La propia historia evolutiva de los elementos de transposición sumada a la ausencia de una presión para la eliminación de elementos deteriorados siempre que los mismos no interfieran con el funcionamiento de genes o de regiones regulatórias, permite la coexistencia de elementos activos (productores de nuevos eventos de transposición) con elementos inactivos (algunos incapaces de movilidad y otros que pueden ser movilizados por formas activas) en un mismo genoma, lo que produce un complejo entramado de interacciones y dinámicas posibles (entre los elementos activos e inactivos, autónomos y no autónomos) dentro de los miembros de una determinada familia. Existen diversos trabajos teóricos que pretenden explicar las dinámicas evolutivas de los elementos de transposición [**Charlesworth & Langley, 1989; Brookfield & Badge, 1997; Badge & Brookfield, 1997; Quesneville & Anxolabehère, 1997; 1998; Biemont et al., 1997**]. Sin embargo, no hay hasta el momento, un panorama claro del proceso de deterioración de los elementos de transposición en un determinado genoma, lo cual es importante para entender de qué manera estos elementos consiguen perpetuarse y mantenerse en los genomas así como de que manera son regulados y eventualmente cómo son eliminados de los genomas. Comprender estos procesos es también importante en relación a la generación de organismos transformados genéticamente. En la concepción de este tipo de abordaje deben considerarse la posible evolución del elemento en cuestión, así como comprender la forma en que este elemento se deteriora hasta su eventual extinción. Una pregunta importante es saber si existe un mecanismo de deterioración único o preponderante para las diferentes clases o familias de elementos que coexisten en un genoma o, si por el contrario, la

deterioración depende del tipo de elemento considerado. Por otra parte, la existencia de subpoblaciones de elementos dentro de una misma familia puede reflejar la existencia de eventos de transposición en diferentes momentos históricos de ese elemento, o por otra parte indicar la existencia de diferentes invasiones genómicas a lo largo del tiempo. La posibilidad de re-invasión de un genoma por un determinado elemento tiene, por su parte, implicaciones en un proyecto de generación de organismos transgénicos utilizando TEs como drivers genéticos ya que estos nuevos "invasores" podrían jugar un papel imprevisto en la dinámica de regulación de una familia ya establecida en el genoma.

Se sabe que los mecanismos de transposición y concomitantemente, los procesos de deterioración de cada clase de TEs son diferentes. Los NLTRs sufren un proceso de deterioración por deleciones en la región 5' debido al propio mecanismo de transposición, generando elementos "*Death on Arrival*" (DoA) que son versiones truncadas de los elementos completos **[Petrov et al., 1996, 1997, 1998]**. Los elementos LTRs pueden perder la región interna por recombinación entre sus LTRs formando lo que se conoce como Solo-LTRs. Por otra parte, los elementos de clase II, más pequeños y con un mecanismo de transposición de tipo "corte y copiado" tienden a incorporar mutaciones en el proceso de copiado de la hebra de DNA dadora.

Previamente, se han descrito una serie de formas aparentemente no autónomas de elementos de transposición en diversas clases y familias de TEs. Por ejemplo, dentro de los LTRS, se han identificado elementos denominados LARD (*Large Retrotransposon Derivatives*) [**Kalendar el al., 2004**] en gramíneas del género *Triticeae*. Estos elementos poseen largos LTRs y un dominio interno

conservado que no posee capacidad codificante. Son, por lo tanto, no autónomos pero, sin embargo, son transcriptos y poseen polimorfismos en relación a sus sitios de inserción, indicando que de alguna forma amplifican su representación en el genoma [**Sabot & Schulman, 2006**]. Los elementos *Morgane* [Sabot et al., 2006] identificados en el genoma del trigo, tampoco poseen capacidad codificante, pero se ha encontrado una alta identidad nucleotídica, con ESTs correspondientes a otros elementos pequeños que se expresan en condiciones de stress. Los elementos conocidos como TRIM (*Terminal Repeat Retroelements in Miniature*) son, por su parte, altamente compactos, poseen pequeños LTRs y tienen regiones internas conservadas, sin capacidad codificante pero con señales para la transicripción reversa [**Witte et al., 2001**].

Dentro de los NLTRs, los elementos denominados SINE (*Short Interspersed Elements*) son versiones no autónomas que utilizan la maquinaria de elementos LINE (*Long Interspersed Elements*) para lograr su propia amplificación. Estos elementos son altamente exitosos en relación al número de copias que consiguen alcanzar, especialmente en algunos genomas (como las secuencias *Alu* en humanos) aún actuando como formas parasíticas de otros elementos.

En la clase II también existen formas no autónomas, deterioradas, sin capacidad codificante, que sin embargo, replican de manera parasítica en el genoma y llegan en algunos casos a tener números de copias muy elevados. Los denominados MITE (*Miniature Inverted Repeat Transposable Elements*) poseen extremos conservados correspondientes a las regiones terminales (TIRs) y un tamaño pequeño, usualmente menor a 500 nucleótidos. Los MITEs utilizan de

forma parasítica a las transposasas codificadas por elementos activos. Recientemente se describieron en el genoma de *Triticeae* elementos denominados SNAC (*Small Non-Autonomous CACTA elements*) [**Wicker et al., 2003**]. Estos elementos poseen pequeños TIRs y regiones internas que no codifican para ninguna proteína.

Todos estos ejemplos representan elementos deteriorados que deben explotar de forma parasítica las enzimas de otros elementos para poder replicar y propagarse en el genoma. La presencia de estos elementos deteriorados, no autónomos que dependen de las maquinarias de los elementos activos pueden afectar el ciclo de vida de una familia de transposones. Por otra parte, el parasitismo tanto de elementos autónomos pero inactivos, como de elementos no-autónomos sobre las copias activas y autónomas también puede afectar la dinámica poblacional de una determinada familia de elementos en el tiempo. Esta variedad de copias inactivas descritas para diferentes familias debe jugar, sin lugar a dudas, algún papel en la dinámica de la o las familias con las cuales interactúan, sea regulando la actividad de las copias activas, compitiendo por la actividad enzimática, o por otros mecanismos. De esta forma al pensar en las familias de elementos de transposición debemos pensar en esta compleja red de interacciones y no en copias individuales.

En esta sección se analizó la evolución y, específicamente, la degradación de diversas familias de elementos representando a las diferentes clases y órdenes de TEs presentes en AnoTExcel.

Este análisis incluyó, por un lado el estudio de la degradación estructural para las diferentes clases y ordenes como un todo (LTRs, NLTRs y Clase II) y por

otro, un estudio más detallado de algunas familias escogidas de AnoTExcel teniendo en cuenta principalmente la presencia de elementos completos y degradados dentro de la misma familia con la intención de poder entender las relaciones entre las diferentes formas de elementos dentro de una familia. En total fueron analizadas siete familias de elementos LTRs (una perteneciente a la superfamilia *Copia*, tres a la *Bel* y tres a la *Gypsy*), siete familias de elementos NLTRs (cuatro pertenecientes a la superfamlia *CR1*, y una perteneciente a cada una de las superfamilias *Jockey*, *Outcast* y *RTE*, respectivamente) y nueve familias de elementos de Clase II (tres pertenecientes a la superfamilia *Tc1-Marriner,* dos a la *hAT,* dos a la *Harbinger*, y una perteneciente a cada una de s superfamilias *Transib, Ikirara* y *Helitron*.

También, se realizaron análisis filogenéticos (*Neighbor Joining*) con el objetivo de detectar la presencia de estructuras poblaciones diferenciables, sea entre secuencias con diferentes grados de deterioración dentro de la misma familia, o entre elementos aparentemente autónomos y no-autónomos de la misma familia. Por ultimo, se propone un análisis de las relaciones entre las diferentes secuencias de la misma familia basado en estudios de redes –*Median Joining Network Analysis*- (Cuadro 1) con el objetivo de entender las relaciones entre las secuencias estudiadas incluyendo la posibilidad de algunas secuencias ser ancestrales de otras. Este tipo de estudio permite, por una parte, entender las relaciones entre los diferentes elementos que componen una familia en términos de ancestralidad y también de estructura poblacional, y por otra, inferir qué tipo de modelo de transposición se ajusta a una determinada familia (Ver Cuadro 2).

Cuadro 1. Análisis de Network y Métodos tradicionales de análisis filogenético

Los métodos tradicionales de análisis filogenético fueron originalmente diseñados para investigar relaciones interespecíficas que tienen la característica de ser jerárquicas ya que son el producto de aislamiento reproductivo durante largos periodos de tiempo. Este tipo de relaciones pueden ser representadas por filogenias estrictamente bifurcadas (donde cada ancestro se divide en dos descendientes). En este caso, todas las secuencias muestreadas ocupan realmente los extremos de las ramas y los ancestros no son muestreados sino reconstruidos.

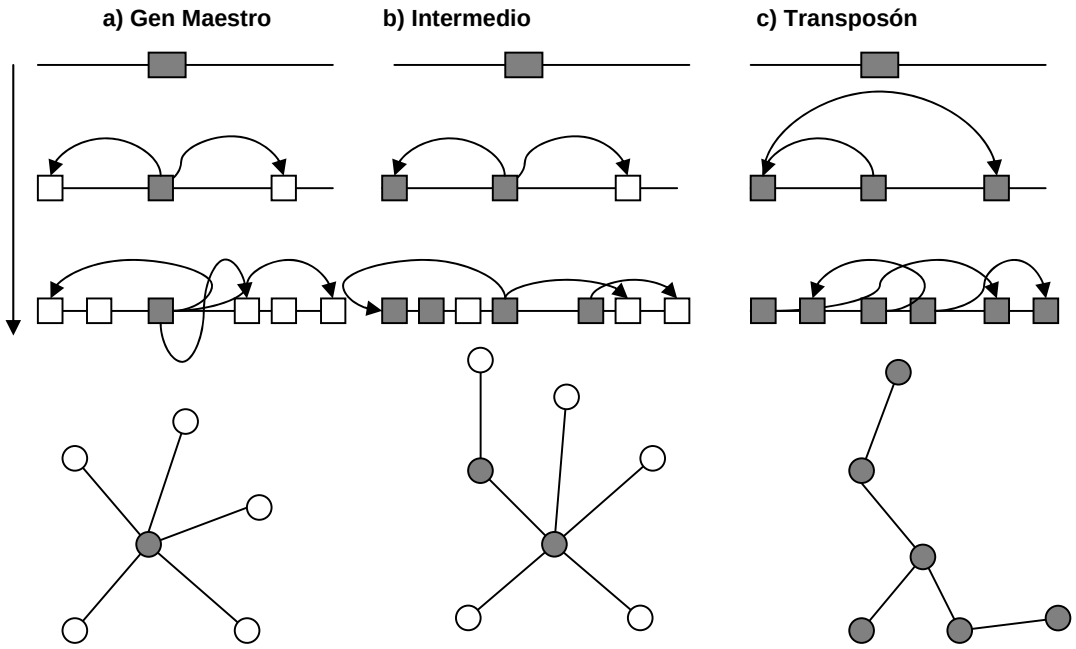
Los métodos basados en redes (*Network Analysis*) para inferencia de relaciones filogenéticas han sido diseñados para investigar las relaciones entre muestras altamente relacionadas además de permitir la coexistencia de nodos ancestrales y multifurcaciones. Situación que refleja la evolución de los elementos de transposición dentro de una especie.

Esta metodología está basada en el principio de parsimonia, es decir que infiere relaciones entre las muestras de forma que sean necesarios el menor número de pasos evolutivos. Sin embargo, contrariamente a los métodos tradicionales esta metodología genera reticulaciones y loops para conectar a las muestras.

Las relaciones entre elementos de transposición de una misma familia están caracterizadas por (a) baja divergencia entre las diferentes secuencias, (b) relaciones multifurcadas y (c) persistencia de los nodos ancestrales. Características contempladas en el análisis de Network.

Cuadro 2. Modelos de Transposición

Existen esencialmente dos modelos contrapuestos para explicar el fenómeno de transposición de un determinado elemento. Por un lado, el denominado Modelo ***Master Gene*** y por otro el modelo **transposón**.
El modelo Master Gene implica que en un determinado momento un único elemento es capaz de ser copiado a nuevas posiciones genómicas y expandir así a dicha familia de elementos. El locus replicativo es denominado de Master gene. En este modelo la existencia de subfamilias corresponde a la evolución de la secuencia de ese gen activo. Este modelo ha sido utilizado para explicar la evolución de algunos elementos de clase I como las secuencias *Alu* y algunos NLTRs [Kass, Batzer & Deininger 1995, Cordaux, et al, 2004]. Este modelo se hace evidente cuando se observa la substitución serial de una subfamilia de elementos por otra subfamilia.
Alternativamente, el modelo llamado transposón [Brookfield & Johnson,2006] implica que en un determinado momento diversos miembros de una familia presentes en diferentes posiciones cromosómicas son capaces de replicar, aunque el potencial de replicación entre esos elementos puede ser variable. Este modelo es típicamente aplicable a los elementos de clase II. Entre medio de estos dos modelos extremos se ha sugerido la existencia de un modelo intermedio.
Estos modelos difieren en la proporción de elementos que son activos en un determinado momento y también en el momento de la inactivación (en la propia transposición o más tarde).



Esquema basado en los modelos de transposición propuestos para secuencias Alu (Cordaux, et al 2004). En gris se representan los elementos activos y en blanco los inactivos

Elementos LTRs

La mayor parte de los elementos LTR encontrados en el genoma de *Anopheles* corresponde a elementos no truncados, de tamaño completo en relación al consenso en Repbase, es decir, secuencias que contienen ambos LTRs flanqueando la región interna de la familia del elemento al cual pertenecen (Figura 22). Por su parte en AnoTExcel no se identificaron elementos LTRs deletados mayores a 20% del tamaño del elemento canónico (Figura 22a). El porcentaje de identidad de las secuencias nucleotídicas de los elementos completos con los respectivos elementos canónicos depositados en RB es de 99,44% (con un rango de 96,66 a 100%).

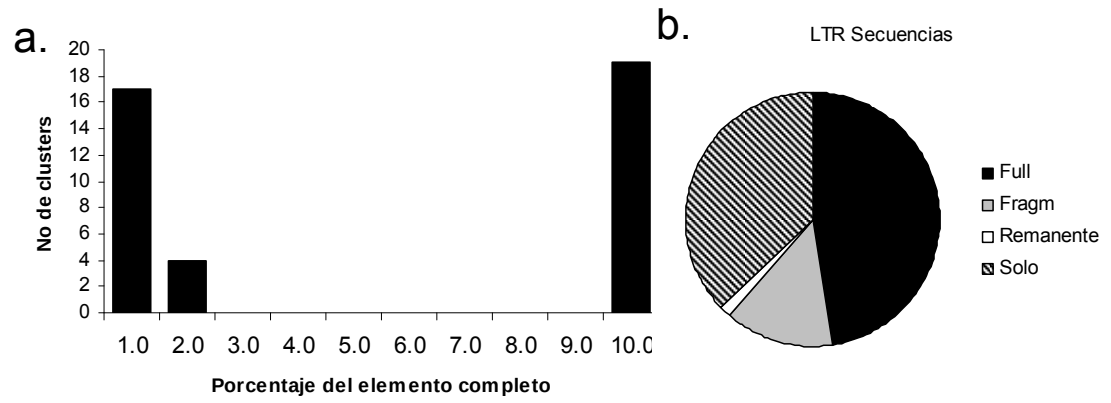


Figura 22. Degradación de elementos LTRs en AnoTExcel a) Tamaño de las secuencias consenso de los elementos LTRs en AnoTExcel en comparación al tamaño de la secuencia canónica según RB. **b)** Proporción de elementos completos (Full), fragmentos (Fragm), remanentes y Solo-LTRs (Solo) en AnoTExcel. Obs. Se consideró a los Solo-LTRs como elementos que corresponden al 10% del elemento completo. Los valores en el eje x corresponden a porcentajes del elemento canónico (de 10-100%).

Los pocos clusters con elementos fragmentados identificados en AnoTExcel, indican que las familias de LTRs se deterioran fundamentalmente por el proceso de recombinación que lleva a la formación de elementos de tipo Solo. En nuestro conjunto de datos los clusters con elementos fragmentados (clusters 211, 225,

167 y 205 correspondientes a la superfamilia *Gypsy* y clusters 53 y 193 a la Pao_*Bel*) poseen una identidad relativamente baja con las secuencias consenso de las familias a las cuales pertenecen 68,78 (64-75%) (Ver figura 23).

Cabe señalar que todas las secuencias del cluster 53 se encuentran en el cromosoma UNKN de *Anopheles gambiae*, es decir corresponde a DNA que no pudo ser localizado en los cromosomas reales, tratándose probablemente de regiones de heterocromatina. El hecho de estas secuencias ser 100% idénticas parece indicar que no se trata de inserciones producidas por diferentes eventos de transposición sino de duplicaciones cromosómicas.

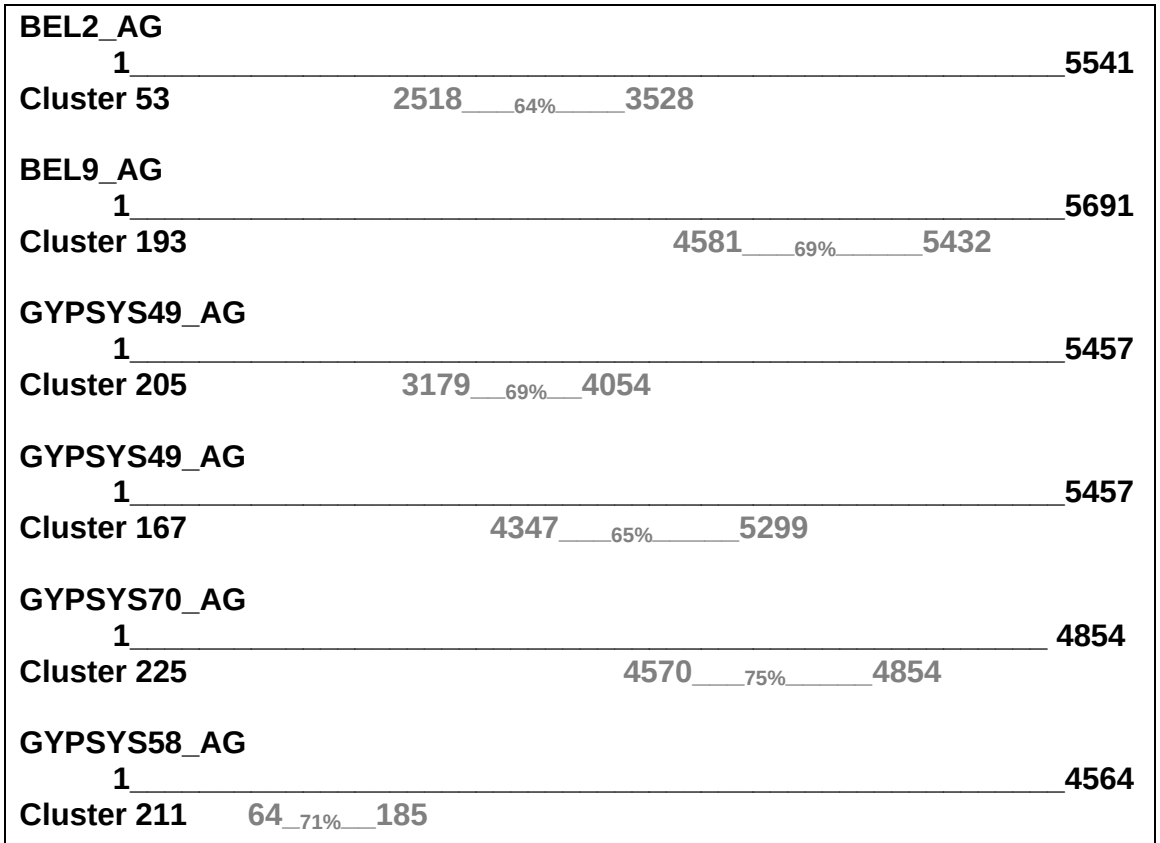


Figura 23. Representación de elementos LTRs deletados. Representación esquemática de los elementos fragmentados del orden LTR presentes en AnoTExcel. El porcentaje indica la identidad de las secuencias presentes en AnoTExcel en relación a los elementos canónicos en RB. Los números representan las posiciones nucleotídicas.

Las regiones deletadas no parecen tener un patrón particular, es decir, no parece existir una forma particular de perdida de DNA de estos elementos. Estos fragmentos pueden haber surgido por perdidas aleatorias de DNA de estos elementos.

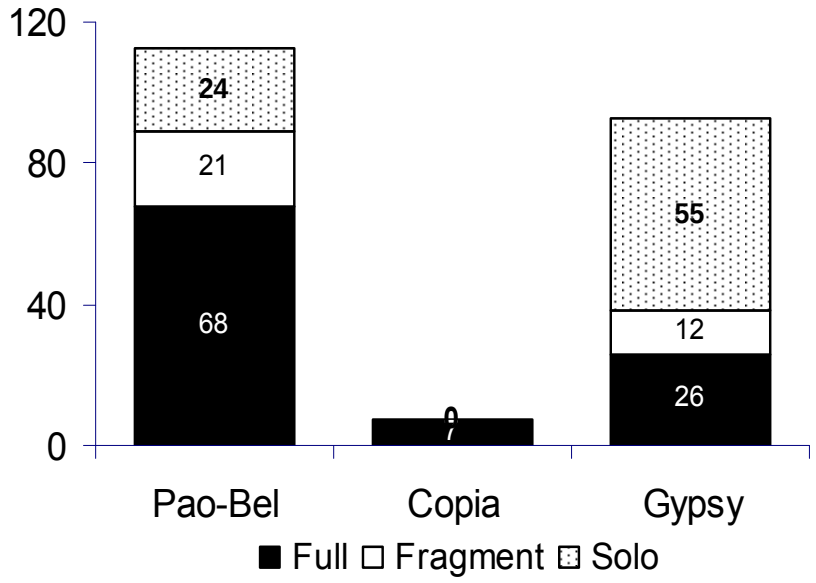


Figura 24. Proporción de elementos completos (Full), fragmentados (Fragment) y Solos-LTRs para los diferentes ordenes de elementos LTRs presentes en AnoTExcel.

Los elementos de tipo Solo-LTRs corresponden a una gran proporción de todos los LTRs detectados. Estas secuencias tienen una similitud con los LTRs de los elementos canónicos de los cuales provienen de 93,67% (con un rango de 73-100%), siendo que 7 de los 15 clusters con secuencias de tipo Solo tienen más de 98% de similitud. Esta elevada proporción de Solo-LTRs fue previamente descripta para los elementos *Gypsy* presentes en *Anopheles gambiae* [**Tubio et al., 2004, 2005**]. Según estos autores, la alta proporción de Solos se debería a un bajo *turn-over* de los elementos LTRs en este genoma, así, cada inserción de

un elemento LTR permanecería en el genoma por un mayor periodo de tiempo, aumentando la probabilidad de intercambio entre los LTRs del mismo elemento y produciendo elementos de tipo Solo. La alta identidad de estas regiones con los LTRs de los elementos completos, sugiere por el contrario, un alto turn-over de los LTRs, es decir, una vez que los elementos LTRs se transponen, rápidamente desaparecerían del genoma por medio de recombinación entre sus LTRs y por este motivo los LTRs mantienen alta identidad. La proporción de elementos de tipo Solo varía dependiendo del genoma analizado, en *Saccharomyces cereviasiae*, 85% de todas las inserciones de LTRs corresponden a Solo-LTRs [**Kim et al., 1998**] mientras que en *Drosophila melanogaster* se ha detectado una baja proporción de los mismos [**Kaminker et al., 2002; Lerat et al., 2003**). En el conjunto de datos presente en AnoTExcel la superfamilia *Gypsy* es la que posee la proporción mayor de elementos de tipo Solo (Figura 24). Las diferencias en las proporciones de elementos completos, fragmentados y Solos observada en esta figura podría también explicarse si los elementos de estas familias tuvieran tiempos de evolución diferentes en el genoma. De ser esta hipótesis verdadera, los elementos *Copia*, que solo presentan secuencias completas serían la superfamilia más nueva seguida de la superfamilia *Pao-Bel* y por último la superfamilia *Gypsy*, que tendría el mayor tiempo de evolución en el genoma por contener una mayor proporción de elementos deteriorados: solos, fragmentos y remanentes (ver AnoTExcel y Figura 24). Como hipótesis alternativa puede pensarse que las diferentes superfamilias de elementos LTRs tienen tendencias diferentes a recombinar y producir elementos de tipo Solos.

Se sabe que en el momento de la transposición de un retroelemento sus LTRs 3' y 5' son idénticos y a partir de ese momento los mismos sufren una evolución independiente que lleva a la incorporación diferencial de substituciones nucleotídicas. Asumiendo la existencia de un reloj molecular en estas secuencias, puede inferirse la edad de las copias actuales, es decir, el tiempo transcurrido a partir del evento de transposición que originó la copia ancestral de las actuales mediante la distancia existente entre ambos LTRs. La relación $T=k/2.\pi$. [**SanMiguel et al., 1998**] permite inferir ese tiempo para LTRs utilizando k como la distancia nucleotídica entre ambos LTRs y π como la tasa de substitución nucleotídica para sitios sinónimo del genoma. Aquí se utilizó la tasa calculada para *Drosophila melanogaster* de 0,0156/1.000.000 de años [**Li, 1997**] para estimar dichos tiempos para los elementos completos LTRs en AnoTExcel (Tabla 6). Para la superfamilia *Gypsy* el calculo de tiempo desde la transposición de los elementos, varía en media entre 0 (indicando actividad actual o reciente) y 1.000.000 años; para la superfamilia *Pao-Bel* entre 54.000 y 750.000 años y para la superfamilia *Copia* el promedio de tiempo oscila entre 0 y 700.000 años. Es claro que el número de familias analizadas no constituye el total de todas las familias descritas como LTRs, pero *a priori* el cálculo de las distancias basado en las distancias de los LTRs confirma la hipótesis de los diferentes tiempos de evolución en el genoma, influyendo en los diferentes grados de deterioración observados en estos elementos.

Por otra parte, varios de los elementos LTRs con secuencias completas tienen indicios de actividad actual o reciente (Tablas 6 y 7). Las identidades nucleotídicas entre las secuencias dentro de cada cluster son mayores a

99,99%, es decir, los elementos son prácticamente idénticos. Por su parte, las secuencias consenso de los clusters en AnoTExcel tienen una identidad promedio del 99,5% con los elementos consenso considerados en RB lo que enfatiza el alto grado de identidad de estas familias. Las distancias nucleotídicas entre el par de LTRs de cada una de estas secuencias es de 0,0000 para todas ellas y mantiene valores bajos para la distancia global de los LTRs considerando todas las secuencias en un mismo cluster.

La mayoría presenta por lo menos un ORF completo con dominios conservados de proteínas de TE activos, así como un *match* positivo contra una base de datos de ESTs en *Anopheles*.

El test de Tajima resultó significativo tan solo para dos de estos clusters (131 y 138, ambos *Gypsy*s). Esto indica que estos elementos evolucionan de manera no neutral. El cluster 131 presenta además *match* positivo tanto con la base de mRNAs como de EST, es decir que este elemento esta siendo expresado y claramente es activo. Por el contrario, el cluster 138, no presenta *match* positivo contra ninguna de las dos bases.

Tabla 6 Características de LTRs identificados en AnoTExcel

ID	Super Familia	#	N	D-p	SD	N-LTRs @	D-p LTRs *	D-p LTRs TODOS	T	Tm
110	BEL16	6	6097	0,0064	0,0006	287	0,0000	0,0017	54487	
191	BEL14	3	6565	0,0068	0,0009	236	0,0000	0,0023	73718	
238	BEL12	3	5896	0,0033	0,0005	221	0,0000	0,0024	76923	
140	BEL12	4	6299	0,0029	0,0004	204	0,0000	0,0029	92949	
199	BEL17	3	6168	0,0009	0,0003	276	0,0000	0,0039	125000	
176	BEL8	3	5879	0,007	0,0009	228	0,0000	0,0047	150641	
98	BEL6	8	6278	0,0116	0,0028	301	0,0000	0,0047	150641	
45	BEL1	21	8554	0,0029	0,002	290	0,0000	0,0059	189103	
101	BEL4	7	6057	0,0115	0,0008	316	0,0000	0,0162	519231	
130	BEL13	4	6236	0,0024	0,0004	240	0,0000	0,0162	519231	
115	BEL11	6	5820	0,0063	0,0007	261	0,0000	0,0232	743590	245047
150	COPIA5	4	4251	0,0011	0,0004	110	0,0000	0,0216	692308	
172	COPIA3	3	1838	0,0024	0,001	112	0,0000	0,0000	0	346154
131	GYPSY1	4	4187	0,0021	0,0005	148	0,0000	0,0000	0	
182	GYPSY3	3	4617	0,0014	0,0004	141	0,0000	0,0000	0	
104	GYPSY32	7	5067	0,001	0,0003	261	0,0000	0,0017	54487	
138	GYPSY29	4	4961	0,0013	0,0004	161	0,0000	0,0054	173077	
119	GYPSY2	5	5918	0,0094	0,0009	369	0,0000	0,0106	339744	
171	GYPSY35	3	4897	0,0095	0,0011	234	0,0000	0,0329	1054487	270299

ID **No de cluster**
**No de secuencias presentes en el cluster**
N **Tamaño Total de las secuencias en Nts**
D-p **Distancia-p entre todas las secuencias**
SD **Desvío Standard**
@ **Tamaño de los LTRs**
$ **Distancia-p entre los LTRs 5-3' de cada secuencias individual en el cluster**
***** **Distancia-p promedio entre todos los LTRs de las secuencias en el cluster**
T **Tiempo desde la Transposición según formula T=k/2(u) k=columna G y u=1,56x10-8**
Tm **Tiempo promedio para los elementos de una misma Superfamilia**

Tabla 7 Características de elementos LTRs identificados en AnoTExcel

ID	ORF 1	D.C	dN ORF1	dS ORF	w ORF1	ORF 2 (Nts/Aa)	D.C	dN ORF2	dS ORF2	w ORF2	Tajima (D)	Sig D	EST	m RNA
110	5253/1751	RT-PepA17	0.0046	0.0132	0.34848	-	-	-	-	-	-0.6495		+	-
191	5703/1900	RT-PepA17-RVE	0.0023	0.0136	0.16912	-	-	-	-	-	n.d.	n.d.	+	-
238	5148/1743	RT-PepA17-RVE	0.0022	0.0062	0.35484	-	-	-	-	-	n.d.	n.d.	+	-
140	5181/1726	RT-PepA17-RVE	0.0022	0.0068	0.32353	-	-	-	-	-	-0.5999		+	-
199	5547/1848	RT-PepA17-RVE	0.0002	0.0002	1.0000	-	-	-	-	-	n.d.	n.d.	+	-
176	5310/1769	RT-PepA17-RVE	0.0053	0.008	0.6625	-	-	-	-	-	n.d.	n.d.	-	-
98	5343/1780	RT-PepA17-RVE	0.0003	0.0003	1.0000	-	-	-	-	-	-0.9421		+	-
45	5316/1799	RT-PepA17-RVE	0.0018	0.0059	0.30508	1530	-	0.0012	0.0085	0.14118	-0.8038		+	+
101	5313/1771	RT-PepA17//-RVE	0.0047	0.0322	0.14596	-	-	-	-	-	-0.8133		+	-
130	5298/1765	RT-PepA17//-RVE	0.0000	0.0026	0.0000	-	-	-	-	-	-0.2078		+	-
115	5145/1714	RT-PepA17-RVE	0.0024	0.009	0.26667	-	-	-	-	-	-0.5706		+	-
150	3987/1322	RVE-RVT2	0.0002	0.0000	1.0000	-	-	-	-	-	-0.8294		-	-
172	1482	-	0.0005	0.0037	0.13514	-	-	-	-	-	n.d.	n.d.	+	-
131	3090/1044	RT_RNAseH-RVE	0.0000	0.001	0.0000	756	-	0.0018	0.0022	0.81818	-1.5118	***	+	+
182	4251/1416	RT-RVE	0.0004	0.0017	0.23529	-	-	-	-	-	n.d.	n.d.	+	-
104	4536/1461	RT-RVE	0.0009	0.0029	0.31034	-	-	-	-	-	-0.9444		+	-
138	4659/1552	RT-RVE	0.0006	0.0002	3.0000	-	-	-	-	-	-1.0809	***	-	-
119	3333/1110	RT_RNAseH-RVE	0.0061	0.0132	0.46212	-	-	0.0084	0.0226	0.37168	-0.4106		+	-
171	4101/1366	RT-RVE	0.0035	0.0057	0.61404	-	-	-	-	-	n.d.	n.d.	+	-

ID **No de cluster**
ORF **Open Reading Frame (marco de lectura abierto)**
dN **Sustituciones no sinónimas en sítios no sinónimos**
dS **Sustituciones sinónimas en sítios sinónimos**
w **dN/dS**
D.C **Dominios Proteicos Conservados. N=Ninguno detectado**
EST **Expressed Sequence Tags**
RT **Transcriptasa Reversa**
PepA17 **Peptidasa A17**
RVT **Integrasa**
******* ***P* < 0.001**
n.d **no determinado**

Análisis evolutivo de elementos LTR

Como ya fue mencionado, los elementos LTRs presentes en AnoTExcel corresponden a elementos completos y en algunos casos activos, y a elementos deteriorados de tipo Solo-LTR o fragmentados. No contamos con familias que posean elementos con diferentes grados de deterioración lo que impide el análisis propuesto previamente. En la Figura 25 se presenta un análisis de la deterioración estructural (como proporción de deleciones en las secuencias pertenecientes a una misma familia a lo largo del alineamiento nucleotidico) que muestra que las siete familias analizadas correspondientes a elementos de las tres principales superfamilias de LTRs poseen pocas inserciones o deleciones (indels) a lo largo de sus secuencias y los indels corresponden a pequeños gaps de alineamiento. Solo el cluster 45 muestra una mayor proporción de deleciones en la región 5'. De cualquier forma, una inspección visual del alineamiento de estas secuencias permitió ver que se trata de cinco secuencias en el alineamiento que presentan tres deleciones de 11, 34 y 115 nucleótidos respectivamente en las posiciones 566, 583 y 669 en relación a la secuencia consenso y que se encuentran fuera del ORF y de los dominios conservados de este elemento. Por lo tanto, teóricamente, no deben interferir con la actividad de este elemento. Para este cluster analizamos las relaciones filogenéticas y el análisis de Network previamente mencionado (Figura 26) quitando del alineamiento las regiones con las deleciones mencionadas. Es interesante el hecho que las cinco secuencias con las deleciones mencionadas no se agrupan en un único clado, sino que conforman dos clados con altos valores de *bootstraping* (Figura 26a).

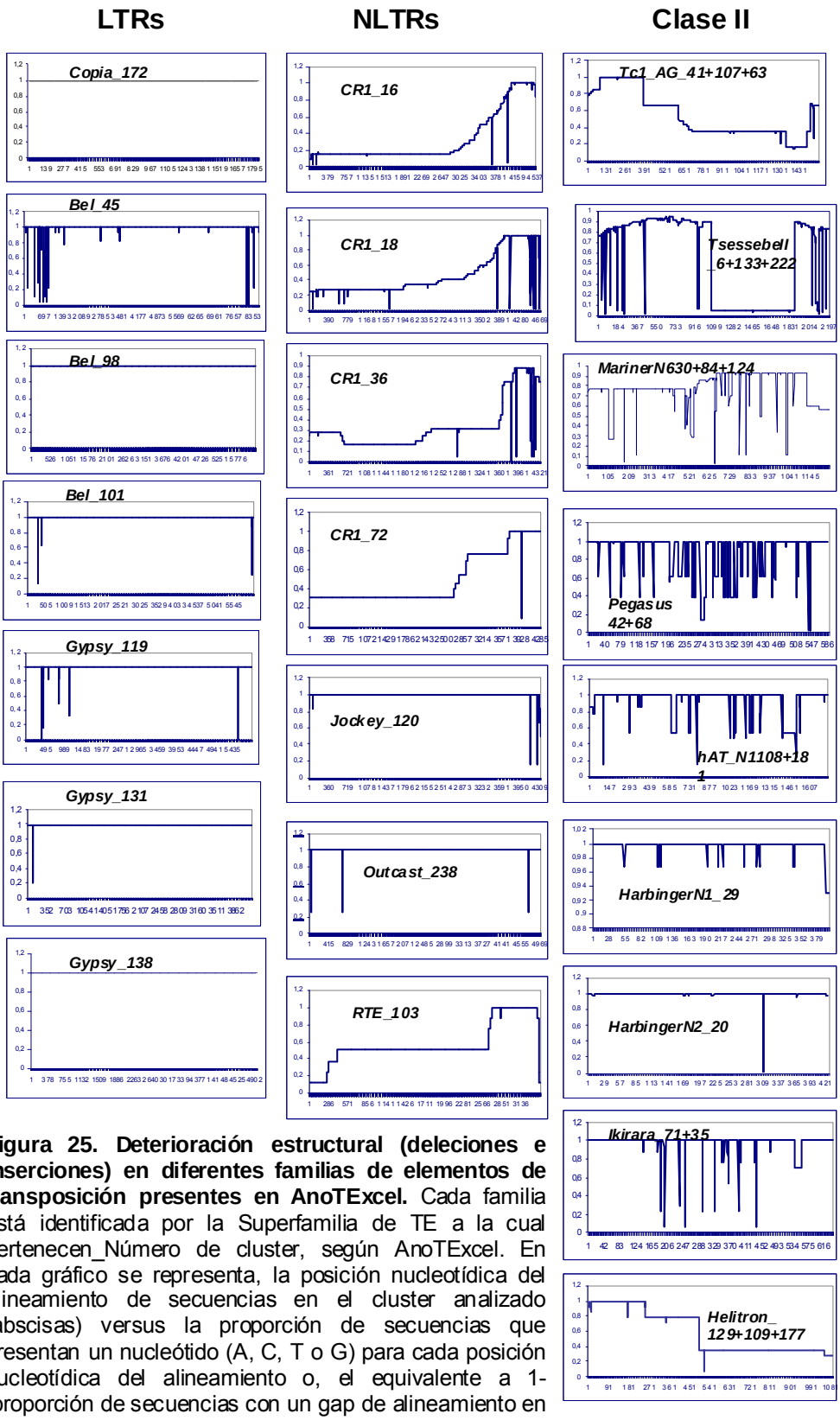


Figura 25. Deterioración estructural (deleciones e inserciones) en diferentes familias de elementos de transposición presentes en AnoTExcel. Cada familia está identificada por la Superfamilia de TE a la cual pertenecen_Número de cluster, según AnoTExcel. En cada gráfico se representa, la posición nucleotídica del alineamiento de secuencias en el cluster analizado (abscisas) versus la proporción de secuencias que presentan un nucleótido (A, C, T o G) para cada posición nucleotídica del alineamiento o, el equivalente a 1-(proporción de secuencias con un gap de alineamiento en dicha posición)

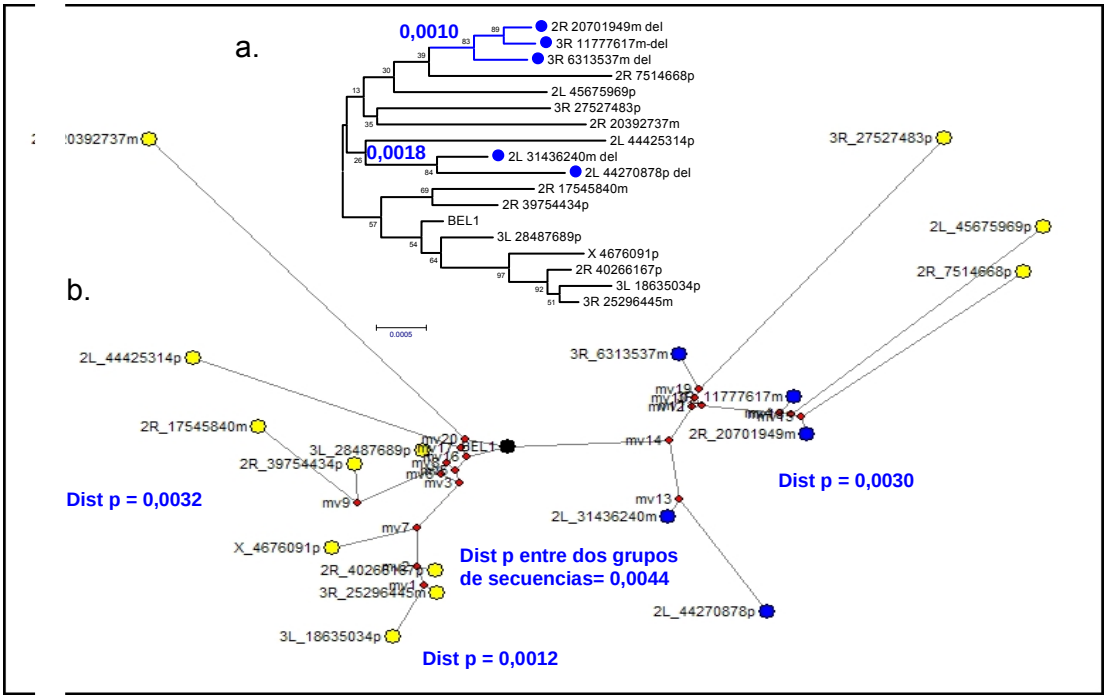


Figura 26. a) Filogenia basada en el método de *Neighbor-Joining* con *bootstraping de 1000* (se presentan los valores en las ramas respectivas). Las distancias fueron inferidas utilizando el método de *Maximum Composite Likelihood*. Se utilizó también el método de deleción completa para tratar las posiciones con gaps en el alineamiento. El alineamiento final consta de 8316 posiciones nucleotídicas. **b)** Análisis de Network. Circulos azules indican las secuencias con una deleción interna en el alineamiento nucleotídico, el circulo negro indica la secuencia del elemento consenso *BEL-1_AG* según RB. Las distancias-p en azul indicadas tanto en el árbol como en el network fueron calculadas utilizando el alineamiento original y el programa MEGA-4.

El análisis de Network (Figura 26b) indica que existen dos grupos de variantes derivados, probablemente, de dos eventos de transposición a partir de una copia ancestral muy semejante (o idéntica) al consenso depositado en Repbase. Las distancias nucleotídicas dentro de cada uno de estos grupos de secuencias a lo largo de los 8358 nucleótidos del alineamiento es de 0,0030 para ambos y la distancia global entre ambos grupos de 0,0044 lo que indica que estas secuencias son muy próximas entre sí. El modelo de transposición que sugiere es un modelo mixto a partir de una copia tipo *Master Gene* [**Clough et al., 1996**].

El resto de las familias de elementos LTRs analizadas en esta sección, contienen pocas copias y están compuestas únicamente por elementos completos. El análisis de network (Figura 27) mostró que en todos los casos el

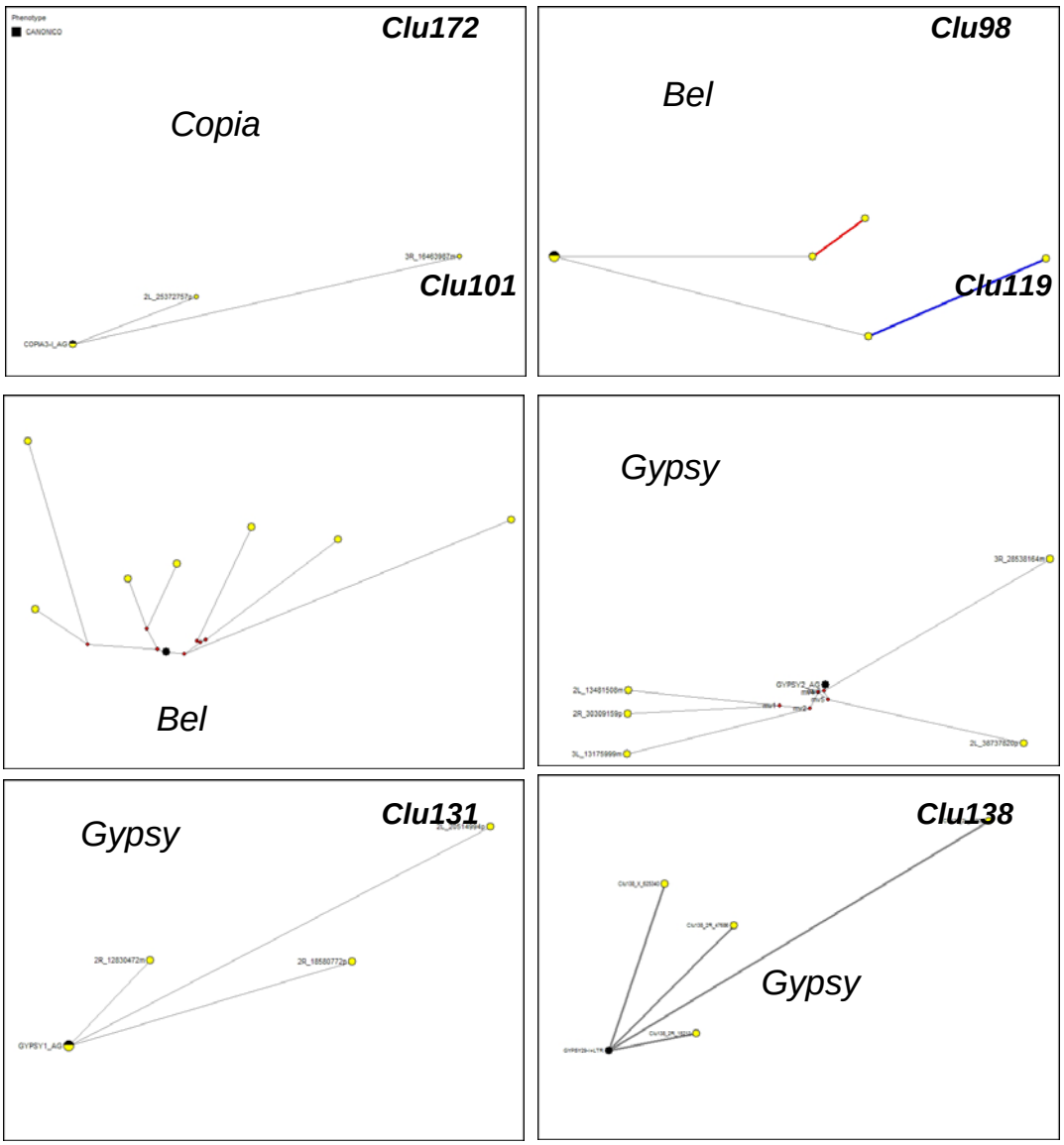


Figura 27. Análisis de Network para elementos *Copia*, *Bel* y *Gypsys*. Se indica el número de cluster en el extremo superior derecho de cada gráfico. Los círculos negros representan las secuencias consenso en RB, los amarillos corresponden a las secuencias dentro de cada cluster. Las líneas equivalen a pasos mutacionales entre las diferentes secuencias en cada análisis.

elemento consenso obtenido de Repbase para la familia analizada ocupa una posición central, es decir, como secuencia a partir de a cual se generan las otras

secuencias del cluster, sugiriendo que el ancestral sería tendría secuencia muy similar a la secuencia consenso depositada en RB. Esto, además de indicar que las secuencias que componen estas familias son homogéneas, en términos de identidad, sugiere que estas familias surgieron a partir de una copia que dio origen al resto de las secuencias. El hecho de estos clusters estar compuestos por tan pocas secuencias no permite un análisis detallado de estos clusters.

Elementos NLTR

Los elementos NLTRs presentes en AnoTExcel fueron clasificados como completos o fragmentados según el tamaño de la secuencia consenso de cada cluster en comparación con el elemento consenso de la familia a la cual pertenecen. A diferencia de los LTRs, estos elementos contienen un grado de deterioración mucho mayor (Figura 28) aunque, en media, los NLTRs completos tienen una identidad nucleotídica elevada, de 98,92% y los fragmentos de 93,58% en las regiones homólogas.

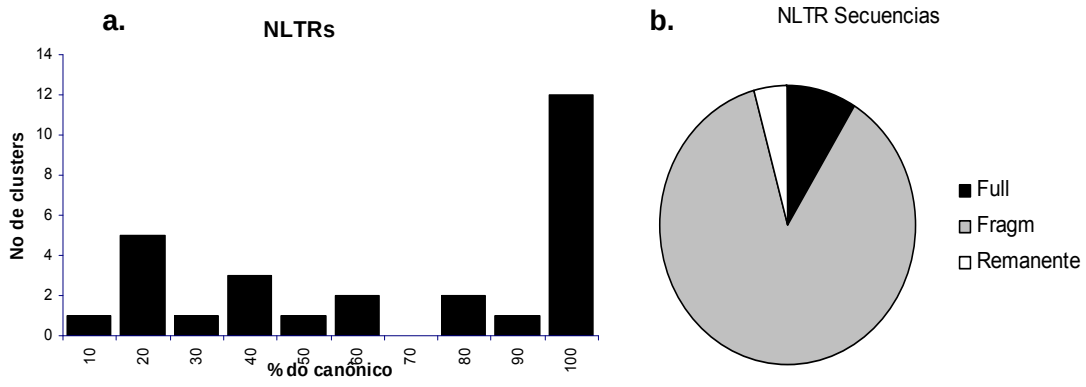


Figura 28. Degradación de elementos NLTRs en AnoTExcel a) Tamaño de las secuencias consenso de los elementos NLTRs en AnoTExcel en comparación al tamaño de la secuencia canónica según RB. **b)** Proporción de elementos completos (Full), fragmentos (Fragm), remanentes en AnoTExcel.

A pesar de haber sido posible generar secuencias consenso que representan al elemento completo para una gran parte de los clusters NLTRs , muchos de éstos están compuestos fundamentalmente por secuencias fragmentadas. Puede observarse en la Figura 28b (realizada a partir del número de secuencias) que en realidad la inmensa mayoría de los elementos NLTRs identificados en AnoTExcel corresponde a secuencias deterioradas.

Los elementos más deteriorados pertenecen a la Superfamilia *CR1*, que por su parte es la familia más numerosa de NLTRs en *Anopheles (*Figura 29). Las familias *Jockey* y *RTE* también se encuentran representadas por clusters conteniendo tanto secuencias completas como fragmentadas.

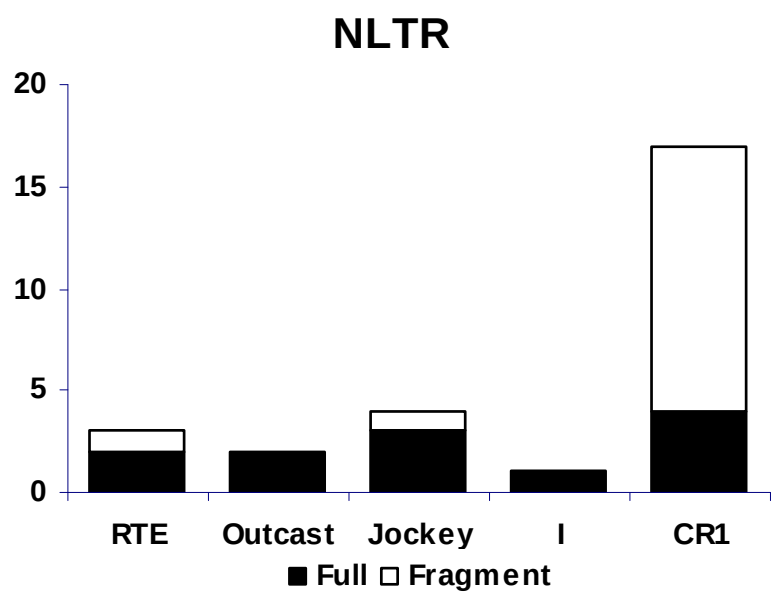


Figura 29. Proporción de elementos NLTRs completos (Full) y deletados (Fragment) en las diferentes superfamilias.

La figura 30 muestra para cada familia de elementos NLTRs identificados en AnoTExcel, el tamaño del elemento consenso para la familia según RB junto al tamaño de las secuencias de mayor y menor tamaño identificadas. Es evidente que todas las familias de la superfamilia *CR1*, y algunas pertenecientes a las familias *Jockey* y *RTE* están conformadas por secuencias completas y fragmentados.

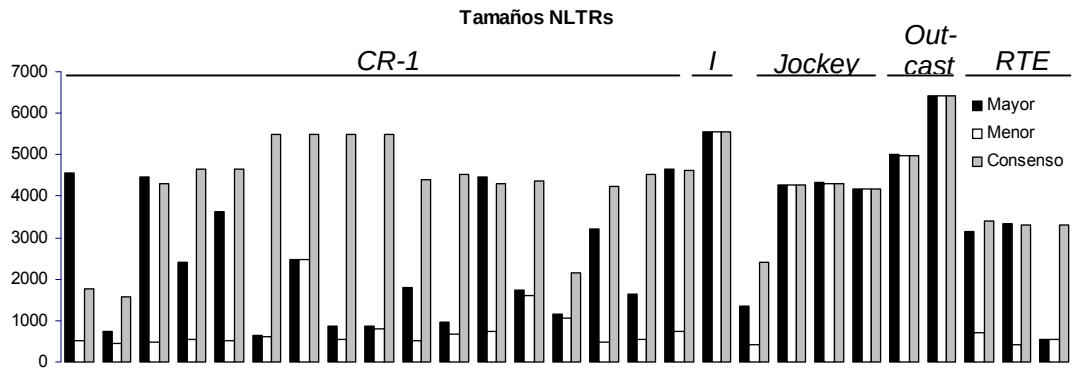


Figura 30. Tamaño de los elementos completos y fragmentados dentro de cada cluster según la Superfamilia a la cual pertenecen en comparación al elemento canónico según RB.

Los diferentes grados de deterioración de estas secuencias permiten estudiar el proceso de deterioración de las familias a las cuales pertenecen. El análisis de actividad de los clusters que presentan por lo menos algunas secuencias completas muestra que en la mayoría de los casos analizados las secuencias presentan más de 99,9% de identidad nucleotídica en las regiones de homología (Tabla 8). Por su parte, las secuencias consenso de los clusters en AnoTExcel tienen una identidad promedio del 98,9% con los elementos canónicos presentes en RB, un poco menor que la identidad que presentan las secuencias dentro de cada cluster. Varios de estos elementos presentan dominios conservados para las proteínas responsables por la replicación (Tabla 9). Para estos ORFs, los ω (dN/dS) son menores que uno indicando la existencia de selección a favor de la conservación proteica, es decir la existencia de selección negativa o purificadora. Todas las familias analizadas aquí con excepción de las *RTE* poseen un segundo ORF sin dominios proteicos conservados que presentan ω menores que uno.

Tabla 8	Características de NLTRs identificados en AnoTExcel				
ID	**Super Familia**	**#**	**N**	**D-p**	**SD**
18	CR1_T1	42	4617	0.0334	0.0017
72	CR1_Ele20	12	4473	0.0043	0.0008
36	CR1_Ele7	24	4398	0.0114	0.0014
16	CR1-Ele13	51	1520	0.008	0.0009
152	I_Ele2	4	5562	0.0007	0.0003
183	Jockey_Ele12	3	4275	0.0026	0.0006
120	Jockey_Ele14	5	4324	0.0036	0.0005
159	Jockey_Ele24	3	4179	0.0227	0.002
237	Outcast_Ele5	3	4993	0.0025	0.0005
188	Outcast_Ele6	3	6417	0.0016	0.0004
103	RTE_Ele1	7 (3)	3117	0.0054	0.0011
2	RTE_Ele2	232	3337	0.0265	0.0018
ID	**identificación de cluster**				
#	**Número de secuencias presentes en el cluster**				
N	**Tamaño Total de las secuencias en Nts**				
D-p	**Distancia-p entre todas las secuencias**				
SD	**Desvío Standard**				

El test de Tajima resultó significativo para tres clusters *CR1* (clusters 16, 36 y 72), todos con *match* significativo contra la base de EST, es decir que son expresados, pero tan solo uno de ellos (cluster 36) tiene *match* positivo contra la base de cDNA. Utilizando la tasa de substituciones sinónimas (dS) se estimó el tiempo desde la transposición para los diferentes elementos NLTRs analizados (Tabla 9). Según estos cálculos, las familias correspondientes a *CR1* y a *RTE* son las familias más antiguas, con más de un millón de años de transposición, mientras que *Jockey* y Outcast serían bastante más recientes. Es interesante destacar que dentro de los datos en AnoTExcel, tanto *CR1* como *RTE* son las superfamilias que presentan mayor grado de deterioración y donde se hace

evidente el modo de deterioración por pérdida de DNA en la región 5' previamente descrito para elementos NLTRs.

Tabla 9 Características de NLTRs identificados en AnoTExcel

ID	ORF 1 (NTs/Aas)	D.C	dN ORF1	dS ORF1	w ORF1	T	Tm	ORF 2 (Nts)	D.C	dN ORF2	dS ORF2	w ORF2	Tajima (D)
18	2826/941	Exo_End//-RT_nLTR	0.0304	0.0836	0.3636	2679487		1227	-	0.0131	0.0555	0.2360	-1.1576
72	2463/820	RT_nLTR	0.0055	0.0137	0.4015	439103		1137	-	0.0051	0.0183	0.2787	-1.6475
36	3009/1003	Exo_Endo/RT_nLTR	0.0070	0.0137	0.5109	439103		-	-	0.0070	0.0137	0.5109	-1.6059
16	1272	-	0.0037	0.0227	0.1630	727564	1071314	996	-	0.0045	0.0177	0.2542	-2.2036
152	3582/1193	Exo_Endo/RT_nLTR/RNAseH	0.0000	0.0013	0.0000	41667		1302	-	0.0000	0.0005	0.0000	-0.5583
183	2652/885	Exo_Endo/RT_nLTR	0.0017	0.0019	0.8947	76923		1308	-	0.0013	0.0024	0.5417	n.d.
120	2697/899	Exo_Endo/RT_nLTR	0.0014	0.0000	-	0		1173	-	0.0012	0.0073	0.1644	-0.9979
159	1137/378	RT_nLTR	0.0070	0.0235	0.2979	753205	276709	840	-	0.0162	0.0397	0.4081	n.d.
237	3645/1214	Exo_Endo/RT_nLTR/RNAseH	0.0008	0.0215	0.0372	689103		957	-	0.0006	0.0006	1.0000	n.d.
188	3594/1199	Exo_Endo/RT_nLTR/RNAseH	0.0000	0.0000	-	0	344551	1296	-	0.0000	0.0005	0.0000	n.d.
103	3075/1025	Exo_Endo/RT_nLTR	0.0204	0.0644	0.3168	2064103		-	-	-	-		-1.3952
2	3303/1101	Exo_Endo/RT_nLTR	0.0205	0.0429	0.4779	1375000	1719551	-	-	-	-		n.d.

ID **No de cluster**
T **Tiempo desde la Transposición según formula T=k/2(u) k=columna K y u=1,56x10-8**
Tm **Tiempo promedio para los elementos de una misma Superfamilia**
ORF **Open Reading Frame (marco de lectura abierto)**
dN **Substituciones no sinónimas en sitios no sinónimos**
dS **Substituciones sinónimas en sitios sinónimos**
w **dN/dS**
D.C **Dominios Proteicos Conservados. N=Ninguno detectado**
Exo-Endo **Exo-endo nucleasa**
RT **Transcriptasa Reversa**
nLTR **Non-LTR**
EST **Expressed Sequence Tags**

Análisis Evolutivo de elementos NLTR

El análisis de deterioración estructural (Figura 25, página 122) de los elementos en la superfamilia *CR1* muestra claramente que los mismos presentan la deterioración por deleciones en la región 5' descrita previamente. Todos los elementos de esta superfamilia presentan este patrón de deterioración (incluidas las secuencias de clusters "fragmentados" no mostrados aquí).

Uno de los clusters pertenecientes a la superfamilia *RTE* (cluster 2) también mostró este patrón. Sin embargo, los clusters representativos de las superfamilias *Jockey* y *Outcast* no mostraron este patrón.

Todos los clusters analizados en esta sección están compuestos por secuencias completas y fragmentadas.

La Figura 31 muestra los resultados de los análisis de network realizados en las familias de elementos NLTRs seleccionadas para este análisis. En ninguno de estos casos las secuencias consenso obtenidas de RB o Tf muestran posiciones centrales, sugiriendo que las mismas no son ancestrales al resto de las secuencias. Por otra parte, en prácticamente todos los casos, se hace necesario incorporar elementos ancestrales en diferentes posiciones (círculos rojos en la figura), es decir, que la muestra de secuencias obtenida para estas familias no contiene secuencias representativas de todos los elementos de esa familia. Las secuencias ancestrales deben haber existido en la ruta evolutiva de los elementos actuales pero pueden haberse degradado a tal punto que no fueron muestreados junto a las familias actuales.

En relación al modelo de transposición que se puede inferir a partir de estos resultados, es posible afirmar que ninguno de estos clusters muestra una

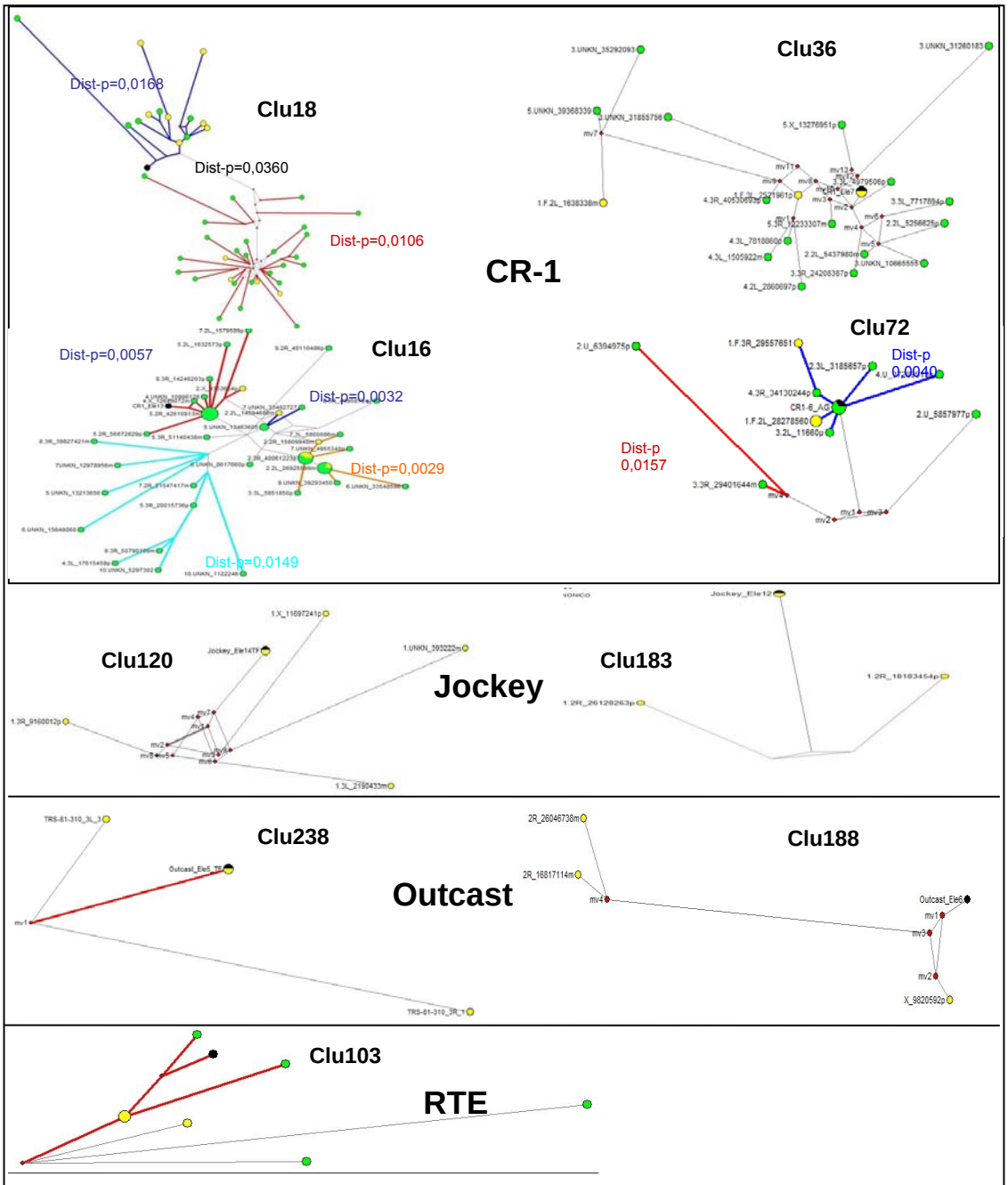


Figura 31. Análisis de Network para elementos NLTRs pertenecientes a diferentes superfamilias. Las ramas indicadas en colores em cada Network, indican grupos de secuencias que en la filogenia (no mostrada aquí) se agrupan con valores de *bootstaping* significativos (>70). Los circulos amarillos, indican secuencias completas, circulos verdes indican secuencias fragmentadas, y los negros corresponden a las secuencias canónicas según RB.

distribución compatible con un modelo de tipo *Master gene*. En todos los casos el modelo parecería corresponder a un modelo Mixto, donde algunas secuencias se multiplican y dan origen a secuencias activas y otras a secuencias inactivas [**Cordaux et al., 2004**].

Se realizaron filogenias por el método de *Neighbor Joining* para todas las familias analizadas aquí (Figuras presentadas en el Anexo 3). En todos los casos en que se agruparon secuencias con altos valores de *bootstraping* en la filogenia convencional, esas mismas secuencias aparecen agrupadas en el análisis de *Network*.

Clase II

La mayor parte de los elementos clase II descriptos en AnoTExcel corresponde a elementos deteriorados (Figura 32). En la figura 32a no fueron considerados los elementos MITE lo que aumentaría aún más la proporción de elementos representando bajos porcentajes de la secuencia consenso.

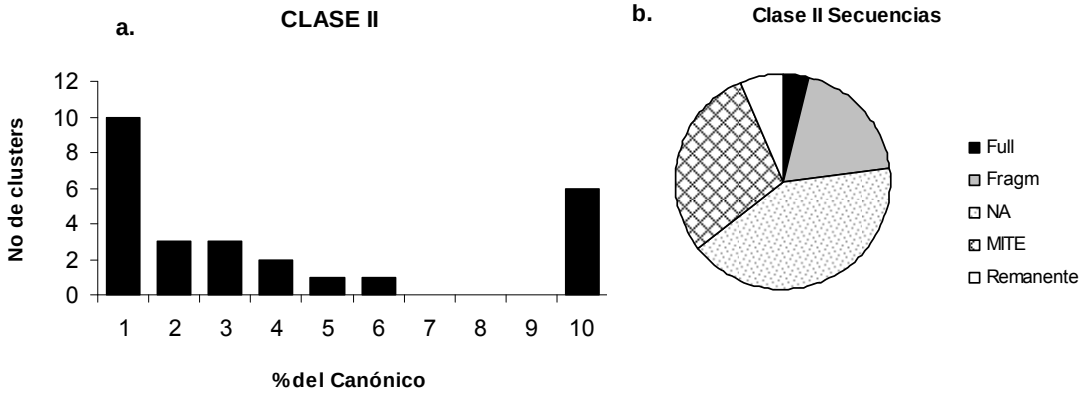
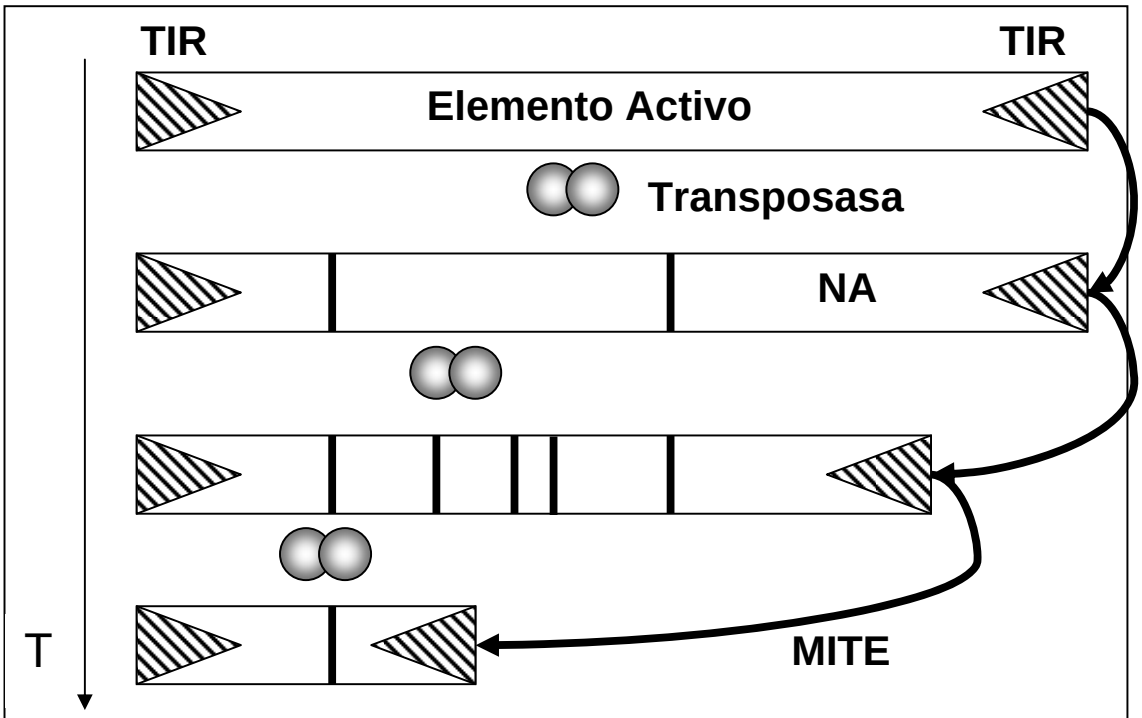


Figura 32. Degradación de elementos Clase II en AnoTExcel a) Proporción de clusters pertenecientes a la Clase II en AnoTExcel según el porcentaje del tamaño que representan en relación a la secuencia consenso. Observación: no se encuentran representados los elementos MITE b) Proporción de elementos completos (Full), y deteriorados: Fragm (fragmentos), NA (No autónomos según caracterización en RB), MITEs y Remanentes en AnoTExcel.

La diversidad de los elementos deteriorados de clase II es mayor que la de los elementos clase I (Figura 32b). En AnoTExcel fueron identificados una parte sustancial de estos elementos, además de una serie de elementos descritos en el capitulo anterior (elementos nuevos). La mayor parte de los elementos deteriorados corresponde a MITEs, que si bien son versiones deletadas y no autónomas de contrapartes activas y completas, tienen potencial de movilización. Los denominados “Fragmentos” corresponden a familias que poseen secuencias con identidad por alguna región de un elemento de clase II y corresponden claramente a elementos no autónomos ya que carecen de cualquier región codificante así como de TIRs. Los elementos denominados “NA”

corresponden a elementos que ya fueron clasificados como tal en RB, que corresponden a fragmentos de elementos completos, no tienen diferencia real con los “Fragmentos” identificados en AnoTExcel y pueden o no, presentar TIRs (en algunos casos secuencias palindrómicas). Esto puede indicar que se trata de elementos de tipo MITE que no han sido descritos previamente. El mecanismo de generación de MITEs no se conoce exactamente. Se sabe que los MITEs son versiones truncadas de sus respectivos elementos completos, con una forma de replicación parasitaria en relación a los elementos de clase II autónomos completos, con los cuales comparten los TIRs. Esto, aparentemente, permite la transcripción en *trans* de los mismos. Es posible pensar un escenario donde la generación de un MITEs comience con la inactivación de un elemento clase II activo por medio de una o más substituciones nucleotídicas en la secuencia que codifica la transposasa. Caso los TIRs de esta secuencia se mantengan conservados, podrá continuar amplificándose en el genoma mediante la actividad en trans de transposasas codificadas por otros elementos. Una vez inactivo, la secuencia sufriría un proceso de deterioración, adquiriendo substituciones y deleciones de forma gradativa, y estas versiones intermedias podrían formar familias de elementos NA, algunos de los cuales posteriormente podrían formar familias de MITEs, más numerosas y “exitosas” –en términos de “capacidad reproductiva”- debido a su pequeño tamaño y a su potenical parasítico. (Cuadro 3).

El análisis de la deterioración por familias (Figura 33) muestra que la superfamilia *Tc1-mariner* es la más diversa en términos del grado de



Cuadro 3. Representación esquemática del modelo propuesto para la generación de MITEs a partir de elementos de clase II activos (Ver Feschotte et al. 2002). NA= No Autónomo. Las lineas verticales indican substituciones nucleotídicas puntuales. Las puntas de flechas a rallas representan las repeticiones terminales invertidas (TIRs). T:Tiempo

deterioración que presentan los elementos que la componen, presentando proporciones similares de elementos Fragmentados, Mites y elementos NA (Figura 33).

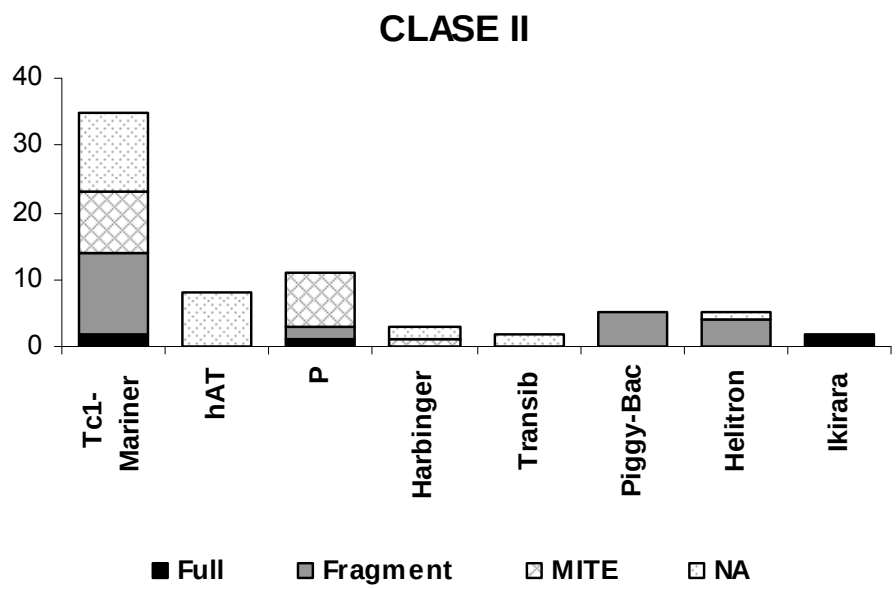


Figura 33. Proporción de elementos completos (Full), fragmentos (Fragment), No autónomos (NA) y MITEs en las diferentes superfamilias de elementos clase II

El análisis de actividad de los clusters con secuencias completas pertenecientes a la clase II muestran un grado de identidad entre sus miembros relativamente elevado (Tabla 10). Por otra parte, la identidad de las secuencias identificadas en AnoTExcel con las respectivas secuencias consenso en RB es de aproximadamente 97%.

Tabla 10 **Características de elementos CLASE II identificados en AnoTExcel**

ID	Super Familia	#	N	D-p	SD	N - TIRs
161	P3	3	4402	0,0005	0,0002	31
41	TC1-1	21 (7 Full)	1611	0,0201	0,0019	28
114	TX-MOS	6	1315	0,0332	0,0033	30
133	TSESSEB	4	2007	0,0129	0,0018	24
135	IKIRARA	4	624	0,0222	0,0043	246
71	IKIRARA	12	629	0,0291	0,0032	141

ID	**Número de cluster**
#	**Número de secuencias presentes en el cluster**
N	**Tamaño Total de las secuencias en Nucleótidos**
D-p	**Distancia-p entre todas las secuencias**
N-TIRs	**Tamaño de los TIRs**
SD	**Desvío Standard**

Sólo cuatro de los seis clusters con secuencias completas presentan ORFs, dos de ellos tienen dominios conservados para la Transposasa (ambos pertenecientes a la superfamilia *Tc1-mariner*). El cluster 41 (elemento Tc1) tiene un $\omega>1$ indicando una presión de selección positiva sobre la Transposasa y el cluster 133 un $\omega<1$ que indica una presión de selección negativa sobre estas secuencias. Además, el elemento 133 que corresponde a la familia *TsessebeII* posee *matches* positivos para las bibliotecas de expresión (EST) y de cDNA, por lo tanto, es claro que este elemento sufrió transposición recientemente en el genoma de *Anopheles gambiae*. El cálculo de edad de las copias basado en la tasa de substitución sinónima de *Drosophila melanogaster* indica un periodo desde la transposición de aproximadamente 150.000 años, bien menor que para los otros dos elementos para los que fue posible realizar este calculo (clusters 41 y 114). El Test de Tajima fue significativo únicamente para los elementos Ikirara, sin embargo estos elementos no poseen ORFs ni dominios proteicos conservados y corresponden a elementos de tipo MITE.

Tabla 11 Características de elementos CLASE II identificados en AnoTExcel

ID	ORF 1 (Nts/Aa)	D.C	dS	dN	w	T	Tajima (D)	Sig D	EST	mRNA
161	1593	-	0,0000	0,0010		0	n.d.	n.d.	+	-
41	1014/336	Transposase_5	0,0243	0,0291	1,19753	778846	-0,1348		+	-
114	636	-	0,0483	0,0315	0,65217	1548077	-0,4961		-	-
133	1017/339	Transposase_5	0,0049	0,0030	0,61224	157051	-0,6761		+	+
135	-	-					-1,0858	***	+	-
71	-	-					-1,7443	#	+	-

ID	**Número de cluster**
T	**Tiempo desde la Transposición según formula T=k/2(u) k=columna K y u=1,56x10-8**
ORF	**Open Reading Frame (marco de lectura abierto)**
dN	**Substituciones no sinónimas en sitios no sinónimos**
dS	**Substituciones sinónimas en sitios sinónimos**
w	**dN/dS**
D.C	**Dominios Proteicos Conservados. N=Ninguno detectado**
EST	**Expressed Sequence Tags**
*******	**p<0,001**
#	**p<0,1**

Análisis evolutivo de elementos clase II

El análisis de deterioración de las familias en esta clase muestra un patrón diferente de los elementos clase I (Figura 25, pagina 122). Los clusters analizados muestran, con excepción de pocos, un grado elevado de deleciones internas a lo largo de los alineamientos sin ninguna región preferencial para la aparición de las mismas.

Al contrario de los elementos clase I, para varias familias de elementos clase II fue posible generar alineamientos entre secuencias de clusters diferentes. Así, para las familias *TC1-AG*, *TsessebeII*, *MarinerNA*, *hAT_Pegasus*, *hATN1* e *Ikirara*, se generaron alineamientos entre secuencias de clusters diferentes junto al elemento consenso según RB. Estos alineamientos representan familias con distintos grados de deterioración y en la medida de lo posible fueron analizadas conjuntamente.

Superfamilia *Tc1-Mariner*

Tc1_AG

La familia *Tc1_AG* en AnoTExcel se encuentra representada por tres clusters diferentes, uno que contiene secuencias completas (cluster 41) y dos que corresponden a fragmentos (clusters 107 y 63). El alineamiento de las secuencias que componen estos tres clusters se encuentra representado esquemáticamente en la Figura 34. Todas las deleciones en los elementos fragmentados preservan uno o ambos extremos del elemento completo. El cluster 63, muestra la estructura génica de un MITE, con ambos TIRs conservados y ninguna capacidad codificante. El cluster 107 esta compuesto por secuencias idénticas todas depositadas en el cromosoma UNKN de *Anopheles*

gambiae, por lo tanto se utilizó una única secuencia para su posterior análisis. Es probable que este cluster esté compuesto por secuencias producto de un evento de duplicación y no por transposición. Por ultimo, el cluster 41 corresponde a secuencias completas, del mismo tamaño que el elemento consenso presente en RB con algunas de ellas presentando una inserción de 150 nucleótidos en la posición 1316.

El alineamiento en la región común a los tres clusters comprende 300 nucleótidos de la región 5', que no incluyen a los 25 nucleótidos correspondientes al TIR 5'. Esta región fue utilizada para el análisis filogenético basado en *Neighbor Joining* (Figura *Filogenias_NJ1000_ClaseII,* en Anexo 3) y el análisis de Network (Figura 35).

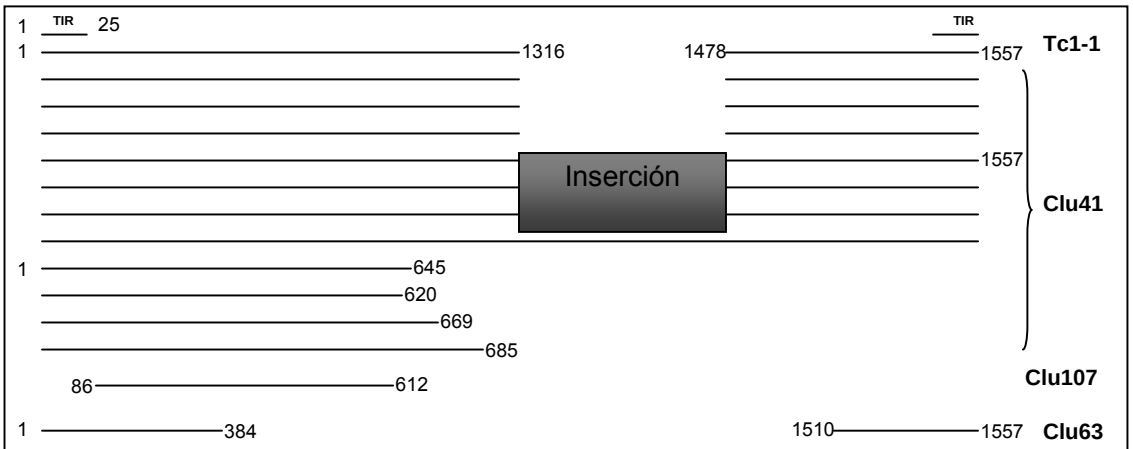


Figura 34. Representación esquemática de las secuencias de los clusters 41, 107 y 63 de AnoTExcel alineadas al elemento Tc1-1 (RB).

Este patrón de deterioración es compartible con el modelo de formación de MITEs sugerido anteriorment (ver Cuadro 3).

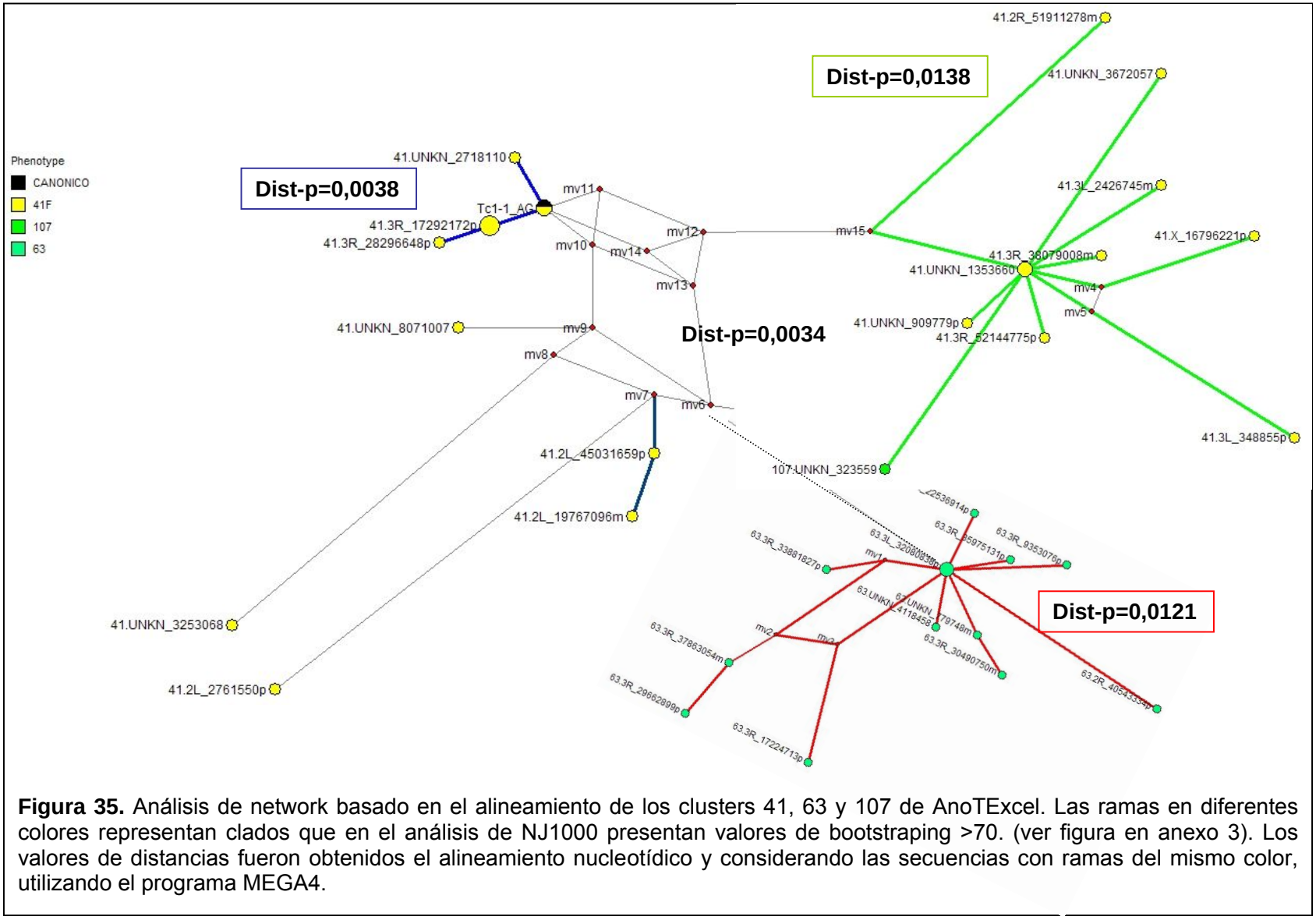


Figura 35. Análisis de network basado en el alineamiento de los clusters 41, 63 y 107 de AnoTExcel. Las ramas en diferentes colores representan clados que en el análisis de NJ1000 presentan valores de bootstraping >70. (ver figura en anexo 3). Los valores de distancias fueron obtenidos el alineamiento nucleotídico y considerando las secuencias con ramas del mismo color, utilizando el programa MEGA4.

En el análisis de Network, las secuencias del cluster 41 aparecen en dos grupos bien definidos (con líneas conectoras en verde y azul) que también pueden observarse en la filogenia basada en el método de *Neighbor Joining* como clados presentando un alto valor de *bootstraping* (Figura en el Anexo 3). El modelo de transposición que sugiere este análisis es Mixto, a pesar de existir nodos centrales con secuencias iguales (representado por círculos de mayor tamaño en la figura 35) a partir de los cuales surgen otras secuencias que componen esta familia. Sorprendentemente, las secuencias del cluster 41 completas que poseen características de actividad (Ver tabla 11) no se agrupan tan claramente en el análisis de *network* y requieren de varios ancestrales intermedios para conectar a las otras secuencias (círculos amarillos a la izquierda). La secuencia consenso presente en RB es idéntica a una de las secuencias del cluster 41 completas y da origen a otras secuencias de ese clado.

TsessebeII

TsessebeII (2026 nucleótidos) pertenece a la superfamilia *Tc1-Mariner*. En AnoTExcel, esta familia se encuentra representada por tres clusters, dos conteniendo secuencias fragmentadas (cluster 6 y 222) y uno con cuatro secuencias completas (cluster 133). El alineamiento de todas estas secuencias se encuentra representado esquemáticamente en la Figura 36. El cluster 6 fue dividido en subgrupos de secuencias con tamaños similares y denominados 6.1;

6.2 hasta 6.10 (ver esquema a continuación). Varias de estas secuencias presentan la misma deleción de 800nts, entre las posiciones 1071 y 1874.

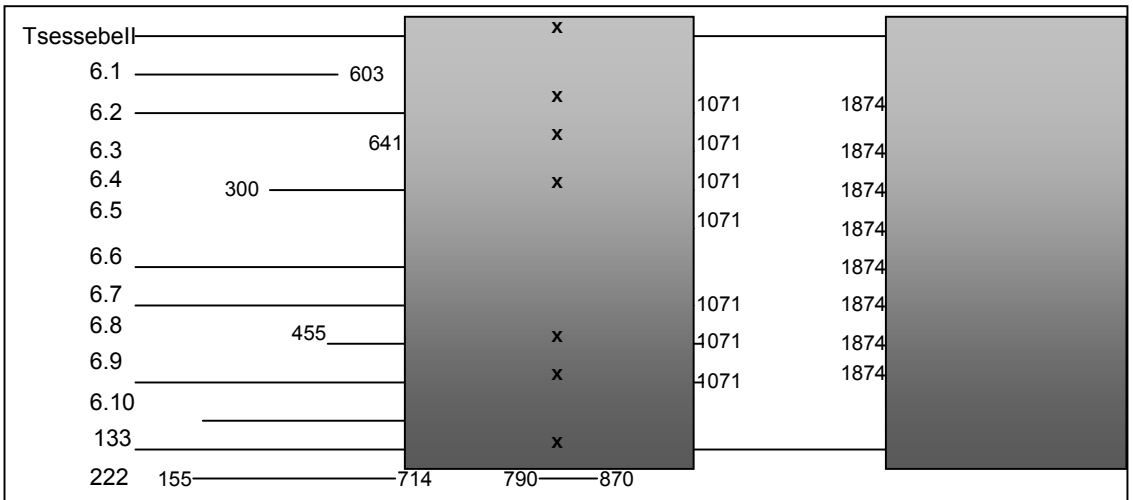


Figura 36. Representación esquemática de las secuencias de los clusters 6, 133 y 222 de AnoTExcel alineadas al elemento TsessebeII (RB). El cluster 6 a su vez está compuesto por secuencias representando diferentes regiones de del elemento completo. Las mismas fueron denominadas 6.1-6.10. Las regiones comunes a la mayor parte de las secuencias, se encuentran sombreadas.

El análisis de Network (Figura 37) y de Neighbor Joining (Figura en Anexo 2) se basa en el alineamiento de las regiones comunes a la mayor parte de las secuencias (regiones sombreadas en la Figura 36 representando 238 posiciones nucleotídicas). En este análisis no se pueden observar secuencias completas (pertenecientes al cluster 133) como ancestros de secuencias fragmentas, por lo tanto puede pensarse que estas secuencias corresponden a eventos de transposición diferentes que evolucionaron independientemente. Es interesante destacar que el cluster 6, compuesto por 107 secuencias es el más diverso en términos de la variedad de secuencias deterioradas que presenta. Algunas de las secuencias dentro de esta familia contienen TIRs a ambos extremos de la secuencia (subgrupos denominados 6.2; 6.6 y 6.9). El grupo denominado 6.2 es

el más numeroso y es probable que estas secuencias se comporten como Mites ya que mantienen sus TIRs conservados. Las otras secuencias poseen deleciones que incluyen a uno o ambos TIRs, en estos casos no es claro cual sería el mecanismo mediante el cual las mismas fueron generadas, pero la explicación más plausible sería por deleciones al azar en secuencias que primero sufrieron la deleción presente en el grupo 6.2. La secuencia consenso de RB se agrupa con las secuencias del cluster 133 (en amarillo), pero no como una secuencia ancestral sino como derivada de una secuencia intermedia común que da lugar a otras completas y activas. No existe en esta muestra un ancestro común a ninguno de estos tres grupos. El modelo de transposición sugerido es un modelo de tipo Mixto con preponderancia de un modelo de tipo transposón.

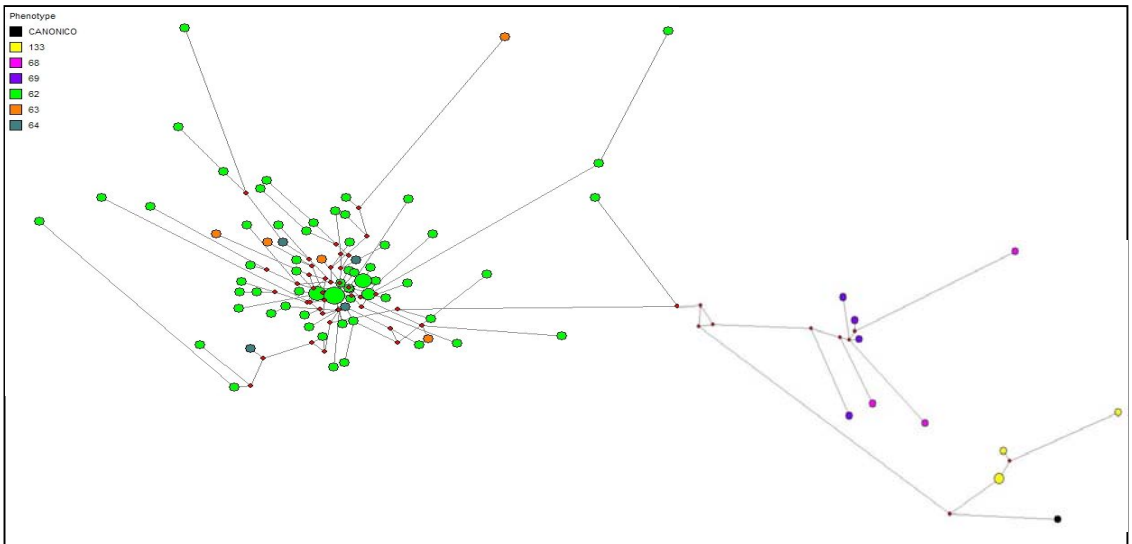


Figura 37. Análisis de *network* basado en el alineamiento de los clusters 133 y 6 de AnoTExcel. Los colores de los círculos representan secuencias pertenecientes a clusters con diferente grado de deterioración. En amarillo, cluster 133, fucsia, violeta, verde, rojo y verde oscuro corresponden a secuencias con diferentes deleciones dentro del cluster 6, respectivamente corresponden a 6.8; 6.9; 6.2; 6.3 y 6.4 según fueron llamadas en la figura 36.

Mariner Elementos No Autónomos (NA)

Se analizaron también elementos no autónomos pertenecientes a la familia Mariner. Los clusters 30, 84 y 124 corresponden a la familia Mariner N6_AG con diferente grado de deterioración y también se analizaron conjuntamente después de alinear todas las secuencias presentes en estos clusters con la secuencia consenso presente en RB. En este caso, al igual que en el caso anterior, se dividió al cluster 84 en dos grupos 84.1 y 84.2, y al 30 en cinco grupos (30.1-30.5) por contener secuencias con diferentes patrones de deleción.

Como puede observarse en la Figura 38 las secuencias presentan varias inserciones y deleciones en relación al elemento consenso Mariner N6, pero sin presentar un patrón específico. Las secuencias se encuentran bastante deterioradas en relación al consenso (75 y 77% de identidad entre los consensos de AnoTExcel y RB para los clusters 124 y 84, respectivamente y 99% para el cluster 30), que además corresponde a una secuencia deteriorada de un Mariner.

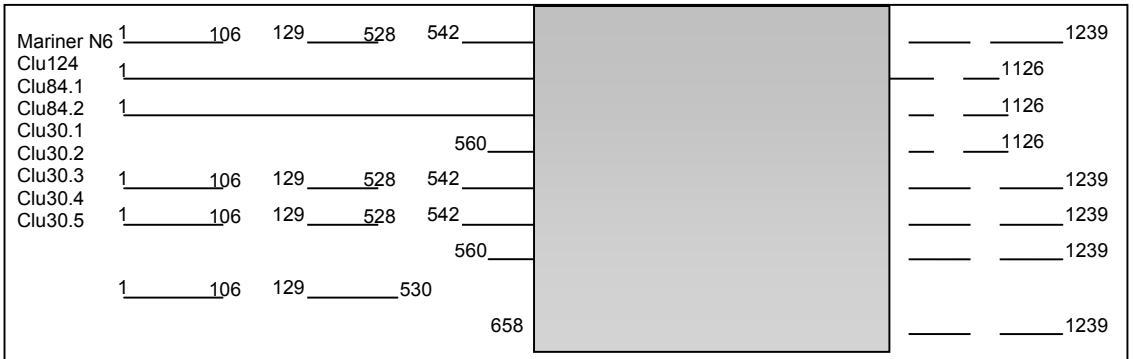


Figura 38. Representación esquemática de las secuencias de los clusters 124, 84 y 30 de AnoTExcel alineadas al elemento *MarinerN6_AG* (RB). Los clusters 84 y 30 están compuestos por secuencias representando diferentes regiones de elemento completo. Las mismas fueron denominadas 84.1 y 84.2 y 30.1-305. La región común a la mayor parte de las secuencias, se encuentra sombreada.

El análisis de Network muestra las secuencias agrupadas en los tres clusters, con distancias nucleotídicas muy próximas entre las diferentes secuencias (Figura 39). La distancia-p entre las secuencias del cluster 124 (en verde) con las secuencias del cluster 30 (azul) es de 0,3357; del cluster 30 con el 84 (azul) de 0,2993 y del 84 con 124 de 0,2353.

La secuencia consenso según RB se localiza en una posición bastante central al grupo de secuencias del cluster 30, y muestra que a partir de esa secuencia se originaron varias otras secuencias de ese grupo. Para el cluster 84 también se puede inferir la existencia de una secuencia que habría dado orígen a las otras secuencias, si bien, algunas de ella, a su vez, serían también activas

El modelo de transposición sugerido por el análisis de Network es también Mixto.

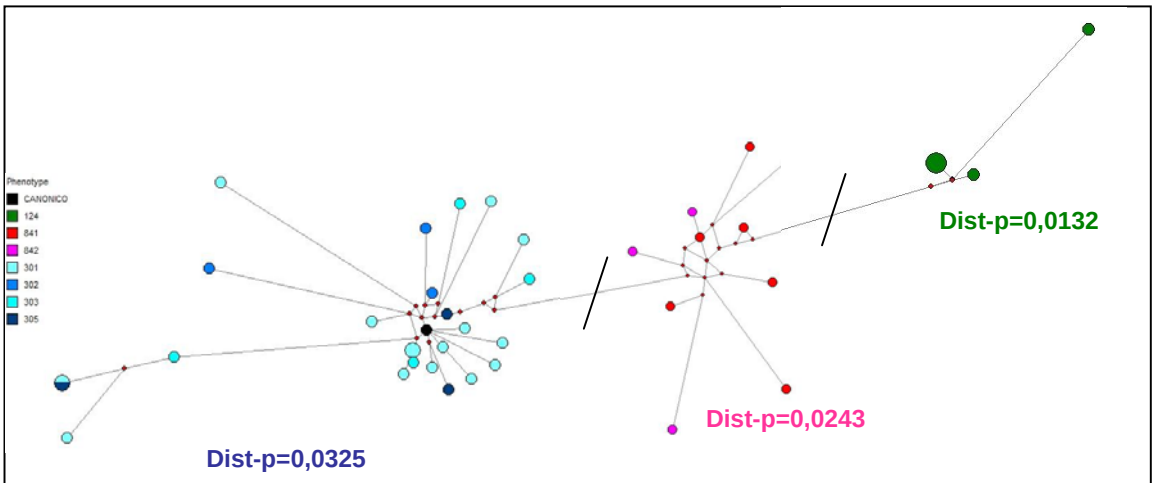


Figura 39. Análisis de *Network* para el alineamiento nucleotídico de los clusters 124, 30 y 84 de AnoTExcel. Las distancias-p fueron calculadas utilizando el programa MEGA4. Los colores de los círculos representan secuencias pertenecientes a clusters con diferente grado de deterioración. En negro se representa el elemento canónico según RB, en verde las secuencias del cluster 124, en rojo las secuencias del cluster 84 y en diversos azules las secuencias con diferentes deleciones presentes en el cluster 30.

Superfamilia *hAT*

Pegasus es una familia de elementos no autónomos pertenecientes a la superfamilia hAT. En Anotexcel, *Pegasus* se encuentra representado por dos clusters, con diferente grado de deterioración, presentando inserciones y deleciones en diferentes regiones en comparación al elemento consenso. El cluster 42 contiene 21 secuencias con 99% de identidad con *Pegasus*, mientras que el cluster 68 presenta 72% de identidad con este elemento consenso. Todas las secuencias presentan TIRs conservados de 8 nucleótidos. Este elemento no se encuentra caracterizado como MITE, pero dado su tamaño, y la presencia de TIRs probablemente se trate de un MITE. Ambos clusters segregan claramente en el *Network* en dos grupos diferentes con distancias nucleotídicas muy bajas entre las secuencias de cada uno de estos clusters (Figura 40) (para las secuencias del cluster 42 la dist-p es de 0,026 y para las del cluster 68 la dist-p es de 0,014). La distancia entre ambos grupo es de 0,2441. En la filogenia ambas subfamilias se presentan como dos clados con 100 de *bootstraping* (Figura en anexo 3)

El análisis de network muestra la división en ambos grupos, uno de ellos, el cluster 68 (representado en verde a la izquierda de la figura 40a) presenta un modelo de transposición mixto, si bien existe un nodo central representando por varias secuencias (circulo verde de mayor tamaño) que da origen a nuevas secuencias pareciendo un modelo tipo *Master gene*. El cluster 42 presenta una identidad alta con la secuencia consenso de Pegasus. Para conectar al conjunto

de secuencias dentro de este clado se hacen necesarios varios nodos

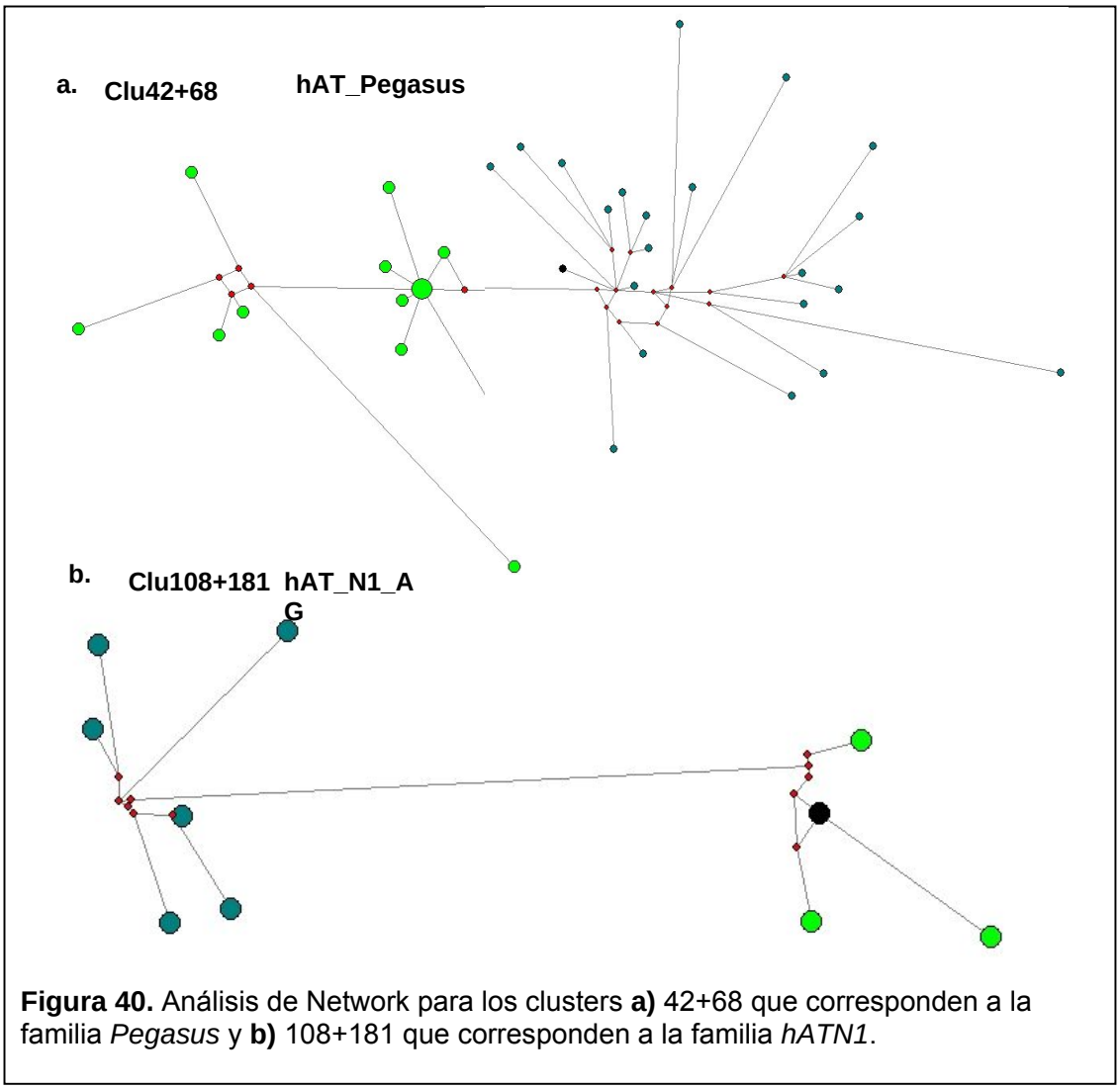


Figura 40. Análisis de Network para los clusters **a)** 42+68 que corresponden a la familia *Pegasus* y **b)** 108+181 que corresponden a la familia *hATN1*.

intermediarios, varios de ellos serian ancestrales de más de una secuencia muestreada en AnoTExcel, así el modelo de transposición sugerido por este modelo es también mixto.

hATN1

hATN1 es una familia de elementos no autónomos perteneciente a la superfamilia *hAT*. En AnoTExcel se presenta como dos clusters diferentes

(clusters 181 y 108) con distinto grado de deterioración (el cluster 108 con 90% de identidad con *hATN1* presenta varias deleciones de diferentes tamaño a lo largo de la secuencia y el cluster 181 con 99%).

La deterioración parece darse por deleciones a lo largo de la secuencia sin un patrón específico (ver figura 25, pagina 122). Las secuencias en estos clusters tienen distancias muy pequeñas entre si (0,0240 y 0,1190, respectivamente). En el análisis de *network* (cluster 108 representado por círculos azules a la izquierda y 181 en verde a la derecha, ver figura 40) sugiere un modelo de transposición Mixto.

Superfamilia *PIF-Harbinger*

La superfamilia *Harbinger* se encuentra representada en AnoTExcel por dos clusters, ambos pertenecientes a elementos no autónomos (cluster 20 y 29 que corresponden respectivamente a *HarbingerN2*_AG y *HarbingerN1*_AG). El análisis de deterioración de los elementos en el cluster 20 muestra que estos no presentan ningún patrón específico y solo son evidentes algunas inserciones o deleciones puntuales a lo largo de algunas secuencias (Ver Figura 25, pagina 122). El análisis de Network (Figura 41) muestra a la secuencia consenso como ancestral de todas las otras secuencias en este cluster. Como en otros casos, el network sugiere un modelo Mixto donde por una parte algunas secuencias dan origen a secuencias que a su vez generan secuencias activas, pero por otra pueden observarse secuencias que actúan como *Master genes*.

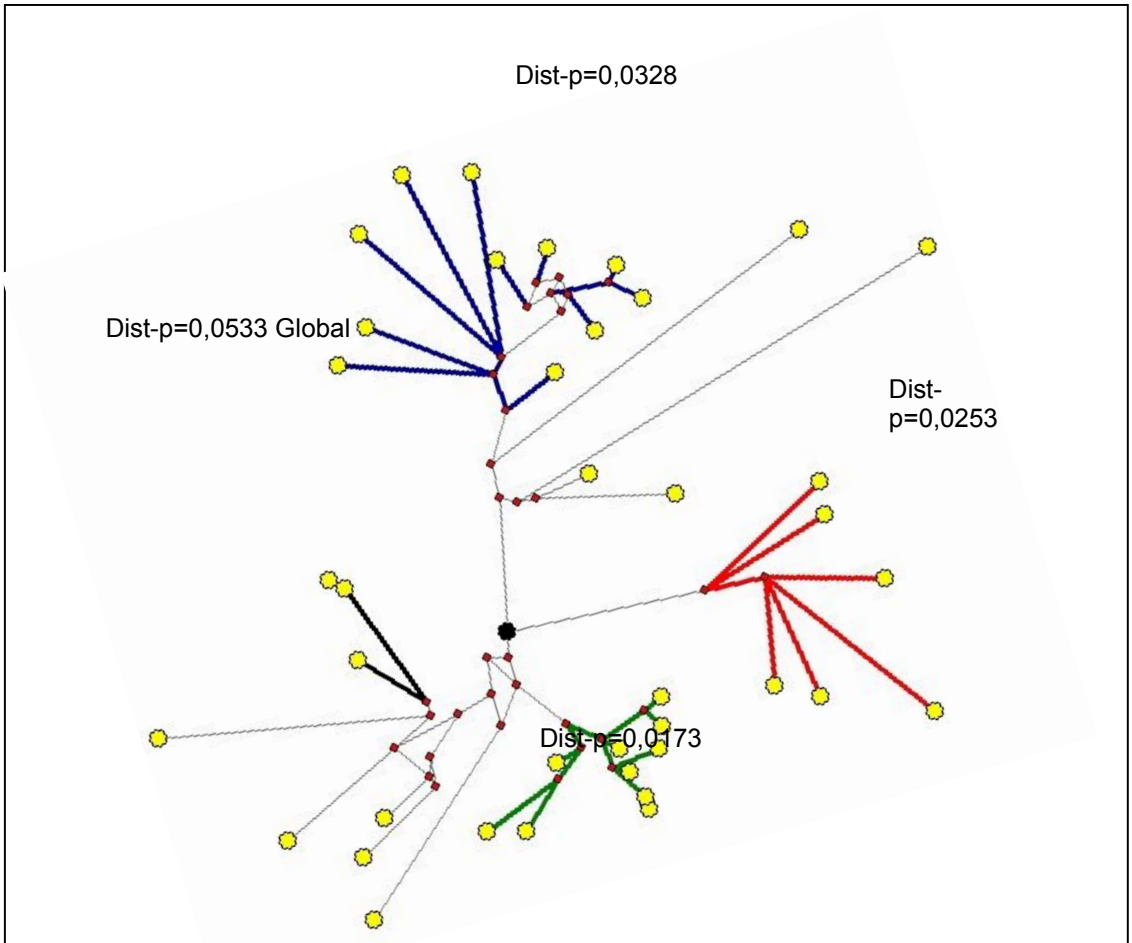


Figura 41. Análisis de network de las secuencias presentes en el cluster 20 (amarillo) y la secuencia consenso HarbingerN2_AG (negro). Los colores de las ramas identifican a grupos de secuencias que en el análisis de NJ1000, presentado en el Anexo 3 presentan valores de *bootstraping* significativos. Las distancias nucleotídicas fueron calculadas a partir del alineamiento de las secuencias indicadas con las ramas en color utilizando el programa MEGA4

Helitron

Estos elementos conforman una subclase de elementos de la clase II que fueron recientemente descubiertos en estudios *in silico* en genomas eucariotas. Se cree que se movilizan de una manera diferente a los clase II, utilizando un mecanismo de circulo rodante (*rolling circle*), ya que sus enzimas transposasas son homologas a los transposones RC de bacterias, que muestran este tipo de replicación [Kapitonov & Jurka, 2007]. En este proceso se ha descrito la captura

de fragmentos de genes del hospedero (principalmente en el maíz) [**Sweredoski et al., 2008**].

Existen dos familias de Helitrones activos identificadas en *Anopheles* (denominados Helitron1_AG y 2_AG) y una familia de elementos no autónomos Helitron2N_AG, derivada de Helitron2_AG. En AnoTExcel se identificaron clusters que corresponden a remanentes de los elementos activos, por tanto formas no autónomas, además de tres clusters correspondientes a la familia Helitron2N (Clusters 109, 129, 177).

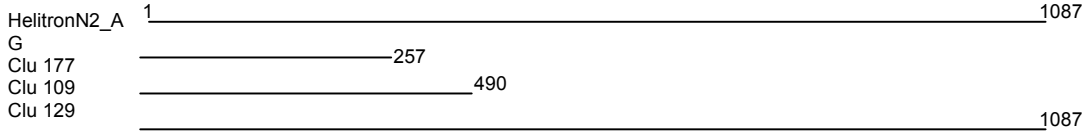


Figura 42. Representación esquemática del alineamiento de las secuencias de los clusters 109, 177 y 129 de AnoTExcel y el elemento HelitronN2 (RB). Los números indican posiciones nucleotídicas.

El análisis de deterioración en relación al elemento NA depositado en RB muestra una pérdida de DNA en la región 3' de algunas secuencias (Figura 25 y 42). Se analizó la región 5' común a todas ellas.

La filogenia (Figura en Anexo 2) agrupó a los tres clusters en dos clados con altos valores de bootstraping, uno conteniendo a las secuencias del cluster 177 y el otro conteniendo a las secuencias de los otros dos clusters que corresponde a las secuencias mayores con una distancia nucleotídica entres estas secuencias de 0,0137.

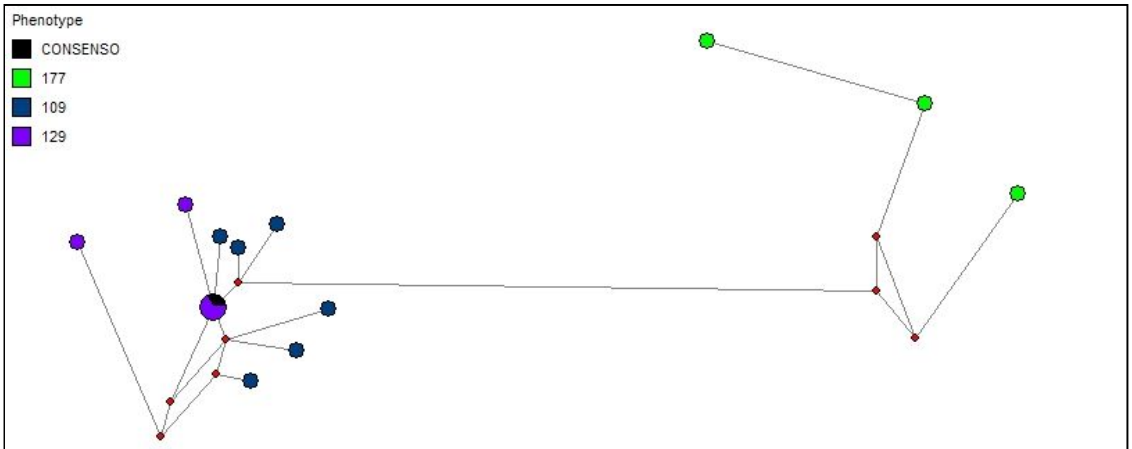


Figura 43. Análisis de Network para las secuencias de los clusters 109, 177 y 129 junto al elemento HelitronN2 (RB)

El análisis de network (Figura 43) sugiere que elementos pertenecientes al cluster 129 (circulo con tamaño mayor en la figura) y con la misma secuencia que el consenso para el elemento *Helitron2N*, dieron origen a otras secuencias del mismo cluster así como al cluster 109 en un modelo de transposición mixto.

Concluyendo, el análisis de deterioración estructural de las diferentes familias de elementos muestra que las tres clases de elementos presentes en el genoma de *Anopheles* se deterioran (pierden DNA) de maneras diferentes (Figura 25). El material disponible en AnoTExcel para análisis no permitió, en todos los casos, analizar familias de elementos que tuviese representadas por elementos individuales en diferentes fases de deterioración, ya que el proceso de clustering utilizado se basó en una identidad mayor al 90% sobre más del 90% del tamaño de los elementos, excluyendo formas truncadas a no ser que existiesen diversos elementos truncados de tamaños diferentes, como en el caso de los NLTRs. Así, los elementos LTRs analizados están formados por elementos completos en

tamaño y activos en algunos casos. Por otra parte, se identificó una alta proporción de elementos de tipo Solo, indicando que esta clase de elementos entran en una fase de deterioración principalmente por la perdida de DNA, provocada por la recombinación entre los LTRs de un mismo elemento.

Varios de los elementos NLTRs analizados aquí presentan deleciones en la región 5' como ya ha sido descrito. Las secuencias en estas familias presentan de cualquier manera, identidades muy elevadas en la región de homología, indicando que las deleciones son la principal causa de deterioración de estos elementos y no las substituciones nucleotídicas.

Los elementos de clase II, presentan una variabilidad en términos de elementos deteriorados mayor a los de la clase I. Además de la existencia de familias que corresponden a fragmentos de elementos completos, existen elementos NA (no autónomos según RB) y MITEs. Por otra parte, los clase II presentan pequeñas deleciones (o *gaps* de alineamientos) a lo largo de todas sus secuencias y presentan en varios casos, deleciones internas, manteniendo únicamente los extremos de estos elementos. Los elementos de tipo MITE serían casos extremos de este tipo de deterioración donde solo permanecen los TIRs y fragmentos de las regiones internas. Se sabe que elementos de tipo MITE pueden reproducirse en el genoma utilizando la maquinaria de replicación de elementos activos. Es posible que estos otros elementos (no clasificados como MITEs) -fundamentalmente por su tamaño mayor- que mantienen TIRs intactos puedan replicar de forma parasitaria. Estos datos permiten especular sobre la existencia de un proceso de deterioración gradual en esta clase de transposones

(Cuadro 3). Un escenario posible comenzaría con la inactivación del gen de la transposasa de un elemento, manteniendo sus TIRs. Este elemento podría replicar de forma parasitaria sobre elementos completos, y podría perder fragmentos de DNA mayores en la región interna sin impedir su transposición y de esta manera propagarse en el genoma y en el tiempo. Estos elementos, más pequeños y con TIRs conservados (MITEs) llevarían una ventaja sobre elementos deteriorados de tamaños mayores (también con relación parasitaria sobre elementos completos) ya que por su tamaño podrían replicar más rápidamente.

En relación a actividad, los elementos LTRs completos presentan varias familias con características compatibles con actividad reciente o actual, presentando dominios conservados para las proteínas necesarias para su replicación y parámetros (ω y D de Tajima) que indican la existencia de selección operando a nivel de sus secuencias nucleotídicas. Los NLTRs presentan familias donde algunos elementos son completos y poseen dominios proteicos conservados con indicios de selección, junto a secuencias claramente deterioradas. Por ultimo, los elementos clase II presentan mayoritariamente familias deterioradas. De las seis familias con secuencias de elementos completas, tan solo dos presentan dominios conservados para la transposasa, es decir solo estas dos familias tendrían capacidad de transposición autónoma. Por otra parte, cinco de estas familias tienen homología con ESTs, lo que indica la posible presencia de elementos que se expresan y que podrían cumplir roles regulatorios dentro de estas familias. El conjunto de secuencias clasificadas como clase II por la

presencia de TIRs, pero aparentemente no emparentadas pueden constituir una red de posibles puntos de reconocimiento y ligación de la transposasa en diferentes puntos del genoma. Estas interacciones podrían ser ventajosas, funcionando como reguladoras de la expansión de las familias clase II.

La presencia de secuencias sin capacidad codificante, pero siendo expresadas y con un alto grado de conservación a nivel nucleotídico podrían indicar la existencia de redes de regulación.

CONCLUSIONES & PERSPECTIVAS FUTURAS

En esta tesis se presentaron los resultados obtenidos a partir de un estudio *in silico* de los elementos de transposición presentes en el genoma de *Anopheles gambiae*. El interés inicial se basó en el potencial uso de estos elementos como drivers genéticos en experimentos de transformación genética de mosquitos del género *Anopheles* como herramienta para el control de la malaria.

La investigación resultante estuvo centrada en tres líneas principales:

(i) Caracterización de los TEs presentes en el genoma del mosquito
(ii) Identificación y Caracterización de elementos nuevos en dicho genoma
(iii) Análisis evolutivo de familias representativas de diferentes clases y órdenes.

(i) En relación al primer punto, se presentó AnoTExcel, que es una planilla de excel con información detallada sobre los elementos de transposición encontrados en el genoma del mosquito. Esta forma de presentación de la información resulta simple y útil, ya que permite la utilización de la información disponible en proyectos relacionados con el análisis de TEs. AnoTExcel, además de presentar información referente a las secuencias de elementos de transposición halladas en el genoma de *Anopheles*, tiene la ventaja principal de ofrecer todas las secuencias identificadas dentro de cada familia así como sus alineamientos globales, lo que permite estudios de dinámicas de esas poblaciones de secuencias. Si bien el objetivo inicial propuesto no era una búsqueda exhaustiva de los elementos de transposición del genoma de

Anopheles, la metodología utilizada permitió identificar y caracterizar una parte sustancial de las familias de TE descritas en el genoma del mosquito.

En AnoTExcel se presentan elementos de transposición pertenecientes a todas las clases y órdenes descritos previamente en el genoma de *Anopheles*. Otra característica importante del material presentado aquí es la presencia de diversos clusters con secuencias deterioradas, lo que permitió el análisis de deterioración propuesto en los objetivos de esta tesis. Así, dentro de los LTRs encontramos elementos completos, algunos de los cuales tienen características compatibles con actividad actual o reciente, elementos truncados o fragmentados así como Solo-LTRs. Por su parte, las familias de NLTRs están compuestas por elementos completos y fragmentados que presentan deleciones de diversos tamaños en la región 5', característica del mecanismo de replicación de los NLTRs. Por último, los elementos clase II, mayoritariamente representados por elementos con grados de deterioración variables: MITEs, elementos NA (no autónomos según Repbase) y fragmentos. Por otra parte, se identificaron también familias de elementos anidados (Nested) que no fueron analizados en esta tesis.

Todas las familias de TEs identificadas en AnoTExcel presentan secuencias con una identidad muy elevada, producto de la forma en que fue configurado y ejecutado el programa PILER en primer lugar y en que fueron agrupadas las secuencias dentro de los clusters obtenidos en AnoTExcel. Otras configuraciones en esta primera etapa permitirían probablemente la recuperación de secuencias de elementos de transposición con mayores grados

de deterioración. Esta perspectiva puede ser interesante para un análisis futuro de elementos de transposición en particular en el genoma, si bien debería también ponderarse la posible recuperación de un número muy grande de secuencias repetitivas que no correspondan a elementos de transposición.

Trabajos futuros, utilizando la plataforma de búsqueda, identificación y caracterización de elementos de transposición utilizada en AnoTExcel en genomas secuenciados de otros organismos puede permitir una comprensión más amplia de las dinámicas de los elementos de transposición, así como del fenómeno de transmisión horizontal.

(ii) En segundo lugar a partir del análisis preliminar de los clusters en AnoTExcel se identificaron elementos no descritos previamente. Se presentó un análisis detallado de familias de elementos que fueron caracterizadas por primera vez en el genoma del mosquito. Ninguna de estas familias habían sido descritas previamente y no se encontraron secuencias con homología en ninguna base de datos previa. Esto enfatiza la importancia de estudios que tengan como objetivo estudiar la diversidad de los elementos de transposición en un genoma. Ya que si bien el genoma de Anopheles ha sido intensamente minerado previamente, este approach permitió la identificación de elementos no descritos hasta el momento. Se identificaron y caracterizaron cuatro familias de elementos LTRs, completos y potencialmente activos y 32 de clase II, de tipo MITE con TIRs similares a elementos de familias ya descritas previamente, en las que no habían sido descritos este tipo de elementos, mostrando la variabilidad de

elementos deteriorados en el genoma. Por otra parte, en esta sección fue posible identificar una serie de clusters conformados por secuencias que comparten alta identidad y mantiene regiones terminales conservadas entre si, pero que no presentan identidad con ningun elemento de clase II previamente descrito ni en sus regiones terminales ni en el gen de la transposas. Tienen secuencias que en algunos casos poseen tamaños mayores a 1Kb. Estas secuencias resultan interesantes, ya que sugieren la existencia de elementos completos, activos, no identificados previamente, y que serían responsables por la amplificación de estas familias. Por otra parte, algunas de estas secuencias poseen alta homología con secuencias nucleotídicas de ESTs, lo que permite pensar que sean elementos con funciones regulatórias o de otro tipo en la dinámica de la familia a la que pertenecen. Algunas de ellas son bastante numerosas, lo que sugiere que tienen o han tenido estrategias de propagación exitosas en el genoma. Un estudio más profundo de estas familias, que intente identificar posibles elementos completos y activos responsables por la mobilización de estos elementos, así como también un análisis de las dinámicas de estas familias resulta interesante desde la perspectiva de la utilización de TEs como drivers genéticos. Elementos de tipo MITE podrían resultar útiles como posibles drivers genéticos dado su éxito de propagación y sus características parasíticas.

(iii) Por último, se realizó un estudio detallado de familias de elementos representando a los diferentes órdenes presentes en AnoTExcel bajo una

perspectiva evolutiva. Se mostró que los elementos de transposición pertenecientes a diferentes clases en el genoma de *Anopheles* se deterioran por procesos diferentes. Primeramente, análisis de deterioración estructural (Figura 25) mostraron que las deleciones se encuentran presentes en todas las clases, pero presentan patrones diferentes dando origen a distinto tipo de elementos deteriorados, en contraposición a lo sugerido previamente [**Quesneville et al., 2003**]. Así, los LTRs producirían fundamentalmente elementos deteriorados de tipo Solo, los NLTRs elementos truncados en la región 5' y los clase II presentan un mecanismo abortivo de reparación de gaps [**Rubin & Levy, 1997**] que produce secuencias con diversas deleciones internas que pueden o no mantener sus TIRs intactos. En este último caso, generando MITEs, que son formas parasíticas, no autónomas que se reproducen con diferente éxito en los genomas. Algunos de las familias analizadas aquí indican la existencia de un proceso gradativo de generación de MITEs.

Las familias analizadas en esta sección fueron también estudiadas filogenéticamente y por el método de network con el objetivo de comprender la relación entre las diferentes secuencias que las conforman. El método de *Median Joining* se utilizó de forma exploratoria ya que las premisas para su uso pueden ser aplicadas a las familias de TEs evolucionando en un genoma, esto es presentan alta identidad y es posible modelar sus interrelaciones a partir de redes o de multifurcaciones. La interpretación biológica que se puede dar al análisis de TEs por esta metodología permite inferir el modelo global de transposición al cual una familia de elementos se adecua (Cuadro 2). Así los

elementos clase I, orden LTR, presentan networks que pueden ser interpretados con modelo de transposición de tipo "*Master gene*" donde un elemento es responsable por la amplificación de las secuencias en esa familia. Los elementos NLTRs poseen networks que muestran relaciones entre ellas de tipo mixta, donde existe una combinación de secuencias que generan otras secuencias algunas de las cuales son capaces de generar nuevos elementos. Los elementos clase II también presentan modelos mixtos o intermediarios.

A partir de este análisis puede claramente observarse que muchas familias de TEs están integradas por secuencias que conforman redes complejas de interacciones. En un determinado momento una familia de TEs puede estar compuesta por una variedad de formas (activas o inactivas; autónomas o no autónomas; y deterioradas en diferente grado) conformando un entramado de posibles interacciones que potencialmente producen redes de regulación y amplificación variadas entre las secuencias que la conforman. La comprensión de la dinámica de deterioración de las familias y de los roles de estos elementos inactivos es fundamental para comprender el cuadro completo del fenómeno de transposición en un genoma.

Si bien el objetivo propuesto inicialmente en esta tesis radicaba en el análisis de elementos de transposición en el contexto de su uso como vectores de transformación genética en el genoma de *Anopheles*, el resultado final resultó una caracterización de los elementos de tansposición en dicho genoma de forma general, lo que sin embargo, puede también aportar al análisis de la adecuación

de un dado elemento de transposición como vector genético. Este estudio muestra la enorme variabilidad de formas deterioradas que existe en las diferentes familias de elementos de transposición en un determinado genoma. Muestra la gran complejidad que está presente en las redes de interacción que presentan estas secuencias entre sí, sugiriendo que la elección de cualquiera de estos elementos como vector de transformación para manipulación genética de insectos requiere de estudios profundos que puedan dar cuenta de las posibles redes de regulación negativa o de activación de estos elementos.

REFERENCIAS

1. Alphey, L. and M. Andreasen: **Dominant lethality and insect population control**. *Mol Biochem Parasitol*, 2002. **121**(2): p. 173-8.
2. Amino, R., R. Menard and F. Frischknecht: **In vivo imaging of malaria parasites--recent advances and future directions**. *Curr Opin Microbiol*, 2005. **8**(4): p. 407-14.
3. Andreasen, M.H. and C.F. Curtis: **Optimal life stage for radiation sterilization of Anopheles males and their fitness for release**. *Med Vet Entomol*, 2005. **19**(3): p. 238-44.
4. Andrieu, O., A.S. Fiston, D. Anxolabehere and H. Quesneville: **Detection of transposable elements by their compositional bias**. *BMC Bioinformatics*, 2004. **5**: p. 94.
5. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al.: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet*, 2000. **25**(1): p. 25-9.
6. Ashburner, M., M.A. Hoy and J.J. Peloquin: **Prospects for the genetic transformation of arthropods**. *Insect Mol Biol*, 1998. **7**(3): p. 201-13.
7. Atkinson, P.W. and K. Michel: **What's buzzing? Mosquito genomics and transgenic mosquitoes**. *Genesis*, 2002. **32**(1): p. 42-8.
8. Badge, R.M. and J.F. Brookfield: **The role of host factors in the population dynamics of selfish transposable elements**. *J Theor Biol*, 1997. **187**(2): p. 261-71.
9. Bandelt, H.J., P. Forster and A. Rohl: **Median-joining networks for**

inferring intraspecific phylogenies. *Mol Biol Evol*, 1999. **16**(1): p. 37-48.

10. Beier, J.C., G.F. Killeen and J.I. Githure: **Short report: entomologic inoculation rates and Plasmodium falciparum malaria prevalence in Africa**. *Am J Trop Med Hyg*, 1999. **61**(1): p. 109-13.
11. Benedict, M.Q. and A.S. Robinson: **The first releases of transgenic mosquitoes: an argument for the sterile insect technique**. *Trends Parasitol*, 2003. **19**(8): p. 349-55.
12. Bergman, C.M. and D. Bensasson: **Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in Drosophila melanogaster**. *Proc Natl Acad Sci U S A*, 2007. **104**(27): p. 11340-5.
13. Bergman, C.M. and H. Quesneville: **Discovering and detecting transposable elements in genome sequences**. *Brief Bioinform*, 2007. **8**(6): p. 382-92.
14. Biedler, J. and Z. Tu: **Non-LTR retrotransposons in the African malaria mosquito, Anopheles gambiae: unprecedented diversity and evidence of recent activity**. *Mol Biol Evol*, 2003. **20**(11): p. 1811-25.
15. Biemont, C., C. Vieira, C. Hoogland, G. Cizeron, C. Loevenbruck, C. Arnault and J.P. Carante: **Maintenance of transposable element copy number in natural populations of Drosophila melanogaster and D. simulans**. *Genetica*, 1997. **100**(1-3): p. 161-6.
16. Boete, C. and J.C. Koella: **Evolutionary ideas about genetically manipulated mosquitoes and malaria control**. *Trends Parasitol*, 2003. **19**(1): p. 32-8.

17. Breman, J.G.: **The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden**. *Am J Trop Med Hyg*, 2001. **64**(1-2 Suppl): p. 1-11.

18. Brookfield, J.F.: **The ecology of the genome - mobile DNA elements and their hosts**. *Nat Rev Genet*, 2005. **6**(2): p. 128-36.

19. Brookfield, J.F. and R.M. Badge: **Population genetics models of transposable elements**. *Genetica*, 1997. **100**(1-3): p. 281-94.

20. Brookfield, J.F. and L.J. Johnson: **The evolution of mobile DNAs: when will transposons create phylogenies that look as if there is a master gene?** *Genetics*, 2006. **173**(2): p. 1115-23.

21. Capy, P., R. Vitalis, T. Langin, D. Higuet and C. Bazin: **Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor?** *J Mol Evol*, 1996. **42**(3): p. 359-68.

22. Carter, R. and K.N. Mendis: **Evolutionary and historical aspects of the burden of malaria**. *Clin Microbiol Rev*, 2002. **15**(4): p. 564-94.

23. Catteruccia, F., T. Nolan, C. Blass, H.M. Muller, A. Crisanti, F.C. Kafatos and T.G. Loukeris: **Toward Anopheles transformation: Minos element activity in anopheline cells and embryos**. *Proc Natl Acad Sci U S A*, 2000. **97**(5): p. 2157-62.

24. Charlesworth, B. and C.H. Langley: **The population genetics of Drosophila transposable elements**. *Annu Rev Genet*, 1989. **23**: p. 251-87.

25. Clough, J.E., J.A. Foster, M. Barnett and H.A. Wichman: **Computer**

simulation of transposable element evolution: random template and strict master models. *J Mol Evol*, 1996. **42**(1): p. 52-8.

26. Coates, C.J., N. Jasinskiene, L. Miyashiro and A.A. James: **Mariner transposition and transformation of the yellow fever mosquito, Aedes aegypti**. *Proc Natl Acad Sci U S A*, 1998. **95**(7): p. 3748-51.
27. Collins, F.H. and S.M. Paskewitz: **Malaria: current and future prospects for control**. *Annu Rev Entomol*, 1995. **40**: p. 195-219.
28. Cordaux, R., D.J. Hedges and M.A. Batzer: **Retrotransposition of Alu elements: how many sources?** *Trends Genet*, 2004. **20**(10): p. 464-7.
29. Coy, M.R. and Z. Tu: **Gambol and Tc1 are two distinct families of DD34E transposons: analysis of the Anopheles gambiae genome expands the diversity of the IS630-Tc1-mariner superfamily**. *Insect Mol Biol*, 2005. **14**(5): p. 537-46.
30. Craig, N.L.: **V(D)J recombination and transposition: closer than expected**. *Science*, 1996. **271**(5255): p. 1512.
31. Curtis, C.F.: **Genetic sex separation in Anopheles arabiensis and the production of sterile hybrids**. *Bull World Health Organ*, 1978. **56**(3): p. 453-4.
32. Curtis, C.F., *Review of Previous Applications of Genetics to Vector Control*, in *Bridging laboratory and field research for genetic control of disease vectors*, B.G.J.L. Knols, C. , Editor. 2006. p. 33-43.
33. Doolittle, W.F. and C. Sapienza: **Selfish genes, the phenotype paradigm and genome evolution**. *Nature*, 1980. **284**(5757): p. 601-3.
34. Edgar, R.C.: **MUSCLE: multiple sequence alignment with high**

accuracy and high throughput. *Nucleic Acids Res*, 2004. **32**(5): p. 1792-7.

35. Edgar, R.C. and E.W. Myers: **PILER: identification and classification of genomic repeats**. *Bioinformatics*, 2005. **21 Suppl 1**: p. i152-8.

36. Elsheikha, H.M. and H.A. Sheashaa: **Epidemiology, pathophysiology, management and outcome of renal dysfunction associated with plasmodia infection**. *Parasitol Res*, 2007. **101**(5): p. 1183-90.

37. Feschotte, C.: **Merlin, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences**. *Mol Biol Evol*, 2004. **21**(9): p. 1769-80.

37b. Feschotte, C, Zhang X and Wessler, SR: **Miniature inverted-repeat transposable elements and their relationship to established DNA transposons**. *In Mobile DNA II 2002 (ed NL Craig, R Craigie, M. Gellert and AM Lambowitz), pp. 1147-1158. Washington, DC:ASM Press*.

38. Feschotte, C.: **Transposable elements and the evolution of regulatory networks**. *Nat Rev Genet*, 2008. **9**(5): p. 397-405.

39. Feschotte, C., N. Jiang and S.R. Wessler: **Plant transposable elements: where genetics meets genomics**. *Nat Rev Genet*, 2002. **3**(5): p. 329-41.

40. Feschotte, C. and E.J. Pritham: **DNA transposons and the evolution of eukaryotic genomes**. *Annu Rev Genet*, 2007. **41**: p. 331-68.

41. Finn, R.D., J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, et al.: **Pfam: clans, web tools and services**. *Nucleic Acids Res*, 2006. **34**(Database issue): p. D247-51.

42. Franz, G. and C. Savakis: **Minos, a new transposable element from Drosophila hydei, is a member of the Tc1-like family of transposons**. *Nucleic Acids Res*, 1991. **19**(23): p. 6646.

43. Fulton, A.M., S.E. Adams, J. Mellor, S.M. Kingsman and A.J. Kingsman: **The organization and expression of the yeast retrotransposon, Ty**. *Microbiol Sci*, 1987. **4**(6): p. 180-5.

44. Gelbart, W.M. and R.K. Blackman: **The hobo element of Drosophila melanogaster**. *Prog Nucleic Acid Res Mol Biol*, 1989. **36**: p. 37-46.

45. Gould, F. and P. Schliekelman: **Population genetics of autocidal control and strain replacement**. *Annu Rev Entomol*, 2004. **49**: p. 193-217.

46. Greenwood, B.M., K. Bojang, C.J. Whitty and G.A. Targett: **Malaria**. *Lancet*, 2005. **365**(9469): p. 1487-98.

47. Grossman, G.L., C.S. Rafferty, M.J. Fraser and M.Q. Benedict: **The piggyBac element is capable of precise excision and transposition in cells and embryos of the mosquito, Anopheles gambiae**. *Insect Biochem Mol Biol*, 2000. **30**(10): p. 909-14.

48. Guerra, C.A., R.W. Snow and S.I. Hay: **Mapping the global extent of malaria in 2005**. *Trends Parasitol*, 2006. **22**(8): p. 353-8.

49. Gupta, S., A.V. Hill, D. Kwiatkowski, A.M. Greenwood, B.M. Greenwood and K.P. Day: **Parasite virulence and disease patterns in Plasmodium falciparum malaria**. *Proc Natl Acad Sci U S A*, 1994. **91**(9): p. 3715-9.

50. Gupta, S., R.W. Snow, C.A. Donnelly, K. Marsh and C. Newbold: **Immunity to non-cerebral severe malaria is acquired after one or two**

infections. *Nat Med*, 1999. **5**(3): p. 340-3.

51. Handler, A.M. and R.A. Harrell, 2nd: **Germline transformation of Drosophila melanogaster with the piggyBac transposon vector**. *Insect Mol Biol*, 1999. **8**(4): p. 449-57.

52. Handler, A.M. and R.A. Harrell, 2nd: **Transformation of the Caribbean fruit fly, Anastrepha suspensa, with a piggyBac vector marked with polyubiquitin-regulated GFP**. *Insect Biochem Mol Biol*, 2001. **31**(2): p. 199-205.

53. Handler, A.M. and S.D. McCombs: **The piggyBac transposon mediates germ-line transformation in the Oriental fruit fly and closely related elements exist in its genome**. *Insect Mol Biol*, 2000. **9**(6): p. 605-12.

54. Hartl, D.L., E.R. Lozovskaya, D.I. Nurminsky and A.R. Lohe: **What restricts the activity of mariner-like transposable elements**. *Trends Genet*, 1997. **13**(5): p. 197-201.

55. Hickey, D.A.: **Selfish DNA: a sexually-transmitted nuclear parasite**. *Genetics*, 1982. **101**(3-4): p. 519-31.

56. Higgins, D.G. and P.M. Sharp: **CLUSTAL: a package for performing multiple sequence alignment on a microcomputer**. *Gene*, 1988. **73**(1): p. 237-44.

57. Holt, R.A., G.M. Subramanian, A. Halpern, G.G. Sutton, R. Charlab, D.R. Nusskern, P. Wincker, A.G. Clark, J.M. Ribeiro, R. Wides, et al.: **The genome sequence of the malaria mosquito Anopheles gambiae**. *Science*, 2002. **298**(5591): p. 129-49.

58. Ito, T., R. Motohashi and K. Shinozaki: **Preparation of transposon**

insertion lines and determination of insertion sites in Arabidopsis genome. *Methods Mol Biol*, 2002. **182**: p. 209-19.

59. Ivics, Z., P.B. Hackett, R.H. Plasterk and Z. Izsvak: **Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells**. *Cell*, 1997. **91**(4): p. 501-10.

60. Jasinskiene, N., C.J. Coates, M.Q. Benedict, A.J. Cornel, C.S. Rafferty, A.A. James and F.H. Collins: **Stable transformation of the yellow fever mosquito, Aedes aegypti, with the Hermes element from the housefly**. *Proc Natl Acad Sci U S A*, 1998. **95**(7): p. 3743-7.

61. Jones, J.M. and M. Gellert: **The taming of a transposon: V(D)J recombination and the immune system**. *Immunol Rev*, 2004. **200**: p. 233-48.

62. Jordan, I.K., I.B. Rogozin, G.V. Glazko and E.V. Koonin: **Origin of a substantial fraction of human regulatory sequences from transposable elements**. *Trends Genet*, 2003. **19**(2): p. 68-72.

63. Joy, D.A., X. Feng, J. Mu, T. Furuya, K. Chotivanich, A.U. Krettli, M. Ho, A. Wang, N.J. White, E. Suh, et al.: **Early origin and recent expansion of Plasmodium falciparum**. *Science*, 2003. **300**(5617): p. 318-21.

64. Jurka, J.: **Repeats in genomic DNA: mining and meaning**. *Curr Opin Struct Biol*, 1998. **8**(3): p. 333-7.

65. Jurka, J.: **Repbase update: a database and an electronic journal of repetitive elements**. *Trends Genet*, 2000. **16**(9): p. 418-20.

66. Jurka, J., V.V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany and J. Walichiewicz: **Repbase Update, a database of eukaryotic repetitive**

elements. *Cytogenet Genome Res*, 2005. **110**(1-4): p. 462-7.

67. Jurka, J. and A. Milosavljevic: **Reconstruction and analysis of human Alu genes**. *J Mol Evol*, 1991. **32**(2): p. 105-21.

68. Kalendar, R., C.M. Vicient, O. Peleg, K. Anamthawat-Jonsson, A. Bolshoy and A.H. Schulman: **Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes**. *Genetics*, 2004. **166**(3): p. 1437-50.

69. Kaminker, J.S., C.M. Bergman, B. Kronmiller, J. Carlson, R. Svirskas, S. Patel, E. Frise, D.A. Wheeler, S.E. Lewis, G.M. Rubin, et al.: **The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective**. *Genome Biol*, 2002. **3**(12): p. RESEARCH0084.

70. Kapitonov V.V., P.A., Jurka J.: **"COPIA1_AG, a family of copia-like LTR retrotransposons from African malaria mosquito."** *Repbase Reports*, 2003. **3**(3): p. 49.

71. Kapitonov, V.V. and J. Jurka: **A universal classification of eukaryotic transposable elements implemented in Repbase**. *Nat Rev Genet*, 2008. **9**(5): p. 411-2; author reply 414.

72. Kass, D.H., M.A. Batzer and P.L. Deininger: **Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution**. *Mol Cell Biol*, 1995. **15**(1): p. 19-25.

73. Kidwell, M.G.: **Transposable elements and the evolution of genome size in eukaryotes**. *Genetica*, 2002. **115**(1): p. 49-63.

74. Kidwell, M.G. and D.R. Lisch: **Transposable elements and host genome**

evolution. *Trends Ecol Evol*, 2000. **15**(3): p. 95-99.

75. Kidwell, M.G. and J.M. Ribeiro: **Can transposable elements be used to drive disease refractoriness genes into vector populations?** *Parasitol Today*, 1992. **8**(10): p. 325-9.

76. Killeen, G.F., U. Fillinger, I. Kiche, L.C. Gouagna and B.G. Knols: **Eradication of Anopheles gambiae from Brazil: lessons for malaria control in Africa?** *Lancet Infect Dis*, 2002. **2**(10): p. 618-27.

77. Killeen, G.F., U. Fillinger and B.G. Knols: **Advantages of larval control for African malaria vectors: low mobility and behavioural responsiveness of immature mosquito stages allow high effective coverage**. *Malar J*, 2002. **1**: p. 8.

78. Kim, J.M., S. Vanguri, J.D. Boeke, A. Gabriel and D.F. Voytas: **Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence**. *Genome Res*, 1998. **8**(5): p. 464-78.

79. Klinakis, A.G., T.G. Loukeris, A. Pavlopoulos and C. Savakis: **Mobility assays confirm the broad host-range activity of the Minos transposable element and validate new transformation tools**. *Insect Mol Biol*, 2000. **9**(3): p. 269-75.

80. Klinakis, A.G., L. Zagoraiou, D.K. Vassilatis and C. Savakis: **Genome-wide insertional mutagenesis in human cells by the Drosophila mobile element Minos**. *EMBO Rep*, 2000. **1**(5): p. 416-21.

81. Knipling, E.F.: **Pest Control: Chemical, Biological, Genetic, and**

Physical Means. *Science*, 1965. **147**(3660): p. 916-918.

82. Knipling, E.F., H. Laven, G.B. Craig, R. Pal, J.B. Kitzmiller, C.N. Smith and A.W. Brown: **Genetic control of insects of public health importance**. *Bull World Health Organ*, 1968. **38**(3): p. 421-38.

83. Kohany, O., A.J. Gentles, L. Hankus and J. Jurka: **Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor**. *BMC Bioinformatics*, 2006. **7**: p. 474.

84. Kokoza, V., A. Ahmed, E.A. Wimmer and A.S. Raikhel: **Efficient transformation of the yellow fever mosquito Aedes aegypti using the piggyBac transposable element vector pBac[3xP3-EGFP afm]**. *Insect Biochem Mol Biol*, 2001. **31**(12): p. 1137-43.

85. Kumar, M.N.a.S., *Molecular Evolution and Phylogenetics*. 2000, New York: Oxford University Press.

86. Kumar, S., K. Tamura and M. Nei: **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment**. *Brief Bioinform*, 2004. **5**(2): p. 150-63.

87. Lampe, D.J., M.E. Churchill and H.M. Robertson: **A purified mariner transposase is sufficient to mediate transposition in vitro**. *Embo J*, 1996. **15**(19): p. 5470-9.

88. Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al.: **Initial sequencing and analysis of the human genome**. *Nature*, 2001. **409**(6822): p. 860-921.

89. Le Rouzic, A. and P. Capy: **The first steps of transposable elements**

invasion: parasitic strategy vs. genetic drift. *Genetics*, 2005. **169**(2): p. 1033-43.

90. Lerat, E. and P. Capy: **Retrotransposons and retroviruses: analysis of the envelope gene**. *Mol Biol Evol*, 1999. **16**(9): p. 1198-207.

91. Lerat, E., C. Rizzon and C. Biemont: **Sequence divergence within transposable element families in the Drosophila melanogaster genome**. *Genome Res*, 2003. **13**(8): p. 1889-96.

92. Levis, R.W., R. Ganesan, K. Houtchens, L.A. Tolar and F.M. Sheen: **Transposons in place of telomeric repeats at a Drosophila telomere**. *Cell*, 1993. **75**(6): p. 1083-93.

93. Li, W.H., *Molecular Evolution*, ed. S. Sinauer Associates. 1997. p 191.

94. Lobo, N., X. Li and M.J. Fraser, Jr.: **Transposition of the piggyBac element in embryos of Drosophila melanogaster, Aedes aegypti and Trichoplusia ni**. *Mol Gen Genet*, 1999. **261**(4-5): p. 803-10.

95. Makalowski, W.: **Genomic scrap yard: how genomes utilize all that junk**. *Gene*, 2000. **259**(1-2): p. 61-7.

96. Malik, H.S. and T.H. Eickbush: **Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons**. *J Virol*, 1999. **73**(6): p. 5186-90.

97. Marrelli, M.T., C. Li, J.L. Rasgon and M. Jacobs-Lorena: **Transgenic malaria-resistant mosquitoes have a fitness advantage when feeding on Plasmodium-infected blood**. *Proc Natl Acad Sci U S A*, 2007. **104**(13): p. 5580-3.

98. McCarthy, E.M. and J.F. McDonald: **LTR_STRUC: a novel search and**

identification program for LTR retrotransposons. *Bioinformatics*, 2003. **19**(3): p. 362-7.

99. McClintock, B.: **The origin and behavior of mutable loci in maize**. *Proc Natl Acad Sci U S A*, 1950. **36**(6): p. 344-55.

100. McClure, M.A.: **Evolution of retroposons by acquisition or deletion of retrovirus-like genes**. *Mol Biol Evol*, 1991. **8**(6): p. 835-56.

101. McDonald, J.F.: **Evolution and consequences of transposable elements**. *Curr Opin Genet Dev*, 1993. **3**(6): p. 855-64.

102. McFadden, J. and G. Knowles: **Escape from evolutionary stasis by transposon-mediated deleterious mutations**. *J Theor Biol*, 1997. **186**(4): p. 441-7.

103. Meyers, B.C., S.V. Tingey and M. Morgante: **Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome**. *Genome Res*, 2001. **11**(10): p. 1660-76.

104. Miller, L.H., D.I. Baruch, K. Marsh and O.K. Doumbo: **The pathogenic basis of malaria**. *Nature*, 2002. **415**(6872): p. 673-9.

105. Miskey, C., Z. Izsvak, R.H. Plasterk and Z. Ivics: **The Frog Prince: a reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells**. *Nucleic Acids Res*, 2003. **31**(23): p. 6873-81.

106. Moreira, L.A., M.J. Edwards, F. Adhami, N. Jasinskiene, A.A. James and M. Jacobs-Lorena: **Robust gut-specific gene expression in transgenic Aedes aegypti mosquitoes**. *Proc Natl Acad Sci U S A*, 2000. **97**(20): p. 10895-8.

107. Moreira, L.A., A.K. Ghosh, E.G. Abraham and M. Jacobs-Lorena: **Genetic transformation of mosquitoes: a quest for malaria control**. *Int J Parasitol*, 2002. **32**(13): p. 1599-605.

108. Nekrutenko, A. and W.H. Li: **Transposable elements are found in a large number of human protein-coding genes**. *Trends Genet*, 2001. **17**(11): p. 619-21.

109. Nene, V., J.R. Wortman, D. Lawson, B. Haas, C. Kodira, Z.J. Tu, B. Loftus, Z. Xi, K. Megy, M. Grabherr, et al.: **Genome sequence of Aedes aegypti, a major arbovirus vector**. *Science*, 2007. **316**(5832): p. 1718-23.

110. Nolan, T., T.M. Bower, A.E. Brown, A. Crisanti and F. Catteruccia: **piggyBac-mediated germline transformation of the malaria mosquito Anopheles stephensi using the red fluorescent protein dsRED as a selectable marker**. *J Biol Chem*, 2002. **277**(11): p. 8759-62.

111. Nosten, F., R. McGready, J.A. Simpson, K.L. Thwai, S. Balkan, T. Cho, L. Hkirijaroen, S. Looareesuwan and N.J. White: **Effects of Plasmodium vivax malaria in pregnancy**. *Lancet*, 1999. **354**(9178): p. 546-9.

112. O'Brochta, D.A., N. Sethuraman, R. Wilson, R.H. Hice, A.C. Pinkerton, C.S. Levesque, D.K. Bideshi, N. Jasinskiene, C.J. Coates, A.A. James, et al.: **Gene vector and transposable element behavior in mosquitoes**. *J Exp Biol*, 2003. **206**(Pt 21): p. 3823-34.

113. O'Brochta, D.A., W.D. Warren, K.J. Saville and P.W. Atkinson: **Hermes, a functional non-Drosophilid insect gene vector from Musca domestica**. *Genetics*, 1996. **142**(3): p. 907-14.

114. O'Hare, K. and G.M. Rubin: **Structures of P transposable elements and their sites of insertion and excision in the Drosophila melanogaster genome**. *Cell*, 1983. **34**(1): p. 25-35.

115. Organization, W.H., *World Malaria Report, 2008*. 2008, WHO: Geneve.

116. Orgel, L.E. and F.H. Crick: **Selfish DNA: the ultimate parasite**. *Nature*, 1980. **284**(5757): p. 604-7.

117. Osta, M.A., G.K. Christophides and F.C. Kafatos: **Effects of mosquito genes on Plasmodium development**. *Science*, 2004. **303**(5666): p. 2030-2.

118. Patterson, R.S., D.E. Weidhaas, H.R. Ford and C.S. Lofgren: **Suppression and elimination of an island population of Culex pipiens quinquefasciatus with sterile males**. *Science*, 1970. **168**(937): p. 1368-70.

119. Pavlicek A., K.V.V., Jurka J: **"COPIA3_AG, a family of nonautonomous, copia-like LTR retrotransposons from African malaria mosquito."** *Repbase Reports*, 2003. **3**(3): p. 51.

120. Peloquin, J.J., S.T. Thibault, R. Staten and T.A. Miller: **Germ-line transformation of pink bollworm (Lepidoptera: gelechiidae) mediated by the piggyBac transposable element**. *Insect Mol Biol*, 2000. **9**(3): p. 323-33.

121. Petrov, D.A. and D.L. Hartl: **Trash DNA is what gets thrown away: high rate of DNA loss in Drosophila**. *Gene*, 1997. **205**(1-2): p. 279-89.

122. Petrov, D.A. and D.L. Hartl: **High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups**. *Mol Biol Evol*,

1998. **15**(3): p. 293-302.

123. Petrov, D.A., E.R. Lozovskaya and D.L. Hartl: **High intrinsic rate of DNA loss in Drosophila**. *Nature*, 1996. **384**(6607): p. 346-9.

124. Quesneville, H. and D. Anxolabehere: **A simulation of P element horizontal transfer in Drosophila**. *Genetica*, 1997. **100**(1-3): p. 295-307.

125. Quesneville, H. and D. Anxolabehere: **Dynamics of transposable elements in metapopulations: a model of P element invasion in Drosophila**. *Theor Popul Biol*, 1998. **54**(2): p. 175-93.

126. Quesneville, H., C.M. Bergman, O. Andrieu, D. Autard, D. Nouaud, M. Ashburner and D. Anxolabehere: **Combined evidence annotation of transposable elements in genome sequences**. *PLoS Comput Biol*, 2005. **1**(2): p. 166-75.

127. Quesneville, H., D. Nouaud and D. Anxolabehere: **Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes**. *J Mol Evol*, 2003. **57 Suppl 1**: p. S50-9.

128. Quesneville, H., D. Nouaud and D. Anxolabehere: **P elements and MITE relatives in the whole genome sequence of Anopheles gambiae**. *BMC Genomics*, 2006. **7**: p. 214.

129. Rho, M., J.H. Choi, S. Kim, M. Lynch and H. Tang: **De novo identification of LTR retrotransposons in eukaryotic genomes**. *BMC Genomics*, 2007. **8**: p. 90.

130. Riehle, M.A., P. Srinivasan, C.K. Moreira and M. Jacobs-Lorena: **Towards genetic manipulation of wild mosquito populations to combat malaria: advances and challenges**. *J Exp Biol*, 2003. **206**(Pt 21): p.

3809-16.

131. Robinson, A.S.: **Genetic sexing strains in medfly, Ceratitis capitata, sterile insect technique programmes**. *Genetica*, 2002. **116**(1): p. 5-13.

132. Robinson, A.S.: **Mutations and their use in insect control**. *Mutat Res*, 2002. **511**(2): p. 113-32.

133. Rohr, C.J., H. Ranson, X. Wang and N.J. Besansky: **Structure and evolution of mtanga, a retrotransposon actively expressed on the Y chromosome of the African malaria vector Anopheles gambiae**. *Mol Biol Evol*, 2002. **19**(2): p. 149-62.

134. Rozas, J., J.C. Sanchez-DelBarrio, X. Messeguer and R. Rozas: **DnaSP, DNA polymorphism analyses by the coalescent and other methods**. *Bioinformatics*, 2003. **19**(18): p. 2496-7.

135. Rubin, E. and A.A. Levy: **Abortive gap repair: underlying mechanism for Ds element formation**. *Mol Cell Biol*, 1997. **17**(11): p. 6294-302.

136. Rubin, G.M. and A.C. Spradling: **Genetic transformation of Drosophila with transposable element vectors**. *Science*, 1982. **218**(4570): p. 348-53.

137. Sabot, F. and A.H. Schulman: **Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome**. *Heredity*, 2006. **97**(6): p. 381-8.

138. Sabot, F., P. Sourdille, N. Chantret and M. Bernard: **Morgane, a new LTR retrotransposon group, and its subfamilies in wheats**. *Genetica*, 2006. **128**(1-3): p. 439-47.

139. Saitou, N. and M. Nei: **The neighbor-joining method: a new method for**

reconstructing phylogenetic trees. *Mol Biol Evol*, 1987. **4**(4): p. 406-25.

140. SanMiguel, P., B.S. Gaut, A. Tikhonov, Y. Nakajima and J.L. Bennetzen: **The paleontology of intergene retrotransposons of maize**. *Nat Genet*, 1998. **20**(1): p. 43-5.

141. Scott, T.W., W. Takken, B.G. Knols and C. Boete: **The ecology of genetically modified mosquitoes**. *Science*, 2002. **298**(5591): p. 117-9.

142. Seberg, O. and G. Petersen: **A unified classification system for eukaryotic transposable elements should reflect their phylogeny**. *Nat Rev Genet*, 2009. **10**(4): p. 276.

143. Sharma, V.P., R.K. Razdan and M.A. Ansari: **Anopheles stephensi: effect of gamma-radiation and chemosterilants on the fertility and fitness of males for sterile male releases**. *J Econ Entomol*, 1978. **71**(3): p. 449-50.

144. Silva, J.C., S.A. Shabalina, D.G. Harris, J.L. Spouge and A.S. Kondrashovi: **Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes**. *Genet Res*, 2003. **82**(1): p. 1-18.

145. Sweredoski, M., L. DeRose-Wilson and B.S. Gaut: **A comparative computational analysis of nonautonomous helitron elements between maize and rice**. *BMC Genomics*, 2008. **9**: p. 467.

146. Tamura, K., J. Dudley, M. Nei and S. Kumar: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0**. *Mol Biol Evol*, 2007. **24**(8): p. 1596-9.

147. Tamura, T., C. Thibert, C. Royer, T. Kanda, E. Abraham, M. Kamba, N. Komoto, J.L. Thomas, B. Mauchamp, G. Chavancy, et al.: **Germline transformation of the silkworm Bombyx mori L. using a piggyBac transposon-derived vector**. *Nat Biotechnol*, 2000. **18**(1): p. 81-4.

148. Tatusova, T.A. and T.L. Madden: **BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences**. *FEMS Microbiol Lett*, 1999. **174**(2): p. 247-50.

149. Thomas, D.D., C.A. Donnelly, R.J. Wood and L.S. Alphey: **Insect population control using a dominant, repressible, lethal genetic system**. *Science*, 2000. **287**(5462): p. 2474-6.

150. Toure, Y.T., A.M. Oduola and C.M. Morel: **The Anopheles gambiae genome: next steps for malaria vector control**. *Trends Parasitol*, 2004. **20**(3): p. 142-9.

151. Tu, Z. and C. Coates: **Mosquito transposable elements**. *Insect Biochem Mol Biol*, 2004. **34**(7): p. 631-44.

152. Tubio, J.M., J.C. Costas and H.F. Naveira: **Evolution of the mdg1 lineage of the Ty3/gypsy group of LTR retrotransposons in Anopheles gambiae**. *Gene*, 2004. **330**: p. 123-31.

153. Tubio, J.M., H. Naveira and J. Costas: **Structural and evolutionary analyses of the Ty3/gypsy group of LTR retrotransposons in the genome of Anopheles gambiae**. *Mol Biol Evol*, 2005. **22**(1): p. 29-39.

154. WHO, *World malaria report 2008.* 2008, World Health Organization: Geneva Switzerland.

155. Wicker, T., R. Guyot, N. Yahiaoui and B. Keller: **CACTA transposons in**

Triticeae. A diverse family of high-copy repetitive elements. *Plant Physiol*, 2003. **132**(1): p. 52-63.

156. Wicker, T., F. Sabot, A. Hua-Van, J.L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, et al.: **A unified classification system for eukaryotic transposable elements**. *Nat Rev Genet*, 2007. **8**(12): p. 973-82.

157. Wicker T, F.S., Aurélie Hua-Van, Jeffrey L. Bennetzen,, B.C. Pierre Capy, Andrew Flavell, Philippe Leroy,, O.P. Michele Morgante, Etienne Paux, Phillip SanMiguel and and A.H. Schulman: **Reply: A unified classification system for eukaryotic transposable elements should reflect their phylogeny**. *Nat Rev Genet*, 2009. **10**(276).

158. Witte, C.P., Q.H. Le, T. Bureau and A. Kumar: **Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes**. *Proc Natl Acad Sci U S A*, 2001. **98**(24): p. 13778-83.

159. Xiong, Y. and T.H. Eickbush: **Origin and evolution of retroelements based upon their reverse transcriptase sequences**. *Embo J*, 1990. **9**(10): p. 3353-62.

160. Yen, J.H. and A.R. Barr: **New hypothesis of the cause of cytoplasmic incompatibility in Culex pipiens L**. *Nature*, 1971. **232**(5313): p. 657-8.

161. Zhang, X. and S.R. Wessler: **Genome-wide comparative analysis of the transposable elements in the related species Arabidopsis thaliana and Brassica oleracea**. *Proc Natl Acad Sci U S A*, 2004. **101**(15): p. 5589-94.

ANEXOS

ANEXO 1

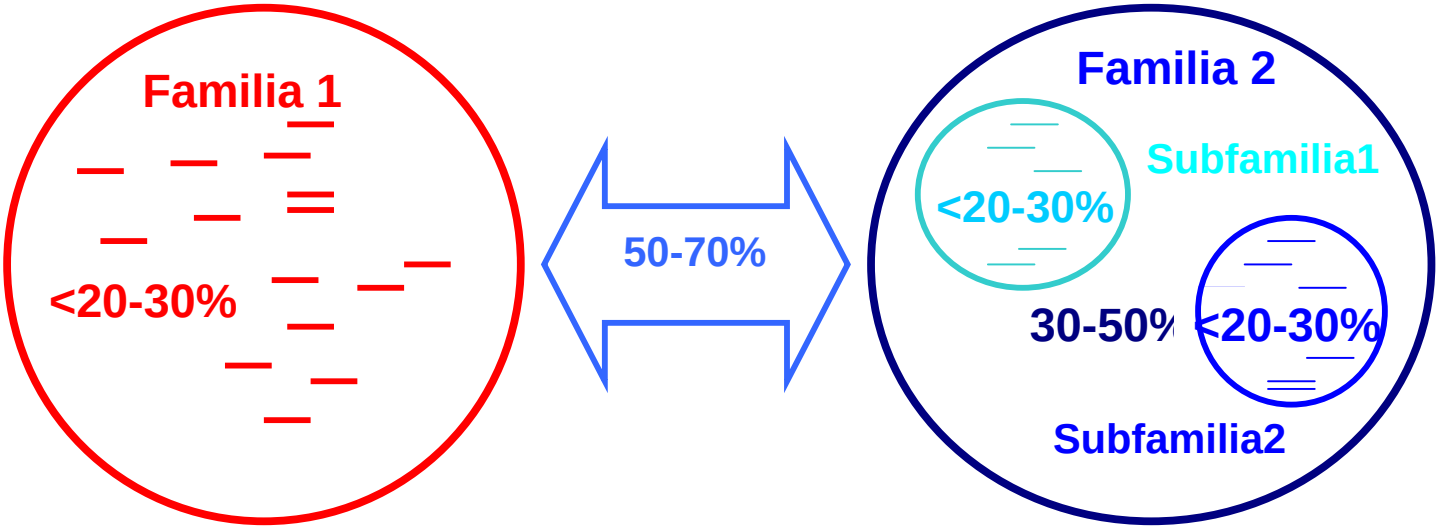
Clase (presencia de intermediarios de RNA en la transposición)

Subclase (para elementos DNA -Clase II- depende del número de copias que son cortadas en el sitio donante)

Orden (Diferencias importantes a nivel del mecanismo de inserción)

Superfamilia (comparten una estrategia de replicación, se distinguen por características comunes ampliamente distribuidas. No hay conservación a nivel nucleotídico, ni a nivel de la estructura génica. Puede haber similitudes limitadas a nivel aminoacídico)

Familia (Conservación a nivel de DNA. Alta similitud a nivel proteico. Criterio de Wicker >80% de identidad; >80% tamaño; >80 nts). Capy considera hasta 30% de distancia como misma familia)



Cuadro. Esquema representando las relaciones y distancias a nivel aminoacídico para caracterizar los diferentes taxa propuestos para clasificación de elementos de transposición basado en Wicker et al 2007 y Capy, 1998.

Tabla 1	*Clase*	*Subclase*	*Orden*	*Superfamilia*	*Familia*	*Subfamilia*	Inserción
Distingue entre ➔	**Diferentes Mecanismos de Transposición**	**Solo para Clase II. Número de hebras de DNA cortadas en el momento de la excisión**	**Diferencias importantes en los mecanismos de inserción**	**Dentro del mismo Orden, comparten estrategias de replicación. No presentan conservación de secuencia a nivel nucleotídico, solo similitud limitada a nivel proteico**	**Conservación a nivel de secuencias de DNA. Alta similitud a nivel proteico entre diferentes familias dentro de la misma superfamilia. Un genoma puede tener miles de familias de TEs. Se utiliza la regla 80-80-80 para asignar secuencias a una familia. i.e. >80% de identidad en >80% largo en secuencias con >80nucleótidos**	**Se define filogenéticamente. Puede distinguir entre poblaciones de elementos autónomos y no autónomos. Este taxon depende de la estructura poblacional de una familia.**	**Copia individual, que correspon de a un evento específico de transposic ión.**
	I		**LTR**	**Copia**			
				Gypsy			
				Pao-Bel			
				Retrovirus			
				ERV			
			DIRS	**DIRS**			
				Nigaro			
				VIPER			
			PLE	**Penelope**			
			LINE	**R2**			
				((R1))			
				RTE			
				((CR1))			
				Jockey			
				L1			
				I			
				?			
				?			
				((RT))			
			SINE	**tRNA**			
				7SL			
				5SL			
				SINEX1-2_AG			
	II	**1**	**TIR**	**Tc1-Mariner**			
				hAT			
				Mutator			
				Merlinn			
				Transib			
				P			
				PiggyBac			
				PIF-Harbinger			
				CACTA			
			Crypton	**Crypton**			
		2	**Helitron**	**Helitron**			
			Maverick	**Maverick**			

ANEXO 2

	BEL12-I_AG	BEL9-I_AG	BEL5-I_AG	BEL13-I_AG	BEL4-I_AG	BEL11-I_AG	BEL15-I_AG	BEL8-I_AG	BEL16-I_AG	BEL6-I_AG	BEL2-I_AG	BEL7-I_AG	BEL10-I_AG	BEL1-I_AG	BEL18-I_AG	BEL3-I_AG	BEL14-I_AG
BEL12-I_AG																	
BEL9-I_AG	0,5115																
BEL5-I_AG	0,6	0,5929															
BEL13-I_AG	0,6212	0,6319	0,6177														
BEL4-I_AG	0,6212	0,6053	0,5965	0,4867													
BEL11-I_AG	0,6195	0,623	0,6088	0,6283	0,5894												
BEL15-I_AG	0,6106	0,6035	0,6124	0,6177	0,6053	0,531											
BEL8-I_AG	0,6195	0,6319	0,6159	0,6018	0,5823	0,5558	0,4389										
BEL16-I_AG	0,6248	0,6319	0,6265	0,623	0,6283	0,6142	0,5947	0,6									
BEL6-I_AG	0,6796	0,6743	0,6496	0,6442	0,6496	0,669	0,6885	0,6761	0,6726								
BEL2-I_AG	0,7876	0,7735	0,7858	0,7788	0,7699	0,7876	0,777	0,7805	0,7752	0,7894							
BEL7-I_AG	0,8212	0,8124	0,8319	0,8159	0,8195	0,8159	0,8124	0,8177	0,8053	0,8389	0,4549						
BEL10-I_AG	0,7664	0,7646	0,7735	0,7681	0,777	0,7788	0,7681	0,7699	0,7717	0,7805	0,4195	0,5611					
BEL1-I_AG	0,7752	0,7841	0,777	0,7823	0,7876	0,7894	0,7823	0,7841	0,7699	0,7876	0,5823	0,6743	0,5699				
BEL18-I_AG	0,7788	0,7735	0,7664	0,7788	0,7982	0,7841	0,7611	0,7681	0,7646	0,7982	0,6442	0,6938	0,6265	0,6354			
BEL3-I_AG	0,7628	0,7735	0,7717	0,7912	0,7929	0,7858	0,7717	0,7841	0,777	0,8035	0,6637	0,7097	0,6566	0,6442	0,554		
BEL14-I_AG	0,7699	0,7699	0,7752	0,7593	0,7929	0,7717	0,7611	0,7788	0,7788	0,7982	0,6531	0,7115	0,646	0,6619	0,5929	0,6177	
BEL17-I_AG	0,7965	0,7735	0,7752	0,7823	0,7947	0,7823	0,7593	0,7717	0,7681	0,7858	0,6566	0,7168	0,6584	0,6814	0,5841	0,6283	0,3646

media	**0,702330065**
MAX	**0,8389**
min	**0,3646**

Matriz_dist-p_Elementos *Pao-Bel*

GYPSY11_AG																																					
GYPSY15_	0,0299							0,235366																													
GYPSY8_A	0,1109	0,1258																																			
GYPSY9_A	0,1578	0,1791	0,1876																																		
GYPSY12_	0,1599	0,1791	0,1812	0,0235																																	
GYPSY16_	0,1983	0,2132	0,2154	0,2196	0,2239																																
GYPSY13_	0,2154	0,2281	0,2388	0,2409	0,2388	0,2431																															
GYPSY14_	0,2068	0,2175	0,2239	0,2345	0,2281	0,2409	0,1684																														
GYPSY10_	0,3092	0,3177	0,307	0,3177	0,3092	0,3156	0,3433	0,3177																													
GYPSY17_	0,3028	0,3113	0,2942	0,3156	0,307	0,3092	0,3348	0,3134	0,0405																												
GYPSY54_	0,5501	0,565	0,5565	0,548	0,548	0,5544	0,5544	0,5565	0,5714	0,565																											
GYPSY46_	0,6311	0,6418	0,6333	0,6311	0,6333	0,6141	0,6269	0,6162	0,6397	0,6354	0,6226																										
GYPSY47_	0,6311	0,6418	0,629	0,6333	0,6269	0,6183	0,6311	0,6183	0,629	0,6183	0,6247	0,1962								0,453569																	
GYPSY44_	0,6482	0,6631	0,6418	0,6482	0,6482	0,6439	0,6482	0,6375	0,6354	0,6397	0,629	0,4222	0,4179																								
GYPSY45_	0,6311	0,6418	0,6354	0,6205	0,6205	0,6418	0,6375	0,6354	0,6354	0,6333	0,6183	0,4264	0,4371	0,2217																							
GYPSY39_	0,5864	0,6034	0,5949	0,5864	0,5842	0,5928	0,5864	0,5864	0,6226	0,6141	0,5778	0,5075	0,5075	0,5011	0,5139																						
GYPSY40_	0,5885	0,6034	0,5949	0,5821	0,58	0,58	0,5949	0,597	0,6162	0,6162	0,5906	0,5139	0,5203	0,5288	0,5288	0,2324																					
GYPSY41_	0,5885	0,6034	0,597	0,597	0,597	0,5928	0,6077	0,6077	0,6098	0,6162	0,6119	0,5245	0,4989	0,5075	0,5139	0,4542	0,4371																				
GYPSY42_	0,6247	0,6354	0,6226	0,6205	0,6183	0,6269	0,6226	0,6205	0,6269	0,6247	0,6183	0,5181	0,5075	0,5053	0,5224	0,4712	0,4691	0,3177																			
GYPSY43_	0,597	0,6098	0,597	0,597	0,5949	0,5906	0,6055	0,6055	0,6077	0,6077	0,597	0,5181	0,5117	0,5053	0,5373	0,4563	0,4584	0,3241	0,2942																		
GYPSY35_	0,7015	0,7079	0,7058	0,7015	0,6994	0,7036	0,6951	0,6738	0,6972	0,6951	0,7122	0,7249	0,71	0,7036	0,7015	0,6994	0,678	0,6823	0,6802	0,6823																	
GYPSY72_	0,7036	0,7079	0,7143	0,71	0,7079	0,693	0,7036	0,6823	0,7036	0,7036	0,7186	0,7271	0,7186	0,7143	0,7058	0,6972	0,6908	0,6823	0,6994	0,6972	0,2964																
GYPSY36_	0,7036	0,7122	0,7207	0,7186	0,7207	0,7036	0,693	0,6759	0,6994	0,6972	0,71	0,7143	0,7164	0,6951	0,6994	0,6972	0,7015	0,6823	0,6823	0,6866	0,3241	0,3049															
GYPSY71_	0,7036	0,7079	0,7015	0,7079	0,7079	0,6972	0,6951	0,678	0,7079	0,7079	0,7058	0,6908	0,6844	0,6695	0,6802	0,6759	0,693	0,6823	0,6738	0,6887	0,5075	0,4968	0,4883														
GYPSY37_	0,6994	0,71	0,7036	0,6972	0,693	0,6994	0,6802	0,661	0,6972	0,6951	0,7143	0,7036	0,6972	0,678	0,6951	0,6759	0,6823	0,661	0,6674	0,6695	0,5032	0,4947	0,4968	0,4264													
GYPSY38_	0,6951	0,7079	0,7058	0,7122	0,7058	0,7058	0,6823	0,6802	0,6866	0,6866	0,7143	0,6994	0,6972	0,6866	0,6994	0,6738	0,6674	0,6567	0,678	0,661	0,5053	0,4797	0,5053	0,4456	0,2857												
GYPSY29_	0,661	0,6716	0,6716	0,6652	0,661	0,6652	0,6652	0,6567	0,6887	0,6844	0,6652	0,6418	0,6183	0,6269	0,661	0,6226	0,6397	0,6461	0,6397	0,6269	0,6055	0,5821	0,5885	0,5522	0,548	0,5885											
GYPSY31_	0,6588	0,6695	0,6631	0,6631	0,6588	0,6652	0,6738	0,6588	0,6866	0,6823	0,6866	0,6375	0,6354	0,6247	0,6461	0,6439	0,6588	0,6439	0,6418	0,6375	0,5991	0,5906	0,5928	0,5544	0,5544	0,5821	0,2367										
GYPSY30_	0,661	0,6695	0,6674	0,6567	0,6546	0,6546	0,6716	0,6525	0,6802	0,6759	0,6887	0,6397	0,6354	0,6397	0,6503	0,6397	0,6461	0,6354	0,6226	0,6333	0,5714	0,5736	0,5842	0,5522	0,5522	0,5757	0,2537	0,2623									
GYPSY33_	0,6588	0,6695	0,6716	0,6631	0,661	0,6695	0,6674	0,6503	0,6738	0,6759	0,6823	0,6695	0,6674	0,6525	0,6738	0,6482	0,6567	0,6482	0,6397	0,6418	0,58	0,5778	0,5906	0,548	0,5544	0,5778	0,2623	0,2644	0,2431								
GYPSY34_	0,6503	0,6588	0,6631	0,6503	0,6482	0,6503	0,661	0,6375	0,6716	0,6738	0,6695	0,6674	0,6631	0,6482	0,6695	0,6226	0,6397	0,6439	0,6439	0,6354	0,5906	0,5842	0,597	0,5458	0,5586	0,5565	0,2729	0,2687	0,2601	0,2196							
GYPSY32_	0,6588	0,6695	0,6716	0,6652	0,6631	0,661	0,6695	0,6567	0,6631	0,661	0,6674	0,6567	0,6567	0,6461	0,661	0,6397	0,6397	0,6588	0,6354	0,6247	0,6013	0,5778	0,5885	0,5501	0,5714	0,565	0,2516	0,2495	0,2452	0,209	0,1663						
GYPSY49_	0,6695	0,678	0,661	0,6716	0,6652	0,6525	0,6695	0,6695	0,6823	0,6802	0,6013	0,6759	0,6652	0,6866	0,6844	0,6525	0,6439	0,6631	0,6674	0,6631	0,6866	0,7058	0,7015	0,7207	0,6994	0,6802	0,6823	0,6844	0,6823	0,6908	0,6695	0,6823					
GYPSY48_	0,6695	0,6738	0,6695	0,6631	0,661	0,661	0,661	0,6674	0,6716	0,6716	0,6183	0,6738	0,678	0,6951	0,6802	0,6567	0,6525	0,6674	0,678	0,6652	0,693	0,71	0,7015	0,7271	0,7186	0,6972	0,6887	0,6908	0,6908	0,6972	0,6631	0,678	0,2431				
GYPSY50_	0,6631	0,6738	0,6652	0,6716	0,6695	0,661	0,6567	0,6631	0,6844	0,678	0,6418	0,7015	0,6994	0,6738	0,6802	0,6546	0,6652	0,6866	0,6994	0,6994	0,6908	0,71	0,693	0,7079	0,6823	0,7058	0,6759	0,6908	0,6695	0,678	0,661	0,6716	0,4478	0,4542			
GYPSY53_	0,6738	0,6802	0,6759	0,678	0,678	0,6652	0,6716	0,6738	0,7015	0,6994	0,6588	0,7164	0,7143	0,693	0,693	0,6951	0,6887	0,7015	0,7058	0,7036	0,7143	0,7143	0,7058	0,7377	0,7335	0,7441	0,7079	0,6972	0,6866	0,6972	0,6716	0,6887	0,4563	0,4542	0,4542		
GYPSY70_	0,6716	0,678	0,6759	0,6823	0,6844	0,6716	0,6759	0,6738	0,6951	0,6908	0,6439	0,7036	0,7164	0,6823	0,6908	0,6866	0,6866	0,6994	0,6994	0,6951	0,7036	0,7079	0,693	0,7207	0,71	0,7271	0,6908	0,6823	0,6866	0,693	0,6716	0,6866	0,4307	0,4478	0,4435	0,2708	
GYPSY52_	0,661	0,6652	0,6652	0,6716	0,6716	0,6525	0,6738	0,6588	0,6759	0,6759	0,6311	0,7036	0,7058	0,6738	0,678	0,6738	0,6759	0,6887	0,6866	0,6823	0,7079	0,7079	0,6972	0,7228	0,7207	0,7292	0,6844	0,6887	0,6887	0,6866	0,6588	0,6674	0,4307	0,4435	0,4456	0,2601	0,2175
GYPSY51_	0,6567	0,661	0,6567	0,6695	0,6695	0,6567	0,6738	0,661	0,6866	0,6823	0,6397	0,6908	0,6951	0,6695	0,6674	0,6759	0,6844	0,6844	0,6844	0,678	0,7122	0,71	0,6972	0,7186	0,7143	0,7207	0,6908	0,6844	0,6844	0,6823	0,6695	0,6759	0,4414	0,4542	0,4627	0,2772	0,2559
GYPSY69_	0,6759	0,6823	0,6738	0,6802	0,6802	0,6738	0,6844	0,6844	0,6844	0,6823	0,6759	0,6951	0,6908	0,7228	0,7143	0,693	0,7036	0,6759	0,693	0,6674	0,7186	0,6887	0,7228	0,693	0,71	0,7122	0,7058	0,7186	0,7122	0,7122	0,7058	0,71	0,7122	0,6844	0,742	0,7292	0,742
GYPSY68_	0,71	0,7207	0,71	0,7143	0,7122	0,7015	0,7036	0,7058	0,7015	0,6994	0,6908	0,7292	0,7186	0,7271	0,7164	0,6823	0,6866	0,6844	0,7058	0,6844	0,7292	0,7463	0,7377	0,7292	0,7186	0,7249	0,7292	0,7271	0,7335	0,7356	0,7313	0,7441	0,7143	0,7122	0,7207	0,7313	0,7271
GYPSY65_	0,6759	0,6823	0,6823	0,6738	0,6716	0,6823	0,6823	0,6802	0,6866	0,6844	0,6759	0,7079	0,7164	0,7292	0,7313	0,6844	0,6972	0,7015	0,7228	0,6866	0,7058	0,7015	0,7079	0,7143	0,71	0,7207	0,6908	0,693	0,6887	0,6802	0,6823	0,6908	0,6866	0,6738	0,6994	0,7207	0,6951
GYPSY20_	0,71	0,7122	0,71	0,7143	0,7143	0,6972	0,7164	0,7271	0,7122	0,7122	0,7122	0,7356	0,7377	0,7271	0,7186	0,6972	0,6994	0,6887	0,7015	0,6908	0,7207	0,7186	0,7271	0,7249	0,7079	0,7335	0,7207	0,7186	0,7015	0,7186	0,7015	0,71	0,693	0,6866	0,7036	0,7015	0,6823
GYPSY23_	0,7207	0,7249	0,7164	0,7164	0,7143	0,7143	0,7164	0,7249	0,7015	0,7015	0,7058	0,7356	0,7356	0,7186	0,7122	0,7015	0,7036	0,6951	0,6866	0,6866	0,7292	0,7228	0,742	0,7313	0,7186	0,7186	0,7228	0,7143	0,7036	0,7164	0,6972	0,7079	0,6994	0,6908	0,7164	0,7079	0,6994
GYPSY22_	0,7164	0,7207	0,7228	0,7207	0,7186	0,6972	0,7079	0,7249	0,71	0,7058	0,71	0,7335	0,7356	0,7079	0,6994	0,6908	0,7164	0,6738	0,6844	0,6738	0,742	0,7249	0,7228	0,7292	0,7058	0,7122	0,7122	0,7143	0,6972	0,7249	0,7036	0,71	0,6887	0,678	0,7036	0,6844	0,6844
GYPSY26_	0,7399	0,7441	0,7313	0,7271	0,7271	0,7164	0,7271	0,7313	0,7122	0,7143	0,7207	0,742	0,7377	0,7015	0,7164	0,7036	0,71	0,7015	0,7143	0,6972	0,7399	0,7228	0,7164	0,7207	0,7079	0,7335	0,7292	0,7228	0,7207	0,7228	0,7058	0,7143	0,7079	0,6994	0,7058	0,7271	0,7207
GYPSY28_	0,742	0,7441	0,7441	0,7271	0,7292	0,7249	0,7356	0,7377	0,7271	0,7292	0,7164	0,7292	0,7292	0,71	0,7122	0,6972	0,7207	0,693	0,7058	0,6866	0,742	0,7207	0,7249	0,7292	0,71	0,7399	0,7463	0,7186	0,7399	0,742	0,7271	0,7335	0,7122	0,7015	0,7122	0,7356	0,7143
GYPSY19_	0,7356	0,7399	0,742	0,7313	0,7292	0,7335	0,7292	0,7335	0,7313	0,7292	0,71	0,7441	0,7463	0,7036	0,7164	0,7079	0,7143	0,6908	0,7058	0,6823	0,7505	0,7335	0,7335	0,7292	0,7335	0,7207	0,7122	0,7186	0,7143	0,7271	0,7058	0,7186	0,7058	0,6887	0,7143	0,7122	0,7079
GYPSY24_	0,7313	0,7356	0,7377	0,7186	0,7164	0,7122	0,7313	0,7377	0,7207	0,7207	0,7143	0,7186	0,7079	0,7036	0,6951	0,6908	0,7015	0,6823	0,7058	0,6802	0,7186	0,7058	0,7143	0,7228	0,7186	0,7207	0,7292	0,7484	0,7228	0,7463	0,7399	0,7377	0,693	0,6951	0,6951	0,7399	0,7228
GYPSY66_	0,7441	0,7484	0,7377	0,7271	0,7271	0,7164	0,7292	0,7313	0,7292	0,7271	0,7271	0,7271	0,7207	0,7143	0,6994	0,6844	0,7058	0,6908	0,7079	0,6908	0,7249	0,6994	0,7186	0,7122	0,7249	0,7292	0,7207	0,7271	0,71	0,7292	0,7207	0,7228	0,6908	0,693	0,6951	0,7335	0,7143
GYPSY18_	0,7143	0,7207	0,7207	0,71	0,7122	0,7015	0,7164	0,7122	0,7079	0,7058	0,6738	0,7292	0,7335	0,7122	0,7228	0,6823	0,693	0,6738	0,6994	0,6866	0,7186	0,7036	0,7143	0,7079	0,6994	0,6908	0,6908	0,6951	0,7122	0,7036	0,6866	0,6866	0,6844	0,6738	0,7036	0,7143	0,6866
GYPSY62_	0,71	0,7207	0,7228	0,7143	0,7143	0,7015	0,7271	0,7249	0,7271	0,7249	0,7079	0,7271	0,7377	0,7036	0,71	0,6866	0,7122	0,7015	0,7164	0,6972	0,7271	0,7079	0,7186	0,7164	0,6994	0,7164	0,693	0,7058	0,7058	0,7143	0,6994	0,7015	0,7079	0,7143	0,7143	0,6972	0,6759
GYPSY63_	0,71	0,7164	0,7207	0,71	0,71	0,6951	0,7186	0,7186	0,71	0,7079	0,6972	0,7249	0,7271	0,6887	0,7036	0,678	0,7143	0,6887	0,7164	0,6887	0,7377	0,7228	0,7313	0,7164	0,7122	0,7164	0,7036	0,7058	0,7143	0,7143	0,7143	0,71	0,6972	0,7036	0,7143	0,6972	0,6738
GYPSY21_	0,7079	0,7143	0,7122	0,6887	0,6908	0,6908	0,7079	0,7122	0,7249	0,7207	0,7122	0,7228	0,7228	0,7079	0,7058	0,6866	0,7058	0,6972	0,7079	0,7058	0,7207	0,7058	0,6994	0,6972	0,6972	0,7015	0,6887	0,6866	0,6866	0,6738	0,693	0,693	0,6844	0,6972	0,6951	0,7036	0,678
GYPSY64_	0,7186	0,7228	0,7228	0,7122	0,7164	0,6951	0,7228	0,7313	0,7207	0,7122	0,6994	0,7335	0,7399	0,7484	0,7356	0,7036	0,7122	0,7079	0,7143	0,6951	0,7164	0,7079	0,7356	0,7271	0,7271	0,7292	0,7079	0,7186	0,7058	0,7058	0,71	0,7122	0,6994	0,6951	0,71	0,7228	0,6994
GYPSY27_	0,7122	0,7228	0,7207	0,7058	0,7058	0,7015	0,71	0,7058	0,7143	0,7122	0,7143	0,7249	0,7335	0,7207	0,7164	0,7015	0,7186	0,7058	0,7015	0,7079	0,6994	0,6887	0,71	0,7058	0,71	0,7122	0,7058	0,7122	0,7228	0,7079	0,7015	0,6994	0,71	0,71	0,7377	0,7164	0,7036
GYPSY61_	0,7249	0,7313	0,7313	0,7143	0,7186	0,6972	0,7207	0,7164	0,7122	0,7079	0,7079	0,7143	0,7122	0,7079	0,7292	0,6844	0,6887	0,6866	0,6844	0,6759	0,71	0,7015	0,7122	0,7164	0,7207	0,7186	0,7079	0,6994	0,7058	0,7079	0,7058	0,6951	0,6908	0,6908	0,7015	0,7036	0,6823
GYPSY55_	0,7164	0,7207	0,7271	0,71	0,7143	0,6994	0,7228	0,7186	0,7249	0,7228	0,6994	0,7399	0,7399	0,7271	0,7292	0,71	0,7228	0,71	0,7271	0,7164	0,7292	0,6994	0,7207	0,7164	0,7313	0,7356	0,7143	0,7079	0,7143	0,7122	0,7015	0,7122	0,7036	0,7058	0,7271	0,6994	0,6802
GYPSY59_	0,7058	0,7122	0,7207	0,7164	0,7186	0,7036	0,7186	0,71	0,7271	0,7228	0,6802	0,7249	0,7313	0,7207	0,7079	0,6802	0,693	0,6823	0,7015	0,6866	0,7122	0,7015	0,7036	0,7164	0,7164	0,7207	0,7164	0,6994	0,6994	0,7079	0,7058	0,6994	0,693	0,6972	0,7143	0,6866	0,6652
GYPSY67_	0,7036	0,71	0,7164	0,7058	0,7122	0,6908	0,7079	0,7143	0,7292	0,7249	0,6738	0,7377	0,7527	0,742	0,7271	0,6887	0,7058	0,6972	0,7036	0,693	0,7249	0,7079	0,7377	0,7164	0,7335	0,7249	0,7164	0,7079	0,693	0,7036	0,6866	0,6972	0,693	0,678	0,7143	0,7058	0,6887
GYPSY56_	0,7207	0,7249	0,7356	0,7164	0,7164	0,71	0,7207	0,7122	0,7164	0,7164	0,6823	0,7313	0,7356	0,7079	0,7228	0,6994	0,7249	0,7015	0,7164	0,7015	0,7186	0,6994	0,6972	0,71	0,7143	0,7292	0,7186	0,71	0,7143	0,7186	0,7143	0,7143	0,7036	0,7015	0,7292	0,693	0,6823
GYPSY60_	0,7186	0,7228	0,7249	0,7271	0,7313	0,7058	0,7143	0,7143	0,7207	0,7143	0,7058	0,7271	0,7292	0,7271	0,7186	0,6951	0,7186	0,6994	0,7122	0,6908	0,7207	0,71	0,7079	0,7228	0,7122	0,7249	0,7036	0,7036	0,7058	0,7058	0,7015	0,693	0,71	0,7036	0,7292	0,71	0,7015
GYPSY25_	0,7143	0,7249	0,7356	0,7164	0,7186	0,7015	0,7079	0,71	0,7313	0,7271	0,7058	0,7335	0,742	0,7058	0,7228	0,7079	0,7271	0,6994	0,6951	0,7036	0,7143	0,6951	0,7164	0,6908	0,6994	0,7186	0,6802	0,7015	0,7036	0,7015	0,6866	0,6887	0,7058	0,7079	0,7335	0,7036	0,6908
GYPSY57_	0,7122	0,7207	0,7313	0,7186	0,7249	0,7122	0,7249	0,7207	0,7249	0,7249	0,7036	0,7292	0,742	0,7228	0,7271	0,7058	0,7143	0,6994	0,7164	0,71	0,7271	0,7164	0,7292	0,7292	0,7249	0,7377	0,71	0,7143	0,7143	0,7271	0,7058	0,7058	0,7271	0,7249	0,7527	0,7079	0,678
GYPSY58_	0,7058	0,7143	0,7271	0,6972	0,7015	0,6951	0,7015	0,6994	0,7058	0,7036	0,7079	0,7356	0,7399	0,7164	0,7228	0,7122	0,7292	0,7079	0,7079	0,7058	0,6951	0,7036	0,7058	0,7186	0,7143	0,7228	0,7058	0,7079	0,7058	0,7249	0,6887	0,6972	0,7079	0,7036	0,7399	0,693	0,6716

Matriz_dist-p_Elementos GYPSY em REPBASE

	COPIA2_AG	Mtanga_Ag	COPIA3_Ag	Clu172	COPIA1_Ag	COPIA4_Ag	Clu134	Clu150	COPIA5_Ag	Clu149
COPIA2_AG					**0,54143**					
Mtanga_Ag	0,5141									
COPIA3_Ag	0,5622	0,548							media	
Clu172	0,5622	0,548	**0**						entre	**0,67265**
COPIA1_Ag	0,6497	0,6299	0,6384	0,63842					COPIAs	
COPIA4_Ag	0,7062	0,6977	0,7401	0,74011	0,70621					
Clu134	0,7147	0,7034	0,7458	0,74576	0,68644	0,6017				
Clu150	0,7288	0,7062	0,7486	0,74859	0,70339	0,6328	0,46045			
COPIA5_Ag	0,7288	0,7062	0,7486	0,74859	0,70339	**0,6328**	**0,46045**	**0**		**0,7646**
Clu149	0,7712	0,7232	0,7571	0,75706	0,74859	0,8023	0,77401	0,7853	0,7853	

COPIA1-5_AG	
media	**0,6608**
MAX	**0,7486**
min	**0,5141**

Matriz_dist-p Elementos COPIA

	GYPSY42_AG	GYPSY43_AG	GYPSY41_AG	GYPSY39_AG	GYPSY40_AG	Gypsy46+47	GYPSY44_AG	GYPSY45_AG	GYPSY54_AG	Gypsy8-15	GYPSY49_AG	GYPSY48_AG	Clu131	GYPSY50_AG	Clu119	Gypsy53+52+51	Clu138	GYPSY29_AG	GYPSY31_AG	Clu104	Gypsy33+34+32	GYPSY30_AG	GYPSY37_AG	GYPSY38_AG
GYPSY42_AG								**0,4511525**																
GYPSY43_AG	0,29247							**0,35269**																
GYPSY41_AG	0,30968	0,32043		**GYPSY_lineage**				0,307526667																
GYPSY39_AG	0,47097	0,45376	0,44946																					
GYPSY40_AG	0,46882	0,45376	0,43656	0,22366																				
Gypsy46+47	0,51398	0,51183	0,51398	0,49892	0,50968																			
GYPSY44_AG	0,50108	0,50323	0,50753	0,49892	0,53118	0,4086																		
GYPSY45_AG	0,52258	0,53548	0,51613	0,50538	0,52473	0,42366	0,22581																	
GYPSY54_AG	0,61505	0,59355	0,60645	0,57419	0,5828	0,6172	0,62796	0,61505	**Mdg1_lineage**	**0,54624**														
Gypsy8-15	0,62151	0,58925	0,58065	0,58495	0,58495	0,63226	0,65806	0,63441	0,54624															
GYPSY49_AG	0,64946	0,64731	0,65376	0,64516	0,63656	0,65376	0,67957	0,67957	0,6043	0,66452						**0,378064**								
GYPSY48_AG	0,66022	0,65376	0,65806	0,64946	0,63871	0,65806	0,68387	0,67312	0,62151	0,66667	0,24301			**CsRn1_lineage**										
Clu131	0,68602	0,69032	0,67097	0,64731	0,65161	0,68817	0,67097	0,67527	0,63871	0,66452	0,44731	0,45591												
GYPSY50_AG	0,68602	0,69032	0,67097	0,64731	0,65161	0,68817	0,67097	0,67527	0,63871	0,66452	0,44731	0,45591	0											
Clu119	0,67097	0,66882	0,67097	0,65806	0,66667	0,67527	0,66022	0,65806	0,64086	0,65591	0,44516	0,46022	0,45806	0,45806										
Gypsy53+52+51	0,67312	0,67527	0,66882	0,65806	0,65806	0,67097	0,66237	0,66452	0,62796	0,65591	0,42151	0,43226	0,42796	0,42796	0,09032									
Clu138	0,62796	0,62796	0,63441	0,61935	0,63656	0,62796	0,62366	0,65591	0,66452	0,67742	0,67742	0,68817	0,68172	0,68172	0,68602	0,68387								
GYPSY29_AG	0,62796	0,62796	0,63441	0,61935	0,63656	0,62796	0,62366	0,65591	0,66452	0,67742	0,67742	0,68817	0,68172	0,68172	0,68602	0,68387	0							
GYPSY31_AG	0,63011	0,63871	0,63011	0,63871	0,65591	0,62796	0,61935	0,64086	0,69032	0,66882	0,67527	0,68602	0,68387	0,68387	0,68172	0,68602	0,23441	0,23441			**Mdg3_lineage**			
Clu104	0,62796	0,62581	0,64731	0,64301	0,64086	0,65161	0,64516	0,66022	0,67312	0,66667	0,67097	0,67527	0,66667	0,66667	0,67312	0,67097	0,24946	0,24946	0,24731					
Gypsy33+34+32	0,63226	0,63656	0,63011	0,63226	0,65376	0,65161	0,63871	0,65806	0,67097	0,65806	0,67097	0,67742	0,66237	0,66237	0,67742	0,66667	0,24516	0,24516	0,24301	0,08602				
GYPSY30_AG	0,60645	0,62581	0,62366	0,63656	0,64731	0,62796	0,63226	0,64516	0,69032	0,67097	0,67097	0,68172	0,66237	0,66237	0,67742	0,68172	0,25806	0,25806	0,26452	0,24516	0,23226			
GYPSY37_AG	0,65376	0,65591	0,64946	0,67097	0,67527	0,69247	0,67312	0,69032	0,70968	0,69032	0,68602	0,70968	0,67742	0,67742	0,70108	0,70108	0,54839	0,54839	0,55484	0,56989	0,55484	0,55054		
GYPSY38_AG	0,66882	0,65161	0,65161	0,66452	0,66022	0,69247	0,68817	0,69462	0,71398	0,68387	0,67097	0,68817	0,70108	0,70108	0,70968	0,70968	0,5871	0,5871	0,5828	0,56559	0,55484	0,57634	0,28387	
GYPSY71_AG	0,66667	0,68602	0,67527	0,67097	0,68817	0,68387	0,67097	0,67527	0,70323	0,69247	0,71613	0,72258	0,70323	0,70323	0,71183	0,70538	0,55484	0,55484	0,55699	0,55054	0,55054	0,55484	0,42366	0,44731
GYPSY36_AG	0,67957	0,68602	0,68172	0,68817	0,69677	0,71183	0,69892	0,69892	0,70968	0,70108	0,69677	0,69677	0,68387	0,68387	0,69032	0,68172	0,5871	0,5871	0,5914	0,58925	0,57849	0,58495	0,49677	0,50538
GYPSY72_AG	0,69247	0,69032	0,67527	0,69677	0,69032	0,72473	0,71613	0,70108	0,72043	0,70538	0,69892	0,70323	0,70538	0,70538	0,69892	0,69462	0,58495	0,58495	0,59355	0,58495	0,57849	0,57634	0,49462	0,48387
Clu171	0,67527	0,67957	0,67312	0,69892	0,67527	0,71613	0,70323	0,69677	0,70968	0,69677	0,68817	0,69677	0,68817	0,68817	0,70108	0,69462	0,60645	0,60645	0,60215	0,60215	0,59355	0,56989	0,50323	0,50753
GYPSY35_AG	0,67527	0,67957	0,67312	0,69892	0,67527	0,71613	0,70323	0,69677	0,70968	0,69677	0,68817	0,69677	0,68817	0,68817	0,70108	0,69462	0,60645	0,60645	0,60215	0,60215	0,59355	0,56989	0,50323	0,50753
GYPSY65_AG	0,71828	0,67957	0,69462	0,68602	0,69462	0,70753	0,73118	0,73118	0,66667	0,67312	0,67957	0,66882	0,68817	0,68817	0,70108	0,69462	0,68387	0,68387	0,68817	0,68387	0,68172	0,67957	0,70108	0,71183
GYPSY68_AG	0,70108	0,68387	0,67957	0,68602	0,69462	0,71828	0,73333	0,72043	0,69462	0,70968	0,71398	0,71613	0,71828	0,71828	0,72688	0,72688	0,73978	0,73978	0,73118	0,74624	0,73333	0,73548	0,70968	0,72258
GYPSY69_AG	0,69247	0,66882	0,67312	0,69032	0,70323	0,67742	0,72903	0,71398	0,67527	0,67957	0,70538	0,68172	0,74409	0,74409	0,73978	0,72688	0,71613	0,71613	0,72688	0,71828	0,70753	0,71398	0,70323	0,70753
GYPSY26_AG	0,71183	0,69677	0,69677	0,69677	0,70968	0,73763	0,70108	0,72043	0,72473	0,72473	0,70753	0,69892	0,70538	0,70538	0,70538	0,71398	0,73118	0,73118	0,71828	0,71613	0,71398	0,72258	0,70323	0,72903
GYPSY28_AG	0,70753	0,69462	0,69247	0,69032	0,71828	0,73333	0,70753	0,71613	0,71183	0,72688	0,70968	0,69892	0,70968	0,70968	0,69892	0,70968	0,74839	0,74839	0,72043	0,73978	0,73333	0,73548	0,70323	0,73118
Gypsy20+22+23	0,69677	0,69247	0,68817	0,70323	0,71183	0,74409	0,73118	0,71613	0,71183	0,70753	0,69247	0,69032	0,70538	0,70538	0,69032	0,69032	0,73763	0,73763	0,72473	0,72258	0,72043	0,71398	0,70968	0,72903
GYPSY19_AG	0,70538	0,68387	0,69462	0,70753	0,72043	0,74624	0,70968	0,72043	0,70968	0,73978	0,70753	0,68817	0,71183	0,71183	0,70968	0,69892	0,72688	0,72688	0,72688	0,73118	0,73118	0,72043	0,73548	0,72473
GYPSY24_AG	0,70753	0,67742	0,68602	0,69247	0,70538	0,70968	0,70538	0,69677	0,70753	0,73548	0,70323	0,70108	0,70323	0,70323	0,72903	0,72043	0,73333	0,73333	0,74839	0,73763	0,73978	0,71613	0,70968	0,71398
GYPSY66_AG	0,70108	0,69032	0,68817	0,68817	0,70968	0,71183	0,70323	0,69462	0,71828	0,72903	0,69247	0,68817	0,69462	0,69462	0,70968	0,70538	0,72258	0,72258	0,72688	0,72043	0,71828	0,70538	0,71613	0,72473
Clu182	0,69247	0,68817	0,66452	0,67312	0,68172	0,72043	0,70968	0,71828	0,66882	0,70108	0,67957	0,66882	0,69462	0,69462	0,69462	0,68817	0,68602	0,68602	0,68817	0,67957	0,68387	0,70538	0,69677	0,68172
Gypsy18+62+63	0,71183	0,69032	0,67742	0,67312	0,69677	0,71398	0,69677	0,70538	0,68817	0,70323	0,69032	0,69032	0,70538	0,70538	0,68602	0,67527	0,69462	0,69462	0,69892	0,69247	0,69032	0,69892	0,69677	0,70108
GYPSY21_AG	0,70538	0,70108	0,68602	0,68387	0,69892	0,71828	0,70538	0,70538	0,70538	0,69462	0,67312	0,69032	0,68817	0,68817	0,69677	0,68387	0,68387	0,68387	0,68172	0,69462	0,68602	0,68172	0,69032	0,68817
GYPSY64_AG	0,71398	0,69462	0,71183	0,69892	0,70968	0,72688	0,74624	0,73333	0,69462	0,71183	0,69247	0,68817	0,70323	0,70323	0,71183	0,69677	0,70753	0,70753	0,71828	0,71398	0,71183	0,70538	0,72903	0,72688
GYPSY27_AG	0,70108	0,70753	0,70753	0,69677	0,71828	0,71613	0,71183	0,71183	0,70323	0,70323	0,70108	0,70538	0,72903	0,72903	0,67527	0,69032	0,70538	0,70538	0,70968	0,69892	0,69677	0,71828	0,71183	0,70538
GYPSY61_AG	0,67527	0,67527	0,68387	0,67957	0,68817	0,70108	0,70108	0,72043	0,69892	0,71183	0,68172	0,68172	0,69462	0,69462	0,67097	0,67312	0,70323	0,70323	0,69677	0,69677	0,69892	0,69677	0,71613	0,70968
GYPSY55_AG	0,72258	0,71183	0,70108	0,70753	0,72473	0,72043	0,72043	0,72903	0,69462	0,70538	0,69677	0,70108	0,72688	0,72688	0,68172	0,67312	0,70753	0,70753	0,70108	0,71183	0,70538	0,70968	0,72258	0,72258
GYPSY25_AG	0,69032	0,70323	0,69677	0,70323	0,72258	0,71828	0,69892	0,72043	0,70323	0,71183	0,69462	0,70108	0,72258	0,72258	0,68172	0,68387	0,67097	0,67097	0,69247	0,67957	0,68602	0,69677	0,69032	0,70753
GYPSY56_AG	0,71613	0,70108	0,69677	0,70108	0,72473	0,72043	0,70323	0,72043	0,67312	0,71398	0,69247	0,69032	0,72043	0,72043	0,67527	0,67527	0,71183	0,71183	0,70323	0,70753	0,70968	0,70538	0,70538	0,71613
Gypsy57+58	0,69462	0,69677	0,69247	0,69462	0,71613	0,71398	0,70968	0,71613	0,70968	0,69892	0,70323	0,70538	0,73333	0,73333	0,68172	0,67742	0,71183	0,71183	0,70968	0,70323	0,70753	0,70108	0,71183	0,72688
GYPSY60_AG	0,70538	0,69032	0,69247	0,69677	0,71828	0,70753	0,71828	0,71613	0,69892	0,71183	0,70108	0,69462	0,72473	0,72473	0,68387	0,68602	0,69892	0,69892	0,69677	0,68602	0,68817	0,69677	0,70108	0,70968
GYPSY59_AG	0,69677	0,68172	0,67742	0,67527	0,69032	0,71613	0,71828	0,70323	0,67527	0,69892	0,68172	0,69032	0,70538	0,70538	0,66667	0,66667	0,71828	0,71828	0,70108	0,70108	0,70108	0,69892	0,70968	0,70968
GYPSY67_AG	0,69677	0,69247	0,69677	0,68817	0,70323	0,73333	0,73333	0,72258	0,66882	0,70108	0,68172	0,66882	0,70753	0,70753	0,67527	0,68172	0,70753	0,70753	0,69892	0,69247	0,69247	0,68602	0,72473	0,70968

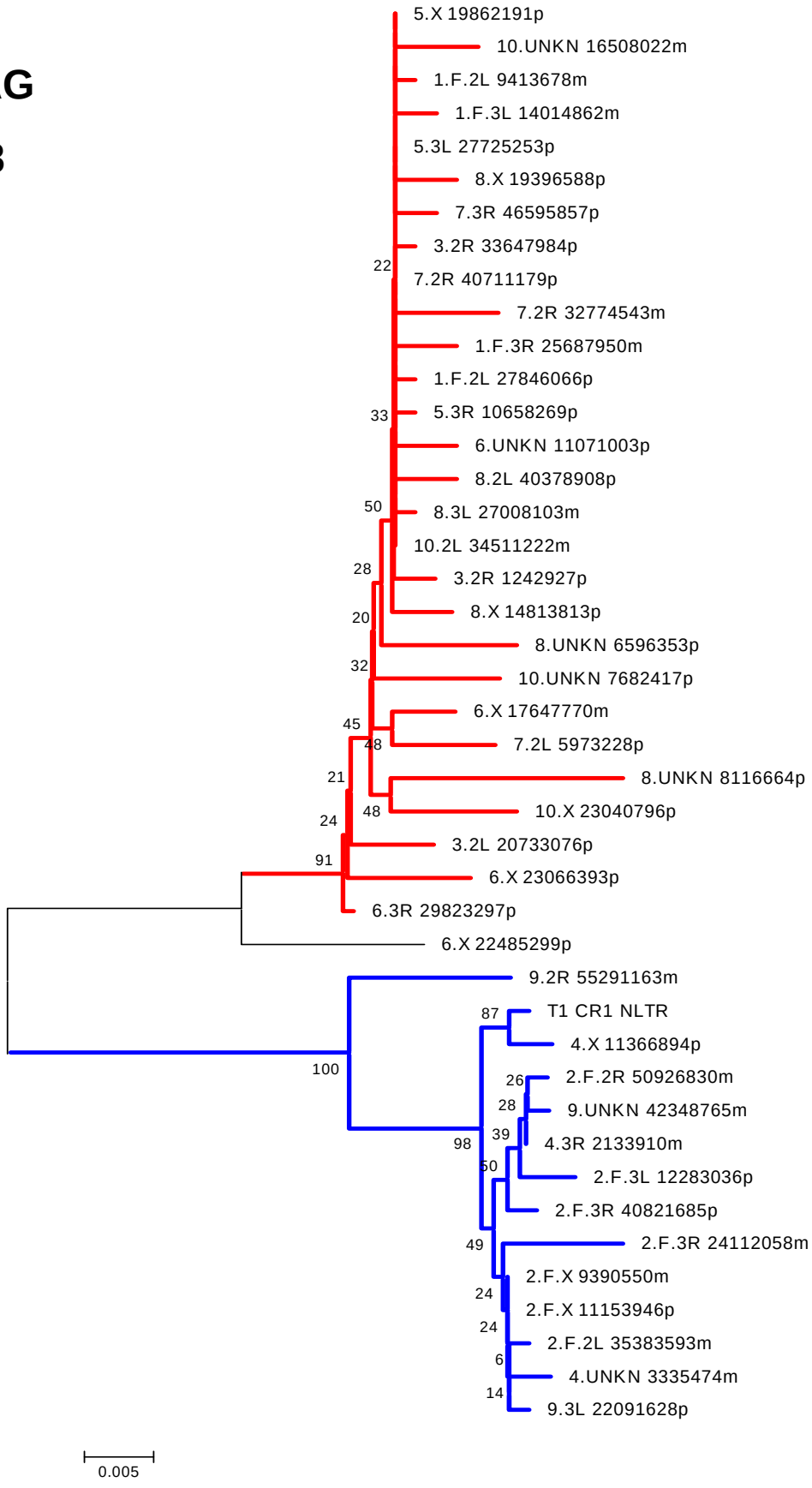
Matriz_dist-p_Elementos GYPSY em AnoTExcel

ANEXO 3

NLTR

T1-CR1_AG

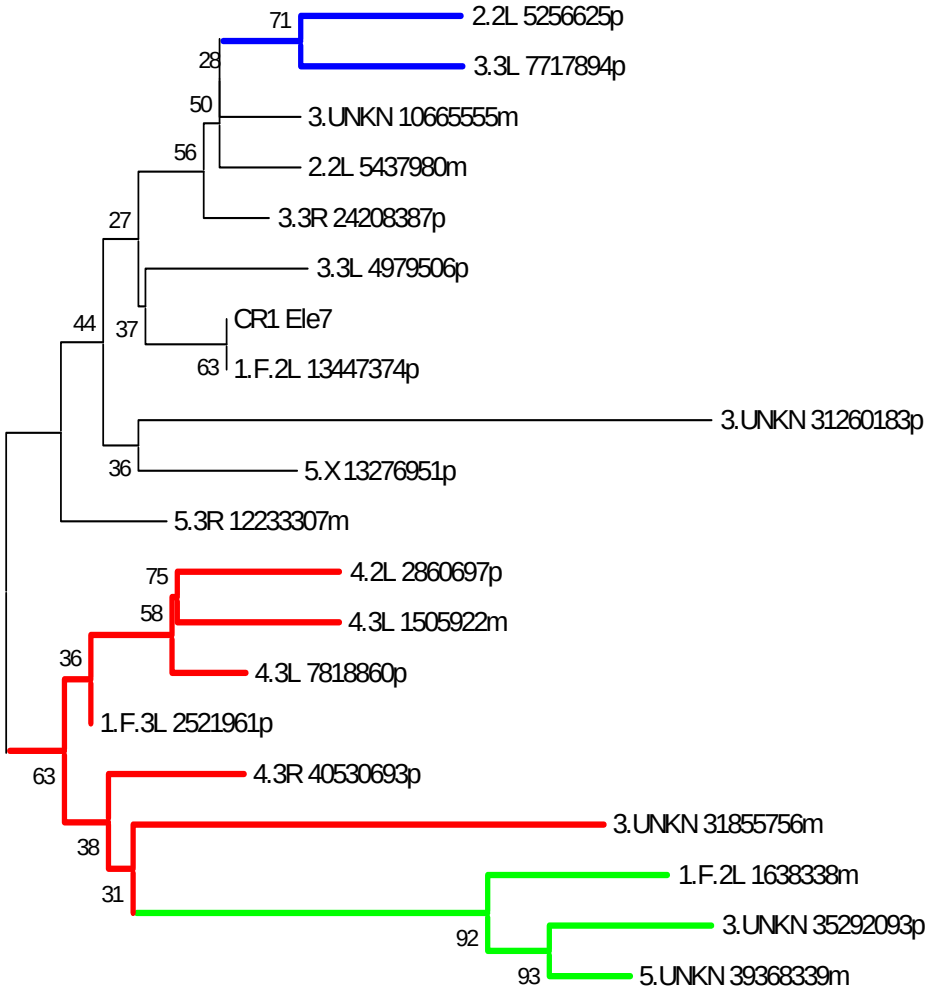
Cluster 18



NLTR

CR1_Ele7

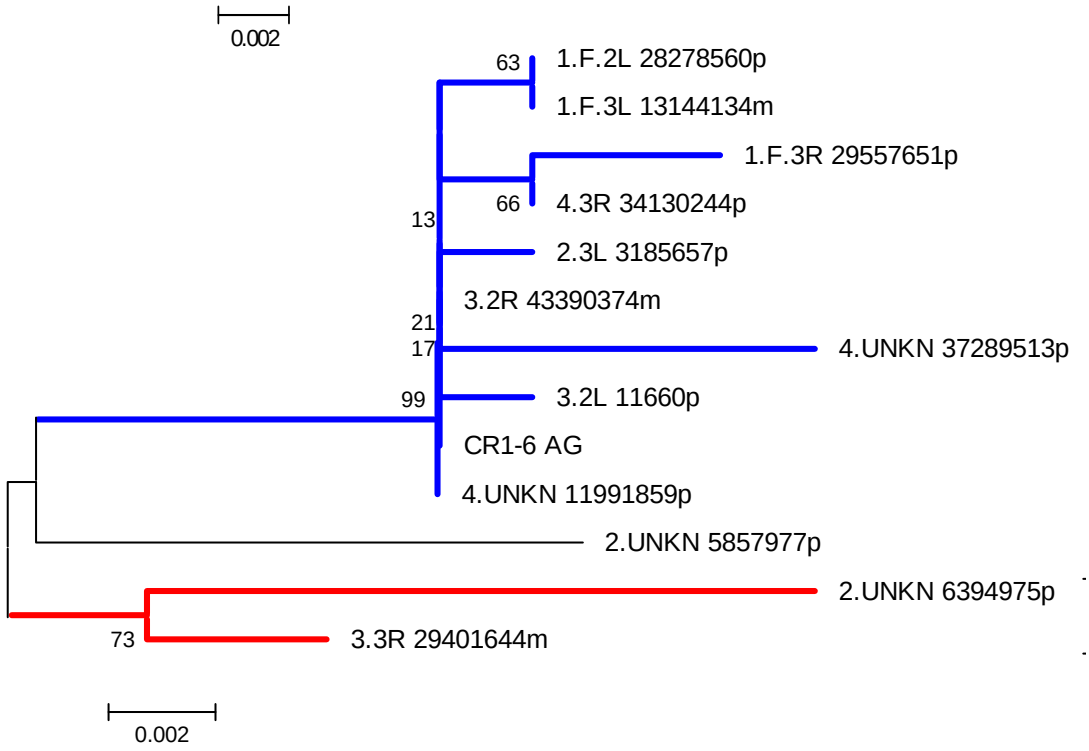
Cluster 36



NLTR

CR1-6_AG

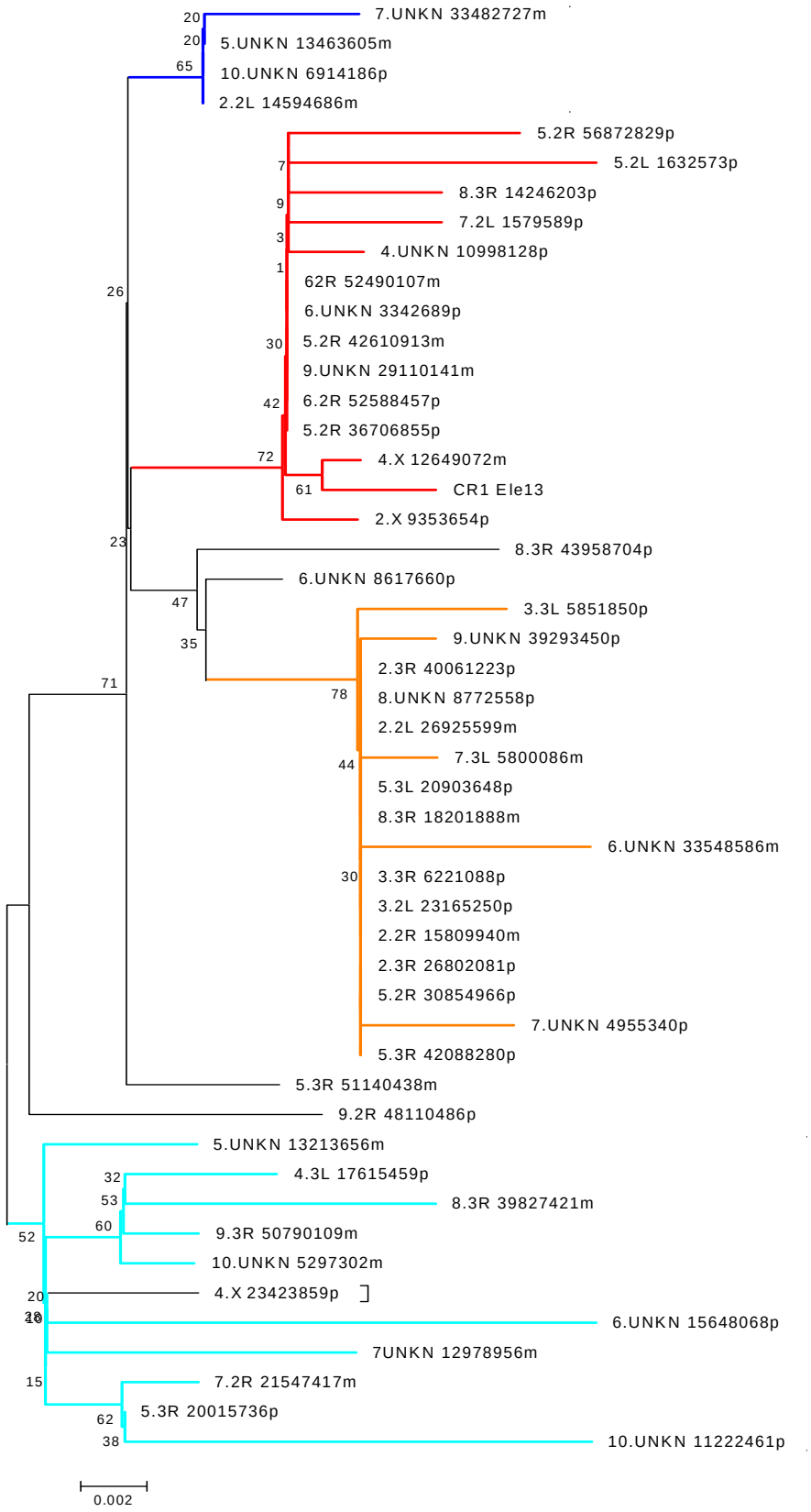
Cluster 72



NLTR

CR1-13_AG

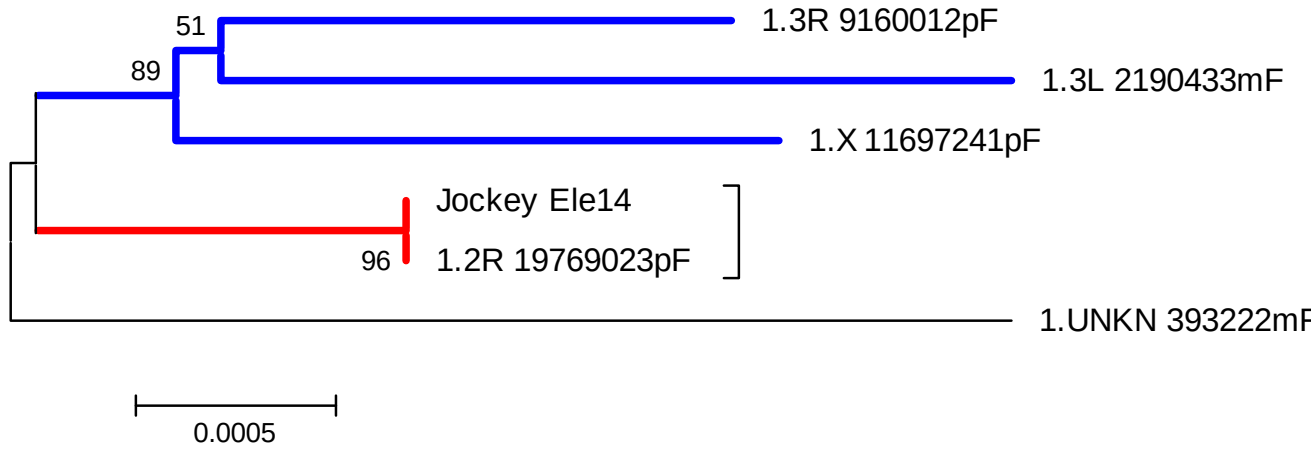
Cluster 16



NLTR

JOCKEY_Ele14

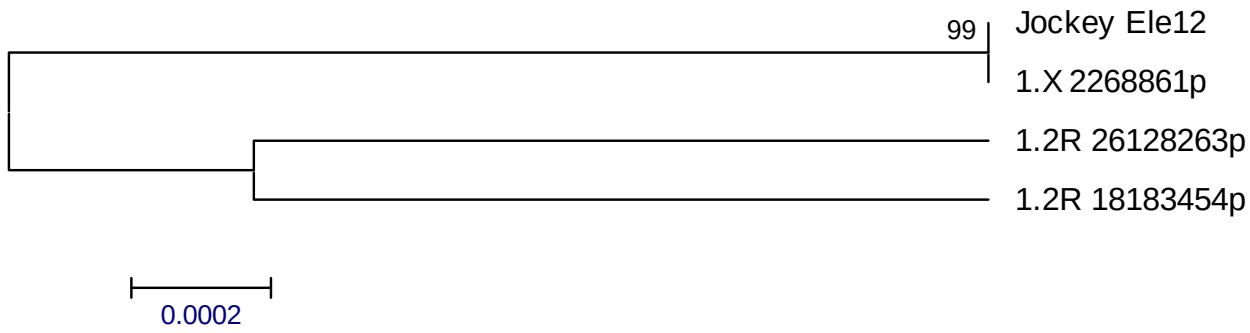
Cluster 120



NLTR

JOCKEY_Ele12

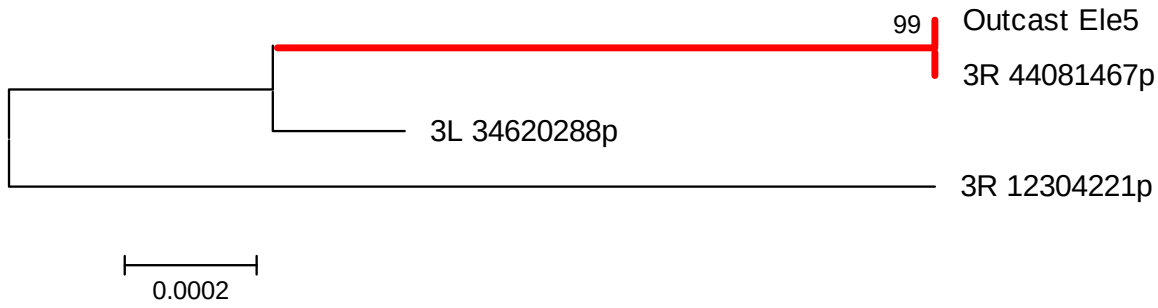
Cluster 183



NLTR

OUTCAST_Ele5

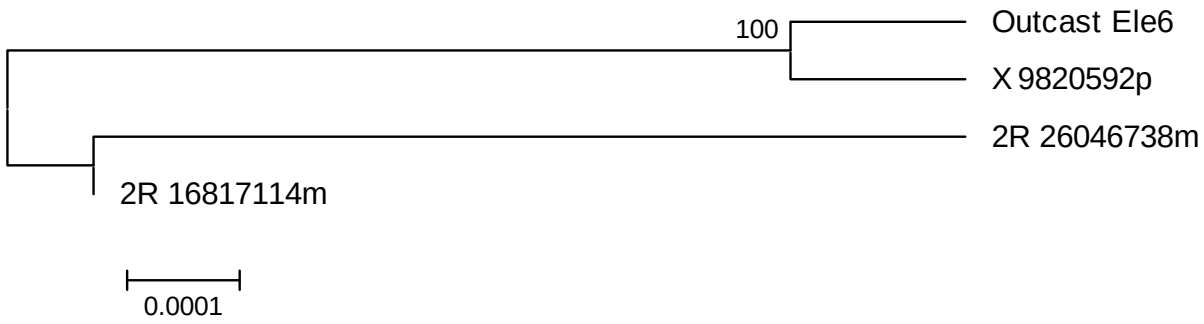
Cluster 238



NLTR

OUTCAST_Ele6

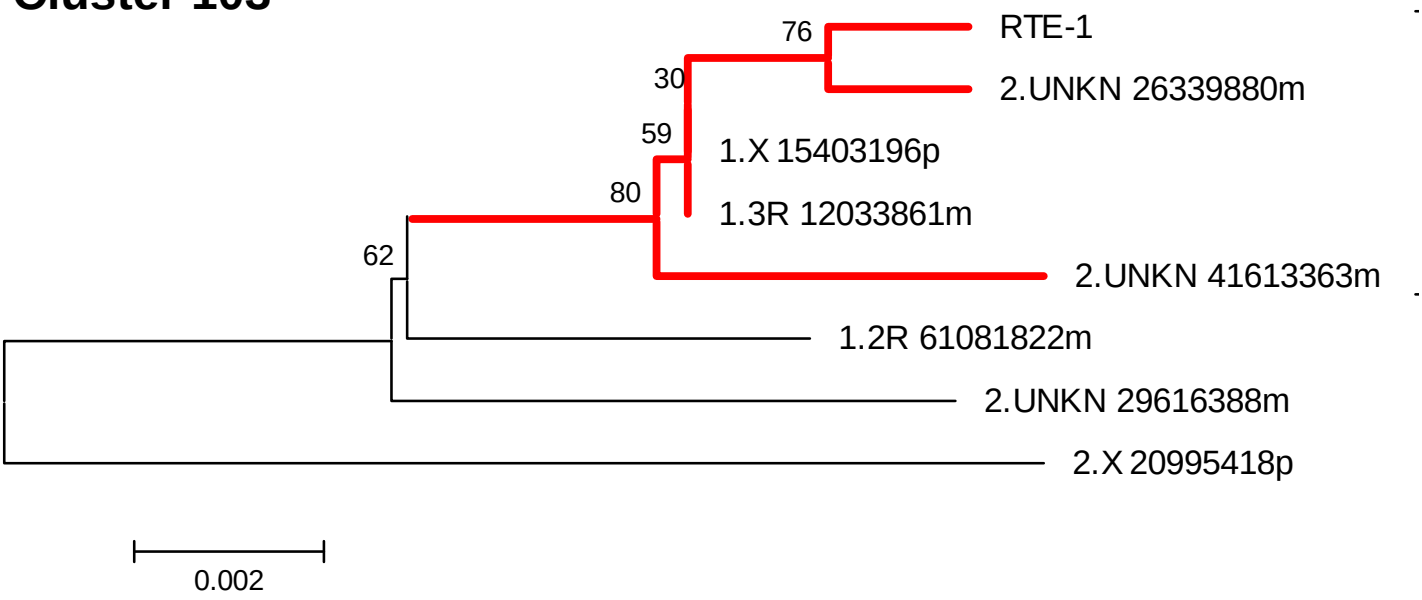
Cluster 188



NLTR

RTE_Ele1

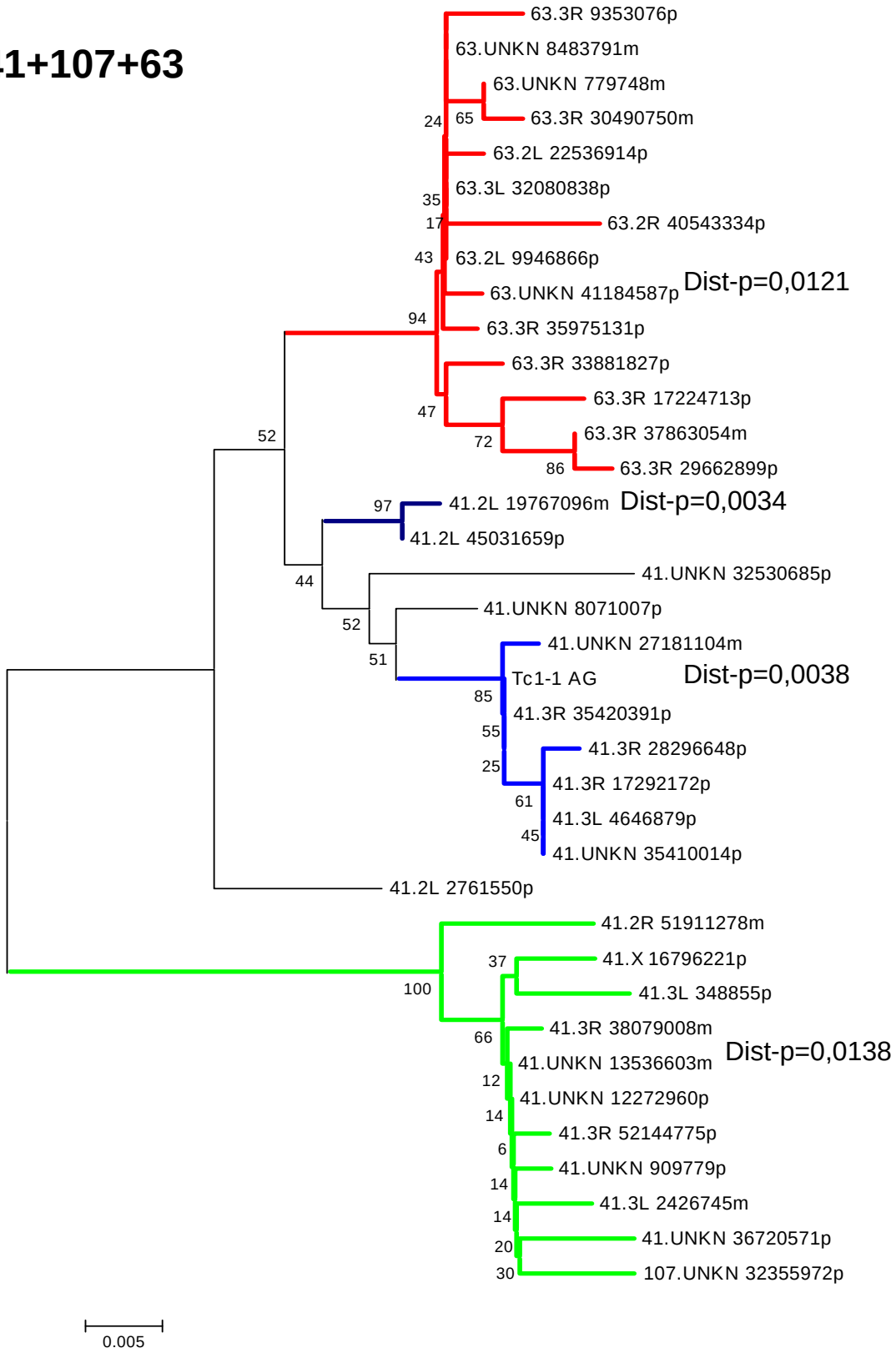
Cluster 103



CLASE II

Tc1_AG

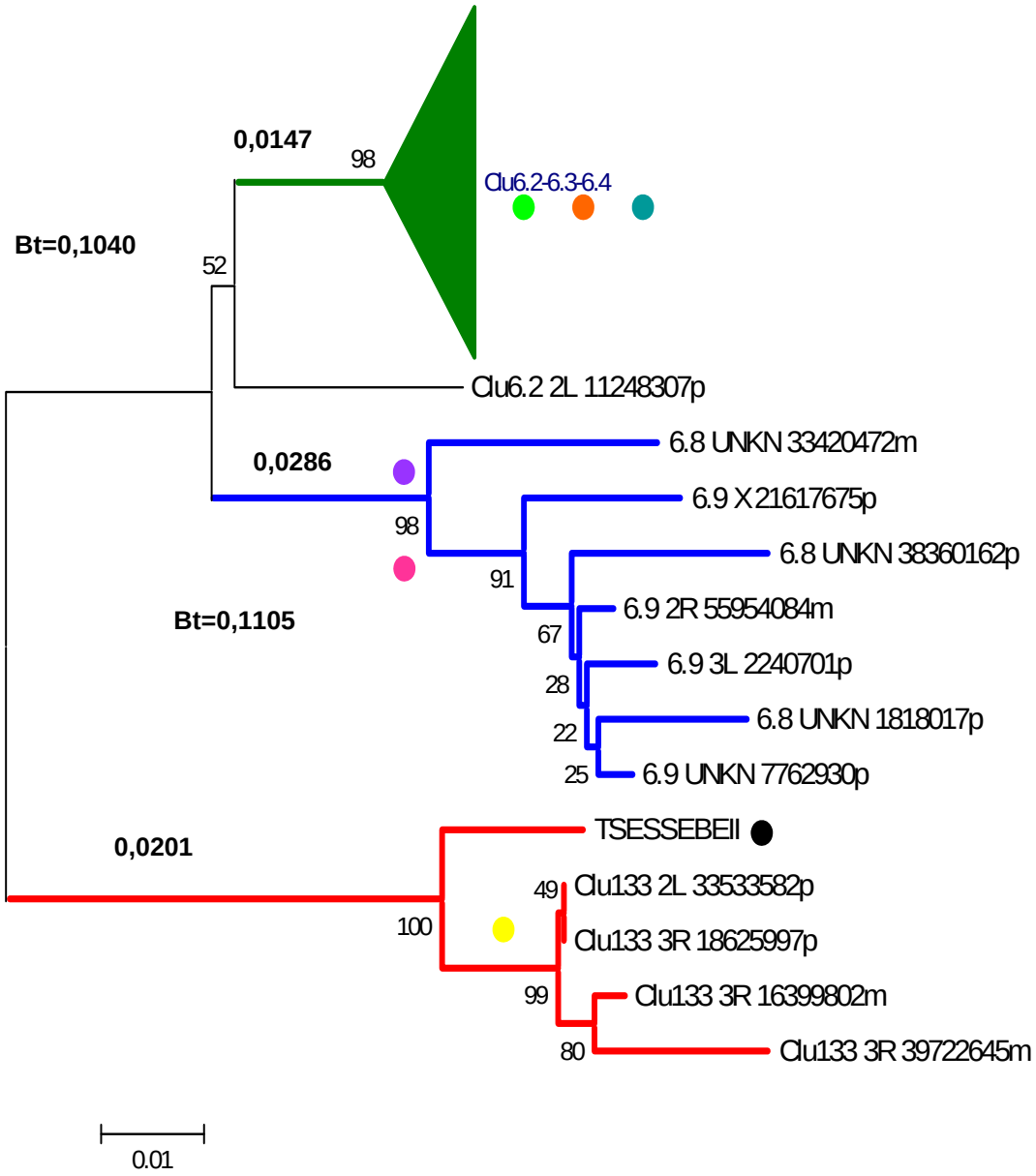
Clusters 41+107+63



CLASE II

TsessebeII

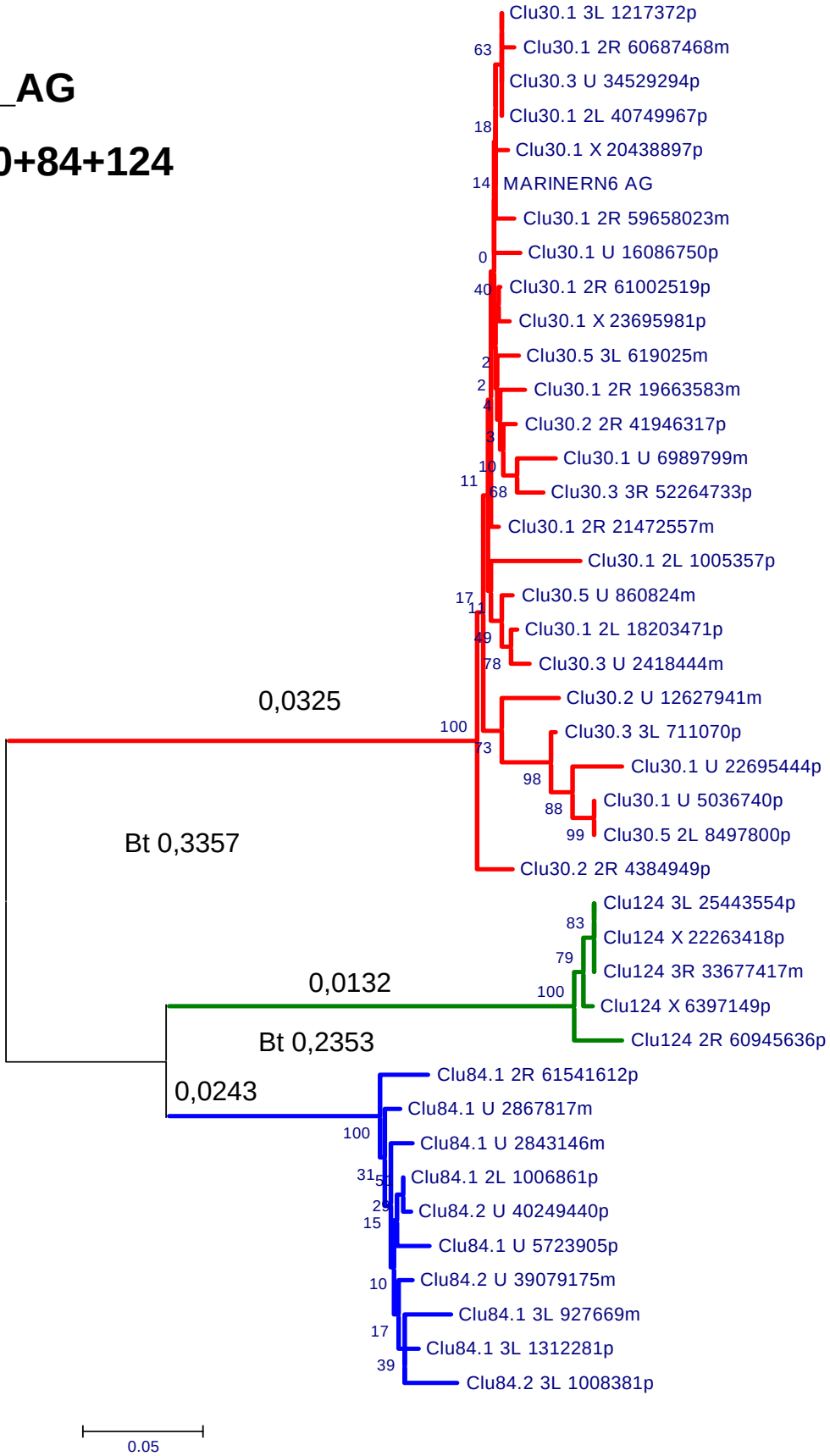
Clusters 6+133+222



CLASE II

MarinerN6_AG

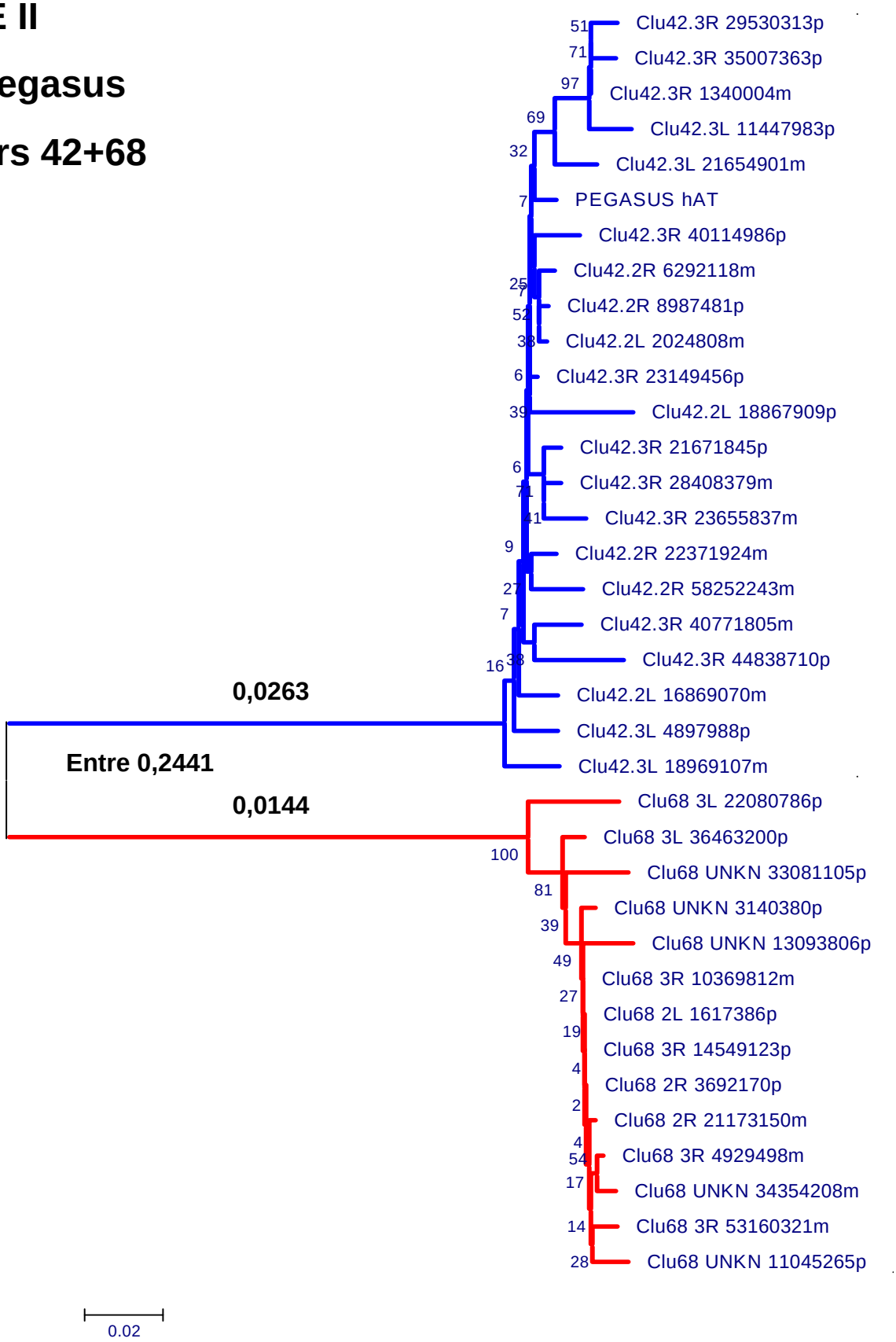
Clusters 30+84+124



CLASE II

hAT_Pegasus

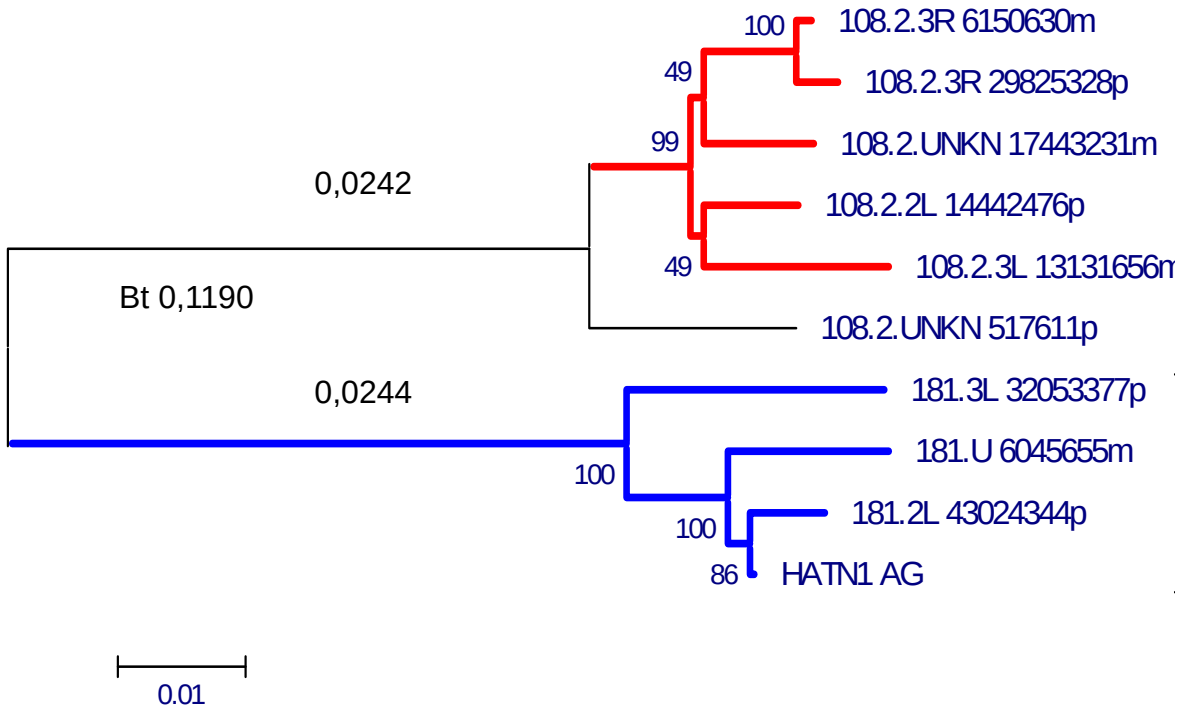
Clusters 42+68



CLASE II

hATN1_AG

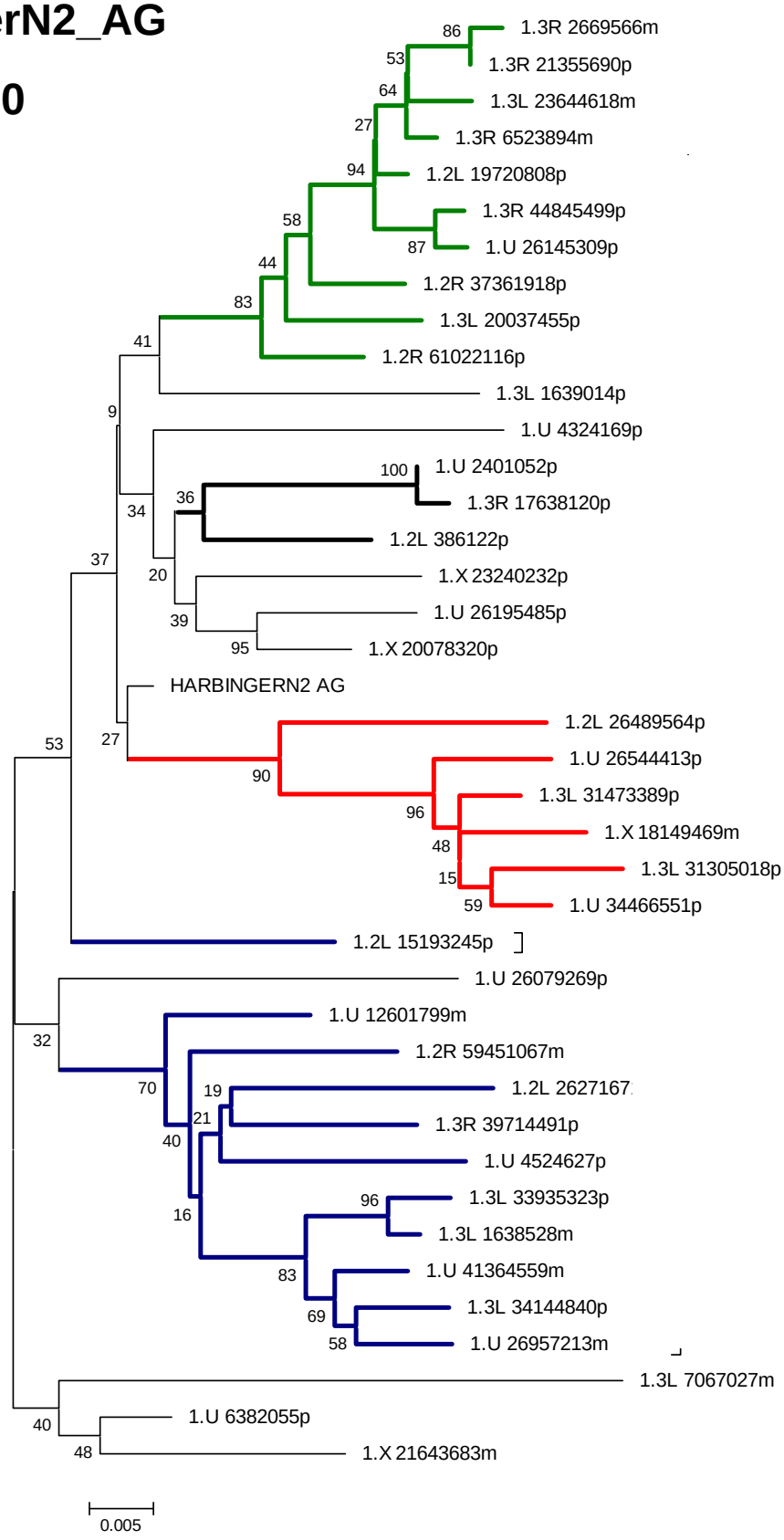
Clusters 108+181



CLASE II

HarbingerN2_AG

Cluster 20



CLASE II

Helitron2N_AG

Clusters 109+129+177