UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

POSGRADO EN CIENCIAS BIOMEDICAS

CENTRO DE CIENCIAS GENÓMICAS

EVOLUCION DE SECUENCIAS DE INSERCION EN POBLACIONES NATIVAS DE *Rhizobium etli*

TESIS

QUE PARA OBTENER EL GRADO ACADEMICO DE

DOCTOR EN CIENCIAS

PRESENTA:

LUIS FERNANDO LOZANO AGUIRRE BELTRAN

DIRECTOR DE TESIS:

DR. VICTOR MANUEL GONZALEZ ZUÑIGA

MEXICO D. F.

2011

A mis padres,

José Antonio y Olga.

A Pepo, Mariana y Fernanda.

A July y Luis.

AGRADECIMIENTOS

A la Universidad Nacional Autónoma de México.

Al Dr. Víctor Manuel González Zuñiga por la dirección del presente trabajo, por su apoyo y amistad.

Al Dr. Mario Soberón Chávez, a la Dra. María del Carmen Gómez Eichelmann, al Dr. Luis Eguiarte Fruns y al Dr. David Rene Romero por sus comentarios y sugerencias.

Al Dr. José Guillermo Dávila Ramos y a la Dra. Valeria Souza por su entusiasmo y ayuda durante todo el doctorado.

A los actuales y antiguos compañeros del laboratorio, Paty, Rosy, Sol, Cesar, José Luis A., José Luis F., Ismael, Humberto, Cinthia, Gama, Orlando, Pablo, Agustín, Gaby, Angeles, America, Deborah, Paco, Ara, Lupita, Guo y Mac.

A José Espíritu que siempre ha estado de nuestro lado.

A los amigos de generación, en especial a Alfredo, Eli y Rodrigo, Pavel, Carlos y Santiago.

A todos los amigos del IBT: Fili, Joel, Flores, Cano, Lety y Mary, Nelly, Alejandro; Rive, Martín, Iván, Héctor, Angel, Javier, Roberto, Sergio, Antonio, Pedro, Eduardo, José, Dago, Oswaldo, Jalil, Arturo, Armando.

A Jalisco, Mariana, Julia, Andrés, Ana, Ricardo, Bernardo, Angeles y Du.

A los compañeros de Maskas, Sideróforos, Fullerenos y Veteranos.

A los amigos de tiempo atrás: Nuri, Emilio, Pancho, Mario, Carlos, Paco y Javier.

Con especial cariño y apoyo a Miguel, Miguelito y Yadira.

INDICE

RESUMEN

Las rizobias son un grupo extenso de bacterias del Orden Rhizobiales, capaces de fijar nitrógeno y de formar nódulos en las raíces de distintas leguminosas. La mayoría de los genes necesarios para esta interacción se encuentran en un plásmido de alto peso molecular, denominado plásmido simbiótico, o bien en regiones cromosomales denominadas islas. El análisis de los genomas de las rizobias ha revelado que los plásmidos y las islas simbióticas tienen gran cantidad de secuencias de inserción. Las Secuencias de Inserción (SIs) son elementos genéticos móviles (<2.5 Kb) con una organización genética sencilla, que tienen la capacidad de insertarse en múltiples sitios del genoma ya que sólo codifican para aquellas funciones relacionadas con su movilidad. El genoma de *Rhizobium etli* CFN42 consta de un cromosoma circular de ~4.4 Mbs y 6 plásmidos de 184 el más pequeño hasta 650 Kbs el más grande. Esta cepa presenta un total de 42 SIs distribuidas en el cromosoma y los plásmidos simbiótico y conjugativo. En este proyecto se analizó la distribución y conservación genética de las SIs halladas en el genoma de *Rhizobium etli* CFN42 en una colección de 87 cepas de *Rhizobium* que corresponden a tres poblaciones de distinto origen geográfico. Primero se obtuvieron perfiles de presencia-ausencia de 39 SIs de *R. etli* CFN42 analizando mediante PCR la conservación de su contexto genómico entre las cepas de las tres poblaciones. En general se obtuvo que las SIs del plásmido simbiótico presentan una mayor conservación de su contexto genómico a diferencia de las SIs del plásmido conjugativo y el cromosoma. Se escogieron dos SIs del plásmido simbiótico para ser secuenciadas en todas las cepas que las presentaron. Así mismo, se secuenciaron el gen nodC del mismo plásmido y dos genes ‘housekeeping’ del cromosoma para todas las cepas. A estos cinco genes se les aplicaron distintas pruebas de genética de poblaciones para conocer su dinámica evolutiva. Los resultados mostraron que las SIs tuvieron una menor diferenciación genética y diversidad nucleotídica, y un menor número de eventos de recombinación en comparación con los genes del cromosoma. Lo anterior siguiere que las poblaciones de *R. etli* divergieron recientemente en México y que el plásmido simbiótico también tiene un origen reciente en las tres poblaciones analizadas.

ABSTRACT

Rhizobia are a group of symbiotic nitrogen- fixing bacteria that belong to the Order Rhizobiales and are able to develop symbiotic structures known as nodules in the roots of several leguminous plants. Most of the genes that participate in this interction are found in the symbiotic plasmid or in chromosomal regions called simbiotyc islands. The analysis of rhizobia genomes has shown that the plasmids and symbiotic islands have several insertion sequence elements. Insertion sequences (IS) are mobile genetic elements (<2.5Kb) with a simple genetic organization that able to move to many different genomic locations because they only codify for the functions related with their transposition. *Rhizobium etli* CFN42 genome consist of one circular chromosome (4.4 Mbs) and six plasmids of distinct molecular sizes (184 to 650 Kbs). The CFN42 strain have 42 IS elements distributed in the chromosome, the symbiotic and the conjugative plasmids. In this work, the distribution and genetic conservation of CFN42 IS elements were studied in a collection of 87 *Rhizobium* strains belonging to three populations with different geographical origin. Presence/absence profiles were obtained for 39 IS elements of CFN42 by means of PCR to analyze their genomic context conservation in the three populations. The IS elements of the symbiotic plasmids have a higher genomic context conservation within the three populations, while the IS of the chromosome and conjugative plasmid were less frequently present. Two IS elements of the symbiotic plasmid were selected and sequenced. Another symbiotic plasmid gene, nodC, and two chromosomal housekeeping genes, glyA and dnaB, were also sequenced. The five genes were used to examine the evolutionary dynamics of the strains of the three populations based on population genetic analysis. The results indicated that the IS elements had a lower genetic diferentiation and nucleotide diversity, and a lower number of recombination events in comparison with the chromosomal housekeeping genes. These suggest that the Rhizobium etli populations diverged recently in Mexico and that the symbiotic plasmid had a recent origin in the three populations.

PRESENTACION Y OBJETIVOS

El presente trabajo se llevó a cabo en el laboratorio de Genómica Evolutiva, del Centro de Ciencias Genómicas del Campus Morelos de la UNAM. Este proyecto surge a partir de la secuenciación del genoma de *Rhizobium etli* CFN42. La tesis describe, a lo largo de cinco capítulos, los aspectos conceptuales y metodológicos, así como los resultados y discusión general del proyecto de doctorado: "Evolución de Secuencias de Inserción en Poblaciones Nativas de *Rhizobium etli*".

El presente trabajo tiene como objetivo general caracterizar la distribución y dinámica evolutiva de las Secuencias de Inserción (SIs) halladas en el genoma de *Rhizobium etli* CFN42 (González, et al. 2006) en tres poblaciones de *Rhizobium* con distinto origen geográfico. El genoma de CFN42 consta de siete replicones, y sus SIs se hallan distribuidas principalmente entre el cromosoma circular y dos plásmidos, el conjugativo y el simbiótico. Las SIs, debido a su capacidad de transposición (Chandler, et al. 2002), han sido consideradas como genes parásitos; aún así, existen ejemplos en la literatura en donde su presencia y actividad pueden favorecer a sus hospederos (Lenski, et al. 2003), razón por la cual se les ha considerado como simbiontes ocasionales. Para poder conocer la dinámica evolutiva de las SIs, se tuvieron como objetivos particulares caracterizar su distribución en 87 cepas de tres poblaciones de *Rhizobium* y determinar la conservación de su contexto genómico y de su secuencia nucleotídica, comparándola con otros genes cromosomales y genes del plásmido simbiótico.

INTRODUCCION GENERAL

Elementos Genéticos Móviles: Secuencias de Inserción

Los genomas procariontes poseen distintos tipos de elementos genéticos móviles (EGMs), como plásmidos conjugativos, islas simbióticas y patogénicas, profagos, transposones, integrones, etc., los cuales participan activamente en la generación de variabilidad genética. Entre estos EGMs se encuentran las Secuencias de Inserción (SIs), las cuales fueron descubiertas inicialmente en experimentos sobre la expresión de genes en *Escherichia coli* y el fago lambda, al hallarse segmentos de ADN en diferentes posiciones y orientaciones en los operones de lactosa y galactosa.

Desde entonces se conoce que la presencia y actividad de SIs en los genomas, provocan diferentes tipos de mutaciones como deleciones, inserciones e inversiones de regiones de ADN, y consecuentemente activan o inactivan la expresion de genes (Chandler y Mahillon, 2002). Generalmente estas mutaciones son poco benéficas para la bacteria, por lo cual las SIs se consideran elementos dañinos para la integridad genómica y supervivencia de la bacteria. Aun así, la secuencia de distintos genomas y varios trabajos experimentales, muestran que estos elementos se encuentran dispersos en un gran número de organismos. Debido a lo anterior, desde hace tiempo se han propuesto diferentes teorías que pretenden dar una explicación a este fenómeno, sin que hoy en día parezca haber una única respuesta.

Características de las SIs

Las Secuencias de Inserción son uno de los EGMs más pequeños y están formadas por uno o varios genes que codifican para una transposasa, la cual por lo común se encuentra bordeada por 2 secuencias inversas repetidas que varían entre 10 y 50 pares de bases (Figura 1). La transposasa contiene la información necesaria para llevar a cabo el evento de transposición de la SI, el cual puede provocar el cambio de localización o la génesis de una nueva copia, dependiendo del tipo de transposición. Las SIs se han clasificado en 25 familias en base a (i) la similitud de secuencia de la transposasa, (ii) el orden y número de genes, (iii) la similitud de las secuencias inversas repetidas y (iv) la formación de secuencias directas repetidas durante el

proceso de transposición (Tabla 1). Las 25 familias se hallan únicamente en bacterias y arqueas, aunque algunas se asemejan a ciertas familias de EGMs presentes en eucariontes, como la familia mariner/Tc presente en *Drosophila*, la familia CACTA en plantas, la familia Ac en plantas e insectos y los helitrones; todos estos elementos presentan semejanzas a nivel de secuencia y en algunos casos con el mecanismo de transposición (Chandler y Mahillon, 2002). Existen aproximadamente 3600 SIs almacenadas en el IS DataBase (Kichenaradja, et al. 2010), la cual es una base de datos especializada en estos elementos. Esta cantidad de SIs es un estimado muy bajo y no representa la distribución que tienen estos elementos entre los procariontes. Analizando el número de SIs que se pueden encontrar en la base de datos no redundante (nr) del NCBI (version nr del 2010), se pueden hallar hasta 15 mil SIs distintas. Así mismo, los genomas secuenciados van incrementado el número de nuevas SIs dentro de cada familia.

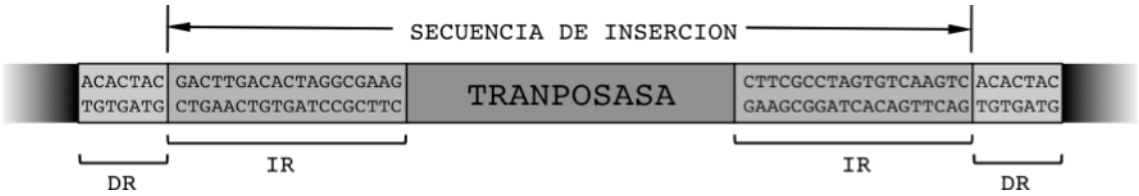

Figura 1. Secuencia de Inserción. Dependiendo la familia puede presentar uno o varios genes que codifican la transposasa; dos secuencias inversas repetidas, IR; y dos secuencias directas repetidas, DR.

El grado de conservación de las SIs, varía dependiendo de cada familia. Mientras que existen familias con pocos representantes conocidos (15 o 20 SIs) y con un grado de identidad en su secuencias proteicas por arriba del 35%, otras pueden llegar a ser tan variables (con identidades por debajo de 15%) que son agrupadas en la misma familia por su semejanza con un solo miembro de la familia o porque tienen inversas repetidas semejantes. Una posible explicación a este fenómeno, se encuentra en el hecho de que las SIs de una misma familia, al hallarse en distintos organismos se encuentran sujetas a diferentes mecanismos que promueven la variación de la secuencia de la transposasa y de la inversa repetida que dependen de la biología, hábitat, tamaño de población, etc. de su hospedero (Chandler y Mahillon, 2002). Actualmente se siguen usando los mismos criterios para clasificar a las familias de SIs empleados por Mahillon y Chandler, posiblemente con la determinación de un mayor numero de SIs se pueda crear un mejor criterio de clasificación.

Tabla 1. Familias de Secuencias de Inserción en Procariontes.

Familias	Subgrupos	Intervalo de Tamaño (bases)	Directas Repetidas	Inversas Repetidas	Genes
IS1	2	740-4600	Si	Si	2
IS1595	9	700-7900	Si	Si	1
IS3	5	1000-1750	Si	Si	2
IS481		950-1300	Si	Si	1
IS4	7	1150-5400	Si	Si	1
IS701		1400-1550	Si	Si	1
ISH3		1225-1500	Si	Si	1
IS1634		1500-2000	Si	Si	1
IS5	6	800-1500	Si	Si	1 - 2
IS1182					
IS6		700-900	Si	Si	1
IS21		1750-2600	Si	Si	2
IS30		1000-1700	Si	Si	1
IS66	2	1350-3000	Si	Si	1 - 3
IS91		1500-2000	No	No	1
IS110	2	1200-1550	No	No	1
IS200/IS605	3	600-2000	No	No	1 - 2
IS607		1700-2500	No	No	2
IS256		1200-1500	Si	Si	1
IS630		1000-1400	Si	Si	1 - 2
IS982		1000	Si	Si	1
IS1380		1550-2000	Si	Si	1
ISAs1		1200-1500	Si	Si	1
ISL3		1300-2300	Si	Si	1
Tn3		>3000	No	Si	>1

Primeros Trabajos: SIs en *E. coli*

Una vez reconocidas las SIs como EGMs, empezaron a ser analizadas en distintos organismos modelo. Estos trabajos se pueden agrupar en tres clases, la primera en los que caracterizan la diversidad y distribución de estos elementos; la segunda, en donde se investigaban los efectos que producen, como los rearreglos y mutaciones; y la tercera, donde se empezó a analizar su variación de secuencia, su presencia en organismos filogenéticamente cercanos y su papel como mediadores de variación genética entre organismos.

La diversidad y distribución de algunas SIs (IS1, IS4, IS5, IS10, IS30, IS150, IS103) fue estudiada principalmente en *E. coli*. Estos trabajos empezaron a analizar el número de copias que presentaban las cepas aisladas y poblaciones naturales de esta bacteria. Posteriormente, investigaron el efecto y las características de las SIs en bacterias entéricas relacionadas con *E. coli*. Inicialmente se demostró que SIs no relacionadas a nivel de secuencia se pueden hallar juntas en distintas cepas de la misma especie y se determinó teóricamente que la presencia de estas SIs en un genoma es resultado de su diseminación por plásmidos altamente transmisibles entre cepas (Hartl 1988). En cepas naturales de *E. coli* de distintos orígenes hay una alta variabilidad en el número y posiciones genómicas de las distintas SIs y la distribución de estos elementos muestra que su origen se puede deber, en algunos casos, a que se encontraban previamente en el ancestro común de estas bacterias, y en otros casos a eventos de transferencia horizontal (TH) (Bisercic 1992). En el mismo trabajo señalan que las SIs entre bacterias entéricas cercanamente relacionadas muestran distintos patrones de evolución (ej. divergencia nucleotídica y su transposición entre genomas) y las diferencias entre dos SIs de distinta familia (IS1 y IS200) se pueden atribuir a las características de cada elemento (ej. especificidad y tasa de transposición) más que a las características de las especies que las acarrean. La variación en la secuencia de DNA entre SIs homólogas en cepas naturales de *E. coli* y otras bacterias entéricas evolutivamente relacionadas tiene una conservación entre un 88 y 98% (Lawrence 1992). Estos mismos autores mencionan que la variación nucleotídica tan baja de las SIs entre cepas de *E. coli*, apoya el modelo de un recambio acelerado en los genomas, junto con una elevada movilidad entre cepas. En otro trabajo, cepas de *E. coli* que han

estado en colecciones de laboratorio durante 30 años fueron analizadas mediante RFLP de SIs, y encontraron que algunas SIs, sobre todo la IS5 y la IS30, han provocado un alta variabilidad genética principalmente mediante eventos de transposición, y aun cuando las condiciones de los cultivos celulares en el laboratorio son distintas que las de las poblaciones naturales, imponen también factores selectivos para estas cepas, dando como resultado que la actividad de SIs puede tener un cierto valor adaptativo (Naas 1994). Otro trabajo analiza la distribución de la IS1 en los genomas de aislados naturales de una colección de laboratorio de *E. coli*, determinando que muchos pares de SIs se encuentran mucho más cercanos de lo que se esperaría al azar y proponen que la explicación es que las SIs presentan un mecanismo conocido como "hopping", en el cual los eventos de transposición se llevan a cabo en posiciones cercanas a la SI original. Por otro lado deducen que la presencia de algunos grupos de IS1 en una región del cromosoma en varias cepas de origen diverso se debe a un origen común mas que a un "hot spot" de transposición (Boyd 1997), lo cual parece mostrar que al menos bajo ciertas condiciones de laboratorio las SIs pueden heredarse de forma vertical y mantenerse entre cepas de origen diverso. Todos los anteriores trabajos realizados *en E. coli* (Hartl 1988; Bisercic 1992; Lawrence 1992; Naas 1994; Boyd 1997), muestran que aún en cepas de una sola especie hay diferencias considerables respecto a la diversidad, distribución y origen de estos elementos.

Diversidad y Distribución de SIs en Genomas de Procariontes

Antes de la secuenciación de genomas completos, se sabía que las SIs estaban dispersas en los genomas de bacterias y arqueas. Inicialmente, la presencia de algunas de las familias de SIs parecía estar restringida a un rango de hospederos filogenéticamente reducido, y en algunos casos se encontró que algunos linajes carecen o parecen inmunes a muchas familias de SIs (Mahillon, Chandler 1999). Tal es el caso de la IS1 que se halla preferentemente en Enterobacterias y la IS66 que se encuentra en bacterias de la rizósfera. Posteriormente, elementos de estas familias de SIs han sido hallados en otro tipo de bacterias, por ejemplo, la familia IS1 ha sido hallada en Cyanobacterias (*Synechocystis sp., Microcystis aeruginosa*), Crenarchaeota (*Sulfolobus solfataricus*), Clamydiae (*Candidatus protochlamydia*), Euryarchaeota (*Methanosarcina acetivorans*), y muchas más, mientras que la familia IS66 ha sido

hallada en Beta proteobacterias (*Burkholderia cenocepacia*) y Gamma-proteobacterias (*Pseudomonas putida*, *Shigella flexneri* y *Escherichia coli*). En cambio, otras SIs tienen un patrón de distribución muy amplio, pudiendo ser halladas en bacterias de ambientes diferentes y filogenéticamente distantes. Lo anterior indica que algunas SIs al parecer son muy adaptables y tienen una elevada capacidad de invadir genomas de todos los clados de procariontes, posiblemente por estar presentes en nichos que presentan una mayor diversidad de clados y se encuentran así mismo presentes en distintos elementos genéticos movibles.

Los genomas de bacterias y arqueas secuenciados muestran diferentes patrones de diversidad y distribución de SIs. Algunos genomas pueden presentar una SI en un elevado número de copias, por ejemplo, 5 cepas de *Shigella* presentan mas de 100 copias de la IS1 y una cepa, Sd197, presenta hasta 273 de la IS1N (Yang, et al. 2005); *Mycobacterium ulcerans* presenta 213 copias de la IS2404 y 91 copias de la IS2606 (Stinear, et al. 2007). Por otro lado, otros genomas tienen SIs de diferentes familias con muy pocas copias. Así mismo, bacterias del mismo género pueden tener distribuciones de SIs muy contrastantes, lo cual podría indicar que la presencia de estos elementos puede llegar a ser cepa específico, por ejemplo, diferentes especies del género *Bordetella* (todas patógenas) pueden tener más 250 SIs o por el contrario no tener ninguna (Parkhill, et al. 2003).

Se ha propuesto que los genomas pequeños deben tener una menor densidad de SIs, porque se esperaría que la transposición resulte más desfavorable cuando en un genoma aumenta la fracción de sitios de inserción deletéreos (por ejemplo, en genes esenciales). Siguier, et al. (2006), demostraron que la densidad de SIs en cromosomas bacterianos es menor al 3% excepto en algunos casos. En el caso de los plásmidos, se observa que cuando son menores a 20 Kb carecen de SIs y cuando son más grandes, el porcentaje de SIs va desde 5 hasta 40% del total de sus genes. Los autores proponen que existe una mínima longitud de DNA para que un plásmido sea viable y cuando son pequeños se vuelven más recalcitrantes a las inserciones, por otro lado dicen que los plásmidos que tienen la capacidad de autotransferirse entre diferentes especies y géneros -y que por ende son más grandes porque presentan todas las funciones para su transferencia- son más propensos a adquirir SIs y otros genes accesorios. Un aspecto que no ha sido analizado, es el número de sitios neutros que

existen en los genomas en los cuales se pueden insertar las SIs. Los cromosomas, aun cuando son más grandes, tienden a presentar una mayor cantidad de genes esenciales, mientras que los plásmidos por lo general presentan genes accesorios; los sitios en los cuales las SIs pueden insertarse (regiones intergénicas, otras SIs, genes duplicados y/o accesorios, etc.) sin afectar la adecuación del organismo, dependerá probablemente de las características de cada replicón, por ejemplo, en *Bacillus subtilis*, se ha determinado que presenta 271 genes esenciales de un total de 4101 genes (Kobayashi, et al. 2003), lo cual implicará que presenta una alta cantidad de genes en los cuales se podrían insertar SIs, aun así hay que considerar que en las condiciones de laboratorio no se analiza la disminución de la adecuación de la bacteria y por ende se desconoce si en el ambiente natural del organismo, la pérdida de un gene no esencial lo pone en desventaja frente a otras bacterias.

Hasta los últimos años, se habían realizado algunos análisis sobre las SIs que hicieran uso de la información disponible en los genomas secuenciados (Filée, et al. 2007; Siguier, et al. 2006; Brügger, et al. 2002). Un análisis realizado recientemente muestra que la ausencia de SIs en los genomas es más común de lo que se pensaba (Touchon y Rocha, 2007). En un total de 262 genomas de bacterias y arqueas, el 24% carecen de SIs completas y un 21% no tiene siquiera remanentes de transposasas. Los genomas que carecen de SIs son filogenéticamente diversos (8 arqueas, 9 chlamydias, 5 cyanobacterias, 5 actinobacterias, 3 firmicutes, 5 mollicutes, 12 α-proteobacterias, 4 ε-proteobacterias, 10 γ-proteobacterias, 3 espiroquetas) y cubren un amplio espectro de estilos de vida: 27% son de vida libre o comensales; 11% son patógenos facultativos; 51% son patógenos obligados y 11% son mutualistas obligados. Una característica en común de estos genomas es que todos son pequeños y sólo 4 alcanzan las 3 Mb. Otro porcentaje alto de genomas tienen menos de 10 SIs, lo cual indica que en la mayoría de los genomas estos elementos se encuentran en cantidades moderadas o ausentes. Así mismo, hay que considerar que los genomas presentan distintas cantidades de genes hipotéticos y entre estos pueden haber SIs que no han sido reconocidas por su baja similitud de secuencia con SIs conocidas actualmente.

Por otro lado, algunos genomas muestran un elevado número de SIs y estos pertenecen a distintos clados. También se observa que algunas familias de SIs son más frecuentes y diversas que otras, aunque aquí existe el problema de que la

clasificación de dichas familias no es del todo adecuada, dado que algunas están definidas de forma menos estricta debido a que carecen de inversas repetidas o no forman directas repetidas cuando transponen (Chandler y Mahillon, 2002).

Así mismo, no existe ninguna tendencia filogenética clara en la distribución del número de SIs en los procariontes (Chandler y Mahillon, 2002). Otra característica es que las cepas que presentan una clase de SI tienden a presentar en su genoma otras diferentes clases de estos elementos. La distribución del número de copias presenta un sesgo que indica que la mayoría de las cepas tienen muy pocas copias de una clase de SI, mientras que pocas cepas presentan un elevado número de estos elementos junto a una alta diversidad de familias (Touchon y Rocha, 2007). Lo anterior podría indicar que las SIs se adquieren continuamente por transferencia horizontal y dado que se pueden presentar en organismos lejanamente relacionados, el hecho de no hallar una SI en un genoma no significa que ese linaje nunca haya presentado estos elementos. Este análisis muestra que mientras más grande sea el genoma, mayor será la densidad de SIs. La abundancia de estos elementos en los genomas también puede depender de la densidad de sitios de transposición neutros.

Los genomas que carecen de SIs presentan a la vez un menor número de genes transferidos horizontalmente. Así mismo los genomas ricos en SIs presentan muchas regiones de origen externo, las cuales por otro lado son más tolerantes a la presencia de SIs. Otra tendencia interesante es que la frecuencia de eventos de TH es un determinante de la presencia de SIs pero no de su abundancia, dado que en los genomas es más común encontrar SIs fuera de las regiones generadas por eventos de TH. La abundancia puede depender más de efectos selectivos sobre estos elementos en los genomas hospederos. Por ejemplo, se ha encontrado que en Mycoplasmas existen niveles más altos de TH de lo que se había considerado, aun así prácticamente carecen de SIs (Sirand-Pugnet, et al. 2007).

Dinámica y Evolución de SIs

Se ha considerado que las SIs persisten en las poblaciones naturales por su capacidad de invadir genomas sin importar el efecto positivo o negativo que tengan en la adecuación de sus hospederos. Por el contrario, si existen efectos deletéreos debido

a los eventos de transposición, sólo aquellas SIs que logren invadir diferentes genomas mediante su presencia en EGMs (plásmidos, profagos, etc.) persistirán en las poblaciones. Esta idea se ve corroborada por el elevado grado de TH de las SIs, sus árboles filogenéticos incongruentes y por la diversidad de SIs que se encuentra entre especies y cepas relacionadas.

Hoy en día, se han propuesto diferentes hipótesis para explicar la variabilidad de SIs en los genomas de bacterias y arqueas. Algunos autores sugieren que las SIs sólo persisten mediante invasiones periódicas a genomas nuevos (Wagner, et al 2006), de forma que se compensan las pérdidas de SIs en genomas donde han provocado efectos deletéreos en la adecuación. Si lo anterior es cierto, se esperaría que el número de SIs estuviera correlacionado positivamente con la tasa de TH. Por otro lado, se ha propuesto que existe una relación entre la patogenicidad y la abundancia de SIs, en especial entre patógenos facultativos o emergentes. Lo anterior se debe a que en poblaciones pequeñas debido a cambios frecuentes de nicho ecológico o en el caso de patógenos que han presentado a cuellos de botella en su historia evolutiva, existe un relajamiento de la selección hacia la presencia de SIs (Parkhill et al. 2003; Moran, Plague 2004). Otra propuesta menciona que las SIs pueden resultar ventajosas para los genomas, por lo cual estos elementos favorecen que algunos organismos puedan invadir un nuevo nicho (Schneider, et al. 2004). Por último, se ha sugerido que los cambios asociados a la actividad humana desde hace aproximadamente 12000 años, provocó que los procariontes asociados a los humanos fueran invadidos por SIs mientras se adaptaban a una creciente población de hospederos humanos (Mira et al. 2006, ver más adelante).

Parásitos o Simbiontes Ocasionales

Inicialmente, al ir caracterizando los efectos que provocaban las SIs en los genomas de distintos organismos, se les consideró como parásitos genómicos, los cuales lograban mantenerse en los genomas debido a una elevada tasa de transposición y mediante procesos de transferencia horizontal. Posteriormente, se fueron encontrando casos en los que se les podía atribuir a las SIs una ventaja selectiva, por esta razón algunos autores propusieron que las SIs tenían

ocasionalmente un papel adaptativo y que debido a esta acción, se mantenían en los genomas.

En un trabajo de Schneider, et al. (2004) se formulan dos hipótesis acerca de la persistencia de las SIs en los genomas. La primera menciona que son parásitos genómicos dañinos para sus hospederos, debido a que ocasionan una tasa elevada de mutaciones deletéreas. Bajo esta perspectiva, las SIs se mantienen debido a la transposición replicativa. La segunda hipótesis plantea que estos elementos pueden ocasionar mutaciones benéficas ocasionalmente, por lo cual se pueden considerar agentes importantes para la evolución adaptativa de sus hospederos y por ende, son mantenidas por la selección sobre estas mutaciones benéficas. Ambas hipótesis reconocen que existe un costo por las mutaciones que pueden ocasionar las SIs. La diferencia radica en si la TH junto con la tasa de transposición o la selección de ciertas mutaciones es la fuerza responsable que mantiene a estos elementos en los genomas. Cabe mencionar que ambas hipótesis no se contraponen, pudiendo llevarse a cabo los dos mecanismos en distintos momentos en un genoma.

En el mismo trabajo se analizan una serie de estudios en los cuales las SIs contribuyen substancialmente a la generación de diversidad genética y demuestran algunos casos donde las mutaciones que generan estos elementos contribuyen a la adaptación de las poblaciones bacterianas en sus respectivos ambientes (Lenski, et al. 2003; Papadopoulus, et al. 1999). Por ejemplo, en poblaciones bacterianas analizadas en el laboratorio se halló que las SIs participan a la adaptación bacteriana en condiciones tanto estresantes como no estresantes. Por otro lado mencionan que algunas de las mutaciones benéficas que estos elementos originan no podrían producirse por otro medio, como mutaciones puntuales, ya que la presencia de SIs promueve rearreglos mediante recombinación homóloga.

En otro trabajo, Zhong, S. et al. (2004), investigan el papel de las SIs en la especialización hacia la utilización de los recursos disponibles mediante la evolución experimental de 50 cepas de *E. coli*. En ambientes constantes, como en los que se encuentran las cepas de laboratorio de este trabajo, la selección se centra intensamente en pocos genes, por esta razón la evolución experimental resulta más fácilmente reproducible, lo cual les permite analizar la adaptación al uso de ciertos recursos

disponibles mediante la actividad de SIs. Se ha encontrado que en aislados naturales de *E. coli* hay un elevado recambio de SIs tanto en número como en posición, aún entre cepas cercanas. Lo anterior sugiere que las SIs son elementos importantes en la generación de variación genética en la cual la selección puede actuar. En este trabajo, analizan los rearreglos que se generan por la presencia y/o transposición de SIs; dado que conocen con anterioridad la posición de las SIs en el genoma ancestral (del cual inició el experimento), pueden determinar qué SIs se han movido a otra localización en los genomas que analizan y saber así cuáles participan en los diferentes rearreglos que favorecen la adaptación a algún recurso. Encuentran que se generan 22 rearreglos (duplicaciones y deleciones), de los cuales sólo un rearreglo surge de dos SIs ya presentes en esa posición en el genoma antes del experimento, mientras que 17 rearreglos surgen de una SI ya presente y una SI que acaba de moverse a otra posición y finalmente, 4 rearreglos surgen a partir de dos SIs que se han movido a una nueva posición. Debido a lo anterior, sugieren que la localización genómica de las SIs tiene una mayor influencia en la evolución genómica, cuando estos elementos son capaces de transponerse constantemente.

Aun cuando estos trabajos muestran que en algunos casos las SIs pueden tener un valor adaptativo, no cabe duda de que su actividad tiene un claro potencial dañino para la integridad genómica de las bacterias. El carácter parasitario o adaptativo de estos elementos puede depender así mismo de las propiedades intrínsecas de cada SI junto con las características genómicas de su hospedero.

Extinción y Expansión de SIs en Genomas Procariontes

La frecuencia de SIs en una población depende de la relación entre su tasa de transposición y la acción negativa de la selección natural sobre estos elementos. Algunos autores han propuesto que la presencia y permanencia de SIs en poblaciones naturales se explica más fácilmente mediante procesos estocásticos, como la deriva génica, y no por efecto de la selección sobre estos elementos, sobre todo cuando dichos elementos se encuentran en poblaciones pequeñas y/o aisladas (Escobar-Páramo, et al. 2005). Lo anterior se debe a que la selección natural por sí sola, limita la dispersión de las SIs, debido a los efectos deletéreos que provoca su transposición.

En un trabajo reciente, Wagner, et al. (2005), propone que las SIs están sujetas a ciclos de extinción/reinfección en los genomas. Analiza la diversidad nucleotídica de 5 familias de SIs (IS4, IS5, IS6, IS30 y IS605/IS200) que codifican su transposasa con un solo gen, y que son las SIs más prominentes en los genomas analizados. De un total de 200 genomas secuenciados, 18 presentan entre 3 y 20 copias de una misma SIs, las cuales presentan una divergencia de DNA muy baja. Wagner encuentra que el 68% de los genes de las transposasas son idénticos dentro de un genoma. Mientras que las SIs de la misma familia pueden tener una elevada divergencia nucleotídica al encontrarlas en genomas de bacterias lejanamente relacionadas, cuando se hallan dentro de un mismo genoma tienen una diversidad muy baja. Con estos resultados propone que las SIs de un genoma son evolutivamente jóvenes y han sido adquiridas recientemente. Así mismo, propone que las SIs deben de presentar extinciones periódicas en los linajes bacterianos, lo cual genera una distribución "parchada" y sesgada de SIs entre los genomas. Otra posibilidad que no se consideró en este trabajo es que algunas cepas bacterianas se extinguen con sus SIs y se genera de esta forma la misma distribución "parchada". Cabe mencionar que cuando se hallan SIs de una misma familia en bacterias cercana o lejanamente relacionadas y que presentan una variación nucleotídica elevada, la mayoría de los autores favorecen la idea de que dichos elementos han llegado a cada organismo por eventos de TH independientes y no porque hayan permanecido en los dos genomas desde el ancestro común y que desde entonces hayan ido acumulando la variación nucleotídica observada.

Se ha observado que los genes que tienen una expresión elevada evolucionan más lentamente (Drummond et al. 2005). Debido a que las transposasas están bajo una estricta regulación (Mahillon, et al. 1999), los niveles de expresión que presentan son muy bajos, razón por la cual deberían evolucionar más rápidamente que otros genes y ser más diversas. Justamente se observa lo contrario al compararlos con otros genes duplicados en los genomas. Por otro lado, se ha calculado que la tasa de transposición de algunas SIs (IS10) es de 10^{-3} por célula por generación, mientras que la tasa de escisión es $<10^{-9}$ (Kleckner 1990). Wagner, et al. (2005), propone que las SIs han entrado al genoma muy recientemente, razón por la cual no se encuentran más SIs en los genomas bacterianos, aún cuando su tasa de escisión es mucho menor que su tasa de transposición. Lo anterior implica que las SIs presentan extinciones periódicas en las poblaciones bacterianas y son reintroducidas por transferencia

horizontal. Si las SIs no sufrieran estas extinciones periódicas tendrían una mayor divergencia intragenómica. Por otro lado, si no fueran reintroducidas mediante TH, los genomas bacterianos carecerían de SIs. A este respecto cabe mencionar que algunos genomas (ej. *Rhizobium etli* CFN42) muestran evidencia de haber tenido un alto número de SIs de una misma familia (IS256) y actualmente todas las copias se encuentran incompletas, lo cual podría indicar que en algún momento esa familia de SI se expandió y posteriormente se ha ido perdiendo. Así mismo, algunos de los primeros trabajos sobre SIs, muestran que tienen una permanencia corta en las poblaciones bacterianas (Sawyer, Hartl, 1986).

Otro aspecto que ha sido recientemente estudiado es el de la expansión de SIs en genomas procariontes. Al analizar los genomas completos disponibles, se ha observado que las especies que se han especializado a un nicho han sufrido un proceso de reducción genómica mediante la eliminación de genes. Mira, A. et al. (2006) comentan que las SIs, entre otros elementos móviles, aumentan en número cuando una bacteria se especializa en un hospedero y participan en el proceso de reducción genómica. Lo anterior se debe a que la eficiencia de la selección en bacterias especializadas se reduce debido a que el tamaño efectivo de población disminuye y entonces actúa la deriva génica. Mencionan que las SIs pueden inactivar genes innecesarios o redundantes al aumentar su número de copias y proponen que el encontrar un elevado número de copias de SIs en bacterias patógenas puede indicar una reciente adaptación al hospedero.

Así mismo, Mira et al. 2006 analizan los genes parálogos, incluyendo las SIs, en 255 genomas procariontes. Encuentran que hay un grupo de genes parálogos con una similitud entre 90 y 100% y que corresponden a las SIs. Existen 89 bacterias en las cuales el 75% de estos parálogos recientes corresponden a SIs y son bacterias asociadas a humanos o a la actividad humana, es decir, asociadas a animales o plantas domesticadas, o son bacterias (simbiontes o patógenos) asociadas a hospederos invertebrados. De lo anterior deducen que la especialización a un nicho se relaciona con la expansión de SIs, lo cual inactiva genes y genera un elevado número de rearreglos genómicos.

En contraposición a estas ideas, se ha observado que no existe una diferencia en el número de SIs entre organismos patógenos y no patógenos, tampoco entre patógenos facultativos y organismos de vida libre, aunque sí hay una menor abundancia en aquellos organismos que son simbiontes obligados, lo cual indica que el tipo de asociación puede influenciar el que un organismo presente o no SIs (Touchon y Rocha, 2007). La razón de esto puede ser que los organismos que viven en hábitats muy variables u oscilan entre diferentes hábitats, pueden verse favorecidos por la presencia de SIs que genera una mayor plasticidad genómica, mientras que los simbiontes obligados no están sujetos por lo común a eventos de TH. Hay que tomar en cuenta que en dicho trabajo no se consideraron otras posibles explicaciones a la variación de la abundancia de SIs en los genomas como son: un sesgo en la tendencia del organismo de producir deleciones, la frecuencia de transposición de cada SI y/o el tamaño efectivo de la población.

Papel de las SIs en la Reducción de Tamaño de Genomas Procariontes

Los genomas más pequeños entre los procariontes pertenecen a bacterias patógenas obligadas o a simbiontes intracelulares de eucariontes (Rickettsiae, Spirochetes, Chlamydiae y algunas gamma-protobacterias asociadas a insectos). Al parecer, los genomas de estos organismos presentan un sesgo mutacional que favorece las deleciones sobre las inserciones. Existen varias propuestas que relacionan la presencia de un elevado número de SIs con este proceso de pérdida de ADN. En algunos organismos filogenéticamente cercanos y con intervalos de tamaño de genoma variables, se ha observado que aquellos genomas con un tamaño intermedio, presentan un elevado número de SIs, mientras que los genomas más pequeños carecen casi por completo de ellas. Una posible explicación a este fenómeno es que las SIs pudieron favorecer los procesos de escisión de ADN y debido a que estos organismos presentan tamaños de población pequeños, estas pérdidas de ADN se fijan por efecto de la deriva génica. Aunado a esto, varios autores apoyan la propuesta de que las SIs presentes justamente en poblaciones pequeñas que se especializan en un nicho pueden acelerar la reducción del genoma, al inactivar genes que son innecesarios o redundantes en dicho hábitat (Parkhill et al. 2003; Moran and Plague 2004).

Papel de la Selección y la Deriva Génica en la Persistencia de SIs

Trabajos recientes prueban que en determinadas circunstancias las SIs muestran evidencia de selección positiva. Por ejemplo, un análisis de 2 SIs en distintas poblaciones de *Helicobacter pylori* muestra que las SIs parecen ser elementos ancestrales del pool génico de estas bacterias ya que han evolucionado aproximadamente a la misma tasa que los genes cromosomales. Por otro lado observan que para una SI hay evidencias de selección positiva lo cual puede significar que se esta seleccionando una actividad de transposición óptima (Kalia et al. 2004). Esta optimización explicaría que no haya tantas copias de la misma SI, evitando así afectar a la bacteria y favoreciendo la permanencia de la SI.

En otro trabajo, analizan la posibilidad de que la deriva génica puede tener un rol importante en la diversificación de *Pyrococcus*, un arquea hipertermofílica presente en distintos sistemas hidrotermales (Escobar-Páramo, et al. 2006). Analizan lo anterior a partir de la presencia de SIs en distintas poblaciones, hallando estos elementos solamente en una población geográficamente aislada. El análisis de 6 SIs de dicha población mostró que las SIs han provocado alteraciones genómicas deletéreas, como la inactivación de genes. Por otro lado, los autores observan que estos elementos se hallan en una frecuencia muy alta (~20 SIs por cepa), razón por la cual proponen que la razón de la permanencia de SIs es la deriva génica.

También mencionan que en poblaciones grandes el número de sitios donde puede llevarse a cabo una transposición es muy grande, por lo que resulta poco probable encontrar SIs en la misma posición. Es decir, la probabilidad de hallar dos individuos con la misma IS en la misma localización es mucho mayor en una población pequeña que en una grande. Luego entonces sugieren que los procesos estocásticos y no la selección natural, son los que dan una mejor explicación a la presencia de SIs en altas frecuencias en poblaciones naturales. Al analizar las SIs de esta población encuentran casos donde algunos de estos elementos se han mantenido en la misma localización genómica y otros casos en los que las SIs se han movido y generado distintos rearreglos. Dada la alta frecuencia de SIs en esta población y sus efectos deletéreos observados, proponen que estas SIs se hallaban en la población ancestral que colonizó este hábitat y permanecen por la acción estocástica de la deriva.

Genoma de *Rhizobium etli*

Rhizobium etli es una bacteria Gram-negativa presente en el suelo capaz de interaccionar con las raíces de *Phaseolus vulgaris*, formando nódulos en los cuales puede fijar nitrógeno. El genoma de esta bacteria esta estructurado en varios replicones: un cromosoma circular y varios replicones que pueden representar hasta un tercio de su genoma (García-de los Santos, et al. 1996). Los genes necesarios para la fijación de nitrógeno y la nodulación se hallan en el plásmido simbiótico. Se ha propuesto que el plásmido simbiótico y conjugativo fueron adquiridos en algún momento posterior a la divergencia de *R. etli*. El plásmido conjugativo es capaz de transferirse entre distintas cepas a una alta frecuencia, mientras que el plásmido simbiótico tiene una baja frecuencia de transferencia, debido a que requiere de cointegrarse con el plásmido conjugativo (Brom, et al 2000). *Rhizobium etli* CFN42, fue la primera cepa totalmente secuenciada de esta especie, su genoma consta de un cromosoma y seis plásmidos entre los que se encuentran el plásmido conjugativo y el simbiótico (González, et al. 2006). Esta cepa presenta un total de 39 SIs distribuidas principalmente en estos dos plásmidos y en el cromosoma. Estas SIs pertenecen a un total de once distintas familias, de las cuales la IS66 (12 copias), la IS630 e IS5 (5 copias cada una) y la IS21 (4 copias) son las más abundantes. Ninguna de estas SIs se encuentra interrumpiendo algún otro gen. Por otro lado, existen un total de 42 pseudoSIs presentes entre los mismos tres replicones principalmente.

METODOLOGIA

Colección de Cepas de Tres Poblaciones

Se emplearon un total de 87 cepas de *Rhizobium* pertenecientes a tres poblaciones previamente caracterizadas. En la Tabla 2 se muestran el numero de cepas por población así como las especies de *Rhizobium* que las conforman.

A) Población Puebla, México

La población denominada 'Puebla' corresponde a cepas de *Rhizobium* aisladas directamente de nódulos de *Phaseolus vulgaris* presentes en una milpa (Silva, et al. 1999). La milpa se hallaba cerca del pueblo de San Miguel Acuexcomac y el muestreo se realizó durante tres años consecutivos (1994, 1995 y 1996). Se seleccionaron un total de 30 cepas pertenecientes a la misma milpa durante los tres años de muestreo, 21 fueron caracterizadas previamente como *Rhizobium etli* y 9 como *Rhizobium gallicum* (Silva, et al. 2003).

B) Población Guanajuato, México

La población denominada 'Guanajuato' consta de 30 cepas de *R. etli* que fueron colectadas de nódulos de *Phaseolus vulgaris* presentes en 3 sitios distintos. Las plantas se clasificaron de acuerdo al sitio en el cual fueron halladas como: Silvestres (12 cepas), Silvestroides (3 cepas) y Domesticadas (8 cepas) (Gasca, J. 2004).

C) Población Andalucía, España

La población denominada 'España' presenta cepas de 5 distintas especies de *Rhizobium*. Las cepas fueron aisladas de suelos donde no se cultiva actualmente *P. vulgaris* (Rodríguez-Navarro, et al. 2000). Las muestras de suelo se colectaron a lo largo del Valle del Río Guadalquivir.

Tabla 2. Colección de Cepas de las tres Poblaciones

Origen	No. Cepas	Especies	Identificador	Referencia
Cepas de *Rhizobium*				
Puebla, Mexico	30	*Rhizobium etli*	1009, 1006, 4771, 1004, 2737, 4815, 4803, 4777, 950, 2704, 4813, 4794, 951, 994, 4837, 4804, 954, 4877, 4795, 2730, 4810	Silva, C. et al. 2003
		Rhizobium gallicum	4868, 4872, 4770, 4845, 988, 2751, 2735, 992, 2703	
Guanajuato, Mexico	30	*Rhizobium etli*	Domesticadas: 6854, 6861, 6857, 6832, 6868, 6867, 6840, 6846, 6850, 6851, 6833, 6845, 6862	Gasca, J. 2004
			Silvestre: 6779, 6778, 6760, 6794, 6805, 6795, 6766, 6784, 6797, 6798, 6776, 6768, 6763	
			Silvestroide: 6815, 6824, 6823, 6813	
Valle Guadalquivir, Espana	27	*Rhizobium etli*	21PR-1, 16C-1, 21NJ-2, 8C-3, 8NJ-2, 14PR-2, 17NJ-2, 16NJ-2, 14C-1, 4PR-2, 17C-2, 6PR-1, 6C-1, 4C-2, GR10, GR62, GR14, GR87, GR56	Rodriguez – Navarro, D. et al. 2000
		Rhizobium gallicum	GR18, GR45, GR60, GR42	
		Rhizobium giardini	GR93, GR03	
		Sinorhizobium fredii	GR64	
		Rhizobium leguminosarum	GR84	
Cepas de Referencia				
Mexico	1	*Rhizobium etli*	CFN42	Gonzalez, V. et al. 2006
Costa Rica	1	*Rhizobium etli*	CIAT652	Flores, M. et al. 2005

Amplificación y Secuenciación

El genoma de *R. etli* CFN42 consta de 39 SIs (González, et al. 2006) de las cuales se seleccionaron 38 SIs (Figura 2) presentes en el cromosoma (11 SIs), el plásmido simbiótico (13 SIs) y el plásmido conjugativo (14 SIs). La primera parte de este trabajo, como se mencionó anteriormente, consistió en buscar las SIs de CFN42 en la misma localización (contexto genómico) en la colección de cepas de las tres poblaciones. Para lograr esto, a cada SI se le diseñó un par de oligos específicos situados en los genes vecinos. De esta forma, si se obtenía un producto de PCR del mismo tamaño que en CFN42 (control), se podía esperar que la SI estuviera presente. Un producto de PCR menor al tamaño esperado, sería indicativo de la ausencia de la SI. Finalmente, la presencia o ausencia de la SI en el mismo contexto genómico, sólo podría ser corroborada mediante la secuenciación del producto de PCR.

Se seleccionaron tres SIs para ser secuenciadas, la ISRel9, la ISRel4 y la ISRel2 del plásmido simbiótico en aquellas cepas que las presentaron. Solamente las secuencias de la ISRel4 y ISRel2 fueron empleadas para el análisis en el presente trabajo. Por otro lado, se seleccionaron otros tres genes para ser secuenciados en todas las cepas de las tres poblaciones:

a) El gen *nodC* que codifica para la N-acetilglucosaminil transferasa relacionada con la formación de factores Nod (lipo-quito-oligosacáridos) que participan en el inicio de la morfogénesis de los nódulos de la raíz en leguminosas. Este gene se encuentra en el plásmido simbiótico.

b) El gen *glyA* que codifica para la proteína serin hidroximetiltransferasa, la cual participa en la biosíntesis de purinas, lípidos y hormonas. Este gen se encuentra localizado en el cromosoma de CFN42.

c) El gen *dnaB* que codifica para una helicasa que participa durante la replicación y elongación al abrir la cadena de DNA, y al igual que *glyA* se encuentra en el cromosoma.

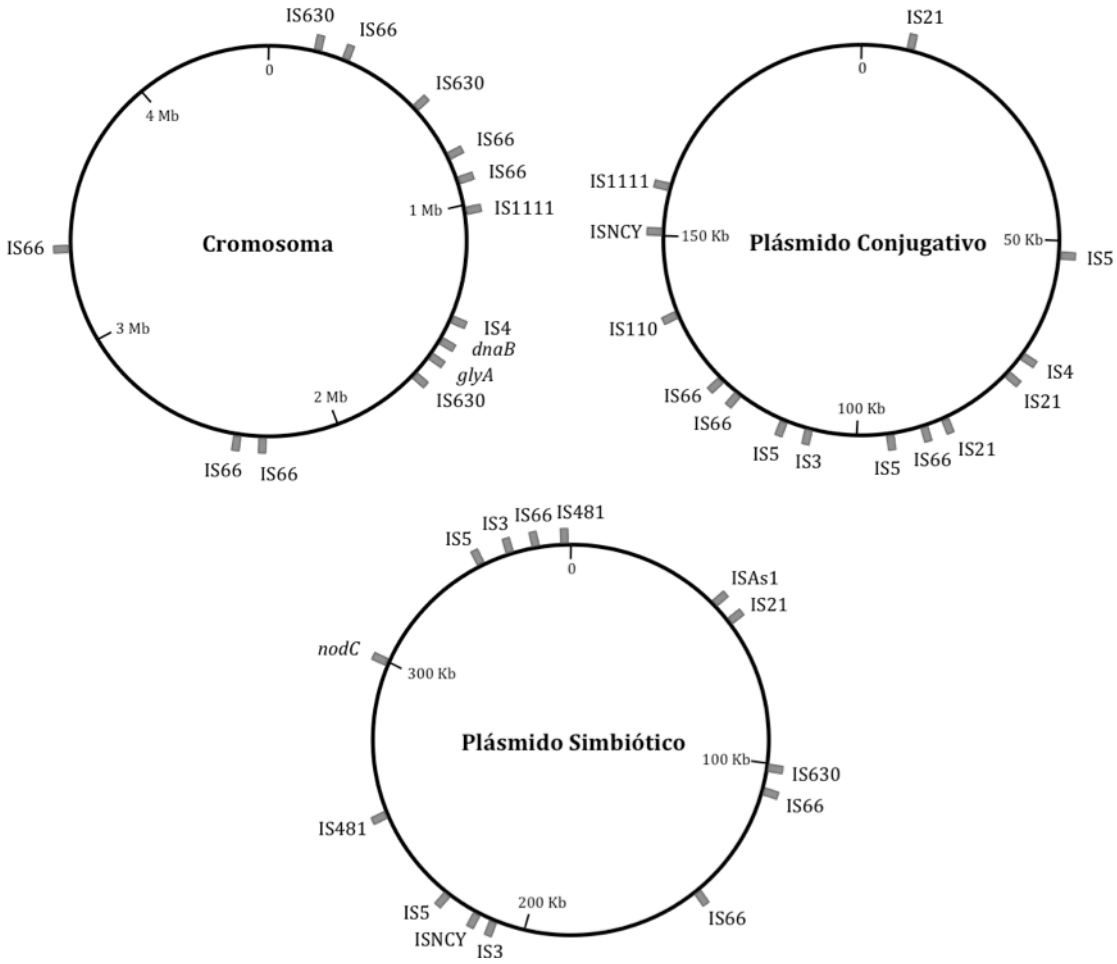

Figura 2. Distribución de SIs en el cromosoma, y los plásmidos simbiótico y conjugativo. Se muestra la localización de los genes *dnaB*, *glyA* y *nodC*.

Análisis de Secuencias y Reconstrucciones Filogenéticas

La calidad de los productos de secuenciación de cada gen secuenciado se analizó mediante el programa SeqScape V2.5 de Applied Biosystems, el cual permite evaluar la calidad de cada base mediante el electroferograma de la secuencia. Todas las secuencias fueron revisadas para evitar incluir bases incorrectas. Posteriormente se realizaron alineamientos múltiples con el programa MUSCLE (Edgar, R. 2004) para los cinco conjuntos de secuencias (ISRel4, ISRel2, *nodC*, *glyA* y *dnaB*). Se realizaron reconstrucciones filogenéticas para los dos genes del cromosoma (*glyA* y *dnaB*) de forma independiente. Por otro lado, se realizó una filogenia para ambos genes concatenando sus respectivos alineamientos múltiples. Todas las reconstrucciones

filogenéticas se realizaron mediante el método de máxima verosimilitud implementado en el programa PHYML (Guidon, S. et al. 2002) con un análisis de bootstrap de 100 replicas por alineamiento.

Genética de Poblaciones y Eventos de Recombinación

Se realizaron distintas pruebas de genética de poblaciones usando el programa DNASP v.4 (Librado, P. et al. 2009) para los cinco genes del presente trabajo, las cuales incluyeron:

a) Diversidad nucleotídica promedio por sitio (π)

La diversidad nucleotídica es una medida de variación genética que mide el grado de polimorfismos dentro de una población. Se denota como π, y se define como el promedio de las diferencias nucleotídicas por sitio entre dos secuencias dentro de una población.

b) Diversidad nucleotídica promedio en sitios sinónimos (πs) y no sinónimos (πns)

Al igual que π, se obtiene el promedio de las diferencias nucleotídicas entre dos secuencias, pero en este caso se analizan por separado los sitios sinónimos y no sinónimos.

c) Diversidad nucleotídica promedio entre poblaciones (Dxy)

La diversidad nucleotídica promedio entre poblaciones es una medida que da información respecto al grado de diferenciación entre las secuencias de un gen entre poblaciones, es decir, mide la divergencia en base a la variación de las secuencias.

d) Flujo genético (F_{st} y Nm)

El flujo genético es la transferencia de uno o varios genes de una población a otra, lo que provoca cambios en la proporción de las distintas variantes de cada gen, lo cual modifica el pool genético de dicha población. Entre los distintas formas de analizar el flujo genético, una de la más empleadas es la medida F_{st} desarrollada por Wright, la cual muestra la variación de las frecuencia alélica entre las poblaciones normalizado por la frecuencia alélica promedio. A partir de la medida F_{st} se puede así

mismo tener un estimado del numero de migrantes que tuvo la población por generación (Nm).

e) Diferenciación Genética (Ks, Kst, Z y Snn)

Las distintas medidas de diferenciación genética como Kst han sido empleadas en gran medida para analizar la subdivisión geográfica de poblaciones. De esta forma, se puede conocer si dos poblaciones son genéticamente diferentes, debido a que permiten evaluar que tan significativos son los resultados de sus respectivas pruebas estadísticas.

Prueba de neutralidad de Tajima (D)

La prueba de neutralidad de Tajima estima la diferencia entre la diversidad nucleotídica promedio (π) y la diversidad esperada en cada sitio (θ) siendo la evolución neutral. El índice de Tajima (D) puede tener tres distintas interpretaciones: si el índice es menor que cero significa que los genes pueden estar bajo el efecto de la selección negativa o purificadora, debido a que π (diversidad observada) es menor a la θ (diversidad esperada); si el índice es igual a cero significa que los genes han evolucionado de forma neutral; si el índice es mayor que cero significa que la diversidad observada es mayor que la diversidad esperada, lo cual significa que los genes pueden estar bajo una selección negativa o purificadora.

Análisis de Recombinación

Se realizaron pruebas de recombinación para los cinco genes empleando el programa RDP3 (Martin, D. P. et al. 2005). RDP3 utiliza de forma simultánea diferentes métodos para la detección de los eventos de recombinación entre un grupo de secuencias alineadas. Entre los métodos empleados en el presente trabajo están: RDP, GENECONV, Bootscan, MaxChi, Chimaera, SiScan, 3SEQ y LARD.

Reconstrucción de Redes Filogenéticas

Por otro lado, se realizaron reconstrucciones de redes filogenéticas para los cinco genes mediante el programa SplitsTree4 (Huson, D. H. et al. 2006). Las reconstrucciones se hicieron con dos métodos: Split decomposition y Neighbor Net. A diferencia de las filogenias realizadas con el programa PHYML con el método de

máxima verosimilitud, las redes filogenéticas muestran de forma gráfica aquellas incompatibilidades entre las secuencias alineadas que no soportan la filogenia original. En el alineamiento de las secuencias existen posiciones que no son compatibles con la filogenia final, al realizar una red filogenética, estas incompatibilidades se representan en forma de redes o retículas en la filogenia. La recombinación homóloga es uno de los mecanismos que pueden generar este tipo de señales conflictivas provocando la formación de retículas en la gráfica. Aunado al análisis de recombinación mencionados anteriormente, la reconstrucción de redes filogenéticas pueden ayudar a determinar qué secuencias han tenido eventos de recombinación.

ARTICULO

Evolutionary Dynamics of Insertion Sequences in Relation to the Evolutionary Histories of the Chromosome and Symbiotic Plasmid Genes of *Rhizobium etli* Populations

Applied and Environmental Microbiology, Oct. 2010, p. 6504–6513 Vol. 76, No. 19
0099-2240/10/$12.00 doi:10.1128/AEM.01001-10

Evolutionary Dynamics of Insertion Sequences in Relation to the Evolutionary Histories of the Chromosome and Symbiotic Plasmid Genes of *Rhizobium etli* Populations▽†

Luis Lozano,[1]* Ismael Hernández-González,[1] Patricia Bustos,[1] Rosa I. Santamaría,[1] Valeria Souza,[2] J. Peter W. Young,[3] Guillermo Dávila,[1] and Víctor González[1]

Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Av. Universidad N/C Col. Chamilpa, Apdo. Postal 565-A, Cuernavaca, Morelos, México[1]; Departamento de Ecología Evolutiva, Instituto de Ecologia, Universidad Nacional Autónoma de México, AP 70-275, CU, Coyoacán 04510, Mexico DF, México[2]; and Department of Biology, University of York, York YO10 5YW, United Kingdom[3]

Received 26 April 2010/Accepted 26 July 2010

Insertion sequences (IS) are mobile genetic elements that are distributed in many prokaryotes. In particular, in the genomes of the symbiotic nitrogen-fixing bacteria collectively known as rhizobia, IS are fairly abundant in plasmids or chromosomal islands that carry the genes needed for symbiosis. Here, we report an analysis of the distribution and genetic conservation of the IS found in the genome of *Rhizobium etli* CFN42 in a collection of 87 *Rhizobium* strains belonging to populations with different geographical origins. We used PCR to generate presence/absence profiles of the 39 IS found in *R. etli* CFN42 and evaluated whether the IS were located in consistent genomic contexts. We found that the IS from the symbiotic plasmid were frequently present in the analyzed strains, whereas the chromosomal IS were observed less frequently. We then examined the evolutionary dynamics of these strains based on a population genetic analysis of two chromosomal housekeeping genes (*glyA* and *dnaB*) and three symbiotic sequences (*nodC* and the two IS elements). Our results indicate that the IS contained within the symbiotic plasmid have a higher degree of genomic context conservation, lower nucleotide diversity and genetic differentiation, and fewer recombination events than the chromosomal housekeeping genes. These results suggest that the *R. etli* populations diverged recently in Mexico, that the symbiotic plasmid also had a recent origin, and that the IS elements have undergone a process of cyclic infection and expansion.

Insertion sequences (IS) are the smallest transposable elements found in prokaryotes (usually less than 3 kb in size). They encode a transposase and may also encode small hypothetical proteins (4, 9). IS are distinguished by their ability to move within prokaryotic replicons, including both the chromosome and plasmids, and to copy themselves into various genomic sites. In this manner, IS elements can inactivate or alter the expression of adjacent genes (4). When IS occur in two or more identical copies within a genome, they can participate in various types of genetic rearrangements (e.g., duplications, inversions, and deletions), suggesting that IS may play an important role in the evolution of their hosts by promoting genomic plasticity (34). Due to these evolutionary dynamics, the diversity and distribution of IS elements differ greatly between taxa and even within strains of the same species (27).

Various theories have been proposed to explain the evolution of IS elements in laboratory model strains and environmental bacterial populations (8, 18, 25, 29). Two main hypotheses seek to explain how these elements are maintained over the long term in their host genomes. The first proposes that they occasionally generate beneficial mutations and therefore may represent a selective advantage to their hosts (34). The second suggests that IS elements are genomic parasites that are maintained by their high rate of transposition and might be disseminated among different bacterial lineages by horizontal gene transfer (HGT). Data supporting the second hypothesis have shown that some IS elements may transpose at high rates upon entering a new host (42). After the initial infection, however, purifying selection may continuously remove these elements from the genome. Thus, IS may undergo an infection-expansion-extinction cycle that allows them to remain in different bacterial populations within the gene pool (42). These two hypotheses are not contradictory, and the evolutionary dynamics and distribution of IS may differ greatly depending on several factors, including (most notably) the rate of transposition and HGT, as well as selective pressures, population size, and the host's habitat (6, 18, 21, 25, 27, 29).

In the nitrogen-fixing symbiotic bacteria of the genera *Rhizobium*, *Sinorhizobium*, *Mesorhizobium*, *Bradyrhizobium* (of the alphaproteobacteria), *Cupriavidus*, and *Burkholderia* (of the betaproteobacteria), IS are particularly abundant in symbiotic plasmids (pSym) and symbiotic chromosomal islands (SI) (2, 12, 14, 19, 20, 43). SI and pSym include most of the genes needed to establish symbiosis in the roots of leguminous plants through nodule formation and nitrogen fixation (11). It is generally believed that these elements entered the rhizobial ge-

* Corresponding author. Mailing address: Centro de Ciencias Genomicas, UNAM, Av. Universidad N/C, Col. Chamilpa, Apdo. Postal 565-A, Cuernavaca, Morelos, Mexico. Phone: 01 777 3291690. Fax: 01 777 3175581. E-mail: llozano@ccg.unam.mx.

† Supplemental material for this article may be found at http://aem.asm.org/.

▽ Published ahead of print on 30 July 2010.

TABLE 1. *Rhizobium* isolates and reference strains used in this study

Origin	No. of strains	Species	Identifier(s)	Reference
Rhizobium isolates				
Puebla, Mexico	30	*R. etli*	1009, 1006, 4771, 1004, 2737, 4815, 4803, 4777, 950, 2704, 4813, 4794, 951, 994, 4837, 4804, 954, 4877, 4795, 2730, 4810	34
		R. gallicum	4868, 4872, 4770, 4845, 988, 2751, 2735, 992, 2703	
Guanajuato, Mexico	30	*R. etli*	Domesticated: 6854, 6861, 6857, 6832, 6868, 6867, 6840, 6846, 6850, 6851, 6833, 6845, 6862 Wild: 6779, 6778, 6760, 6794, 6805, 6795, 6766, 6784, 6797, 6798, 6776, 6768, 6763 Weedy: 6815, 6824, 6823, 6813	12
Valle Guadalquivir, Spain	27	*R. etli*	21PR-1, 16C-1, 21NJ-2, 8C-3, 8NJ-2, 14PR-2, 17NJ-2, 16NJ-2, 14C-1, 4PR-2, 17C-2, 6PR-1, 6C-1, 4C-2, GR10, GR62, GR14, GR87, GR56	31
		R. gallicum	GR18, GR45, GR60, GR42	
		R. giardinni	GR93, GR03	
		R. fredii	GR64	
		R. leguminosarum	GR84	
Rhizobium reference strains				
Mexico	1	*R. etli*	CFN42	13
Costa Rica	1	*R. etli*	CIAT652	10

nomes through HGT (39, 40). Comparative genomic analyses have shown that both pSym and SI are highly variable, with the exception of a common set of genes encoding factors critical to nitrogen fixation (*nif*) and nodulation (*nod*) (5, 14). SI and pSym have been found to have lower GC contents and different codon usages than the corresponding chromosomal and nonsymbiotic plasmid sequences, suggesting that they were recently acquired by HGT.

Some of these symbiotic elements, such as in pSym of *Rhizobium etli* CFN42 and the SI of *Mesorhizobium loti*, are conjugative and mobile (30, 32). Genomic analysis of *R. etli* CFN42 revealed the presence of 39 IS belonging to different families (14); they were found in the chromosome (11 IS); the 371-kb symbiotic plasmid (13 IS); the smaller 192-kb conjugative plasmid, p42a (13 IS); and two other plasmids, p42b and p42c (2 IS). Interestingly, this particular strain shows no evidence of IS disrupting open reading frames (ORFs) or having transpositional activity. However, another 42 incomplete IS may be found in the chromosome, pSym, and the conjugative plasmid; these incomplete sequences are truncated or contain stop codons in their coding sequences.

Here, we focused on the dynamics and distribution of IS in different populations of the nitrogen-fixing symbiont *R. etli*. Since the maintenance of IS in bacterial species might depend on their transpositional activities and horizontal transfer rates, the identification of IS in the same genomic contexts across different strains of the same species could provide new insights into their persistence and divergence over short evolutionary periods. To examine the evolutionary dynamics of IS in natural populations of *R. etli*, we characterized the distributions, genomic contexts, and sequence variations of IS in isolates of *R. etli* from three populations with different origins, as well as in some other *Rhizobium* species. More specifically, we used PCR to generate presence/absence profiles of the 39 IS found in *R. etli* CFN42 in a collection of 87 strains representing different geographical sites and a gradient of domestication of the bacterial host, the common bean (*Phaseolus vulgaris*). We also evaluated whether the IS were conserved in the same genomic context relative to their position in *R. etli* CFN42 and determined the nucleotide sequences of two IS found in most of the isolates. Several population genetic tests applied to these IS, another pSym gene (*nodC*), and two chromosomal housekeeping genes (*dnaB* and *glyA*) suggest that these two IS elements have been inherited vertically and represent recent components of the *R. etli* gene pool. Finally, the present study strongly suggests that symbiotic plasmids have a recent origin within the *R. etli* populations.

MATERIALS AND METHODS

Bacterial strains, growth conditions, and DNA extraction. The 87 *Rhizobium* strains used in this study correspond to three different collections (Table 1). Two of the strain collections are from Mexico. The first collection (36) was derived from two plots in a traditional milpa system of native bean landraces (San Miguel Acuexcomac, Puebla, Mexico); the collection consists of 30 strains that were the dominant strains for several years. The second collection (13) comes from the Michoacan-Guanajuato area, which is the reported center of origin of bean domestication (24); the 30 strains in this collection include a bean plant domestication gradient from wild, nondomesticated *P. vulgaris* to milpa landrace cultivars, as well as wild bean plants that are probably the descendants of cultivated plants. The third collection (33) represents *R. etli* from Spain and includes 27 strains obtained from 21 soil samples collected along the Guadalquivir River Valley. They are believed to represent either original native rhizobia or a subsample of New World rhizobia that traveled along with the original bean seeds (Table 1). Some strains from the Puebla and Spanish collections were previously characterized as *Rhizobium gallicum*, *Rhizobium giardini*, *Rhizobium fredii*, and *Rhizobium leguminosarum* (Table 1).

The various *Rhizobium* strains were grown in peptone-yeast (PY) medium for 24 to 48 h. Genomic DNA was extracted with a GenomicPrep DNA Isolation Kit (Amersham Biosciences), and the DNA concentration was determined with a DyNA Quant 200 fluorometer (Hoefer). Two completely sequenced *R. etli* strains, CFN42 and CIAT652, were used as references strains.

PCR amplification and DNA sequencing. We first localized each of the 39 selected IS within the CFN42 genome, thereby "anchoring" our examination of whether these IS conserved their genomic locations (i.e., their synteny) across the wild *R. etli* isolates from different geographical regions. These 39 IS elements represent 11 different families, and some of these families have identical or nearly identical copy numbers: 6 and 4 copies for 2 different elements of the IS*66* family, 5 copies for IS*630*, 2 for IS*1111*, and 2 for IS*21*. We then designed specific PCR primers spanning the immediate neighborhoods of the 39 studied IS. These

primers allowed us to test for conservation (i.e., amplification of a fragment of a size similar to that predicted in CFN42). In cases where we found two contiguous IS elements, we designed primers in the neighboring genes of both IS. There were three possible results: (i) no PCR product, indicating that there were no identical priming sites in the wild isolate (this was not an absolute positive or negative result for the presence of the IS); (ii) a small DNA fragment equivalent to the distance between the two sequences near the insertion site but lacking the IS (a negative result for the IS); and (iii) a large DNA fragment of a size similar to that in CFN42 (a positive result for the IS). The PCRs were performed in a DNA thermal cycler (Gene Amp 9700; Applied Biosystems) in a 25-μl reaction mixture containing approximately 10 ng genomic template DNA, 2 mM $MgCl_2$, 1 mM deoxynucleoside triphosphates (dNTPs), 5 pmol of each primer, and 2 U of *Taq* polymerase (AltaEnzymes). The mixtures were subjected to 5 min of denaturation at 94°C followed by 30 cycles of 1 min at 94°C, 1 to 3 min at 58 to 62°C, and 3 min at 72°C.

For a comparison of evolutionary dynamics, we performed direct sequencing on the following: two IS elements, IS*Rel4* and IS*Rel2* from pSym, which were present in the highest proportion of test samples (see Results and Fig. 2); *nodC*, which is a pSym gene encoding an *N*-acetylglucosaminyltransferase that participates in the nodulation process; and two chromosomal housekeeping genes, *glyA* (serine hydroxymethyltransferase) and *dnaB* (replicative DNA helicase), which have been proposed to serve as predictors of genome relatedness because their sequence divergence rates reflect the overall rate of genome divergence (44). Internal PCR primers were designed for the last three genes to obtain partial sequences for each gene. The sequencing reaction mixtures contained approximately 10 ng genomic template DNA, 1 mM $MgCl_2$, 1 mM dNTPs, 5 pmol of each primer, and 1 U of r*Tth* Polymerase XL (Applied Biosystems). The mixtures were subjected to 5 min of denaturation at 92°C, followed by 30 cycles of 30 s at 92°C and 1 to 6 min at 58 to 62°C. The PCR products were purified with an Exo/SAP kit (Affymetrix) and sequenced using a Dye-terminator cycle-sequencing kit (Perkin Elmer Applied Biosystems). The sequencing reactions were run on an ABI 3730 sequencer (Applied Biosystems).

Sequence analysis, alignments, and phylogenetic reconstruction. Sequence quality analysis, assembly, and comparison were performed using SeqScape software V2.5 (Applied Biosystems). For IS characterization, we compared all of the putative transposases in the complete annotated genome sequence of *R. etli* CFN42 to the nr Database of NCBI and the Insertion Sequence Database (35) with BLASTp (1).

To define a complete and (most likely) functional IS, we used the same criteria applied by the authors of the Insertion Sequence Database as follows: (i) similar organizations of the genes, (ii) sequence similarity of the transposases and inverted repeats, and (iii) presence of direct repeats (14). Gene alignments for the assessed genes and IS elements were done using the MUSCLE program (7). For each alignment, the best substitution model was determined using FindModel (31). Phylogenetic reconstruction was performed using the maximum-likelihood method implemented in PHYML (16). The analysis was carried out with a nonparametric bootstrap analysis of 100 replicates for each alignment. The complete chromosome and plasmid sequences of *R. etli* CFN42 and CIAT652 were compared with the Mummer program (23). The genomic contexts of the IS in the two strains were examined using Perl programs. The rate of synonymous and nonsynonymous substitutions, the dn/ds ratio, was evaluated using SNAP (22).

Population genetic parameter estimation. DNASP version 4 (26) was used to assess the following parameters: the average nucleotide divergence between populations (D_{xy}), the average nucleotide diversity within populations, the average nucleotide divergence per site (π), the average nucleotide diversity at synonymous (π_s) and nonsynonymous (π_{ns}) sites, gene flow estimates (F_{st} and N_m), genetic differentiation estimates (K_s, K_{st}, Z, and S_{nn}), the numbers of shared haplotypes and fixed and shared polymorphisms, and Tajima's D neutrality test.

Recombination analysis. Recombination tests were performed with RDP3 (28) and SplitsTree4 (17) for the five genes studied in each of the collections. The RDP3 software was applied to detect and analyze recombination signals using eight different methods (RDP, GENECONV, Bootscan, MaxChi, Chimaera, SiScan, 3SEQ, and LARD), with 100 permutation steps, a Bonferroni correction for multiple tests, and a P value threshold of 0.05. Recombination events and breakpoints were accepted if they were detected by two or more methods that used different approximations to detect recombination. The recombination events were then confirmed independently for each recombinant strain, as suggested in the literature (RDP website user manual [http://darwin.uvigo.es/rdp/rdp.html]) (28). Specific parameter modifications were used for each gene alignment depending on the number of sequences in the analysis. Split decomposition and neighbor net analyses were performed with SplitsTree4 software. Both phylogenetic networks represent incompatibilities within and between data sets by the use of splits. For the neighbor net analysis, we applied a split filter based on the network weight and a threshold value of 95% (3).

Nucleotide sequence accession numbers. The relevant sequence data have been deposited in the GenBank database under accession numbers GUO84443 to GUO84796.

RESULTS

Differentiation of *R. etli* populations. To study the evolutionary dynamics of the IS of *R. etli*, we first asked if the three studied collections of *R. etli* strains, which were obtained from the root nodules of *P. vulgaris* plants located in different areas, were geographically structured. The phylogenies constructed based on the chromosomal housekeeping genes, *glyA* and *dnaB*, were relatively similar to each other, with a few differences in the groupings of some strains (data not shown). In order to obtain a phylogenetic reconstruction that more accurately represented the evolutionary relationships among the 87 strains contained within the three collections, we concatenated the sequence alignments (2,238 bp) of both genes. We obtained three large *R. etli* clusters that broadly corresponded to the three different geographical areas from which the strains were collected. The Puebla collection formed one group with some intermingled strains from the Michoacan-Guanajuato and Spanish collections (Fig. 1). The second group consisted of the Michoacan-Guanajuato strains with three intermingled sequences belonging to the Spanish collection. The third *R. etli* group represented the Spanish collection with a single intermingled sequence from the Puebla collection. The Michoacan-Guanajuato and Spanish collections were more closely related to one another than to the Puebla collection (Fig. 1). A distant fourth group, which was later used as the outgroup, contained several *R. gallicum* strains from Puebla and Spain. Since this phylogenetic reconstruction clearly separated the three different collections according to their geographic origins, we considered them to be different populations in the subsequent analyses.

IS profiles of the *R. etli* populations. We next examined which IS from *R. etli* CFN42 were also maintained in the same genomic context among the *R. etli* isolates from the three different populations. Our PCR-based profiling (see Materials and Methods) yielded either positive or negative results for 24 of the 39 IS and showed marked differences in the conservation of the various IS among the tested populations (Fig. 2). IS*Rel4* and IS*Rel2* (from pSym of *R. etli* CFN42) were the most common IS, present in more than 50% of the individual strains. Sequencing experiments showed that IS*Rel9* (also from pSym) in most cases was actually a hypothetical protein (hyp304) not related to any IS element and similar in size to IS*Rel9*. Only five isolates (all from Puebla) had an intact IS*Rel9* in the expected genomic context; these elements had an average identity of 100%. IS*Rel5* and IS*Rel12* of pSym were relatively frequent, as they were present in 25% of the tested strains. The other seven IS from pSym were not common. The IS from p42a were found in the Puebla and Michoacan-Guanajuato populations, but not in the collection from Spain, while the chromosomal IS were found only in the Puebla strains. Interestingly, the CFN42 reference strain was found to be close to the Michoacan-Guanajuato population (Fig. 1).

DNA sequence divergence of IS*Rel4* and IS*Rel2*. The observation that some IS occur in the same genomic context suggests

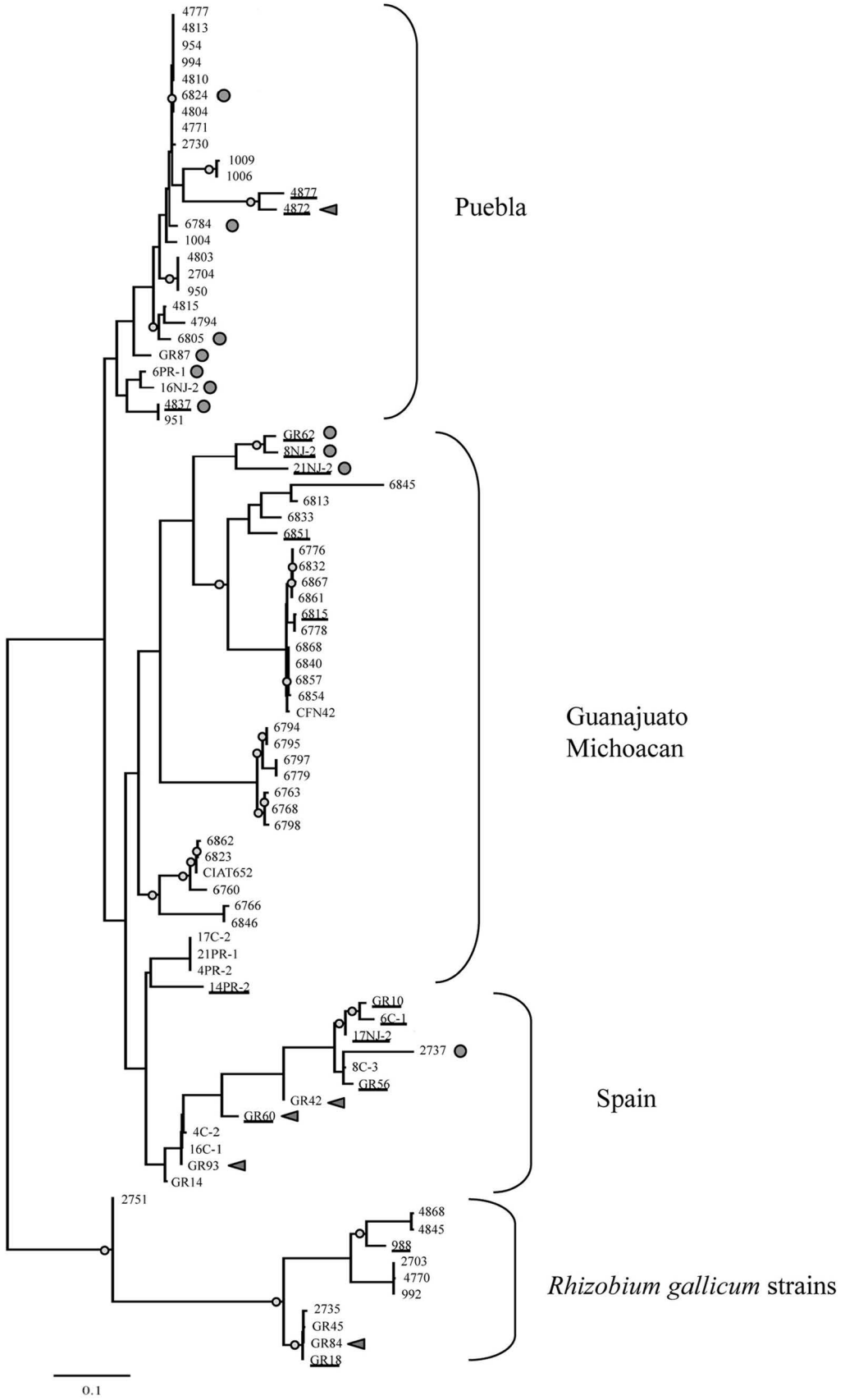

FIG. 1. Phylogenetic reconstruction representing the evolutionary relationships across the three collections. The circles inside the branches indicate bootstrap support of >70%. The external circles indicate intermingled *R. etli* strains from one of the other collections. The arrowheads indicate other intermingled species (4872, GR42, and GR60 are *R. gallicum*; GR93 is *R. giardinii*; and GR84 is *R. leguminosarum*). The underlined strains represent recombinant strains for *dnaB* or *glyA* between the three populations and the *R. gallicum* strain.

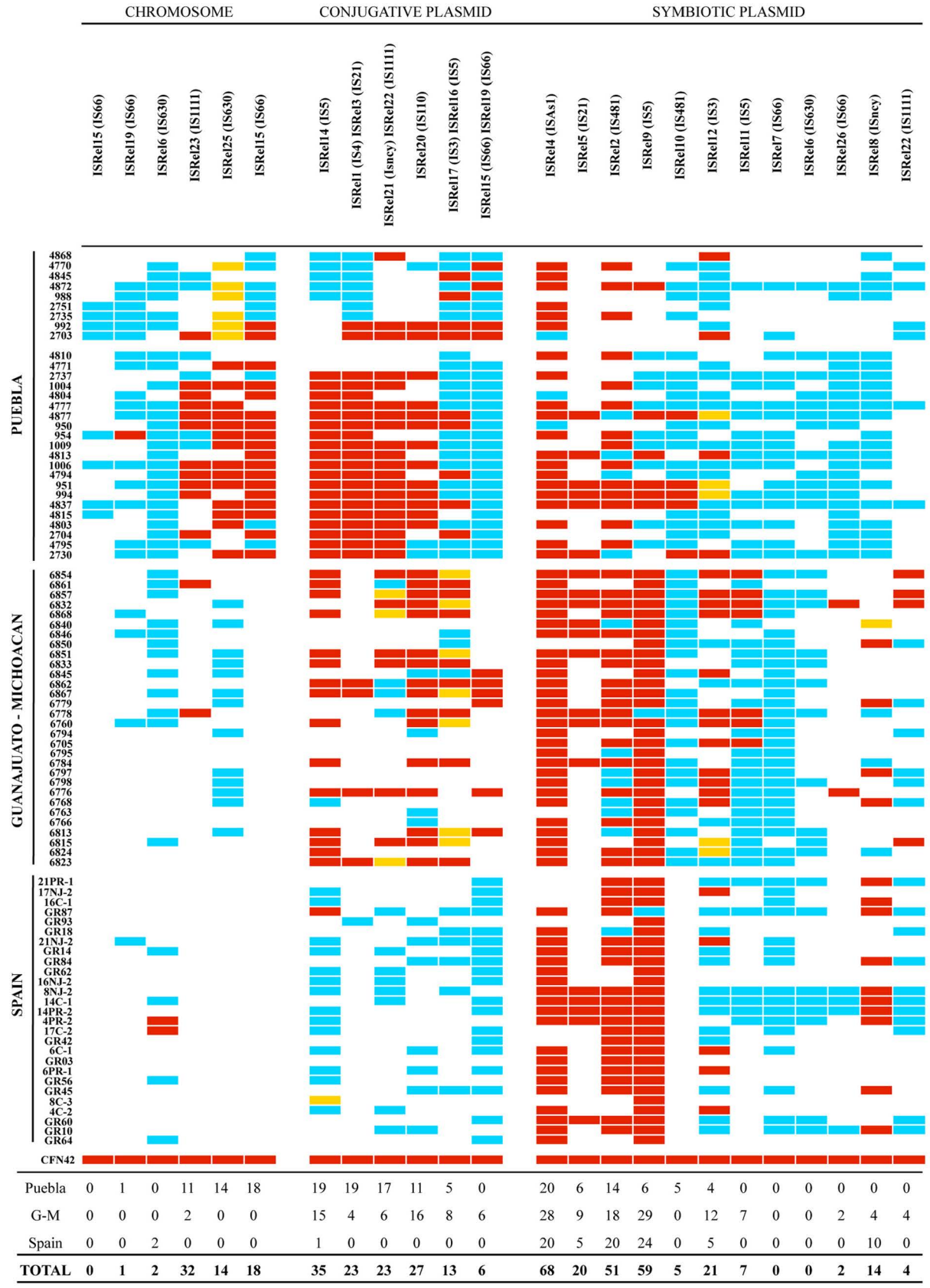

FIG. 2. Presence/absence profile of the IS elements present in the chromosome and conjugative and symbiotic plasmids across the three populations. The bottom line represents the CFN42 control strain. The red bars indicate positive confirmation of the IS element (a DNA fragment of similar size to that in CFN42), the blue bars indicate a negative result for the presence of the IS element (a DNA fragment equivalent to the distance between the two sequences near the insertion site but lacking the IS), no color indicates no PCR product, and the yellow bars indicate DNA fragments that were bigger than expected.

TABLE 2. DNA divergence in *R. etli* IS elements, plasmid genes, and chromosomal genes

Gene	No. of polymorphic sites[a]	dn/ds	Θ[b]	Nucleotide diversity (π)	Nucleotide diversity at synonymous sites (π_s)	Nucleotide diversity at nonsynonymous sites (π_{ns})	Tajima's *D*
dnaB	263 (23.6)	0.019	0.047	0.066	0.237	0.009	1.39
glyA	278 (24.7)	0.029	0.070	0.088	0.122	0.075	0.84
nodC	19 (3.3)	0.355	0.007	0.014	0.028	0.009	0.72
IS*Rel2*	6 (0.6)	0.267	0.001	0.002	0.001	0.002	2.72
IS*Rel4*	13 (1.3)	0.325	0.002	0.002	0.003	0.007	1.50

[a] The percentages of polymorphic sites per sequence length are shown in parentheses.
[b] Population mutation rate (per bp).

that these IS may have been present since the arrival of pSym in the populations, as it is far less likely that independent transposition events could have inserted the IS into the same genomic sites. It is important to mention that IS*Rel4* and IS*Rel2* belong to IS families, IS*As1* and IS*481*, respectively, that do not have specific target sequences for transposition. To investigate the evolutionary dynamics of the two IS elements, IS*Rel4* and IS*Rel2*, that show conservation of their genomic contexts, we looked for differences in the degree of diversification and selection pressures among these IS elements, as well as in other symbiotic and chromosomal housekeeping genes (see Table S1 in the supplemental material). The DNA sequence conservation was very high for IS*Rel4* and IS*Rel2*, with average identities of 98.7% and 99.4%, respectively. This is surprisingly high compared to the sequence identities for the housekeeping genes *glyA* and *dnaB*, which were 75.3% and 76.4%, respectively. Given that the diversification of chromosomal genes could differ from that on the symbiotic plasmid, we sequenced the *nodC* gene, which encodes an *N*-acetylglucosaminyltransferase involved in the synthesis of sugar backbones for the production of lipooligosaccharides (critical for nodulation signaling). This was done in order to have another gene to trace the diversification differences between pSym and the chromosome. The average nucleotide identity among the 44 analyzed *nodC* sequences was 96.7%. The nucleotide diversity (π) values for the pSym sequences (IS*As1*, IS*481*, and *nodC*) were very low compared with those of *glyA* and *dnaB* (Table 2). The total average π values of *dnaB* and *glyA* were 0.066 and 0.088, respectively, whereas those for IS*Rel4*, IS*Rel2*, and *nodC* were 0.002, 0.002, and 0.014, respectively (Table 2). Since the number of polymorphic sites and the π values for IS*Rel4*, IS*Rel2*, and *nodC* were very low, we hypothesize that these sequences may have entered the gene pools of the three populations relatively recently.

Recent origins of IS*Rel4*, IS*Rel2*, and pSym. To evaluate the hypothesis that IS*Rel4* and IS*Rel2* may have originated recently, we measured their average nucleotide diversities at synonymous sites (π_s) and nonsynonymous sites (π_{ns}) and compared these values with those for *nodC*, *glyA*, and *dnaB*. The π_s value reflects the age of an allele in the genetic pool; genes with lower π_s values are generally considered to have a recent common ancestor. Our results revealed that the π_s values were low for IS*Rel4*, IS*Rel2*, and *nodC* but high for *dnaB* and *glyA*, whereas the π_{ns} values were low for all five sequences (Table 2). Similar results were obtained when these π_s and π_{ns} values were measured within each population (see Table S2 in the supplemental material). These findings indicate that both the IS elements and *nodC* might have recently entered the gene pools of the three populations.

Overall, the dn/ds ratio was lower for the chromosomal genes than for the pSym genes. To test what this might mean in terms of pSym evolution, we compared the shared haplotypes for chromosomal and pSym genes among the three populations. We did not find any shared *dnaB* and *glyA* (i.e., chromosomal) haplotypes between the populations. In contrast, we identified a number of shared haplotypes for the pSym genes (*nodC*, IS*Rel2*, and IS*Rel4*) (data not shown). These results suggest either that pSym has a relatively recent origin or that there is high gene flow among pSym sequences; in any case, it seems that the pSym sequences have a different evolutionary history than the chromosomal genes. To further test the possible recent origin of pSym sequences in the gene pools of the three populations, we compared the mean genetic divergence values between populations (D_{xy}) to the π values describing the mean genetic divergence within the populations. Higher values of D_{xy} indicate increasing time since population divergence. Table 3 shows that the D_{xy} values were higher than the mean within-group divergence (π) for the chromosomal genes, potentially reflecting phylogenetic differentiation among the three populations. In the case of the pSym sequences, however, the D_{xy} and π values were very low and did not show clear phylogenetic differentiation between the populations. This further supports the idea that the pSym sequences have a relatively recent origin in the gene pools of the three populations and have followed a different evolutionary history than the chromosomal genes.

We applied additional genetic-differentiation tests (K_s^*, K_{st}, Z, Z^*, and S_{nn}) to the pSym and chromosomal sequences to test whether the two IS and *nodC* have a lower level of genetic differentiation than the chromosomal sequences, which would further support a recent origin within the three populations. The results of the genetic-differentiation tests for the two housekeeping genes were highly significant ($P < 0.001$), supporting the idea of genetic differentiation among the three populations (Table 4; see Table S3 in the supplemental material). In contrast, the three pSym sequences had much lower levels of genetic differentiation across the three populations. For example, comparison of the pSym sequences from the two Mexican populations (those from Puebla and Guanajuato-Michoacan) yielded K_{st} values that were either nonsignificant or just barely significant at a P value of <0.05 (Table 4). Although the chromosomal housekeeping genes showed evidence of genetic differentiation, we did not find evidence of fixed polymorphisms; this may indicate that the three populations also

TABLE 3. Comparison of mean genetic divergence between populations (D_{xy}) and mean genetic divergence (π) within each population

Gene and origin[a]	D_{xy}	π
dnaB		
Pue-Gto	0.072	0.056–0.052
Pue-Spain	0.068	0.056–0.064
Gto-Spain	0.073	0.052–0.064
glyA		
Pue-Gto	0.097	0.080–0.066
Pue-Spain	0.100	0.080–0.077
Gto-Spain	0.089	0.066–0.077
IS*Rel2*		
Pue-Gto	0.003	0.002–0.002
Pue-Spain	0.002	0.002–0.001
Gto-Spain	0.002	0.002–0.001
IS*Rel4*		
Pue-Gto	0.002	0.001–0.003
Pue-Spain	0.002	0.001–0.002
Gto-Spain	0.002	0.003–0.002
nodC		
Pue-Gto	0.003	0.001–0.005
Pue-Spain	0.026	0.001–0.008
Gto-Spain	0.025	0.005–0.008

[a] Pue, Puebla; Gto, Guanajuato.

TABLE 4. Genetic differentiation and gene flow tests in *R. etli* IS elements, plasmid genes, and chromosomal genes

Gene and origin[a]	K_{st}[b]	F_{st}	N_m
dnaB			
Pue-Gto	0.146***	0.252	1.48
Pue-Spain	0.064***	0.118	3.72
Gto-Spain	0.118***	0.208	1.90
glyA			
Pue-Gto	0.144***	0.249	1.51
Pue-Spain	0.142***	0.246	1.54
Gto-Spain	0.129***	0.226	1.71
IS*Rel2*			
Pue-Gto	0.007	0.025	19.75
Pue-Spain	0.652**	0.845	0.09
Gto-Spain	0.607**	0.750	0.17
IS*Rel4*			
Pue-Gto	0.056	0.217	1.80
Pue-Spain	0.310**	0.634	0.29
Gto-Spain	0.156**	0.270	1.35
nodC			
Pue-Gto	0.184*	0.374	0.84
Pue-Spain	0.118*	0.258	1.44
Gto-Spain	0.225**	0.357	0.90

[a] Pue, Puebla; Gto, Guanajuato.
[b] The asterisks represent the probabilities obtained by a permutation test with 1,000 replicates (*, $0.01 < P < 0.05$; **, $0.001 < P < 0.01$; ***, $P < 0.001$).

diverged recently. Given that the pSym sequences had low π_s values, their D_{xy} and π values were not well differentiated and had several shared alleles, and the genetic-differentiation tests were not significant, it is not surprising that these sequences also lacked fixed polymorphisms.

Next, we analyzed whether there was some degree of gene flow across the three populations. If an F_{st} value above 0.25 and an N_m value of >1 were taken to represent a significant between-population gene flow (37), most of the analyzed genes could be considered to have evidence of gene flow. The pSym sequences often had higher F_{st} values than the chromosomal genes (Table 4). A comparison between the Mexican and Spanish populations also yielded values indicative of gene flow. Given the geographical distance between these populations, it seems conceivable that these values could reflect a recent divergence of the three populations and a recent origin for the pSym genes.

There is published evidence indicating that genetic exchange occurred within *R. etli* populations from a single field; this process involved both plasmid and chromosomal loci (36). Accordingly, we hypothesized that the numbers of recombination events would be different between the chromosomal and pSym genes if their divergence times were clearly distinct. To assess this, we used the RDP3 software package to implement eight different recombination tests to search for recombination events across the three populations (see Table S4 in the supplemental material). The chromosomal genes, *dnaB* and *glyA*, showed evidence of four and two intragenic recombination events involving 10 and 5 strains, respectively. In contrast, only one pSym sequence (IS*Rel4*) showed a recombination event, even though the low sequence divergence made it harder to detect recombination. Interestingly, some of the same recombination events were found in strains of both *R. etli* and *R. gallicum*; moreover, four recombination events found between populations could explain some intermingled strains in the phylogeny of the three populations (Fig. 1; see Table S4 in the supplemental material).

Next, we used a phylogenetic-network approach, consisting of split decomposition and neighbor net analyses, to test whether the inconsistencies in the phylogenetic reconstructions could be due to recombination events. The split decomposition analysis of the two chromosomal genes showed a partition in the data between sequences from the *R. etli* and *R. gallicum* species. In the case of *glyA*, we found some splits within the *R. etli* strains; these could come from the detected recombination events. IS*Rel4* was the only pSym sequence for which we obtained a split that clearly represented the unique recombination event we detected in the *R. etli* and *R. gallicum* strains. The neighbor net analysis, which is more sensitive to the conflicting sequence data that may represent recombination events, showed that the two chromosomal genes had several splits, some of which were related to the detected recombination events (see Fig. S1 in the supplemental material). IS*Rel4* yielded the same split we had found in the split decomposition analysis. The other two pSym sequences failed to reveal splits with either analysis; this was consistent with the lack of any recombination events detected by our RDP3 analysis of these sequences. The relative lack of splits and recombination events in the pSym sequences supports our hypothesis that these genes (and probably the symbiotic plasmid itself) originated relatively recently in the three populations.

Another possibility that could explain the high DNA con-

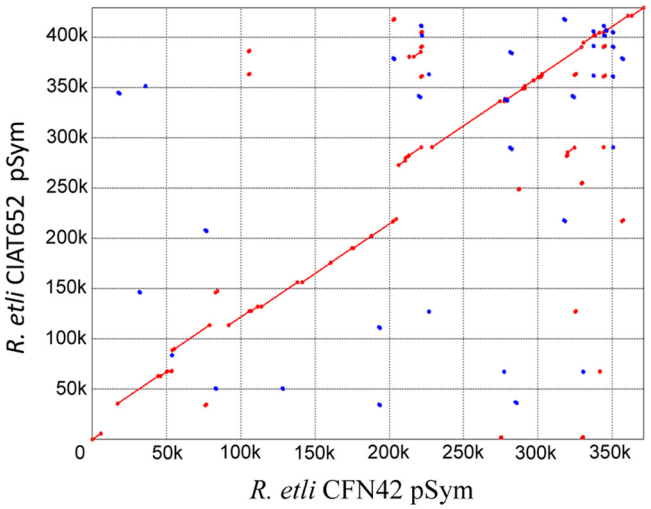

FIG. 3. Dot plot graphic of pSym from *R. etli* CFN42 and CIAT652. The diagonal red lines represent DNA regions that were aligned between the two replicons. The blue and red points represent small DNA regions, in many cases repeated sequences, that were aligned in different regions of each replicon.

servation of the pSym sequences and the conserved genomic contexts of IS*Rel2* and IS*Rel4* is that selection pressures may inhibit variation at these sequences. However, Tajima's *D* neutrality tests showed that most of the observed nucleotide substitution patterns in the chromosomal and pSym sequences from the three populations were under a neutral equilibrium model (i.e., all statistical results were nonsignificant) (Table 2; see Table S2 in the supplemental material). The neutral equilibrium model was rejected only for *nodC*. However, this could be due to the effect of a decrease in population size and/or balancing selection.

IS elements in *R. etli* symbiotic plasmids. We recently sequenced another *R. etli* strain, the Costa Rican strain CIAT652, which has one chromosome, three plasmids, and 22 IS elements distributed mainly in the chromosome and the symbiotic plasmid (15). Comparative analysis of the CFN42 and CIAT652 symbiotic plasmids showed that their sequences were highly conserved, with a nucleotide identity value of 99% (Fig. 3). In contrast, the identities of the chromosome and other plasmid sequences ranged from 90 to 95%. The two pSyms differed mainly in the presence/absence of certain genomic regions, the diversity of IS families, and the presence/absence of the IS elements in a given genomic context. This indicates that the major evolutionary events modifying these two symbiotic plasmids were the gains/losses of genomic regions and the losses/translocations of IS elements. In contrast, the chromosomes and other plasmids did not contain IS in the same genomic contexts. Comparison of the IS elements in the chromosomes and plasmids indicated that the symbiotic plasmids may have originated externally and harbored IS that thereafter transposed to the chromosomes. Both CFN42 and CIAT652 harbored IS elements that were 100% identical in sequence between the chromosome and pSym. If these plasmids have a recent origin (as suggested by our population genetic analyses), this would seem to indicate that the IS had recently moved to the chromosome in both strains, in the manner of an infection-expansion-extinction cycle (42).

DISCUSSION

In order to examine the evolutionary dynamics of IS elements in relation to the evolutionary history of the chromosomal and symbiotic plasmid genes, we compared the distributions and genetic diversity of these sequences in three *R. etli* populations. First, we used the concatenated alignment of the chromosomal *glyA* and *dnaB* genes to perform a phylogenetic reconstruction and found that the three tested *R. etli* populations were almost completely differentiated according to their geographic origins. Moreover, they were clearly differentiated from the *R. gallicum* strains, which formed a fourth clade. The close phylogenetic relationship between the Michoacan-Guanajuato and Spanish populations was highlighted by the individual phylogenies of *dnaB* and *glyA*, especially the one based on *glyA*, where several strains appeared to be intermingled in both populations. This pattern might be explained by recombination events, which were assessed using the RDP3 software (see Table S4 in the supplemental material). However, another possibility is that the intermingled strains from Guanajuato could be closely related to migrant strains isolated from the Spanish population. The phylogeny made with the concatenated alignment of *dnaB* and *glyA* was used as the reference for mapping the IS distributions.

Comparison of the IS presence/absence profiles of the *R. etli* populations to that of the CFN42 reference strain showed that the most conserved IS were on the symbiotic plasmid. Because the majority of the IS elements did not maintain their genomic contexts, the presence/absence profiles were inconsistent with the housekeeping gene phylogeny; some strains that were closely related in the phylogeny had very different IS profiles. For example, most of the Puebla strains had IS profiles more similar to that of the model strain, CFN42, which belonged to the Guanajuato-Michoacan clade. These data demonstrate that the distribution of IS elements between and within populations is strain specific, in concordance with the HGT and transpositional dynamics of IS (4).

We analyzed two of the IS elements, IS*Re12* and IS*Rel4* (both from pSym), in detail, as they were the most conserved, in terms of their genomic contexts, across the three *R. etli* populations. This pattern may indicate that some selective advantage is conferred by the presence of these IS elements, or it could be a consequence of the apparently recent origin of the pSym plasmid. Other studies on the evolutionary dynamics of IS elements have offered different explanations for the maintenance of such elements in populations or strains of the same species. For example, a report on IS in strains of *Helicobacter pylori* from different geographical origins suggested that IS are ancient components of the *H. pylori* gene pool and that they evolve at approximately the same rate as normal chromosomal genes (18). A paper on IS within a single population of the hyperthermophilic archaeon *Pyrococcus* suggested that the high frequencies of some IS were due to genetic drift (8). Neither of these explanations seems to be applicable to the IS in *R. etli* examined here. Instead, we think that the IS have undergone rapid turnover in the population and that the con-

textual and sequence conservation of IS*Rel2* and IS*Rel4* should be viewed within the evolutionary history of pSym.

Due to the transpositional nature of IS elements, they are not expected to be found at the same genomic site in populations of different geographical origin. However, here, we observed such conservation for IS*Rel2* and IS*Rel4*. One explanation for our observation is that that the IS might have been frequently found in the same genomic context as a consequence of the small population size of the *R. etli* collections analyzed in this work. The probability of finding several strains with the same IS in the same location is lower for large populations than for small populations (38). In our case, however, this may not be the best explanation, given the characteristics of the three populations (see Materials and Methods) and the geographical distances separating the collection sites. Another possibility is that these IS elements were recently acquired by the studied *R. etli* strains. Comparisons among these IS elements, *nodC*, and the chromosomal genes suggested that the pSym sequences recently entered the gene pools of the three *R. etli* populations. Population genetic analyses of the pSym sequences showed that they had the following features: (i) low π values, (ii) D_{xy} and π values that imply no clear differentiation, (iii) several different shared alleles, and (iv) nonsignificant results from their genetic differentiation tests. In contrast, opposing results were obtained for the chromosomal genes, *glyA* and *dnaB*. Thus, the differences noted in our population genetic analyses suggest that the chromosome and pSym have different evolutionary histories in the studied *R. etli* populations. These findings, along with the high sequence conservation found in the pSym sequences of CFN42 and CIAT652, support the proposition that pSym could have a recent origin in *R. etli* (14). However, the pSym sequences from CFN42 and CIAT652 have clear differences in terms of the presence/absence of certain genomic regions (some as large as 50 kb) and IS elements; notably, the gains and losses of IS elements appear to be the main events differentiating these two pSyms.

A prior analysis of the complete genomes of *R. etli* CFN42 and *R. etli* CIAT652 demonstrated that the IS do not disrupt genes (with one exception) and that these strains appear to harbor the same pSym plasmid (15). Horizontal transfer of this plasmid to *R. etli* CFN42 and *R. etli* CIAT652 could explain the asymmetric distribution of IS, which were found only in these plasmids and in the chromosome. Some IS elements probably moved from the plasmids to the chromosome, but they do not appear to have moved to other plasmids. The asymmetrical distribution of IS in the chromosome and plasmids could be due to the sizes of the different replicons in the *R. etli* CFN42 genome (41). In the chromosome of CFN42, there are six probable recent transposition events of IS elements (100% identity between the chromosomal IS and another IS in the pSym or conjugative plasmid). This implies a transposition rate of 1 IS per 730 kb, which is greater than the size of any other replicon of CFN42 (14), so the asymmetrical distribution could be the product of the chromosome size.

The complete genomes of the two strains of *R. etli*, CFN42 and CIAT652, revealed the presence of different IS family members in different copy numbers. It has been proposed that when new IS elements enter the genome, they actively transpose to different genomic positions (42). This hypothesis may help explain the behavior of some of the IS in *R. etli* CFN42. For instance, members of the IS*66* family, which is a very common IS family in rhizobial species (27), were found in several copies in the pSym plasmids of CFN42 and CIAT652, and IS*66* copies with 100% nucleotide sequence identity could also be found in the chromosome. The most plausible explanation for this finding is that the IS originally entered the genome via plasmids and were then transposed to the chromosome.

The present work provides useful new insights into differences of the distribution and genomic context maintenance of IS elements in natural populations of *R. etli*. Although the pSym in these populations seems to be of relatively recent origin, the harbored IS appear to be active elements that may participate in the genomic plasticity of these organisms. Our comparisons of the two genomes of *R. etli* and across the natural populations provide further evidence that different IS can rapidly expand when they arrive in a new host genome, as recently proposed by Wagner (43).

ACKNOWLEDGMENTS

We thank Santiago Castillo for critical reading of the mnuscript. We thank Susana Brom (Universidad Nacional Autónoma de México) for providing the Spanish strains, Antonio Cruz for providing those from Puebla and Michoacan-Guanajuato, and J. Espíritu for technical and computational assistance.

This work was supported by grants from CONACyT (grant U4633), PAPIIT-UNAM (grant IN223005), and the Natural Environment Research Council. L.L. was a recipient of a CONACyT scholarship.

L.L. was responsible for data analysis and manuscript preparation. L.L. and V.G. were responsible for the experimental design. I.H.-G. was responsible for bioinformatics analysis. P.B. and R.I.S. were responsible for DNA sequencing. L.L., V.S., J.P.W.Y., G.D., and V.G. were responsible for discussion of the data.

REFERENCES

1. **Altschul, S., T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Millar, and D. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
2. **Amadou, C., G. Pascal, S. Mangenot, M. Glew, C. Bontemps, D. Capela, S. Carrere, S. Cruveiller, C. Dossat, A. Lajus, M. Marchetti, V. Poinsot, B. Rouy, Z. Servin, M. Saad, C. Schenowitz, V. Barbe, J. Batut, C. Medigue, and C. Masson-Boivin.** 2008. Genome sequence of the beta-rhizobium Cupriavidus taiwanensis and comparative genomics of rhizobia. Genome Res. **18:**1472–1483.
3. **Bailly, X., I. Olivieri, S. De Mita, J. C. Cleyet-Marel, and G. Bena.** 2006. Recombination and selection shape the molecular diversity pattern of nitrogen-fixing Sinorhizobium sp. associated to Medicago. Mol. Ecol. **15:**2719–2734.
4. **Chandler, M., and J. Mahillon.** 2002. Insertion sequences revisited, p. 305–366. *In* L. Craig et al. (ed.), Mobile DNA II. ASM Press, Washington, DC.
5. **Crossman, L. C., S. Castillo-Ramírez, C. McAnnula, L. Lozano, G. S. Vernikos, J. L. Acosta, Z. F. Ghazoui, I. Hernández-González, G. Meakin, A. W. Walker, M. F. Hynes, J. P. Young, J. A. Downie, D. Romero, A. W. Johnston, G. Dávila, J. Parkhill, and V. González.** 2008. A common genomic framework for a diverse assembly of plasmids in the symbiotic nitrogen fixing bacteria. PloS ONE **3:**e2567.
6. **Doolittle, W. F., and C. Sapienza.** 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature **284:**601–603.
7. **Edgar, R.** 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform. **5:**113.
8. **Escobar-Páramo, P., S. Ghosh, and J. DiRuggiero.** 2005. Evidence for genetic drift in the diversification of a geographically isolated population of the hyperthermophilic archaeon *Pyrococcus*. Mol. Biol. Evol. **22:**2297–2303.
9. **Filée, J., P. Siguier, and M. Chandler.** 2007. Insertion sequence diversity in Archea. Microbiol. Mol. Biol. Rev. **71:**121–157.
10. **Flores, M., L. Morales, M. Avila, P. Bustos, V. González, D. Garcia, Y. Mora, X. Guo, J. Collado-Vides, D. Piñero, G. Davila, J. Mora, and R. Palacios.** 2005. Diversification of DNA sequences in the symbiotic genome of Rhizobium etli. J. Bacteriol. **187:**7185–7192.
11. **Freiberg, C., R. Fellay, A. Bairoch, W. J. Broughton, A. Rosenthal, and X. Perret.** 1997. Molecular basis of symbiosis between Rhizobium and legumes. Nature **387:**394–401.

12. **Galibert, F., T. M. Finan, S. R. Long, A. Puhler, P. Abola, F. Ampe, F. Barloy-Hubler, M. J. Barnett, A. Becker, P. Boistard, G. Bothe, M. Boutry, L. Bowser, J. Buhrmester, E. Cadieu, D. Capela, P. Chain, A. Cowie, R. W. Davis, S. Dreano, N. A. Federspiel, R. F. Fisher, S. Gloux, T. Godrie, A. Goffeau, B. Golding, J. Gouzy, M. Gurjal, I. Hernandez-Lucas, A. Hong, L. Huizar, R. W. Hyman, T. Jones, D. Kahn, M. L. Kahn, S. Kalman, D. H. Keating, E. Kiss, C. Komp, V. Lelaure, D. Masuy, C. Palm, M. C. Peck, T. M. Pohl, D. Portetelle, B. Purnelle, U. Ramsperger, R. Surzycki, P. Thebault, M. Vandenbol, F. J. Vorholter, S. Weidner, D. H. Wells, K. Wong, K. C. Yeh, and J. Batut.** 2001. The composite genome of the legume symbiont Sinorhizobium meliloti. Science **293:**668–672.
13. **Gasca, J.** 2004. Genética de poblaciones de *Rhizobium etli* asociado a *Phaseolus vulgaris* en dos localidades del bajío mexicano. B.S. thesis. Istituto de Ecología, Universidad Nacional Autónoma de México, Mexico DF, Mexico.
14. **González, V., R. Santamaría, P. Bustos, I. Hernandez-González, A. Medrano-Soto, G. Moreno-Hagelsieb, S. Janga, M. Ramirez, V. Jiménez-Jacinto, J. Collado-Vides, and G. Davila.** 2006. The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. Proc. Natl. Acad. Sci. U. S. A. **103:**3834–3839.
15. **González, V., J. L. Acosta, R. Santamaría, P. Bustos, J. L. Fernandez, I. Hernandez-González, R. Diza, M. Flores, R. Palacios, J. Mora, and G. Davila.** 2010. Conserved symbiotic plasmid DNA sequences in the multireplicon pangenomic structure of Rhizobium etli. Appl. Environ. Microbiol. **76:**1604–1614.
16. **Guindon, S., and O. Gascuel.** 2002. Efficient biased estimation of evolutionary distances when substitution rates vary across sites. Mol. Biol. Evol. **19:**534–543.
17. **Huson, D. H., and D. Bryant.** 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. **23:**254–267.
18. **Kalia, A., A. K. Mukhopadhyay, G. Dailide, Y. Ito, T. Azuma, B. C. Wong, and D. E. Berg.** 2004. Evolutionary dynamics of insertion sequences in Helicobacter pylori. J. Bacteriol. **186:**7508–7520.
19. **Kaneko, T., Y. Nakamura, S. Sato, E. Asamizu, T. Kato, S. Sasamoto, A. Watanabe, K. Idesawa, A. Ishikawa, K. Kawashima, T. Kimura, Y. Kishida, C. Kiyokawa, M. Kohara, M. Matsumoto, A. Matsuno, Y. Mochizuki, S. Nakayama, N. Nakazaki, S. Shimpo, M. Sugimoto, C. Takeuchi, M. Yamada, and S. Tabata.** 2000. Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti. DNA Res. **7:**331–338.
20. **Kaneko, T., Y. Nakamura, S. Sato, K. Minamisawa, T. Uchiumi, S. Sasamoto, A. Watanabe, K. Idesawa, M. Iriguchi, K. Kawashima, M. Kohara, M. Matsumoto, S. Shimpo, H. Tsuruoka, T. Wada, M. Yamada, and S. Tabata.** 2002. Complete genomic sequence of nitrogen-fixing symbiotic bacterium Bradyrhizobium japonicum USDA110. DNA Res. **9:**189–197.
21. **Kidwell, M. G., and D. R. Lisch.** 2002. Transposable elements as sources of genomic variation, p. 305–366. *In* L. Craig et al. (ed.), Mobile DNA II. ASM Press, Washington, DC.
22. **Korber, B.** 2000. HIV signature and sequence variation analysis, p. 55–72. *In* A. G. Rodrigo and G. H. Learn (ed.), Computational analysis of HIV molecular sequences. Kluwer Academic Publishers, Dordrecht, Netherlands.
23. **Kurtz, S., A. Phullippy, A. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. Salzberg.** 2004. Versatile and open software for comparing large genomes. Genome Biol. **5:**R12.
24. **Kwak, M., and P. Gepts.** 2009. Structure of genetic diversity in the two major gene pools of common bean (Phaseolus vulgaris L., Fabaceae). Theor. Appl. Genet. **118:**979–992.
25. **Lenski, R., C. Winkwirth, and M. Riley.** 2003. Rates of DNA sequence evolution in experimental populations of *Escherichia coli* during 20,000 generations, J. Mol. Evol. **56:**498–508.
26. **Librado, P., and J. Rozas.** 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics **25:**1451–1452.
27. **Mahillon, J., C. Léonard, and M. Chandler.** 1999. IS elements as constituents of bacterial genomes. Res. Microbiol. **150:**675–687.
28. **Martin, D. P., C. Williamson, and D. Posada.** 2005. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics **21:**260–262.
29. **Papadopoulos, D., D. Schneider, J. Meier-Eis, W. Arber, R. E. Lenski, M. Blot, et al.** 1999. Genomic evolution during a 10,000-generation experiement with bacteria. Proc. Natl. Acad. Sci. U. S. A. **96:**3807–3812.
30. **Pérez-Mendoza, D., A. Dominguez-Ferreras, S. Munoz, M. J. Soto, J. Olivares, S. Brom, L. Girard, J. A. Herrera-Cervera, and J. Sanjuan.** 2004. Identification of functional mob regions in Rhizobium etli: evidence for self-transmissibility of the symbiotic plasmid pRetCFN42d. J. Bacteriol. **186:**5753–5761.
31. **Posada, D., and K. A. Crandall.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14:**817–818.
32. **Ramsay, J. P., J. T. Sullivan, G. S. Stuart, I. L. Lamont, and C. W. Ronson.** 2006. Excision and transfer of the Mesorhizobium loti R7A symbiosis island requires an integrase IntS, a novel recombination directionality factor RdfS, and a putative relaxase RlxS. Mol. Microbiol. **62:**723–734.
33. **Rodríguez-Navarro, D., A. Buendia, M. Camacho, M. Lucas, and C. Santamaria.** 2000. Characterization of *Rhizobium* spp. Bean isolates from southwest Spain. Soil Biol. Biochem. **32:**1601–1613.
34. **Schneider, D., and R. Lenski.** 2004. Dynamics of insertion sequence elements during experimental evolution of bacteria. Res. Microbiol. **155:**319–327.
35. **Siguier, P., J. Filée, and M. Chandler.** 2006. Insertion sequences in prokaryotic genomes. Curr. Opin. Microbiol. **9:**1–6.
36. **Silva, C., P. Vinuesa, L. Eguiarte, E. Martínez-Romero, and V. Souza.** 2003. *Rhizobium etli* and *Rhizobium gallicum* nodulate common bean (*Phaseolus vulgaris*) in a traditionally managed milpa plot in Mexico: population genetics and biogeographic implications. Appl. Environ. Microbiol. **69:**884–893.
37. **Silva, C., P. Vinuesa, L. Eguiarte, V. Souza, and E. Martínez-Romero.** 2005. Evolutionary genetics and biogeographic structure of Rhizobium gallicum sensu lato, a widely distributed bacterial symbiont of diverse legumes. Mol. Ecol. **14:**4033–4050.
38. **Slatkin, M.** 1985. Genetic differentiation of transposable elements under mutation and unbiased gene conversion. Genetics **110:**145–158.
39. **Sullivan, J. T., and C. W. Ronson.** 1998. Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. Proc. Natl. Acad. Sci. U. S. A. **95:**5145–5149.
40. **Sullivan, J. T., J. R. Trzebiatowski, R. W. Cruickshank, J. Gouzy, S. D. Brown, R. M. Elliot, D. J. Fleetwood, N. G. McCallum, U. Rossbach, G. S. Stuart, J. E. Weaver, R. J. Webby, F. J. De Bruijn, and C. W. Ronson.** 2002. Comparative sequence analysis of the symbiosis island of Mesorhizobium loti strain R7A. J. Bacteriol. **184:**3086–3095.
41. **Touchon, M., and E. Rocha.** 2007. Causes of insertion sequences abundance in prokaryotic genomes. Mol. Biol. Evol. **4:**969–981.
42. **Wagner, A.** 2006. Periodic extinctions of transposable elements in bacterial lineales: evidence from intragenomic variation in multiple genomes. Mol. Biol. Evol. **23:**723–733.
43. **Young, J. P., L. C. Crossman, A. W. Johnston, N. R. Thomson, Z. F. Ghazoui, K. H. Hull, M. Wexler, A. R. Curson, J. D. Todd, P. S. Poole, T. H. Mauchline, A. K. East, M. A. Quail, C. Churcher, C. Arrowsmith, I. Cherevach, T. Chillingworth, K. Clarke, A. Cronin, P. Davis, A. Fraser, Z. Hance, H. Hauser, K. Jagels, S. Moule, K. Mungall, H. Norbertczak, E. Rabbinowitsch, M. Sanders, M. Simmonds, S. Whitehead, and J. Parkhill.** 2006. The genome of Rhizobium leguminosarum has recognizable core and accessory components. Genome Biol. **7:**R34.
44. **Zeigler, D. R.** 2003. Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int. J. Syst. Evol. Microbiol. **53:**1893–1900.

RESULTADOS

Análisis Filogenético de las Poblaciones de *Rhizobium*

En el presente proyecto se trabajó con tres colecciones de cepas obtenidas en tres distintos puntos geográficos, dos en México (Puebla y Guanajuato-Michoacán) y una de España (ver Metodología). Las tres colecciones consisten principalmente de cepas de *Rhizobium etli*, aunque en algunos casos se colectaron cepas de otras especies (ver tabla 2). El primer objetivo fue determinar si las cepas de las tres colecciones se hallan estructuradas geográficamente, es decir, si las tres poblaciones corresponden a tres grupos poblacionales bien diferenciados. Para ello los genes cromosomales, *glyA* y *dnaB* se secuenciaron en todas las cepas y se realizaron filogenias para cada uno por separado. En ambas filogenias (figuras 3 y 4) se observa de forma general la separación de las cepas en tres poblaciones, aun cuando existen algunas cepas que aparecen situadas en otra población a la que no pertenecen. Estas poblaciones corresponden a las tres regiones geográficas de su aislamiento. Así mismo, se observa un cuarto grupo formado por aquellas cepas de las tres colecciones que representan especies distintas de *R. etli*. En la filogenia del gen *glyA* (figura 3) se observan un mayor número de cepas que aparecen más cercanas a cepas de otra colección. Por otro lado, las relaciones entre las tres poblaciones son distintas entre las dos filogenias, mientras que en la filogenia del gen *dnaB* (figura 4), las poblaciones de Puebla y Guanajuato-Michoacán están más cercanamente relacionadas, en la filogenia del gen *glyA*, España y Puebla se hallan más cercanas entre si.

Para obtener un mejor panorama de las relaciones filogenéticas entre las tres poblaciones, se concatenaron los alineamientos múltiples de los dos genes y se realizó otra filogenia (ver figura 1 del Artículo). Esta filogenia muestra que las poblaciones de España y Guanajuato-Michoacán están más relacionadas, y la población de Puebla presenta un mayor número de cepas que no corresponden a dicha población.

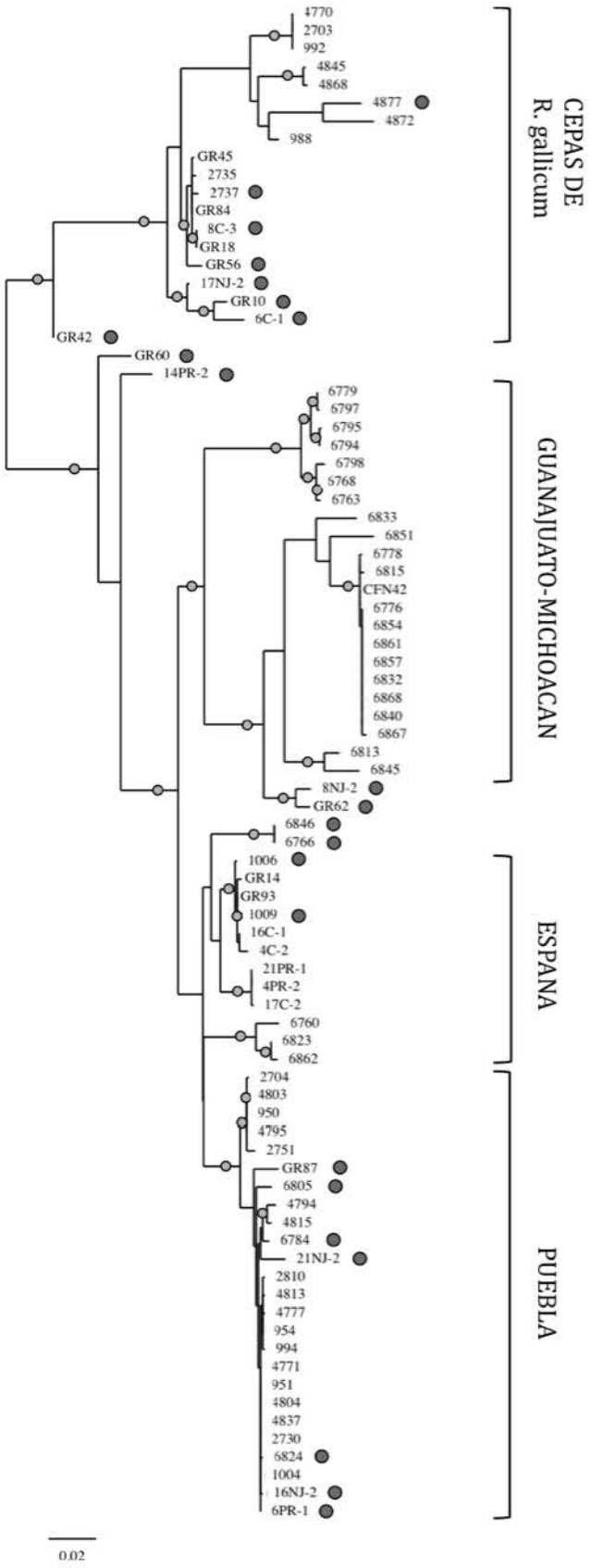

Figura 3. Filogenia realizada con secuencias del gen *glyA* (Método: Máxima Verosimilitud; Modelo de evolución: GTR). Los círculos internos (gris claro) representan nodos con valores de bootstrap >70. Los círculos externos muestran las cepas se situaron fuera de su población.

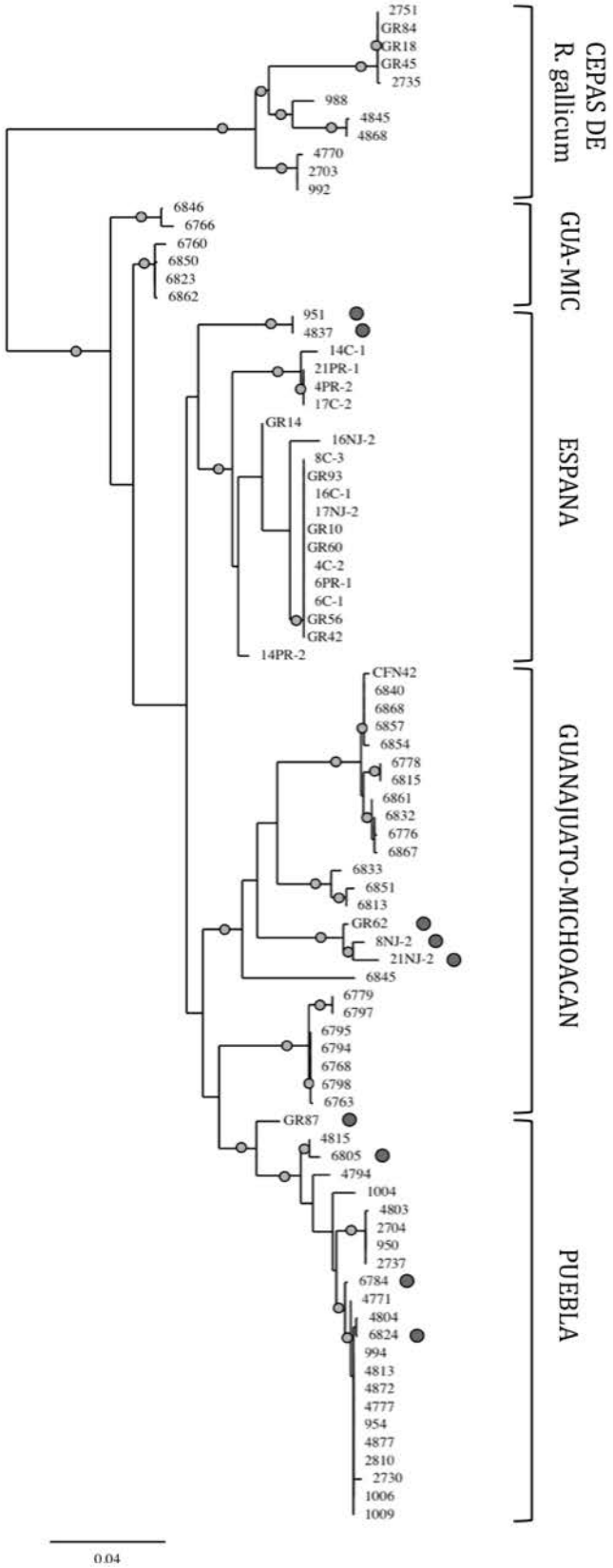

Figura 4. Filogenia realizada con secuencias del gen *dnaB* (Método: Máxima Verosimilitud; Modelo de evolución: GTR). Los círculos internos (gris claro) representan nodos con valores de bootstrap >70. Los círculos externos muestran las cepas se situaron fuera de su población.

Debido a la diferencia entre las filogenias de *glyA* y *dnaB*, se realizó un prueba de congruencia entre los dos árboles filogéneticos. La idea es conocer si las dos filogenias son topológicamente congruentes, aun cuando existen claras diferencias entre ellos como se mencionó arriba. Para este propósito se obtuvo el índice I_{cong} (M. De Vienne, D. et al. 2007) que permite comparar dos árboles sin importar el número de hojas (en este caso especies o cepas) que contenga cada uno. En el caso de las filogenias de *glyA* y *dnaB*, sólo se diferencian por una cepa de *R. etli* (*glyA* con 84 cepas y *dnaB* con 85). El índice I_{cong} se basa en encontrar el árbol filogénetico que contenga el mayor número de hojas y que sea compatible con las dos filogenias. El resultado obtenido, I_{cong}=1.79, P-value=5.94e-08, indica que las dos filogenias son congruentes, pero cabe resaltar que el árbol con el mayor número de hojas compatibles con ambas filogenias fue de 24 hojas. Lo anterior se debe a que las diferentes poblaciones se relacionaron de forma distinta en cada árbol (como se mencionó más arriba).

Perfil de Presencia/Ausencia de SIs en las Poblaciones de *Rhizobium*

El siguiente paso para conocer la dinámica de las SIs en las tres poblaciones consistió en analizar si las SIs de la cepa CFN42 se encuentran conservadas en el mismo contexto genómico en todas las cepas. Se obtuvo un perfil de presencia/ausencia de SIs (ver figura 2 del Artículo) el cual muestra las cepas tienen conservadas las SIs de CFN42 en el mismo contexto genómico. Aun cuando los perfiles de SIs resultan ser cepa específico, existen semejanzas dentro de cada población, sobre todo entre cepas de la población de Puebla y entre las cepas de la población Guanajuato-Michoacán. Lo anterior se puede observar al realizar un análisis de similitud entre perfiles de las distintas cepas (ver Apéndice 1). El perfil de SIs sirvió finalmente para escoger las Si mas representadas entre las tres poblaciones, que resultaron ser la ISRel2 y la ISRel4. Cabe mencionar que el perfil de presencia/ausencia muestra otra SI altamente representada entre las tres poblaciones, principalmente en España y Guanajuato-Michoacán. Esta SI de la familia IS5 no fue considerada ya que en la mayoría de los casos en que fue secuenciada resultó ser un gen hipotético y no la SI esperada. Este gen hipotético se halla a alrededor de 100 Kbs de distancia junto a otra SI de la misma familia, lo cual puede indicar que hubo un evento de recombinación anterior a la dispersión de esta región en las cepas.

Origen Reciente de las SIs y del Plásmido Simbiótico

Generalmente las SIs no se hallan por mucho tiempo en el mismo sitio de un genoma sin acumular cambios en su secuencia a una tasa más elevada que genes esenciales, a menos que exista alguna presión de selección. Al tener ISRel2 e ISRel4 una mayor conservación de secuencia y una diversidad nucleotídica (π) mucho menor que *glyA* y *dnaB,* y debido al hecho de que *nodC* también presenta una alta conservación de secuencia (ver tabla 2 del Artículo), resulta necesario aplicar otras pruebas de genética de poblaciones que permitan averiguar si estos tres genes del plásmido simbiótico realmente tienen un origen reciente en las tres poblaciones.

Para probar lo anterior, se analizó la diversidad nucleotídica promedio en sitios sinónimos (πs) y sitios no sinónimos (πns). En particular, πs resulta útil debido a que es un parámetro que refleja que tan antiguo es un alelo en el pool genético de una población; valores altos de πs indican que el gen en cuestión ha estado presente el suficiente tiempo en una población como para acumular un elevado número de cambios sinónimos diferentes; cuando se obtienen valores bajos de πs, se considera que el gen es muy reciente en la población, razón por la cual no ha acumulado cambios sinónimos. Los valores de πs para los tres genes del plásmido simbiótico son mucho menores que los de los genes cromosomales, en especial aquellos de las dos SIs (ver tabla 2 del Artículo). En el caso particular de *nodC*, el valor de πs es mayor que el de las dos SIs y por otro lado su relación con su valor de πns es semejante al de los genes cromosomales, es decir, el valor de πs es claramente mayor que el de πns. Lo anterior podría indicar que *nodC* tiene una dinámica si bien no similar a la de un gen que lleva mucho tiempo en el pool genético, si la de un gen que tiene una función importante para la bacteria, como puede ser el proceso simbiótico. Por otro lado, si comparamos los valores de πs y πns en cada población por separado para los cinco genes se observan los mismos resultados (Tabla 3).

Otra forma de determinar si los genes ISRel2, ISRel4 y *nodC* son más recientes es comparar la divergencia nucleotídica promedio entre dos poblaciones (Dxy) contra la divergencia nucleotídica promedio (π) de la población ‘x’ y de la

población ’y’. En otras palabras, si la divergencia nucleotídica promedio entre poblaciones, Dxy, es mayor que la divergencia nucleotídica promedio de cada una, π, significa que ha pasado suficiente tiempo para que ambas poblaciones se diferencien. Por el contrario, si los valores de Dxy y π de cada población son semejantes, significa que no ha pasado suficiente tiempo para hallar evidencias de diferenciación genética entre las dos poblaciones. En el Artículo (tabla 3) se observa que los valores de Dxy entre las poblaciones de los genes *glyA* y *dnaB* son mayores que los valores de π, es decir, cada gen por separado muestra que ha pasado suficiente tiempo para que haya evidencia de divergencia nucleotídica entre las dos poblaciones (ya sea, Pue-Gto, Pue-Esp o Gto-Esp). Lo contrario se observa para el caso de las dos SIs, no existe evidencia de una mayor diferenciación genética entre las poblaciones que de cada una por separado, esto apoya la idea de que estos genes tienen un origen reciente, dado que no tienen la misma divergencia que muestran los dos genes cromosomales. En el caso de *nodC*, no hay diferencia entre las poblaciones de México, Pue-Gto, pero cuando se comparan contra la población de España si hay evidencia de divergencia entre las poblaciones, esto se puede deber a que las condiciones a las que esta sujeto el gen *nodC* en la población de España son distintas, ya que proviene de una región donde no hay evidencia de cultivo de *Phaseolus vulgaris* (ver metodología). Aun así, habría que analizar más detenidamente el que en *nodC* se presente cierto grado de divergencia para con otras poblaciones de España y en el caso de las SIs no se observe divergencia entre las poblaciones.

Otro dato interesante es la presencia de haplotipos compartidos entre las tres poblaciones. Cuando ha pasado suficiente tiempo desde la divergencia entre dos poblaciones, es muy probable que dejen de compartir los mismos haplotipos que compartían antes de su divergencia. Al analizar los haplotipos para cada uno de los cinco genes, no se encontraron haplotipos compartidos para *glyA* o *dnaB* entre las tres poblaciones. Por el contrario, los tres genes del plásmido simbiótico presentan, por separado, haplotipos compartidos entre las tres poblaciones. Posteriormente, se analizó si existe evidencia de subdivisión geográfica entre las poblaciones con lo cual se busca conocer de forma significativa si las poblaciones son genéticamente diferentes. Para esto se uso la prueba de subdivisión geográfica de Hudson, Kst (Hudson, R., et al. 1992). El Artículo muestra los valores de Kst comparando las tres

poblaciones de cada gen. En lo que respecta a los genes cromosomales los valores de Kst resultaron ser altamente significativos lo cual indica que estos dos genes, *glyA* y *dnaB* muestran que existe diferenciación genética entre las tres poblaciones. Por el contrario los genes del plásmido simbiótico tienen valores menos significativos, en especial las comparaciones entre las poblaciones de Puebla y Michoacán-Guanajuato.

Posteriormente se analizó si existía evidencia de flujo génico entre las tres poblaciones, esperando hallar diferencias entre los genes cromosomales y simbióticos, de forma que se apoyara el origen reciente de las SIs y *nodC*. El índice F_{st} es una medida de la diversidad genética que se obtiene analizando las diferencias de las frecuencias alélicas entre poblaciones, en el cual se analiza la variabilidad genética dentro y entre dichas poblaciones. Si se considera que un valor de F_{st} mayor de 0.25 y un Nm (número de migrantes por generación) mayor a 1 son valores que indican que existe flujo génico entre las poblaciones, en el Artículo (tabla 4) se muestra que en general estos valores son un poco menores para las genes *glyA* y *dnaB* pero cercanos a 0.25 de F_{st} y 1 de Nm, así mismo se ve que para los genes del plásmido los valores de F_{st} tienden a ser altos, excepto dos casos: Pue-Gto de los genes ISRel2 e ISRel4. Debido a que las pruebas anteriores han mostrado evidencia de que los genes simbióticos parecen tener un origen reciente en el pool genético de las tres poblaciones, estos resultados de F_{st} se pueden deber al mismo fenómeno, es decir, al haber llegado recientemente estos genes, se pueden obtener valores altos de flujo génico.

En resumen, las dos SIs y el gen *nodC* muestran una dinámica distinta a la de *glyA* y *dnaB*, ya que tienen valores bajos de diversidad nucleotídica en sitios sinónimos (πs); sus valores de Dxy y π no están claramente diferenciados; comparten varios alelos y otras distintas pruebas de diferenciación genética aplicadas resultaron ser poco significativas. Debido a estas similitudes entre los tres genes simbióticos y a que a diferencia de *nodC*, las dos SIs son elementos en los cuales se esperaría que tuvieran mayores diferencias en los tests utilizados, se podría plantear la posibilidad de que regiones del plásmido simbiótico o el mismo plásmido en su totalidad es un replicón reciente en las tres poblaciones.

Tabla 3. Divergencia del ADN de los genes cromosomales y simbióticos entre las tres distintas poblaciones.

Gen	Θ	Diversidad nucleotidica π	Diversidad nucleotidica en sitios sinonimos π_S	Diversidad nucleotidica en sitios no sinonimos π_{NS}	Tajima D
dnaB	**0.047**	**0.066**	**0.237**	**0.009**	1.39
Puebla	0.040	0.056	0.191	0.011	1.48
Guanajuato	0.040	0.052	0.194	0.005	1.08
Spain	0.042	0.064	0.227	0.010	2.15*
glyA	**0.070**	**0.088**	**0.122**	**0.075**	0.84
Puebla	0.065	0.080	0.118	0.066	0.88
Guanajuato	0.057	0.066	0.091	0.057	0.57
Spain	0.070	0.077	0.098	0.061	0.07
nodC	**0.007**	**0.014**	**0.028**	**0.009**	2.72**
Puebla	0.001	0.001	0.002	0	-1.01
Guanajuato	0.007	0.005	0.012	0.002	-1.16
Spain	0.010	0.008	0.015	0.005	-0.87
ISRel2	**0.001**	**0.002**	**0.001**	**0.002**	0.72
Puebla	0.001	0.002	0.002	0.001	0.70
Guanajuato	0.001	0.002	0.001	0.002	0.29
Spain	0.001	0.001	0	0.001	0.52
ISRel4	**0.002**	**0.002**	**0.003**	**0.002**	-0.62
Puebla	0.001	0.001	0	0.001	-1.13
Guanajuato	0.003	0.003	0.005	0.002	-0.14
Spain	0.002	0.002	0.003	0.001	-0.65

Considerando que las diferencias halladas muestran que los tiempos de divergencia son distintos entre los genes cromosomales y simbióticos, el número de eventos de recombinación identificables en cada uno de ellos debe ser distinto. Por consiguiente se analizaron mediante distintos métodos los eventos de recombinación en cada gen (ver metodología). La tabla 4 muestra que *glyA* y *dnaB* han tenido por lo menos dos y cuatro eventos de recombinación intragénica identificables que involucran a 11 y 5 cepas respectivamente. En contraste, solamente ISRel4 tuvo un evento identificable y en ISRel2 y *nodC* ningún método fue capaz de registrar evento alguno. Cabe mencionar que algunos de los eventos hallados en *glyA* y *dnaB* podrían explicar el porque algunas cepa en la filogenia de estos dos genes (ver Artículo) se hallan en poblaciones que no les corresponden. Además, algunos de los eventos de recombinación intragénica fueron identificables entre cepas de *R. etli* y de *R. gallicum*.

Tabla 4. Eventos de recombinación presentes en los genes cromosomales y del plásmido simbiótico. Cada línea representa un evento de recombinación distinto. Las cepas subrayadas representan cepas que en la filogenia (figura 5) aparecen situadas en una población a la que no pertenecen.

Gen	Cepas Recombinantes	Cepas Parentales	Métodos	P-value
dnaB	4877, 4872, 988	4868 - 4795	3	0.564
	6851	6760 - 6778	4	0.509
	14PR-2, GR60	6PR-1 - 4872	3	0.634
	GR18, 17NJ-2, GR10, 6C-1, GR56	988 – 6PR-1	4	0.621
glyA	4837	4803 – GR42	2	0.479
	6815, 8NJ-2, 21NJ-2, GR62	GR14 - 4803	2	0.510
ISRel4	954	GR87 - 4795	4	0.549

Posteriormente, se hicieron reconstrucciones de redes filogenéticas para evaluar si algunos de los eventos de recombinación hallados (cepas recombinantes y parentales) se representaban en forma de retículas en la gráfica. La formación de retículas es producto de señales conflictivas entre las secuencias, las cuales pueden ser generadas por eventos de recombinación. La figura 5 muestra las reconstrucciones para *glyA* y *dnaB* mediante el método de Neighbor Net, en ambos casos se muestran subrayadas las cepas que tuvieron eventos de recombinación. Si bien la mayoría de estas cepas se localizan en regiones reticuladas que muestran el conflicto filogenético entre los sitios informativos que presentan las secuencias en su alineamiento múltiple, muchas cepas más generan estas señales filogenéticas incompatibles, aún cuando no pudieron ser identificadas como cepas que participaran en eventos de recombinación. Por otro lado, para el caso de ISRel4 (figura 6c) la red filogenética presenta una reticulación la cual concuerda con el único evento de recombinación. El análisis de recombinación muestra que existe una cepa recombinante, 954, la cual se sitúa justamente en la reticulación, mientras que las dos cepas parentales, 4795 y GR87, se hallan a los lados. Finalmente, los otros dos genes simbióticos, *nodC* e ISRel2, no produjeron ninguna reticulación en sus respectivas redes filogenéticas, esto se debe probablemente a que no ha pasado el tiempo suficiente para que sus secuencias adquieran modificaciones que representen señales incompatibles, como eventos de recombinación.

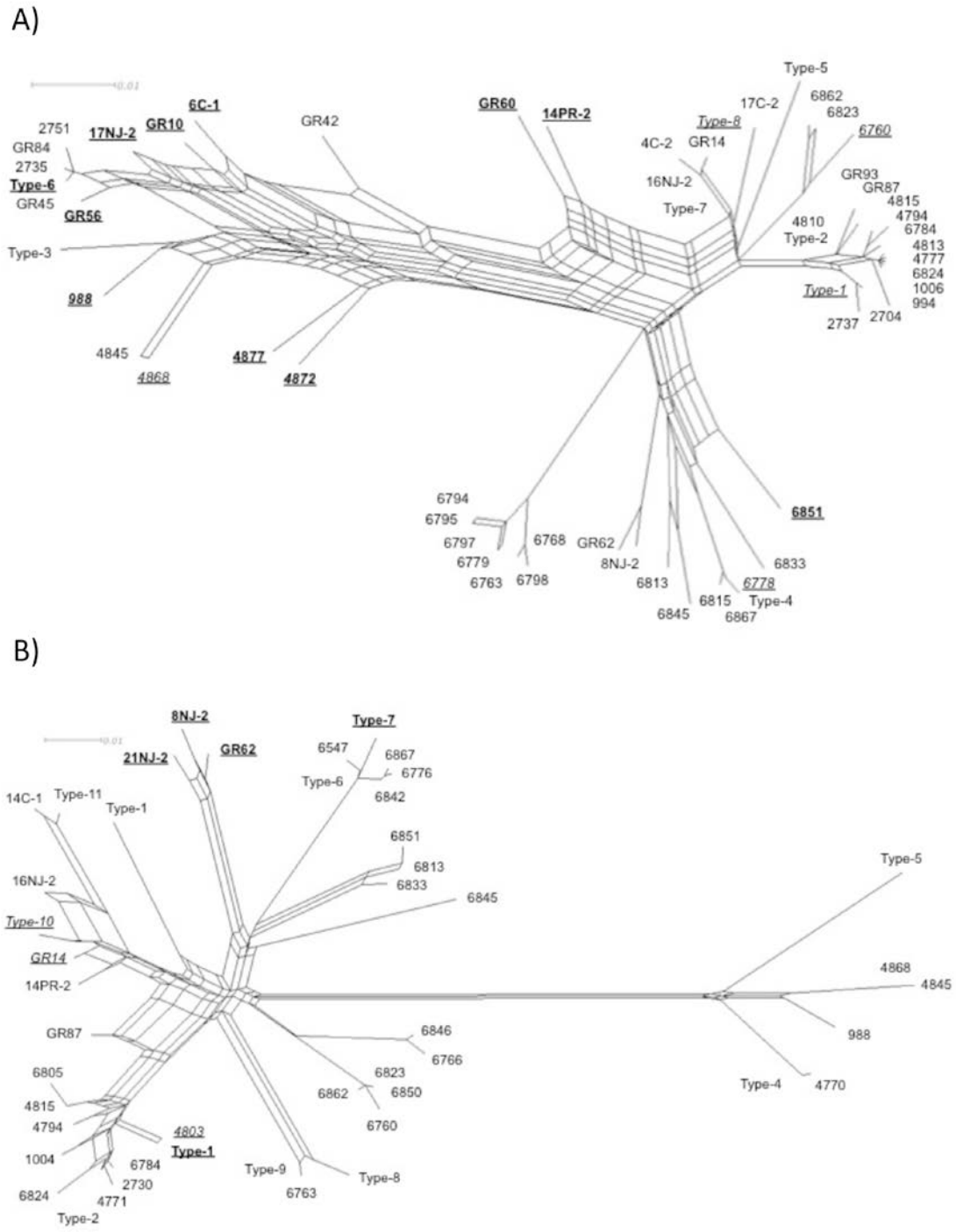

Figura 5. Reconstrucción de redes filogenéticas para los genes A) *dnaB* y B) *glyA*. Las cepas subrayadas representan aquellas identificadas como participes de eventos de recombinación mediante RDP3 (tabla 7), las cepas en negritas representan cepa recombinantes y aquellas en itálicas representan cepas parentales. Los grupos (denominados Type) representan varias cepas en el mismo nodo (Ver Apéndice).

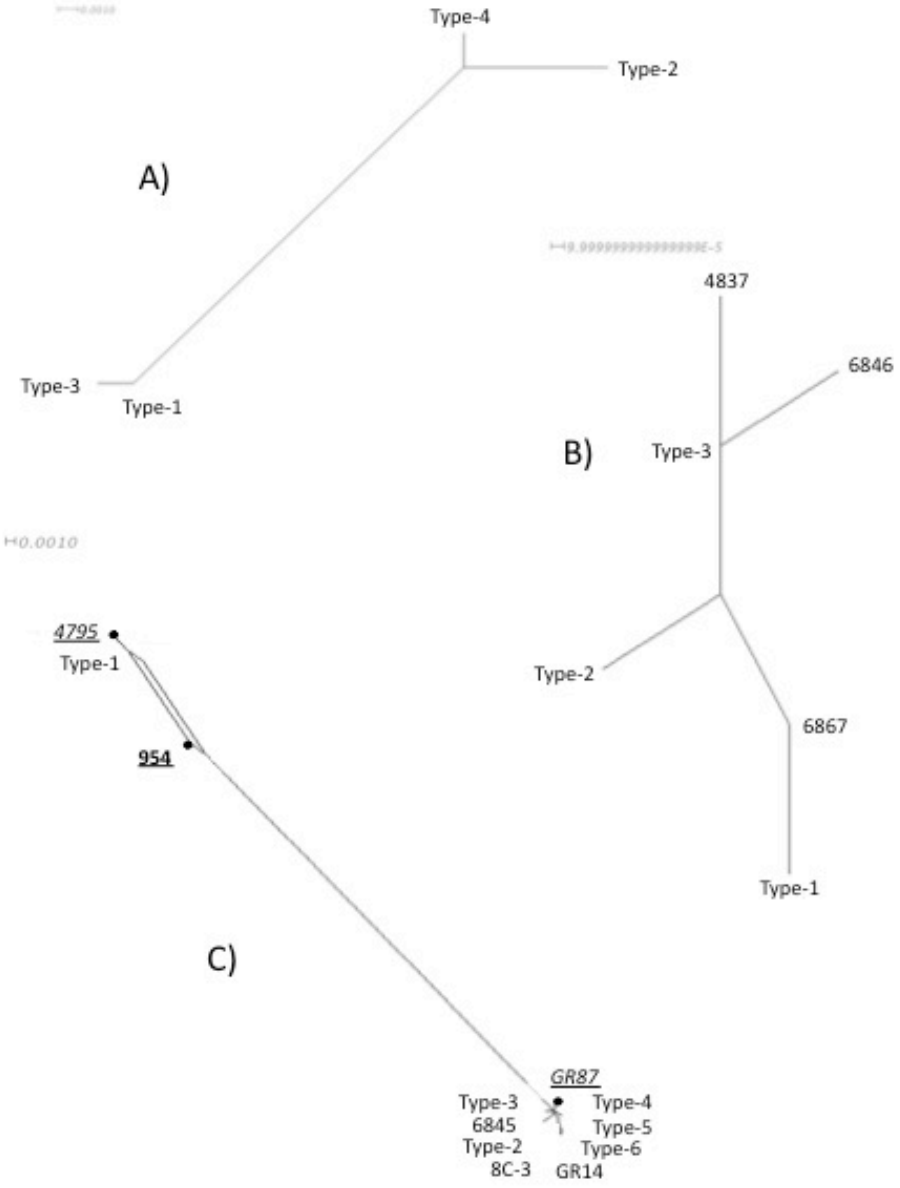

Figura 6. Reconstrucción de redes filogenéticas para los genes A) *nodC*, B) ISRel2 y C) ISRel4. Las cepas subrayadas representan aquellas identificadas como participes de eventos de recombinación mediante RDP3 (tabla 7), las cepas en negritas representan cepa recombinantes y aquellas en itálicas representan cepas parentales. Los grupos (denominados Type) representan varias cepas en el mismo nodo (Ver Apéndice).

Finalmente, existe la posibilidad de que los genes simbióticos estén bajo el efecto de una presión de selección que explique la conservación de sus secuencias y así mismo explique la conservación del contexto genómico de las SIs. Por consiguiente, se analizó si los patrones de sustitución de nucleótidos de los cinco genes se encuentran bajo un modelo de equilibrio neutral mediante la D de Tajima. En el Artículo se muestra que para todas las poblaciones por separado los resultados no son significativos, lo cual indica que se hallan bajo un modelo neutral, sin algún efecto claro de la selección. Solamente cuando se consideran todas las secuencias de *nodC* se obtiene un resultado significativo. Al parecer para el caso de las SIs no hay evidencia de que alguna presión de selección favorezca su conservación de secuencia y de contexto genómico.

DISCUSION

En el presente trabajo se analizó la dinámica evolutiva de las SIs de *R. etli* CFN42 en tres poblaciones de distinto origen geográfico. Al seleccionar SIs que se encuentran conservadas en el mismo contexto genómico, se analizaron SIs que presentan el mismo origen común, lo cual permite comparar su dinámica con genes de distintos replicones.

Las relaciones filogenéticas entre las tres poblaciones (figura 5) mostraron que las poblaciones de Michoacán-Guanajuato y España están más relacionadas entre ellas que con la población de Puebla, cuando lo esperado sería que fueran las dos poblaciones mexicanas las más cercanamente relacionas debido a la distancia geográfica entre ellas. Por un lado esto se puede deber al hecho de que estas tres poblaciones divergieron hace poco y usar únicamente dos genes no es suficiente para trazar sus relaciones filogenéticas. Esto se puede evidenciar por la cantidad de cepas que se situaron en poblaciones que no les correspondían. Diferencias claras se observan entre las filogenias de *glyA* (figura 3) y *dnaB* (figura 4), donde las relaciones filogenéticas entre las tres poblaciones son distintas. Idealmente se requieren un mayor número de marcadores, es decir, un mayor número de sitios informativos para tener una mejor reconstrucción filogenética. Aún así, como se vio en el análisis de genética de poblaciones estos dos genes cromosomales, *glyA* y *dnaB*, resultaron útiles para evaluar las diferentes dinámicas evolutivas de las SIs y así mismo, analizar las diferentes historias evolutivas del cromosoma y el plásmido simbiótico. Por otro lado, aún cuando los análisis filogenéticos entre ambos son distintos, los análisis de genética de poblaciones muestran que estos dos genes “housekeeping” tienen la mismas características.

Hay que tener en consideración que se analizaron tres poblaciones de características distintas. Las cepas de Puebla pertenecen a una milpa, Michoacán-Guanajuato a una milpa y cepas que fueron halladas en plantas de *P. vulgaris* cercanas, mientras que las cepas de España fueron recolectadas a lo largo de una

región de varios kilómetros. Estas diferencias se reflejan en las filogenias, por ejemplo, la población de Puebla forma el clado mejor diferenciado en las tres filogenias (*glyA*, *dnaB* y *glyA-dnaB*) con pocas cepas situadas en otras poblaciones. Por el contrario, España es la población con más cepas presentes en otras poblaciones en las tres filogenias. De forma que idealmente, se requiere trabajar con colecciones que presenten las mismas características. Cabe mencionar que las poblaciones de Puebla y España tienen *Rhizobium* de otras especies, lo cual tiene un efecto entre otras cosas en los eventos de recombinación, ya que tres eventos de recombinación fueron identificables entre cepas de *R. etli* y de *R. gallicum*.

Las SIs por lo general, no permanecen mucho tiempo dentro de un genoma (Wagner, A., 2006), y debido a esto, el perfil de presencia/ausencia de las SIs de CFN42 no corresponde con la filogenia de *glyA* y *dnaB*. Por otro lado, dicho perfil hace evidente que pocas SIs permanecen en el mismo contexto genómico, aun cuando haya pasado poco tiempo desde la divergencia de las tres poblaciones. Aquí es importante mencionar que se están usando las SIs de una cepa actual para averiguar si en otras cepas se hallan en el mismo sitio, lo cual no significa que el ancestro o población ancestro a estas tres poblaciones tuviera el mismo perfil de SIs que CFN42. Aun así, se halló evidencia de dos SIs que mantenían su contexto genómico. Los análisis de genética de poblaciones apuntan claramente a que estas dos SIs han permanecido en el mismo contexto por el origen reciente del plásmido simbiótico. En este sentido la elevada sintenia y conservación de las secuencias de los dos plásmidos simbióticos, de CFN42 y CIAT652, apoyan esta idea. Otras SIs se mantienen en el mismo contexto de algunas poblaciones y/o solamente en algunos replicones como es el caso de algunas SIs del cromosoma y plásmido conjugativo de la población de Puebla.

Como se mencionó en la introducción, existen otros trabajos que han analizado la dinámica evolutiva de las SIs de otras especies. Por un lado, el análisis de varias cepas de *Helicobacter pylori* de distinto origen geográfico (Kalia, et al. 2004) sugiere que sus SIs son elementos ancestrales del pool genético de esta bacteria y que han evolucionado a la misma tasa que otros genes cromosomales. Por otro lado, un trabajo realizado en una población de *Pyrococcus*, un arquea hipertermofílica, sugiere que la deriva génica ha ocasionado que se encuentren algunas SIs con una alta

frecuencia (Escobar-Páramo, P. et al. 2007). Como se muestra en el presente trabajo, para el caso de las dos SIs, ISRel2 e ISRel4, analizadas en las tres poblaciones, su conservación parece estar relacionada con la historia evolutiva del plásmido simbiótico. Tomando en consideración las semejanzas de los datos de genética de poblaciones entre las dos SIs y el gen *nodC*, y comparándolos con los resultados para los genes *glyA* y *dnaB*, es que se propone que el plásmido simbiótico tiene un origen reciente. A diferencia de los genes cromosomales, los tres genes del plásmido simbiótico tienen: (i) valores bajos de diversidad nucleotídica en sitios sinónimos (πs); (ii) sus valores de Dxy y π no están claramente diferenciados; (iii) los tres genes simbióticos comparten varios alelos; (iv) las distintas pruebas de diferenciación genética aplicadas resultaron ser poco significativas; (v) carecen prácticamente de eventos de recombinación y (vi) los plásmidos simbióticos de CFN42 y CIAT652 son altamente sinténicos y tienen una identidad muy alta.

Los análisis de SIs en distintos grupos bacterianos parecen indicar que la dinámica de estos elementos depende en gran medida de las características de las especies y hasta de las cepas (Filée, J. et al. 2007; Lenski, R. et al. 2003; Mira, A. et al. 2006; Parkhill, J. et al. 2003; Schneider, D. et al. 2004). En el caso de *R. etli*, debido a que presenta un plásmido simbiótico capaz de cointegrarse con el plásmido conjugativo (Brom, S. et al. 2000), se podría esperar que las SIs presentes en estos plásmidos fueran las mas distribuidas entre las cepas. El perfil de presencia/ausencia de SIs de este trabajo no puede reflejar esto directamente, aún así se observa que las SIs que conservan en mayor medida su contexto genómico se hallan en estos dos plásmidos. Lamentablemente la razón por la cual algunas SIs conservan su contexto y otras lo pierden no es del todo claro con los análisis realizados en este trabajo.

Finalmente, con ayuda de los genomas completos de CFN42 y CIAT652, se pudo observar que varias SIs, en especial de la familia IS66, se hallaban en varias copias 100% idénticas entre el plásmido simbiótico y el cromosoma. Esto nos daría evidencia de que algunas SIs presentan eventos de transposición recientes entre los dos replicones. Por otro lado, las SIs que se encuentran en plásmidos capaces de movilizarse entre cepas, como los plásmidos conjugativos y simbióticos de *Rhizobium*, podrían estar en un ciclo de infección y expansión en estos genomas,

como se ha propuesto (Wagner, A. 2006). Si el plásmido simbiótico tiene un origen reciente, como se propone en este trabajo, sería factible que estas SIs hayan tenido diversos eventos de transposición recientes del plásmido simbiótico y conjugativo hacia el cromosoma. Con el análisis de un mayor número de genomas y poblaciones de *Rhizobium* se podrá determinar de mejor forma qué tan diseminado se encuentra este plásmido simbiótico y cómo varía la dinámica de sus distintas SIs.

PERSPECTIVAS

Colección de Cepas de *Rhizobium* representativa de México.

En el presente estudio se trabajó con tres colecciones de *Rhizobium* de características diferentes como se detalla en la metodología. Actualmente existe una colección de cepas de *Rhizobium* de distintas especies que representan de mejor forma la distribución y diversidad de este género en México. Dicha colección está resguardada en el Laboratorio de Evolución Molecular y Experimental del Instituto de Ecología de la UNAM. Esta colección permite tener un mejor conocimiento del origen y características de las cepas de esta especie. Esto da la oportunidad de trazar de mejor forma la distribución de las distintas *Rhizobium* en relación a la presencia y diversidad de las SIs. Analizando esta colección se puede obtener un perfil más detallado de las SIs de la cepa CFN42 en cepas que presenten las mismas características. Por ejemplo, se podrían seleccionar cepas provenientes únicamente de milpas o por el contrario, cepas colectadas en regiones donde no es tan habitual el cultivo de frijol. De esta forma se podría evaluar la conservación del contexto genómico de las SIs relacionándolo con las características de las cepas seleccionadas.

Movilidad de SIs en Cepas de *Rhizobium etli*.

Actualmente no se tiene evidencia directa de la movilidad de SIs en cepas de *Rhizobium etli*. Para poder entender de mejor forma la dinámica de las SIs es necesario conocer que tan activos son estos elementos en los genomas. Existen trabajos que han estudiado la movilidad de SIs en otras especies diferentes de *Rhizobium* (Hernández-Lucas, et al. 2006; Waskar, et al. 2000). Estos trabajos emplean plásmidos que 'atrapan' a las SIs, lo cual permite posteriormente secuenciarlas y analizarlas. En relación al presente trabajo, inicialmente se pueden seleccionar las cepas CFN42 y CIAT652 que tienen su genoma completamente secuenciado. De esta forma se podría determinar la localización inicial de la SI que fuera activa. Por un lado resulta interesante determinar la actividad de las dos SIs, ISRel2 e ISRel4, de este trabajo, lo que permitiría relacionar su movilidad con la conservación de su contexto genómico. Por otro lado, se puede analizar la actividad

de las SIs que al parecer se han estado copiando entre los distintos replicones, como es el caso de las IS66 de las dos cepas.

Análisis de SIs en Genomas Completos y Parciales de *Rhizobium etli* y otras Especies de *Rhizobium*.

Actualmente existen dos genomas completos y siete genomas parciales de *Rhizobium etli* en la base de datos de GenBank del NCBI. Por otro lado existen genomas de otras especies como *Rhizobium leguminosarum,* y actualmente en el laboratorio de Genómica Evolutiva, en colaboración con otros laboratorios del Centro de Ciencias Genómicas, se están ensamblando y analizando genomas de otras cepas (*R. etli* bv. Mimosae y *R. etli* IE4771) y otras especies (*Rhizobium gallicum*). Estos genomas dan la posibilidad de realizar diversos análisis relacionados con la diversidad y conservación de SIs, así como de los efectos de su presencia y/o actividad. Por otro lado, se puede evaluar si los eventos de expansión de la IS66 en las cepas CFN42 y CIAT652 de *R. etli* se repiten en otros genomas.

Los genomas de *R. etli* permiten analizar con mas detalle la relación de las SIs con los replicones que las presentan, principalmente las diferencias halladas entre los plásmidos conjugativo y simbiótico. En esta tesis se analizaron únicamente las SIs de una cepa, CFN42, pero haciendo uso de estos genomas el análisis se puede extender a SIs que no estén conservadas en el mismo contexto genómico. Un ejemplo se encuentra en la ISRel2 de este trabajo, la cual no se encontró en el mismo contexto genómico en la cepa IE4771, pero un análisis del ensamble de su genoma muestra que efectivamente se encuentra en otra región con una identidad de 99.6%. Cabe mencionar que esta región presenta genes homólogos a genes del plásmido simbiótico de otras cepas. Otro ejemplo se presenta en el genoma de *R. etli* bv. Mimosae. El análisis del ensamble de esta cepa muestra que existen diferencias entre su plásmido simbiótico y el de otras cepas secuenciadas. Al parecer *R. etli* bv. Mimosae tiene un plásmido simbiótico distinto, aun así, al buscar la ISRel2 se encuentra en el mismo contexto genómico que CFN42 con una identidad de 99.4%. Una posible explicación seria que el plásmido simbiótico de *R. etli* bv. Mimosae tuviera una región homóloga al plásmido simbiótico de CFN42. Lo anterior no resultaría extraño ya que existen ejemplos de plásmidos quiméricos formados de varias regiones de distintos replicones

(Cervantes, et al. 2011). De esta forma, se pueden rastrear regiones de plásmidos y/o cromosomas que muestren señales de homología en busca de las mismas SIs. Por otro lado, se puede analizar la dinámica de estos elementos considerando su movilidad y participación en eventos como rearreglos. Así mismo, se obtendrían un mayor número de genes adicionales para analizar su historia evolutiva y compararlos con las SIs como se realizó con los genes cromosomales de este trabajo.

APENDICE I

El perfil de presencia/ausencia de SIs (ver figura 2 del Artículo) muestra que las cepas presentan SIs de CFN42 conservadas en el mismo contexto genómico. En general, los perfiles resultan ser cepa específico, aun así existen semejanzas dentro de cada población, sobre todo entre cepas de la población de Puebla y entre las cepas de la población Guanajuato-Michoacán. Lo anterior se puede observa al realizar un análisis de las similitudes entre los distintos perfiles de SIs obteniendo las distancias de Sorensen entre las cepas y realizando posteriormente un dendograma que relaciona las cepas con perfiles más semejantes.

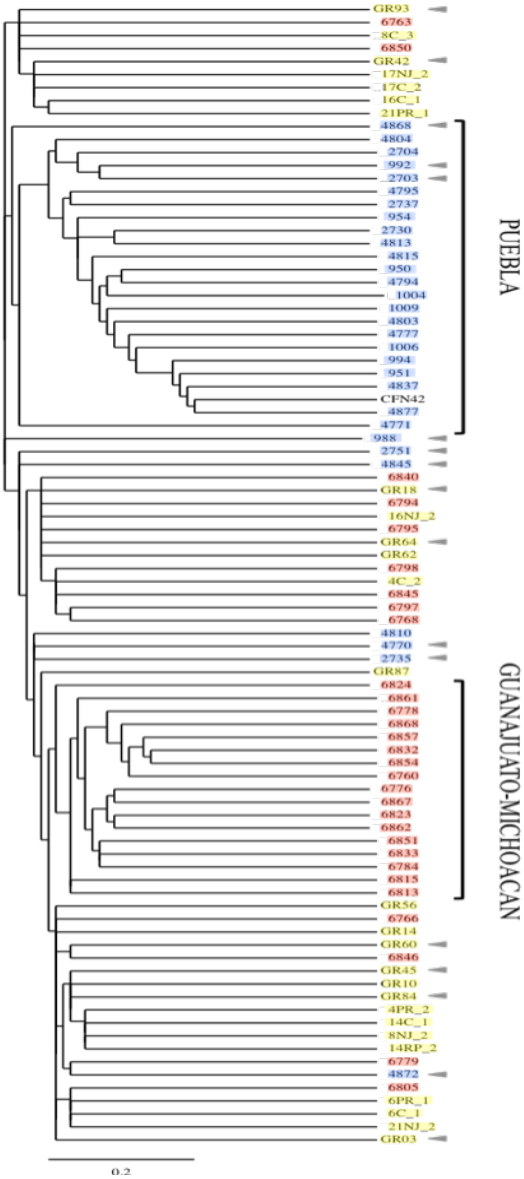

Figura X. Dendograma que muestra las similitudes en perfiles de SIs de las cepas de las poblaciones (azul:Puebla; rojo:Guanajuato-Michoacán; amarillo:España). Las flechas indican cepas distintas de *R. etli*.

APENDICE II

Relación de los grupos (type) de las redes filogenéticas con las cepas de las tres poblaciones.

dnaB

Type-1 : 4803, 950, 4795

Type-2 : 951, 4804, 4837, 2730, 1004, 1009, 4771

Type-3 : 4770, 2703, 992

Type-4 : 6840, 6868, 6832, 6857, 6861, 6854, 6776

Type-5 : 6846, 6766

Type-6 : 8C-3, GR18

Type-7 : GR93, 6PR-1

Type-8 : 21PR-1, 4PR-2

glyA

Type-1 : 951, 4837

Type-2 : 4813, 4872, 4777, 954, 4877, 4810, 1006, 1009, 994

Type-3 : 2704, 950, 2737

Type-4 : 2703, 992

Type-5 : 2751, GR84, GR18, GR45

Type-6 : 6840, 6868, 6857

Type-7 : 6778, 6815

Type-8 : 6779, 6797

Type-9 : 6795, 6794, 6768, 6798

Type-10 : 8C-3, GR93, 16C-1, 17NJ-2, GR10, GR60, 4C-2, 6PR-1, 6C-1, GR56, GR42

Type-11 : 21PR-1, 4PR-2, 17C-2

nodC

Type-1 : CFN42, 4813,4837, 4810, 4872, 951, 994, 6854, 6868, 6833, 6776, 6760, 6815, 6824, 6832, 6846, 6857, 6778, 6867, 6850, 6862, GR14

Type-2 : 17C-2 GR18, GR10, 21PR-1, 14C-1, GR60, 8NJ-2, GR45, 8C-3, GR56, GR84, GR03, GR42

Type-3 : 6766, 6823

Type-4 : 2730, 6861, 6845, 6784, 6813, 6840, 21NJ-2

ISRel2

Type1 : CFN42, 4C-2, 8C-3, 6C-1, 17NJ-2, 6868, 6854, 6778, 6823, 6760, 6766, 6857, 6832, 6805

Type2 : 6824, 6862, 2730, 4813, 994

Type3 : 951, 6833, GR10, 8NJ-2, 6PR-1, 14C-1, GR56, GR87, 21PR-1, GR03, 6784, GR42, GR60, 14PR-2, GR45, GR14, 4PR-2, GR84, 17C-2, 21NJ-2

ISRel4

Type1 : 2737, 4872

Type2 : 6779, 6795, 6794, 6768

Type3 : 4813 2730, 951, 994, 4877

Type4 : 6805, 6C-1, 6PR-1, 4C-2

Type5 : 6857, 6776, 6832, 6778, GR18, GR62, 8NJ-2, 4PR-2, GR03, 14PR-2, GR45, GR60, GR84, 18NJ-2, GR10, 14C-1, GR64, GR56

Type6 : 4803, 6840, 6851, 6862, 6861, 6867, 6846, 6868, 6784, 6823, 6813

REFERENCIAS

1. Bisercic, M. y H.Ochman. 1993. The ancestry of insertion sequences common to *Escherichia coli* and *Salmonella typhimurium*. J. Bacteriol. 175 (24):7863-7868.
2. Blot, M. 1994. Transposable elements and adaptation of host bacteria. Genetica. 93 (1-3):5-12.
3. Boyd, E. y D. Hartl. 1997. Nonrandom location of IS1 elements in the genomes of natural islolates of Escherichia coli. Mol. Biol. Evol. 14 (7):725-732.
4. Brom, S., García-de los Santos, A., Cervantes, L., Palacios, R. y Romero D. 2000. In *Rhizobium etli* symbiotic plasmid transfer, nodulation competitivity and cellular growth require interaction among different replicons. Plasmad. 44:34-43.
5. Brügger, K., et al. 2002. Mobile elements in archaeal genomes. FEMS Microbiol. Lett. 206:131-141.
6. Cervantes, L., Bustos, P., Girard, L., Santamaría, R., Dávila, G., Vinuesa, P., Romero, D. y S. Brom. 2011. The conjugative plasmid plasmid of a bean-nodulating *Sinorhizobium fredii* strain is assembled from sequences of two *Rhizobium* plasmids and the chromosome of a *Sinorhizobium* strain. BMC Microbiology. 11:149-159.
7. Chandler, M., y J. Mahillon. 2002. Insertion Sequences revisited. Pp.305-366. In Craig, L. et al. (Eds). Mobile DNA II, ASM Press, Washington D. C.
8. Edgar, R. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5:113.
9. Edwards, R. y J. Brookfield. 2003. Transiently beneficial insertions could maintain mobile DNA sequences in variable environments. Mol. Biol. Evol. 20(1):30-37.
10. Escobar-Páramo, P., Ghosh, S. y J. DiRuggiero. 2005. Evidence for genetic drift in the diversification of a geographically isolated population of the hyperthermophilic arcaeon *Pyrococcus*. Mol. Biol. Evol. 22 (11):2297-2303.

11. Filée, J., Siguier, P. y M. Chandler. 2007. Insertion sequence diversity in Archea. Microbiol. Mol. Biol. Rev. 71 (1): 121-157.
12. Flores, M., Morales, L., Avila, M., Bustos, P., González, V., Garcia, D., Mora, Y., Guo, X., Collado-Vides, J., Piñero, D., Davila, G Mora, J. y R. Palacios. 2005. Diversification of DNA sequences in the symbiotic genome of Rhizobium etli. J. Bacterial. 187(21):7185-92.
13. García-de los Santos, A., Brom ,S., and Romero, D. (1996). Rhizobium plasmids in bacteria-legume interactions. World. Microbiol. Biotechnol. 12:119-125.
14. Gasca, J. 2004. Genética de poblaciones de *Rhizobium etli* asociado a *Phaseolus vulgaris* en dos localidades del bajío mexicano. Tesis de Licenciatura. Instituto de Ecología. UNAM. 102 pag.
15. González, V., Santamaría, R., Bustos, P., Hernandez-González, I., Medrano-Soto, A., Moreno-Hagelsieb, G., Janga, S., Ramirez, M., Jiménez-Jacinto, V., Collado-Vides, J. and G. Davila. 2006. The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. PNAS 103(10):3834-3839.
16. Guindon, S. and O. Gascuel. 2002. Efficient biased estimation of evolutionary distances when substitution rates vary across sites. Mol Biol Evol. (4):534-43.
17. Hartl, D., & S. Sawyer. 1988. Why do unrelated insertion sequences occur together in the genome of *Escherichia coli*? Genetics 118:537-541.
18. Hernández-Lucas, I., Ramírez-Trujillo, J., Gaitán, M., Guo, X., Flores, M., Martínez-Romero, E., Pérez-Rueda, E. Y P. Mavingui. 2006. Isolation of functional insertion sequences of rhizobia. FEMS Microbiol. Lett. 261:26-31.
19. Huson, D. H. and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23(2): 254-67.
20. Kalia, A. et al. 2004. Evolutionary dynamics of insertion sequences in *Helicobacter pylori*. J. Bacterial. 186 (22):7508-7520.
21. Kichenaradja P., Siguier, P., Percochon, J. and M. Chandler. ISbrowser: an extensión of ISfinder for visualizing insertion sequences in prokaryotic genomes. Nucleic Acids Res. 38:D62-68.
22. Kleckner, N. 1990. Regulation of transposition in bacteria. Annu. Rev. Cell. Biol. 6:297-327.

23. Kobayashi, K. et al. 2003. Essential *Bacillis subtilis* genes. PNAS 100 (8):4678-4683.
24. Lawrence, J., Ochman, H., & D. Hartl. 1992. The evolution of insertion sequences within enteric bacteria. Genetics 131:9-20.
25. Lenski, R., Winkwirth, C., & M. Riley. 2003. Rates of DNA sequence evolution in experimental populations of *Escherichia coli* during 20,000 generations, J. Mol. Evol. 56:498-508.
26. Librado, P. and J. Rozas. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25(11): 1451-2.
27. M. De Vienne, Giraud, T & Olivier C.M. 2007. A congruente index for testing topological similarity between trees. Bioinformatics. 23(23):3119-3124.
28. Mahillon, J., Léonard, C., & M. Chandler. 1999. IS elements as constituents of bacterial genomes. Res Microbiol. 150:675-687.
29. Martin, D. P., Williamson, C. and D. Posada. 2005. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics 21(2): 260-2.
30. Mira, A., Pushker, R., & F. Rodríguez-Valera. 2006. The neolithic revolution of bacterial genomes. Trends Microbiol. 14 (5):200-206.
31. Moran, N. & G. Plague. 2004. Genomic changes following host restriction in bacteria. Curr Opin Genet Dev. 14(6):627-33.
32. Naas, T., Blot, M., Fitch, W., & W. Arber. 1994. Insertion sequence-related genetic variation in resting *Escherichia coli* K-12. Genetics 136:721-730.
33. Naas, T., Blot, M., Fitch, W., & W. Arber. 1995. Dynamics of IS-related genetics rearrangements in resting *Escherichia coli* K-12. Mol. Biol. Evol. 12 (2):198-207.
34. Papadopoulus. D. et al. 1999. Genomic evolution during a 10,000-generation experiement with bacteria. Proc. Natl. Acad. Sci. 96:3807-3812.
35. Parkhill, J. et al. 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat. Genet. 35:32-40.
36. Rodríguez-Navarro, D., Buendia, A. Camacho, M., Lucas M. and C. Santamaria. 2000. Characterization of *Rhizobium* spp. Bean isolates from south-west Spain. Soil Biol. Biochem. 32:1601-1613.

37. Sawyer, S. & D. Hartl. 1986. Distribution of transposable elements in prokaryotes. Theor. Popul. Biol. 30:1-16.
38. Schneider, D., Duperchy, E. Coursange, E. Lenski, R. & M. Blot. 2000. Long-term experimental evolution in Escherichia coli IX. Characterization of insertion sequence-mediated mutations and rearrangements. Genetics. 165 (2):477-488.
39. Schneider, D., & R. Lenski. 2004. Dynamics of insertion sequence elements during experimental evolution of bacteria. Res. Microbiol. 155:319-327.
40. Siguier, P., Filée, J., & M. Chandler. 2006. Insertion sequences in prokaryotic genomes. Curr. Opin. Microbiol. 9:1-6.
41. Siguier, P., et al. 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 43:D32-D36.
42. Silva, C., Eguiarte, L. E. y Souza,V. 1999. Reticulated and epidemic population genetic structure of Rhizobium etli biovar phaseoli in a traditionally managed locality in Mexico. Mol. Ecol. 8:277-287.
43. Silva, C., Vinuesa, P., Eguiarte, L. Martínez-Romero, E. and V. Souza. 2003. *Rhizobium etli* and *Rhizobium gallicum* nodulate common bean (*Phaseolus vulgaris*) in a traditionally managed milpa plot in Mexico: Population genetics and Biogeographic implications. Appl. Env. Microbiol. 69:884-893.
44. Sirand-Pugnet, P. et al. 2007. Being pathogenic, plastic, and sexual while living with nearly minimal bacterial genome. PLOS Genetics 3 (5):744-758.
45. Stinear, T., et al. 2007. Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. Genome Res. 17:192-200.
46. Touchon, M., & E.Rocha. 2007. Causes of insertion sequence abundante in prokariotyc genomes. Mol. Biol. Evol. 24(4):969-981.
47. Wagner, A. 2006. Periodic extinctions of transposable elements in bacterial lineales: Evidence from intragenomic variation in multiple genomes. Mol. Biol. Evol. 23 (4):723-733.
48. Wagner, A., Lewis, C., & M. Bichsel. 2007. A survey of bacterial insertion sequences using IScan.. Nucleic Acids Res. 1-10.
49. Waskar, M., Kumar, D., Kumar, A. & R. Srivastava. 2000. Isolation of novel insertion sequence from *Mycobacterium fortuitum* using a trap vector based on inactivation of a lacZ reporter gene. Microbiology 146:1157-1162.

50. Yang, F. et al. 2005. Genome dynamics and diversity of Shigella species, the etologic agents of bacillary dysentery. Nucleic Acids Res. 33 (19):6445-6458.
51. Zhong, S. et al. 2004. Evolutionary genomics of ecological specialization. PNAS. 101 (32):11719-11724.