UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

CENTRO DE CIENCIAS GENÓMICAS
DOCTORADO EN CIENCIAS BIOMÉDICAS

IMPACTO DE LA RECOMBINACIÓN HOMÓLOGA EN GENOMAS DE *Rhizobium etli*

T E S I S

QUE PARA OBTENER EL GRADO DE

DOCTOR EN CIENCIAS

P R E S E N T A:

JOSÉ LUIS ACOSTA RODRÍGUEZ

DIRECTOR DE TESIS: DR. VÍCTOR MANUEL GONZÁLEZ ZÚÑIGA

CUERNAVACA, MOR. MARZO DEL 2012

AGRADECIMIENTOS

Un especial y sincero agradecimiento, así como todo mi respeto y gratitud al Dr. Víctor Manuel González Zúñiga, quien me brindo la oportunidad de estar en su grupo de trabajo. Además de apoyarme intelectual, material y emocionalmente durante la realización del doctorado.

Agradezco a mi comité tutor, integrado por el Dr. José Guillermo Dávila Ramos y el Dr. Luis Enrique Eguiarte Fruns, por toda la ayuda brindada, desde la concepción hasta el término del proyecto.

Al comité revisor, integrado por el Dr. Miguel Ángel Cevallos Gaos, Dra. Susana Magallón Puebla, Dr. Arturo Carlos II Becerra Bracho y Dr. Jesús Silva Sánchez por la meticulosa revisión de esta tesis y los comentarios brindados para mejorarla.

Al personal del laboratorio de Genómica Evolutiva, Ing. José Espíritu, I.Q. Patricia Bustos, M. en C. Rosa Isela y Lupita. De igual manera, agradezco a todos mis compañeros de laboratorio, especialmente a Luis Lozano y Napoleón González por compartir agradables momentos.

Toda mi gratitud a mi familia por su apoyo incondicional, especialmente a Fabiola Mendoza por haberme apoyado en momentos difíciles y por darle nuevos bríos a mi vida.

Finalmente, agradezco el apoyo económico brindado por CONACYT y PAPIIT-UNAM, así como a todo el personal del Centro de Ciencias Genómicas, UNAM por su atención y ayuda brindada.

ÍNDICE

INDICE DE FIGURAS

PRESENTACION

El proyecto de doctorado se realizó en las instalaciones del Centro de Ciencias Genómicas UNAM, campus Cuernavaca, en el laboratorio de Genómica Evolutiva bajo la dirección del Dr. Víctor Manuel González Zúñiga. El presente trabajo es la culminación de dos estudios previos en los que participé como coautor y que se describen en los antecedentes y se anexan los artículos correspondientes en el apéndice.

El comité tutor que evaluó el avance del presente proyecto de investigación estuvo integrado por el Dr. Víctor Manuel González Zúñiga, el Dr. José Guillermo Dávila Ramos y el Dr. Luis Enrique Eguiarte Fruns.

El jurado de la réplica oral de este trabajo de investigación se integró por el Dr. Miguel Ángel Cevallos Gaos, el Dr. Víctor Manuel González Zúñiga, la Dra. Susana Magallón Puebla, el Dr. Arturo Becerra Bracho y el Dr. Jesús Silva Sánchez.

El trabajo experimental de esta tesis lo financió PAPIIT-UNAM (IN215908 y IN223005) y CONACyT (CB131499).

Durante la realización de esta tesis José Luis Acosta Rodríguez recibió una beca de doctorado otorgada por CONACyT (200018) del periodo de Agosto 2005 a Julio del 2010.

Correo: jlacosta@ccg.unam.mx;acostarojo@gmail.com

PROLOGO

La taxonomía del genero *Rhizobium* esta en continuos cambios, esto es debido al progreso de las técnicas empleadas para su clasificación. Cabe destacar que la técnica que ha evolucionado constantemente es la secuenciación de DNA genómico. A la fecha, se puede secuenciar un genoma bacteriano en un par de días, siempre y cuando se utilice una corrida completa de un aparato de secuenciación, ya sea 454 (*FLX* y *Titanium*) de Roche, SOLiD de AppliedBiosystems o Illumina (*GA* y *Hiseq2000*) de Solexa. Sin embargo, generalmente se utiliza una parte de la corrida para secuenciar un genoma bacteriano, dando como resultado un genoma fragmentado o incompleto.

Los análisis genómicos y filogenéticos de genomas incompletos representan un verdadero reto, ya que las inferencias obtenidas a partir de estos pueden estar subestimadas por las bajas coberturas de secuenciación. Por lo tanto, es de suma importancia desarrollar herramientas y algoritmos que reduzcan el margen de error generado por las bajas coberturas.

El genero *Rhizobium* se caracteriza por incluir a especies bacterianas fijadoras de nitrógeno. Esta fijación se realiza mediante una relación simbiótica con plantas leguminosas. Los genomas de estas bacterias tienen la propiedad de ser multireplicones, es decir, que poseen un cromosoma (o dos cromosomas) y plásmidos. Estos replicones (plásmidos), se ganan y se pierden en las poblaciones y un porcentaje importante de estos incluye gran cantidad de elementos móviles.

Aunado a esto, los genomas de *Rhizobium*, presentan una cantidad importante de DNA redundante (familias de parálogos) y endémico (especifico para cada cepa).

La estructura genómica de *Rhizobium etli* es dinámica y está moldeada por fuerzas evolutivas como la mutación y la recombinación (homóloga y no homóloga). La contribución de ambas fuerzas no está del todo clara, previos estudios indican que existe recombinación dentro de las poblaciones de *R. etli*. Sin embargo, no se ha hecho una valoración a nivel de genoma.

La mayoría de genomas secuenciados de *R. etli* están fragmentados (seis genomas) y sólo se cuenta con dos genomas completos. La heterogeneidad de las coberturas de secuenciación, anudado a la complejidad y dinámica genómica, representan un doble reto al tratar de valorar la contribución de la mutación y la recombinación en la evolución de la especie. El presente trabajo tiene dos aportaciones: El metodológico y el evolutivo. La primera aportación es el desarrollo de metodologías para la determinación de segmentos ortólogos, identificación de la variación nucleotídica y la detección de recombinación homóloga. Estas metodologías están enfocadas y adaptadas para trabajar con genomas incompletos. La segunda aportación consta de propuestas para tratar de delimitar a la especie de *R. etli*, estudios filogenéticos y la valoración de la recombinación homologa en la evolución de los genomas de *R. etli.*

CAPITULOS

La tesis describe, a lo largo de seis capítulos, los aspectos conceptuales y metodológicos, así como los resultados y discusión general del mi proyecto de doctorado.

El capítulo I explica y define la relación simbiótica rizobio-leguminosa, detallando las características de su interacción y evolución entre los socios. Además, también se mencionan los pormenores que sufre la taxonomía de *Rhizobium*. Por otra parte, se define la recombinación homóloga, así como su dinámica y detección. Por último, introducimos el término de especie genómica, el cual trata de sortear las discrepancias provocadas por clasificaciones taxonómicas.

El capítulo II disecta la estructura genómica de *R. etli*, esto abarca desde los genes encargados de la simbiosis hasta nivel de replicón. En esta parte, incluimos los artículos de coautoría. Estos análisis son comparaciones genómicas en primera instancia entre especies de *Rhizobium* y una subsecuente comparación dentro de la especie de *R. etli*.

El capítulo III presenta el artículo aceptado en BMC evolutionary biology cuyo titulo es: “Genomic lineages of *Rhizobium etli* revealed by the extent of nucleotide polymorphisms and low recombination”.

El capitulo IV lista los genomas usados, describe las metodologías diseñadas, así como su valoración. Además incluye metodologías que no están en

el artículo publicado. Estos nuevos análisis están adaptados para utilizar genomas incompletos, sin importa la metodología con que se hayan secuenciado.

El capitulo V reseña los resultados obtenidos en el artículo, así como también los resultados de los análisis adicionales, los cuales son una base importante para las inferencias realizadas.

El capitulo VI proporciona las discusiones hechas en el artículo, además de estar soportadas por resultados adicionales, los cuales brindan una mejor comprensión sobre el rol de la recombinación homóloga en la diversificación genómica de *R. etli*.

OBJETIVOS

Este estudio tiene como objetivo general, conocer el papel de la recombinación homóloga en la generación de diversidad nucleotídica en ocho genomas de *Rhizobium etli* provenientes de cepas de distinto origen geográfico. Los genomas utilizados se secuenciaron por la metodología de secuenciación al azar (*Shotgun sequencing*), con la plataforma Sanger (secuenciación capilar). La muestra consta de 2 genomas completos y 6 genomas parciales, estos últimos tienen una cobertura aproximada de 1X (una base se ha cubierto al menos una vez). Debido a la naturaleza fragmentada de los genomas parciales se generaron los siguientes objetivos particulares: (1) Diseño y desarrollo de una metodología confiable para la identificación de mutaciones puntuales (SNPs), (2) Caracterización de SNPs, (3) Implementación de una medida de diferenciación genómica y finalmente, (4) Desarrollo de una metodología para la detección de recombinación homóloga. Estos objetivos se enfocaron para trabajar con genomas fragmentados y las metodologías generadas en este trabajo pueden aplicarse para los genomas producidos por la segunda generación de tecnologías de secuenciación masiva (*Illumina* Hiseq, *454* (paired-end) y *SOLiD4*), ya que generalmente esos genomas no llegan a terminarse con la debida solidez para considerarlos como genomas “completos”.

RESUMEN

En este trabajo, se estudió el impacto de la recombinación homóloga en genomas de *R. etli* obtenidos de cepas de distinto origen geográfico. Con este propósito se utilizó la secuencia de dos genomas completos (*CFN42* y *CIAT652*) y seis genomas incompletos (*BRASIL5*, *CIAT894*, *GR56*, *IE4771*, *KIM5* y *8C-3*) con una cobertura aproximada de 1X. Debido a la naturaleza fragmentada de los genomas incompletos, diseñamos y valoramos una metodología para la identificación de SNPs. Además cuantificamos y caracterizamos la variación intragénica en los genomas de *R. etli* y para ello propusimos una medida de variación promedio. Para detectar y analizar la recombinación homóloga implementamos un análisis de cuartetos que permite utilizar la información contenida en genomas fragmentados. Nuestros análisis identificaron una alta variación intragénica, tanto en comparaciones pareadas como en tercias (dos genomas de referencia vs genoma incompleto). La divergencia promedio que encontramos entre las comparaciones pareadas fue entre el 4% y 6%. Por otra parte, los eventos de recombinación estimados afectan del 3% al 10% de la muestra genómica analizada y además la diversidad nucleotídica (π) fue mayor en los segmentos recombinantes que en los no recombinantes. Sin embargo, la valoración filogenética indicó que la recombinación no tuvo la fuerza suficiente para interrumpir la congruencia del árbol filogenético de la muestra. En base a estos resultados, se discute el aporte de la recombinación en la formación de la diversidad genómica y en la evolución de los genomas analizados.

ABSTRACT

In this work, we studied the impact of homologous recombination in *R. etli* genomes obtained from strains of different geographical origin. Two complete genomes (*CFN42* and *CIAT652*) and six incomplete genomes (*Brazil5*, *CIAT894*, *GR56*, *IE4771*, *KIM5* and *8C-3*) with approximately 1X coverage were used. Due to the fragmented nature of the incomplete genomes, we design and value a methodology for identifying SNPs. Besides, we quantifying and characterizing intragenic variation in the genomes of *R. etli* and it proposed a measure of average change. To detect and analyze homologous recombination implement an analysis of quartets, which allows extending the search to incomplete genomes. Our analysis identified a high intragenic variation in both paired comparisons and in thirds (two reference genomes genome vs. incomplete). The average difference found was 4% and 6% for paired comparisons. Moreover, the estimated recombination events affecting 3% to 10% of the sample analyzed and further genomic nucleotide diversity (π) was higher in recombinant segments that non recombinants. However, phylogenetic assessment indicated that recombination was not strong enough to disrupt the congruence of the phylogenetic tree of the sample. Based on these results, we discuss the contribution of recombination in the formation of genomic diversity and evolution of genomes analyzed.

CAPITULO I.- INTRODUCCION GENERAL

I.1 Relación simbiótica

I.1.1 Interacción leguminosa-rizobios

Los rizobios son bacterias del suelo que ocasionalmente establecen una relación simbiótica con plantas leguminosas mediante la formación de nódulos en las raíces y dentro de estos (nódulos) se lleva a cabo la fijación de nitrógeno atmosférico. Los principales géneros que conforman los rizobios son: *Sinorhizobium*, *Mesorhizobium*, *Bradyrhizobium*, *Allorhizobium* y *Rhizobium* [1]. En el caso de las interacciones *Rhizobium*-leguminosas, la simbiosis se inicia mediante un dialogo molecular, en el cual los flavonoides de la planta interactúan con la proteína activadora *NodD*, para encender las señales necesarias para la producción del factor Nod por los rizobios, que permite la infección en el cortex de la raíz y dispara el desarrollo del nódulo [2-4]. Sin embargo, la simbiosis *Rhizobium*-leguminosa varía en especificidad, tanto en el rango de los huéspedes, como también en la diversidad de especies bacterianas nodulantes [5].

La diversidad en rizobios se da a niveles de especies y cepas [6] lo que incrementa la oportunidad para un huésped (leguminosa) de encontrar un rizobio compatible en un suelo particular [7]. No obstante, antes de infectar a las plantas leguminosas, los rizobios emplean una variedad de estrategias que permiten que existan en el suelo y se adapten a diversas condiciones ambientales [8]. Dichas estrategias dependen de genes que intervienen en tolerancia al estrés, los cuales son una importante propiedad para una cepa inoculante, por ejemplo, mientras

ocurre la desecación de una semilla, o si ésta se encuentra en un lugar hostil, dando como resultado rizobios endémicos capaces de nodular a huéspedes locales, aunque estos no sean muy eficientes en la fijación de nitrógeno [9]. Muchos de estos genes también son importantes para la competitividad entre rizobios y generalmente se acarrean a través de plásmidos (compartimentos extracromosomales) de gran tamaño (> 100kb-1500kb). En el caso de los genes involucrados en la nodulación y fijación de nitrógeno, también se localizan en plásmidos (pSyms) o en islas simbióticas [10] y en algunos casos, la interacción funcional entre el plásmido simbiótico y los demás replicones (plásmidos) es necesaria no solamente para la competitividad en la nodulación, sino también para la utilización de diferentes fuentes de carbono [11], lo cual podría ser importante para la adaptación al estilo de vida libre.

La presencia de ciertos tipos de plásmidos es determinante para establecer la simbiosis mutualista, dicha simbiosis frecuentemente se basa en la integración funcional, estructural y genética de los socios, lo que da como resultado el desarrollo de nuevos rasgos fenotípicos [12]. Sin embargo, las leguminosas seleccionan los rizobios más cooperativos, es decir, los que tienen mayor capacidad de fijar nitrógeno [13], ya que una sola leguminosa puede estar infectada por diferentes linajes bacterianos [14]. Esta ventaja selectiva incrementa la frecuencia genotípica de los rizobios huésped-especifico de los no específicos, impactando en la integridad funcional y en la eficiencia ecológica de la interacción leguminosa-rizobio [15].

I.1.2 Taxonomía de Rizobios

La taxonomía de los rizobios está en un constante cambio en la medida que se incluyen en ella nuevos avances tecnológicos y nuevos criterios tanto morfológicos, como bioquímicos, fisiológicos y últimamente el análisis de secuencias [16]. En la década del setenta la taxonomía se enfocó en la variación de las secuencias de los genes que codifican para funciones esenciales, los cuales podrían servir como marcadores taxonómicos [17]. En la actualidad el uso de secuencias de 16S rRNA [18] ha permitido identificar a 44 especies reconocidas de bacterias que nodulan a leguminosas dentro de 11 géneros, nueve de los cuales pertenecen a las alfa-proteobacterias: *Allorhizobium, Azorhizobium, Bradyrhizobium, Devosia, Mesorhizobium, Methylobacterium, Ochrobactrum, Rhizobium y Sinorhizobium* de acuerdo a la lista de nombres de procariotas http://www.bacterio.cict.fr/pub/ [19]. Sin embargo, existe una incongruencia en los resultados filogenéticos obtenidos con 16S rRNA, generada por eventos de transferencia lateral de genes (TLG) y por eventos recombinación [20-23].

Con el fin de disminuir el efecto de la TLG en el marcador 16S rRNA, en secuencias de *Agrobaterium* y *Rhizobium,* se utilizó un valor de corte de identidad nucleotídica mayor a 93 %, dando como resultado una relación estrecha entre ambos géneros [22]. Por otra parte, al tomar secuencias de diferentes regiones del operón *rrn*, tales como: 16S rRNA, 23S rRNA e ITS (regiones intergénicas) de varios miembros del grupo de rizobios, se observó que las topologías generadas con 16S fueron significativamente diferentes de las obtenidas con 23S y ITS [23].

Ante los efectos evidentes de la TLG y la recombinación dentro de los marcadores ribosomales, la selección de genes de mantenimiento celular (*housekeeping*) de diversos *loci* cromosomales aunado a la hibridación DNA-DNA, mejoran indudablemente los criterios para la identificación de las bacterias a nivel de especie [24].

Para los procariotas en general los genes *rpoA, thdF* y *recN* han demostrado tener la misma precisión que las hibridaciones DNA-DNA [25]. Sin embargo, para el análisis en rizobias existe un conjunto de genes que comúnmente se utilizan con ese propósito, como es el caso de genes cromosomales *gluA, gyrB* y *recA* y para genes simbióticos *nodA, nodB, nifH* y *nifK* [26]. No obstante, estos marcadores no están exentos del impacto de la TLG, en especial los marcadores simbióticos (*nif*), ya que en algunas especies de *Bradyrhizobium* se encontró evidencia de recombinación entre sus linajes [27]. Por lo tanto, las posibles inferencias filogenéticas están en función de la elección de buenos marcadores moleculares. Ante estos hechos, la filogenómica (la comparación de una gran número de ortólogos) puede ser una solución, ya que permite localizar la duplicación y la pérdida de genes, reconocer parálogos difíciles de identificar y finalmente coadyuva en la elección de un buen grupo de genes de referencia [28].

I.1.3 Evolución de la interacción leguminosa-rizobios

La simbiosis nódulo-raíz (NR) se pudo derivar de la simbiosis micorrizas-arbusculares (MA), ya que esta es probablemente la interacción entre plantas y bacterias más dispersa, tanto en el contexto ecológico como en el filogenético [29, 30]. Además de ser más antigua que la simbiosis RN, (aproximadamente en 400 millones de años, en el periodo Devoniano [31]), la simbiosis MA se genera a través de estructuras altamente ramificadas dentro de las células de la raíz, también llamados arbúsculos [32], mientras que la simbiosis NR implica la morfogénesis de nódulos (descrito anteriormente). Estas estructuras especializadas (nódulos) probablemente sean el resultado de la coevolución de plantas y rizobios ancestrales, es decir, procesos evolutivos que llevaron a la radiación y divergencia de las leguminosas y de los rizobios. Por ende la planta seleccionó a la bacteria constantemente y contribuyó ligeramente en la evolución de ésta [33]. Por ejemplo, en las poblaciones de *Phaseolus vulgaris,* que comprenden tres grandes pozas génicas distribuidas geográficamente a lo largo del continente Americano (la primera se distribuye en México, Centro América y Colombia, la segunda en el sur de los Andes y la tercera en Ecuador y norte de Perú) [34-39], se encontró una nodulación preferencial del frijol hacia linajes de *Rhizobium etli* relacionados geográficamente mediante alelos del gen *nodC* y además dicha nodulación fue independiente del tipo de suelo utilizado [40]. Sin embargo, *Rhizobium etli* no es el único simbionte del frijol, ya que existen otros linajes de rizobios que lo nodulan, por ejemplo *Rhizobium gallicum [41]*, *Rhizobium*

tropici [42], y *Rhizobium giardinii* [43], entre otras especies. Este reemplazo de simbiontes, se pudo haber debido a la gran plasticidad genética bacteriana, que resultó en una gran capacidad de *Rhizobium* para adaptarse a las leguminosas [33].

La coevolución puede ser detectada (en caso de endosimbiontes), mediante la congruencia y reflejo de las filogenias de ambos socios, tanto del huésped como del endosimbionte bacteriano [44], pero algunas veces esta congruencia filogenética puede deberse a una historia biogeografía común [45]. En contraparte, la transferencia horizontal de genes que permiten realizar la simbiosis bacteriana entre huéspedes individuales sin relación alguna, pueden distorsionar la señal de la coevolución, ejemplos de éstos huéspedes son: plantas de diferentes especies cómo maíz o fríjol, así como la importación de una planta de fríjol de otra región biogeográfica [46]. En el caso de *Rhizobium,* la transferencia lateral de genes *nod* permite la creación de nuevos simbiontes, que no han coevolucionado con sus leguminosas huésped [47, 48] pero si contienen información simbiótica. Por lo tanto, la transferencia lateral de genes y la recombinación pueden ser vistas también como precursoras de la coevolución planta-bacteria [49].

I.2 Recombinación

I.2.1 Recombinación homóloga

El intercambio genético, anteriormente visto como algo poco común en la reproducción asexual bacteriana, ahora se reconoce como una importante fuerza

en la evolución en la mayoría de especies bacterianas [50]. De hecho la carencia de intercambio genético entre las bacterias está confinado a pocos linajes patógenos genéticamente monomórficos [51]. Este intercambio incluye la adquisición, la perdida, el reordenamiento y la transferencia de genes [52]. La transmisión de genes entre y dentro de poblaciones bacterianas ocurre por medio de tres mecanismos: la conjugación, la transducción y la transformación [53]. Estos procesos juegan un papel importante para generar fuentes de variación, así como la modificación de genomas en muchas especies [54-57]. Mediante eventos de recombinación ilegitima (aquellos no mediados por RecA), las especies bacterias puede obtener genes cromosomales y grupos de genes (islas), así como genes plasmídicos, transposones y profagos, los cuales han sobrevivido con éxito los rigores de la selección natural en otra población [58].

Aparte de la recombinación ilegitima, existe evidencia convincente que la recombinación homóloga (mediada por RecA), con frecuencia intercambia pequeñas regiones genómicas entre miembros de una misma población (especie) o entre especies estrechamente relacionadas [59] y muchas veces estos pequeños fragmentos pueden tener un origen ilegitimo (resultado de recombinación ilegitima) y por lo consiguiente, ser dispersados dentro de la población vía recombinación homóloga [60, 61].

I.2.2 Dinámica de la recombinación en la evolución

Existe gran variación en las tasas de recombinación homóloga entre las bacterias y, en algunos casos, la recombinación puede contribuir más que las mutaciones puntuales a la diversificación de especies, mientras que, en otras especies, la recombinación parece ser rara [62]. Por otra parte, la recombinación ilegitima brinda una fuente de variación para la evolución adaptativa en muchas especies de bacterias, por ejemplo, la expansión de nicho, la resistencia a antibióticos, la virulencia, la resistencia al estrés ambiental y otras características. Mientras que la recombinación homologa bajo ciertas condiciones puede promover la tasa de evolución adaptativa dentro de algunas poblaciones [58]. Por lo tanto, las variaciones en las frecuencias de la recombinación homóloga y la recombinación ilegitima podrían explicar la mayoría de variación genética que se encuentra en las poblaciones bacterianas.

En algunas especies como *E. coli*, un importante componente de diversidad genética se encontró dentro de los genes. Esta variabilidad está compuesta por pequeños segmentos de DNA altamente variables que posiblemente se generaron por recombinación [63]. Esta diversidad contribuye a la variación fenotípica y dependiendo de la tasa de recombinación, un particular conjunto genes (asociados a un cierto genotipo) puede ser predominante. Sin embargo, los genotipos pueden aparecer o perderse por azar dentro de una población, donde el nivel de la recombinación es bajo y el tamaño poblacional pequeño [64]. Contario a esto,

cuando la tasa de recombinación es alta, los genotipos ya no podrían divergir, ya que serían constantemente reabsorbidos en la población general, por la acción cohesiva de la recombinación [65]. Por otra parte, la intensidad en la cual la recombinación afecta la estructura de la población está en función de la divergencia genética [66], es decir, a mayor divergencia menor recombinación. En organismos muy relacionados es difícil discriminar entre recombinación homologa y la recombinación ilegitima y más aún, discriminar la contribución relativa de éstas en la diversificación bacteriana.

I.2.3 Detección de la recombinación

La recombinación puede detectarse a través de marcadores genéticos, Electroforesis de Enzimas Multiloci (*MLEE*, por sus siglas en inglés), Polimofismos en longitud de los fragmentos amplificados (*AFLP*, por sus siglas en inglés) y Análisis de secuencia de multiloci (*MLST*, por sus siglas en inglés) [67, 68]. Estos métodos generalmente examinan genes de mantenimiento central para analizar la disminución de la asociación entre alelos de diferentes *loci*, es decir, se mide el desequilibrio de ligamiento (LD). Por ejemplo, para cepas de *E. coli* se ha podido demostrar un fuerte LD, reflejando infrecuente mezcla genética dentro de poblaciones locales [69]. Hoy en día, la disponibilidad de secuencias de genomas completos permite calcular y recalcular las frecuencias de recombinación. Esto tiene el potencial para revelar mucho más que los análisis con siete genes (*MLST*), porque un número limitado de genes de mantenimiento central no pueden

ser representativos del genoma completo [70]. Sin embargo, el número de genomas secuenciados de una especie generalmente varía entre 5 a 20 genomas y frecuentemente no son representativos de una población bacteriana, a excepción de *E. coli*, de la cual hay más de 30 genomas secuenciados, tanto completos como incompletos (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). En *E. coli* se encontró que las tasas de recombinación son mayores que las mutaciones puntuales, pero no lo suficiente para distorsionar la señal filogenética, es decir, incongruencia filogenética entre los distintos marcadores. A pesar de la frecuente recombinación entre las cepas, los genes puede coexistir en un genoma organizado, lo que resulta en una plasticidad cromosomal que puede acelerar la adaptación de *E. coli* a una variedad de ambientes [71]. Desafortunadamente, estos análisis no son aplicables para todas las especies, ya que el promedio de genomas secuenciados por especie es menor de cinco, por lo tanto, una propuesta genómica auxiliada con análisis de genética de poblaciones podría brindar una mejor resolución en la determinación de las frecuencias de recombinación.

I.3 Especie Genómica

Los agrupamientos (*clusters*) obtenidos en función de la secuencia completa de genomas y del contenido de genes, se da principalmente en patógenos obligatorios (los cuales son preferentemente secuenciados), los cuales tienen un nicho ecológico muy restringido y conforme este nicho se expande o se

sobrelapan, existe una disminución de las identidades nucleotídicas promedios y por lo tanto, la coherencia en los agrupamientos disminuye [72]. Esta disminución de la coherencia podría ser provocada por el contenido génico único, es decir, la cantidad de DNA que no se comparte con ninguna otra cepa, por ejemplo, aproximadamente el 30 % de genes en *E. coli O157:H7* (un patógeno peligroso) son únicos comparados con la cepa K12 [73]. Sin embargo, al aumentar la muestra a 20 genomas de *E. coli*, los "*backbone*" o bloques conservados entre todos los genomas, son 98.3 % idénticos a nivel de secuencia [71] y los genes que no están en los *backbone* están en un constante flujo resultado de eventos de recombinación. Esta dinámica en el contenido de genes da como resultado una plasticidad genómica dentro de una especie, también llamada "especie genómica" o "pan-genoma", la cual consta de tres componentes: un genoma core (nucleo), que incluye los genes de mantenimiento y metabolismo central, así como aquellos responsables para la transcripción, translación, replicación y homeostasis de energía; un genoma de carácter constituido por genes que intervienen en la adaptación a un nicho ambiental particular y finalmente el genoma accesorio, que lo conforman los genes cepa-especifica y cuya función a menudo se desconoce [74].

El genoma accesorio parece ser un componente importante en algunas especies: en el caso de *Streptococcus,* el análisis del pan-genoma se llevo a nivel de género, partiendo de seis especies (26 genomas) y se pudo determinar que su pan-genoma consiste en 6000 genes, así como de un genoma core formado por

600 genes, y un conjunto de genes accesorios, el cual conforma un porcentaje importante de contenido génico, además de ser muy variable. Por otra parte, se encontró evidencia de altos niveles de recombinación en el genoma core, así como diferentes tamaños de pan-genoma en cada especie [75], esto podría sugerir que los tamaños de pan-genomas y las tasas de recombinación probablemente reflejen diferencias en nichos y estilos de vida de diferentes especies. Por ejemplo, se necesitaría grandes pan-genomas para especies que habitan en una gran diversidad de nichos. La asociación entre huéspedes cercanos puede generar diversidad a través de la recombinación entre compañeros de colonización de un nicho en particular [76].

CAPITULO II.- ANTECEDENTES

II.1 *Rhizobium etli*

R. etli es el simbionte más común encontrado en los nódulos de *P. vulgaris*, tanto para las variedades silvestres como para las cultivables en aéreas nativas (no cultivables) [77-79], además de ser la especie más competitiva para nodular al frijol [80]. Esta capacidad de nodulación esta codificada en los genes simbióticos, los cuales se encuentran en plásmidos o en islas simbióticas cromosomales. Generalmente, estos genes simbióticos tienen historias evolutivas diferentes a los genes cromosomales [81-84], debido en parte, a la transferencia lateral, la recombinación y los rearreglos en los plásmidos. Este flujo de genes dentro de los plasmidos genera estructuras de mosaicos, las cuales contienen numerosos elementos móviles, familias de DNA repetidas y carecen de sintenia [85]. Al comparar seis regiones ortólogas de aproximadamente 5kb del plásmido simbiótico de diferentes cepas (distinto origen geográfico) de *R. etli* [86], se identificaron regiones altamente polimórficas distribuidas asimétricamente entre el plásmido simbiótico. Dichas regiones presentan el mismo tipo de perfil de polimorfismos entre dos o más cepas, sugiriendo que las mutaciones se distribuyen principalmente por recombinación. Un panorama similar se observó en aislados de un mismo sitio geográfico de cepas de *Sinorhizobium medicae*, donde la recombinación fue mayor en los plásmidos y en los megaplásmidos en comparación al cromosoma [87].

II. 2 Genoma de *R. etli*

El genoma de *R. etli CFN42*, está compuesto de seis grandes plásmidos (p42a, p42b, p42c, p42d, p42e y p42f) que en conjunto comprenden un tercio del total de genoma de esta cepa. El cromosoma codifica para la mayoría de funciones necesarias para que la célula crezca, mientras que pocos genes esenciales o vías metabólicas completas están localizados en los plásmidos. Tal es el caso del plásmido p42e, que contiene genes que codifican para funciones esenciales como: metabolismo primario, funciones de biosíntesis y degradación [88]. La sintenia cromosomal la interrumpen algunos genes relacionados a secuencias de inserción, fagos, plásmidos y componentes de superficie celular. Por otra parte, los plásmidos no muestran sintenia y comparte sus ortólogos con otros replicones accesorios de especies cercanas con genomas multiparticionados [89].

El plásmido simbiótico (p42d), es una molécula circular de 371 kilobases que contiene a la mayoría de los genes encargados de la nodulación y de la fijación de nitrógeno, en una región de 124 kb. Este plásmido posee numerosas secuencias relacionadas a elementos móviles, dispersos a lo largo de este. En algunos casos los elementos móviles están flanqueados por bloques de secuencias con funciones relacionadas, estos pueden indicar un rol de transposición. De hecho, el plásmido simbiótico frecuentemente esta tiroleado por eventos de recombinación y transferencia lateral de genes, dando como resultado

una estructura de mosaico [90]. Existe evidencia que sugiere que los plásmidos simbiótico y el p42a tienen un origen externo [91]. Finalmente, una característica importante de los genomas de *R. etli*, es su redundancia metabólica y en el caso de *CFN42* consta de: 23 genes que codifican factores sigma putativos, numerosos genes que codifican isoenzimas y un gran número de familias de parálogos, todos estos elementos generan una plasticidad genoma necesaria tanto para el estilo de vida libre como para el simbiótico.

II.3 Comparaciones genómicas entre especies de *Rizhobium*

Crossman, L. C., S. Castillo-Ramírez, et al. (2008). "A common genomic framework for a diverse assembly of plasmids in the symbiotic nitrogen fixing bacteria." PLoS One 3(7): e2567. (Articulo anexado al apéndice)

La estructura multipartita del genoma de *R. etli CFN42*, es decir compuesta de un cromosoma y varios otros replicones de alto peso molecular, lo comparten 3 especies más: *Sinorizhobium meliloti*, *Agrobacterium tumefaciens* y *Rizhobium leguminosarum* [92]. Estos cuatro genomas comparten aproximadamente un genoma core (definido previamente) de 2,000 familias de genes y aproximadamente 300 de ellas son endémicas de cada genoma. En lo que respecta a las comparaciones pareadas entre los cuatro genomas, la comparación de *R. etli (CFN42)* con *R. leguminosarum biovar viciae (3841),* presentó un genoma core estable, más un componente accesorio variable. Los cromosomas de ambas cepas presentan una alta sintenia, mientras que los plásmidos se relaciona entre sí por pocos bloques sinténicos. Los pares de plásmidos que

presentan mayor similitud son las parejas p42f-pRL12, p42e-pRL11 y p42b-pRL9. Cabe destacar la extensiva región sinténica común entre p42c-pRL10 (plásmidos simbióticos), la cual representa aproximadamente el 59% de longitud de p42c. Por lo tanto, ya sea que pRL10 haya ganado una larga inserción que incluya las funciones de nodulación y fijación de nitrógeno, o que p42c haya sufrido una gran deleción de estos genes, ya que la región *nif-nod* puede representar un potencial cassette simbiótico bordeado por secuencias repetidas. Además, las diferencias estructurales entre los plásmidos simbióticos de *R. etli* (p42d) y *R. leguminosarum* (pRL10) sugieren que tienen caminos evolutivos distintos. Los plásmidos de *R. etli* son más pequeños que los de *R. leguminosarum* y aproximadamente del 44% a 58% de su longitud están organizados en bloques sinténicos con respecto a *R. leguminosarum*. Por otra parte, los plásmidos p42a, p42d, pRL7 y pRL8 se aprecian como externos o atípicos con respecto al resto de los replicones que comparten muchas regiones comunes y estas podrían ser parte de un replicón ancestral. No obstante la relación filogenética entre plásmidos permanece obscura. Un mejor panorama de la evolución de los genomas particionados se podría llevar a cabo con la comparación de un conjunto de plásmidos de cepas adicionales y representativas de *R. etli* y de *R. leguminosarum*.

II.4 Comparaciones genómicas dentro de la especie de *R. etli*

González, V., J. L. Acosta, et al. (2010). "Conserved symbiotic plasmid DNA sequences in the multireplicon pangenomic structure of Rhizobium etli." Appl Environ Microbiol **76**(5): 1604-1614. (Articulo anexado al apéndice)

Al extender las comparaciones con otro genoma completo *(CIAT652)* y seis genomas incompletos *(BRASIL5, CIAT894, GR56, IE4771, KIM5 Y C3-8)* de distinto origen geográfico y de aproximadamente una cobertura 1X de *Rizhobium etli* [93], se observó que la mayoría de los bloques conservados están localizados dentro del cromosoma y que algunas regiones dentro de los plásmidos son comunes entre las especies (Figura 1), por ejemplo, los plásmidos pA (*CIAT652*), p42e (*CFN42*) y pRL11 (*R. leguminosarum 3481*) comparten gran número de familias de genes, otro ejemplo son los plásmidos pC (*CIAT652*), p42f (*CFN42*) y pRL12 (*R. leguminosarum 3481*). En lo que respecta al plásmido simbiótico, existe una región compartida (bloques sinténicos) entre los plásmidos pB, p42d y pRL10 de los genomas *CIAT652*, *CFN42* y *R. leguminosarum,* respectivamente. Sin embargo, también hay pérdidas dentro de las familias de genes (Figura 1), dando como resultado bloques conservados solamente en dos genomas, por ejemplo, los plásmidos p42c (*CFN42*) y pRL10 (*R. leguminosarum*). El caso particular del plásmido pC de *CIAT652*, que comparte varios bloques conservados con varios replicones de *CFN42* y *R. leguminosarum*, es un ejemplo de fusiones o deleciones de bloques conservados entre los plásmidos de *Rhizobium*. Los bloques cromosomales sinténicos incluyeron un total de 3,971 familias (tanto unicopia, como multicopias) de genes que se comparten entre los genomas completos (*CFN42, CIAT652 y R. leguminosarum biovar viciae 3841*). A pesar del gran número de familias ortólogas, también existen una cantidad considerable de

familias que son únicas de cada genoma completo: 698 (*CFN42*), 994 (*CIAT652*) y 2,071 (*R. leguminosarum*).

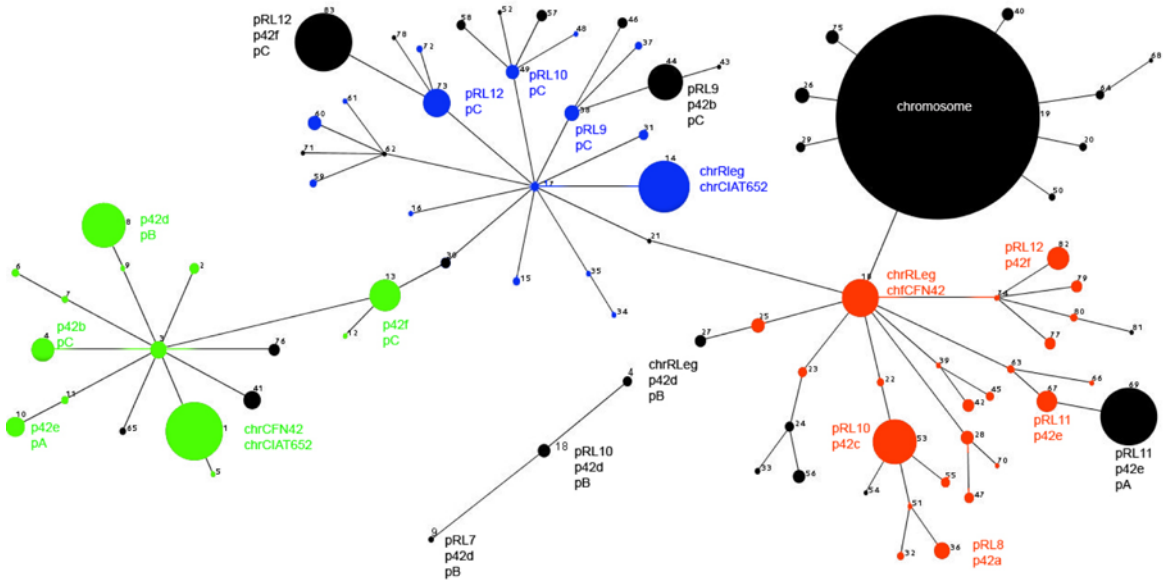

Figura 1. Perfiles de familias de proteínas y su distibución en replicones de *Rhizobium*. Las familias de proteínas de un sólo miembro por cada genoma (*CFN42*, *CIAT652* y *R. leguminosarum 3841*) fueron utilizadas para construir perfiles de presencia y ausencia para los replicones de los tres genomas (González, Acosta et al. 2010). El tamaño de las esferas representa la frecuencia de los perfiles en un replicón dado. El color negro indica la presencia de la familia en los tres replicones. El color verde son las familias de proteínas con una pérdida en el genoma de *R. leguminosarum*, el color naranja indica una pérdida en el genoma de *CIAT652* y finalmente, el color azul una pérdida en *CFN42*.

En el caso de los plásmidos simbióticos (tanto *CFN42* y *CIAT652*), las regiones compartidas (regiones homólogas) entre los genomas incompletos presentaron una identidad nucleotídica del 92 a 98 %, salvo la cepa *IE4771* (89 %). En contraparte, de 400 a 750 kb de los genomas incompletos no es compartido por ningún otro genoma de *R. etli*, ni por ninguna otra secuencia de la base de datos no redundante (obtenida del NCBI ftp://ftp.ncbi.nih.gov/blast/db/). Esta gran cantidad de DNA endémico, refleja una estructura pangenómica abierta,

es decir, no existe un número determinado de genes para la población (especie), ya que siempre hay la inclusión de nuevos genes (no compartidos con las demás cepas) conforme se añade un nuevo individuo (genoma). Esta estructura pangenómica abierta es mantenida por la aportación de aproximadamente 99 nuevos genes, los cuales brindan una plasticidad genómica, además de ser, materia prima para la transferencia de DNA, la recombinación y la conversión génica, entre y dentro las poblaciones de *R. etli*.

CAPITULO III.- Genomic lineages of *Rhizobium etli* revealed by the extent of nucleotide polymorphisms and low recombination

III.1 Resumen

En este trabajo determinamos y valoramos la variación a nivel nucleotídico en ocho genomas (dos genomas completos y seis genomas parciales) de cepas provenientes de distinto origen geográfico de *R. etli*. Aunado a esto, detectamos y cuantificamos los porcentajes de recombinación homóloga en fragmentos ortólogos y extendimos la búsqueda mediante cuartetos (tres genomas completos y un genoma parcial). Como resultado de estos análisis, identificamos altos niveles de DNA polimorfico en *R. etli* y determinamos una divergencia nucleotídica promedio de 4% a 6% entre los genomas muestreados. Por otra parte, los eventos de recombinación detectada se estimaron dentro de un amplio rango de 3% a 10% y en la mayoría de los casos, la diversidad nucleotídica (π) fue mayor en los cuartetos recombinantes que los no recombinantes. Al valorar la señal filogenética de los segmentos ortólogos, no encontramos incongruencias filogenéticas lo suficientemente fuertes (en frecuencia) para interrumpir la topología del árbol de la muestra. Además de que la mayoría de eventos de recombinación detectada se localizaron en el cromosoma y en menor grado en los plásmidos p42b, p42e y p42f (tomando a *CFN42* como genoma de referencia). Por último, se discuten las implicaciones de haber encontrado bajas frecuencias de recombinación homóloga y una gran variación entre los genomas de *R. etli*.

Acosta et al. BMC Evolutionary Biology 2011, 11:305
http://www.biomedcentral.com/1471-2148/11/305

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Genomic lineages of *Rhizobium etli* revealed by the extent of nucleotide polymorphisms and low recombination

José L Acosta[1*], Luis E Eguiarte[2], Rosa I Santamaría[1], Patricia Bustos[1], Pablo Vinuesa[1], Esperanza Martínez-Romero[1], Guillermo Dávila[1] and Víctor González[1]

Abstract

Background: Most of the DNA variations found in bacterial species are in the form of single nucleotide polymorphisms (SNPs), but there is some debate regarding how much of this variation comes from mutation versus recombination. The nitrogen-fixing symbiotic bacteria *Rhizobium etli* is highly variable in both genomic structure and gene content. However, no previous report has provided a detailed genomic analysis of this variation at nucleotide level or the role of recombination in generating diversity in this bacterium. Here, we compared draft genomic sequences versus complete genomic sequences to obtain reliable measures of genetic diversity and then estimated the role of recombination in the generation of genomic diversity among *Rhizobium etli*.

Results: We identified high levels of DNA polymorphism in *R. etli*, and found that there was an average divergence of 4% to 6% among the tested strain pairs. DNA recombination events were estimated to affect 3% to 10% of the genomic sample analyzed. In most instances, the nucleotide diversity (π) was greater in DNA segments with recombinant events than in non-recombinant segments. However, this degree of recombination was not sufficiently large to disrupt the congruence of the phylogenetic trees, and further evaluation of recombination in strains quartets indicated that the recombination levels in this species are proportionally low.

Conclusion: Our data suggest that *R. etli* is a species composed of separated lineages with low homologous recombination among the strains. Horizontal gene transfer, particularly via the symbiotic plasmid characteristic of this species, seems to play an important role in diversity but the lineages maintain their evolutionary cohesiveness.

Background

Bacterial species typically contain large amounts of genetic variation in the form of single nucleotide polymorphisms (SNPs), which originate by mutation and have dynamics that depend on the balance between natural selection and genetic drift [1,2]. There is some debate on whether or not most of these polymorphisms are selectively neutral at the molecular level [3]. Species have been genetically defined through the analysis of DNA variation using comparative techniques such as hybridization, the sequencing of gene markers, and (more recently) complete genome sequences [4,5]. It has been proposed that similarity values greater than 70% obtained in DNA-DNA hybridization experiments are sufficient to define a coherent group of organisms as belonging to the same species [6]. These estimates are very rough, subject to experimental variation, and they only indirectly measure similarity (i.e. via hybridization efficiency) [7]. A comparative analysis of complete genomes minimizes most of these limitations. Several measures of genomic relatedness, such as the Average Nucleotide Identity (ANI) and the Maximal Unique Matches (MUM) have been proposed for such analyses [8,9]. Both ANI and MUM are based on pairwise nucleotide comparisons of complete genomes, and several reports have shown good correlations between the results from these analyses and other measures of genetic relatedness, such as those based on Multilocus Sequencing Typing (MLST), 16S sequencing, and gene

* Correspondence: jlacosta@ccg.unam.mx
[1]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Av. Universidad N/C Col. Chamilpa, Apdo. Postal 565-A, Cuernavaca 62210, México
Full list of author information is available at the end of the article

content [10]. However, these comparative methods rely on the availability of complete genome sequences and are affected by the quality of the DNA sequencing data, which in the case of draft genomes might not be optimal [10]. The latter issue has not been thoroughly addressed in past studies. One exception was the comparisons made by Richter and Roselló-Mora [10], who suggested that low genome sequence coverage can be sufficient for inferring DNA similarity values comparable to ANI obtained with complete genomes.

Bacterial species have mechanisms for gene exchange (transformation, conjugation and transduction), and genetic recombination is believed to play a prominent role in diversifying species by distributing variation and generating new allele combinations [11]. Horizontal gene transfer is an important source of genomic variation within and between species [12-16], and homologous recombination frequently results in the exchange of small genomic regions between members of the same or closely related species [17]. The estimated rates of homologous recombination vary widely among bacteria; in some instances, recombination seems to have contributed to species diversification to a greater extent than even point mutations, whereas in other species homologous recombination appears to be rare [18].

Recombination has typically been assessed by molecular techniques such as Multilocus Enzyme Electrophoresis (MLEE), Amplified Fragment Length Polymorphism (AFLP), or Multi Locus Sequence Typing MLST [19-21]. These methods primarily measure linkage disequilibrium (LD), and are based on the degree of allele association at different housekeeping loci. For example, *E. coli* strains show strong LD, reflecting infrequent genetic mixing within local populations [22]. More recently, the availability of complete genomic sequences has allowed recombination to be assessed more accurately [23]. Interestingly, genomic sequencing combined with analyses of population genetics have shown that the recombination rates within *E. coli* are higher than the mutation rates, but not to the extent that the phylogenetic signal is distorted [24]. Despite frequent recombination between strains, therefore, the genes seem to coexist in an organized genome, resulting in a chromosomal plasticity that accelerates the adaptation of *E. coli* to various environments.

In this work, we studied the intraspecific variability and recombination in *Rhizobium etli*, a soil bacterium that associates with bean roots to fix nitrogen. Previous studies have noted that this species has a variable gene content and high genomic divergence [24], as well as a low rate of recombination (in housekeeping genes) among isolates from the same geographical site [22,25,26]. However, in isolates (from the same geographical site) of *Sinorhizobium medicae*, it was found that frequency of recombination was higher in plasmids and megaplasmids, as compared to the chromosome [27]. The first purpose of this work was to perform a detailed genomic analysis of the nucleotide variation in this species. Accordingly, we used stringent methods to identify SNPs from a set of complete and draft genomes of *R. etli*, assessed the value of draft genomes and low coverage data when seeking to obtain global measures of genetic relatedness, and then examined the nucleotide differences among various strains of *R. etli*. The second purpose was to assess the role of recombination in generating genomic diversity in *R. etli*. Our results confirm and extend the previous estimations on the genomic diversity of *R. etli*, and indicate that recombination might play only a minor role in generating such diversity. Therefore, we conclude that the species *R. etli* is composed of separate genomic lineages that share a low rate of recombination but have a common symbiotic phenotype.

Results

Nucleotide variation assessment in complete and draft genomes

Since accurate SNP identification relies largely on the quality of the sequence data, the use of draft genome sequences could potentially introduce errors into the variation estimates. Therefore, stringent parameters (see Methods) were used to identify high-quality SNPs in a set of two complete *R. etli* genomes, CFN42 and CIAT652, isolated from México and Costa Rica respectively, and six draft genome sequences from strains isolated in different places of the world: BRASIL5 (Brazil), CIAT894 (Colombia), GR56, IE4771 (México), KIM5 (USA), and 8C-3 and GR56 (Spain) [24]. All the Sanger reads were collected from the draft genomes (about 13,000 reads of 1000 nucleotides in length per genome on average) were aligned against the predicted ORFs of the CFN42 or CIAT652 genomes, and the alignments were evaluated using Polybayes (additional file 1 Figure S1), which determined the probability that a nucleotide site was polymorphic, based on the Phred quality of the read. A Phred value of Q20 and a probability greater than 0.90 are generally considered acceptable for the detection of SNPs [28]. Most of the SNPs in our data set had probability scores > 0.975, indicating that more than 100,000 SNPs per genome had Phred qualities over Q45 (additional file 1 Figure S1). To avoid the possible inclusion of false positives (in average 27,000 SNPs by each strain), we used only SNPs with a minimum Phred score of Q45 and the highest Bayesian probabilities (> 0.99) throughout this work [29].

Additional errors in SNP determination might arise from poorly aligned regions. Since *R. etli* genomes have a high proportion of paralogous sequences [24,30], a

stringent identification of orthologous segments of genes was performed. We aligned the contigs of each draft genome sequence against the ORFs from the complete genomes of either CFN42 or CIAT652, using both ungapped and gapped alignments, along with the reciprocal best hit criteria. We considered DNA gene segments as being orthologous to the reference sequence if they had nucleotide identities higher than 85% and coverage higher than 60% of the reference gene. Various numbers of orthologous segments were identified from the draft genomes, covering about 40% of the total gene contents of the reference strains. The total amount of data collected by this procedure is about 2 to 2.5 Mb per draft genome (additional file 1 Table S1).

To determine the robustness of the above-described procedure, we simulated a draft assembly by using Sanger read samples of the complete genomes of different *E. coli* strains at low coverage (1x) (see Methods). The contigs of the simulated assembly were aligned with the genome of *E. coli* K12, and SNPs were detected as described above. On average, the obtained nucleotide variation ranged from about 1% to 2% (SNPs/alignment length) (Figure 1). There was no significant difference (p-value lower at 0.05, according to Mann-Whitney and Kolmogorov-Smirnov tests obtained from Predictive Analytics Software PASW Statistics 18 (SPSS Inc., Chicago, IL)) when we compared the results obtained at 1× coverage versus those obtained with the complete genome assembled at about 10× coverage, indicating that 1× coverage of the genome sequence could be considered a robust proxy of full variation at the genomic level in this species.

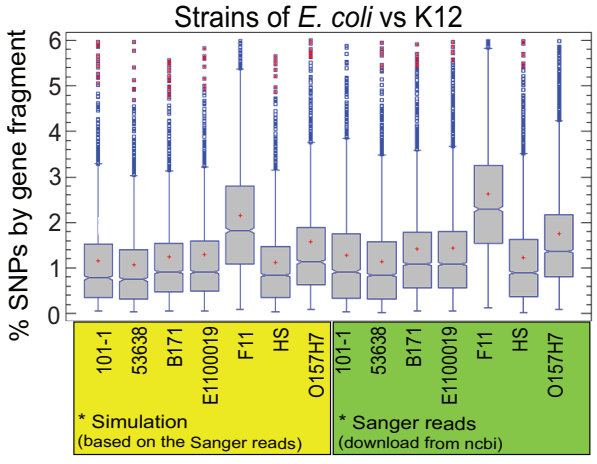

Figure 1 SNP assessment on lower coverage. Paired comparisons between *E. coli* strains and *K12* strain as reference genome. We used two set of *E. coli* strains: Complete genomes obtained from Sanger reads (green box) and Sanger reads simulated from complete genomes with coverage approximate of 1X (yellow box). For each alignment, we determined the percentage of SNPs by gene fragment (SNP number/length of contig; Y axis) by our methodology. Boxes inside the graphic include the median values (middle line) and the first and third quartiles (lower and upper lines) of the distribution. Abscissa: *E. coli* strains.

SNP frequencies among the *R. etli* strains

We quantified the SNPs in *R. etli* by computing the pairwise nucleotide differences between individual draft genomes versus the complete genomes of strains CFN42 or CIAT652. More SNPs were found in comparisons made versus the CFN42 genome (Figure 2, gray boxes) than the CIAT652 genome (Figure 2, blue boxes). For example, the BRASIL5 strain had a median of 5% SNPs per aligned fragment when compared with CFN42 but only 2% compared to CIAT652, indicating that BRASIL5 is more closely related to CIAT652 than CFN42. Similarly, variance was higher when BRASIL5 was compared with CFN42 rather than CIAT652 (Figure 2). A very similar pattern was found for strain 8C-3. The other strains showed similar levels of variation, on the order of 6% (CFN42) and 4% (CIAT652), with the latter comparison always showing a lower variance. Comparison between the complete genomes of CFN42 and CIAT652 (Figure 2, red box) result in a median variation of 9%, that is high but still lower than the comparisons between CFN42 and *R. leguminosarum* bv *viciae* 3841 (Figure 2 green box). Moreover, when we compared *R. leguminosarum bv. viciae* 3841 with all of the *R. etli* strains (complete and draft genomes) (additional file 1

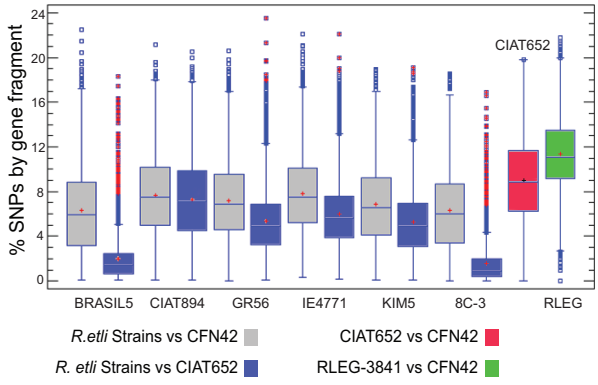

Figure 2 Paired comparisons between *R. etli* strains. We performed four paired comparisons used our methodology: draft genomes of *R. etli* against CFN42 (gray boxes), draft genomes of *R. etli* against CIAT652 (blue boxes), CIAT652 against CFN42 (red boxes) and finally *R. leguminosarum bv viciae 3841* against CFN42 (green boxes). For all comparisons the Y axis is the percentage of SNPs by gene fragment (SNP number/length of contig). Boxes inside the graphic include the median values (middle line) and the first and third quartiles (lower and upper lines) of the distribution. Abscissa: *R. etli* and *R. leguminosarum bv viciae 3841* strains.

Figure S2), the greatest difference in SNP percentage (median 11%) was seen in the comparison with strain CFN42 (Figure 2 green boxes, and discussion section).

Average nucleotide variation

We sought to obtain a single measure of the nucleotide variation across the whole set of genomes. To this end, we averaged the medians of the SNP distributions for each alignment (*i.e.*, the number of SNPs/alignment length of each draft genome with respect to CFN42 or CIAT652) and generated average confidence interval (obtained and adjusted by distribution of genes size medians) using Predictive Analytics Software PASW Statistics 18 (SPSS Inc., Chicago, IL). This statistical test of proportions compares the observed proportions of an event (here, SNPs) in k samples (here, strains), uses a chi-squared test to seek significant differences among the proportions, and subsequently adjusts the confidence intervals for each sample. The generated measure, herein called the average nucleotide variation (ANV), might represent the species-level variation. We obtained ANV values of 4% and 6% when we compared all the analyzed strains against CIAT652 and CFN42, respectively (Figure 3). Although the largest numbers of SNPs were found in comparisons with the CFN42 genome, all strains were similarly divergent according to the 95% confidence intervals with respect to the median (blue lines in Figure 3). This observation indicates that CFN42 is almost equally divergent with respect to all other strains. Comparisons with the CIAT652 genome showed that strains BRASIL5 and 8C-3 were closer to this strain than to CFN42. Moreover, the CIAT894 strain yielded the highest number of SNPs, causing its average SNP proportion to fall outside the average confidence interval (red lines in Figure 3). Strains CIAT894 and IE4771 showed greater divergences than the rest of the strains, regardless of the reference strain (CFN42 or CIAT652) used in the comparison.

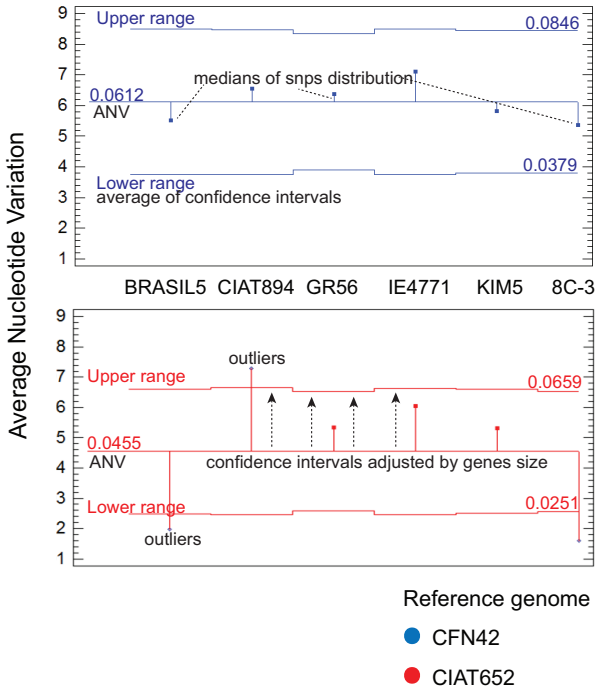

Figure 3 Average nucleotide variation. We calculated the Average nucleotide variation (middle lines of each graphic) from the median SNP percentages (dots indicates by dashes lines) for each aligned comparison (Y axis) of test strain versus the reference strains, CFN42 (blue) or CIAT652 (red). Average Confidence interval was adjusted (arrows with dashes lines) to the medians of the length distributions of the aligned fragments (genes). The medians SNPs that exceed the average confidence interval were outliers. Abscissa: BRASIL5, CIAT894, GR56, IE4771, KIM5 and 8C-3.

Nucleotide variation profiles in homologous genomic segments from different *R. etli* strains

To explore how SNPs are distributed in the *R. etli* genomes, we first identified orthologous segments for which we had sequence information in all eight studied strains (Figure 4). A total of 240 segments with a median size of 275 bp were common to all strains, and spanned a total of about 71,630 bp that represent about 1% of the genome length. These sequences mapped mainly to the chromosomes of CFN42 and CIAT652 (92%), with a lower proportion (8%) distributing to plasmids. We generated a concatenated alignment of these shared segments according to the gene order found in the CFN42 genome, and then inferred a consensus sequence and computed the number of nucleotide differences across windows of 250 bp. Using this procedure, we detected the patterns of shared and unique (singleton) SNPs particular to each strain. As shown in Figure 4, we were able to distinguish two classes of shared SNPs: biallelic SNPs (Figure 4 gray smoothed areas), which showed only one nucleotide difference with respect to the consensus; and polyallelic (Figure 4, white bars), which showed multiple differences at the same nucleotide site with respect to the consensus. Some of these SNP patterns were shared in some strains but not others. For example, as shown in Figure 4, pattern A was shared by strains CIAT652, CIAT894 and 8C-3, whereas pattern B was found in strains GR56, IE4771 and Kim5. Further shared patterns were identified through a careful inspection of the plot. In addition, a large number of polymorphisms were not shared, but instead appeared to be strain-specific variants. Interestingly, strain CFN42 was found to have the greatest number of differences with respect to the consensus (Figure 4, black bars). Even thought this approach is limited by the amount of common segments among the eight strains, we were able to cover 3.7% (223) of the total gene content (5,963) of the

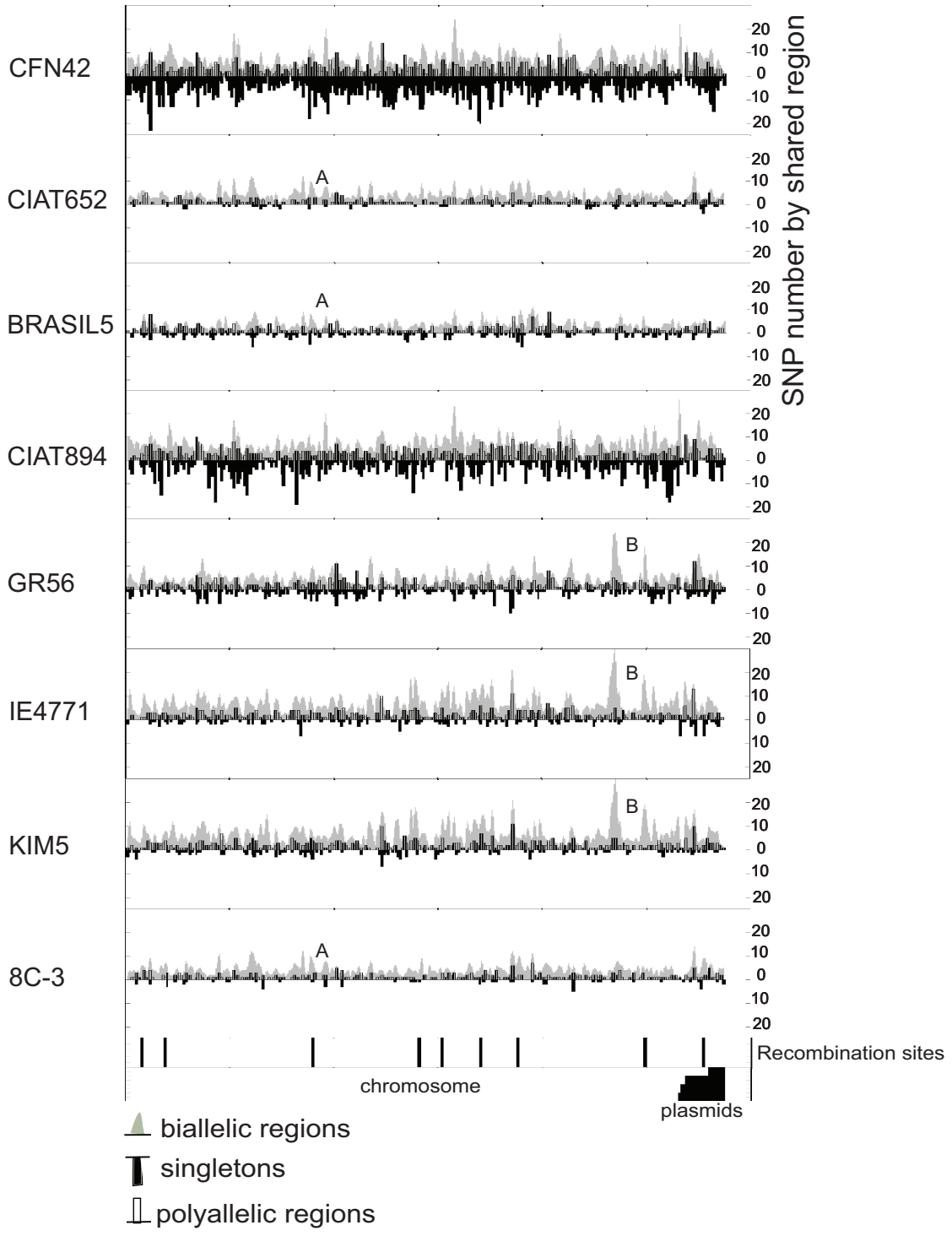

Figure 4 SNP distribution profiles. Alignments were performed on a total of 240 sequence segments available for all tested strains of *Rhizobium etli*. Each nucleotide position in the alignment is represented by a consensus. In instances where half of the strains had the same nucleotide and the other half a different nucleotide, the consensus was defined as the nucleotide present in *R. etli CIAT652*. Common segments were concatenated according to the gene order found in the CFN42 genome (chromosome and after plasmids), yielding 71,630 aligned base pairs. The numbers of nucleotides differing from the consensus are plotted as bars, across independent windows of 250 nucleotides. The black bars (running downwards) show SNPs present in a single strain; the gray areas indicate when the same SNP pattern was present in at least two strains at the same position within the alignment (patterns A and B); and the white bars indicate polymorphic sites where at least three alleles were present in at least two strains, again within the alignment. Segments showing significant recombination events are indicated by bars at the bottom of the plot, and with bars indicating the genomic location of segments with respect to CFN42 (chromosome, white; plasmids, black).

CFN42 reference strain that include the main COG categories and subcategories (see Methods). For instance, metabolism (transport and metabolism of sugar, amino acids, and carbohydrates); cellular processes and signaling (envelope biogenesis, signal transduction); information storage and processing (transcription, replication, and recombination); and poorly characterized proteins (function unknown). A detailed annotation of the gene segments can be seen in additional file 2 Table S1.

Phylogenetic congruence

Since recombination can distort phylogenetic trees such a way that no two individual trees are topologically equivalent, we decided to perform phylogenetic reconstructions using a) a neighbor-joining network [31]; and b) a comparison of a consensus tree with individual trees constructed using the 187 segments common to the eight studied *R. etli* genomes and *R. leguminosarum bv viciae 3841* (RLEG). The consensus trees obtained from the concatenated alignments had identical topologies when constructed by maximum likelihood, Bayesian, and neighbor joining network methods (see Methods). Only the tree based on neighbor joining network is shown in Figure 5. This tree was found to contain six internal branches (denoted by split numbers). There are two main clusters in the tree, separated by branches 2 and 3 that group the most closely related strains: one containing KIM5, IE4771, and GR56 (branch 2) and another grouping BRASIL5, 8C3, and CIAT652 (branch 3). These branches are internal in relation to branch 5, which separates CFN42, CIAT894, and RLEG that are the strains with the longest branches (greatest number of nucleotide substitution per site). A few inconsistencies were found among the topologies recovered from reconstructions based on individual gene segments (187), as compared to the topology of the consensus tree (not shown). These alternative topologies are mainly due to the position of CIAT894 and RLEG, whereas the splits 2, 3, and 5 where consistently recovered. Thirty out of 187 trees supported the placement of RLEG as the most distant strain, 39 trees supported placement of CIAT894 as the external strain, whereas the most frequent topology shows that these strains are equally distant to the rest of strains (Figure 5). These alternative topologies could be the result of shared ancestral polymorphisms, as suggested by the long branches coupled with low frequency of recombination. Altogether, the phylogenetic reconstructions suggested that the levels of recombination were insufficient to erase the phylogenetic signal, thus allowing for the identification of the most probable strain tree. Consistent with this conclusion, only nine (3.75%) of the 223 gene segments common among the eight *R. etli* strains (Figure 4) showed at least one recombination event.

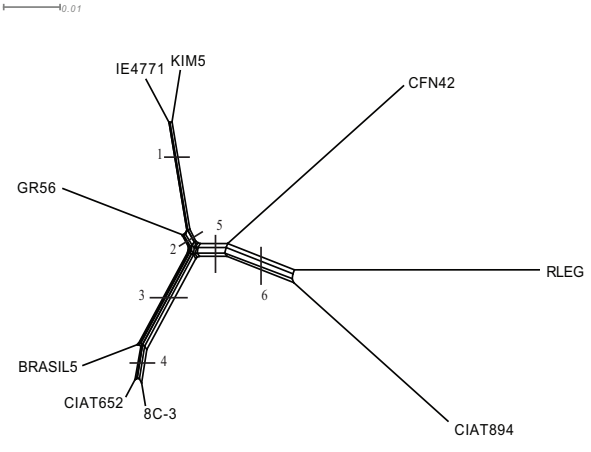

Figure 5 Genetic relatedness. Network joining network phylogeny inferred from 186 concatenated regions shared among strains of *Rhizobium etli* and *Rhizobium leguminosarum bv viciae* 3841 (see Methods). The tree is unrooted and has six internal branches, indicated by split numbers on each internal branch. The scale bar denotes the expected number of nucleotide substitutions per site.

Extent of recombination

To evaluate the extent of the probable recombination events among strains of *R. etli*, we performed a recombination analysis in orthologous quartets (see Methods). We aligned the shared gene segments from each draft genome with the corresponding segments of the ORFs from CFN42, CIAT652, and the *R. leguminosarum bv viciae 3841* complete genomes, yielding six different groups of quartets (one group for each incomplete genome; Figure 6). The proportion of aligned segments varied across the six groups of quartets, from ~2,781 segments in the group containing BRASIL5, to ~3,672 in the group containing CIAT894. The segments ranged from 200 to 4651 bp in length and covering approximately 50% of the genome (additional file 1 Table S2). For each group of quartets, we performed four different recombination tests (see Methods), and determined the number of recombination events (only those that were detected by at least two methods) for each quartet (describe above) (Figure 6). The lowest proportions of recombination events were detected for the quartets containing strains BRASIL5 and 8C-3, which showed 4.42% (123 out 2781) and 3.57% (102 out 2854) recombination events, respectively. The other groups showed approximately twice as many recombination events, with frequencies ranging from 8.67% (KIM5 quartets) to 10.86% (GR56). In addition, for each group of recombinant quartets, we determined the number of events of recombination between pairs of strains (Figure 6). In

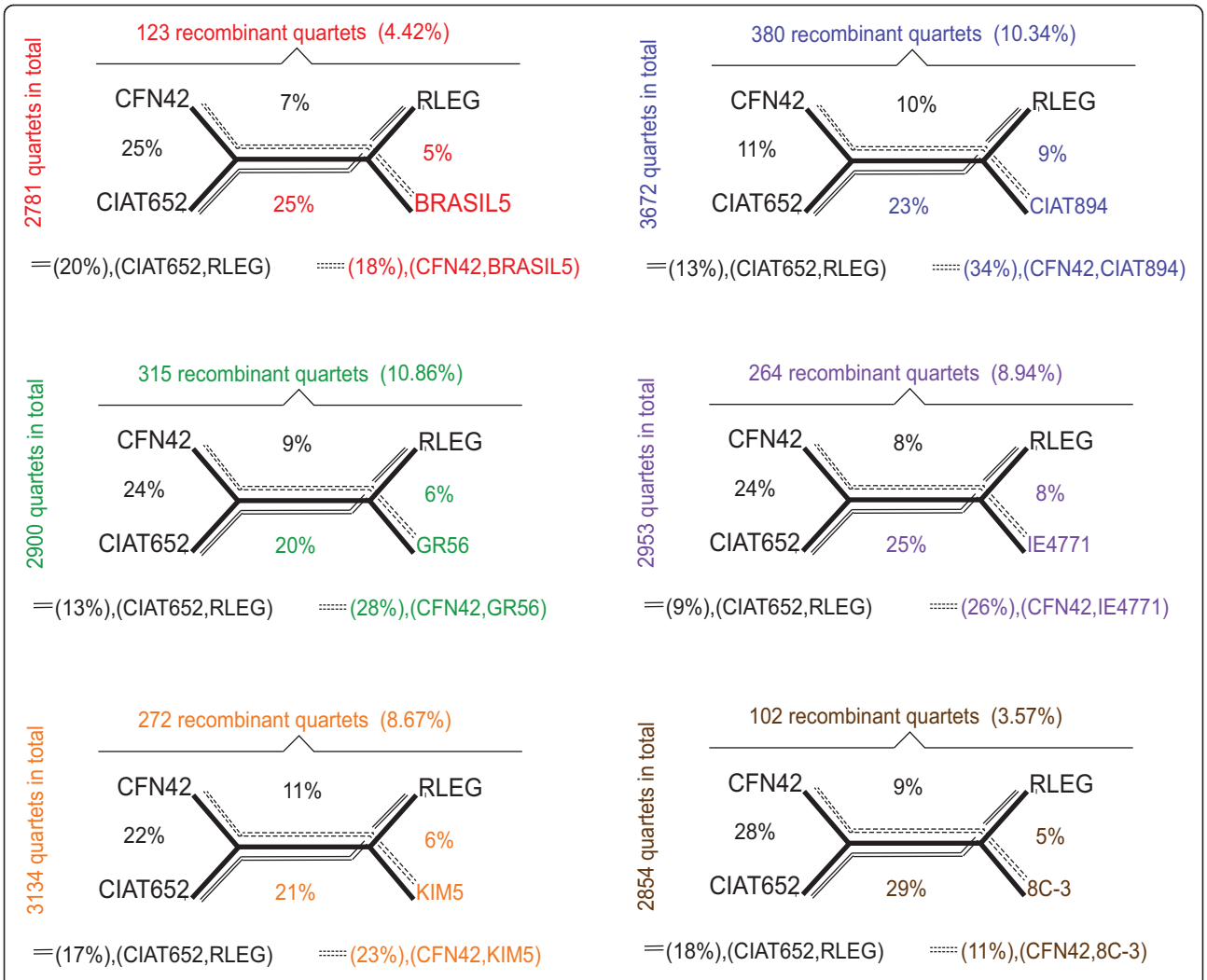

Figure 6 Detection of recombination in quartets. Six groups of quartets of orthologous segments were built with the Mauve program; they included shared sequences from CFN42, CIAT652, *Rhizobium leguminosarum bv viciae 3841*, and each one of the six test strains (incomplete genomes). Test of recombination were performed for all the quartets as described in Methods section. The number (as well as the percentage) of recombinant segments predicted for each group of quartets is indicated above the quartet diagram. The total number of quartets analyzed in each group is indicated at the left side of the diagrams. The percentages of recombinant quartets between pairs of strains are shown inside the diagram and by dashed and continuous lines defined below the diagram. To facilitate searching each test strains o incomplete genome has its own color.

general, recombination events were more frequently predicted between *R. etli* strains pairs than between any given *R. etli* strain and *R. leguminosarum bv viciae 3841* (Figure 6). For instance, in the group of quartets containing BRASIL5, the percentage of recombinant segments is about 7% in CFN42-RLEG, 5% in BRASIL5-RLEG, and 20% in CIAT652-RLEG pairs, whereas recombinant segments were detected more frequently between pairs of *R. etli* strains: 18% (CFN42-BRASIL 5), 25% (CFN42-CIAT652), and 25% (CFN42-CIAT652). The same pattern was seen for the other five groups of quartets. This effect is because homologous recombination depends on a high nucleotide identify, and greater divergence is associated with less homologous recombination [32]. Therefore, recombination might be more frequent between strains (populations) that are closely related. Indeed, we observed the same recombination events in different groups of quartets (of different strains), as indicated by a presence/absence matrix. In general, the number of common recombination events (small number of events) was related to the phylogenetic proximity of the strains, for instance BRASIL5 and 8C-3 share the most recombination events in common (data not shown).

To explore whether the recombination is particularly acting on some classes of genes, we assigned the

recombinant segments to COGs (see Methods), as shown additional file 1 Figure S3. All the functional classes annotated in the CFN42 genome are present in the draft genomes but they are represented unevenly in the recombinant segments. For instance, the categories: amino acid transport and metabolism, carbohydrate transport and metabolism, energy production and conversion, lipid transport and metabolism, general function prediction only and function unknown appear overrepresented among the recombinant segments. In counterpart, some other categories like transcription and signal transduction mechanisms are in lower frequency among the recombinant segments than in CFN42. Even though we performed a chi-square and Range tests [33] to assess the significance of these differences, the incomplete nature of draft genomes does not allow to conclude about some bias toward recombination in certain classes of genes.

Genetic diversity

Together the above-described data suggest that recombination may not be a major driver of genomic diversification in *R. etli*, but rather might have relatively limited effects. To directly examine this point, we estimated the mean nucleotide diversity per nucleotide site (π) for the recombinant and non-recombinant gene segments of each strain (Figure 7a). In general, recombinant segments showed higher π values than non-recombinant segments. These differences were significant only for strains CIAT894, GR56, IE4771 and KIM5 (Student's *t*-test, $p < 0.001$), but the combined data for the π values of the 240 recombinant and non-recombinant gene segments common to the eight strains showed the lowest π values (0.06 on average). Although there was no significant difference between recombinant (red circles) and non-recombinant segments (blue circles) with regard to the regions common to all eight strains (Figure 7b), most of the recombinant segments had higher-than-average π values and generally showed the highest transition/transversion ratios (indicated by the size of the circles in Figure 7b). Since the probability of transitions is higher than transversions [34], high ratios of transition/transversion suggest that they were under strong purifying selection, because transitions at the third 'wobble' position are more likely to be synonymous than transversions [35].

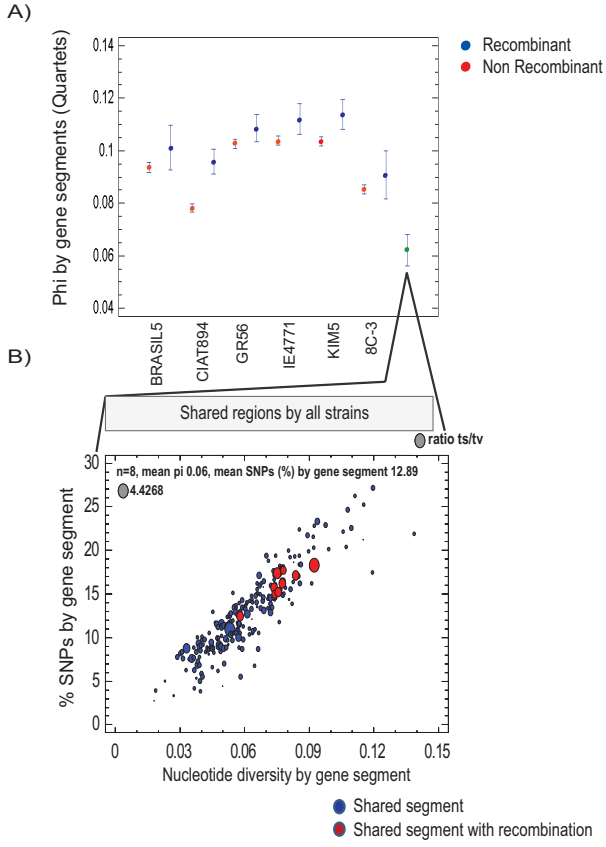

Figure 7 Genetic diversity in recombinant segments. A) For each homologous segment (quartet; regardless of evidence of recombination), we calculated the nucleotide diversity phi (Y axis; see Methods). The dots indicate the distribution means and the bars represent the 95% confidence intervals. Blue and red dots indicate recombinant and non-recombinant segments, respectively. Moreover, we determined the nucleotide diversity of the sequence regions shared across all of the tested strains of *R. etli* (green dot). Abscissa: BRASIL5, CIAT894, GR56, IE4771, KIM5, 8C-3 and shared regions. B) Magnification of the results from the 240 common sequence segments shared by all tested strains. The average percentage of SNPs was 12.89% per segment. The sphere sizes indicate the proportions of ts/tv for both recombinant segments (red circles) and non-recombinant segments (blue circles). The Y axis denotes the SNP percentage by gene segment, while the X axis shows the nucleotide diversity.

Discussion

In the present work, we used a genomic approach to detect and measure variation in the form of SNPs, and to analyze the contribution of recombination to the genomic diversification of *R. etli* strains. Our results demonstrated that draft genomic sequences samples representing ~1× of the genome can be used to measure variation at the whole-genome level in this species. In *R. etli* we found a great amount of variation (more than 161,998 SNPs) when any draft genome was compared to the complete genomes of CFN42 and CIAT652. To assess the reliability of this method for identifying SNPs, we quantified the SNPs in *E. coli* genomes at 1× and in complete genomes assembled at about 10× coverage. We found the same variation level using either draft or complete *E. coli* genomes, indicating that draft genomes produced estimations of DNA variability comparable to those generated using complete genomes even at only

1× coverage. Richter and Roselló-Mora [10] previously reported on the use of partial sequences representing about 20% of the genomes of several bacterial species to infer reliable values of DNA divergence between strains. The authors of the prior paper showed that ANI values obtained with these samples correlated well with the DDH values, indicating that draft genome sequences are an acceptable data source. At present, the rapid improvement of DNA sequencing technology is allowing researchers to use multiplex sequencing to simultaneously process an increasing number of genomic sequences. These experiments will produce additional draft genome sequences of different qualities, and the approach proposed herein should prove useful for their early analysis.

We identified a higher proportion of SNPs in *R. etli* strains than in *E. coli* strains, and the differences between the various *R. etli* strains and *Rhizobium leguminosarum bv viciae 3841* ranged from 7% to 11% (median; additional file 1 Figure S2), with the latter figure corresponding to the CFN42 comparison. *R. etli* and *R. leguminosarum* are different species according to 16S comparison; however, they share a common genomic core and are distinguished by variable accessory components (e.g., plasmids) [24,36,37]. Therefore, an ANV range of 7-11% might be a good indicator of speciation within *Rhizobium*. Despite of the variability in ANV among the tested strains of *R. etli* (about 4-6%), none had ANV values comparable to those obtained with respect to *R. leguminosarum*. The levels of ANV were higher for comparisons using CFN42 than those done with CIAT652. For taxonomic purposes, CFN42 is the type strain of *R. etli* [38]. In the present analysis, however, we found that CFN42 was the most differentiated of the studied samples, had the highest proportion of unique SNPs, and clustered as a divergent independent branch when the strain phylogeny was explored. We recently re-sequenced strain CFN42 using Solexa-Illumina technology and compared it with the former complete genome sequence. Very few indels and SNPs (less than 20 SNPs) and no rearrangements were found. Therefore, very small variation can be expected from an in vitro lifestyle. In contrast, most of the strains analyzed were more closely related to CIAT652 than to CFN42. A prior study noted that CIAT652 and CFN42 have a low ANI value (90.44%) [10] and suggested that CIAT652 is improperly classified as *R. etli*. We previously showed that CFN42 and CIAT652 share a very conserved symbiotic plasmid, but have high divergence throughout the rest of their genomes [24]. Given that all isolates of *R. etli* have been recovered from nitrogen-fixing bean nodules, this characteristic would be expected to dominate the classification criteria. The genomic divergence described herein is thus consistent with a model in which the species *R. etli* is composed of divergent genomic lineages that share the symbiotic phenotype conferred by the symbiotic plasmid [24], which is called a common symbiovar [39]. Indeed, our analysis suggests that in some instances, the use of type strains could lead to misleading taxonomic classifications, especially when gene transfer mechanisms are active. *R. etli* is known to have mobile elements such as conjugative plasmids, insertion sequences and bacteriophages [40-42]. Therefore, gene flow and recombination among strains of *R. etli* might be important to the production of genomic diversity, as reflected in its pangenomic structure [24]. However, no prior study has assessed the role of homologous recombination in promoting the genomic diversity of *R. etli*. Earlier works using MLEE or MLST concluded that *R. etli* populations are essentially clonal, with low recombination even in sympatric populations [22,25,26,43]. More recently, Flores *et al.* [44] showed that despite the high conservation of the symbiotic plasmid pSym sequences from a collection of different strains of *R. etli*, some regions shared identical SNP distribution profiles. This observation was interpreted as evidence of homologous recombination. Here, we obtained similar findings for a set of common genomic DNA segments, mainly chromosomal in origin, belonging to eight strains of *R. etli*. Quantification of probable recombination events and the extrapolation of our findings to the whole genome suggested that a minimum of 260 recombination events had occurred in the genome of each strain. Strains CFN42 and CIAT894 were the more variable in terms of SNPs, and the latter also showed the most evidence for recombinant events in our quartet analysis (within the orthologous segments). Even though there were some discrepancies within the clades of the various phylogenetic trees we generated, most of the trees were congruent with the consensus tree. Moreover, although the estimated recombination was correlated with genetic diversity (Figure 7a), it was low overall (3-10%). In comparison, the whole-genome recombination estimates reported for *Rickettsia* and *Streptococcus* were on the order of 18-37% and 28%, respectively [45,46]. These data suggest that only a minor fraction of the *R. etli* genome has undergone recombination, which thus accounts for only a low proportion of the polymorphism in this species.

In bacteria, the frequency of RecA-mediated homologous recombination depends on the level DNA identity, and small DNA fragments are often introduced into the cell via conjugation, transformation or transduction. Consequently, only a fraction of the genome might be targeted by recombination [47]. Several other factors might account for the low recombination frequency detected in the isolate of *R. etli* studied here. Among them, the ample degree of divergence among the studied

R. etli strains, their distant geographical origins (USA, México, Costa Rica, Colombia, Brazil, and Spain) [24], and the small number of sampled strains. Recently, Bailly et al., reported a population genomics analysis of sympatric strains of *Sinorhizobium medicae* [27]. They found very low levels of polymorphism and recombination in the chromosome in comparison with the megaplasmids. Future studies using our methodology on *R. etli* isolates from single sites could be used to improve our understanding of how recombination impacts the diversification of this species.

Conclusion

In summary, our results and the previous reports on *R. etli* support a model in which the species is composed of evolutionarily independent lineages that share a symbiotic phenotype but have low levels of recombination among the various lineages. However, although genetic barriers imposed by divergence or other barriers such as geographical distance might preclude homologous recombination among the strains, gene flow (e.g., by plasmids and chromosomal islands) is an ongoing process that shapes the genomic and pangenomic structures of *R. etli*.

Methods

Genomes used

Complete genome sequences were downloaded from GenBank as follows: for *R. etli CFN42*: chromosome [GenBank:NC_007761], and plasmids pCFN42a [GenBank:NC_007762], pCFN42b [GenBank:NC_007763], pCFN4c [GenBank:NC_007764], pCFN42d [GenBank:NC_004041], pCFN42e [GenBank:NC_007765], and pCFN42f [GenBank:NC_007766]; for *R. etli CIAT652*: chromosome [GenBank:NC010994], and plasmids pCIAT652a [GenBank:NC010998], pCIAT652b [GenBank:NC010996], and pCIAT652c [GenBank:NC010994]; and for *R. leguminosarum 3841*: chromosome [GenBank:NC_008380], and plasmids pRL7 [GenBank:NC_008382], pRL8 [GenBank:NC_008383], pRL9 [GenBank:NC_008379], pRL10 [GenBank:NC_008381], pRL11 [GenBank:NC_008384], and pRL12 [GenBank:NC_008378]. We also used reads and contigs from the draft genomes of *R. etli* strains 8C-3 [GenBank:NZ_ABRA00000000], BRASIL5 [GenBank:NZ_ABQZ00000000], CIAT894 [GenBank:NZ_ABRD0000000], GR56 [GenBank:NZ_AABRD0000000], IE4771 [GenBank:NZ_ABRD00000000], and KIM5 [GenBank:NZ_ABQY0000000].

Determination of SNPs and pairwise nucleotide differences

Paired alignments between the draft genomes (contigs) and the ORFs from the genomes of CFN42 or CIAT652 were performed using the Dds2 program [48], which produces ungapped alignments of fragments having similarities greater than 80%. Each duplicated paired alignment (*i.e.*, segments for which paralogous existed in the reference genome) was filtered using the reciprocal best hits option of the Fil program [48] under the following parameter set: coverage > 60% with respect to a reference gene and a percentage differential score cutoff < 10%. When two alignments had the same coverage, we selected the alignment with the higher score. Once the results were filtered, we created a gapped alignment using the Gap22 program [48] on segments for which the identity was greater than 85%. Both sequences were extracted using an *ad hoc* Perl script (homemade) formed for each paired alignment. To avoid frameshifts, we realigned each pair using cross-match [49] with the following parameters: discrep_lists masklevel, 0; tags gap_init, 3; gap_ext, 2; ins_gap_ext, 2; del_gap_ext, 2; minmatch, 14; maxmatch, 14; max_group_size, 20; minscore, 30; bandwidth, 14; and indexwordsize, 10. Finally, for each alignment, we determined the probability that a site was polymorphic using the Polybayes program [50], with the probability set at greater than 0.99 and a minimum Phred of Q45 [51,52].

Assessment of methodological accuracy at low coverage

To determine if differences in coverage among the studied strains affected the reliability of the variability estimations, we took readings representing ~1× sequence coverages of seven *E. coli* genomes and complete-genome readings (about 10×) of the same genomes from GenBank (ftp://ftp.ncbi.nih.gov/genomes/) and assembled these readings using the Celera assembler [53]. The above-described analysis was applied to both the 1× and 10× coverage datasets, and the results were compared using the Mann-Whitney and Kolmogorov-Smirnov tests [33]. The utilized *E. coli* draft genomes were: 101-1 [GenBank:NZ_AAMK00000000], 53638 [GenBank:NZ_AAKB00000000], B171 [GenBank:NZ_AAJX0000000], E1100019 [GenBank:NZ_AAJW0000000], F11 [GenBank:NZ_AAJU0000000], HS [GenBank:NC_009800], and *O157_H7_ec4024* [GenBank:NZ_ABJT0000000]. The complete genome sequence of *E. coli* K12 [GenBank:NC_000913] was used as the reference.

Determination of triplets (homologous segments)

For the comparisons between all ORFs of the reference genomes (both CFN42 and CIAT652 were used throughout the work) and each incomplete genome (the contigs), we obtained the coordinates of all homologous segments (triplets) using the Mauve program [54]. Our analysis was standardized by aligning p42F (CFN42) against (*R. leguminosarum bv viciae 3841*) pRL12 using

the following parameters: backbone-size = 100; max-backbone-gap = 50; weight = 90; island-size = 100. These plasmids were chosen because they contain shared syntenic blocks [36]. Sequence extraction, realignment of each conserved segment (backbone) and SNP determination were all performed as described above (see determination of SNPs section).

Determination of quartets (orthologous segments)

To detect recombination events among DNA sequences, at least four sequences are required for the analysis [55]. Here, we first identified SNPs that distinguished each draft genome from the two reference genomes (CFN42 and CIAT652), and then determined the fragments that were shared between each draft genome and the ORFs from CFN42 and CIAT652, together with all replicons of *R. leguminosarum 3841* (chosen because of its extensive synteny with CFN42) [36]. Sharing was determined using the Mauve program [54] (see determination of triplets section) and the shared fragments were realigned with the Muscle program (default parameters) [56]. To eliminate any large gaps within the alignments (rare in orthologous fragments), we used the Gblocks program under its default parameters [57].

Detection of recombination

A variety of methods for detection of recombination have been reported in the literature [58], but no one strategy performs optimally under all evolutionary scenarios [59]. Therefore, a reasonable approach is to employ multiple methods and consider recombination events predicted by at least two methods as being the most reliable. Here, we used this strategy and considered recombination events that were detected by at least two of the following four programs [46]:

A) Geneconv [60]: Using this program, we ran 100,000 simulations for each quartet with the following parameters chosen: Dumptab; Dumpjseq; Dumpfrag; Annotate; WideCol; ShowBlast; Indel_blocs; ShowBcPwKaPvals; SortGfragsBySeq; Show_maxmeansims; ShowUnal; Gscale = 1; ListPair; ListBest; Bcsims; Allouter; Numsim = 100000/sp. This allowed us to detect possible genetic conversion events.

B) Pist [61]: With this program, we first identified the best-fit DNA substitution model for each shared fragment using the Akaike information criterion. We then used the best model to reconstruct the phylogeny using a maximum likelihood method (Phyml [62]) with 100 non-parametric bootstrap replicates. We next determined the invariant sites, alpha values, ts/tv ratios, base frequencies, and constant sites using the PAML program [63] and the GTR model. Finally, we ran Pist with the REV model and 10,000 permutations. Pist uses parsimony-informative sites to detect recombination events and is robust for highly divergent genes.

C) PhiPack [64]: We used the parameters of 10,000 permutations and a window size of 25 nt, and implemented the Pairwise Homoplasy Index, Maximum X2, and the Neighbor Similarity Score.

D) Hyphy program [65]: We used the routine GARD, which enables automated phylogenetic detection of recombination. We employed the GTR model and beta-gamma rate variation.

To determine if a recombinant gene was present in two different quartets or strains, we constructed a binary presence/absence matrix (1/0) for each gene that was found in two or more strains. These profiles were hierarchically clustered using the Cluster program [66].

Phylogenetic analysis

Regions shared among all strains of *R. etli* and *R. leguminosarum bv viciae 3841* were identified using the Mauve program [54], realigned by Muscle using the default parameters [56], and filtered for long gaps with Gblocks [57]. We then obtained the phylogeny of each region using a maximum likelihood approach employed by the Phyml program [62] (with 1,000 non-parametric bootstrap replicates) and the best nucleotide substitution model identified by the Akaike information criterion [67,68]. We used three methods to construct the phylogeny from the concatenated dataset, in order to determine the species tree. The first was the RAxML program (maximum likelihood) [69], in which we ran the GTR nucleotide substitution model and a GAMMA +P-Invar estimation of rate heterogeneity. This analysis yielded a Maximum Likelihood ML estimate of the alpha parameter and 1,000 distinct randomized Maximum Parsimony trees. The second program used was Phyml (maximum likelihood) [62], running 1,000 non-parametric replicates and the GTRG model. Finally, we employed the MrBayes program (Bayesian analysis) [70] running the Nucmodel 4by4 for DNA. The number of rate categories for the gamma distribution was set at four, with an allowance for a proportion of invariable sites. Because of the high computational burden, we performed two runs with four chains, for 500,000 generations in total. Trees were sampled every 500 generations, 25% of all samples were removed as reflecting burn-ins, and a consensus was obtained. Moreover, to assess differences in topology among the probable strain trees and individual gene trees, we used the Consel program [71], which calculated expected likelihood weighting and performed the Shimodaira-Hasegawa SH test [72]. Finally, a neighbor-net network was generated using the concatenated sequences and the Splits tree4 program [31].

Nucleotide diversity and ts/tv ratios

For each shared fragment (quartet), we determined the nucleotide diversity and segregating sites using *R. leguminosarum 3841* as an outgroup and employing the libsequence library [73]. The transition/transversion ts/tv ratios were determined for each quartet by using the PAML program [63] and applying the best model of nucleotide substitution obtained from each orthologous segment (see determination of quartet).

Functional assignment

We used the COGs database [74] to undertake functional annotation across the four broad categories and sub categories to shared regions (all strains) as well as recombinant quartets. Quartets that had not been functionally assigned within the COG database were placed in the "Poorly Characterized" category. For assignation to a category, we used the reciprocal best hits technique with an E-value < 1×10^{-7}.

Additional material

Additional file 1: Strategy for Determining SNPs. The additional file (in .pdf format) includes text and figures delineating our process for determining SNPs (parameters, paired comparisons and SNP differences). Also include the distribution of functional classes (COGs) of recombinant quartets of each draft genome and your comparison against distribution of CFN42.

Additional file 2: Table of genes presented in 240 shared regions. The additional file (in .xls) includes the tables of genes and your features, as name, coordinates, gi, COGs and other.

Acknowledgements

We thank José Espiritu, Ismael L Hernández, and José L Fernández for their help with technical and computational resources, and Miguel A Cevallos for critical reading of the manuscript.

This work was supported by grants from CONACyT (CB131499 and U4633) and PAPIIT-UNAM (IN215908 and IN223005). JLA received a Ph. D. fellowship from CONACyT. We thank the Doctorate in Biomedical Sciences and specially to National Autonomous University of México.

LEE contributed during a sabbatical visit to the University of California Irvine (UCI) in the laboratory of Brandon Gaut; this work was supported by UC-MEXUS, CONACYT, and DGAPA, UNAM.

Author details

[1]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Av. Universidad N/C Col. Chamilpa, Apdo. Postal 565-A, Cuernavaca 62210, México. [2]Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, CU, AP 70-275 Coyoacán, 04510 México, DF, México.

Authors' contributions

JLA conceived, designed, and performed the experiments. JLA and VG analyzed the data and wrote the manuscript. RIS and PB were responsible for the genomic sequencing. PV, LEE and EM-R discussed the data. GD and VG contributed materials. VG edited the manuscript and is the Ph. D. thesis advisor of JLA. All authors read and approved the final manuscript.

Received: 1 June 2011 Accepted: 17 October 2011
Published: 17 October 2011

References

1. Kimura M, Takahata N: **Selective constraint in protein polymorphism: study of the effectively neutral mutation model by using an improved pseudosampling method.** *Proc Natl Acad Sci USA* 1983, **80(4)**:1048-1052.
2. Nei M: **Selectionism and neutralism in molecular evolution.** *Mol Biol Evol* 2005, **22(12)**:2318-2342.
3. Wagner A: **Neutralism and selectionism: a network-based reconciliation.** *Nat Rev Genet* 2008, **9(12)**:965-974.
4. Konstantinidis KT, Ramette A, Tiedje JM: **The bacterial species definition in the genomic era.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361(1475)**:1929-1940.
5. Rossello-Mora R, Amann R: **The species concept for prokaryotes.** *FEMS Microbiol Rev* 2001, **25(1)**:39-67.
6. Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, Moore WEC, Murray RGE, Stackebrandt E, Starr MP, *et al*: **Report of the ad hoc committee on reconciliation of approaches to bacterial systematics.** *Int J Syst Bacteriol* 1987, **37**:463-464.
7. Konstantinidis KT, Ramette A, Tiedje JM: **The bacterial species definition in the genomic era.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361(1475)**:1929-1940.
8. Konstantinidis KT, Tiedje JM: **Towards a genome-based taxonomy for prokaryotes.** *J Bacteriol* 2005, **187(18)**:6258-6264.
9. Deloger M, El Karoui M, Petit MA: **A genomic distance based on MUM indicates discontinuity between most bacterial species and genera.** *J Bacteriol* 2009, **191(1)**:91-99.
10. Richter M, Rossello-Mora R: **Shifting the genomic gold standard for the prokaryotic species definition.** *Proc Natl Acad Sci USA* 2009, **106(45)**:19126-19131.
11. Didelot X, Maiden MC: **Impact of recombination on bacterial evolution.** *Trends Microbiol* **18(7)**:315-322.
12. Bergstrom CT, Lipsitch M, Levin BR: **Natural selection, infectious transfer and the existence conditions for bacterial plasmids.** *Genetics* 2000, **155(4)**:1505-1519.
13. Campbell AM: **Lateral gene transfer in prokaryotes.** *Theor Popul Biol* 2000, **57(2)**:71-77.
14. Gal-Mor O, Finlay BB: **Pathogenicity islands: a molecular toolbox for bacterial virulence.** *Cell Microbiol* 2006, **8(11)**:1707-1719.
15. Posada D, Crandall KA: **The effect of recombination on the accuracy of phylogeny estimation.** *J Mol Evol* 2002, **54(3)**:396-402.
16. Thomas CM, Nielsen KM: **Mechanisms of, and barriers to, horizontal gene transfer between bacteria.** *Nat Rev Microbiol* 2005, **3(9)**:711-721.
17. Touzain F, Denamur E, Medigue C, Barbe V, El Karoui M, Petit MA: **Small variable segments constitute a major type of diversity of bacterial genomes at the species level.** *Genome Biol* 2010, **11(4)**:R45.
18. Spratt BG, Hanage WP, Feil EJ: **The relative contributions of recombination and point mutation to the diversification of bacterial clones.** *Curr Opin Microbiol* 2001, **4(5)**:602-606.
19. Jiggins FM: **The rate of recombination in Wolbachia bacteria.** *Mol Biol Evol* 2002, **19(9)**:1640-1643.
20. Jolley KA, Wilson DJ, Kriz P, McVean G, Maiden MC: **The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in Neisseria meningitidis.** *Mol Biol Evol* 2005, **22(3)**:562-569.
21. Tettelin H, Riley D, Cattuto C, Medini D: **Comparative genomics: the bacterial pan-genome.** *Curr Opin Microbiol* 2008, **11(5)**:472-477.
22. Souza V, Nguyen TT, Hudson RR, Pinero D, Lenski RE: **Hierarchical analysis of linkage disequilibrium in Rhizobium populations: evidence for sex?** *Proc Natl Acad Sci USA* 1992, **89(17)**:8389-8393.
23. Abby S, Daubin V: **Comparative genomics and the evolution of prokaryotes.** *Trends Microbiol* 2007, **15(3)**:135-141.
24. Gonzalez V, Acosta JL, Santamaria Rl, Bustos P, Fernandez JL, Hernandez Gonzalez IL, Diaz R, Flores M, Palacios R, Mora J, *et al*: **Conserved symbiotic plasmid DNA sequences in the multireplicon pangenomic structure of Rhizobium etli.** *Appl Environ Microbiol* 2010, **76(5)**:1604-1614.
25. Pinero D, Martinez E, Selander RK: **Genetic diversity and relationships among isolates of Rhizobium leguminosarum biovar phaseoli.** *Appl Environ Microbiol* 1988, **54(11)**:2825-2832.
26. Silva C, Le E, Souza V: **Reticulated and epidemic population genetic structure of *Rhizobium etli* biovar *phaseoli* in a traditionally managed locality in Mexico.** *Molecular Ecology* 1999, **8**:277-287.

27. Bailly X, Giuntini E, Sexton MC, Lower RP, Harrison PW, Kumar N, Young JP: **Population genomics of *Sinorhizobium medicae* based on low-coverage sequencing of sympatric isolates.** *ISME J* 2011.
28. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB: **Quality scores and SNP detection in sequencing-by-synthesis systems.** *Genome Res* 2008, **18(5)**:763-770.
29. Hubisz MJ, Lin MF, Kellis M, Siepel A: **Error and error mitigation in low-coverage genome assemblies.** *PLoS One* 2011, **6(2)**:e17034.
30. Gonzalez V, Bustos P, Ramirez-Romero MA, Medrano-Soto A, Salgado H, Hernandez-Gonzalez I, Hernandez-Celis JC, Quintero V, Moreno-Hagelsieb G, Girard L, *et al*: **The mosaic structure of the symbiotic plasmid of Rhizobium etli CFN42 and its relation to other symbiotic genome compartments.** *Genome Biol* 2003, **4(6)**:R36.
31. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol* 2006, **23(2)**:254-267.
32. Fraser C, Hanage WP, Spratt BG: **Recombination and the nature of bacterial speciation.** *Science* 2007, **315(5811)**:476-480.
33. Dean CB, Nielsen JD: **Generalized linear mixed models: a review and some extensions.** *Lifetime Data Anal* 2007, **13(4)**:497-512.
34. Keller I, Bensasson D, Nichols RA: **Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes.** *PLoS Genet* 2007, **3(2)**:e22.
35. Yang Z, Bielawski JP: **Statistical methods for detecting molecular adaptation.** *Trends Ecol Evol* 2000, **15(12)**:496-503.
36. Crossman LC, Castillo-Ramirez S, McAnnula C, Lozano L, Vernikos GS, Acosta JL, Ghazoui ZF, Hernandez-Gonzalez I, Meakin G, Walker AW, *et al*: **A common genomic framework for a diverse assembly of plasmids in the symbiotic nitrogen fixing bacteria.** *PLoS One* 2008, **3(7)**:e2567.
37. Young JP, Crossman LC, Johnston AW, Thomson NR, Ghazoui ZF, Hull KH, Wexler M, Curson AR, Todd JD, Poole PS, *et al*: **The genome of Rhizobium leguminosarum has recognizable core and accessory components.** *Genome Biol* 2006, **7(4)**:R34.
38. Segovia L, Pinero D, Palacios R, Martinez-Romero E: **Genetic structure of a soil population of nonsymbiotic Rhizobium leguminosarum.** *Appl Environ Microbiol* 1991, **57(2)**:426-433.
39. Rogel MA, Ormeno-Orrillo E, Martinez Romero E: **Symbiovars in rhizobia reflect bacterial adaptation to legumes.** *Syst Appl Microbiol* 2011, **34(2)**:96-104.
40. Tun-Garrido C, Bustos P, Gonzalez V, Brom S: **Conjugative transfer of p42a from rhizobium etli CFN42, which is required for mobilization of the symbiotic plasmid, is regulated by quorum sensing.** *J Bacteriol* 2003, **185(5)**:1681-1692.
41. Lozano L, Hernandez-Gonzalez I, Bustos P, Santamaria RI, Souza V, Young JP, Davila G, Gonzalez V: **Evolutionary dynamics of insertion sequences in relation to the evolutionary histories of the chromosome and symbiotic plasmid genes of Rhizobium etli populations.** *Appl Environ Microbiol* 2010, **76(19)**:6504-6513.
42. Perez-Mendoza D, Dominguez-Ferreras A, Munoz S, Soto MJ, Olivares J, Brom S, Girard L, Herrera-Cervera JA, Sanjuan J: **Identification of functional mob regions in Rhizobium etli: evidence for self-transmissibility of the symbiotic plasmid pRetCFN42d.** *J Bacteriol* 2004, **186(17)**:5753-5761.
43. Silva C, Vinuesa P, Eguiarte LE, Martinez-Romero E, Souza V: **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* 2003, **69(2)**:884-893.
44. Flores M, Morales L, Avila A, Gonzalez V, Bustos P, Garcia D, Mora Y, Guo X, Collado-Vides J, Pinero D, *et al*: **Diversification of DNA sequences in the symbiotic genome of Rhizobium etli.** *J Bacteriol* 2005, **187(21)**:7185-7192.
45. Lefebure T, Stanhope MJ: **Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition.** *Genome Biol* 2007, **8(5)**:R71.
46. Wu J, Yu T, Bao Q, Zhao F: **Evidence of extensive homologous recombination in the core genome of rickettsia.** *Comp Funct Genomics* 2009, 510270.
47. Retchless AC, Lawrence JG: **Phylogenetic incongruence arising from fragmented speciation in enteric bacteria.** *Proc Natl Acad Sci USA* 2010, **107(25)**:11453-11458.
48. Wang J, Huang X: **A method for finding single-nucleotide polymorphisms with allele frequencies in sequences of deep coverage.** *BMC Bioinformatics* 2005, **6**:220.
49. Gordon D, Desmarais C, Green P: **Automated finishing with autofinish.** *Genome Res* 2001, **11(4)**:614-625.
50. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, Gish WR: **A general approach to single-nucleotide polymorphism discovery.** *Nat Genet* 1999, **23(4)**:452-456.
51. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8(3)**:186-194.
52. Lawrence CB, Solovyev VV: **Assignment of position-specific error probability to primary DNA sequence data.** *Nucleic Acids Res* 1994, **22(7)**:1272-1280.
53. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al*: **The genome sequence of Drosophila melanogaster.** *Science* 2000, **287(5461)**:2185-2195.
54. Darling AC, Mau B, Blattner FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements.** *Genome Res* 2004, **14(7)**:1394-1403.
55. Posada D, Crandall KA: **The effect of recombination on the accuracy of phylogeny estimation.** *J Mol Evol* 2002, **54(3)**:396-402.
56. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5**:113.
57. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Syst Biol* 2007, **56(4)**:564-577.
58. Posada D: **Evaluation of methods for detecting recombination from DNA sequences: empirical data.** *Mol Biol Evol* 2002, **19(5)**:708-717.
59. Posada D, Crandall KA, Holmes EC: **Recombination in evolutionary genomics.** *Annu Rev Genet* 2002, **36**:75-97.
60. Sawyer S: **Statistical tests for detecting gene conversion.** *Mol Biol Evol* 1989, **6(5)**:526-538.
61. Worobey M: **A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria.** *Mol Biol Evol* 2001, **18(8)**:1425-1434.
62. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52(5)**:696-704.
63. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24(8)**:1586-1591.
64. Bruen TC, Philippe H, Bryant D: **A simple and robust statistical test for detecting the presence of recombination.** *Genetics* 2006, **172(4)**:2665-2681.
65. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD: **GARD: a genetic algorithm for recombination detection.** *Bioinformatics* 2006, **22(24)**:3096-3098.
66. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95(25)**:14863-14868.
67. Posada D, Crandall KA: **Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1).** *Mol Biol Evol* 2001, **18(6)**:897-906.
68. Posada D, Crandall KA: **Selecting the best-fit model of nucleotide substitution.** *Syst Biol* 2001, **50(4)**:580-601.
69. Stamatakis A, Ludwig T, Meier H: **RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees.** *Bioinformatics* 2005, **21(4)**:456-463.
70. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19(12)**:1572-1574.
71. Shimodaira H, Hasegawa M: **CONSEL: for assessing the confidence of phylogenetic tree selection.** *Bioinformatics* 2001, **17(12)**:1246-1247.
72. Strimmer K, Rambaut A: **Inferring confidence sets of possibly misspecified gene trees.** *Proc Biol Sci* 2002, **269(1487)**:137-142.
73. Thornton K: **Libsequence: a C++ class library for evolutionary genetic analysis.** *Bioinformatics* 2003, **19(17)**:2325-2327.
74. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution.** *Nucleic Acids Res* 2000, **28(1)**:33-36.

doi:10.1186/1471-2148-11-305
Cite this article as: Acosta *et al.*: Genomic lineages of *Rhizobium etli* revealed by the extent of nucleotide polymorphisms and low recombination. *BMC Evolutionary Biology* 2011 **11**:305.

CAPITULO IV.- METODOLOGIA

IV. 1 Diseño y valoración de una metodología para la determinación de SNPs en genomas fragmentados

Se diseñaron dos metodologías para la determinación de SNPs, la primera toma en cuenta un genoma de referencia y la segunda toma dos o más genomas de referencia al mismo tiempo.

IV. 1.1 Genomas utilizados

La secuencia de los genomas utilizados en este trabajo se obtuvieron del GenBank en el siguiente orden: el genoma de *R. etli CFN42*: el cromosoma [GenBank:NC_007761], y los plásmidos pCFN42a [GenBank:NC_007762], pCFN42b [GenBank:NC_007763], pCFN4c [GenBank:NC_007764], pCFN42d [GenBank:NC_004041], pCFN42e [GenBank:NC_007765] y pCFN42f [GenBank:NC_007766]; el genoma de *R. etli CIAT652*: el cromosoma [GenBank:NC010994] y los plásmidos pCIAT652a [GenBank:NC010998], pCIAT652b [GenBank:NC010996], y pCIAT652c [GenBank:NC010994]; y finalmente el genoma de *R. leguminosarum 3841*: el cromosoma [GenBank:NC_008380], y los plásmidos pRL7 [GenBank:NC_008382], pRL8 [GenBank:NC_008383], pRL9 [GenBank:NC_008379], pRL10 [GenBank:NC_008381], pRL11 [GenBank:NC_008384] y pRL12 [GenBank:NC_008378]. Las lecturas y los contigs de los siguientes genomas parciales fueron utilizados, los genomas de *R. etli* son los siguientes: cepa *8C-3*

[GenBank:NZ_ABRA00000000], *BRASIL5* [GenBank:NZ_ABQZ00000000], *CIAT894* [GenBank:NZ_ABRD0000000], *GR56* [GenBank:NZ_AABRD0000000], *IE4771* [GenBank:NZ_ABRD00000000] y *KIM5* [GenBank:NZ_ABQY0000000].

IV. 1.2 Un genoma de referencia

Se basa en comparaciones pareadas entre cada uno de los genomas incompletos (contigs) y los ORFs (marcos de lectura abiertos) del genoma de referencia (*CIAT652* o *CFN42*). Las comparaciones se hicieron con el programa Dds2 [94], el cual produce alineamientos sin *gaps* (vacios), siempre y cuando tenga una identidad nucleotídica mayor al 80 %. Cada alineamiento duplicado (parálogos en el genoma de referencia) fue filtrado usando el *reciprocal best hits* (relación homóloga reciproca entre dos genes) mediante el programa Fil [94] y se retuvieron los alineamientos que presentaron una cobertura mayor al 60% con respecto al gen de referencia y un porcentaje diferencial de la calificación (*score*) del alineamiento menor al 10%. Cuando dos alineamientos tuvieron la misma cobertura, seleccionamos aquellos con la calificación más alta. Una vez filtrados los posibles parálogos, hicimos re-alineamientos con *gaps* a partir de los segmentos que pasaron los filtros de cobertura y de calificación (descritos arriba) usando para esto el programa Gap22 [94]. Solamente se almacenaron los alineamientos con una identidad mayor al 85%. Para cada alineamiento almacenado, extrajimos las secuencias del segmento homólogo (tanto para el genoma parcial, como para el genoma de referencia) con la mejor calificación

mediante un programa de *Perl* (hecho en casa). Con el fin de evitar cambios de marco de lectura (*frameshifts*), realineamos cada par de segmentos homólogos mediante el programa Cross-match [95], usando los siguientes parámetros: discrep_lists masklevel, 0; tags gap_init, 3; gap_ext, 2; ins_gap_ext, 2; del_gap_ext, 2; minmatch, 14; maxmatch, 14; max_group_size, 20; minscore, 30; bandwidth, 14; and indexwordsize, 10. Finalmente, para cada alineamiento (resultado del Cross-match), determinamos la probabilidad de que un sitio sea polimórfico usando el programa Polybayes [96], tomando siempre las mutaciones con una probabilidad mayor a 0.99 y un valor mínimo de Phred de Q45 [97, 98].

IV.1.3 Dos o más genomas de referencia (tercias)

Para las comparaciones entre tres genomas, dos genomas completos (*CIAT652* y *CFN42*) y un genoma parcial (contigs), obtuvimos los *backbone* (segmentos ortólogos sin *gaps*) mediante el programa Mauve [99]. Con el fin de encontrar los mejores parámetros para nuestra muestra, estandarizamos el programa Mauve alineando el plásmido p42F de *R. etli CFN42* contra el plásmido pRL12 de *R. leguminosarum bv viciae 3841*, usando los siguientes parámetros: backbone-size = 100; max-backbone-gap = 50; weight = 90; island-size = 100. Estos plásmidos se eligieron por qué comparten varios bloques sinténicos [100]. La extracción y el realineamiento de cada segmento conservado (*backbone*), así como la determinación de SNPs (tomando las variaciones de cada genoma parcial), se realizó como está descrito en la sección de un genoma de referencia.

IV. 2 Valoración de la metodología a bajas coberturas

Para determinar si las coberturas 1X afectan la exactitud en la determinación de la variación nucleotídica, utilizamos siete genomas de *E. coli*, con coberturas aproximadas de 10X (casi terminados), obtenidos del GenBank (ftp://ftp.ncbi.nih.gov/genomes/). Para cada genoma, simulamos un nuevo genoma con una cobertura aproximada de 1X, en otras palabras, tomamos de forma aleatoria un número de lecturas (en promedio 12,000) que se ensamblaron con el ensamblador Celera [101]. De esta forma, se obtuvieron dos bases de datos con genomas a diferentes coberturas: 1X y 10X. Para cada grupo de genomas se aplicó la metodología descrita previamente y tomando siempre como genoma de referencia a *E. coli K12.* Los resultados de ambas determinaciones se compararon mediante las pruebas estadísticas: The Mann-Whitney y Kolmogorov-Smirnov [102]. Los genomas utilizados de *E. coli* fueron: *101-1* [GenBank:NZ_AAMK00000000], *53638* [GenBank:NZ_AAKB00000000], *B171* [GenBank:NZ_AAJX0000000], *E1100019* [GenBank:NZ_AAJW0000000], *F11* [GenBank:NZ_AAJU0000000], *HS* [GenBank:NC_009800], and *O157_H7_ec4024* [GenBank:NZ_ABJT0000000] y finalmente el genoma completo de *E. coli K12* [GenBank:NC_000913].

IV. 3 Implementación de una metodología para la detección de Recombinación homóloga en genomas fragmentados

La implementación consta de dos partes: la construcción de grupos de cuartetos (número mínimo de cepas o genomas para identificar eventos de recombinación) y la aplicación de múltiples métodos independientes para la detección de la recombinación dentro de cada grupo de cuartetos.

IV.3.1 Construcción de Cuartetos

Para llevar a cabo el análisis de detección de eventos de recombinación entre secuencias de DNA, se necesita al menos cuatro secuencias [103]. Por lo tanto, construimos cuartetos a partir de los genomas completos de: *R. etli CFN42, CIAT652* y *R. leguminosarum bv viciae 3481* y un genoma parcial. Se eligió el genoma de *R. leguminosarum* por que este comparte una extensiva sintenia tanto a nivel cromosoma, como en plásmidos [100]. Por cada genoma parcial obtuvimos un grupo de cuartetos, es decir, un grupo para el genoma de *BRASIL5*, otro grupo para *CIAT894*, etc. Antes de realizar el análisis, determinamos y filtramos los SNPs de buena calidad mediante la metodología descrita en la sección de “Dos o más genomas de referencia”, esta propuesta dió como resultado segmentos ortólogos (tercias) integradas por los genomas completos de *R. etli CFN42, CIAT652* y cada genoma parcial, de esta manera, tenemos un grupo de tercias por cada genoma parcial. Los cuartetos se obtuvieron al alinear cada grupo de tercias contra el genoma de *R. leguminosarum*, usando el programa Mauve [99] y cada cuarteto determinado se realineo con el programa Muscle [104]. Finalmente con el

fin de eliminar los *gaps* largos dentro de cada alineamiento (raros en fragmentos ortólogos), usamos el programa Gblocks [105] con los parámetros default.

IV. 3.2 Métodos independientes para la detección de recombinación

Se han reportado un gran variedad de métodos para la detección de Recombinación [106], pero la utilización de una sola estrategia no es suficiente para una identificación óptima ante todos los posibles escenarios evolutivos [107]. Por lo tanto, utilizamos cuatro diferentes métodos con sus subsecuentes pruebas estadísticas para la identificación de los eventos de recombinación dentro de cada cuarteto (descrito en la sección previa) y consideramos solamente los eventos de predichos por al menos dos métodos independientes. Los métodos utilizados fueron los siguientes:

a) Geneconv [108].- Este método nos permite identificar los posibles eventos de conversión génica y corrimos 100,000 simulaciones para cada cuarteto con los siguientes parámetros: Dumptab; Dumpjseq; Dumpfrag; Annotate; WideCol; ShowBlast; Indel_blocs; ShowBcPwKaPvals; SortGfragsBySeq; Show_maxmeansims; ShowUnal; Gscale = 1; ListPair; ListBest; Bcsims; Allouter; Numsim = 100000/sp.

b) Pist [109]: Esta metodología usa los sitios parsimoniosamente informativos para detectar recombinación. La metodología es robusta ante genes altamente divergentes. Antes de correr este programa, identificamos el mejor modelo de sustitución nucleotídico para cada cuarteto usando para

esto el criterio de “Akaike information”, el cual nos indica cómo se ajusta cada modelo evolutivo a una determinada secuencia. Una vez identificado el mejor modelo de sustitución nucleotídica, hicimos la reconstrucción filogenética usando el método de máxima verosimilitud incluido en el programa Phyml [110] utilizando 100 replicas con bootstrap no paramétrico. Además determinamos el número de sitios invariantes, valores de alpha, tasas de ts/tv (transiciones/trasversiones), frecuencias de las bases (cuatro nucleótidos) y sitios constantes tomando como base el modelo GTR (modelo general de sustitución nucleotídica). Estos parámetros se determinaron usando el programa Paml [111]. Finalmente, corrimos Pist con el modelo REV y 10,000 permutaciones.

c) PhiPack [112]: Este metodología esta implementada en el programa Splitstree 4 [113] y lo corrimos usando 10,000 permutaciones y un tamaño de palabra de 25 nucleótidos. Esta propuesta está basada en los siguientes métodos: Pairwise Homoplasy Index, Maximun X^2 y Neighboor Similarity Score.

d) Hyphy [114]: Utilizamos la rutina GARD, la cual está diseñada para detección de recombinación mediante la identificando de las incongruencias filogenéticas dentro de un grupo de secuencias. Empleamos el modelo GTR y el parámetro de tasa beta-gama para correr el programa.

Una vez identificados los cuartetos que presentaron posibles eventos de recombinación y dado que cada cuarteto pertenece a un grupo de segmentos obtenidos a partir de cada genoma incompleto, hicimos comparaciones pareadas de cada grupo de cuartetos recombinantes para identificar los cuartetos compartidos entre dos o más cepas. Esta comparación se llevó a cabo mediante la construcción de una matriz binaria de presencia/ausencia (1/0), con el fin de obtener perfiles de presencia/ausencia para cada cuarteto. Finalmente, los perfiles se agruparon jerárquicamente usando el programa Cluster [115].

IV.4 Análisis Filogenéticos

Para llevar a cabo la inferencia filogenética, primero obtuvimos las regiones compartidas entre los ocho genomas de *R. etli*, tanto de los genomas completos como de los incompletos, esto se realizó con el programa Mauve [99]. Los parámetros utilizados fueron: backbone-size = 100; max-backbone-gap = 50; weight = 90; island-size = 100. El resultado del alineamiento es una lista con las coordenadas de cada backbone (segmento ortólogo) por genoma. Por lo tanto, tuvimos que extraer la secuencia de cada segmento ortólogo mediante un script de *Perl* (hecho en casa). Posteriormente, cada segmento ortólogo se realineó con el programa Muscle [104] con parámetros default y finalmente los *gaps* largos dentro de cada segmento se filtraron con el programa Gblocks [105] con parámetros default.

Para cada segmento ortólogo, obtuvimos la filogenia usando el método de máxima verosimilitud empleando el programa Phyml [110]. La filogenia se soportó con 1,000 replicas con *boostrap* no paramétrico y el mejor modelo de sustitución nucleotídica se identificó para cada segmento ortólogo mediante el criterio de información Akaike [116, 117], a través del programa modeltest [118] con parámetros default. Finalmente el árbol consenso de los arboles individuales se obtuvo con el programa Consense implementado en la suite Phylip [119].

Con el fin de determinar el posible árbol de la muestra (especie), concatenamos todos los segmentos ortólogos (descritos arriba) y empleamos tres programas para inferir la filogenia: El primero fue RAxML (máxima verosimilitud) [120] y se corrió con el modelo de sustitución nucleotídica GTR y la estimación de los parámetros: tasa de heterogeneidad GAMMA+P-invar y alpha. Además, hicimos 1,000 distintos de arboles aleatorios bajo el método de máxima parsimonia. El segundo programa que usamos fue Phyml (máxima verosimilitud) [110] y se corrió con los siguientes parámetros: 1,000 replicas no paramétricas y usamos el modelo GTR. Finalmente, empleamos el programa MrBayes (análisis bayesiano) [121], utilizando el modelo Nucmodel 4by4 DNA. El número de categorías para la distribución gamma fue de cuatro con una asignación para los sitios invariables. Debido al costo computacional, hicimos dos corridas con cuatro cadenas y 500,000 generaciones en total. Los arboles se muestrearon cada 500 generaciones y el 25 % de toda muestra se removió como reflejo del trabajo de la cadena caliente y finalmente se obtuvo un consenso.

Cada árbol individual, de cada segmento ortólogo, se comparó con el mejor árbol de la muestra (descrito arriba), de esta manera valoramos las diferentes topologías entre los arboles individuales y los posibles arboles de la especie. Lo anterior se hizo con el programa Consel [122], el cual calcula el *expected likelihood weighting* (pesaje de la probabilidad esperada) e implementa la prueba Shimodaira-Hasegawa [123]. Finalmente, generamos una red de tipo *neighbor-net* con la secuencia concatenada (todos los segmentos ortólogos) y el programa Splits Tree4 [113].

IV. 5 Diversidad nucleotídica y asignación funcional

Partiendo de los grupos de cuartetos construidos (descrito en Construcción de Cuartetos), determinamos la diversidad nucleotídica y los sitios segregantes para cada cuarteto usando a *R. leguminosarum 3841* como grupo externo. Este análisis se llevó a cabo utilizando la librería *libsequence* [124]. Por otra parte, las tasas de transición/transversión se determinaron empleando el programa Paml [111], con la previa identificación del mejor modelos de sustitución nucleotídica para cada segmento ortólogo (descrito en construcción de cuartetos).

Para determinar la clase funcional de cada grupo de cuartetos, utilizamos la base de datos de los COGs [125]. La asignación se hizo a las cuatro principales categorías y subcategorizas por para cada grupo de cuartetos. Los cuartetos que no tuvieron una asignación dentro de los COGs, se colocaron en la categoría

principal de *Poorly Characterized*. Este análisis se hizo mediante la técnica de BBH (*reciprocal best hits*), usando un E-value de $1x10^{-7}$.

IV. 6 Identificación de operones

La obtención de los operones se realizó para los genomas *CFN42, CIAT652 y R. leguminosarum bv viciae 3841*, usando el programa UberPredictor [126]. Sin embargo, antes de correr el programa utilizamos la base de datos Operondb [127], con el fin de obtener los directones (conjunto de genes consecutivos sobre una cadena de DNA la cual no es interrumpida por genes RNA o genes que estén en la cadena opuesta) del conjunto de genomas completos: *CFN42*, *CIAT652*, *R. leguminosarum trifolli WSM2304* (NC_011369), *R. leguminosarum bv viciae 3841* y *S. meliloti 1021* (NC_003078). Estos genomas se eligieron por su cercanía filogenética. Otro requisito para el programa UberPredictor, son los genes ortólogos que se obtienen mediante la técnica del *reciprocal best hits*, los cuales se determinan mediante comparaciones pareadas todos *vs* todos de los proteomas de los genomas descritos previamente, usando para esto el programa Blastp [128], con los siguientes parámetros: e-value 1x10-7, -m9, -S T y –Sm F. Finalmente, los ortólogos se determinaron mediante el programa MCL [129].

CAPITULO V.- RESULTADOS

V. 1 Implementación y valoración de la metodología para la determinación de SNPs en genomas completos e incompletos

En esta sección diseñamos e implementamos una metodología para el manejo de genomas parciales y valoramos su exactitud en la determinación de SNPs. Esta propuesta se aplicó primero a una base de datos de genomas de *E. coli* y finalmente se aplicó a los ocho genomas de *R. etli*.

V.1.1 Diseño

La exactitud en la identificación de SNPs, está determinada por la calidad de la secuenciación de DNA, es decir, de los valores de calidad producidos por el secuenciador (Secuenciación Capilar), dichos valores se denotan como *Qphred*. Un Q20 significa un error en cada 100 pares de bases, Q30 un error en cada 1,000 pares de bases y así sucesivamente. El valor estándar que generalmente se utiliza es de Q20 y las mutaciones con valores de calidad menores a éste, son consideradas más bien como errores potenciales de secuenciación. Debido a este problema, usamos parámetros estrictos (ver Metodología) para identificar SNPs de alta calidad en un conjunto de ocho genomas de *R. etli*, dos genomas completos: *CFN42* y *CIAT652*, aislados de México y de Costa Rica respectivamente y seis genomas parciales de cepas provenientes de diferentes lugares del mundo: *BRASIL5* (Brasil), *CIAT894* (Colombia), *GR56* (España), *IE4771* (México), *KIM5* (USA) y *8C-3* (España) [130]. Los genomas parciales tienen una profundidad

aproximada a 1X y la secuencia de cada uno de los genomas (Lecturas) se obtuvieron del sitio de descarga del NCBI (National Center for Biotechnology Information) http://www.ncbi.nlm.nih.gov/Ftp/. Descargamos aproximadamente un promedio de 13,000 lecturas Sanger, con una media en longitud de 1,000 nucleótidos por cada genoma incompleto.

Las lecturas de cada genoma incompleto se alinearon contra cada genoma de referencia (tanto con *CFN42*, como con *CIAT652*), con el fin de evaluar los valores de calidad de la secuencia. Dichos alineamientos se hicieron con el programa Cross-match [131] y la evaluación de los SNPs se realizaron con el programa Polybayes [132], el cual determina la probabilidad de que un sitio sea polimórfico, basándose en la calidad *Phred* (Q) de la lectura. En todas las comparaciones pareadas la mayoría de SNPs (mayor a 10,000 mutaciones) tuvieron una probabilidad superior a 0.975 y además aproximadamente 90% (100,000 SNPs) de éstos tienen valores de calidad mayor a Q45. Aunque los valores de Q20 y una probabilidad de sitio polimórfico mayor a 0.90 son los que generalmente se utilizan para la identificación de SNPs [133]. Identificamos un promedio de 27,000 falsos positivos por genoma y muchos de estos falsos positivos tuvieron valores de calidad de Q20. Para evitar la posible inclusión de errores de secuencia utilizamos un valor de probabilidad mayor a 0.99 y valores por encima de Q45. Estos parámetros coinciden con otro estudio dónde evaluaron los posibles errores de secuenciación [134].

Las lecturas se ensamblaron con el programa Phrap [131], dando como resultados contigs (consensos). Debido a que cada posición dentro de un contig está soportada por más de una lectura, los valores de calidad aumentan en función del número de lecturas que soporten dicha posición, en otras palabras, a mayores lecturas mayor confiabilidad en la secuenciación del nucleótido. De esta manera, al trabajar con contigs y no lecturas aumentamos la confiabilidad en la determinación de SNPs.

Otro problema adicional y por lo tanto implementado en la metodología, es la estructura genómica de *R. etli*. Estos genomas tienen una alta proporción de redundancia metabólica y por ende muchas familias de parálogos, aproximadamente el 30 % del genoma tiene duplicaciones [91, 130]. El impacto de esta redundancia radica en la sobreestimación de SNPs en regiones alineadas incorrectamente. Para evitar esto, se realizó una identificación conservadora de los segmentos ortólogos dentro de los genes (de los genomas de referencia). Partiendo de los alineamientos pareados entre los contigs de cada genoma incompleto y los ORFs (marcos de lectura abierta) de cada genoma completo, ya sea *CFN42* o *CIAT652*, identificamos los segmentos ortólogos (ver Metodología). Dichos alineamientos fueron en primera instancia sin *gaps* y se retuvieron solamente los alineamientos con una cobertura mayor al 60% del gen de referencia y una identidad mayor al 80%. Los alineamientos que pasaron los filtros descritos arriba se realinearon permitiendo la inclusión de *gaps*, siempre y cuando el valor de identidad sea mayor al 85%. De esta manera se identificaron varios

segmentos ortólogos por cada genoma incompleto, cubriendo aproximadamente el 40% del total de genes del genoma de referencia. En lo que respecta a secuencia, la cantidad total colectada por esta metodología fue de 2 a 2.5 Mb por genoma incompleto.

Se hizo una modificación a la metodología cuando se toman dos genomas de referencia al mismo tiempo, para este caso los segmentos ortólogos se obtuvieron a través del programa Mauve [99] previamente estandarizado (ver Metodología). De esta manera podemos excluir los SNPs endémicos de cada genoma completo y por lo tanto, identificar los SNPs de cada genoma incompleto. Por ejemplo, para una posición dada dentro de un alineamiento (una tercia): *CFN42* presenta una “T”, *CIAT652* presenta una “A” y finalmente para el genoma incompleto (en este caso *BRASIL5*) presenta una “A”. Por lo tanto, el SNPs pertenece a *CFN42*. De esta forma, si sólo comparamos contra el genoma de *CFN42*, estaríamos sobreestimando la variación nucleotídica, ya que esta depende de la cercanía filogenética del genoma de referencia con respecto al genoma incompleto. Este artefacto se minimiza cuando utilizamos ambos genomas de referencia al mismo tiempo.

V.1.2 Valoración

Debido a que parte de nuestra muestra (genomas) está secuenciada a una cobertura 1X (aproximadamente el 60% del genoma), implementamos una metodología para una confiable determinación de SNPs. Sin embargo, antes de

utilizarla en nuestros datos, valoramos primero su efectividad en genomas de *E. coli* (ver Metodología). Estos genomas están secuenciados con la tecnología Sanger (Secuenciación Capilar), con una profundidad aproximada de 10X y se obtuvieron de la base de datos (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) del Gen Bank. Para cada genoma simulamos una cobertura 1X, esto se realizó en dos pasos: a) tomando lecturas al azar (en promedio 12,000) y b) re-ensamblando cada grupo de lecturas. Los contigs obtenidos se agruparon en dos conjuntos de datos, genomas con coberturas de 10X y de 1X. A cada conjunto de datos (contigs) aplicamos nuestra metodología y determinamos los SNPs. En promedio, el rango de variación nucleotídica obtenido va de 1% a 2%, tanto para los genomas 10X, como para los 1X. El porcentaje se definió por: número de SNPs/longitud de alineamiento. Al comparar ambos conjuntos, no se encontró diferencia significativa de acuerdo a las pruebas Mann-Whitney y Kolmogorov-Smirnov (p-value menor a 0.05). Este resultado nos indica, que la variación nucleotídica encontrada a partir de genomas 1X, es representativa de la variación total y puede utilizarse para hacer inferencias a nivel de genoma.

V. 2 Caracterización de SNPs en genomas de *R. etli*

En esta parte, identificamos las mutaciones puntuales (SNPs) de los genomas de referencia de *CFN42* y *CIAT652*, ya sea al mismo tiempo o por separados usando nuestra propuesta. Analizamos las diferencias obtenidas de la variación nucleotídica y propusimos una medida de variación promedio.

Finalmente, la variación identificada fue mayor al tomar los dos genomas de referencia al mismo tiempo y se cubrió más del 50% del contenido de genes en ambos genomas.

V. 2.1 Determinación de SNPs en base a un genoma de referencia

Cuantificamos los SNPs en *R. etli*, computando las diferencias entre los alineamientos pareados de los genomas parciales *versus* el genoma completo de *CFN42* o *CIAT652*. En general, la mayor cantidad de SNPs se encontró cuando comparamos contra el genoma de *CFN42*. Sin embargo, nosotros observamos diferencias significativas en los genomas de *BRASIL5* y *8C-3*. El primero presentó una mediana de 5% en la distribución de SNPs con *CFN42* y 2% cuando se alineó contra *CIAT652*, esto indica que el genoma de *BRASIL5* está estrechamente relacionado al de *CIAT652* que al de *CFN42*. En lo que respecta a la varianza en las distribuciones de SNPs, las comparaciones con *CFN42* tuvieron la mayor amplitud. Ésto nos puede sugerir que no existe homogeneidad (poca varianza) dentro de las mutaciones. En lo que respecta al genoma de *8C-3,* el comportamiento de las varianzas fue similar al de *BRASIL5*. Las cepas restantes mostraron niveles similares de variación cuando se alinearon a los genomas completos *CFN42* y *CIAT652*, en promedio 6% y 4% respectivamente, siendo esta última comparación la que presentó el menor nivel de varianza. Por otra parte, las comparaciones entre los genomas completos *CFN42* versus *CIAT652*, tuvieron una mediana de 9% en la distribución de SNPs, este valor es alto en comparación a los obtenidos con los genomas incompletos, sin embargo, es menor al obtenido

(aproximadamente una mediana de 11%) entre la comparación de *CFN42* y *R. leguminosarum viciae 3841*. Esta última comparación fue la que presentó la mayor variabilidad entre los genomas de *R. etli versus R. leguminosarum bv viciae 3841.*

V. 2. 2 Variación nucleotídica promedio

Un objetivo particular importante en este trabajo, fue la obtención de una medida de variación nucleotídica única que represente la variabilidad en los ocho genomas de *R. etli*. Con este fin, nosotros proponemos la variación nucleotídica promedio (VNP), la cual se obtiene promediando las medianas de las distribuciones de SNPs, es decir, número de SNPs/longitud del alineamiento de cada genoma incompleto con respecto a *CFN42* o *CIAT652.* Un componente importante de la VNP a determinar, son los intervalos de confianza, los cuales se generaron y se ajustaron por la distribución de las medianas del tamaño de los genes, ya que no es lo mismo 10 SNP en un gen de 2,000 pares de bases a un gen de 500 pares de bases. La prueba estadística de proporciones se llevó a cabo con el programa PASW Statistics 18 (SPSS Inc., Chicago, IL). Esta prueba compara las proporciones observadas de un evento (en este caso, SNPs) in k muestras (cepas), usando una chi-cuadrada para ver las diferencias significantes entre las proporciones y un subsecuente análisis ajusta los intervalos de confianza para cada muestra. Esta medida puede representar la variación a nivel de especie. La VNP que se obtuvo utilizando los genomas de referencia de *CFN42* y *CIAT652* fue de 4% y 6%, respectivamente, aunque la mayoría de SNPs se obtuvieron en las comparaciones con la *CFN42*. Todos los genomas incompletos son

similarmente divergentes con un intervalo de confianza del 95% con respecto a las medianas. Estas observaciones indican que *CFN42* es igualmente divergente con respecto a las demás cepas. En contraste, las comparaciones con *CIAT652* mostraron que las cepas *BRASIL5* y *8C-3* están más estrechamente relacionadas a *CIAT652* que a *CFN42*. Por otra parte, la cepa que presentó el mayor número de SNPs fue *CIAT894* y como resultado de esta divergencia su mediana se localizó fuera del intervalo de confianza superior, lo cual puede sugerir una lejanía filogenética con respecto a las demás cepas y posiblemente esté relacionada con otra especie. Un panorama diferente se observó en las cepas *BRASIL5* y *8C-3*, las cuales presentaron una VNP aproximada de 1% y sus promedios se localizan por debajo del intervalo inferior, lo cual es un fuerte indicativo de que estas tres cepas (*BRASIL5*, *8C-3* y *CIAT652*), estén dentro de un mismo cluster filogenético. Finalmente, otro ejemplo de gran divergencia entre las comparaciones pareadas, ya sea con el genoma de *CFN42* o *CIAT652*, fue la cepa *IE4771*, sin embargo, ésta siempre se mantuvo dentro de los intervalos de confianza.

V. 2.3 Determinación de SNPS con dos genomas de referencias

La variación nucleotídica promedio fluctúa en razón del genoma de referencia, por lo tanto, determinamos la variabilidad (SNPs) tomando en cuenta dos genomas de referencia al mismo tiempo (ver Metodología). Mediante la construcción de tercias, conformadas por dos genomas completos (*CFN42* y *CIAT652*) y el genoma incompleto en cuestión, abarcamos más de 3,200 fragmentos de genes, los cuales tienen un rango de tamaño de 200 a 4,600 pares

de bases, ésto representa aproximadamente el 60% del contenido génico del genoma de *R. etli*. En lo que respecta a la cobertura, abarcamos más de 2,000,000 de pares de bases, representando aproximadamente el 50% del genoma de *CFN42*. Partiendo de los grupos de segmentos ortólogos (un grupo por cada genoma incompleto), cuantificamos los SNPs en cada tercia (segmento ortólogo) y determinamos los porcentajes de SNPs para cada genoma incompleto. El menor porcentaje lo presentó la cepa *GR56* aproximadamente 6.93%. Por otra parte, las cepas *CIAT894*, *IE4771* y *KIM5* mostraron porcentajes similares que van de 7.28% a 7.45%. Finalmente los mayores porcentajes los tuvieron las cepas *BRASIL5* y *8C-3*, 8.19% y 8.64% respectivamente. Aunque estos valores son mayores en comparación a los obtenidos con un genoma de referencia, presentan una mayor homogeneidad y además no rebasan la mayor variación encontrada de 11% entre *CFN42* y *R. leguminosarum bv viciae 3841*. La variación nucleotídica promedio obtenida a partir de las tercias (Figura 2), fue de 8% y en comparación de las VNP obtenidas de cada genoma de referencia (*CFN42* o *CIAT652*), mostró mayor cohesividad entre todas las cepas, siendo además robusta a la lejanía filogenética de alguno de los genomas de referencia. El genoma de referencia que mostró la mayor variabilidad en todas las comparaciones fue *CFN42.*

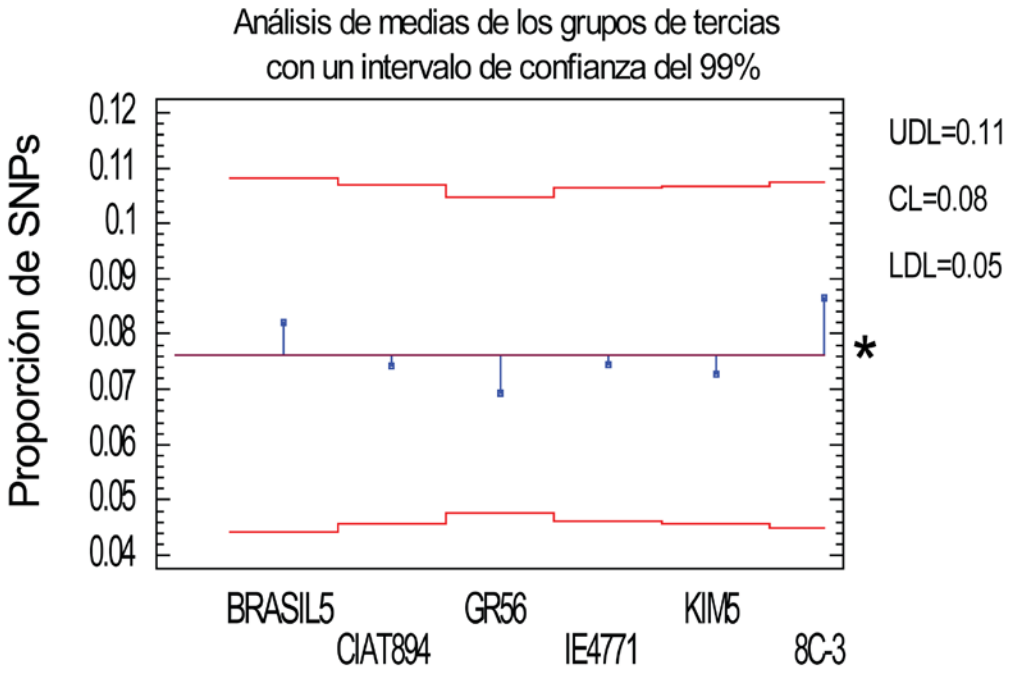

Figura 2. Variación nucleotídica promedio de las tercias (asterisco). Las lineas rojas denotan los intervalos de confianza y los puntos azules son las medianas de las distribuciones de SNPs para cada genoma incompleto. El eje de las Y son las proporciones de SNPs y el eje de las X son los genomas incompletos.

V. 3 Perfiles de variación nucleotídica, detección de recombinación homóloga e inferencia filogenética de regiones compartidas entre genomas de *R. etli*

Este apartado, describe la variación que se encontró en las regiones compartidas de todos los genoma de *R. etli.* A partir de esta variación, se construyeron patrones de SNPs para cada genoma incompleto. La recombinación homologa se valoró dentro de los patrones y además la inferencia filogenética se realizó para cada segmento ortólogo.

V. 3. 1 Patrones de SNPs en regiones compartidas

Dado que los análisis previos, han sido el resultado de alineamientos pareados (*CFN42* o *CIAT62*) o de tercias (*CFN42*, *CIAT652* versus cada genoma incompleto), analizamos las regiones compartidas entre las ocho cepas; de esta manera observamos cómo están distribuidos los SNPs en los genomas de *R. etli*. Partiendo de los segmentos ortólogos identificados por el programa Mauve [99], el siguiente paso, consistió en refinamientos sucesivos para cada segmento ortólogo (ver Metodología), lo que dio como resultado 240 segmentos comunes en todas las cepas, con una mediana en el tamaño de 275 pb y con un tamaño total de 71,630 pb, lo cual representa alrededor del 1% de la longitud del genoma. Estas secuencias se localizaron principalmente en el cromosoma de *CFN42* y *CIAT652* (92% de 240 segmentos) y la proporción restante se localizó en los plásmidos p42b, p42d, p42e y p42f pertenecientes a la *CFN42*. Con el fin de observar los patrones de SNPs, concatenamos los 240 segmentos ortólogos y entonces inferimos la secuencia consenso y computamos el número de diferencias nucleotídicas a través de ventanas independientes de 250 pb. En dado caso que el consenso este conformado por dos tipos de bases en un porcentaje igual al 50% tomamos como referencia la base contenida en el genoma de *CIAT652* (ya que este mostró mayor cohesión en las comparaciones previas). Encontramos tres tipos de patrones: bialélicas, polialélicas y mutaciones únicas. Los patrones bialélicos incluyen solamente una mutación con respecto al consenso y estos patrones fueron los más comunes en toda la región concatenada. Los patrones

polialélicos son aquellos que presentaron múltiples diferencias en un mismo sitio nucleotídico con respecto al consenso, los cuales pueden ser indicativos de *hotspot* de mutación. Finalmente, las mutaciones únicas fueron aquellas que estuvieron presentes en una sola cepa. Estos patrones en altas frecuencias pueden ser indicativos de cepas atípicas o filogenéticamente lejanas del grupo, tal fue el caso, del genoma de *CFN42* y en menor grado el genoma de *CIAT894*. Ambas cepas pueden considerarse como linajes independientes al resto del grupo. En general, se observaron patrones compartidos entre dos o más cepas, lo cual puede ser una valoración cualitativa de posibles eventos de recombinación, esta misma estrategia se utilizó en regiones compartidas con el plásmido simbiótico [135]. Sin embargo, la valoración cualitativa de dichos patrones de SNPs compartidos no es práctica, ya que al aumentar el número de genomas, aumenta la combinación de patrones a lo largo de los genomas.

V. 3. 2 Detección de Recombinación en regiones compartidas

Para cada uno de los 240 segmentos ortólogos obtenidos de las ocho cepas, identificamos los eventos de recombinación homóloga dentro de cada segmento, siempre y cuando dicho evento lo avale al menos dos algoritmos independientes para detectar recombinación (ver Metodología). Solamente nueve segmentos ortólogos presentaron eventos de recombinación. Ocho de nueve eventos ocurrieron en el cromosoma y el evento restante se ubicó en el plásmido p42e con referencia a *CFN42*. Cabe destacar, que los nueve eventos de

recombinación coinciden con segmentos ortólogos que tienen patrones de SNPs compartidos por más de dos cepas (previamente descritos), ésto reafirma que cada evento de recombinación tiene un soporte cualitativo y cuantitativo. Estos eventos de recombinación representan una baja frecuencia de recombinación y aunque esta propuesta está limitada por la cantidad de segmentos entre las ocho cepas, cubrimos aproximadamente el 3.7% (223) del contenido total de genes (5,963) en el caso del genoma de referencia *CFN42*. Además los segmentos ortólogos están distribuidos entre las cuatro principales categorías y subcategorías de la base de datos de los COGs (ver Metodología). Por ejemplo, metabolismo (transporte y metabolismo de azúcar, amino ácidos y carbohidratos); procesos celulares y señalización (biogénesis y transducción de señales); almacenaje de información y procesamiento (transcripción, replicación y recombinación) y finalmente, proteínas pobremente caracterizadas (función desconocida). Por lo tanto, los segmentos ortólogos que se obtuvieron no tuvieron algún sesgo hacia una clase funcional, ni son genes ribosomales o genes de mantenimiento central, los cuales son generalmente utilizados en análisis en poblaciones bacterianas [136].

V. 3. 3 Inferencia filogenética

Uno de los efectos de la recombinación, filogenéticamente hablando, se observa en las topologías resultantes de diferentes genes de una misma especie. Si existe diferencia entre ellas, se dice que existe incongruencia filogenética, dando como resultado diferentes escenarios evolutivos para cada gen. Por lo

tanto, decidimos hacer las construcciones filogenéticas para cada una de las 187 regiones compartidas entre los ocho genomas de *R. etli* y el genoma de *R. leguminosarum bv viciae 3841*. Este último se utilizó cómo grupo externo, debido a la gran sintenia compartida con *CFN42*. Las construcciones filogenéticas para cada segmento se hicieron con el método de máxima verosimilitud y con el fin de determinar el posible árbol de la muestra, hicimos un consenso a partir de cada una de las topologías (ver Metodología). Otra propuesta utilizada para obtener el posible árbol de la especie, es la concatenación de todos los genes (segmentos ortólogos compartidos). De hecho, nosotros utilizamos tres métodos para hacer la inferencia filogenética del concatenado. En primer lugar, utilizamos el método de máxima verosimilitud (*maximum likelihood*), seguido por el método bayesiano (*bayesian*) (Figura 3A) y por último, utilizamos una red de vecinos cercanos (*neighbor joining network*). Tanto el árbol consenso, como los arboles obtenidos mediante los concatenados (máxima verosimilitud, bayesiano o red de vecinos) presentaron la misma topología y para fines prácticos utilizamos como base la filogenia obtenida a través de red de vecinos (Figura 3C). Este árbol consistió de seis nodos internos y mediante estos nodos observamos dos principales grupos, el primero constó de las cepas de *KIM5*, *IE4771* y *GR56* y el otro grupo incluyó a *BRASIL5*, *8C-3* y *CIAT652*. Por otra parte, las ramas internas que separan estos dos grupos con *CFN42*, *CIAT894* y *RLEG* (*R. leguminosarum*) son más largas que el resto (presentan un gran número de sustituciones por sitio) y son filogenéticamente equidistantes, es decir, de una longitud similar. Cabe resaltar,

que las inconsistencias encontradas entre las topologías de cada uno de los segmentos ortólogos (regiones compartidas entre todas las cepas) y la red de vecinos fueron mínimas (Figura 3B). Estas topologías alternativas, son resultantes del reposicionamientos de las cepas *CIAT894* y *RLEG* dentro de la filogenia. 30 segmentos soportan a *RLEG* como la cepa más distante, mientras que 39 segmentos colocan a *CIAT894* como grupo externo. Sin embargo, las incongruencias filogenéticas encontradas podrían ser el resultado de polimorfismos ancestrales, debido en parte, a las largas ramas terminales y a las bajas frecuencias de recombinación. Por lo tanto, podemos sugerir que los niveles de recombinación identificados son insuficientes para borrar la señal filogenética y solamente el 3.75% de 223 segmentos de genes comunes entre las cepas de *R. etli* presentaron eventos de recombinación.

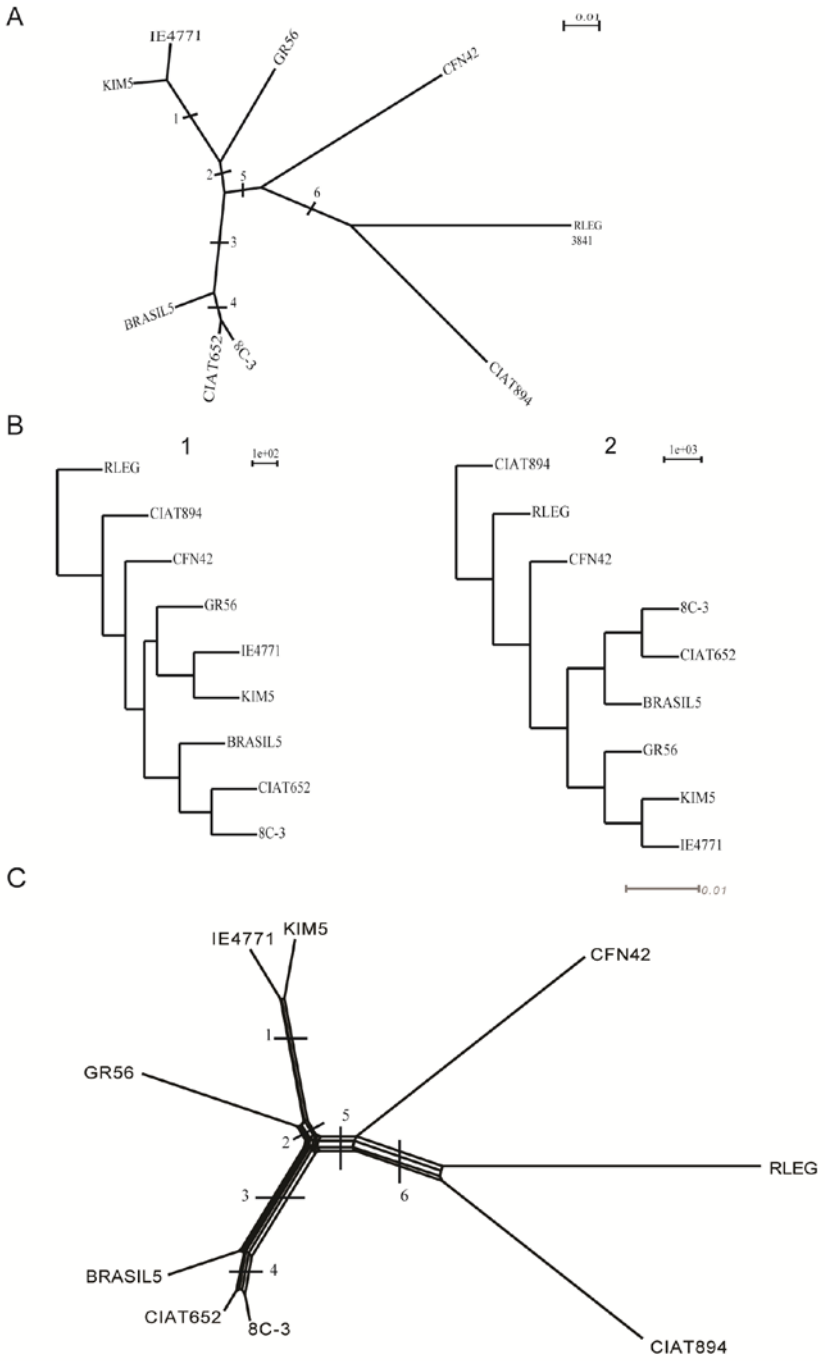

Figura 3. Inferencia filogenética. A) Topología inferida a partir de 186 regiones compartidas entre los genomas muestreados. El análisis filogenético se realizó a través del método bayesiano. El árbol está sin raíz y se obtuvieron 6 biparticiones mostradas por lineas de separación indicadas por números dentro de las ramas internas. B) Topologías incongruentes de los árboles de genes individuales contra la topología de la muestra. Cada árbol individual se construyó usando el método de máxima verosimilitud. C) Análisis de redes de las 186 regiones (descritas en A). Cada bipartición se localizó mediante números en las ramas internas. En todas las topologías la barra de escala denota el número esperado de substituciones nucleotídicas por sitio.

V. 4 Extensión en la detección de la recombinación

En esta sección, diseñamos una propuesta para ampliar la búsqueda de los eventos de recombinación para cada uno de los genomas incompletos. Además, caracterizamos y determinamos la diversidad nucleotídica tanto para los segmentos recombinantes como para los no recombinantes. Finalmente localizamos cada segmento recombinante en los genomas de referencia, desde replicones hasta nivel de operón.

V. 4. 1 Construcción de cuartetos y detección de recombinación

Los eventos de recombinación detectados previamente se analizaron en base a los segmentos de genes ortólogos (regiones compartidas entre todos los genomas de *R. etli*) principalmente distribuidos en el cromosoma y en algunas regiones de plásmidos. Los segmentos ortólogos representaron el 3% del contenido génico y además la mayoría de clases funcionales de los COGs estuvieron representadas. Sin embargo, para ampliar la muestra dentro de los genomas incompletos, construimos grupos de cuartetos (segmentos de genes alineados), tomando como fondo genético a los genomas de referencia de *CFN42* y *CIAT652* de *R. etli* y el genoma de *3841* de *R. leguminosarum* (Ver Metodología). Se obtuvieron seis grupos de cuartetos, uno por cada genoma incompleto, por ejemplo, para el genoma de *BRASIL5* en promedio se construyó 2,781 cuartetos y para el genoma de *CIAT894* fueron 3,672. Los tamaños de los cuartetos fueron de 200 a 4651 pares de bases y cubren aproximadamente el 50%

del genoma. Para cada grupo de cuartetos, detectamos los eventos de recombinación dentro de cada segmento ortólogo (cuarteto), siempre y cuando el evento de recombinación esté avalado por dos algoritmos. Los porcentajes de cuartetos recombinantes son heterogéneos, por ejemplo la menor proporción fue de 4.42% (123 de 2,781) para *BRASIL5* y 3.57% (102 de 2,854) para *8C-3* respectivamente. Los genomas restantes mostraron aproximadamente el doble de proporciones. En el caso de los cuartetos de *KIM5* se encontraron 8.67% y 10.86% para los cuartetos de GR56. Por otra parte, cuantificamos la participación de cada una de las cepas en los eventos de recombinación, ya que las proporciones de cuartetos recombinantes no reflejan la participación del genoma incompleto que se utilizó para su construcción, por ejemplo, en un cuarteto derivado de la cepa *BRASIL5*, se detectó un evento de recombinación y en éste participaron las cepas de *CFN42* y *CIAT652*, por lo tanto, reordenamos los eventos de recombinación en función a quienes participan (par de cepas). En general la recombinación fue mayor entre las cepas de *R. etli*, sin embargo se detectó participación de la cepa *3841* de *R. leguminosarum*. En el caso de los cuartetos obtenidos de la cepa *BRASIL5*, armamos los pares de cepas siguientes: *CFN42-BRASIL5*, *CFN42-CIAT652*, *BRASIL5-CIAT652*, *CFN42-RLEG*, *BRASIL5-RLEG y CIAT652-RLEG*. Los menores porcentajes los presentaron los tres últimos pares, 7%, 5% y 20%, respectivamente y los pares restantes presentaron valores mayores al 18%. Observamos patrones similares en los cinco grupos restantes (*CIAT894*, *IE4771*, *GR56*, *KIM5* y *8C-3*). Este resultado puede deberse a que la

recombinación homóloga depende de una alta identidad nucleotídica, es decir, a mayor divergencia, menor recombinación y viceversa [66]. Por lo tanto, la recombinación puede ser más frecuente dentro de las poblaciones que entre ellas. De hecho, encontramos a un mismo gen recombinante en dos o más grupos de cuartetos y en general, el número de eventos de recombinación comunes fue pequeño. Las cepas que presentaron la mayoría de eventos en común fueron *BRASIL5* y *8C-3*, esto es debido al resultado de su cercanía filogenética. Este análisis se llevó a cabo mediante una matriz de presencia/ausencia entre los cuartetos recombinantes.

V. 4. 2 Distribución de cuartetos recombinantes en el genoma

Al igual que las regiones compartidas entre todos los genomas, los cuartetos recombinantes se localizaron principalmente en el cromosoma y en los replicones de p42d, p42e, p42f del genoma de referencia *CFN42* (Figura 3). El mismo sesgo cromosomal se observó, cuando tomamos como referencia al cromosoma de *CIAT652*, pero a diferencia de *CFN42*, sólo se encontraron dos plásmidos (pA y pC). La distribución de los cuartetos recombinantes fue homogénea (a lo largo del replicón), tanto en el cromosoma, como para los plásmidos, ya sea tomando a *CFN42* o *CIAT652* como genomas de referencia.

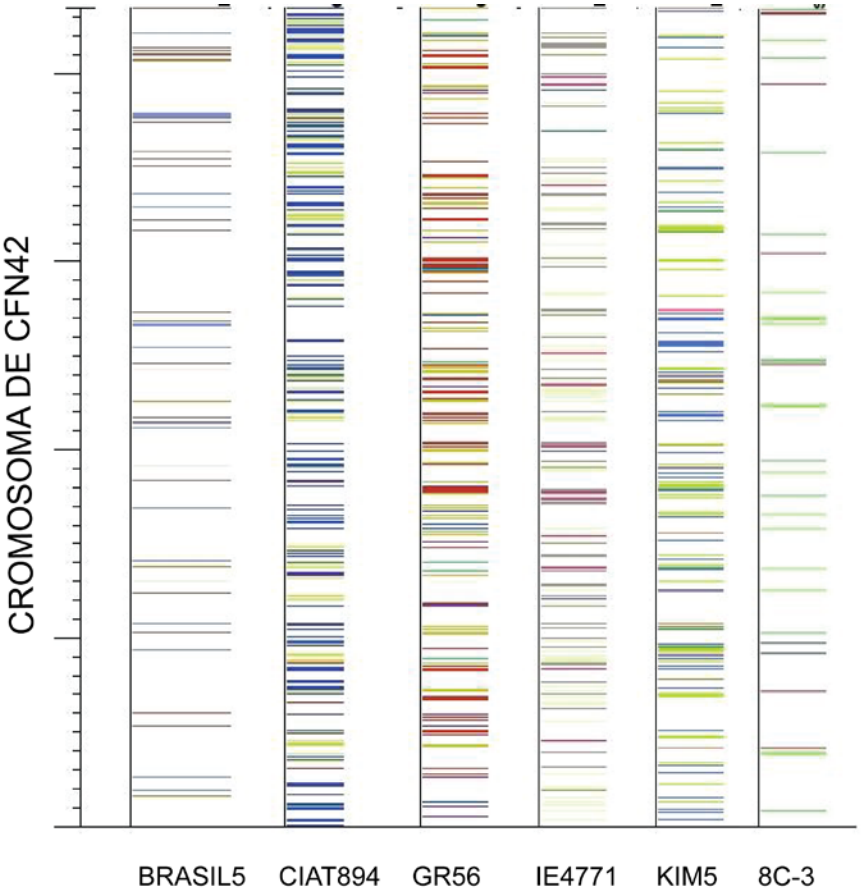

Figura 4. Cuartetos recombiantes localizados en el cromosoma de *CFN42*. Cada línea equivale a un evento de recombinación. El eje de las Y son las coordenadas del cromosoma y el eje de las X son los genomas incompletos.

Sin embargo, existe evidencia que la recombinación puede observarse en regiones específicas, también llamadas *hostspots* de recombinación o en bloques conservados de hasta 2000 pares de bases y además al disminuir el tamaño de los bloques aumenta la frecuencia de recombinación [137]. En el caso de *R. etli*, el

tamaño promedio de los cuartetos recombinantes fue menor a 1,000 pares de bases y además los eventos de recombinación se disgregaron a lo largo de los replicones. Debido a estas diferencias, decidimos valorar el alcance de la recombinación a nivel de operón, ya que representan unidades funcionales y se ha señalado que los genes que lo componen evolucionan de la misma manera o tienden a conservarse juntos. En primer lugar, determinamos la estructura de operón de los genomas de referencia: *CFN42*, *CIAT652* y *Leguminosarum bv viciae 3841,* usando el programa Uber-operon (ver Metodologia) [126]. La base de datos utilizada para predecir operones incluyó a los genomas de: *CFN42*, *CIAT652*, *R. leguminosarum trifolii WSN2304*, *R. Leguminosarum bv viciae 3841* y *S. meliloti 1012.* En segundo lugar, cuantificamos el número de genes localizados en las estructuras de operon predichas para cada genoma de referencia. Para realizar la cuantificación identificamos los socios recombinantes (aquellas cepas que participaron en un evento de recombinación): *CFN42-CIAT652, CFN42-RLEG* y *CIAT652-RLEG*. Nosotros encontramos que el 40% de los genes de los genomas de *CFN42*, *CIAT652* y *RLEG* están organizados en operones (Figura 4A). En el caso de los genomas incompletos, no es posible predecir la estructura de operon debido por su naturaleza fragmentada, en su defecto, identificamos los eventos de recombinación donde haya participado dicho genoma incompleto. Una vez identificados los genes recombinantes, los mapeamos contra los operones de *CFN42* (Figura 4B). Los genes recombinantes en los genomas completos, se localizaron en un operón, siempre y cuando el operón esté conservado en ambos

genomas. Del total de los segmentos recombinantes, solo se mapearon aproximadamente el 23%, tanto en los genomas completos como en los genomas incompletos.

A)

Genomas	Operones	Genes	Porcentaje de genes en operones
CFN42	505	2468	41
CIAT652	545	2541	42
RLEG	620	3101	43

B)

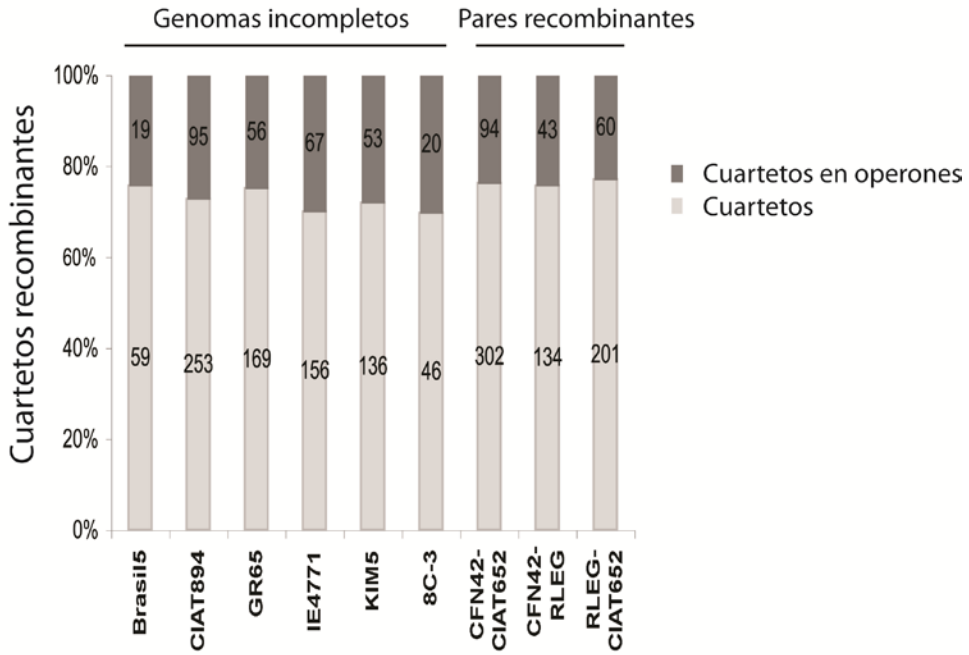

Figura 5. Determinación y cuantificación de operones. A) Identificación de operones en los genomas de referencia. B) Mapeo de cuartetos recombinantes. El eje de las Y denota los porcentajes de cuartetos recombinantes y el Eje de las X se divide en dos partes, genomas incompletos y pares recombinantes (solamente para genomas completos). Los cuartetos que No están en operones son denotados por las barras de color gris claro, en contraparte, las barras de color gris obscuro son los cuartetos que están incluidos dentro de un operón. Los números dentro de las barras son la frecuencia de los cuartetos recombinantes.

Las proporciones de genes recombinantes dentro de operones fueron menores a lo esperado en los genomas de referencia. Cabe destacar, que los genomas de *IE4771* y *8C-3* presentaron un pequeño incremento en la proporción promedio de operones mapeados. Este resultado podría sugerir que la estructura de operón es menos propensa a presentar eventos de recombinación, debido tal vez a una restricción funcional, por ejemplo, genes del mantenimiento central o adaptativos.

V. 4. 3 Clases funcionales en cuartetos recombinantes

Al examinar las clases funcionales (COGs) en las regiones compartidas entre todas las cepas de *R. etli*, encontramos que su distribución fue heterogénea (representadas en todas las clases funcionales). De igual manera, hicimos la valoración de las clases funcionales para los grupos de cuartetos recombinantes. Sin embargo, debido al diseño de la construcción de los cuartetos, utilizamos las clases funcionales presentes en el genoma de *CFN42* para determinar si existe diferencia significativa contra los genes recombinantes. Aunque todas las clases funcionales estuvieron representadas en los genomas incompletos, encontramos sesgos en las categorías de: transporte de amino ácidos y metabolismo, transporte de carbohidratos y metabolismo, producción de energía y conversión, transporte de lípidos y metabolismo, predicción de función general y finalmente función desconocida. En contraparte, las categorías subrepresentadas fueron: transcripción y transducción de señales. A pesar de que las diferencias se evaluaron mediante las pruebas estadísticas de Ji-cuadrada y Rangos [102], la

naturaleza fragmentada de los genomas incompletos no permite concluir que la recombinación actúa preferentemente en algunas clases funcionales, pero cabe destacar el sesgo aparente a las categorías de metabolismo a pesar de la muestra.

V. 4. 4 Diversidad Nucleotídica en cuartetos

Los análisis descritos arriba pueden sugerir que la recombinación no juega un rol importante en el manejo de la diversidad genómica en *R. etli*, sin embargo esta puede tener un efecto relativo limitado, ya que a bajas frecuencias, la recombinación puede tener un carácter diversificador [138]. Por lo tanto, para examinar este punto, estimamos el promedio de la diversidad nucleotídica por sitio (π) para cada grupo de cuartetos, tanto para los segmentos recombinantes, cómo para los no recombinantes. En general, los segmentos recombinantes mostraron valores mayores de π en un rango de 0.09 a 0.115. Estas diferencias fueron significativas solamente para las cepas *CIAT894*, *GR56*, *IE4771* y *KIM5*, de acuerdo a la prueba estadística t-student´s, con un $p < 0.001$, pero los datos de diversidad obtenidos de las 240 regiones concatenadas, entre ellas recombinantes y no recombinantes de los ocho genoma de *R. etli* fue menor que los cuartetos (en promedio 0.06). Además las regiones compartidas no presentaron diferencias significativas entre los segmentos recombinantes y los no recombinantes, pero cabe destacar que los promedios de diversidad fueron mayores en los segmentos recombinantes que en los no recombinantes. Aunado a esto, calculamos las proporciones de transiciones/transversiones (ts/tv), y dado que la probabilidad de

transiciones es mayor a las de transversiones [139], las altas proporciones de ts/tv pueden sugerir la presencia de selección purificadora, resultado de transiciones en la tercera posición de un codón, que generalmente son mutaciones sinónimas [140].

V. 4. 5 Presión selectiva en cuartetos recombinantes

La mayoría de los cuartetos recombinantes incluyeron genes con funciones metabólicas, sin embargo, sólo el 23% se encuentran dentro de operones. Además el tamaño promedio de los segmentos recombinantes fue menor a 1,000 pares de bases. La dispersión de los eventos recombinantes en todo el genoma y la exclusión de los operones, podría sugerir una fuerte selección en todo el genoma. Una comparación previa tomando los genomas de *CFN42* y *R. leguminosarum bv viciae 3841* mostró una fuerte selección purificadora [100], pero el incremento de la diversidad nucleotídica en los cuartetos recombinantes, podría sugerir variaciones en el tipo de selección. Por lo tanto, determinamos las tasas de sustitución sinónima (dS) y no sinónima (dN), tanto para los cuartetos recombinantes como para los no recombinantes de cada genoma incompleto. En este análisis, exploramos si el incremento de la diversidad nucleotídica se acompaña de un incremento en las tasas dN, ya que esta relación podría sugerir una selección adaptativa. Alternativamente, un incremento en las tasas dN y dS podría indicar un incremento de la deriva génica. Otro posible escenario, es la disminución de las tasas dN y dS lo que pudiera indicar una selección relajada [141]. En la figura 5, se observa un sobrelape de los segmentos recombinantes

(esferas rojas) y los no recombinantes (esferas azules), dando como resultado un agrupamiento compacto de ambas tasas (dN y dS). Este sobrelape sugiere que todos los cuartetos están sujetos al mismo tipo de selección, con un fuerte sesgo hacia dS. Sin embargo, en base a los valores de dS, se obtuvieron dos grupos: aquellos con un rango mayor de 0.54 a 0.56 (*GR56*, *KIM5* y *IE4771*) y aquellos con un rango menor a 0.47 a 0.50 (*CIAT894*, *8C-3* y *BRASIL5*). Cabe destacar que la compactación forma un agrupamiento principal, identificado en la Figura 5 mediante un círculo con líneas cortadas, los cuartetos que se extienden o se alejan del agrupamiento principal (centro de sobrelape), pueden sugerir variaciones en el tipo de selección principal delimitado por los promedios de ambas tasas (dS y dN). La dirección de las desviaciones implica incrementos o disminuciones en las tasas dN y dS, dando como resultado un tipo de selección local, el cual existe a pesar del dominio de la selección purificadora. Por ejemplo, la cepa *BRASIL5* tiene un centro de sobrelape en las coordenadas X (0.48-0.53) y coordenadas Y (0.0347-0.042) y observamos claramente dos grupos de genes, el primero tiene un incremento en la tasa dN y un decremento en la tasa dS, lo cual refleja un evento de selección adaptativa y el segundo presentó solamente incremento en la tasa dS. En algunos casos, las desviaciones tienden ir acompañadas por disminuciones en los valores de diversidad nucleotídica (en la Figura 5, se puede ver como una disminución en el área de la esferas), esto refleja la presencia de selección purificadora. Ejemplo de ésto, lo presentó la cepa

CIAT894, donde un grupo de genes se separó del agrupamiento principal, mostrando un incremento en la tasa dS y una caída en la diversidad nucleotídica.

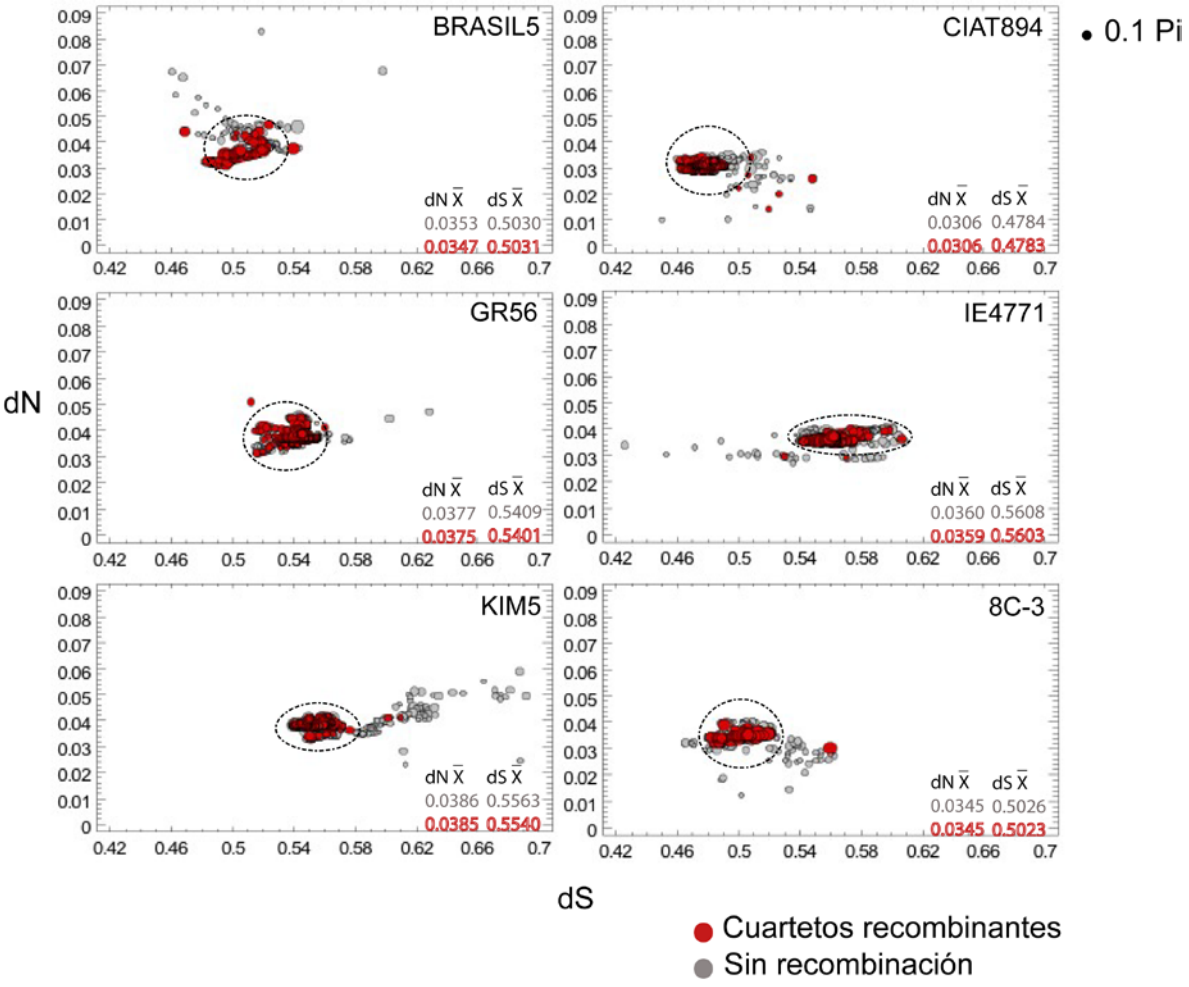

Figura 6. Selección y diversidad nucleotídica. Las tasas de sustitución no sinónima (dN, eje de las Y) y sinónima (dS, eje de las X) fueron determinadas para cada grupo de cuartetos recombinantes, obtenidos de los alineamientos entre cada genoma incompleto y de los genomas de *CFN42*, *CIAT652* y *R. leguminosarum bv viciae 3841* (Ver Metodología). Tanto para cuartetos recombinantes (esferas rojas) como no recombinantes (esferas gris) determinamos la diversidad nucleotídica Pi, denotada por el tamaño de las esferas. Los promedios de ambos tipos de tasas (dS y dN) se muestran en la parte inferior derecha de cada grafica.

CAPITULO VI.- DISCUSION

VI. 1 Valoración de la metodología

En el presente trabajo, usamos una propuesta genómica para detectar los SNPs y medir su variación, además valoramos la contribución de la recombinación en la diversificación genómica en los genomas de *R. etli*. Nuestros resultados demostraron, que las muestras de secuencias genómicas a bajas coberturas (aproximadamente de 1X del genoma) pueden usarse para medir la variación a nivel de genoma completo en esta especie. En *R. etli*, encontramos en los genomas incompletos una gran cantidad de variación nucleotídica (más de 161,998 SNPs), tomando como referencias los genomas de referencias *CFN42* y *CIAT652*. Con el fin de valorar la confiabilidad de nuestro método para identificar SNPs, cuantificamos los SNPs en genomas de *E. coli* a una cobertura de 1X y 10X. En dicha cuantificación encontramos el mismo nivel de variación, tanto en los genomas incompletos de *E. coli* como en los completos. Este resultado indica que las estimaciones de la variabilidad que se obtuvieron en los genomas incompletos reflejan la variabilidad total del genoma. Richter y Rosselló-Mora [142], reportaron que pueden usarse secuencias genómicas parciales de varias especies bacterianas para inferir valores confiables de divergencia de DNA, siempre y cuando las secuencias parciales cubran alrededor del 20% del genoma de referencia. Además, mostraron que los valores de identidad nucleotídica promedio (por sus siglas en inglés ANI) obtenidos de esas muestras correlacionan con los

valores de hibridación DNA-DNA (por sus siglas en inglés DDH), e indican que las secuencias de genomas incompletos son un recurso aceptable para medir la variación nucleotídica. A la fecha, el rápido desarrollo de las tecnologías de secuenciación masiva de DNA, ya sea *454-Roche*, *Illumina*, o *SOLiD4* permite a los investigadores el uso de secuenciación *multiplex* (meter varias muestras dentro de una línea de secuenciación) con el fin de abarcar un mayor número de cepas y aunque estos experimentos dan una gran cobertura, los genomas generalmente quedan fragmentados (multiples contigs), y con valores de calidad heterogéneos. Por lo tanto, nuestra propuesta metodológica también puede utilizarse para el análisis de la variación nucleotídica, en estos genomas incompletos.

VI. 2 Variación nucletídica promedio

La variación identificada en los genomas de *R. etli* fue mayor a la que se encontró en los genomas de *E. coli*. Hecho que puede interpretarse que la primera especie tiene mayores tasas de mutación o es un linaje más antiguo, y no a un artefacto de secuenciación o de cobertura. La divergencia entre las cepas de *R. etli* y de *R. leguminosarum bv viciae 3841* tuvo un rango del 7% al 11% (medianas de las distribución de SNPs), este último valor se obtuvo con el genoma de *CFN42*. *R. etli* y *R. leguminosarum* son especies distintas, de acuerdo a las comparaciones hechas con el 16S ribosomal. Sin embargo, estas cepas comparten un genoma core común y además poseen un componente accesorio variable (plásmidos) [100, 130, 143]. Por lo tanto, un rango de variación nucleotídica promedio (VNP) del 7% al 11% puede ser un buen indicador de

especiación dentro de *Rhizobium*, a pesar de las diferencias de VNP (4%-6%) entre las cepas muestreadas de *R. etli*. Los niveles de VNP fueron mayores en la cepa *CFN42* (la cepa tipo) [78, 144], que en la *CIAT652*. En el presente análisis, encontramos que la cepa *CFN42* fue la más divergente entre las cepas estudiadas, ya que tiene la mayor proporción de SNPs únicos y se agrupó como una rama independientemente divergente cuando se exploró la filogenia. Recientemente resecuenciamos la cepa *CFN42* usando tecnología de secuenciación masiva *Solexa-Illumina* y comparamos la secuencia obtenida contra el genoma completo: encontramos muy pocos indeles y menos de 20 SNPs, así como ningún rearreglo genómico. Por lo tanto, muy pocas variaciones pueden ser esperadas por el manejo *in vitro* dentro del laboratorio. En contraste, la mayoría de las cepas analizadas fueron más cercanas filogenéticamente a la cepa *CIAT652* que a la *CFN42*. Un estudio reciente indica que las cepas *CIAT652* y *CFN42* tienen valores bajos de identidad nucleotídica promedio: alrededor de 90.44% [142] y sugiere que la cepa *CIAT652* está erróneamente caracterizada como *R. etli*. Sin embargo, previamente mostramos que *CFN42* y *CIAT652* comparten regiones altamente conservadas del plásmido simbiótico, pero tienen alta divergencia entre el resto de sus genomas [130]. Dado que todos los aislados de *R. etli* se colectaron a través de nódulos de frijol que fijan nitrógeno, esta característica pudiera ser la dominante en el criterio de clasificación. La divergencia descrita aquí es por lo tanto consistente con un modelo en el cual la especie de *R. etli* está compuesta por linajes genómicos divergentes que

comparten el fenotipo simbiótico conferido por los plásmidos simbióticos [130], el cual es llamado como un simbiovar común [145]. De hecho, nuestros análisis sugieren, que en algunos casos el uso de la cepa tipo (*CFN42*) podría conducir a una clasificación taxonómica engañosa, especialmente cuando los mecanismos de la transferencia de genes están activos. *R. etli* claramente posee elementos móviles como plásmidos conjugativos, secuencias de inserción y bacteriófagos [146-148]. Por lo tanto, el flujo de genes y la recombinación entre la cepas *R. etli*, pueden ser importantes para la producción de una diversidad genómica que se refleja claramente en su estructura pangenómica [130].

VI. 3 Recombinación homóloga e inferencia filogenética

En trabajos recientes concluyen que poblaciones de *R. etli* son esencialmente clonales con pocos eventos de recombinación, incluso en poblaciones simpátricas, usando *MLEE* o *MLST* [149-152]. Por otra parte, Flores et *al.* [135] mostraron que a pesar de la alta conservación de 6 regiones del plásmido simbiótico (cepas colectadas de distinto origen geográfico) algunas regiones compartieron perfiles idénticos de SNPs. Esta observación se interpretó como evidencia de recombinación homóloga. En este trabajo, obtuvimos resultados similares para un conjunto de segmentos de DNA genómicos comunes entre las ocho cepas de *R. etli*, distribuidos principalmente en el cromosoma y en algunos plásmidos. La cuantificación de los probables eventos de recombinación y la extrapolación de nuestros resultados al nivel de genoma completo sugieren que un mínimo de 260 eventos de recombinación podrían ocurrir en el genoma de

cada cepa. Sin embargo, las cepas *CFN42* y *CIAT894* fueron las más variables en término de SNPs y ésta última también mostró la mayoría de eventos recombinantes en nuestro análisis de cuartetos (dentro de los segmentos ortólogos).

A pesar de que existieron discrepancias dentro de algunos nodos de varios de los árboles filogenéticos que generamos, la mayoría de arboles fueron congruentes con el árbol consenso. Independientemente de las incongruencias entre las cepas *CIAT894* y *R. leguminosarum bv viciae 3841*, en todas las topologías se mostraron relacionadas a pesar del hecho de que la cepa *CIAT894* se clasificó como *R. etli* [150] y además, la cepa *CIAT894* presentó la mayor proporción de recombinación con *R. leguminosarum*, lo que sugiere un flujo de genes entre ambas especies. Esta hipótesis se soporta a nivel de población, ya que existe evidencia de recombinación en el gen 16S ribosomal entre aislados de *R. etli* y de *R. leguminosarum* en distintas localidades de México y Colombia [153]. Esta recombinación ribosomal también se detectó en aislados de *Bradyrhizobium* y la frecuencia de recombinación entre cepas fue baja [154]. Algunas veces, esta recombinación puede impedir una identificación confiable de las especies, como es el caso de *R. gallicum* y de *R. mongolense* [155].

VI. 4 Contribución de la recombinación homologa

La contribución de la recombinación homóloga en la evolución bacteriana no está del todo clara. En algunos casos, la recombinación puede promover la

diversidad nucleotídica [66]. A nivel de especies, la mayoría de diversidad es intragénica, contenida en pequeños segmentos de genes distribuidos en todo el genoma y algunos de los *loci* que muestran microdiversidad pueden originarse por recombinación [156]. En *R. etli*, aunque la recombinación estimada fue proporcional a la diversidad nucleotídica (mayor frecuencia de recombinación mayor diversidad), las frecuencias de recombinación fueron bajas (3%-10%) y además el tamaño promedio de los segmentos recombinantes fue de 623 pares de bases, distribuidos de manera homogénea a través de todo el genoma. Esto correlaciona con los sitios de microdiversidad encontrado en otras bacterias. Las frecuencias de recombinación que se determinaron para *R. etli* fueron pequeñas en comparación a las obtenidas en los géneros *Rickettsia* y *Streptococcus*, los cuales tuvieron un rango de 18%-37% y 28%, respectivamente [157, 158]. Aunque estos análisis no son comparables debido a que la detección de la recombinación se realizó dentro del genoma core, nuestros datos sugieren que solamente una fracción menor del genoma de *R. etli* está bajo recombinación, la cual representa solamente una pequeña proporción de los polimorfismos totales en esta especie. Sin embargo, esta pequeña porción recombinante, presentó un sesgo aparente en las categorías funcionales de “transporte de lípidos, aminoácidos y carbohidratos”, así como “metabolismo”. Este sesgo funcional podría promover adaptación a nuevos ambientes y en el caso de *R. etli* podría promover el incremento de los fenotipos simbióticos o habilidad competitiva. Esto se ha visto en los genomas de cepas patogénicas de *E. coli*, que mostraron evidencia de un extensivo

intercambio de elementos móviles, que generó complejos multiclonales capaces de adaptarse a nuevos huéspedes [159]. Un panorama similar se presentó en genomas de *Bacillus*, Alcaraz et *al.* encontraron que las diferencias funcionales promueven la diferenciación de grupos taxonómicos y además la mayoría de la variación ocurren en los genes que se requieren para responder al estrés ambiental [160]. De hecho, la mayoría de segmentos recombinantes no están en operones ya que más del 76% forma parte de genes individuales. Estos segmentos pequeños de genes favorecen la disgregación de la microdiversidad en todo genoma. Este “tiroleado” de diversidad, aun en frecuencia baja, está limitado por una selección purificadora que actúa en todo el genoma, pero también existe variaciones en los tipos de selección ya sea en los genes recombinantes o en los no-recombinantes.

Finalmente, la frecuencia de recombinación homóloga mediada por *RecA* depende del nivel de identidad de DNA en las bacterias. Los pequeños fragmentos de DNA son a menudo introducidos en la célula vía conjugación, transformación o transducción. Estos fragmentos recombinantes, presentan una divergencia mayor al promedio del genoma, ocasionando que solamente una fracción del DNA total pueda ser alcanza por la recombinación homóloga [161]. Los genomas de *R. etli* presentaron una variación nucleotídica promedio del 8%, esta gran divergencia es un factor importante para delimitar la acción de la recombinación. Otros factores pueden influir las bajas frecuencias de recombinación detectadas en los aislados de *R. etli*, como son: su origen biogeográfico distante (USA, México, Costa Rica,

Colombia, Brasil y España) [130] y el pequeño número de cepas analizadas. Bailly et *al.* obtuvieron resultados similares en análisis de genómica poblacional de cepas simpátricas de *S. medicae* [162]. Ellos encontraron niveles muy bajos de polimorfismos y de recombinación en el cromosoma en comparación a los megaplásmidos. Nuestros resultados y algunos reportes previos en *R. etli* apoyan un modelo en cual la especie está compuesta por linajes evolutivos independientes, con bajos niveles de recombinación entre ellos, pero que comparten un fenotipo simbiótico. Sin embargo, aunque las barreras genéticas impuestas por la divergencia nucleotídica, así como las barreras biogeográficas y la selección purificadora evitan o regulan la recombinación entre las cepas, el flujo de genes (entre poblaciones) y la transferencia lateral (mediante plásmidos e islas cromosomales) podrían ser determinantes en la estructura pangenómica de *R. etli*.

REFERENCIAS

1. **Willems A:** The taxonomy of rhizobia: an overview**.** ***Plant Soil*** **2006,** 287**:3-14.**

2. **Demont N, Debelle F, Aurelle H, Denarie J, Prome JC:** Role of the *Rhizobium meliloti* nodF and nodE genes in the biosynthesis of lipo-oligosaccharidic nodulation factors**.** ***J Biol Chem*** **1993,** 268**(27):20134-20142.**

3. **Cooper JE:** Early interactions between legumes and rhizobia: disclosing complexity in a molecular dialogue**.** ***J Appl Microbiol*** **2007,** 103**(5):1355-1365.**

4. **Schultze M, Kondorosi A:** Regulation of symbiotic root nodule development**.** ***Annu Rev Genet*** **1998,** 32**:33-57.**

5. **Tian CF, Young JP, Wang ET, Tamimi SM, Chen WX:** Population mixing of *Rhizobium leguminosarum bv. viciae* nodulating Vicia faba: the role of recombination and lateral gene transfer**.** ***FEMS Microbiol Ecol*** **2010,** 73**(3):563-576.**

6. **Bala A, Giller KE:** Symbiotic specificity of tropical tree rhizobia for host legumes**.** ***New Phytol*** **2001,** 149**:212-217.**

7. **Lindstrom K, Murwira M, Willems A, Altier N:** The biodiversity of beneficial microbe-host mutualism: the case of rhizobia**.** ***Res Microbiol*** **2010,** 161**(6):453-463.**

8. **Janczarek M, Kutkowska J, Piersiak T, Skorupska A:** *Rhizobium leguminosarum bv. trifolii* rosR is required for interaction with clover, biofilm formation and adaptation to the environment**.** ***BMC Microbiol*** **2010,** 10**(1):284.**

9. **Franche C, Lindström K, Elmerich C:** Nitrogen-fixing bacteria associated with leguminous and non-leguminous plants**.** ***Plant Soil*** **2009,** 321**:35-59.**

10. Ding H, Hynes MF: Plasmid transfer systems in the rhizobia**.** ***Can J Microbiol*** **2009,** 55**(8):917-927.**

11. Brom S, García-de los Santos A, Cervantes L, Palacios R, Romero D: In *Rhizobium etli* symbiotic plasmid transfer, nodulation competitivity and cellular growth require interaction among different replicons**.** ***Plasmid*** **2000,** 44**(1):34-43.**

12. Tikhonovich IA, Provorov NA: From plant-microbe interactions to symbiogenetics: a universal paradigm for the interspecies genetic integration**.** ***Ann Appl Biol*** **2009,** 154**:341-350.**

13. Kiers ET, Rousseau RA, West SA, Denison RF: Host sanctions and the legume-*rhizobium* mutualism**.** ***Nature*** **2003,** 425**(6953):78-81.**

14. Denison RF: Legume Sanctions and the Evolution of Symbiotic Cooperation by Rhizobia**.** ***Am Nat*** **2000,** 156**(6):567-576.**

15. Provorov NA, Vorobyov NI: Simulation of evolution implemented in the mutualistic symbioses towards enhancing their ecological efficiency, functional integrity and genotypic specificity**.** ***Theor Popul Biol*** **2010,** 78**(4):259-269.**

16. Broughton WJ: Roses by other names: taxonomy of the *Rhizobiaceae***.** ***J Bacteriol*** **2003,** 185**(10):2975-2979.**

17. Janda JM, Abbott SL: Bacterial identification for publication: when is enough enough? ***J Clin Microbiol*** **2002,** 40**(6):1887-1891.**

18. Olsen GJ, Woese CR, Overbeek R: The winds of (evolutionary) change: breathing new life into microbiology**.** ***J Bacteriol*** **1994,** 176**(1):1-6.**

19. Euzeby JP: List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet**.** ***Int J Syst Bacteriol*** **1997,** 47**(2):590-592.**

20. Rainey FA, Ward-Rainey NL, Janssen PH, Hippe H, Stackebrandt E: *Clostridium paradoxum DSM 7308T* contains multiple 16S rRNA genes with heterogeneous intervening sequences**.** ***Microbiology*** **1996,** 142 (Pt 8)**:2087-2095.**

21. Amann G, Stetter KO, Llobet-Brossa E, Amann R, Anton J: Direct proof for the presence and expression of two 5% different 16S rRNA genes in individual cells of *Haloarcula marismortui***.** ***Extremophiles*** **2000,** 4**(6):373-376.**

22. Young JM, Kuykendall LD, Martínez-Romero E, Kerr A, Sawada H: A revision of *Rhizobium Frank 1889,* with an emended description of the genus, and the inclusion of all species of *Agrobacterium* Conn 1942 and Allorhizobium undicola de Lajudie et al. 1998 as new combinations: *Rhizobium radiobacter, R. rhizogenes, R. rubi, R. undicola* and *R. vitis***.** ***Int J Syst Evol Microbiol*** **2001,** 51**(Pt 1):89-103.**

23. van Berkum P, Terefework Z, Paulin L, Suomalainen S, Lindstrom K, Eardly BD: Discordant phylogenies within the rrn loci of Rhizobia**.** ***J Bacteriol*** **2003,** 185**(10):2988-2998.**

24. Stackebrandt E, Breymann S, Steiner U, Prauser H, Weiss N, Schumann P: Re-evaluation of the status of the genus *Oerskovia*, reclassification of Promicromonospora enterophila (Jager et al. 1983) as *Oerskovia enterophila* comb. nov. and description of *Oerskovia jenensis sp.* nov. and *Oerskovia paurometabola sp.* nov**.** ***Int J Syst Evol Microbiol*** **2002,** 52**(Pt 4):1105-1111.**

25. Zeigler DR: Gene sequences useful for predicting relatedness of whole genomes in bacteria**.** ***Int J Syst Evol Microbiol*** **2003,** 53**(Pt 6):1893-1900.**

26. Sahgal M, Johri BN: Taxonomy of rhizobia: Current status**.** ***Current Science*** **2006,** 90**(4):486-487.**

27. Vinuesa P, Silva C, Werner D, Martinez-Romero E: Population genetics and phylogenetic inference in bacterial molecular systematics: the roles of migration and recombination in Bradyrhizobium species cohesion and delineation**.** ***Mol Phylogenet Evol*** **2005,** 34**(1):29-54.**

28. Roy SW: Phylogenomics: gene duplication, unrecognized paralogy and outgroup choice**.** ***PLoS One*** **2009,** 4**(2):e4568.**

29. Kistner C, Parniske M: Evolution of signal transduction in intracellular symbiosis**.** ***Trends Plant Sci*** **2002,** 7**(11):511-518.**

30. Bonfante P, Genre A: Mechanisms underlying beneficial plant-fungus interactions in mycorrhizal symbiosis**.** ***Nat Commun*** **2010,** 1**(4):1-11.**

31. Kawaguchi M, Minamisawa K: Plant-microbe communications for symbiosis**.** ***Plant Cell Physiol*** **2010,** 51**(9):1377-1380.**

32. Parniske M: Arbuscular mycorrhiza: the mother of plant root endosymbioses**.** ***Nat Rev Microbiol*** **2008,** 6**(10):763-775.**

33. Martínez-Romero E: Coevolution in *Rhizobium*-legume symbiosis? ***DNA Cell Biol*** **2009,** 28**(8):361-370.**

34. Becerra Velásquez VL, Gepts P: RFLP diversity of common bean (*Phaseolus vulgaris*) in its centres of origin**.** ***Genome*** **1994,** 37**:256-263.**

35. Beebe S, Rengifo J, Gaitán E, Tohme J: Diversity and Origin of Andean Landraces of Common Bean**.** ***Crop Sci*** **2001,** 41**:854-862.**

36. Singh SP, Gepts P, Debouck DG: Races of common bean (*Phaseolus vulgaris* L., Fabaceae)**.** ***Econ Bot*** **1991,** 45**:379-396.**

37. Tohme J, González DO, Beebe S, Duque MC: AFLP Analysis of Gene Pools of a Wild Bean Core Collection**.** ***Crop Sci*** **1996,** 36**(5):1375-1384.**

38. Kami J, Velasquez VB, Debouck DG, Gepts P: Identification of presumed ancestral DNA sequences of phaseolin in *Phaseolus vulgaris*. ***Proc Natl Acad Sci U S A*** **1995,** 92**(4):1101-1104.**

39. Geffroy V, Sicard D, de Oliveira JC, Sevignac M, Cohen S, Gepts P, Neema C, Langin T, Dron M: Identification of an ancestral resistance gene cluster involved in the coevolution process between *Phaseolus vulgaris* and its fungal pathogen *Colletotrichum lindemuthianum*. ***Mol Plant Microbe Interact*** **1999,** 12**(9):774-784.**

40. Aguilar OM, Riva O, Peltzer E: Analysis of Rhizobium etli and of its symbiosis with wild Phaseolus vulgaris supports coevolution in centers of host diversification. ***Proc Natl Acad Sci U S A*** **2004,** 101**(37):13548-13553.**

41. Shamseldin A, Vinuesa P, Thierfelder H, Werner D: *Rhizobium etli* and *Rhizobium gallicum* Nodulate *Phaseolus vulgaris* in Egyptian Soils and Display Cultivar-Dependent Symbiotic Efficiency. ***Symbiosis*** **2005,** 38**:145-161.**

42. Martínez-Romero E, Segovia L, Mercante FM, Franco AA, Graham P, Pardo MA: *Rhizobium tropici*, a novel species nodulating *Phaseolus vulgaris* L. beans and *Leucaena sp*. trees. ***Int J Syst Bacteriol*** **1991,** 41**(3):417-426.**

43. Amarger N, Macheret V, Laguerre G: *Rhizobium gallicum* sp. nov. and *Rhizobium giardinii* sp. nov., from *Phaseolus vulgaris* nodules. ***Int J Syst Bacteriol*** **1997,** 47**(4):996-1006.**

44. Futuyma DJ: Evolution**, 2nd edition edn: Sinauer Associates, Sunderland, MA; 2005.**

45. Smith CI, Godsoe WK, Tank S, Yoder JB, Pellmyr O: Distinguishing coevolution from covicariance in an obligate pollination mutualism: *asynchronous* divergence in Joshua tree and its pollinators. ***Evolution*** **2008,** 62**(10):2676-2687.**

46. Crespi BJ: The evolution of social behavior in microorganisms**.** ***Trends Ecol Evol*** **2001,** 16**(4):178-183.**

47. Sullivan JT, Patrick HN, Lowther WL, Scott DB, Ronson CW: Nodulating strains of *Rhizobium loti* arise through chromosomal symbiotic gene transfer in the environment**.** ***Proc Natl Acad Sci U S A*** **1995,** 92**(19):8985-8989.**

48. Nandasena KG, O'Hara GW, Tiwari RP, Sezmis E, Howieson JG: In situ lateral transfer of symbiosis islands results in rapid evolution of diverse competitive strains of mesorhizobia suboptimal in symbiotic nitrogen fixation on the pasture legume *Biserrula pelecinus* L**.** ***Environ Microbiol*** **2007,** 9**(10):2496-2511.**

49. Provorov NA: Coevolution of rhizobia with legumes: facts and hypothesis**.** ***Symbiosis*** **1998,** 24**:337-368.**

50. Didelot X, Maiden MC: Impact of recombination on bacterial evolution**.** ***Trends Microbiol*** **2010,** 18**(7):315-322.**

51. Achtman M: Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens**.** ***Annu Rev Microbiol*** **2008,** 62**:53-70.**

52. Ochman H, Lawrence JG, Groisman EA: Lateral gene transfer and the nature of bacterial innovation**.** ***Nature*** **2000,** 405**(6784):299-304.**

53. Thomas CM, Nielsen KM: Mechanisms of, and barriers to, horizontal gene transfer between bacteria**.** ***Nat Rev Microbiol*** **2005,** 3**(9):711-721.**

54. Bergstrom CT, Lipsitch M, Levin BR: Natural selection, infectious transfer and the existence conditions for bacterial plasmids**.** ***Genetics*** **2000,** 155**(4):1505-1519.**

55. Campbell AM: Lateral gene transfer in prokaryotes**.** ***Theor Popul Biol*** **2000,** 57**(2):71-77.**

56. Gal-Mor O, Finlay BB: Pathogenicity islands: a molecular toolbox for bacterial virulence**.** ***Cell Microbiol*** **2006,** 8**(11):1707-1719.**

57. Posada D, Crandall KA, Holmes EC: Recombination in evolutionary genomics**.** ***Annu Rev Genet*** **2002,** 36**:75-97.**

58. Levin BR, Cornejo OE: The population and evolutionary dynamics of homologous gene recombination in bacterial populations**.** ***PLoS Genet*** **2009,** 5**(8):e1000601.**

59. Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE *et al*: Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences**.** ***Proc Natl Acad Sci U S A*** **2001,** 98**(1):182-187.**

60. de Vries J, Wackernagel W: Integration of foreign DNA during natural transformation of Acinetobacter sp. by homology-facilitated illegitimate recombination**.** ***Proc Natl Acad Sci U S A*** **2002,** 99**(4):2094-2099.**

61. Prudhomme M, Libante V, Claverys JP: Homologous recombination at the border: insertion-deletions and the trapping of foreign DNA in Streptococcus pneumoniae**.** ***Proc Natl Acad Sci U S A*** **2002,** 99**(4):2100-2105.**

62. Spratt BG, Hanage WP, Feil EJ: The relative contributions of recombination and point mutation to the diversification of bacterial clones**.** ***Curr Opin Microbiol*** **2001,** 4**(5):602-606.**

63. Touzain F, Denamur E, Medigue C, Barbe V, El Karoui M, Petit MA: Small variable segments constitute a major type of diversity of bacterial genomes at the species level**.** ***Genome Biol*** **2010,** 11**(4):R45.**

64. Felsenstein J: The evolutionary advantage of recombination**.** ***Genetics*** **1974,** 78**(2):737-756.**

65. Felsenstein J, Yokoyama S: The evolutionary advantage of recombination. II. Individual selection for recombination**.** ***Genetics*** **1976,** 83**(4):845-859.**

66. Fraser C, Hanage WP, Spratt BG: Recombination and the nature of bacterial speciation**.** ***Science*** **2007,** 315**(5811):476-480.**

67. Jiggins FM: The rate of recombination in Wolbachia bacteria**.** ***Mol Biol Evol*** **2002,** 19**(9):1640-1643.**

68. Jolley KA, Wilson DJ, Kriz P, McVean G, Maiden MC: The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in Neisseria meningitidis**.** ***Mol Biol Evol*** **2005,** 22**(3):562-569.**

69. Souza V, Nguyen TT, Hudson RR, Pinero D, Lenski RE: Hierarchical analysis of linkage disequilibrium in Rhizobium populations: evidence for sex? ***Proc Natl Acad Sci U S A*** **1992,** 89**(17):8389-8393.**

70. Abby S, Daubin V: Comparative genomics and the evolution of prokaryotes**.** ***Trends Microbiol*** **2007,** 15**(3):135-141.**

71. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O *et al*: Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths**.** ***PLoS Genet*** **2009,** 5**(1):e1000344.**

72. Konstantinidis KT, Ramette A, Tiedje JM: The bacterial species definition in the genomic era**.** ***Philos Trans R Soc Lond B Biol Sci*** **2006,** 361**(1475):1929-1940.**

73. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T *et al*: Complete genome sequence of enterohemorrhagic *Escherichia coli O157:H7* and genomic comparison with a laboratory strain K-12**.** ***DNA Res*** **2001,** 8**(1):11-22.**

74. Lapierre P, Gogarten JP: Estimating the size of the bacterial pan-genome**.** ***Trends Genet*** **2009,** 25**(3):107-110.**

75. Lefebure T, Stanhope MJ: Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition**.** ***Genome Biol*** **2007,** 8**(5):R71.**

76. Bentley S: Sequencing the species pan-genome**.** ***Nat Rev Microbiol*** **2009,** 7**(4):258-259.**

77. Pinero D, Martinez E, Selander RK: Genetic diversity and relationships among isolates of Rhizobium leguminosarum biovar phaseoli**.** ***Appl Environ Microbiol*** **1988,** 54**(11):2825-2832.**

78. Segovia L, Young JP, Martínez-Romero E: Reclassification of American *Rhizobium leguminosarum biovar phaseoli* type I strains as *Rhizobium etli* sp. nov**.** ***Int J Syst Bacteriol*** **1993,** 43**(2):374-377.**

79. Martínez-Romero E: Diversity of *Rhizobium-Phaseolus vulgaris* symbiosis: overview and perspectives**.** ***Plant Soil*** **2003,** 252**:11-23.**

80. Martínez-Romero E, Rosenblueth M: Increased Bean (*Phaseolus vulgaris L.*) Nodulation Competitiveness of Genetically Modified *Rhizobium* Strains**.** ***Appl Environ Microbiol*** **1990,** 56**(8):2384-2388.**

81. Haukka K, Lindstrom K, Young JP: Three phylogenetic groups of nodA and nifH genes in *Sinorhizobium* and *Mesorhizobium* isolates from leguminous trees growing in Africa and Latin America**.** ***Appl Environ Microbiol*** **1998,** 64**(2):419-426.**

82. Toledo I, Lloret L, Martínez-Romero E: *Sinorhizobium americanus* sp. nov., a new *Sinorhizobium* species nodulating native Acacia spp. in Mexico**.** ***Syst Appl Microbiol*** **2003,** 26**(1):54-64.**

83. Lloret L, Ormeno-Orrillo E, Rincón R, Martínez-Romero J, Rogel-Hernández MA, Martínez-Romero E: *Ensifer mexicanus sp*. nov. a new species nodulating *Acacia angustissima* (Mill.) Kuntze in Mexico**.** ***Syst Appl Microbiol*** **2007,** 30**(4):280-290.**

84. Rincón-Rosales R, Lloret L, Ponce E, Martínez-Romero E: Rhizobia with different symbiotic efficiencies nodulate *Acaciella angustissima* in Mexico, including *Sinorhizobium chiapanecum sp*. nov. which has common symbiotic genes with *Sinorhizobium mexicanum***.** ***FEMS Microbiol Ecol*** **2009,** 68**(2):255.**

85. Gonzalez V, Bustos P, Ramirez-Romero MA, Medrano-Soto A, Salgado H, Hernandez-Gonzalez I, Hernandez-Celis JC, Quintero V, Moreno-Hagelsieb G, Girard L *et al*: The mosaic structure of the symbiotic plasmid of Rhizobium etli CFN42 and its relation to other symbiotic genome compartments**.** ***Genome Biol*** **2003,** 4**(6):R36.**

86. Flores M, Morales L, Avila A, Gonzalez V, Bustos P, Garcia D, Mora Y, Guo X, Collado-Vides J, Pinero D *et al*: Diversification of DNA sequences in the symbiotic genome of Rhizobium etli**.** ***J Bacteriol*** **2005,** 187**(21):7185-7192.**

87. Bailly X, Giuntini E, Sexton MC, Lower RP, Harrison PW, Kumar N, Young JP: Population genomics of Sinorhizobium medicae based on low-coverage sequencing of sympatric isolates**.** ***ISME J*** **2011.**

88. Landeta C, Dávalos A, Cevallos MA, Geiger O, Brom S, Romero D: Plasmids with a chromosome-like role in rhizobia**.** ***J Bacteriol*** **2011,** 193**(6):1317-1326.**

89. Gonzalez V, Santamaria RI, Bustos P, Hernandez-Gonzalez I, Medrano-Soto A, Moreno-Hagelsieb G, Janga SC, Ramirez MA, Jimenez-Jacinto V, Collado-Vides J *et al*: The partitioned Rhizobium etli genome: genetic and

metabolic redundancy in seven interacting replicons**.** ***Proc Natl Acad Sci U S A*** **2006,** 103**(10):3834-3839.**

90. González V, Bustos P, Ramírez-Romero MA, Medrano-Soto A, Salgado H, Hernández-González I, Hernández-Celis JC, Quintero V, Moreno-Hagelsieb G, Girard L *et al*: The mosaic structure of the symbiotic plasmid of *Rhizobium etli CFN42* and its relation to other symbiotic genome compartments**.** ***Genome Biol*** **2003,** 4**(6):R36.**

91. González V, Santamaría RI, Bustos P, Hernández-González I, Medrano-Soto A, Moreno-Hagelsieb G, Janga SC, Ramírez MA, Jiménez-Jacinto V, Collado-Vides J *et al*: The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons**.** ***Proc Natl Acad Sci U S A*** **2006,** 103**(10):3834-3839.**

92. Crossman LC, Castillo-Ramirez S, McAnnula C, Lozano L, Vernikos GS, Acosta JL, Ghazoui ZF, Hernandez-Gonzalez I, Meakin G, Walker AW *et al*: A common genomic framework for a diverse assembly of plasmids in the symbiotic nitrogen fixing bacteria**.** ***PLoS One*** **2008,** 3**(7):e2567.**

93. Gonzalez V, Acosta JL, Santamaria RI, Bustos P, Fernandez JL, Hernandez Gonzalez IL, Diaz R, Flores M, Palacios R, Mora J *et al*: Conserved symbiotic plasmid DNA sequences in the multireplicon pangenomic structure of Rhizobium etli**.** ***Appl Environ Microbiol*** **2010,** 76**(5):1604-1614.**

94. Wang J, Huang X: A method for finding single-nucleotide polymorphisms with allele frequencies in sequences of deep coverage**.** ***BMC Bioinformatics*** **2005,** 6**:220.**

95. Gordon D, Desmarais C, Green P: Automated finishing with autofinish**.** ***Genome Res*** **2001,** 11**(4):614-625.**

96. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, Gish WR: A general approach to single-nucleotide polymorphism discovery**.** ***Nat Genet*** **1999,** 23**(4):452-456.**

97. Ewing B, Green P: Base-calling of automated sequencer traces using phred. II. Error probabilities**.** ***Genome Res*** **1998,** 8**(3):186-194.**

98. Lawrence CB, Solovyev VV: Assignment of position-specific error probability to primary DNA sequence data**.** ***Nucleic Acids Res*** **1994,** 22**(7):1272-1280.**

99. Darling AC, Mau B, Blattner FR, Perna NT: Mauve: multiple alignment of conserved genomic sequence with rearrangements**.** ***Genome Res*** **2004,** 14**(7):1394-1403.**

100. Crossman LC, Castillo-Ramírez S, McAnnula C, Lozano L, Vernikos GS, Acosta JL, Ghazoui ZF, Hernández-González I, Meakin G, Walker AW *et al*: A common genomic framework for a diverse assembly of plasmids in the symbiotic nitrogen fixing bacteria**.** ***PLoS One*** **2008,** 3**(7):e2567.**

101. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al*: The genome sequence of *Drosophila melanogaster***.** ***Science*** **2000,** 287**(5461):2185-2195.**

102. Dean CB, Nielsen JD: Generalized linear mixed models: a review and some extensions**.** ***Lifetime Data Anal*** **2007,** 13**(4):497-512.**

103. Posada D, Crandall KA: The effect of recombination on the accuracy of phylogeny estimation**.** ***J Mol Evol*** **2002,** 54**(3):396-402.**

104. Edgar RC: MUSCLE: a multiple sequence alignment method with reduced time and space complexity**.** ***BMC Bioinformatics*** **2004,** 5**:113.**

105. Talavera G, Castresana J: Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments**.** ***Syst Biol*** **2007,** 56**(4):564-577.**

106. Posada D: Evaluation of methods for detecting recombination from DNA sequences: empirical data**.** ***Mol Biol Evol*** **2002,** 19**(5):708-717.**

107. Posada D, Crandall KA, Holmes EC: Recombination in evolutionary genomics**.** ***Annu Rev Genet*** **2002,** 36**:75-97.**

108. Sawyer S: Statistical tests for detecting gene conversion**.** ***Mol Biol Evol*** **1989,** 6**(5):526-538.**

109. Worobey M: A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria**.** ***Mol Biol Evol*** **2001,** 18**(8):1425-1434.**

110. Guindon S, Gascuel O: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood**.** ***Syst Biol*** **2003,** 52**(5):696-704.**

111. Yang Z: PAML 4: phylogenetic analysis by maximum likelihood**.** ***Mol Biol Evol*** **2007,** 24**(8):1586-1591.**

112. Bruen TC, Philippe H, Bryant D: A simple and robust statistical test for detecting the presence of recombination**.** ***Genetics*** **2006,** 172**(4):2665-2681.**

113. Huson DH, Bryant D: Application of phylogenetic networks in evolutionary studies**.** ***Mol Biol Evol*** **2006,** 23**(2):254-267.**

114. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD: GARD: a genetic algorithm for recombination detection**.** ***Bioinformatics*** **2006,** 22**(24):3096-3098.**

115. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns**.** ***Proc Natl Acad Sci U S A*** **1998,** 95**(25):14863-14868.**

116. Posada D, Crandall KA: Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1)**.** ***Mol Biol Evol*** **2001,** 18**(6):897-906.**

117. Posada D, Crandall KA: Selecting the best-fit model of nucleotide substitution**.** ***Syst Biol*** **2001,** 50**(4):580-601.**

118. Posada D, Crandall KA: MODELTEST: testing the model of DNA substitution**.** ***Bioinformatics*** **1998,** 14**(9):817-818.**

119. Felsenstein J: PHYLIP (Phylogeny inference Package) version 3.6**.** ***Department of Genome Sciences University of Washington, Seattle*** **2005.**

120. Stamatakis A, Ludwig T, Meier H: RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees**.** ***Bioinformatics*** **2005,** 21**(4):456-463.**

121. Ronquist F, Huelsenbeck JP: MrBayes 3: Bayesian phylogenetic inference under mixed models**.** ***Bioinformatics*** **2003,** 19**(12):1572-1574.**

122. Shimodaira H, Hasegawa M: CONSEL: for assessing the confidence of phylogenetic tree selection**.** ***Bioinformatics*** **2001,** 17**(12):1246-1247.**

123. Strimmer K, Rambaut A: Inferring confidence sets of possibly misspecified gene trees**.** ***Proc Biol Sci*** **2002,** 269**(1487):137-142.**

124. Thornton K: Libsequence: a C++ class library for evolutionary genetic analysis**.** ***Bioinformatics*** **2003,** 19**(17):2325-2327.**

125. Tatusov RL, Galperin MY, Natale DA, Koonin EV: The COG database: a tool for genome-scale analysis of protein functions and evolution**.** ***Nucleic Acids Res*** **2000,** 28**(1):33-36.**

126. Che D, Li G, Mao F, Wu H, Xu Y: Detecting uber-operons in prokaryotic genomes**.** ***Nucleic Acids Res*** **2006,** 34**(8):2418-2427.**

127. Pertea M, Ayanbule K, Smedinghoff M, Salzberg SL: OperonDB: a comprehensive database of predicted operons in microbial genomes**.** ***Nucleic Acids Res*** **2009,** 37**(Database issue):D479-D482.**

128. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**.** ***Nucleic Acids Res*** **1997,** 25**(17):3389-3402.**

129. Enright AJ, Van Dongen S, Ouzounis CA: An efficient algorithm for large-scale detection of protein families**.** ***Nucleic Acids Res*** **2002,** 30**(7):1575-1584.**

130. González V, Acosta JL, Santamaría RI, Bustos P, Fernández JL, Hernández González IL, Díaz R, Flores M, Palacios R, Mora J *et al*: Conserved symbiotic plasmid DNA sequences in the multireplicon pangenomic structure of *Rhizobium etli***.** ***Appl Environ Microbiol*** **2010,** 76**(5):1604-1614.**

131. Gordon D, Desmarais C, Green P: Automated finishing with autofinish**.** ***Genome Res*** **2001,** 11**(4):614-625.**

132. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, Gish WR: A general approach to single-nucleotide polymorphism discovery**.** ***Nat Genet*** **1999,** 23**(4):452-456.**

133. Brockman W, Álvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB: Quality scores and SNP detection in sequencing-by-synthesis systems**.** ***Genome Res*** **2008,** 18**(5):763-770.**

134. Hubisz MJ, Lin MF, Kellis M, Siepel A: Error and error mitigation in low-coverage genome assemblies**.** ***PLoS One*** **2011,** 6**(2):e17034.**

135. Flores M, Morales L, Ávila A, González V, Bustos P, García D, Mora Y, Guo X, Collado-Vides J, Pinero D *et al*: Diversification of DNA sequences in the symbiotic genome of *Rhizobium etli***.** ***J Bacteriol*** **2005,** 187**(21):7185-7192.**

136. Silva C, Vinuesa P, Eguiarte LE, Souza V, Martínez-Romero E: Evolutionary genetics and biogeographic structure of *Rhizobium gallicum sensu lato*, a widely distributed bacterial symbiont of diverse legumes**.** ***Mol Ecol*** **2005,** 14**(13):4033-4050.**

137. Didelot X, Lawson D, Darling A, Falush D: Inference of homologous recombination in bacteria using whole-genome sequences**.** ***Genetics*** **2010,** 186**(4):1435-1449.**

138. Spratt BG, Hanage WP, Feil EJ: The relative contributions of recombination and point mutation to the diversification of bacterial clones**.** ***Curr Opin Microbiol*** **2001,** 4**(5):602-606.**

139. Keller I, Bensasson D, Nichols RA: Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes**.** ***PLoS Genet*** **2007,** 3**(2):e22.**

140. Yang Z, Bielawski JP: Statistical methods for detecting molecular adaptation**.** ***Trends Ecol Evol*** **2000,** 15**(12):496-503.**

141. Kuo CH, Moran NA, Ochman H: The consequences of genetic drift for bacterial genome complexity**.** ***Genome Res*** **2009,** 19**(8):1450-1454.**

142. Richter M, Rossello-Mora R: Shifting the genomic gold standard for the prokaryotic species definition**.** ***Proc Natl Acad Sci U S A*** **2009,** 106**(45):19126-19131.**

143. Young JP, Crossman LC, Johnston AW, Thomson NR, Ghazoui ZF, Hull KH, Wexler M, Curson AR, Todd JD, Poole PS *et al*: The genome of *Rhizobium leguminosarum* has recognizable core and accessory components**.** ***Genome Biol*** **2006,** 7**(4):R34.**

144. Segovia L, Piñero D, Palacios R, Martínez-Romero E: Genetic structure of a soil population of nonsymbiotic *Rhizobium leguminosarum*. ***Appl Environ Microbiol*** **1991,** 57**(2):426-433.**

145. Rogel MA, Ormeno-Orrillo E, Martínez Romero E: Symbiovars in rhizobia reflect bacterial adaptation to legumes**.** ***Syst Appl Microbiol*** **2011,** 34**(2):96-104.**

146. Tun-Garrido C, Bustos P, González V, Brom S: Conjugative transfer of p42a from *Rhizobium etli CFN42*, which is required for mobilization of the symbiotic plasmid, is regulated by quorum sensing**.** ***J Bacteriol*** **2003,** 185**(5):1681-1692.**

147. Pérez-Mendoza D, Domínguez-Ferreras A, Munoz S, Soto MJ, Olivares J, Brom S, Girard L, Herrera-Cervera JA, Sanjuan J: Identification of functional mob regions in *Rhizobium etli*: evidence for self-transmissibility of the symbiotic plasmid pRetCFN42d**.** ***J Bacteriol*** **2004,** 186**(17):5753-5761.**

148. Lozano L, Hernández-González I, Bustos P, Santamaría RI, Souza V, Young JP, Dávila G, González V: Evolutionary dynamics of insertion sequences in relation to the evolutionary histories of the chromosome and symbiotic plasmid genes of *Rhizobium etli* populations**.** ***Appl Environ Microbiol*** **2010,** 76**(19):6504-6513.**

149. Souza V, Nguyen TT, Hudson RR, Piñero D, Lenski RE: Hierarchical analysis of linkage disequilibrium in *Rhizobium* populations: evidence for sex? ***Proc Natl Acad Sci U S A*** **1992,** 89**(17):8389-8393.**

150. Piñero D, Martínez E, Selander RK: Genetic diversity and relationships among isolates of *Rhizobium leguminosarum biovar phaseoli***. *Appl Environ Microbiol* 1988,** 54**(11):2825-2832.**

151. Silva C, LE E, Souza V: Reticulated and epidemic population genetic structure of *Rhizobium etli* biovar *phaseoli* in a traditionally managed locality in Mexico**. *Molecular Ecology* 1999,** 8**:277-287.**

152. Silva C, Vinuesa P, Eguiarte LE, Martínez-Romero E, Souza V: *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* 2003,** 69**(2):884-893.**

153. Eardly BD, Wang FS, Whittam TS, Selander RK: Species limits in *Rhizobium* populations that nodulate the common bean (*Phaseolus vulgaris*)**. *Appl Environ Microbiol* 1995,** 61**(2):507-512.**

154. Vinuesa P, Silva C, Werner D, Martínez-Romero E: Population genetics and phylogenetic inference in bacterial molecular systematics: the roles of migration and recombination in *Bradyrhizobium* species cohesion and delineation**. *Mol Phylogenet Evol* 2005,** 34**(1):29-54.**

155. Young JM, Park DC, Weir BS: Diversity of 16S rDNA sequences of *Rhizobium spp.* implications for species determinations**. *FEMS Microbiol Lett* 2004,** 238**(1):125-131.**

156. Touzain F, Denamur E, Medigue C, Barbe V, El Karoui M, Petit MA: Small variable segments constitute a major type of diversity of bacterial genomes at the species level**. *Genome Biol* 2010,** 11**(4):R45.**

157. Lefebure T, Stanhope MJ: Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition**. *Genome Biol* 2007,** 8**(5):R71.**

158. Wu J, Yu T, Bao Q, Zhao F: Evidence of extensive homologous recombination in the core genome of *Rickettsia*. ***Comp Funct Genomics*** **2009:510270.**

159. Abu-Ali GS, Lacher DW, Wick LM, Qi W, Whittam TS: Genomic diversity of pathogenic *Escherichia coli* of the EHEC 2 clonal complex. ***BMC Genomics*** **2009,** 10**:296.**

160. Alcaraz LD, Moreno-Hagelsieb G, Eguiarte LE, Souza V, Herrera-Estrella L, Olmedo G: Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics. ***BMC Genomics*** **2010,** 11**:332.**

161. Retchless AC, Lawrence JG: Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. ***Proc Natl Acad Sci U S A*** **2010,** 107**(25):11453-11458.**

162. Bailly X, Giuntini E, Sexton MC, Lower RP, Harrison PW, Kumar N, Young JP: Population genomics of *Sinorhizobium medicae* based on low-coverage sequencing of sympatric isolates. ***ISME J*** **2011.**

GLOSARIO

El orden de las definiciones está relacionado con su aparición en la tesis.

Fijación de Nitrogeno.- Es el proceso por el cual el nitrógeno se toma de su forma molecular natural, relativamente inerte (N_2) en la atmósfera y convertido en compuestos del nitrógeno (por ejemplo amoníaco, nitrato y dióxido del nitrógeno). La fijación de nitrógeno es realizada naturalmente por un número de diferentes procariotes, incluyendo bacterias, actinobacteria, y ciertos tipos de anaerobio bacterias.

Islas simbióticas.- Agrupamientos de genes que son requeridos para la interacción simbiótica con un huésped eucarionte, de origen ilegitimo (transferidos lateralmente) y se encuentran flanqueados por elementos de inserción, como IS (secuencias de inserción) o transposones. Generalmente se encuentran dentro de plásmidos y en menor grado en cromosomas.

Marcadores taxonómicos.- Están basados en segmentos de DNA, ya sea regiones codificantes (genes) o no codificantes y proporcionan información para permitir una rápida clasificación taxonómica preliminar y ayudar en casos donde existen dudas o los datos morfológicos no son claros

Hibridaciones DNA-DNA.- Técnica para medir el grado de similitud genética entre dos genomas, mediante la medición de calor requerido para romper los enlaces de hidrógeno entre pares de bases que forman uniones entre las dos hebras de la doble hélice del DNA dúplex.

Incongruencia filogenética.- Es un camino para valorar la disminución de la compatibilidad entre un árbol de referencia (árbol de una especie) y un conjunto de arboles a través del conteo de biparticiones o cuartetos topológicamente distintos al árbol individual muestreado.

Filogenómica.- Utiliza datos genómicos (genomas) para estudiar el árbol de la vida o para algún taxón en particular. Se basa en 4 propuestas: i) a partir de un lote de arboles de genes individuales, ii) concatenando un grupo de genes para construir un súper árbol, iii) construir un único árbol basado en el contenido de genes dentro de genomas ya sean ortólogos u homólogos e iv) investigar el orden de genes dentro de un grupo de regiones compartidas entre todos los genomas muestreados.

Evolución adaptativa.- Es aquel cambio provocado por la selección natural en el fenotipo de los organismos dirigido a incrementar el fitness en un ambiente determinado.

Electroforesis de Enzimas Multiloci (*MLEE*).- Es un método para caracterizar organismos a través de las movilidades relativas de un largo numero de enzimas intracelulares bajo electroforesis. Estas diferencias en la movilidad están relacionadas a mutaciones en un locus (gene) que causa una sustitución de un amino acido en la enzima codificada por el gen. Diferencias en las cargas electroestáticas entre el amino acido original y el amino acido sustituido pueden afectar la carga neta de la enzima y por lo tanto su movilidad electroforética. Así,

es posible relacionar diferencias en la movilidad con diferentes alelos para un locus en particular. Estas variantes en movilidad se denominan electroferotipos.

Polimofismos en longitud de los fragmentos amplificados (*AFLP*).- Son un tipo de marcador molecular que está basado en la restricción del DNA genómico mediante enzimas de restricción y en la subsecuente amplificación de algunos de esos fragmentos mediante la reacción en cadena de la polimerasa. Son una herramienta poderosa de análisis del genoma dado que poseen un alto poder de detección de la variabilidad genética. El ensayo de AFLP combina la especificidad, resolución y poder de muestreo de la digestión con enzimas de restricción con la velocidad y practicidad de la detección de polimorfismos mediante PCR, sin necesidad de disponer de información previa del genoma a estudiar.

Análisis de secuencia de multiloci (*MLST*).- El objetivo del *MLST* es proporcionar un sistema de tipificación portátil, preciso y muy exigente que puede ser utilizado para la mayoría de las bacterias y algunos otros organismos. El *MLST* se basa en los conceptos probados de la electroforesis de enzimas multilocus (*MLEE*) y se han adaptado para que los alelos en cada locus se definan directamente, mediante la secuenciación de nucleótidos, en lugar de indirectamente de la movilidad electroforética de los productos de sus genes.

Desequilibrio de ligamiento.- Ocurre cuando dos alelos se presentan juntos más a menudo de lo que puede explicarse por el azar. También indica que los dos alelos están físicamente cerca sobre una misma cadena de DNA.

Identidad nucleotídica.- Denota el porcentaje de nucleótidos idénticos entre un par de secuencias alineadas. Este porcentaje disminuye por la presencia de mutaciones o inserciones dentro del alineamiento.

Contenido génico.- El número total de genes incluidos en un genoma.

Pan-genoma abierto.- El pangenoma se define como los genes totales de una especie y está formado por el genoma core (genes compartidos por todas las cepas), el genoma dispensable (genes presentes en dos o más cepas) y finalmente los genes únicos (pertenecientes a una cepa especifica). Se denomina un pangenoma abierto cuando el tamaño de este tiende al infinito, es decir, que conforme se va añadiendo un nuevo genoma a la especie se provee nuevos genes al pangenoma y su crecimiento nunca tiende a la asíntota.

Rearreglos genómicos.- Son cambios que afecta la estructura del genoma y se define como el reacomodo (cambio de lugar) de segmentos genómicos, ya sea dentro del cromosoma o en plásmidos. Su tamaño va desde grandes porciones a cientos o miles de pares de bases.

Inversiones genómicas.- Una inversión es cuando un segmento genómico cambia de orientación dentro de un cromosoma o plásmido. Para que se produzca este suceso es necesario una doble rotura y un doble giro de 180º del segmento formado por las roturas. En el caso de organismo diploides, hay dos tipos de inversiones según su relación con el centrómero: i) Pericéntricas: Incluyen al centrómero. Se detectan fácilmente al microscopio óptico pues implican un cambio

en la forma del cromosoma. ii) Paracéntricas: No incluyen al centrómero y por tanto tampoco afectan a la forma del cromosoma.

Plásmido con estructuras de mosaico.- Son plásmidos con distintas regiones adquiridas de múltiples genomas. Estas regiones pueden ir flanqueadas por elementos móviles como: secuencias de inserción, transposones y elementos de bacteriófagos.

Sintenia genómica.- Se define como la localización paralela entre dos segmentos genómicos de diferentes organismos. Esta sintenia puede contener desde operones hasta grandes porciones de cromosomas o plásmidos.

Genes homólogos.- Un gen relacionado a otro gene mediante descendencia de una secuencia de DNA ancestral. Este término también se puede aplicar a la relación entre genes separados por un evento de especiación.

Genes ortólogos.- Son genes de diferentes especies que evolucionaron de un gen ancestral común por especiación. Normalmente, los genes ortólogos retienen la misma función en el transcurso de la evolución. La identificación de los genes ortólogos es crucial para una confiable predicción de la función de genes en genomas recién secuenciados.

Genes parálogos.- Son genes generados por una duplicación dentro de un genoma. Generalmente estos genes evolucionan a nuevas funciones.

SNPs.- Es una variación en una secuencia de DNA que ocurre en un solo nucleótido (A,T,G,C), esta variación puede ocurrir dentro de un cromosoma, plasmido o segmentos de estos. Generalmente esta mutación puntual difiere entre miembros de una especie biológica o cromosomas pareados (organismos diploides). Dentro de una población, los SNPs pueden ser asignados con una frecuencia alélica mínima.

Factores sigma.- Es un factor de iniciación de la transcripción bacteriana que permite la unión específica de la RNA polimerasa a los promotores de los genes. Diferentes factores sigma se activan en respuesta a diferentes condiciones ambientales. Cada molécula de RNA polimerasa contiene exactamente una subunidad factor sigma.

Familias de genes.- Son grupos de genes homólogos que comparten funciones similares. En muchos casos, los genes en una familia comparten similar secuencia de DNA representada por bloques a lo largo de la secuencia. Las diferencias en el tamaño de las familias son debido a duplicaciones o pérdidas de genes en linajes-específicos.

Redundancia metabólica.- Es el resultado de una alta duplicación de genes que comparten una misma función. Por lo tanto, las familias de genes parálogos generan una redundancia en funciones que son necesarias para el fitness de la bacteria.

Cobertura 1X.- Es el número promedio de veces que un determinado nucleótido es representado en la secuencia, por ejemplo, una cobertura 1X significa que un nucleótido se ha secuenciado una sola vez, 2X denota que ha sido secuenciado dos veces y así sucesivamente.

Contigs.- Es una secuencia continua de DNA que fue ensamblada a partir del sobrelape de fragmentos de DNA clonados. La longitud de un contig es la longitud de la secuencia consenso.

ORFs.- Es una secuencia de DNA que no contiene un stop codón en un determinado marco de lectura. Existen 6 marcos de lectura, 3 en la cadena principal y 3 en la cadena secundaria.

Reciprocal best hits.- También conocido como el criterio boomerang, para determinar ortología entre secuencias. El *best hits* puede ser encontrado mediante el programa blast o por un algún otro reciente algoritmo a parir de una comparación entre dos genomas. Este criterio se cumple siempre y cuando el best hit del gen A en el genoma A´ sea el gen B del genoma B´ y viceversa. Sin embargo, en algunos casos el *Reciprocal best hits* no es suficiente para satisfacer la condición de ortología.

Gap.- Las secuencias homólogas pueden tener diferentes longitudes. Estas diferencias pueden ser explicadas por inserciones o deleciones. Así, una letra o un conjunto de letras pueden ser emparejadas con guiones en otras secuencias dentro de un alineamiento, dando como resultado final un bloque de secuencias

con la misma longitud. Estos *gaps* reflejan una inserción o deleción. Ya que una inserción en una secuencia da como resultado siempre una deleción en otra secuencia, a esto se le denomina indel.

***Score* de un alineamiento**.- Se obtiene de un código de substitución que es elegido para cada par de letras alineadas. El conjunto de estos códigos se denomina matrix de substitución. Al comparar dos secuencias, la igualdad en nucleótidos refleja un alto *score* (sumatoria) ponderado por el código de substitución, a mayor *score*, mayor identidad entre secuencias. Por lo tanto, el *score* de un alineamiento es la suma de los *scores* específicos para cada par de letras alineadas, los indeles también son tomados en cuenta en el *score*. Un mayor *score* denota un mejor alineamiento.

Frameshifts.- Es una mutación que inserta o borra un simple nucleótido en una secuencia de ADN. Debido a la naturaleza de la expresión genética, en forma de triplete, la inserción o borrado puede alterar la agrupación de codones, dando como resultado una traducción completamente diferente del original.

Valor de calidad *Qphred*.- Es una probabilidad de error log-transformada de que un nucleótido se ha secuenciado correctamente, un Q20 denota un error en cada 100 pares de bases, un Q30, un error en cada 1000 pares de bases y así sucesivamente.

Probabilidad de sitio polimórfico.- Probabilidad bayesiana de que un sitio dentro de un alineamiento sea una mutación y no sea producto de un error de secuencia.

Se obtiene del valor de calidad de las secuencias vecinas al sitio que presenta la mutación putativa.

Diversidad nucleotídica.- Es la distancia promedio entre dos secuencias de DNA elegidas al azar de una población. Se utiliza para medir el grado de polimorfismo en una población.

APENDICE

Artículos de coautoría

A Common Genomic Framework for a Diverse Assembly of Plasmids in the Symbiotic Nitrogen Fixing Bacteria

Lisa C. Crossman[1]*, Santiago Castillo-Ramírez[2], Craig McAnnula[3], Luis Lozano[2], Georgios S. Vernikos[1], José L. Acosta[2], Zara F. Ghazoui[4], Ismael Hernández-González[2], Georgina Meakin[5], Alan W. Walker[1], Michael F. Hynes[6], J. Peter W. Young[4], J. Allan Downie[3], David Romero[2], Andrew W. B. Johnston[5], Guillermo Dávila[2], Julian Parkhill[1], Víctor González[2]*

1 The Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom, **2** Universidad Nacional Autónoma de México, Cuernavaca, México, **3** John Innes Centre, Norwich, United Kingdom, **4** Department of Biology, University of York, York, United Kingdom, **5** School of Biological Sciences, University of East Anglia, Norwich, United Kingdom, **6** Department of Biological Sciences, University of Calgary, Calgary, Canada

Abstract

This work centres on the genomic comparisons of two closely-related nitrogen-fixing symbiotic bacteria, *Rhizobium leguminosarum* biovar *viciae* 3841 and *Rhizobium etli* CFN42. These strains maintain a stable genomic core that is also common to other rhizobia species plus a very variable and significant accessory component. The chromosomes are highly syntenic, whereas plasmids are related by fewer syntenic blocks and have mosaic structures. The pairs of plasmids p42f-pRL12, p42e-pRL11 and p42b-pRL9 as well large parts of p42c with pRL10 are shown to be similar, whereas the symbiotic plasmids (p42d and pRL10) are structurally unrelated and seem to follow distinct evolutionary paths. Even though purifying selection is acting on the whole genome, the accessory component is evolving more rapidly. This component is constituted largely for proteins for transport of diverse metabolites and elements of external origin. The present analysis allows us to conclude that a heterogeneous and quickly diversifying group of plasmids co-exists in a common genomic framework.

Citation: Crossman LC, Castillo-Ramírez S, McAnnula C, Lozano L, Vernikos GS, et al. (2008) A Common Genomic Framework for a Diverse Assembly of Plasmids in the Symbiotic Nitrogen Fixing Bacteria. PLoS ONE 3(7): e2567. doi:10.1371/journal.pone.0002567

Editor: Richard R. Copley, Wellcome Trust Centre for Human Genetics, United Kingdom

Received February 19, 2008; **Accepted** May 6, 2008; **Published** July 2, 2008

Funding: CONACyT 46333-Q and PAPIIT-UNAM IN223005-5 grants supported work on Rhizobium etli. Work on Rhizobium leguminosarum was supported by the Wellcome Trust and the BBSRC under grants 104/P16988, 208/BRE13665, and 208/PRS12210.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: lcc@sanger.ac.uk (LCC); vgonzal@ccg.unam.mx (VC)

Introduction

Rhizobium etli and *Rhizobium leguminosarum* bv *viciae* (henceforth called *R. leguminosarum*) are closely related species which are able to fix atmospheric nitrogen in symbiosis with specific leguminous plants. The common bean is the natural host of *R. etli* whereas *R. leguminosarum* interacts with peas, lentils, vetches and *Lathyrus* spp. Recently, we reported the complete genome sequences of a strain of *R. etli* and a strain of *R. leguminosarum* [1,2], but no comprehensive genome comparison between these species had been carried out. To date, several other complete genome sequences of symbiotic nitrogen fixing bacteria have been published: *Mesorhizobium loti*, *Bradyrhizobium japonicum*, *B. spp. ORS278*, *B. spp. BTAi1* and *Sinorhizobium meliloti* [3–6]). Our comparisons of *R. etli* and *R. leguminosarum* show that: 1) *Rhizobium* genomes are composed of "core" and "accessory" components; 2) the chromosomes are markedly conserved in gene content (despite differences in size) and amongst the closest species gene order is also conserved; 3) the plasmids are heterogeneous in size and gene content and in some cases no synteny can be seen even in comparison with phylogenetic neighbours.

Rhizobium field isolates have the unusual feature of harbouring several plasmids, ranging in size from 100 kb to >1,000 kb and the plasmid profiles of a particular isolate can be used to type strains reliably [7]. Since *R. etli* CFN42 and *R. leguminosarum* 3841 are the most closely-related rhizobial species yet sequenced and both strains have six large plasmids, a detailed genome comparison between them may help us interpret the evolutionary history of these prototypical accessory elements. Indeed, whole genome comparisons allowed us to discern the distinctive properties of the core genome, and also to highlight the genetic differences between these species.

Results

Main features of the compared species

Both *R. etli* CFN42 and *R. leguminosarum* 3841 have large genomes composed of a circular chromosome and six large plasmids [1,2]. The six CFN42 plasmids, pRetCFN42a-f, will be referred to as p42a-f throughout this article, whilst the six 3841 plasmids (sometimes known as pRL7JI-pRL12JI) are termed pRL7-12. The total size of the *R. etli* CFN42 genome is 1,221,081 bp shorter than that of *R. leguminosarum* 3841 (Table S1). The two smaller plasmids of *R. etli* are substantially larger than the two smallest plasmids of *R. leguminosarum*, whilst the opposite is the case for the other four plasmids (Table S1). *R. leguminosarum* plasmids comprise 34.8% of the total genome, whilst *R. etli* plasmids comprise an equivalent 32.9%. The two smallest *R. leguminosarum* plasmids are of lower than average GC content, whilst in *R. etli* the major nitrogen fixation plasmid (pSym; p42d)

and the smallest plasmid (p42a) are the only plasmids of significantly lower GC content. The largest plasmids in both genomes resemble their corresponding chromosomes both in GC content and dinucleotide signatures. Symbiotic functions, specified by the *nod*, *nol*, *nif* and *fix* genes, are mainly encoded by a single plasmid (p42d in *R. etli* and pRL10 in *R. leguminosarum*), but other symbiosis-related genes are located on other plasmids and in the chromosome [1,8]. The *R. etli* plasmid p42a is transferable at high frequencies and can help the mobilization of p42d [9–11] and p42d is also self-transmissible by conjugation [12] although its transfer ability is tightly repressed [13]. In *R. leguminosarum*, pRL7 and pRL8 are transmissible by conjugation, although neither carries a full set of *tra* genes [2].

Phylogenomic relatedness between *R. etli* and *R. leguminosarum*

R. leguminosarum and *R. etli* are closely related species, judged by 16S rRNA comparisons and other molecular criteria (Figure S1). We first tested the consistency of these traditional phylogenies with genome phylogenies obtained with all individual proteins included in quartops (QUARtet of Orthologous Proteins). To do this, we incorporated two other species of the Rhizobiaceae family, *S. meliloti* and the non-nitrogen-fixing *Agrobacterium tumefaciens*, whose complete genomes are also available.. A total of 33% and 39% of *R. leguminosarum* and *R. etli* proteins, respectively, were present in the Quartops; this equates to 2,392 predicted proteins representing core genes that are common to these four organisms (Table 1). Most of these predicted proteins are chromosomally encoded (2,054) but 338 belong to plasmids pRL9, pRL11 and pRL12. Three of the plasmids (pRL7, pRL8 and pRL10) do not have any proteins in Quartops. A total of 2,241 (85% of all proteins included in quartops) supports the phylogenetic relationship that proposes

Table 1. Quartops analysis with *R. leguminosarum*, *R. etli*, *A. tumefaciens and S. meliloti*.

	Total proteins	*No in quartops*	*Percentage in quartops*	*Rl-Re*	*Rl-At*	*Rl-Sm*
Chr	4736	2054	43.4	1951	25	23
pRL12	790	96	12.1	71	5	6
pRL11	635	147	23.1	136	-	5
pRL10	461	-	-	-	-	-
pRL9	313	95	30.3	83	4	4
pRL8	140	-	-	-	-	-
pRL7	188	-	-	-	-	-

doi:10.1371/journal.pone.0002567.t001

R. leguminosarum and *R. etli* are the most closely related. However, the high numbers of proteins absent from Quartops suggests that gene losses and gains might significantly have driven the diversification of the fast growing rhizobia. To investigate this area, we clustered all the predicted proteins of *R. etli*, *R. leguminosarum*, *S. meliloti* and *A. tumefaciens* into families by means of the MCL algorithm [14]. About 28% of the protein families identified (1,965 out 6,827) are shared by the four species, whereas about 10% (668) are only present in three species (Figure 1, bars 1–6). The rest of the protein families (13% or 908) occur in just two species. Most of these families (443) belong to the *R. etli-R. leguminosarum* pair, giving further support to the quartop phylogeny and the recent divergence of these two species (Figure 1, bars 7–11). Moreover, an appreciable number of families were particular to individual genomes. They belong to known and hypothetical

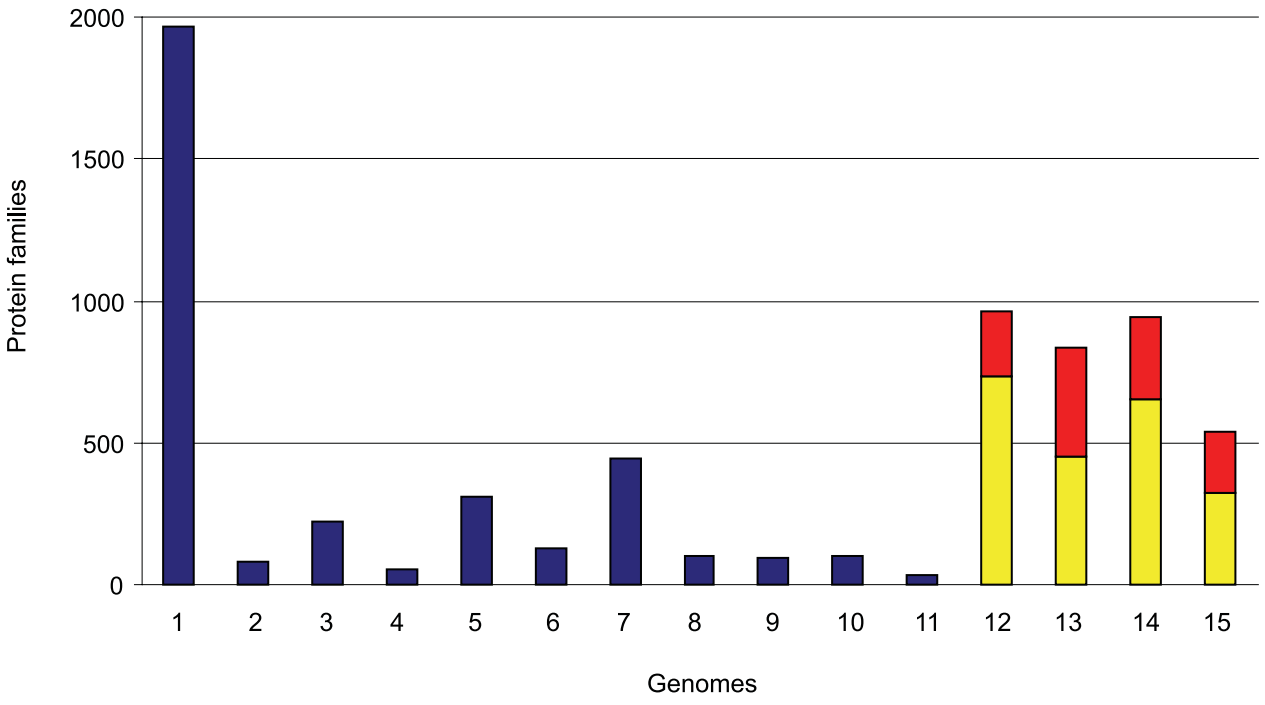

Figure 1. Distribution of protein families in the genomes of *S. meliloti* (S), *A. tumefaciens* (A), *R. leguminosarum* (L) and *R. etli* (E). - Bar number indicates the assignation of the protein families to the corresponding genome according to the following letters code: 1, SALE; 2, SAL; 3, ALE; 4, SAE; 5, SLE; 6. SA; 7, LE; 8, AL; 9, SE; 10, SL; 11, AE; 12, S; 13, A; 14, L; 15, E. Bars 12-15 show in red the proportion of orphan genes compared with those which match with known or hypothetical proteins present in the nr database of Genbank (yellow).
doi:10.1371/journal.pone.0002567.g001

families already present in the Genbank or they are orphan genes (Figure 1, bars 12–15). This confirms the previous findings that the coding potential of the rhizobial species is very variable while maintaining a stable common core.

Genome synteny

To investigate whether the evolutionary relationship between *R. etli* and *R. leguminosarum* is also maintained at the level of gene order, the whole genomes were compared using ACT and Nucmer softwares [15,16]. A clear syntenic pattern is distinguished between both chromosomes but it is also noticeable for some pairs of plasmids: (p42f-pRL12), (p42e-pRL11) and (p42b-pRL9) as well as large parts of p42c woth pRL10, suggesting a common origin (Figure 2). These observations are supported by the similarity of the replication genes, *repABC*, of those pairs of plasmids, as well as experimental demonstration of incompatibility between the plasmid pairs (Clark, Mattson, Garcia and Hynes, in preparation). Plasmids pRL7 and pRL8 appear to be unique to *R. leguminosarum* whilst p42a is peculiar to *R. etli* (see below). A more accurate measure of synteny between the genomes was obtained by calculating the length and number of colineal blocks (CBs). To do this, we employed a whole alignment obtained by Nucmer [16], then individual matches were clustered in CBs taking all the continuous segments separated by gaps less than 1kb. In total, 4,557,466 bp (70%) of the *R. etli* genome is contained in CBs with nucleotide identity about 85–95% (to *R. leguminosarum*). In the total genome of *R. leguminosarum*, 4,931,491 bp (63%) are contained in CBs. A total of 353 CBs >1 kb were recognized. The largest and most abundant (221) CBs are located on the chromosome and the rest on plasmids. Figure 3 shows that 81% and 74% of the chromosomes of *R. etli* and *R. leguminosarum* respectively are contained in CBs. Three of the *R. etli* plasmids have 44–58% of their genetic information in CBs that also occur in *R. leguminosarum*. Plasmids with fewer CBs are p42a, p42d, pRL7 and pRL8. Some of the plasmid pairs can be functionally identified by the presence of specific genes. For example, p42f and pRL12 carry some genes for flagellar biosynthesis (*flgLKE)* and for oxidative stress protection (*oxyR* and *katG*); p42e and pRL11 harbor cell division genes (*minCDE*), as well as thiamin, cobalamin, NAD biosynthetic genes (*thiMED*, *cobFGHIJKLM*, *nadABC*), and an isolated flagellin (*fla*) gene, as well as a rhamnose catabolism operon[17]. In some cases, *e.g. thiMED*, these genes are functionally interchangeable between these species [18]. A duplication of the *fixNOQP* operon in p42f [19], in *R. leguminosarum* is located in pRL9, a plasmid with

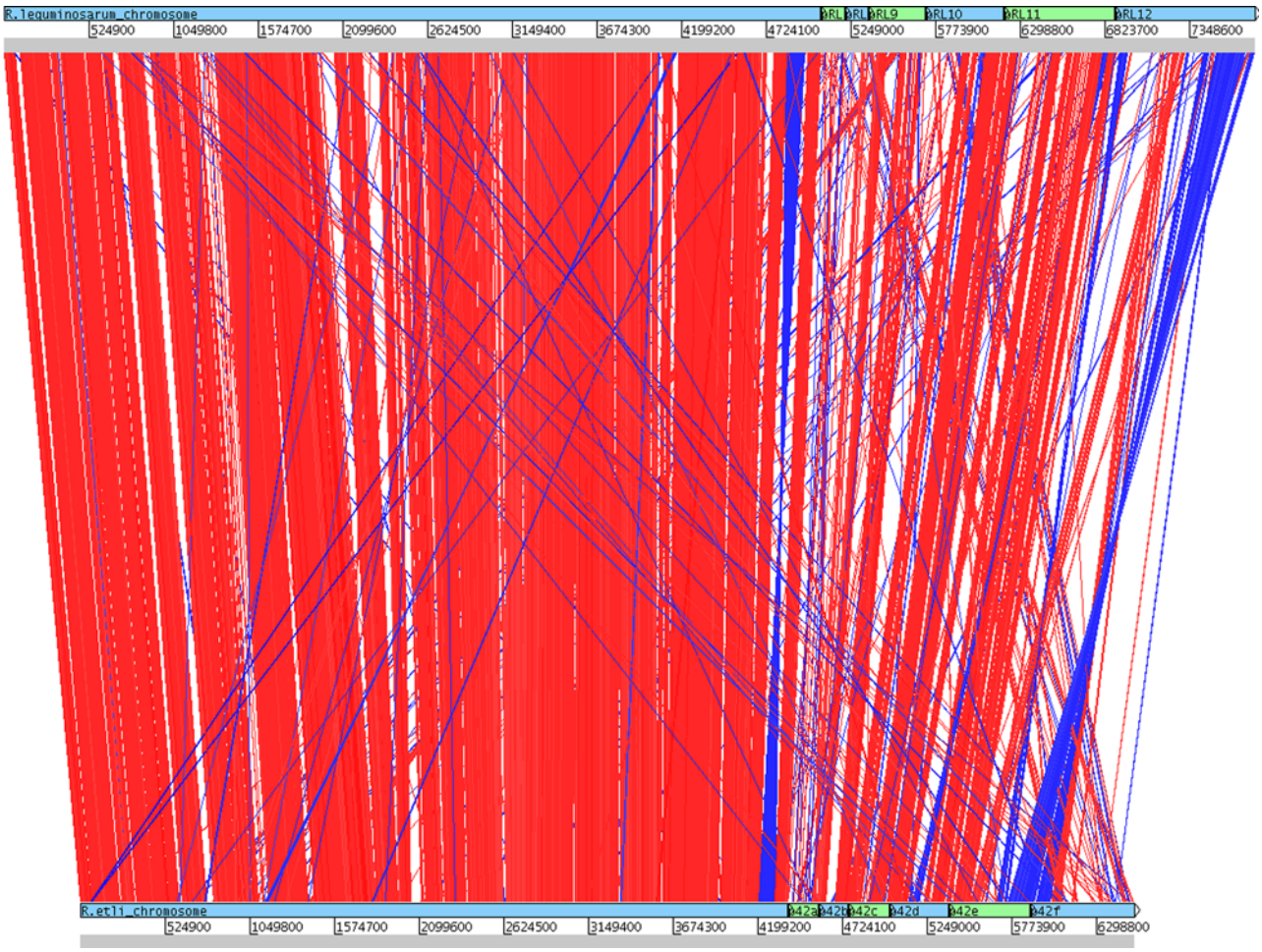

Figure 2. ACT View of Chromosome and Plasmids. The chromosomal and plasmid DNAs have been laid end-to-end and analysed using the Artemis comparison tool (ACT) [15]. Red bars represent close matches, whilst blue bars represent inverted close matches. The *R. leguminosarum* genome is at the top of the figure with replicons in the order Chromosome, pRL7, pRL8, pRL9, pRL10, pRL11, pRL12 whilst the *R. etli* genome is shown at the bottom of the figure in order Chromosome, p42a, p42b, p42c, p42d, p42e, p42f.
doi:10.1371/journal.pone.0002567.g002

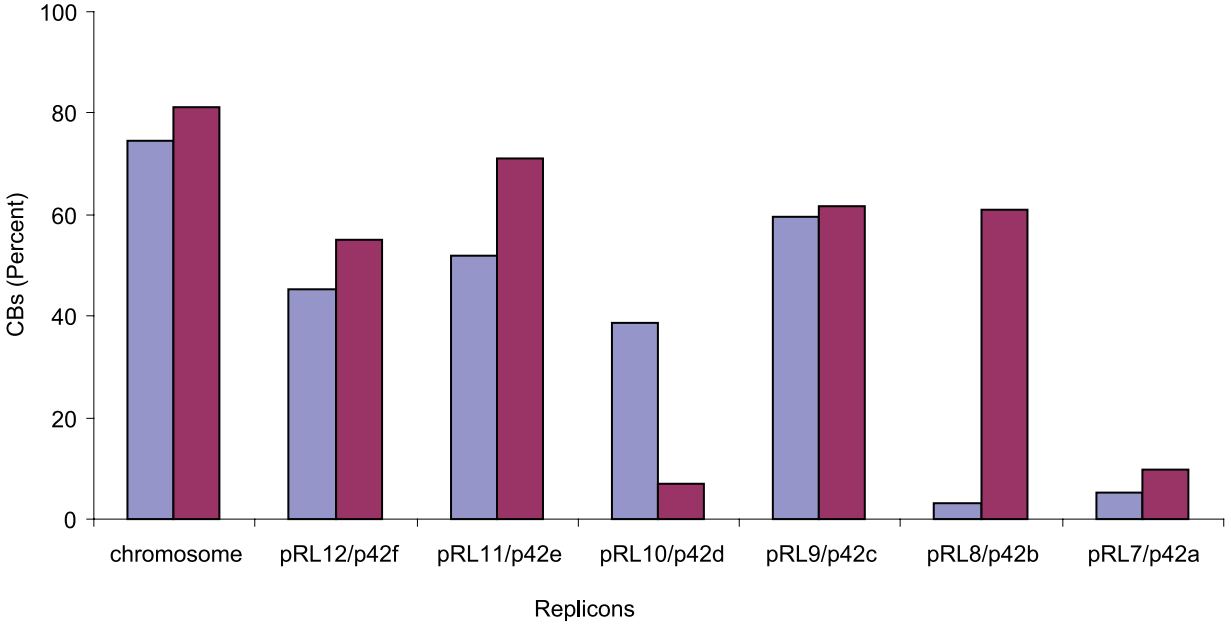

Figure 3. The proportion of synteny in the *R.etli* genome as compared to the *R. leguminosarum* genome. - The proportion of synteny is expressed as the percentage of the total DNA in CBs (Y axis) considering the total length of the pairs of replicons (X axis). Blue color *R. etli* CFN42; magenta, *R. leguminosarum* 3841.
doi:10.1371/journal.pone.0002567.g003

homologous segments to p42b. Conjugative plasmids pRL8, pRL7 and p42a, which are otherwise unrelated to each other, have homologous *tra-trb* systems.

Core genome composition and evolution

R. etli and *R. leguminosarum* share 5,470 genes with approximately 89–100% similarity (see methods). A significant fraction of these common genes (3,359 or 62%) is solely present in both chromosomes (Chromosomal Only, CHR-O). The rest are situated either in the chromosome or plasmids or exclusively in the plasmids (Non-Chromosomal, N-CHR). Using the Riley classification scheme [20], CHR-O genes are overrepresented in the categories corresponding to small and macromolecule metabolism, structural elements, regulators and hypothetical conserved genes. In contrast, the N-CHR group tends to contain genes implicated in processes like chemotaxis, chaperones, transport, and elements of external origin (Figure 4a). A detailed classification using COGs [21] reveals other differences between CHR-O and N-CHR groups. Some of the COGs that are overrepresented in N-CHR are COG K (replication, recombination and repair) and the COGs related with predicted transport and metabolism of carbohydrates, amino acids, lipids and inorganic ions (COG G, E, I, and P) (Figure 4b), but not COGs related to information storage and processing.

Differences between the CHR-O and the N-CHR gene compartments were also detected in regard to rates of evolution. To do this, we calculated the rates of nucleotide substitution per synonymous (Ks) and non-synonymous sites (Ka), for a subset of 2,917 single copy homologues (see methods; Figure 5). It is clear that both CHR-O and N-CHR homologous groups are under negative selection. Nevertheless, as seen by the slopes of the regression lines, the CHR-O group seems to be under stronger negative selection than the N-CHR group. However, many genes of the N-CHR group show higher Ka (>0.19) and Ks (>2.0) values than those of the CHR-O group. Therefore, negative selection is acting on the whole genome, but overall, the N-CHR gene compartment is less constrained.

Mosaic replicons

Despite the high level of genome conservation, it is reasonable to expect that some degree of intra-genomic recombination has occurred since these two strains *R. etli* and *R. leguminosarum* had a common ancestor. This was substantiated by comparing the locations of the N-CHR group of genes in the different replicons of both genomes. Approximately 7% of the chromosomal genes of *R. leguminosarum* are represented in the plasmids of *R. etli*, and 10% of the chromosomal genes of *R. etli* are located in the *R. leguminosarum* plasmids. As shown before, some pairs of plasmids are likely equivalent in terms of their global similarity, but they are mosaic replicons that contain genes from the other replicons. For instance, pRL12 has significant similarity with p42f, but also possesses genes that in *R. etli* are chromosomal or on another plasmid (Figure 6). A similar pattern is observed in the other replicons (Figure 6). Such heterogeneous composition of the plasmids has precluded any attempt to make a reliable plasmid phylogeny. One way to assess the phylogenetic relatedness among plasmids is to compare their RepABC proteins that are essential components for plasmid replication [22].. However, we observed here that only the p42c-pRL10, p42d-pRL11 and p42f-pRL12 pairs carry closely related replication systems. They share nucleotide identities greater than 82% in the three proteins, whereas the RepABC proteins of the other plasmids are poorly related. Therefore, the replication genes might have been shuffled several times among the distinct plasmids, perhaps to allow a number of plasmids to coexist in the same cell.

A potential symbiosis cassette

A comparison of the major symbiotic plasmids (pSyms) pRL10 and p42d shows that the *nif-nod* region in pRL10 is compacted into 60 kb, whereas in p42d it encompasses 125 kb. As many as 20

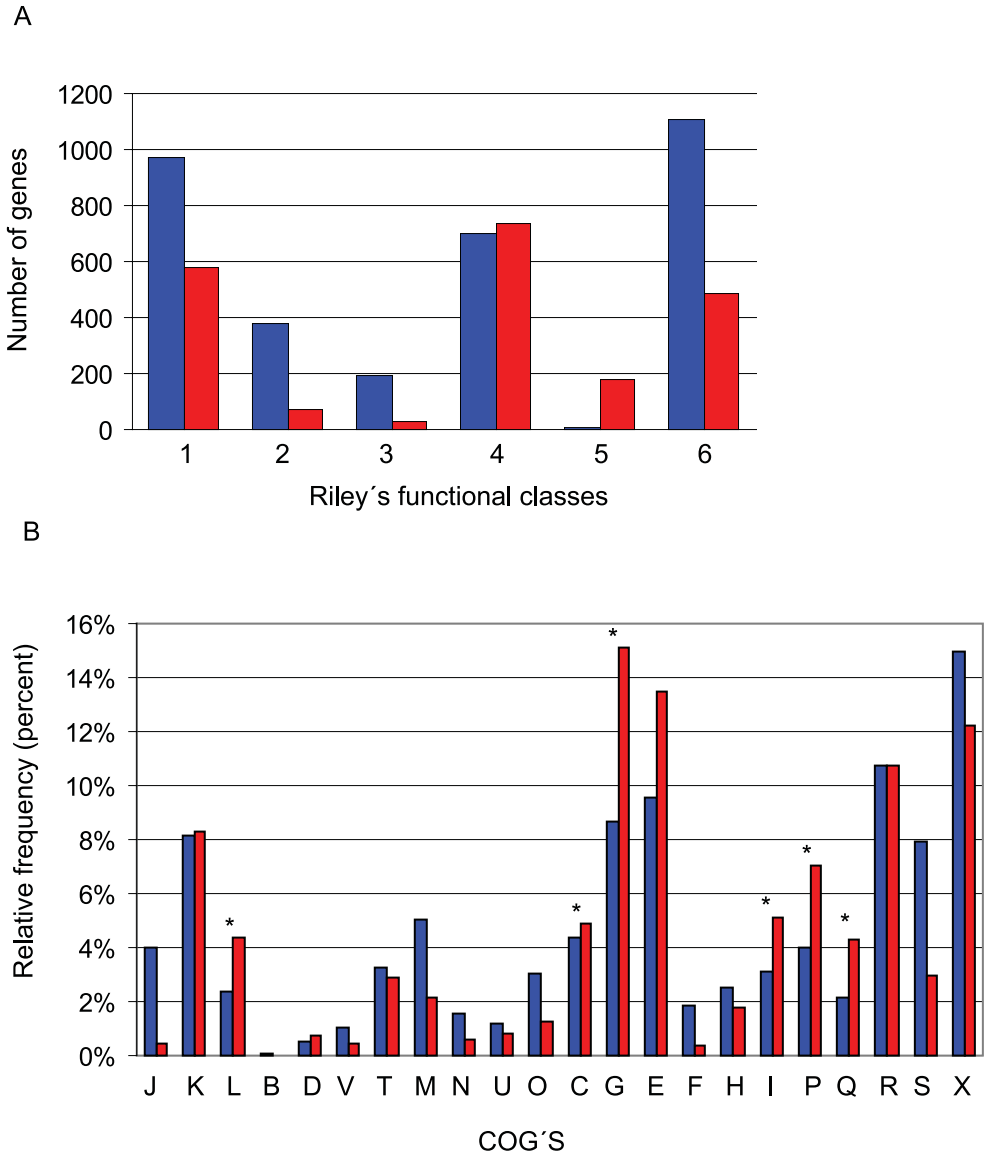

Figure 4. Functional bias in CHR-O (chromosomal only) and N-CHR (non-chromosomal) classes of homologues. - Figure 4a) Rileys categories: 1. small molecule metabolism. 2. Macromolecule metabolism. 3. Structural elements. 4. Cell process. 5. External origin. 6. Miscellaneous. 4b) COGs functional classification. Bars indicate the relative frequency for each COG J, Translation, ribosomal structure and biogenesis; K, Transcription; L, Replication, recombination and repair; B, chromatin structure and dynamics; D, Cell cycle control; V, Defense mechanisms; T, Signal transduction mechanisms; M, Cell wall, membrane envelope biogenesis; N, Cell motility; U, Intracellular trafficking and secretion; 0, Postranslational modification and chaperones; C, Energy production and conversion; G, Carhohydrate transport and metabolism; E, Amino acid transport and metabolism; F, Nucleotide transport and metabolism; H, Coenzyme transport and metabolism; I, Inorganic ion transport and metabolism; P, inorganic ion transport and metabolism; Q, Secondary metabolites biosynthesis, transport and catabolism; R, General function prediction; S, function unknown; X, No COG.
doi:10.1371/journal.pone.0002567.g004

common *nod* and *nif* genes have been identified in comparisons among complete sequences of pSyms and symbiotic islands of different rhizobia [8]. The plasmid pRL10 contains 18 of these genes and has a particularly enhanced set of nodulation genes, including genes that lack homologs in *R. etli*, such as *nodTNMLEF* and *rhiABCR*. In contrast, pRL10 has a restricted set of genes for nitrogenase maturation, lacking *nifS*, *nifW*, *nifZ*, *nifX*, *iscN*, and *nifU*, which are present in *R. etli* and in other rhizobia. Besides the common nodulation genes, the *R. etli* pSym possesses *nolT*, *nolL*, *nolR*, *noeI*, *noeJ*, and a Type III secretion system.

The symbiotic genes of *R. leguminosarum* may have been acquired by horizontal gene transfer, since an *in silico* analysis of pRL10 with the Alien Hunter program [23] reveals that its symbiotic gene cluster, which includes the *nif*, *nod*, *rhi* and *fix* genes, is located in a short potentially mobile region of DNA (~63.5 kb). Internal to this region are the *nifNEKDH* genes that are found bounded by two identical IS element repeat regions. The *rhi* and *nod* gene cluster, together with *fixABCX*, lie adjacent on this potential genomic island and are potentially bounded by 20 bp repeats, whilst the *fixNOPQ* and *fixGHIS* genes lie immediately downstream on a separate putative genomic island of approximately 11,000 bp, potentially bounded by 18 bp repeats (Figure 7). It is possible that the *fixNOPQ*, *fixGHIS* island represents a second acquisition of DNA as an independent event. These adjacent symbiotic nitrogen

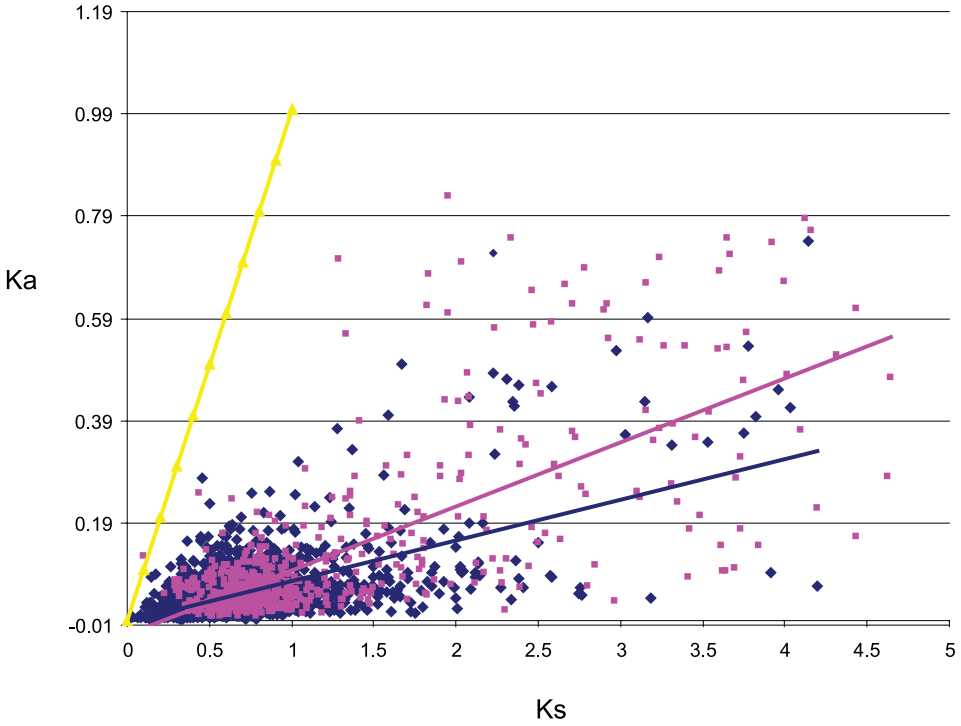

Figure 5. Rates of synonymous (Ks) and non-synonymous substitutions (Ka) in orthologous genes of *R. etli* and *R. leguminosarum*. - Neutrality line (Ka = Ks) is indicated in yellow. Linear regressions for CO class (blue color line and diamonds) and NC class (rose color line and diamonds) are indicated. As neutrality assumes equal nucleotide substitutions rates per synonymous and non-synonymous sites, points under the neutrality line indicate negative selection. Strong selective constraints are acting on genes of the CHR-O class ($R2 = 0.6124$; $P \ll 0.001$) but are slightly less intense for some genes of the N-CHR class ($R2 = 0.5094$), as can be seen by the dispersion of the rose color diamonds.
doi:10.1371/journal.pone.0002567.g005

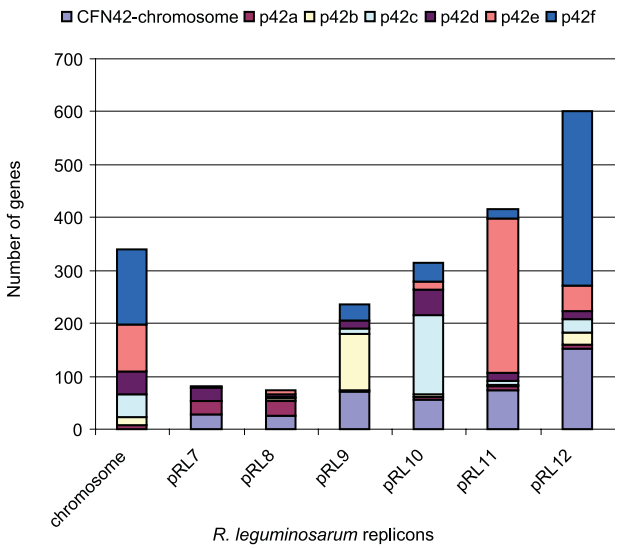

Figure 6. Composition of the *R. leguminosarum* and *R.etli* genomes according to N-CHR homologues. - The composition of the *R. leguminosarum* genome compared to the replicons of *R. etli* is shown. The replicon name is given at the base of the figure and color key to the right of the figure. Genes on the *R. etli* chromosome may be elsewhere on the *R. leguminosarum* genome (as shown in pale blue), genes from *R. etli* p42a (burgundy), p42b (cream), p42c (cyan), p42d (purple), p42e (salmon) and p42f are shown in royal blue.
doi:10.1371/journal.pone.0002567.g006

fixation gene clusters are located in one particular region of the plasmid with six other short potentially horizontally transferred areas. The remainder of the pRL10 plasmid is highly similar to the p42c plasmid of *Rhizobium etli*. By contrast, the symbiotic nitrogen fixation genes are scattered throughout 125 kb of the p42d plasmid of *R. etli*. However, this region is surrounded by insertion sequences, which prompted the idea that it might be transposable [8]. When plasmid p42d was analysed by the Alien Hunter program 16 regions were detected as atypical. These regions contain the Type III transport system genes, *nod* genes, genes for virulence and conjugation (*vir* and *tra*), as well as cytochrome and chemotaxis genes (Figure S2). They are bordered by repeated sequences that might represent potential composite transposons when the repeats are homologous insertion sequences. Alternatively, the chimeric structure of p42d might have been the result of multiple gene exchanges and rearrangements.

Physiological differences

The consequences of the evolutionary process of gain and losses are reflected in some physiological differences. For example, no candidate genes for respiratory nitrate reductases have been identified in the *R. etli* or *R. leguminosarum* genomes, however, the *nirK* gene for the respiratory nitrite reductase is present on *R. etli* p42f (RE1PF0000526). This gene appears to participate in nitrite detoxification [24]. Nitric oxide (NO) removal is encoded by *R. etli* as a predicted *norECBD* operon on the p42f plasmid located in proximity to the *nirKV* and probable regulators. These genes are absent in *R. leguminosarum*, although there are possible alternative NO consumption systems. One of such pathways encoded chromosomally by both *R. etli* and *R. leguminosarum* is *via* the assimilatory nitrite reductase. Another difference is the presence of erythritol catabolic

Figure 7. Diagram of the *R. leguminosarum* major nitrogen fixation gene cluster. - This cluster represents a potentially laterally transferred region of DNA. Major nitrogen fixation genes are represented as blocks and are as shown in the color key.
doi:10.1371/journal.pone.0002567.g007

genes, possibly originating from a horizontal transfer event, on pRL12 [25]. This gene cluster is absent from the CFN42 genome.

Since the lifestyles of *R. etli* and *R. leguminosarum* bv. *viciae* are similar, they may have similar responses to environmental stimuli. Thus, they may respond similarly with respect to environmental stimuli. For instance, population-density-dependent gene induction by N-acyl homoserine lactones (AHLs) influence symbiotic functions such as nodulation, nitrogen fixation, and surface polysaccharide production as well as several aspects of growth including plasmid transfer and stationary phase adaptation [8,26] for reviews). Comparative analysis with known AHL regulators shows that there are 11 LuxR-type regulators in *R. etli* and 9 in *R. leguminosarum*. Some of them are known AHL regulators (CinR, TraR and RhiR) with associated AHL synthases (CinI, TraI, and RhiI) but there are also three other regulators, ExpR, AvhR and AsaR, for which there are no matching AHL synthases. In addition, we identified three LuxR-like sequences in *R. leguminosarum* (RL0606, RL0607 and RL3528) that matched the LuxR family over their entire length, but they could not be identified using protein domain searches. Two of these (RL0606 and RL0607) are highly conserved in *R. etli*, and in each case, they are located within a cluster of genes associated with bacterial motility, chemotaxis and flagella biosynthesis. Two related genes, *visN* and *visR* from *S. meliloti* strain RU10/406 act as global regulators of flagellar motility and chemotaxis, their products probably functioning as a heterodimer [27]. Although the third regulator (RL3258) also appears to be conserved in *R. etli* (CH03080) it has no known function. Remarkably, RhiR regulates the *rhiABC* operon that plays an undefined role in legume infection in *R. leguminosarum*, although this regulator is not present in *R. etli*, [28].

Discussion

Rhizobium genomes consist of single circular chromosomes and several large plasmids. It is not understood why these genomes are so large and divided. Young *et al.* (2006) proposed that microbial life in the soil, a very heterogeneous environment, selects for a versatile genomes that encode multiple capabilities [2]. Therefore, genome comparisons between closely related *Rhizobium* species may indicate how variable these capabilities could be, as well as establishing whether they are distributed throughout the genome or in particular replicons. The comparative analysis presented here allows us to conclude that most of the differences between *R. etli* and *R. leguminosarum* tend to be in the plasmids. Previous genomic comparisons of *S. meliloti*, *A. tumefaciens*, and *R. etli* have shown that chromosomes are well conserved both in gene content and gene order, whereas plasmids have few common regions (*nif-nod*, *tra-trb*, *vir*, and others) and a lack of synteny [8]. These comparisons indicate that the plasmids in those three species are not closely related phylogenetically or that they have undergone many recombination events. Our analysis reveals many syntenic blocks exist between some pairs of plasmids of *R. etli* and *R. leguminosarum* (p42f-pRL12, p42e-pRL11 and p42b-pRL9 as well large parts of p42c with pRL10) suggesting a common origin. Plasmids of *R. etli* are smaller than those of *R. leguminosarum*, and 44–58% of their length is contained in CBs common to *R. leguminosarum*. Nonetheless, the phylogenetic relationships among the plasmids remain obscure.

A particular case of the mosaic structure of *Rhizobium* plasmids is shown by comparison of the symbiotic plasmids. In *R. leguminosarum* the pSyms are variable in size and also differ in *repC* group [2]. It has been noted that pRL10 and pRL1 (a pSym of 200 kb in *R. leguminosarum*) have a virtually identical *nod-nif* region, but the remainder of these plasmids appear to be dissimilar [2]. Speculatively, the entire symbiotic region may be a mobile element in *R. leguminosarum*, as has been proposed for the symbiotic region of p42d [8]. Although direct evidence for this scenario is still lacking, it is plausible given the observed recombinational plasticity displayed by rhizobial plasmids (reviewed by [29,30]. Nevertheless, the overall structure of pRL10 more closely resembles p42c than p42d of *R. etli* (Figure 2). Extensive syntenic regions are common between pRL10 and p42c, accounting for

59% of the length of p42c (Figure 2). Thus, either pRL10 has gained a large insertion carrying the symbiotic nitrogen fixation functions, or p42c has suffered a large deletion of these genes. We show here that the former possibility could be plausible since the *nif-nod* region is a potential symbiotic cassette surrounded by repeated sequences. Furthermore, the structural differences between the pSyms of *R. etli* and *R. leguminosarum*, prompt us to suggest that they have evolved differently. In *R. leguminosarum* the Sym region resembles an specific "cassette", whereas in *R. etli* the partial nucleotide sequence of different pSyms suggests that their diversification is driven by general recombination [31].

Some authors have proposed that bacterial genomes consist of "core" and "accessory" components [2,32]. The "Core" component, exemplified by the chromosome, is more stable and changes more slowly over time than the "accessory" component. Plasmids are prototypical accessory elements composed of genes from different genomic contexts and evolutionary origin. As shown here, *R. etli* and *R. leguminosarum* are good models to study the evolution of plasmid ("accessory") versus chromosome ("core") evolution. Their chromosomes are nearly identical and harbor a distinct collection of plasmids that have evolved at different rates to the chromosome. It is tantalizing to speculate that these organisms can recruit plasmids from a pool in their soil environment [32]. Plasmids p42a, p42d, pRL7 and pRL8, in particular, seem to be the outliers. Other plasmids share many common regions and might have been part of the ancestral chromosome. Shuffling of the *repABC* genes might be a strategy to allow many plasmids to coexist in the same bacterium, and might explain the amazing plasmid diversity of *Rhizobium*. A more comprehensive picture of the evolution of the partitioned genomes can only be reached by comparing the respective plasmid pool of additional strains of *R. etli* and *R. leguminosarum* to describe how they are able to function in a common genomic framework.

Methods

Phylogenetic analysis

The 16S rRNA sequences were downloaded from EMBL for *R. leguminosarum* bv viciae Rlv3841, *R. etli* CFN42, *Agrobacterium tumefaciens* C58, *Sinorhizobium meliloti* 2011, *Mesorhizobium loti* MAFF303099, *Bradyrhizobium japonicum* USDA 110 and *Escherichia coli* T10. We first aligned the sequences using ClustalX [33] and generated a maximum likelihood tree using the PHYLIP package [34].

Genome Comparisons

The complete nucleotide sequences of the *R. etli* CFN42 and *R. leguminosarum* Rlv3841 were obtained from Genbank (Accession numbers: *R. etli*, NC_007761-NC_007766, and NC_004041; *R. leguminosarum* NC_008378-NC008384). The sequences of the replicons for each genome were concatenated and used in a global comparison using ACT [15] and the Nucmer application of the Mummer package [16], with the default settings. To calculate the CBs, we took the nucmer.delta output and then parsed it with the show-coords utility. Syntenic segments >1 kb and separated by >1 kb were curated with *ad hoc* perl scripts and manual editing.

Clustering of protein families. First we did BLAST-P comparisons of "all versus all" complete proteomes of *R. etli*, *R. leguminosarum*, *S. meliloti* and *A. tumefaciens*. Clustering was achieved with MCL using an e-value of 10^{-7} and an inflation parameter of 1.5 [14].

Homolog grouping and analysis of evolutionary rates

The most probable set of homologous proteins shared by *R. etli* and *R. leguminosarum* was identified using a reciprocal best-hit criterion. To that end, all *R. etli* predicted proteins were searched against the *R. leguminosarum* predicted proteome and *vice versa* using BLAST with cutoff e value of 10^{-12} and employing the Blosum-80 matrix [35]. In addition to this criterion, to be included in a homolog group the difference in length between the subject protein and query protein had to be $<10\%$, the alignment region had to be at least of 80%, and there had to be a at least 50% similarity of both query and target sizes. We identified 5,470 homolog groups. The whole set was divided into two subdivisions. The first subdivision contains all the homolog groups in which there was only one protein per genome (unique bidirectional hits or possible orthologs, 2,917). The second subdivision contains homolog groups in which there is more than one protein in at least one genome, that is, possible paralogs (2,533). Further classification of the homolog groups was based on their localization. The "chromosomal-only" group (CHR-O) of homologs is present only in the chromosomes of both genomes, whereas the non-chromosomal group (N-CHR) was located either in chromosome or in plasmids, or exclusively in plasmids. Exclusive genes were recorded as those with no hits in the genomes at e-value of $<10^{-6}$. The number of nucleotide substitutions per synonymous site "Ks" and the number of nucleotide substitutions per non-synonymous site "Ka" were determined with yn00 from PAML13.14 [36]

Identification of genes involved in quorum sensing

We identified LuxI homologues using homology searches and independently determined proteins matching InterPro family IPR001690 (Autoinducer synthase). Both methods gave identical results. LuxR homologues were identified using homology searches as a guide, but were not by themselves used to identify likely LuxR proteins since the C-terminal DNA-binding domain in LuxR is also present at the C-terminus of a number of other proteins. Proteins containing the InterPro domain IPR005143 were identified, which corresponds to the N-terminal autoinducer-binding domain.

Identification of horizontally acquired regions

Potentially horizontally acquired areas of DNA were identified with the Alien Hunter program, available from http://www.sanger.ac.uk/Software/analysis/alien_hunter.

Supporting Information

Table S1 General features of the Genomes of *R.etli* and *R.leguminosarum*. A comparison of the main features of the genomes of *Rhizobium leguminosarum* and *Rhizobium etli*. Each replicon is described in terms of length in base pairs, %G+C content and number of coding sequences (CDS).
Found at: doi:10.1371/journal.pone.0002567.s001 (0.04 MB DOC)

Figure S1 Phylogenetic tree. Maximum likelihood phylogenetic tree showing bacteria related to *R.etli* and *R.leguminosarum*
Found at: doi:10.1371/journal.pone.0002567.s002 (0.08 MB TIF)

Figure S2 Chimeric structure of *R.etli* plasmid p42d. The circles show (outermost to innermost): 1. Atypical regions as bars of degraded colour (red to pale rose) according to the scores obtained from Alien Hunter (red, highest score 73 over a threshold of 32). 2. The 125 kb nif-nod region. 3, CDS of p42d according to the following colour code: blue, nodulation genes; yellow, nif genes; red, energy transfer genes (*fix* genes); green, insertion sequences; pink, transfer and replication genes; brown, hypotheticals; grey, transport (*vir* and *tss*III genes); sky blue, regulators. 4. Insertion sequences 5. Repeats from 100 to 300 identical nucleotides (black lines); repeats higher than 300 nucleotides (red lines).
Found at: doi:10.1371/journal.pone.0002567.s003 (0.54 MB EPS)

Acknowledgments

The UNAM group wishes to thank the assistance of Patricia Bustos, Rosa I. Santamaría and José L. Fernández. Critical reading of the manuscript by Xianwu Guo is also appreciated.

Author Contributions

Analyzed the data: GV LC JY AW GM JD CM JA ZG IH DR VG SC. Contributed reagents/materials/analysis tools: JP LC ZG VG. Wrote the paper: JP AJ LC GM JD CM MH GD VG.

References

1. González V, Santamaría RI, Bustos P, Hernández-González I, Medrano-Soto A, et al. (2006) The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. Proc Natl Acad Sci U S A 103: 3834–3839.
2. Young JP, Crossman LC, Johnston AW, Thomson NR, Ghazoui ZF, et al. (2006) The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. Genome Biol 7: R34.
3. Kaneko T, Nakamura Y, Sato S, Asamizu E, Kato T, et al. (2000) Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. DNA Res 7: 331–338.
4. Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, et al. (2002) Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110 (supplement). DNA Res 9: 225–256.
5. Galibert F, Finan TM, Long SR, Pühler A, Abola P, et al. (2001) The composite genome of the legume symbiont *Sinorhizobium meliloti*. Science 293: 668–672.
6. Giraud E, Moulin L, Vallenet D, Barbe V, Cytryn E, et al. (2007) Legumes symbioses: absence of Nod genes in photosynthetic bradyrhizobia. Science 316: 1307–1312.
7. Jumas-Bilak E, Michaux-Charachon S, Bourg G, Ramuz M, Allardet-Servent A (1998) Unconventional genomic organization in the alpha subgroup of the Proteobacteria. J Bacteriol 180: 2749–2755.
8. González V, Bustos P, Ramírez-Romero MA, Medrano-Soto A, Salgado H, et al. (2003) The mosaic structure of the symbiotic plasmid of *Rhizobium etli* CFN42 and its relation to other symbiotic genome compartments. Genome Biol 4: R36.
9. Brom S, García-de los Santos A, Cervantes L, Palacios R, Romero D (2000) In *Rhizobium etli* symbiotic plasmid transfer, nodulation competitivity and cellular growth require interaction among different replicons. Plasmid 44: 34–43.
10. Tun-Garrido C, Bustos P, González V, Brom S (2003) Conjugative transfer of p42a from *Rhizobium etli* CFN42, which is required for mobilization of the symbiotic plasmid, is regulated by quorum sensing. J Bacteriol 185: 1681–1692.
11. Brom S, Girard L, Tun-Garrido C, García-de los Santos A, Bustos P, et al. (2004) Transfer of the symbiotic plasmid of *Rhizobium etli* CFN42 requires cointegration with p42a, which may be mediated by site-specific recombination. J Bacteriol 186: 7538–7548.
12. Pérez-Mendoza D, Domínguez-Ferreras A, Muñoz S, Soto MJ, Olivares J, et al. (2004) Identification of functional mob regions in *Rhizobium etli*: evidence for self-transmissibility of the symbiotic plasmid pRetCFN42d. J Bacteriol 186: 5753–5761.
13. Pérez-Mendoza D, Sepúlveda E, Pando V, Muñoz S, Nogales J, et al. (2005) Identification of the *rctA* gene, which is required for repression of conjugative transfer of rhizobial symbiotic megaplasmids. J Bacteriol 187: 7341–7350.
14. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–1584.
15. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, et al. (2005) ACT: the Artemis Comparison Tool. Bioinformatics 21: 3422–3423.
16. Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 30: 2478–2483.
17. Richardson JS, Hynes MF, Oresnik IJ (2004) A genetic locus necessary for rhamnose uptake and catabolism in *Rhizobium leguminosarum* bv. *trifolii*. J Bacteriol 186: 8433–8442.
18. Karunakaran R, Ebert K, Harvey S, Leonard ME, Ramachandran V, et al. (2006) Thiamine is synthesized by a salvage pathway in *Rhizobium leguminosarum* bv. *viciae* strain 3841. J Bacteriol 188: 6661–6668.
19. Girard L, Brom S, Davalos A, López O, Soberón M, et al. (2000) Differential regulation of *fixN*-reiterated genes in *Rhizobium etli* by a novel *fixL-fixK* cascade. Mol Plant Microbe Interact 13: 1283–1292.
20. Riley M (1993) Functions of the gene products of *Escherichia coli*. Microbiol Rev 57: 862–952.
21. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631–637.
22. Cevallos MA, Porta H, Izquierdo J, Tun-Garrido C, García-de-los-Santos A, et al. (2002) *Rhizobium etli* CFN42 contains at least three plasmids of the *repABC* family: a structural and evolutionary analysis. Plasmid 48: 104–116.
23. Vernikos GS, Parkhill J (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics 22: 2196–2203.
24. Bueno E, Gómez-Hernández N, Girard L, Bedmar EJ, Delgado MJ (2005) Function of the *Rhizobium etli* CFN42 *nirK* gene in nitrite metabolism. Biochem Soc Trans 33: 162–163.
25. Yost CK, Rath AM, Noel TC, Hynes MF (2006) Characterization of genes involved in erythritol catabolism in *Rhizobium leguminosarum* bv. *viciae*. Microbiology 152: 2061–2074.
26. Sánchez-Contreras M, Bauer WD, Gao M, Robinson JB, Allan Downie J (2007) Quorum-sensing regulation in rhizobia and its role in symbiotic interactions with legumes. Philos Trans R Soc Lond B Biol Sci 362: 1149–1163.
27. Sourjik V, Muschler P, Scharf B, Schmitt R (2000) VisN and VisR are global regulators of chemotaxis, flagellar, and motility genes in *Sinorhizobium (Rhizobium) meliloti*. J Bacteriol 182: 782–788.
28. Rosemeyer V, Michiels J, Verreth C, Vanderleyden J (1998) luxI- and luxR-homologous genes of *Rhizobium etli* CNPAF512 contribute to synthesis of autoinducer molecules and nodulation of *Phaseolus vulgaris*. J Bacteriol 180: 815–821.
29. Palacios R, Flores M (2005) Genome dynamics in rhizobial organisms. In: Newton WE, Palacios R, eds. In Genomes and Genomics of Nitrogen-fixing Organisms Springer. pp 183–200.
30. Romero D, Brom S (2004) The symbiotic plasmids of the *Rhizobiaceae*. In: (Chapter 12), pp 271–290. In: Phillips G, Funnell BE, eds. Plasmid Biology: American Society for Microbiology.
31. Flores M, Morales L, Avila A, González V, Bustos P, et al. (2005) Diversification of DNA sequences in the symbiotic genome of *Rhizobium etli*. J Bacteriol 187: 7185–7192.
32. Reanney D (1976) Extrachromosomal elements as possible agents of adaptation and development. Bacteriol Rev 40: 552–590.
33. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876–4882.
34. Felsenstein J (2005) "Phylip (Phylogeny Inference Package) version 3.6." Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
36. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556.

Applied and Environmental Microbiology, Mar. 2010, p. 1604–1614 Vol. 76, No. 5
0099-2240/10/$12.00 doi:10.1128/AEM.02039-09

Conserved Symbiotic Plasmid DNA Sequences in the Multireplicon Pangenomic Structure of *Rhizobium etli*[▽]

Víctor González,* José L. Acosta, Rosa I. Santamaría, Patricia Bustos, José L. Fernández, Ismael L. Hernández González, Rafael Díaz, Margarita Flores, Rafael Palacios, Jaime Mora, and Guillermo Dávila

Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Av. Universidad N/C Col. Chamilpa, Apdo. Postal 565-A, Cuernavaca 62210, Mexico

Received 24 August 2009/Accepted 23 December 2009

Strains of the same bacterial species often show considerable genomic variation. To examine the extent of such variation in *Rhizobium etli*, the complete genome sequence of *R. etli* CIAT652 and the partial genomic sequences of six additional *R. etli* strains having different geographical origins were determined. The sequences were compared with each other and with the previously reported genome sequence of *R. etli* CFN42. DNA sequences common to all strains constituted the greater part of these genomes and were localized in both the chromosome and large plasmids. About 700 to 1,000 kb of DNA that did not match sequences of the complete genomes of strains CIAT652 and CFN42 was unique to each *R. etli* strain. These sequences were distributed throughout the chromosome as individual genes or chromosomal islands and in plasmids, and they encoded accessory functions, such as transport of sugars and amino acids, or secondary metabolism; they also included mobile elements and hypothetical genes. Sequences corresponding to symbiotic plasmids showed high levels of nucleotide identity (about 98 to 99%), whereas chromosomal sequences and the sequences with matches to other plasmids showed lower levels of identity (on average, about 90 to 95%). We concluded that *R. etli* has a pangenomic structure with a core genome composed of both chromosomal and plasmid sequences, including a highly conserved symbiotic plasmid, despite the overall genomic divergence.

It is becoming clear that bacterial genomes of strains of the same species vary widely both in size and in gene composition (39). An unexpected degree of genomic diversity has been found by comparing whole genomes (39). For instance, in *Escherichia coli* strains, differences of up to 1,400 kb account for some strain-specific pathogenic traits (5, 56). The extent of intraspecies genome diversity varies in different bacterial lineages. Some species have a wide range of variation; these species include *E. coli* (42), *Streptococcus agalactiae* (53), and *Haloquadratum walsbyi* (34). Other bacteria display only limited gene content diversity; an example is *Ureaplasma urealyticum* (1, 54). Tettelin and colleagues have suggested that bacterial species can be characterized by the presence of a pangenome consisting of a core genome containing genes present in all strains and a dispensable genome consisting of partially shared and strain-specific genes (53, 54). This concept is rooted in the earlier ideas of Reanney (43) and Campbell (7) concerning the structure of bacterial populations, and it indicates both that there is a pool of accessory genetic information in bacterial species and that strains of the same or even different species can obtain this information by horizontal transfer mechanisms (7, 43).

Genome size and diversity are related to bacterial lifestyle. Small genomes are typical of strict pathogens such as *Rickettsia prowazekii* (2) and endosymbionts such as *Buchnera aphidicola* (44a). In contrast, free-living bacteria, such as *Pseudomonas syringae* and *Streptomyces coelicolor*, have large genomes (4, 6). The bacteria with the largest genomes are common inhabitants of heterogeneous environments, such as soil, where energy sources are limited but diverse (32). An increase in genome size is attributable mainly to expansion of functions such as secondary metabolism, transport of metabolites, and gene regulation. All these features are common to the nitrogen-fixing symbiotic bacteria of legumes, which are collectively known as rhizobia, and their close relative the plant pathogen *Agrobacterium*. The genomes of such bacterial species have diverse architectures with circular chromosomes that are different sizes or linear chromosomes, like that in *Agrobacterium* species, and the organisms contain variable numbers of large plasmids (31, 49). Comparative genomic studies have highlighted the conservation of gene content and order among the chromosomes of some species of rhizobia (22, 23, 25, 40). Furthermore, Guerrero and colleagues (25) observed that most essential genes occur in syntenic arrangements and display a higher level of sequence identity than nonsyntenic genes. In contrast, plasmids, including symbiotic plasmids and symbiotic chromosomal islands (like those in *Mesorhizobium loti* and *Bradyrhizobium japonicum*) are poorly conserved in terms of both gene content and gene order (21). It is not clear what evolutionary advantage, if any, is provided by multipartite genomes, but some authors have speculated that such genomes may allow further accumulation of genes independent of the chromosome. Recently, Slater and coworkers (46) proposed a model for the origin of secondary chromosomes. Their idea is based on the notion of intragenomic gene transfers that might occur from primary chromosomes to ancestral plasmids of the

* Corresponding author. Mailing address: Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Av. Universidad N/C Col. Chamilpa, Apdo. Postal 565-A, Cuernavaca 62210, Mexico. Phone: 52 777 3291690. Fax: 52 777 3276120. E-mail: vgonzal@ccg.unam.mx.

▽ Published ahead of print on 4 January 2010.

TABLE 1. Genome features and sequence data for the *R. etli* strains used in this work

Strain	Origin	No. of plasmids	No. of readings > Q20	No. of contigs	No. of sequenced bases in contigs	Predicted genome length (bp)	Coverage
CFN42	México	6		7	6,530,228	6,530,230	9×
CIAT652	Costa Rica	3	71,605	4	6,448,138	6,498,420	7.9×
Kim5	USA	2	14,359	2,749	4,416,663	5,982,980	0.96×
Brasil5	Brazil	4	13,920	2,734	3,762,131	5,958,880	0.93×
8C-3	Spain	3	12,866	2,732	3,760,107	6,007,740	0.86×
GR56	Spain	4	12,628	2,315	4,472,919	5,034,290	1×
IE4771	México	4	14,613	2,904	4,317,800	6,335,090	0.92×
CIAT894	Colombia	4	14,081	2,852	3,998,701	6,232,930	0.9×

repABC type. Observations of conservation of clusters of genes in secondary chromosomes or in large plasmids that retain synteny with respect to the main chromosome support this hypothesis (46).

We have been studying *Rhizobium etli* as a multipartite genome model species (23). This organism is a free-living soil bacterium that is able to form nodules and fix nitrogen in the roots of bean plants. The genome of *R. etli* is partitioned into several replicons, a circular chromosome, and several large plasmids. In the reference strain *R. etli* CFN42, the genome is composed of a circular chromosome consisting of about 4,381 kb and 6 large plasmids whose total size is 2,148 kb (23). A 371-kb plasmid, termed pSym or the symbiotic plasmid, contains most of the genes required for symbiosis (21). Previous studies have described the high level of genetic diversity among geographically different *R. etli* isolates (41). The strains are also variable with respect to the number and size of plasmids. Nevertheless, there has been no direct measurement of diversity at the genomic level, nor have comparative studies of shared and particular genomic features of *R. etli* strains been reported. Therefore, to assess the degrees of genomic difference and genomic similarity in *R. etli*, we obtained the complete genomic sequence of an additional *R. etli* strain and partial genomic sequences of six other *R. etli* strains isolated worldwide. Our results support the concept of a pangenomic structure at the multireplicon level and show that a highly conserved symbiotic plasmid is present in divergent *R. etli* isolates.

MATERIALS AND METHODS

Bacterial strains. Seven strains of *R. etli* were chosen for this study (Table 1), and in our analysis we also used the complete genomic sequences of *R. etli* CFN42 (referred to below as RetCFN42) and *R. leguminosarum* biovar viciae 3841 (referred to below as Rlv3841) described previously (23, 58). *R. etli* strains were previously classified by using the following taxonomic criteria: the ability to nodulate and fix nitrogen in common bean plants, the presence of reiterations of the *nifHDK* operon encoding nitrogenase, and demonstration of species-specific growth characteristics. The strains were isolated from distinct locations worldwide and had different plasmid contents (Table 1).

DNA sequencing. Random genomic sequences were obtained from the seven strains by the shotgun method, and this was followed by capillary DNA sequencing using an ABI3730XL automatic DNA sequencing machine (Applied Biosystems, Foster City, CA). We determined the complete genomic sequence for one strain, *R. etli* CIAT652 (referred to below as RetCIAT652), and partial genomic sequences (coverage, about 1×) for the other six strains (Table 1). Assemblies were obtained using Phred-Phrap-Consed software (15, 16, 24). Gaps were filled in by use of appropriate PCR amplification. Partial assemblies were obtained for the low-coverage genomes using only high-quality readings. An *ad hoc* perl script was used to trim at least the first 20 low-quality bases at the 5′ and 3′ ends of each read. Reads with Phred values lower than Q20 were discarded.

Genome size prediction. We used the formula of Fraser and Fleischmann (19) to estimate genome size (L) ($L = -nw/\ln(c/n)$ where n is the number of clones, w is the average read length, and c is the number of contigs). To predict the fraction (percentage) (p) of DNA which was not sequenced and to estimate genome coverage, we employed the formula of Lander and Waterman ($p = e^{-nw/L}$) (33). To evaluate the accuracy of our predictions, we assembled 13,000 random reads for the complete genomes of RetCFN42 and RetCIAT652 and then predicted the genome sizes of the other strains using the formula described above and the partial genomic sequences. We found good agreement between the predictions and the experimentally determined lengths of the complete genomes (Table 1).

Protein families. We performed pairwise comparisons of the whole proteomes of RetCFN42, RetCIAT652, and Rlv3841 using BLASTp with an E value of <1e-7 without filtering for low complexity and enabling the algorithm of Smith and Waterman (47). Orthologs were defined by constructing a similarity matrix using OrthoMCL (35). Similarity matrices were used to run the MCL program (14) to cluster orthologs (and recent paralogs) into families using the following parameters: inflation (I) = 1.5, initial iteration (i) = 1.4, initial inflation (l) = 5, scheme (K) = 7, and centering (c) = 1.2. Proteins that were not grouped into any family were classified as single-member families.

Distribution profiles of single-copy protein families. Genes encoding members of single-member protein families were located in the replicons of RetCFN42 (seven replicons), RetCIAT652 (four replicons), and Rlv3841 (seven replicons). To identify replicons with common gene contents in the three genomes, we constructed a profile for each family using genome localization. The profiles were given the following numbers: 1 for the chromosomes of the three genomes; 2 to 7 for plasmids of RetCFN42 and Rlv3841, starting with the smallest plasmid; and 2 to 5 for plasmids of RetCIAT652, starting with the smallest plasmid. The absence of a gene in any replicon was encoded 0. As expected, linked genes in a replicon had the same profile. For instance, the profile 1-1-1 reflected genes present exclusively in chromosomes, and the profile 1-1-2 meant that the gene was located in the chromosome in two species but in plasmid 2 in the other species. To visualize the profiles, we used the program E-Burst V3 (17, 50).

Annotation and comparative genomics. The RetCIAT652 genome was annotated manually by following the gene model constructed for the previously reported genomic sequence of RetCFN42 (23). Open reading frames were predicted by using GLIMMER 2.0 (10, 44), and annotations were obtained by analysis of BLASTx hits with the nonredundant databases of GenBank and Interpro. To compare partial genomic sequences with the nonredundant database of GenBank, BLASTx searches were performed, and the top hits were classified with respect to organisms with which they matched. Additional comparisons of the complete genomes of RetCFN42, RetCIAT62, and Rlv3841 and the collection of shotgun genomic sequences of strains of *R. etli* were performed using either BLASTn, BLASTx, or Mummer (12). To be considered homologous contigs, genes, or proteins, alignments had to be 30% (60% in the case of contigs) identical over 60% of the length of the largest contig, gene, or protein. Chromosomal islands were defined as contiguous groups of genes that were present in one strain but not in another. Predictions of chromosomal islands were obtained with the aid of the Alien-Hunter program (55). Pangenomic prediction was performed using the power law fit method as described by Tettelin and colleagues (54).

Genome tree construction. *R. etli* complete genomes and the contigs of partial genomes were compared all versus all by BLASTn. Then we used the median percentage of identity for bidirectional best hits between pairs of genomes as a similarity measure, making no distinction between coding and noncoding regions (48). Further, this value was used to estimate the evolutionary distances based on

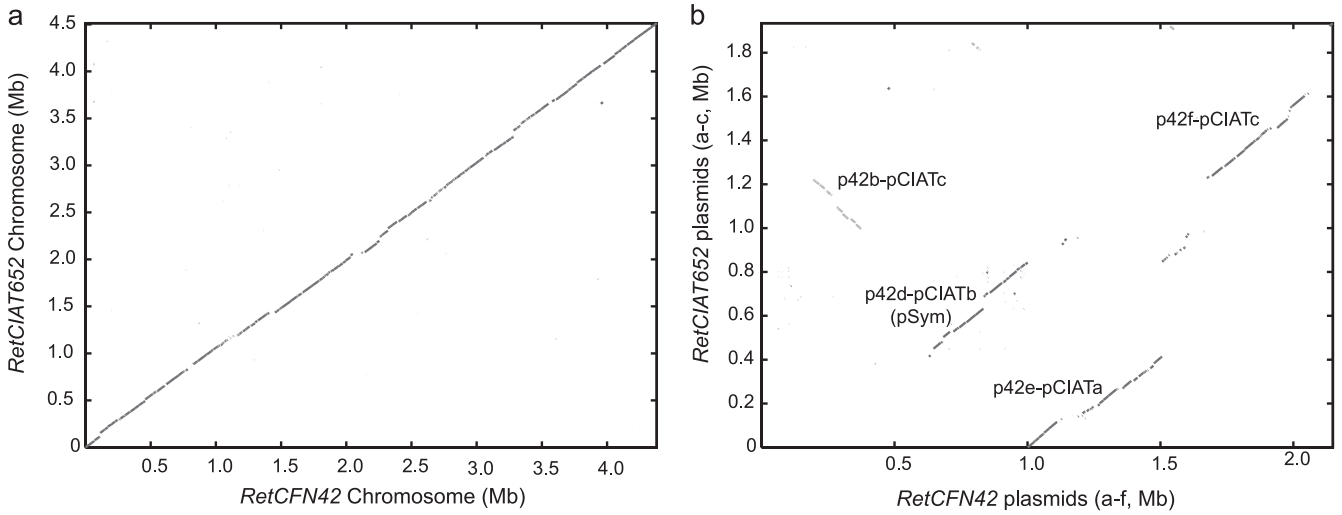

FIG. 1. Synteny relationships between RetCIAT652 and RetCFN42. (a) Nucmer plots of the chromosomes. (b) Nucmer plots of the concatenated plasmids.

the numbers of substitutions per site, to construct a distance matrix, and subsequently to construct an unrooted tree by the neighbor-joining method (57). For estimation of evolutionary distances we used the Poisson distance (d), determined as follows: $d = -\ln\mu$, where μ is $(q - 0.05)/0.95$ and q is the percentage of identity.

Nucleotide sequence accession numbers. The complete sequences of the RetCIAT652 chromosome (accession number NC_010994) and plasmids pCIAT652a (NC_010998), pCIAT652b (NC_010996), and pCIAT652c (NC_010997) have been deposited in the GenBank database. Draft genomes of *R. etli* strains 8C-3 (NZ_ABRA00000000), Brasil5 (NZ_ABQZ00000000), CIAT894 (NZ_ABRD00000000), GR56 (NZ_AABRD00000000), IE4771 (NZ_ABRD00000000), and Kim5 (NZ_ABQY00000000) have also been deposited in the GenBank database.

RESULTS

General features of the genome of *R. etli* CIAT652. *R. etli* strains have diverse genomic architectures highlighted by disparities in plasmid size and number (Table 1). To study the degree of intraspecies genomic similarity and divergence, we obtained the complete genomic sequence of RetCIAT652, a strain isolated in Costa Rica, and the partial sequences of six other *R. etli* strains isolated from various sites worldwide (23). RetCIAT652 had a circular chromosome consisting of about 4,513 kb and three plasmids (designated pCIAT652a, pCIAT652b, and pCIAT652c) consisting of about 414 kb, 429 kb, and 1,091 kb, respectively. The chromosome of RetCIAT652 is 131 kb larger than the chromosome of RetCFN42. In RetCIAT652, most of the genes required for symbiosis were carried on the pCIAT652b plasmid, which is 58 kb larger than the equivalent plasmid of *R. etli* CFN42, p42d. Annotation of the CIAT652 genome yielded 4,072 protein-encoding sequences (CDS) in functional classes and a substantial number (about 2,220) of hypothetical and orphan CDS. Compared with the CFN42 genome, the CIAT652 genome contained 473 more CDS with unknown functions and 215 fewer CDS for which functional annotations were available. Like the CFN42 genome, the CIAT652 genome contained a large number of CDS involved in transport and transcriptional regulation. There were 20 sigma subunits encoded in the CIAT652 genome, compared with the 23 sigma subunits encoded in the CFN42 genome.

Structural correspondence among RetCIAT652 and RetCFN42 replicons. To establish structural correspondence among the replicons of the two *R. etli* strains, the chromosomal and plasmid sequences were aligned using Mummer (11). The chromosomes of both strain showed a straight line of synteny interrupted by several gaps of different sizes but without inversions or any other large rearrangements (Fig. 1). The plasmids were structurally heterogeneous, but some of them seemed to be equivalent. For instance, pCIAT652a had several large segments in common with p42e, as did pCIAT652b with p42d (pSym), and pCIAT652c with p42f and p42b. There were no matches with plasmids p42a and p42c of RetCFN42, indicating that these plasmids were not present in RetCIAT652.

Previous genomic comparisons of RetCFN42 and its close relative Rlv3841 showed that there is extensive chromosomal synteny and, to a lesser degree, synteny between some pairs of plasmids (9). A similar result was obtained when RetCIAT652 and Rlv3841 were compared (data not shown). Despite the divergence of these species, plasmid pCIAT652a showed conservation with pRL11 and plasmid pCIAT652c showed conservation with pRL9 and pRL12. The symbiotic plasmids of *R. etli* (p42d and pCIAT652b) are not related to any replicon in R1v38411, except for 20 common genes required for symbiosis (9). Recently, the complete genome sequences of two strains of *R. leguminosarum* biovar trifoli (RtrWSM1325 and RtrWSM2304) were deposited in the GenBank database; for RtrWSM1325 the accession number for the chromosome was NC_012850 and the accession numbers for the plasmids were NC_12848, NC_12858, NC_12853, NC_12852, and NC_12854, and for RtrWSM2304 the accession number for the chromosome was NC_011369 and the accession numbers for the plasmids were NC_11366, NC_011368, NC_011370, and NC_011371. Strain RtrWSM1325 has 5 plasmids, and strain RtrWSM2304 contains 4 plasmids. We looked for plasmid equivalence between these strains and RetCFN42.

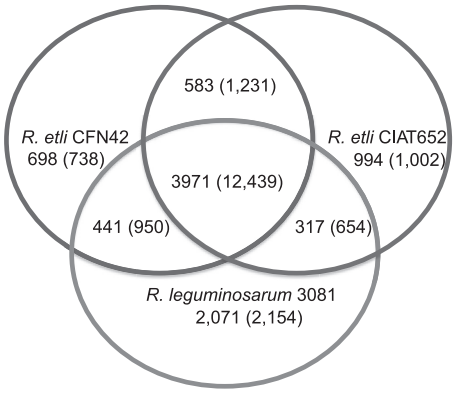

FIG. 2. Common protein families in *R. etli* and Rlv3841. A total of 19,085 predicted proteins encoded by the three genomes were clustered into families using the OrthoMCL and MCL algorithms (14, 35) (see Materials and Methods). The total numbers of families are indicated, and the total numbers of proteins are indicated in parentheses. Both single-member families and paralogous families are included. BLASTx searching of the GenBank nonredundant database for homologs of proteins unique to RetCFN42, RetCIAT652, and Rlv3841 resulted in identification of 170, 350, and 291 orphan genes, respectively.

Nucmer comparisons showed that the plasmids of the RtrWSM1325 and RtrWSM2304 strains have large syntenic regions in common with plasmids p42b, p42c, p42e, and p42f but not with the pSym (p42d) plasmid. This observation suggests that there may be a common plasmid pool for *R. leguminosarum* and *R. etli*.

Core genome of *Rhizobium*. Synteny relationships among the replicons of the two *R. etli* strains and the single Rlv3841 strain indicate that the core genome of *Rhizobium* might be not confined to the chromosome but may extend to some plasmids. To test this possibility, we used a clustering method to group the whole predicted proteomes encoded by the three complete genomes into protein families and then examine the distributions of these families in replicons. We found that a set of 3,971 protein families was encoded in the three genomes; 3,753 of these protein families were protein products of single genes, whereas 218 families corresponded to families with two or more protein homologs encoded in the genomes (Fig. 2). The genes encoding the 3,753 single-protein families were localized in the replicons of the three genomes by constructing presence-absence profiles, as described in Materials and Methods. Genes encoding core protein families were found predominantly in chromosomes but were also present in plasmids common to the three genomes (Fig. 3). There were two main clusters that contained plasmid-encoded proteins. One cluster corresponded to p42f, pCIAT652c, and pRL12 genes encoding 242 proteins and represented 42% of the coding capacity of p42f, 23% of the coding capacity of pCIAT652c, and 30% of the coding capacity of pRL12. The second cluster consisted of 237 proteins encoded by genes in p42e (51% of the total coding capacity), pCIAT652a (59%), and pRL11 (37%). Fur-

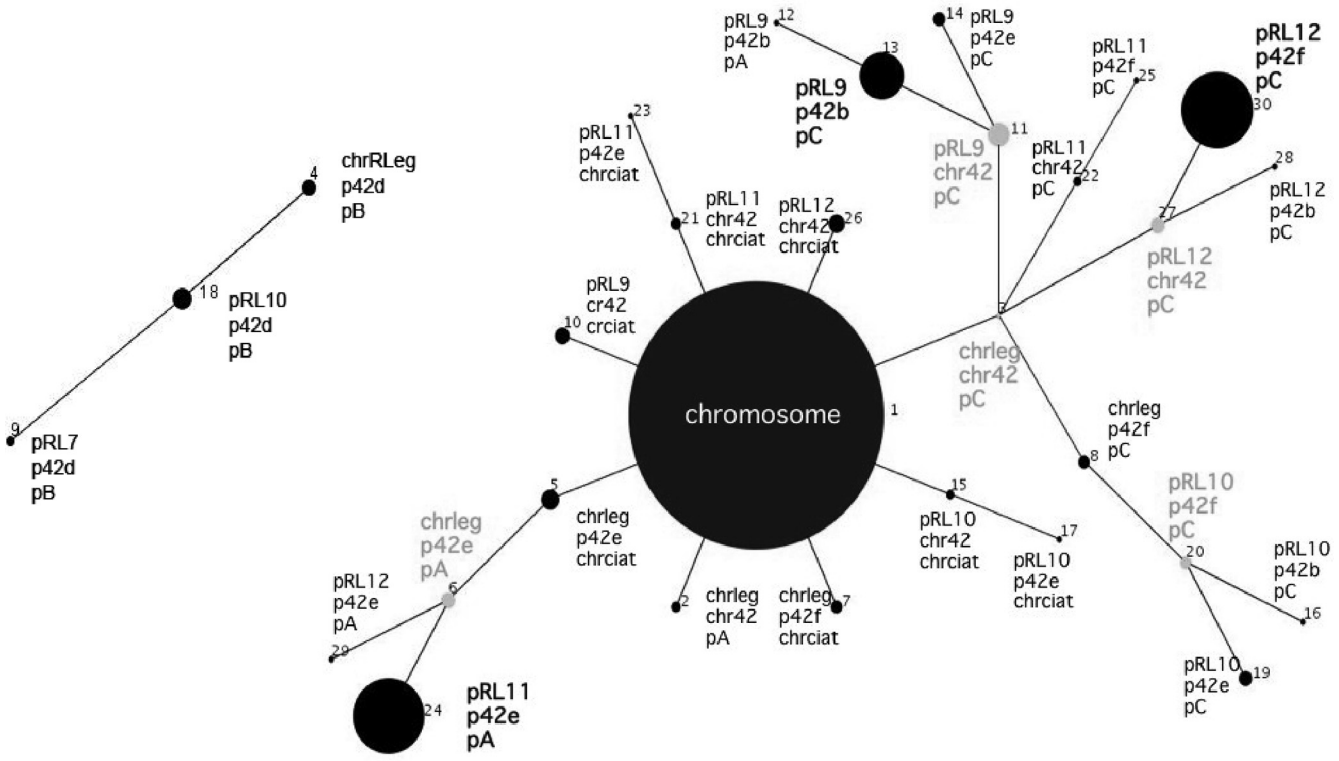

FIG. 3. Core genome profiles and their distribution in *Rhizobium* replicons. Single-member protein families were used to construct presence-absence profiles for replicons of the three genomes. Large circles represent the most common profiles, and small circles represent profiles containing few proteins. Plotting of profile distributions was accomplished by using the e-burst program (17, 50). Profile numbers were assigned arbitrarily.

thermore, most genes encoding these proteins were arranged in syntenic segments common to the three plasmids (data not shown). Another cluster consisted of 91 proteins common to plasmid pCIAT652c, the largest plasmid in the CIAT652 genome, and plasmids p42b and pRL9. Since, as shown above, plasmid pCIAT652c is related p42f and pRL12, this cluster shows that pCIAT652c might represent a chimeric structure that originated by interreplicon recombination. In addition, several other minor profiles appeared in the analysis, indicating that intragenomic recombination also involves small DNA segments (Fig. 3, smallest circles).

Core genome stability is affected by other recombination processes, like gene duplications and gene loss. For construction of profiles we only used single-member proteins; thus, we were unable to observe the effect of gene duplications. However, gene loss was illustrated well by the presence of profiles that included proteins encoded by only two genomes in common clusters. Proteins encoded by only two genomes are subgroups of core replicons (data not shown). For example, proteins represented by the pCIAT642c-pCFN42f, pCIAT652c-pRL12, and pCFN42f-pRL12 profiles, as well as proteins encoded by two chromosomes, fall into this category (data not shown). A particular but important case in this category is the proteins encoded by the symbiotic plasmids that have a unique profile for the two *R. etli* strains and are grouped separate from the symbiotic plasmid pRL10 of Rlv3841 together with *R. etli* plasmid p42c (data not shown) (9). These data suggest that the genes that encode proteins involved in symbiosis are not part of the core genome of *Rhizobium* (Fig. 3). Quite a few of the symbiosis genes that have been identified are common among *Rhizobium* species (9, 22). In RetCIAT652, RetCFN42, and Rlv3841, only 11 genes were maintained in the symbiotic plasmids (cluster 18) (Fig. 3). These genes are *nodA*, *nodB*, *nodC*, *fdxB*, *nodI*, *nodJ*, *fixX*, *fixA*, *fixB*, *fixC*, and *nifN*.

Estimation of the size of the core genome of *Rhizobium* was performed by using the methods of Tettelin et al. (54) and the three complete *Rhizobium* genome sequences available (RetCFN42, RetCIAT652, and Rlv4841 sequences). This estimation yielded 3,220 core genes (with a 99% confidence interval) and predicted that about 99 new genes might be added to the pangenome by every new complete genome sequence (data not shown). Although the analysis was limited by the small number of complete *Rhizobium* genome sequences available, the estimate for the α parameter was 0.6. Since α represents the proportion of new genes discovered as more genomes are sampled, the fact that in this case α is <1 suggests that *Rhizobium* might have an open pangenome.

Accessory genome of *Rhizobium*. The difference between *R. etli* and *R. leguminosarum* can be measured by estimating the numbers of species-specific proteins. Rlv3841 has the highest number of individual protein families (2,071), whereas the two *R. etli* strains were significantly different for this characteristic (Fig. 2). There were 698 protein families of RetCFN42 that were not present in RetCIAT652 and 994 protein families in RetCIAT652 that were not present in RetCFN42. Most members of these families are hypothetical conserved proteins with unknown functions or orphans (23% and 35% for RetCFN42 and RetCIAT652, respectively). In contrast to core genes, which have an average G+C content of 61%, genes unique to each genome had low G+C contents (on average, 57 to 58%).

To further examine genomic differences between *R. etli* strains, we analyzed samples of genome sequences from six *R. etli* strains having distinct geographical origins. The data were obtained by random shotgun sequencing of whole genomes with coverage of about 1× that of the predicted genome length (Table 1). This coverage allowed estimation of the genome lengths for these strains of *R. etli*. The genome size varied over a wide range; strain GR56, an isolate from Spain, had the smallest genome (about 5,000 kb), and RetCFN42 and RetCIAT652 had the largest genomes. Thus, the differences in genome sizes among the strains of *R. etli* examined here were on the order of hundreds of kilobases to 1,500 kb (Table 1).

To determine genomic relationships among *R. etli* strains, BLASTn analysis was used to compare the contigs of each partial genomic sequence with the whole-genome sequences of RetCFN42 and RetCIAT652 and the GenBank nonredundant database. As expected, most sequences of *R. etli* strains were present in the complete genomes (Fig. 4, red and orange bars). These sequences were equivalent to approximately 3,000 kb of common DNA and thus represented less than 50% of the total genome of RetCFN42 or RetCIAT652. In addition, there was about 1,000 kb in each strain for which no homologous sequences were detected in the complete RetCFN42 genome. A smaller, but still substantial, amount of extra DNA was found when similar comparisons were performed using the RetCIAT652 genome as the reference (Fig. 4). A proportion of this extra DNA showed matches to sequences of other organisms deposited in the GenBank database. However, most of the extra DNA did not match any other sequence in the database. A small proportion of the extra DNA was similar to sequences present in at least one other *R. etli* strain (Fig. 4, dark blue and light blue bars).

Chromosomal islands in *R. etli*. When the collection of shotgun genomic sequences was aligned with the sequence of the chromosome of either RetCFN42 or RetCIAT652, the distribution of sequences along the chromosomes was found to be essentially random. Almost every chromosomal region of RetCFN42 and RetCIAT652 contained sequences present in at least one strain. Nevertheless, some chromosomal regions in RetCFN42 and RetCIAT652 contained sequences with no matches with any sequence from either incompletely or completely sequenced *R. etli* strains or Rlv3841 (Fig. 5). A prediction of the Alien Hunter program suggested that many of these regions are chromosomal islands that were acquired by horizontal transfer. Such regions differ from the average with respect to nucleotide composition and codon usage. We analyzed 13 chromosomal islands that were present exclusively in RetCFN42 and 12 chromosomal islands that were unique to RetCIAT652. As Fig. 5 shows, these islands were for the most part not present in the other *R. etli* strains examined and Rlv3841. The chromosomal islands were variable in length (range, about 8 kb to 69 kb), and the largest chromosomal island was found in the RetCFN42 chromosome. The islands were dispersed throughout the chromosomes, and only three locations appeared to be preferentially used for island insertion. Islands 3, 4, and 6 of RetCFN42 had the same locations as islands in the RetCIAT652 chromosome (islands 3, 4, and 5), but the islands differed in gene composition. Island 3, between the $tRNA^{His}$ and $tRNA^{Gln}$ genes in the RetCFN42 chromosome, contains genes already described as the α-*lps*

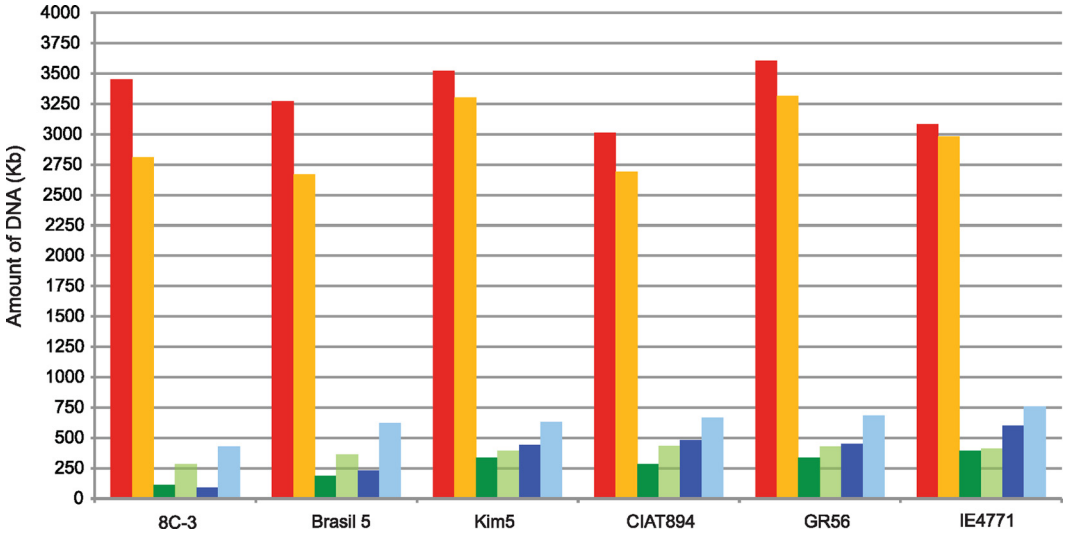

FIG. 4. DNA common to or unique in *R. etli* strains. DNA readings from each partial genome sequence assembled into contigs were compared, using BLASTn, with the complete genomic sequences of RetCIAT652 (red bars) and RetCFN42 (orange bars), using the parameters described in Materials and Methods. Contigs without matches to RetCIAT62 or RetCFN42 sequences were compared with the GenBank nonredundant database (dark green and light green bars, respectively). The dark blue bars indicate the remaining DNA in contigs without matches with either RetCIAT652 or the nonredundant database, and the light blue bars indicate the remaining DNA in contigs without matches with RetCFN42 or the nonredundant database. These contigs contain sequences found only in *R. etli* and orphan genes.

region, which is involved in synthesis, maturation, and transport of the O antigen (8, 13). Although mutations in some genes of this locus affect the symbiotic capabilities of *R. etli* CE3 (a streptomycin-resistant strain derived from RetCFN42), the presence of this island is not widespread in *R. etli*. In contrast, an island at the same position was found in the chromosome of RetCIAT652, and it also seemed to contain genes involved in polysaccharide synthesis and transport; however,

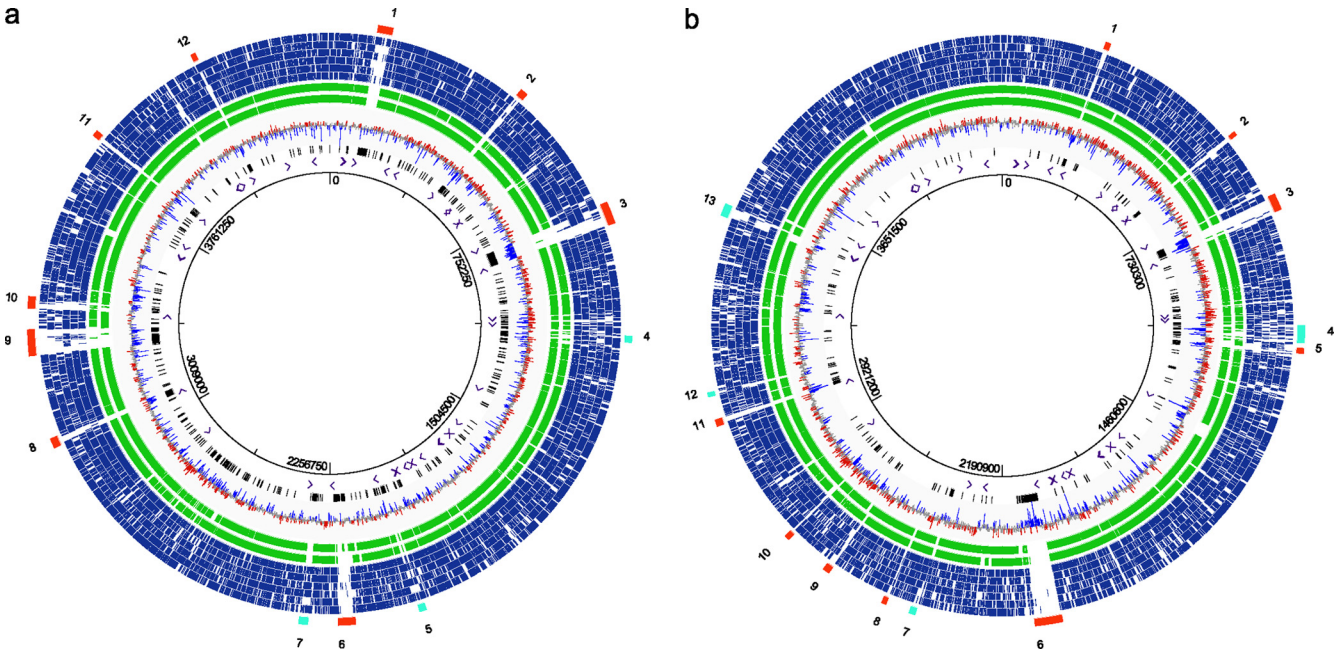

FIG. 5. Chromosomal islands in *R. etli*. Readings for every partial genomic sequence were aligned, using BLASTn, with the chromosomes of RetCIAT652 (a) and RetCFN42 (b). The outer blue circles represent sequences from strains 8C-3, Brasil 5, CIAT894, GR56, IE4771, and Kim5, respectively, from the outside toward the center. The inner green circles represent alignments with the chromosomes of Rlv3841 and RetCFN42 (a) or with the chromosome of RetCIAT652 (b). The next circles represent the G+C content, which was either above average (blue peaks) or below average (red peaks). The two black circles represent genes unique to RetCIAT652 (a) or RetCFN42 chromosomes (b), and the distribution of tRNA genes along the chromosomes is also shown (> and <). The numbers outside the circles indicate putative islands in the chromosomes of RetCIAT652 (a) and RetCFN42 (b), predicted by the Alien Hunter program (55), that agree with regions without any sequence representation (red bars) or heterogeneous sequence representation (turquoise bars) in other *R. etli* strains.

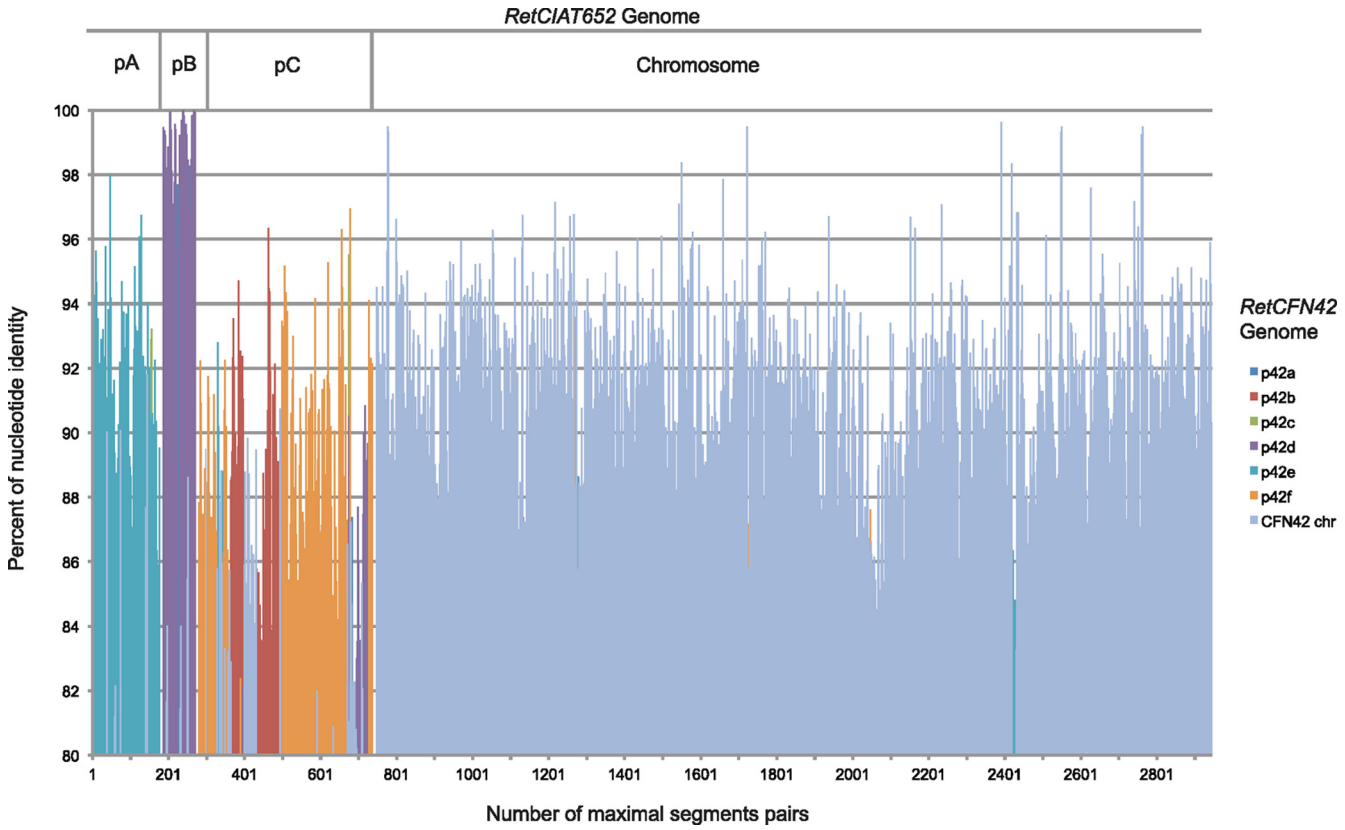

FIG. 6. Nucleotide identities for the replicons of RetCIAT652 and RetCFN42. Local alignments constructed by BLASTn for the two genomes were computed using individual maximal segment pairs (MSPs). All MSPs more than 200 bp long with levels of nucleotide identity of >79% (the lowest value found) were recorded. The average length of MSPs was 2,234 bp. The MSPs for the pCIAT652b-pCFN42d (pSym) pair were the longest MSPs (average, 3,974 bp; longest match, 42,006 bp) and exhibited the highest levels of nucleotide identity (98 to 100%).

none of these genes was homologous to genes of the α-*lps* loci. Island 4 of both chromosomes was highly variable in terms of genetic content, whereas island 6 in RetCFN42 (island 5 in RetCIAT652) was located near the putative terminus of replication, a region of the bacterial chromosome thought to undergo frequent genetic rearrangement. The other islands harbored most genes unique to RetCFN42 or RetCIAT652. These genes are genes related to a variety of enzyme activities, genes with unknown functions, and mobile elements (insertion sequences and genes of phage and plasmid origin). Genomic comparisons of *B. japonicum* strains using macroarrays of the reference strain *B. japonicum* USDA110 showed the presence of 14 genomic islands, some of which were associated with the symbiotic performance of strain USDA110 (29). Thus, the presence of genomic islands in *R. etli* might also be related to some symbiotic capabilities, but this possibility was not addressed here.

DNA conservation among *R. etli* pSym plasmids. We showed that the two complete *R. etli* genome sequences displayed a high degree of conservation at the chromosomal level and that large syntenic segments were present in plasmids. To evaluate the level of DNA divergence between homologous replicons of the two completely sequenced *R. etli* genomes, we compiled all local alignments made by BLASTn. The aligned regions of plasmids pCIAT652a and pCIAT652c, as well as those of the chromosomes, had levels of nucleotide identity of about 85 to 95% (Fig. 6). In marked contrast, sequences that the pSym plasmid of RetCFN42 (pCFN42d) and the pSym plasmid of RetCIAT652 (pCIAT652b) had in common showed the highest levels of nucleotide identity for these genomes (98 to 100%) (Fig. 6). Mummer alignments of the collection of partial genomic sequences with the complete genome sequence of RetCFN42 showed a similar pattern of nucleotide identity in the pSym plasmids, in contrast to the rest of the genomic sequences (Table 2). An exception to this pattern was strain IE4771, an isolate from Puebla, México, which had the most divergent pSym sequences compared with the pSym plasmid sequences of both RetCFN42 and RetCIAT652. Strain IE4771 also had a low proportion of pSym sequences (less than 10%) compared with other strains, which clearly contained at least 50% of the pSym sequence (Table 2). This indicates that the IE4771 strain lost pSym or that a different type of pSym plasmid is present. The latter suggestion seems more plausible because BLASTx comparisons of the partial sequence of strain IE4771 yielded matches with some pSym genes of RetCFN42 or RetCIAT652, including *nifH*, *nifA*, *fixN*, *fixA*, and *fixB*.

When the *R. etli* CIAT652 genome was used as a reference in a comparison of the partial genome sequences, other conservation patterns were observed. Strains 8C-3 and Brasil 5 exhibited a high level of nucleotide conservation compared to RetCIAT652, and there were no appreciable differences in nucleotide identities among the pSym sequences and the rest

TABLE 2. Nucleotide identities of pSym sequences[a]

Strain	CFN42				CIAT652			
	Genome		pSym		Genome		pSym	
	% Identity	% Coverage	% Identity	% Coverage	% Identity	% Coverage	% Identity	% Coverage
8C-3	90.3 ± 3.3	29	97.9 ± 3.7	54	98.1 ± 2.7	37	96.8 ± 3.8	51
Brasil5	90.2 ± 3.3	26	97.1 ± 3.6	50	96.6 ± 2.5	32	97.3 ± 4	55
CIAT894	89.1 ± 2.9	29	96.6 ± 4	44	89.7 ± 3.1	31	97.7 ± 3.5	46
GR56	90.2 ± 3.1	35	97.2 ± 3.9	58	92.5 ± 2.9	40	96.4 ± 4	61
IE4771	89 ± 3.1	33	89.2 ± 4.5	10	92 ± 2.9	37	89.3 ± 4.4	9
Kim5	89.9 ± 3.1	36	96.7 ± 4.4	54	92 ± 2.9	40	92.6 ± 4	52

[a] The values are averages ± standard deviations. The genome data are the results for the chromosome and plasmids of the strain without the sequences that match the pSym sequences of CFN42 or CIAT652.

of their genomes. Two strains (IE4771 and Kim5) had the lowest levels of nucleotide identity in the genome as a whole, and only strains CIAT894 and GR56 had pSym sequences that were more conserved than the chromosome (Table 2).

To evaluate if there is a relationship between the overall genomic divergence in *R. etli* strains and the conservation of pSym, we constructed a distance matrix tree based on BLASTn comparisons performed in an all-versus-all manner, using both the two complete genomes and the partial genomic sequences of *R. etli* and including Rlv3841 as the outgroup (Fig. 7). This tree showed that RetCIAT652, 8C-3, and Brasil 5 are very closely related but the rest of the strains have diverged to different degrees. In particular, RetCFN42 appeared to have no close relatives. Based on all of these observations, we concluded that pSym sequences are highly conserved and dispersed in variable genomic backgrounds.

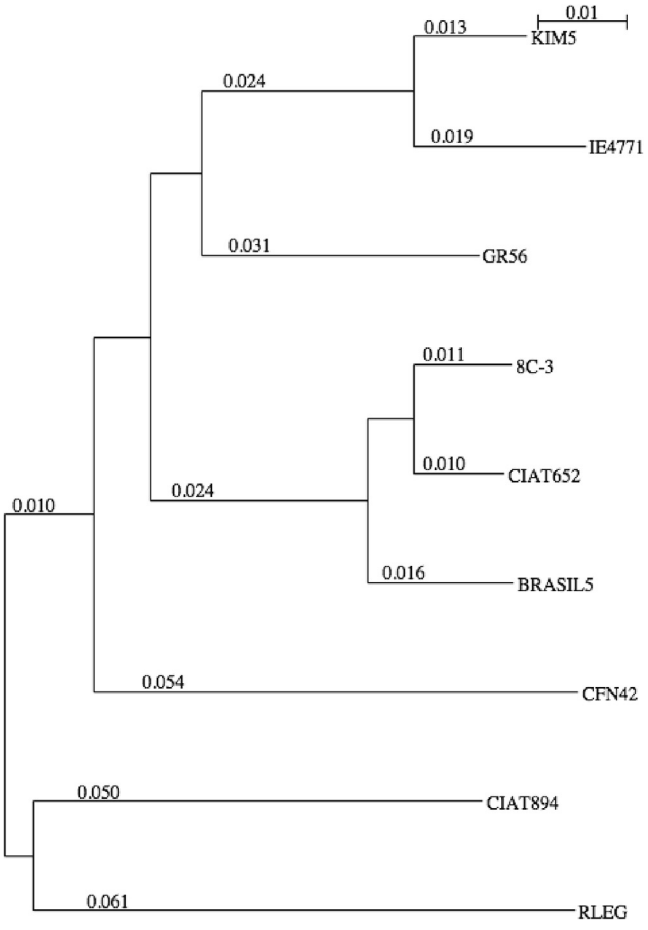

FIG. 7. Genetic relatedness of *R. etli* strains. A distance matrix derived from all-versus-all BLASTn alignments was used to estimate the degrees of divergence among *R. etli* strains and R1v3841. Next, an unrooted tree was constructed using the neighbor-joining method of the PHYLIP program (see Materials and Methods). The numbers at the branch points indicate the numbers of substitutions per site.

DISCUSSION

Previous theories of bacterial evolution emphasized that bacteria have evolved a "strategy to expand the effective genome size of the species without imposing on each individual the burden of reproducing the entire genome" (7). Campbell and other researchers hypothesized that a bacterial genome is composed of an "euchromosome" and "accessory elements" (7, 43). The terms have been changed in the modern genomic era and are now the "core genome" and "pangenome" (3, 37, 52). The core genome is defined as the set of genes shared by all members of a monophyletic group (1). In contrast, the pangenome is an expanded version of the repertoire of genes found in a species, and accessory genes are not present in all individual bacteria. Genome sequences of a number of strains of species such as *S. agalactiae*, *E. coli*, *Haemophilus influenzae*, and *Streptococcus pneumoniae* have provided support for this concept (27, 28, 42, 52).

We have previously sequenced the complete genome of RetCFN42, the reference strain. Here we approached an understanding of the pangenomic structure of *R. etli* by sequencing another complete genome and by low-coverage sequencing of six other genomes of strains of *R. etli*. We found a set of genes that might represent the core genome of *Rhizobium* by comparing sequences of the two complete *R. etli* genomes and the sequence of the close relative Rlv3841. Furthermore, we found that in *R. etli* (and Rlv3841) the core genome is not limited to the chromosome but also extends to some large plasmids. It is known that members of the rhizobia have multipartitioned genomes composed of the chromosome and a variable number of plasmids (31, 49). Genes of the core genome are commonly carried on the chromosome and are maintained in syntenic blocks in closely related species (25, 51). In

contrast, neither symbiotic plasmids nor other plasmids are conserved, except for the presence of a few common genes (21, 23). We found here that plasmids p42f, pCIAT652c, and pRL12 are highly related in terms of gene content, as are plasmids p42e, pCIAT652a, and pRL11 and might be considered part of the core genome. The large plasmids and the linear chromosome of *Agrobacterium*, chromosome II of *Brucella*, and the pSymB plasmid of *Sinorhizobium meliloti* might also be viewed as secondary chromosomes (46). The sizes of the replicons, the G+C content, and the presence of certain classes of genes normally found in primary chromosomes make this possibility plausible. Recently, Slater and colleagues suggested that secondary chromosomes originated via intragenomic gene transfers from primary chromosomes to an ancestral *repABC* replicon (46). Evidence for this hypothesis comes from the conservation of several syntenic blocks of genes, such as *minCDE* (cell division proteins), *hutIGU* (histidine biosynthesis), and *pcaGHID* (protocatechuate biosynthesis), in secondary chromosomes and plasmids across members of the rhizobia (46). These three blocks of genes are located in the p42e-pCIAT652a-pRL11 cluster that is included in the set of genes encoding core proteins. Our comparison in the present work revealed that a substantial proportion of about 479 single-member protein families are encoded in the largest plasmids common to the genomes of *R. etli* and Rlv3841. Furthermore, most genes encoding these proteins were arranged in common syntenic segments (data not shown).

We found that a substantial proportion of DNA in the newly partially sequenced strains of *R. etli* was not present in the model strain RetCFN42 and in RetCIAT652, whose new complete genome sequence is reported here. There were 738 and 1,002 different protein-encoding genes in RetCFN42 and RetCIAT652, respectively, for which complete genome sequences are available. Similar amounts of extra DNA were detected when partial genomic sequences of various *R. etli* strains were compared to sequences of the complete genomes of RetCFN42 and RetCIAT652. The extra DNA represents the accessory component of the *R. etli* pangenome. This DNA has a low G+C content and contains numerous hypothetical genes and mobile elements that are also common in the accessory components of other bacterial species. Strain-specific chromosomal islands, which were shown here to be present in the chromosomes of RetCFN42 and RetCIAT652, are some of the locations of such extra DNA. The plasmid pool of *R. etli* is variable and also contributes importantly to the extra DNA, but it is not known to what extent.

A striking result of the present work was the high level of nucleotide identity in homologous segments of pSym of *R. etli*, in contrast to the more divergent sequences seen in the rest of the genome. In these segments there are 210 very conserved CDS (98 to 100% nucleotide identity), which represent 60% of the coding capacity of the pSym plasmid of RetCFN42, including known symbiosis genes (*nif*, *nod*, *fix*, *fdx*), as well as other genes not involved in symbiosis (*vir*, *tra*) and hypothetical genes. Only strain IE4771 displayed a low level of nucleotide identity and a low level of coverage compared with pSym sequences of RetCFN42 and RetCIAT652. According to 16S RNA gene data and nodulation tests, strain IE4771 belongs to an *R. etli* group that is distinct because it has two copies, not three copies, of *nifH* (three copies are usual in the more common *R. etli* strains) (45). These data suggest that at least two symbiotic plasmids may exist in the *R. etli* population. One plasmid would be highly conserved, often found in *R. etli* isolates, and prototypically defined by three *nifH* reiterations. This plasmid is exemplified by the pSym plasmids of RetCFN42 and RetCIAT652. The other, more divergent pSym plasmid, which is present in isolates from Puebla, México, has not been characterized at the genomic level yet. A recent origin of pSym is the simplest explanation for the high level of conservation of pSym in very divergent *R. etli* isolates. Alternatively, nodulation performance might provide strong selection pressure, selecting against any variation in pSym. The latter hypothesis seems improbable, as pSym genes with roles in nodulation and genes having unknown functions both have identical nucleotide sequences. In previous work, we analyzed the patterns of single-nucleotide polymorphisms in DNA sequences of the pSym plasmids of several *R. etli* strains (some of which were included in this study) (18). The data indicate that most of the nucleotide substitutions are spread over the population by recombination and that the contribution of mutations to polymorphism is relatively low. In agreement with this model, very few nucleotide variations were found in the pSym sequences compared here.

Several years ago Palacios and colleagues (38) asked how many genotypes would be capable of conferring the *R. etli* phenotype. Our data indicate that a unique pSym genotype (or perhaps very few pSym genotypes) might be responsible for the ability of *R. etli* to nodulate the common bean. Other comparative genomics studies using microarrays have shown that the symbiotic plasmid pSymA is the most variable replicon in strains of *S. meliloti* (20, 26). This result contrasts with the pSym conservation in the *R. etli* strains compared here. Since it was demonstrated that the ability to nodulate is encoded by plasmids (30), it has been common to call these plasmids "symbiotic." New genome sequencing technologies and comparative genomics have revealed that a variety of mechanisms have led to symbiosis with legumes (36) and that pSym plasmids in rhizobial species are not comparable despite the fact that they have some common *nod* and *nif* genes. It should be emphasized that as genome sequence technology is becoming more accessible, it is now feasible to analyze many more *R. etli* genomes to understand diversification and evolution in *R. etli* pSym plasmids and the genome of the species as a whole. Future work should also help determine more accurately the sizes of the core and accessory genomes, as well as the size of the plasmid pool of *R. etli*. Lastly, a clear picture of the evolutionary relationships among the different genome components of *Rhizobium* should emerge from studies performed with this kind of approach.

ACKNOWLEDGMENTS

We thank José Espíritu and Víctor del Moral for help with computational resources and Miguel A. Cevallos for critical reading of the manuscript. We also thank the anonymous reviewers for their valuable suggestions.

This work was supported by grants from CONACyT (grant U4633) and PAPIIT-UNAM (grants IN215908 and IN223005).

V.G. was responsible for the experimental design and manuscript preparation; J.L.A. was responsible for the comparative genomic analysis; R.I.S. and P.B. were responsible for genome sequencing and annotation; I.L.H.G. was responsible for bioinformatic analysis; J.L.F.

and R.D. were responsible for genome sequencing; M.F., R.P., and G.D. were responsible for discussion of the data and provision of supporting materials; and J.M. was responsible for genome sequencing and participated in discussions. All authors read and approved the final manuscript.

REFERENCES

1. **Abby, S., and V. Daubin.** 2007. Comparative genomics and the evolution of prokaryotes. Trends Microbiol. **15:**135–141.
2. **Andersson, S. G., A. Zomorodipour, J. O. Andersson, T. Sicheritz-Ponten, U. C. Alsmark, R. M. Podowski, A. K. Naslund, A. S. Eriksson, H. H. Winkler, and C. G. Kurland.** 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature **396:**133–140.
3. **Bentley, S.** 2009. Sequencing the species pan-genome. Nat. Rev. Microbiol. **7:**258–259.
4. **Bentley, S. D., K. F. Chater, A. M. Cerdeno-Tarraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C. H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M. A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill, and D. A. Hopwood.** 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature **417:**141–147.
5. **Bergthorsson, U., and H. Ochman.** 1995. Heterogeneity of genome sizes among natural isolates of *Escherichia coli*. J. Bacteriol. **177:**5784–5789.
6. **Buell, C. R., V. Joardar, M. Lindeberg, J. Selengut, I. T. Paulsen, M. L. Gwinn, R. J. Dodson, R. T. Deboy, A. S. Durkin, J. F. Kolonay, R. Madupu, S. Daugherty, L. Brinkac, M. J. Beanan, D. H. Haft, W. C. Nelson, T. Davidsen, N. Zafar, L. Zhou, J. Liu, Q. Yuan, H. Khouri, N. Fedorova, B. Tran, D. Russell, K. Berry, T. Utterback, S. E. Van Aken, T. V. Feldblyum, M. D'Ascenzo, W. L. Deng, A. R. Ramos, J. R. Alfano, S. Cartinhour, A. K. Chatterjee, T. P. Delaney, S. G. Lazarowitz, G. B. Martin, D. J. Schneider, X. Tang, C. L. Bender, O. White, C. M. Fraser, and A. Collmer.** 2003. The complete genome sequence of *Arabidopsis* and the tomato pathogen *Pseudomonas syringae* pv. tomato DC3000. Proc. Natl. Acad. Sci. U. S. A. **100:**10181–10186.
7. **Campbell, A.** 1981. Evolutionary significance of accessory DNA elements in bacteria. Annu. Rev. Microbiol. **35:**55–83.
8. **Carlson, R. W., B. Reuhs, T. B. Chen, U. R. Bhat, and K. D. Noel.** 1995. Lipopolysaccharide core structures in *Rhizobium etli* and mutants deficient in O-antigen. J. Biol. Chem. **270:**11783–11788.
9. **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.
10. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641.
11. **Delcher, A. L., A. Phillippy, J. Carlton, and S. L. Salzberg.** 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. **30:**2478–2483.
12. **Delcher, A. L., S. L. Salzberg, and A. M. Phillippy.** 2003. Using MUMmer to identify similar regions in large sequence sets. Chapter 10, unit 10.3. Curr. Protoc. Bioinformatics **2003:**10.3.1–10.3.18. doi:10.1002/0471250953.bi1003s00.
13. **Duelli, D. M., A. Tobin, J. M. Box, V. S. Kolli, R. W. Carlson, and K. D. Noel.** 2001. Genetic locus required for antigenic maturation of *Rhizobium etli* CE3 lipopolysaccharide. J. Bacteriol. **183:**6054–6064.
14. **Enright, A. J., S. Van Dongen, and C. A. Ouzounis.** 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. **30:**1575–1584.
15. **Ewing, B., and P. Green.** 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:**186–194.
16. **Ewing, B., L. Hillier, M. C. Wendl, and P. Green.** 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:**175–185.
17. **Feil, E. J., B. C. Li, D. M. Aanensen, W. P. Hanage, and B. G. Spratt.** 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol. **186:**1518–1530.
18. **Flores, M., L. Morales, A. Avila, V. González, P. Bustos, D. García, Y. Mora, X. Guo, J. Collado-Vides, D. Piñero, G. Dávila, J. Mora, and R. Palacios.** 2005. Diversification of DNA sequences in the symbiotic genome of *Rhizobium etli*. J. Bacteriol. **187:**7185–7192.
19. **Fraser, C. M., and R. D. Fleischmann.** 1997. Strategies for whole microbial genome sequencing and analysis. Electrophoresis **18:**1207–1216.
20. **Giuntini, E., A. Mengoni, C. De Filippo, D. Cavalieri, N. Aubin-Horth, C. R. Landry, A. Becker, and M. Bazzicalupo.** 2005. Large-scale genetic variation of the symbiosis-required megaplasmid pSymA revealed by comparative genomic analysis of *Sinorhizobium meliloti* natural strains. BMC Genomics **6:**158.
21. **González, V., P. Bustos, M. A. Ramírez-Romero, A. Medrano-Soto, H. Salgado, I. Hernández-González, J. C. Hernández-Celis, V. Quintero, G. Moreno-Hagelsieb, L. Girard, O. Rodríguez, M. Flores, M. A. Cevallos, J. Collado-Vides, D. Romero, and G. Dávila.** 2003. The mosaic structure of the symbiotic plasmid of *Rhizobium etli* CFN42 and its relation to other symbiotic genome compartments. Genome Biol. **4:**R36.
22. **González, V., L. Lozano, S. Castillo-Ramírez, I. Hernández-González, P. Bustos, R. I. Santamaría, J. L. Fernández, J. L. Acosta, and G. Dávila.** 2008. Evolutionary genomics of the nitrogen-fixing symbiotic bacteria, p. 183–198. *In* P. Dion and C. S. Nautiyal (ed.), Molecular mechanisms of plant and microbe coexistence, vol. 15. Springer, Heidelberg, Germany.
23. **González, V., R. I. Santamaría, P. Bustos, I. Hernández-González, A. Medrano-Soto, G. Moreno-Hagelsieb, S. C. Janga, M. A. Ramírez, V. Jimenez-Jacinto, J. Collado-Vides, and G. Dávila.** 2006. The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. Proc. Natl. Acad. Sci. U. S. A. **103:**3834–3839.
24. **Gordon, D., C. Abajian, and P. Green.** 1998. Consed: a graphical tool for sequence finishing. Genome Res. **8:**195–202.
25. **Guerrero, G., H. Peralta, A. Aguilar, R. Díaz, M. A. Villalobos, A. Medrano-Soto, and J. Mora.** 2005. Evolutionary, structural and functional relationships revealed by comparative analysis of syntenic genes in Rhizobiales. BMC Evol. Biol. **5:**55.
26. **Guo, H., S. Sun, B. Eardly, T. Finan, and J. Xu.** 2009. Genome variation in the symbiotic nitrogen-fixing bacterium *Sinorhizobium meliloti*. Genome **52:**862–875.
27. **Hiller, N. L., B. Janto, J. S. Hogg, R. Boissy, S. Yu, E. Powell, R. Keefe, N. E. Ehrlich, K. Shen, J. Hayes, K. Barbadora, W. Klimke, D. Dernovoy, T. Tatusova, J. Parkhill, S. D. Bentley, J. C. Post, G. D. Ehrlich, and F. Z. Hu.** 2007. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. J. Bacteriol. **189:**8186–8195.
28. **Hogg, J. S., F. Z. Hu, B. Janto, R. Boissy, J. Hayes, R. Keefe, J. C. Post, and G. D. Ehrlich.** 2007. Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol. **8:**R103.
29. **Itakura, M., K. Saeki, H. Omori, T. Yokoyama, T. Kaneko, S. Tabata, T. Ohwada, S. Tajima, T. Uchiumi, K. Honnma, K. Fujita, H. Iwata, Y. Saeki, Y. Hara, S. Ikeda, S. Eda, H. Mitsui, and K. Minamisawa.** 2009. Genomic comparison of *Bradyrhizobium japonicum* strains with different symbiotic nitrogen-fixing capabilities and other Bradyrhizobiaceae members. ISME J. **3:**326–339.
30. **Johnston, A. W. B., J. L. Beynon, A. V. Buchanan-Wollaston, S. M. Setchell, P. R. Hirsh, and J. E. Beringer.** 1978. High frequency transfer of nodulating ability between strains and species of *Rhizobium*. Nature **276:**634–636.
31. **Jumas-Bilak, E., S. Michaux-Charachon, G. Bourg, M. Ramuz, and A. Allardet-Servent.** 1998. Unconventional genomic organization in the alpha subgroup of the Proteobacteria. J. Bacteriol. **180:**2749–2755.
32. **Konstantinidis, K. T., and J. M. Tiedje.** 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc. Natl. Acad. Sci. U. S. A. **101:**3160–3165.
33. **Lander, E. S., and M. S. Waterman.** 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics **2:**231–239.
34. **Legault, B. A., A. López-López, J. C. Alba-Casado, W. F. Doolittle, H. Bolhuis, F. Rodríguez-Valera, and R. T. Papke.** 2006. Environmental genomics of "*Haloquadratum walsbyi*" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. BMC Genomics **7:**171.
35. **Li, L., C. J. Stoeckert, Jr., and D. S. Roos.** 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **13:**2178–2189.
36. **Masson-Boivin, C., E. Giraud, X. Perret, and J. Batut.** 2009. Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes? Trends Microbiol. **17:**458–466.
37. **Medini, D., C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli.** 2005. The microbial pan-genome. Curr. Opin. Genet. Dev. **15:**589–594.
38. **Palacios, R., M. Flores, S. Brom, E. Martínez, V. González, S. Frenk, C. Quinto, M. A. Cevallos, L. Girard, D. Romero, A. Garciarrubio, D. Piñero, and G. Dávila.** 1986. Organization of the *Rhizobium phaseoli* genome, p. 151–156. *In* D. P. S. Verma and N. Brisson (ed.), Molecular genetics of plant-microbe interactions. Martinus Nijhoff Publishers Organization, Dordrecht, the Netherlands.
39. **Pallen, M. J., and B. W. Wren.** 2007. Bacterial pathogenomics. Nature **449:**835–842.
40. **Paulsen, I. T., R. Seshadri, K. E. Nelson, J. A. Eisen, J. F. Heidelberg, T. D. Read, R. J. Dodson, L. Umayam, L. M. Brinkac, M. J. Beanan, S. C. Daugherty, R. T. Deboy, A. S. Durkin, J. F. Kolonay, R. Madupu, W. C. Nelson, B. Ayodeji, M. Kraul, J. Shetty, J. Malek, S. E. Van Aken, S. Riedmuller, H. Tettelin, S. R. Gill, O. White, S. L. Salzberg, D. L. Hoover, L. E. Lindler, S. M. Halling, S. M. Boyle, and C. M. Fraser.** 2002. The *Brucella suis* genome reveals fundamental similarities between animal and

plant pathogens and symbionts. Proc. Natl. Acad. Sci. U. S. A. **99:**13148–13153.

41. **Piñero, D., E. Martinez, and R. K. Selander.** 1988. Genetic diversity and relationships among isolates of *Rhizobium leguminosarum* biovar *phaseoli*. Appl. Environ. Microbiol. **54:**2825–2832.
42. **Rasko, D. A., M. J. Rosovitz, G. S. Myers, E. F. Mongodin, W. F. Fricke, P. Gajer, J. Crabtree, M. Sebaihia, N. R. Thomson, R. Chaudhuri, I. R. Henderson, V. Sperandio, and J. Ravel.** 2008. The pangenome structure of *Escherichia coli:* comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J. Bacteriol. **190:**6881–6893.
43. **Reanney, D.** 1976. Extrachromosomal elements as possible agents of adaptation and development. Bacteriol. Rev. **40:**552–590.
44. **Salzberg, S. L., A. L. Delcher, S. Kasif, and O. White.** 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. **26:**544–548.

44a. **Shinegobu, S., H. Watanabe, M. Hattori, Y. Sakaki, and H. Ishikawa.** 2000. Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature **407:**81–86.

45. **Silva, C., P. Vinuesa, L. E. Eguiarte, V. Souza, and E. Martinez-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.
46. **Slater, S. C., B. S. Goldman, B. Goodner, J. C. Setubal, S. K. Farrand, E. W. Nester, T. J. Burr, L. Banta, A. W. Dickerman, I. Paulsen, L. Otten, G. Suen, R. Welch, N. F. Almeida, F. Arnold, O. T. Burton, Z. Du, A. Ewing, E. Godsy, S. Heisel, K. L. Houmiel, J. Jhaveri, J. Lu, N. M. Miller, S. Norton, Q. Chen, W. Phoolcharoen, V. Ohlin, D. Ondrusek, N. Pride, S. L. Stricklin, J. Sun, C. Wheeler, L. Wilson, H. Zhu, and D. W. Wood.** 2009. Genome sequences of three *Agrobacterium* biovars help elucidate the evolution of multichromosome genomes in bacteria. J. Bacteriol. **191:**2501–2511.
47. **Smith, T. F., and M. S. Waterman.** 1981. Identification of common molecular subsequences. J. Mol. Biol. **147:**195–197.
48. **Snel, B., M. A. Huynen, and B. E. Dutilh.** 2005. Genome trees and the nature of genome evolution. Annu. Rev. Microbiol. **59:**191–209.
49. **Sobral, B. W., R. J. Honeycutt, and A. G. Atherly.** 1991. The genomes of the family Rhizobiaceae: size, stability, and rarely cutting restriction endonucleases. J. Bacteriol. **173:**704–709.
50. **Spratt, B. G., W. P. Hanage, B. Li, D. M. Aanensen, and E. J. Feil.** 2004. Displaying the relatedness among isolates of bacterial species—the eBURST approach. FEMS Microbiol. Lett. **241:**129–134.
51. **Tamames, J.** 2001. Evolution of gene order conservation in prokaryotes. Genome Biol. **2:**RESEARCH0020.
52. **Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser.** 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae:* implications for the microbial "pan-genome." Proc. Natl. Acad. Sci. U. S. A. **102:**13950–13955.
53. **Tettelin, H., V. Masignani, M. J. Cieslewicz, J. A. Eisen, S. Peterson, M. R. Wessels, I. T. Paulsen, K. E. Nelson, I. Margarit, T. D. Read, L. C. Madoff, A. M. Wolf, M. J. Beanan, L. M. Brinkac, S. C. Daugherty, R. T. DeBoy, A. S. Durkin, J. F. Kolonay, R. Madupu, M. R. Lewis, D. Radune, N. B. Fedorova, D. Scanlan, H. Khouri, S. Mulligan, H. A. Carty, R. T. Cline, S. E. Van Aken, J. Gill, M. Scarselli, M. Mora, E. T. Iacobini, C. Brettoni, G. Galli, M. Mariani, F. Vegni, D. Maione, D. Rinaudo, R. Rappuoli, J. L. Telford, D. L. Kasper, G. Grandi, and C. M. Fraser.** 2002. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc. Natl. Acad. Sci. U. S. A. **99:**12391–12396.
54. **Tettelin, H., D. Riley, C. Cattuto, and D. Medini.** 2008. Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol. **11:**472–477.
55. **Vernikos, G. S., and J. Parkhill.** 2006. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics **22:**2196–2203.
56. **Welch, R. A., V. Burland, G. Plunkett III, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S. R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. Mobley, M. S. Donnenberg, and F. R. Blattner.** 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. **99:**17020–17024.
57. **Wolf, Y. I., I. B. Rogozin, N. V. Grishin, R. L. Tatusov, and E. V. Koonin.** 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol. Biol. **1:**8.
58. **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.