UNAM

INSTITUTO DE BIOTECNOLOGÍA

POSGRADO EN CIENCIAS BIOQUÍMICAS

Análisis de la congruencia evolutiva de los genes ortólogos de *Rhizobium etli* CFN42

TESIS

QUE PARA OBTENER EL GRADO DE ACADÉMICO DE

DOCTOR EN CIENCIAS BIOQUÍMICAS

PRESENTA

SANTIAGO CASTILLO RAMÍREZ

DIRECTOR DE TESIS: DR. VÍCTOR GONZÁLEZ ZÚÑIGA

ÍNDICE

NOTA PRELIMINAR

A lo largo de mi doctorado participé en varios proyectos pero esta tesis sólo presenta aquel al cual dedique más tiempo y por el cual fui evaluado. A continuación mencionaré de manera muy escueta los demás proyectos. Una vez que el Dr. Xian Wo secuenció el cloroplasto de *Phaseolus vulgaris* cv Negro Jamapa, comparé los genomas de los cloroplastos de la familia *Fabacea* por medio de filogenias, tanto a nivel de todo el genoma como de los genes individuales, y 'relative rate tests' para establecer si algunos genes y/o especies tenían tasas de evolución aceleradas. Esto fue publicado en BMC Genomics en 2007 (ver anexo artículos). En otro proyecto, en el cual participó más gente del Programa de Genómica Evolutiva y gente del Sanger Institute (Cambrige, Inglaterra), se compararon los genomas de *Rhizobium etli* CFN42 y *Rhizobiun leguminosarum* 3841. Para ese proyecto definí grupos de homólogos y de posibles ortólogos, asigné categorías funcionales, establecí las posibles equivalencias entre los plásmidos de estas dos especies, y determiné las tasas de substitución sinónimas y no sinónimas para los grupos de homólogos. Dicho trabajo fue publicado el año pasado en Plos ONE (ver anexo artículos). Durante mi proceso de formación realicé dos estancias de investigación en el laboratorio del Dr. Scott V. Edwards, del departamento de biología evolutiva del la Universidad de Harvard. Como resultado de éstas dos estancias actualmente estoy escribiendo junto con Scott V. Edwards, Dennis Pearl y Liu Liang el capítulo de un libro que trata del proceso de inferencia del árbol de las especies. En la parte final de mi doctorado, en colaboración con el Dr. Miguel Angel Cevallos, realicé un estudio evolutivo con el fin de ver si los genes del operon *repABC* (que es el sistema de partición y segregación de muchos replicones secundarios de las *Alfaproteobacterias*) presentaban una historia común y restricciones funcionales similares. Este estudio está actualmente sometido en la revista *BMC genomics* (anexo artículos).

RESUMEN

Los genes ortólogos deberían ser congruentes entre ellos mismos y reflejar la historia de las especies. Este proyecto determinó si lo anterior realmente ocurre y para ello se escogió el orden *Rhizobiales*, un grupo de bacterias con distancias filogenéticas moderadas. La mayoría de los genes ortólogos no reflejan exactamente la historia de las especies e inesperadamente la topología más común no fue la de la del árbol de la especies. Aunque los genes ortólogos no reflejan exactamente la historia de las especies, las topologías coinciden, en promedio, 70% con el árbol de las especies. Uno de los factores que afecta la concordancia entre el árbol de las especies y los genes ortólogos es el error de muestro; sin embargo, éste no afecta de manera uniforme a los genes ortólogos de las diferentes categorías funcionales. La separación incompleta de linajes génicos es otro factor que ha afectado a los genes ortólogos. A pesar de la amplia variedad de topologías, la restricción funcional organiza a los genes ortólogos en unos cuantos grupos. La mayoría de los genes ortólogos son más propensos a desempeñar funciones relacionadas con las categorías de "Almacenamiento y procesamiento de la información" o "Procesos celulares y señalización"; además, están más conservados y son más refractarios a las causas de discordancia. Por otro lado, los genes ortólogos pertenecientes a la categoría "Pobremente caracterizados" fueron los menos abundantes, tuvieron los mayores grados de divergencia, y se vieron más afectados por las causas de discordancia.

ABSTRACT

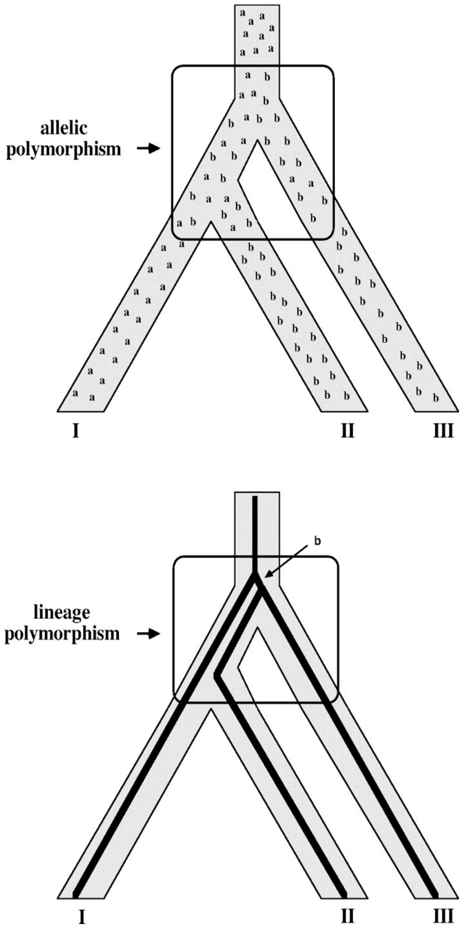
As originally defined, orthologous genes implied a reflection of the history of the species. In recent years, many studies have examined the concordance between orthologous gene trees and species trees in bacteria. These studies have produced contradictory results that may have been influenced by orthologous gene misidentification and artefactual phylogenetic reconstructions. Here, we address the question of whether bacterial orthologous genes really reflect the history of the species, using a method that allows the exclusion of false positives during identification of orthologous genes. We identified a set of 370 orthologous genes from the bacterial order *Rhizobiales*. Although manifesting strong vertical signal, almost every orthologous gene had a distinct phylogeny, and the most common topology among the orthologous gene trees did not correspond with the best estimate of the species tree. However, each orthologous gene tree shared an average of 70% of its bipartitions with the best estimate of the species tree. Stochastic error related to gene size affected the concordance between the best estimated species tree and the orthologous gene trees, although this effect was weak and distributed unevenly among the functional categories. The nodes showing the greatest discordance were those defined by the shortest internal branches in the best estimated of the species tree. Moreover, a clear bias was evident with respect to the function of the orthologous genes, and the degree of divergence among the orthologous genes appeared to be related to their functional classification. Orthologous genes do not reflect the history of the species when taken as individual markers, but they do when taken as a whole. Stochastic error affected the concordance of orthologous genes with the species tree, albeit weakly. We conclude that two important biological causes of discordance among orthologous genes are incomplete lineage sorting and functional restriction.

INTRODUCIÓN Y ANTECEDENTES

Fitch acuñó el término de genes ortólogos en 1970 refiriéndose a todos aquellos genes cuyas filogenias reflejan la historia evolutiva de las especies [1, 2]. De acuerdo con esta definición, éstos genes pueden ser utilizados para rastrear los eventos de especiación ocurridos en el conjunto de especies de las cuales fueron secuenciados; más aún, cuando uno está seguro de la ortología de un gen, éste puede ser usado para inferir la historia de las especies[2].

Las causas por las cuales la historia de un gen puede diferir de la historia de las especies son: la transferencia horizontal de genes entre diferentes especies; la duplicación de genes con la subsiguiente pérdida de alguna de las copias en una o varias especies; el proceso de la separación incompleta de linajes génicos (en ingles "incomplete lineage sorting"), en el cual las genealogías de los loci pueden parecer incorrectas o no informativas respecto a las especies, debido a la retención y el arreglo estocástico de polimorfismos ancestrales (la figura A muestra una representación grafica de este proceso); y, por último, la restricción funcional, en la cual la selección purificadora es actor fundamental. Las dos primeras causas son sumamente comunes en las bacterias [3]. De hecho algunos investigadores afirman que la transferencia horizontal es tan abundante en las bacterias que no tiene ningún sentido construir un árbol de las especies [3]. Así mismo, se sabe que la duplicación y la pérdida de genes son procesos bastante frecuentes en las bacterias. Por su parte la separación incompleta de linajes génicos parece ocurrir sólo en los eucariotes; hay varios estudios que han demostrado que la retención de polimorfismos ancestrales ha afectado eventos de especiación reciente en vertebrados.

FIGURA A



Robinson T. J. et.al. PNAS 2008:105:14477-

Incluso cuando la historia del gen es igual a la historia de las especies puede haber ciertos factores que distorsionen la filogenia del gen si no son tomados en cuenta. Estos son los llamados sesgos sistemáticos. Dentro de éstos está el sesgo en el contenido de GC, heterogeneidad en la frecuencia de aminoácidos en las diferentes especies, la variación en la tasa de sustitución de los sitios en los diferentes linajes, etcétera [4].

Con el advenimiento de las nuevas tecnologías de secuenciación se ha obtenido un gran número de genomas secuenciados. Esto es especialmente cierto para los procariotes (actualmente se cuentan con más de 650 genomas bacterianos secuenciados). Tanto en eucariotes como en procariotes se han llevado a cabo estudios en los que, utilizando genomas completamente secuenciados, se ha tratado de inferir el árbol de las especies para diferentes clados [5-8]. En los procariotes los dos clados más analizados han sido las *Alfaproteobacterias* y las *Gamaproteobacterias*. Lerat *et al.* en un estudio llevado a cabo en *Gamaproteobacterias* se encontró concordancia en 203 de las 205 familias génicas, las cuales eran supuestos ortólogos [9]. Sin embargo, Bapteste *et al.* usando las mismas familias génicas encontró que el 10% de ellas habían sufrido transferencias horizontales y que el resto tenían poca señal filogenética [5]. En un estudio más reciente, Comas *et al.* encontró que solo 3 de 200 genes ortólogos dieron una filogenia con una topología igual a la del árbol de las especies y que 29% de éstos rechazó el árbol de las especies [6]. Fitzpatrick *et al.* encontró, usando a las *Alfaproteobacterias,* que el 77% de las filogenias inferidas para genes individuales no tuvieron diferencias significativas con el súper árbol propuesto, el cual fue inferido con todas las filogenias individuales [8]. En otro trabajo cuyo objetivo era obtener un árbol para las *Alfaproteobacterias*, Williams *et al.* determinó que aunque el concatenado de todos los alineamientos individuales de 107 genes ortólogos dio una filogenia robusta, ninguna de las filogenias individuales fue igual a otra [10]. La presencia de falsos positivos, es decir, genes que han sufrido transferencias horizontales y/o duplicaciones con subsiguiente pérdida del gen ortólogo, pudo haber afectado a los trabajos antes mencionados, ya que ellos establecieron ortología usando: 1) mejores "hits" bidireccionales entre pares de genomas; o, 2) familias génicas que sólo presentan un gen por genoma. Ninguna de estas dos estrategias está exenta de falsos positivos. Por otra parte, la mayoría de estos estudios no realizó selección de modelos y sólo utilizó un tipo de matriz de aminoácidos para construir sus filogenias. Esto puede ser una falla considerable, pues según los resultados de una investigación reciente, las familias génicas presentes en las *Proteobacterias* seleccionaron diferentes matrices de aminoácidos [11].

PLANTEAMIENTO DEL PROBLEMA Y JUSTIFICACIÓN.

Dada su definición, los genes ortólogos deberían ser congruentes entre ellos mismos y reflejar sólo una historia: la historia de las especies. Mi proyecto de doctorado trato de determinar si lo anterior realmente ocurre y para ello se plantearon las siguientes preguntas: ¿Los genes ortólogos son congruentes entre ellos mismos? ¿Reflejan los genes ortólogos la historia de las especies de manera precisa?

A diferencia de los trabajos mencionados en la introducción, en los que no se analizó si los genes ortólogos eran congruentes entre ellos mismos, sino si se podía inferir un supuesto árbol robusto de las especies, aquí se analizó si de manera individual los genes ortólogos reflejaban la historia de las especies, lo cual implicaría que son congruentes. En principio sólo el error de muestro (esto es, que los genes sólo tienen un número finito de sitios donde pueden registrar la historia de las especies, por lo tanto, los genes de menor tamaño tendrían menor posibilidad de reflejar de manera precisa dicha historia) y la pérdida de la señal filogenética por múltiples substituciones deberían modificar la posibilidad de que un gen ortólogo refleje la historia de las especies.

Para evitar el problema de saturación ocasionado por múltiples substituciones, en este estudio se escogió un grupo de bacterias con distancias filogenéticas menores en comparación con las distancias de los grupos usados en trabajos previos. El grupo de estudio estuvo constituido por 19 genomas del orden *Rhizobiales* y el genoma de *Caulobacter crescentus*, como grupo externo (se ha estimado que *Caulobacter crescentus* divergió del orden *Rhizobiales* hace 1.5 mil millones de años mientras que las *Alfaproteobacteria* y las *Gamaproteobacteria* divergieron hace más de 2 mil millones de años [12]). Por otra parte, para evitar los efectos de los sesgos sistemáticos, en cada una de las filogenias inferidas se llevó a cabo la selección de modelos (lo cual incluye selección de matrices de aminoácidos, correcciones para la variación en la tasa de substitución a lo largo de los sitios, correcciones para las heterogeneidad en la frecuencia de aminoácidos). Con el fin de establecer un criterio de ortología robusto, fueron inferidas aproximaciones del posible árbol de las especies usando las técnicas de súper matrices (superalineamientos) y árboles consenso. Éstas aproximaciones del posible árbol de las especies fueron usadas para eliminar falsos positivos y para tener un referente de la historia de las especies.

OBJETIVOS

General: Determinar si los genes ortólogos son congruentes con el árbol de las especies.

Particulares.

1) Identificar un conjunto de genes ortólogos confiables.

2) Establecer el posible árbol de las especies.

3) Comparar las filogenias de los genes ortólogos y el árbol de las especies.

4) Determinar los factores que afectan la congruencia entre el árbol de las especies y las filogenias de los genes ortólogos.

HIPÓTESIS

Dada la restricción funcional de cada gen y el hecho de que presenta un número finito, y en general pequeño, de sitios informativos, éste no reflejara de manera exacta la historia de las especies.

METODOLOGÍA

Identificación de los genes ortólogos.

Esta primera parte, la cual define el material de estudio, fue a la que se dedicó mayor tiempo. Aquí la intención fue obtener un material de estudio lo más puro posible, para lo cual se establecieron dos criterios que minimizaban la presencia de falsos positivos. Lo anterior no es intrascendente ya que los grupos de ortólogos fueron utilizados para construir el árbol de las especies. Dado que *Bartonella quintana* cepa Toulouse tiene el proteoma más pequeño de las especies aquí utilizadas (Anexo 1), se buscaron sus homólogos en el resto de las especies para cada una de las proteínas de *B. quintana* cepa Toulouse. Para llevar acabo lo anterior se utilizo BLAST [13], con un E-value menor a 1.0e-12, y sólo se consideraron los casos en los que los hits y la proteína usada como semilla alineaban por lo menos en 50% del total de sus tamaños. Se consideraron como grupos de genes ortólogos potenciales todos aquellos casos donde hubo una relación de mejores hits bidireccionales entre los homólogos presentes en el resto de las especies y el homólogo de *B. quintana*. Posteriormente, para eliminar falsos positivos, se eliminaron todos aquellos grupos de genes que rechazaron la topología del árbol de las especies (ver abajo). Esto fue hecho con la prueba "expected likelihood weights" –esta sirve para determinar si un alineamiento es incompatible con ciertas topologías- implementado en PUZZLE [14]. Por último, para depurar aun más este conjunto de datos, se eliminaron todos aquellos grupos de genes ortólogos potenciales cuyas filogenias no presentaron la relación de hermandad de grupos mostrada por el árbol de las especies (ver figura 1b).

Alineamientos, selección de modelos y filogenias de los genes ortólogos.

En esta sección se describe la construcción de las historias evolutivas individuales de los genes ortólogos. La idea fue realizar filogenias lo más depuradas posibles. Para ello, no sólo se checo el contenido de señal filogenética para cada gen, sino que además se llevó a cabo la selección de modelos para evitar posibles artefactos en la reconstrucción filogenética. Además, se utilizó uno de los métodos más robustos, "máxima verisimilitud", para construir las filogenias individuales. Para cada grupo de genes ortólogos se construyó un alineamiento múltiple con MUSCLE [15]. Después, se hizo la selección de modelos para cada alineamiento múltiple. Esto fue hecho por medio de PROTEST [16]. Las filogenias individuales se realizaron con "máxima verisimilitud", aquí se permitió variación en la tasa de substitución a lo largo de los sitios (distribución gamma) y correcciones para las heterogeneidad en la frecuencia de aminoácidos (cuando se requirió). El programa PHYML [17] fue usado para construir las filogenias, usando la matriz de aminoácidos especificada por PROTEST. La técnica de "likelihood mapping analysis" fue aplicada para determinar el contenido de señal filogenética presente en los grupos de genes ortólogos, se utilizando PUZZLE para dicho fin.

Árbol de las especies.

Para tener mayor certeza en el árbol de las especies se usaron dos aproximaciones, esto se debe a las cuestiones que menciono a continuación. Primero la inferencia de árboles de especies no es una cuestión trivial en si y, segundo, los propios genes ortólogos (lo cuales fueron el objeto de estudio de este trabajo) fueron utilizados para inferir el árbol de las especies. En la primera aproximación se concatenaron todos los alineamientos individuales y posteriormente se uso ese superalineamiento para construir una filogenia Bayesiana. Las filogenias Bayesianas se construyeron con el programa MrBayes [18], permitiendo que se exploraran todas la matrices de aminoácidos que contiene este programa. El número de categorías de la distribución gamma fue 4 y se permitió una proporción de sitios invariables. Dada la excesiva carga computacional, sólo se realizó una sola corrida por 500, 000 generaciones y cada 500 generaciones se tomó un árbol con todos sus parámetros (topología, matriz de aminoácidos, largo de ramas, etc.). El primer 25% del total de las generaciones se tomó como “burn-in” y se descartó. Posteriormente, se resumió el restante 75% para establecer un árbol con todos sus parámetros.

La otra aproximación consistió en la construcción de un árbol consenso a partir de las filogenias individuales de los genes ortólogos. Esto se llevó acabo con la aplicación CONSENSE contenida en el software PHYLIP [19].

Comparación entre el árbol de las especies y la filogenias de los genes ortólogos.

Para saber que tan parecidas eran las filogenias respecto a el árbol de las especies y a ellas mismas se utilizó la distancia de Robison y Fould, implementada en la aplicación TREEDIST que se encuentra en el paquete PHYLIP. Dicha distancia indica el número de biparticiones que son exclusivas de una u otra topología. Cuando dos topologías son iguales la distancia de Robison y Fould es cero.

Categorías funcionales y grados de divergencia de los genes ortólogos.

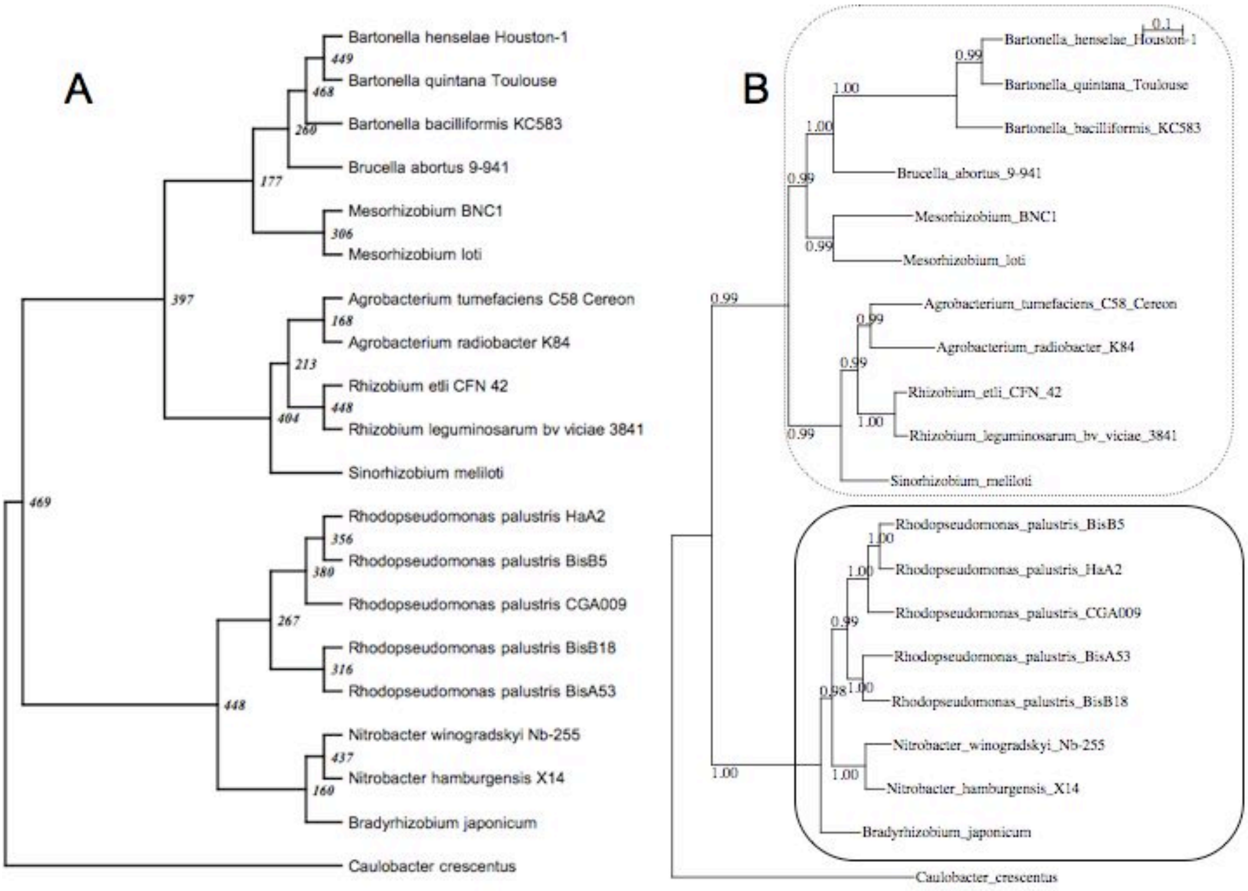
La base de datos "COG" [20] fue empleada para dividir los grupos de genes ortólogos en cuatro categorías funcionales. La idea de esta división fue tratar de establecer si los genes de alguna categoría en particular son más propensos a ser genes ortólogos. Estas categorías son: "Almacenamiento y procesamiento de la información", "Procesos celulares y señalización", "Metabolismo" y "Pobremente caracterizados". Los pocos genes que no pudieron ser asignados a ninguna de estas categorías fueron asignados a la categoría "Pobremente caracterizados". Por otra parte, dependiendo de la función de un gen ortólogos este puede acumular cambios con mayor o menor facilidad, para determinar eso se necesita alguna medida del grado de divergencia de los diferentes genes ortólogos. Como una medida del grado de divergencia se utilizó el largo total de cada una de las filogenias individuales.

RESULTADOS

1. Identificación de los genes ortólogos.

Con el método de mejores "hits" bidireccionales, un conjunto de 469 genes ortólogos potenciales fueron determinados. Con ellos se obtuvo una aproximación de lo que podría ser el árbol de las especies, utilizando los métodos de árboles consenso y superalineamiento. Las filogenia Bayesiana obtenida del superalineamineto se muestra en la figura 1B. El árbol consenso de las 469 filogenias inferidas por máxima verosimilitud, no fue idéntico a la filogenia del superalineamineto (figura 1A). La diferencia radica en la posición de la especie *Bradyrhizobium japonicum*: mientras que en la filogenia Bayesiana la especie *B. japonicum* queda excluida del grupo formado por los géneros *Nitrobacter* y *Rhodopseudomonas*, en el árbol consenso la especie *B. japonicum* y el género *Nitrobacter* forman un grupo y excluyen al género *Rhodopseudomonas*. El grupo antes mencionado fue el menos soportado en el árbol consenso y sólo estuvo presente en 160 de las 469 filogenias.

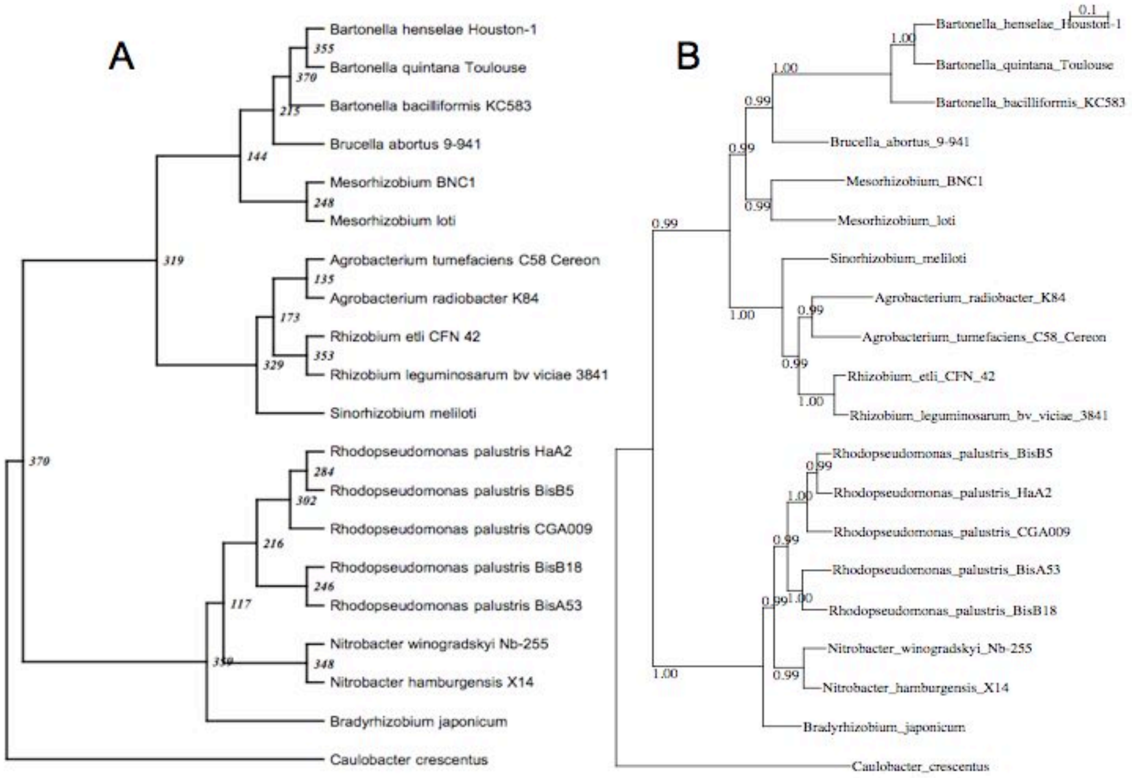
Figura 1


A
Bartonella henselae Houston-1
449
Bartonella quintana Toulouse
468
Bartonella bacilliformis KC583
260
Brucella abortus 9-941
177
Mesorhizobium BNC1
306
Mesorhizobium loti
397
Agrobacterium tumefaciens C58 Cereon
168
Agrobacterium radiobacter K84
213
Rhizobium etli CFN 42
448
Rhizobium leguminosarum bv viciae 3841
404
Sinorhizobium meliloti
469
Rhodopseudomonas palustris HaA2
356
Rhodopseudomonas palustris BisB5
380
Rhodopseudomonas palustris CGA009
267
Rhodopseudomonas palustris BisB18
316
Rhodopseudomonas palustris BisA53
448
Nitrobacter winogradskyi Nb-255
437
Nitrobacter hamburgensis X14
160
Bradyrhizobium japonicum
Caulobacter crescentus
B
0.1
Bartonella_henselae_Houston-1
0.99
Bartonella_quintana_Toulouse
1.00
Bartonella_bacilliformis_KC583
1.00
Brucella_abortus_9-941
0.99
Mesorhizobium_BNC1
0.99
Mesorhizobium_loti
0.99
Agrobacterium_tumefaciens_C58_Cereon
0.99
Agrobacterium_radiobacter_K84
0.99
Rhizobium_etli_CFN_42
1.00
Rhizobium_leguminosarum_bv_viciae_3841
0.99
Sinorhizobium_meliloti
Rhodopseudomonas_palustris_BisB5
1.00
Rhodopseudomonas_palustris_HaA2
1.00
Rhodopseudomonas_palustris_CGA009
0.99
Rhodopseudomonas_palustris_BisA53
0.98
1.00
Rhodopseudomonas_palustris_BisB18
Nitrobacter_winogradskyi_Nb-255
1.00
1.00
Nitrobacter_hamburgensis_X14
Bradyrhizobium_japonicum
Caulobacter_crescentus

2. Eliminando falsos positivos y determinando el contenido de señal filogenética.

Dos filtros, que fueron la prueba “expected likelihood weights” –esta sirve para determinar si un alineamiento es incompatible con ciertas topologías- y la relación de hermandad de grupos mostrada en la figura 1B, se aplicaron para eliminar todos aquellos grupos de ortólogos potenciales que pudieron haber sido afectados por eventos de transferencia horizontal y/o duplicación. Alrededor de un 20% del grupo de genes ortólogos potenciales fueron eliminados por uno o ambos filtros; 370 grupos de ortólogos quedaron después de aplicar ambos filtros. Con esos 370 grupos se hizo un nuevo superalineamiento, el cual se utilizó para construir una nueva filogenia Bayesiana y con las filogenias individuales un nuevo árbol consenso fue determinado. Como era de esperarse, si el proceso de remoción de falsos positivos fue efectivo, tanto la nueva filogenia Bayesiana como el árbol consenso produjeron la misma topología (figura 2) y ésta fue igual a la filogenia Bayesiana determinada con el superalineamiento hecho con los 469 grupos de ortólogos potenciales. Puesto que es muy probable que en los 370 grupos de ortólogos no haya falsos positivos, la mejor aproximación del árbol de las especies es la filogenia Bayesiana inferida del superalineamiento de estos 370 grupos. La técnica de "likelihood mapping analysis" fue aplicada para determinar el contenido de señal filogenética de los 370 grupos de ortólogos. La media del número de cuartetos resueltos de las filogenias de los 370 grupos fue de 90.9%, con un error estándar de 1.32%, lo cual sugiere que en general los ortólogos tienen muy buena señal filogenética, ya que el caso perfecto es aquel en el que se tiene un 100% de cuartetos resueltos. Para cada filogenia individual de los 370 ortólogos, se sacó la mediana de los valores de "bootstrap" de los diferentes nodos; la media de esas medianas fue 77 con un error estándar de 4. El dato anterior indica que las filogenias individuales están bien soportadas en la mayoría de sus nodos.

Figura 2


A
Bartonella henselae Houston-1
355
Bartonella quintana Toulouse
370
Bartonella bacilliformis KC583
215
Brucella abortus 9-941
144
Mesorhizobium BNC1
248
Mesorhizobium loti
319
Agrobacterium tumefaciens C58 Cereon
135
Agrobacterium radiobacter K84
173
Rhizobium etli CFN 42
353
329
Rhizobium leguminosarum bv viciae 3841
Sinorhizobium meliloti
370
Rhodopseudomonas palustris HaA2
284
Rhodopseudomonas palustris BisB5
302
Rhodopseudomonas palustris CGA009
216
Rhodopseudomonas palustris BisB18
246
Rhodopseudomonas palustris BisA53
117
Nitrobacter winogradskyi Nb-255
350
348
Nitrobacter hamburgensis X14
Bradyrhizobium japonicum
Caulobacter crescentus
B
0.1
Bartonella_henselae_Houston-1
1.00
Bartonella_quintana_Toulouse
1.00
0.99
Bartonella_bacilliformis_KC583
Brucella_abortus_9-941
0.99
Mesorhizobium_BNC1
0.99
Mesorhizobium_loti
0.99
Sinorhizobium_meliloti
Agrobacterium_radiobacter_K84
1.00
0.99
Agrobacterium_tumefaciens_C58_Cereon
0.99
Rhizobium_etli_CFN_42
1.00
Rhizobium_leguminosarum_bv_viciae_3841
Rhodopseudomonas_palustris_BisB5
0.99
Rhodopseudomonas_palustris_HaA2
1.00
Rhodopseudomonas_palustris_CGA009
0.99
Rhodopseudomonas_palustris_BisA53
0.99
1.00
Rhodopseudomonas_palustris_BisB18
Nitrobacter_winogradskyi_Nb-255
1.00
0.99
Nitrobacter_hamburgensis_X14
Bradyrhizobium_japonicum
Caulobacter_crescentus

3. Gran variedad de topologías pero pocas biparticiones distintas.

Hubo 346 topologías diferentes en los 370 ortólogos y 93% de éstos tuvieron topologías que no compartieron con ningún otro ortólogo. Sólo dos genes presentaron la topología de la filogenia Bayesiana hecha con los 370 grupos y éstos codifican la proteasa Lon dependiente de ATP y la subunidad beta de la RNA polimerasa. Inesperadamente, la topología más frecuente, presentada por 6 ortólogos, fue la obtenida por el árbol consenso sacado de los 469 grupos de ortólogos potenciales. Esto pone de manifiesto que hay una gran variedad de topologías y que la más frecuente no fue la del árbol de las especies, lo cual deja claro que la inmensa mayoría de genes ortólogos no refleja de manera exacta el árbol de las especies. Pero por otra parte, al analizar las biparticiones de las filogenias –particiones no triviales; es decir, aquellas particiones que ocurren sólo en las ramas internas de las filogenias-, 72% de las biparticiones totales concuerdan con las biparticiones de la filogenia Bayesiana sacada con el superalineamiento de los 370 grupos de ortólogos. Más aún, en promedio cada filogenia individual comparte 71.76% de sus biparticiones con la filogenia Bayesiana del superalineamiento, de tal suerte que en conjunto y a nivel individual las filogenias de los genes ortólogos comparten más del 70% de sus biparticiones con la filogenia Bayesiana.

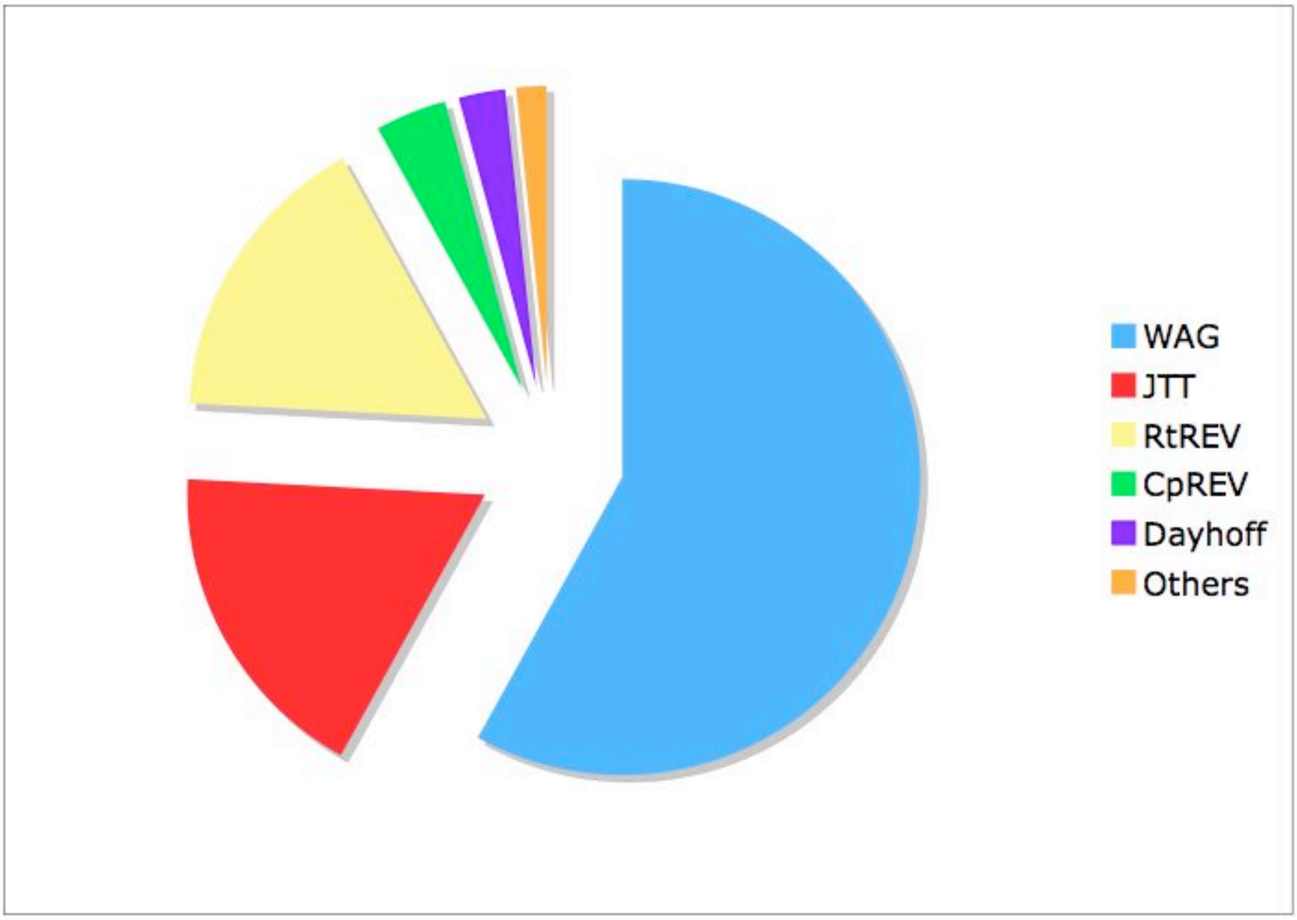
4. Error de muestreo.

El error de muestro debido al tamaño del gen está afectando a los genes ortólogos. Una correlación significativa, aunque no fuerte, ($p < 0.00001$, con coeficientes de correlación y determinación de 0.39 y 0.15) fue encontrada entre el tamaño de los genes ortólogos y el número de biparticiones en común entre la filogenia de cada gen ortólogo y la filogenia Bayesiana inferida con el superalineamiento de los 370 grupos de ortólogos. Por otra parte, se encontró una correlación significativa (coeficiente de correlación de 0.624 y coeficiente de determinación de 0.384) entre el largo de la ramas internas de la mejor aproximación del árbol de las especies y la ocurrencia de la biparticiones definidas por esas ramas en el conjunto de los genes ortólogos. Esto es, las biparticiones definidas por ramas más largas en árbol de las especies son más propensas a ocurrir en las filogenias de los genes ortólogos.

5. Diferentes modelos evolutivos fueron escogidos.

La selección de modelos se realizó para evitar artefactos en la construcción de las filogenias, pero a su vez ésta puede arrojar luz acerca del modo en que están evolucionando los genes ortólogos. La matriz de aminoácidos más escogida por los 370 grupos fue la WAG, con un 58%; la segunda matriz mas requerida fue la JTT escogida por un 18% de los genes ortólogos (ver figura 3). Sólo cuatro de las ocho matrices seleccionadas fueron escogidas por más de 10 genes ortólogos (ver figura 3). Todos los genes ortólogos necesitaron tomar en cuenta la variación en la tasa de substitución a lo largo de los sitios al inferir sus filogenias y 68% requirió correcciones para la frecuencia de aminoácidos. Cerca de 40% necesitó tener una proporción de sitios invariables. Los resultados de la selección de modelos indican que hubo diversidad, aunque no muy amplia, en la descripción del modo de evolución de los genes ortólogos.

Figura 3



6. Los genes ortólogos presentaron un sesgo en el tipo de funciones que realizan.

La base de datos COG (acrónimo del inglés "Cluster of Orthologous Groups") fue utilizada para contrastar las funciones de los genes ortólogos identificados en este trabajo. La distribución de las cuatro categorías más inclusivas difirió significativamente entre la base de datos COG y los genes ortólogos aquí localizados (chi-cuadrada $p < 0.0005$, Tabla 1). Mientras que la categoría más frecuente para los genes ortólogos fue "Almacenamiento y procesamiento de la información", con un 34%, para la base de datos COG la categoría más abundante fue la categoría "Pobremente caracterizados", con un 40% (Tabla 1). En el lado opuesto, la categoría con menor cantidad de ortólogos fue "Almacenamiento y procesamiento de la información", teniendo un 15%, en la base de datos COG, mientras que la menos frecuente para los genes ortólogos de este trabajo fue la "Pobremente caracterizados", la cual tiene un 12%. Las dos categorías que están relacionadas con el funcionamiento básico de la célula procariótica tuvieron una mayor representación en el conjunto de datos.

Tabla 1

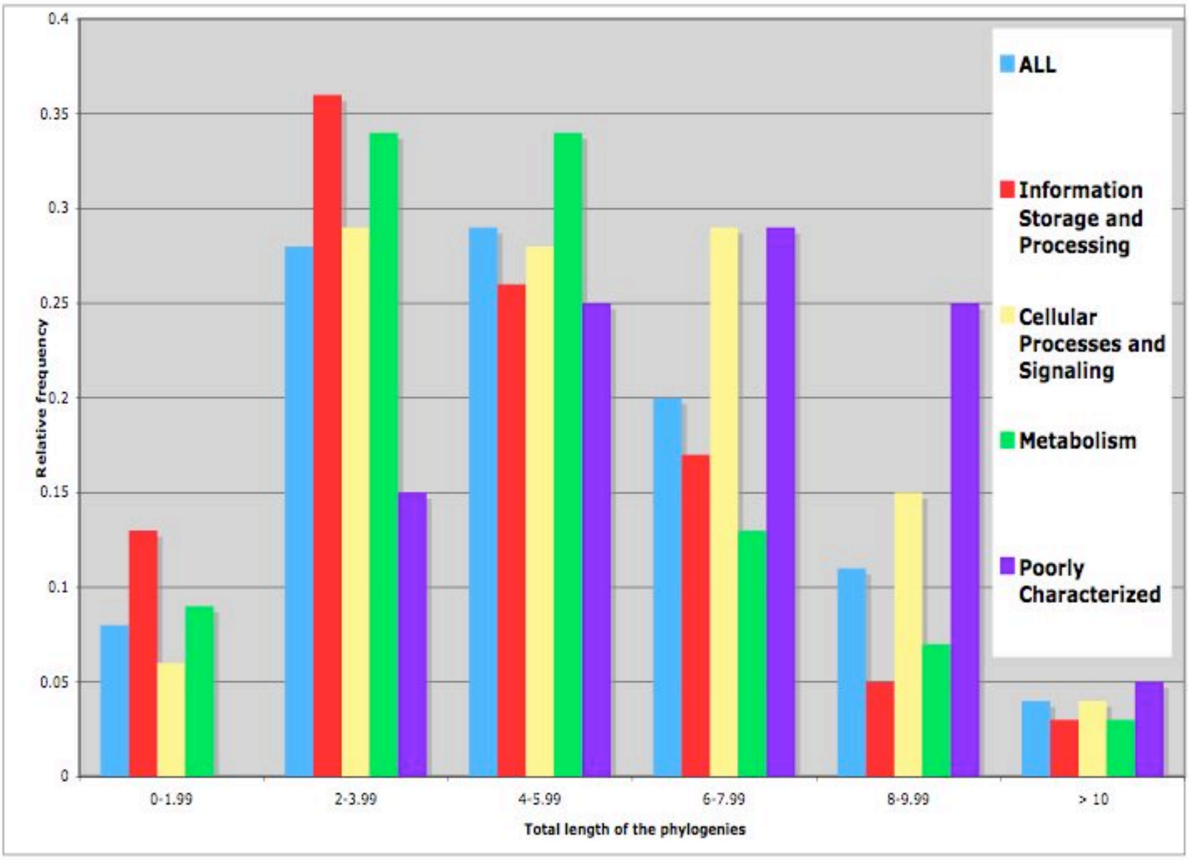
Categoría	Genes ortólogos	Base de datos COG	Error de muestreo
Almacenamiento y procesamiento de la información	34%	15%	0.36
Procesos celulares y señalización	24%	18%	0.37
Metabolismo	31%	28%	0.15*
Pobremente caracterizados	12%	40%	0.49

* correlación no significativa

7. Los genes ortólogos de las diferentes categorías funcionales presentaron diferentes grados de divergencia.

El largo total de las filogenias fue usado como una medida del grado de divergencia. Hubo una gran variación en los grados de divergencia en los genes ortólogos, el coeficiente de variación fue de 53%. Más aún, las categorías funcionales presentaron diferentes grados de divergencia (Kruskal-Wallis, $p < 0.0005$, figura 4). La categoría "Pobremente caracterizados" tuvo los ortólogos más divergentes: la mayoría de sus genes tuvo entre 6 y 8 substituciones por sitio por filogenia (figura 4, barras moradas). Las categorías con los ortólogos menos divergentes fueron "Almacenamiento y procesamiento de la información" y "Metabolismo", cuyos genes, en su mayoría, tuvieron entre 2 y 4 substituciones por sitio por filogenia (barras rojas y verdes en la figura 4). Los resultados anteriores sugieren que los genes ortólogos no sólo tienen una gran variación en sus niveles de divergencia sino que además estos grados de divergencia se estructuran de acuerdo a las categorías funcionales.

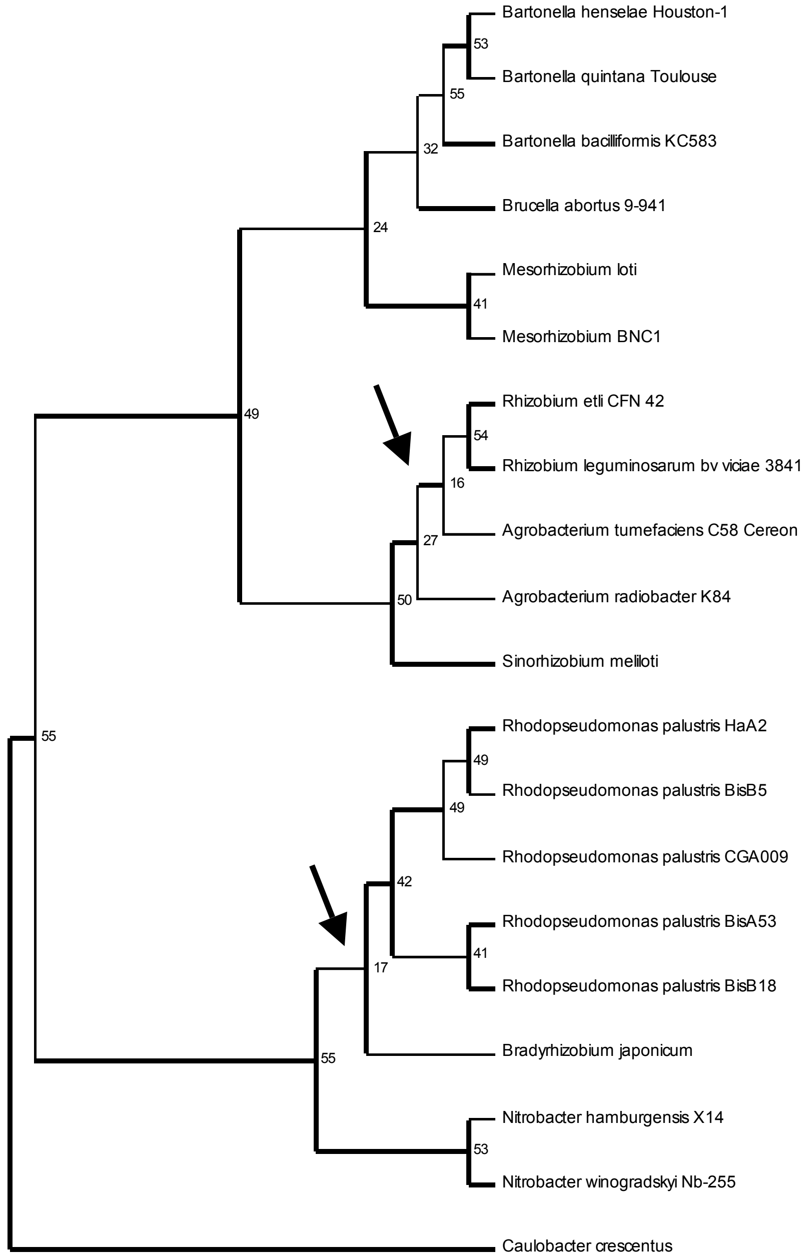
Figura 4



8. Congruencia y error de muestreo en las diferentes categorías funcionales.

Para analizar la congruencia al interior de las categorías funcionales, filogenias Bayesianas -usando superalineamientos- y árboles consenso fueron establecidos para cada una de las cuatro categorías funcionales. Las filogenias Bayesianas de las cuatro categorías dieron la misma topología que la mejor aproximación del árbol de las especies. Pero no fue así para los árboles consenso. Los árboles consenso de las categorías "Almacenamiento y procesamiento de la información" y "Procesos celulares y señalización" si tuvieron la misma topología las filogenias de los superalineamientos. El árbol consenso de la categoría "Metabolismo" fue igual que el árbol consenso que consideró los 469 grupos de ortólogos potenciales –la discordancia de esta topología consiste en la posición de *B. japonicum* antes mencionada-. Por su parte, el árbol consenso de la categoría “Pobremente caracterizados” además de una discordancia en la posición de *B. japonicum*, no recuperó al género *Agrobacterium* como un grupo monofilético (figura 5, los puntos de incongruencia están señalados con flechas), interesantemente un artículo publicado en 2001 ya había dicho que el género *Agrobacterium* no es un grupo natural [21]. Así mismo, se analizó el error de muestro para cada categoría. Las categorías tuvieron diferentes coeficientes de correlación; todas las correlaciones, salvo la categoría “Metabolismo”, fueron significativas (Tabla 1). La categoría que tuvo una correlación más fuerte fue “Pobremente caracterizados” con un coeficiente de casi 0.5, en tanto que “Almacenamiento y procesamiento de la información” y “Procesos celulares y señalización” tuvieron coeficientes de correlación muy parecidos. De lo anterior se puede colegir que hay categorías funcionales con menor congruencia y que el error de muestro no se distribuye de manera uniforme en las categorías funcionales.

Figura 5


Bartonella henselae Houston-1
53
Bartonella quintana Toulouse
55
Bartonella bacilliformis KC583
32
Brucella abortus 9-941
24
Mesorhizobium loti
41
Mesorhizobium BNC1
49
Rhizobium etli CFN 42
54
Rhizobium leguminosarum bv viciae 3841
16
Agrobacterium tumefaciens C58 Cereon
27
Agrobacterium radiobacter K84
50
Sinorhizobium meliloti
55
Rhodopseudomonas palustris HaA2
49
Rhodopseudomonas palustris BisB5
49
Rhodopseudomonas palustris CGA009
42
Rhodopseudomonas palustris BisA53
41
Rhodopseudomonas palustris BisB18
17
Bradyrhizobium japonicum
55
Nitrobacter hamburgensis X14
53
Nitrobacter winogradskyi Nb-255
Caulobacter crescentus

DISCUSIÓN Y CONCLUSIONES

Sólo dos de los 370 grupos de ortólogos presentaron la topología del mejor aproximado del árbol de las especies. Así, la mayoría de los genes ortólogos no reflejan exactamente la historia de las especies. De manera inesperada la topología más común no fue la de la mejor aproximación del árbol de la especies, aunque es muy parecida (más adelante discuto su posible explicación). Esto no tiene que ver con la falta de señal filogenética puesto que este grupo de ortólogos, en general, tuvo un porcentaje alto de cuartetos resueltos (lo cual es indicativo de un conjunto de datos con buena señal filogenética), así como un buen valor medio de "bootstrap", lo que sugiere que tampoco hay problemas con el soporte de los nodos de las filogenias. La imposibilidad de reflejar exactamente la historia de las especies tampoco parece estar relacionada con los sesgos sistemáticos, pues para cada una de las filogenias aquí construidas se utilizó la selección de modelos.

Si bien los genes ortólogos, de manera individual, no reflejan exactamente la historia de las especies, gran parte de sus biparticiones coinciden con el mejor aproximado del árbol de las especies. De hecho, en promedio, 70% de sus bipariciones reflejan dicha historia. La situación cambia cuando los genes ortólogos son tomados en conjunto, ya que en este caso se ve potenciada la capacidad de reflejar la historia de las especies, pues las dos aproximaciones usadas para inferir el árbol de las especies dieron como resultado la misma topología.

Uno de los factores que afecta la concordancia entre el árbol de las especies y los genes ortólogos individuales es el error de muestro; es decir, dado que los genes sólo tienen un número finito de sitios donde pueden registrar la historia de las especies, los genes de menor tamaño tendrían menor posibilidad de reflejar de manera precisa dicha historia. Sin embargo éste no afecta de manera uniforme a todos los genes ortólogos. De hecho los ortólogos de la categoría funcional de “Metabolismo” no parecen ser afectados mientras que la categoría “Pobremente caracterizados” fue la que se vio afectada de manera más drástica.

Por otro lado, las biparticiones del árbol de las especies que son menos reflejadas en la filogenias individuales son las ramas internas más cortas. Estas ramas no sólo son difíciles de resolver por la poca acumulación de caracteres que las definen, sino que además pudieron haber sido afectadas por el proceso de la separación incompleta de linajes génicos. La separación incompleta de linajes génicos puede ser un factor de discordancia, particularmente cuando las ramas internas del árbol de las especies son muy cortas, de tal manera que la coalescencia de los genes antecede al evento de especiación [22, 23]. Un trabajo analítico mostró que ramas internas cortas, con posiciones profundas en árboles de las especies, conteniendo 5 o más especies, pueden tener filogenias anómalas, es decir, que dichas filogenias que no empatan con el árbol de las especies [22]. Incluso la filogenia más probable puede tener una topología diferente del árbol de las especies si algunas ramas tienen un largo muy pequeño en unidades de coalescencia (esto es denominado AGT, acrónimo del inglés “Anomalous Gene Tree”) [22]. A continuación enumero dos hechos que hacen pensar que la separación incompleta de linajes génicos ha afectado al conjunto de ortólogos estudiados. Primero, las dos biparticiones menos presentes en las filogenias de los genes individuales involucraron 2 de las 3 ramas internas más cortas en árbol de las especies. La topología más común, que es un claro ejemplo de AGT, difirió del árbol de las especies en una de las dos ramas internas más cortas antes discutidas.

Los genes ortólogos no forman un conjunto uniforme: presentaron una gran variedad de topologías e incluso diferencias en los modos en que han evolucionado. A pesar de esta amplia diversidad, la restricción funcional organiza a los genes ortólogos en unos cuantos grupos. Por un lado, la mayoría de los genes ortólogos son más propensos a desempeñar funciones relacionadas con las categorías "Almacenamiento y procesamiento de la información" y "Procesos celulares y señalización" y, al mismo tiempo, dichos genes presentan menores grados de divergencia, es decir, están más conservados y son más refractarios a las causas de discordancia. Por otro lado, los genes ortólogos pertenecientes a la categoría "Pobremente caracterizados" no sólo fueron lo menos abundantes, sino además tuvieron los mayores grados de divergencia y el error de muestreo tuvo un mayor impacto. Además esta categoría fue la que presentó mayor discordancia.

PERSPECTIVAS

Este estudio utilizó como objeto de estudio los genes ortólogos de un solo clado bacteriano, por lo que es incierto hasta que punto lo encontrado en este trabajo puede extrapolarse a otros clados bacterianos. En ese sentido convendría extrapolar la estrategia experimental aquí planteada a otros clados bacterianos y determinar si las reglas aquí encontradas se aplican a la mayoría de los clados bacterianos. En principio uno esperaría que ciertos factores estuvieran presentes en cualquier clado bacteriano, tal es el caso del error de muestro o la organización de los genes ortólogos de acuerdo a su restricción funcional. Pero a su vez, hay otros factores que *a priori* no tendrían porque ser ubicuos. Tal es el caso de la separación incompleta de los linajes génicos o el hecho de que muy pocos genes ortólogos reflejen la topología del árbol de las especies.

Uno de los resultados más inesperados de este proyecto fue la presencia de la separación incompleta de linajes génicos como un factor de discordancia relevante. De hecho, este factor nunca había sido descrito como un elemento de discordancia en el dominio de las bacterias. Convendría explorar hasta que punto este factor afecta no solo a los genes ortólogos sino a cualquier grupo de homólogos en los diferentes clados bacterianos.

BIBLIOGRAFÍA

1. Fitch, W.M., *Distinguishing homologous from analogous proteins.* Syst Zool, 1970. **19**(2): p. 99-113.
2. Fitch, W.M., *Homology a personal view on some of the problems.* Trends Genet, 2000. **16**(5): p. 227-31.
3. Zhaxybayeva, O., P. Lapierre, and J.P. Gogarten, *Genome mosaicism and organismal lineages.* Trends Genet, 2004. **20**(5): p. 254-60.
4. Rodriguez-Ezpeleta, N., et al., *Detecting and overcoming systematic errors in genome-scale phylogenies.* Syst Biol, 2007. **56**(3): p. 389-99.
5. Bapteste, E., et al., *Do orthologous gene phylogenies really support tree-thinking?* BMC Evol Biol, 2005. **5**(1): p. 33.
6. Comas, I., A. Moya, and F. Gonzalez-Candelas, *From phylogenetics to phylogenomics: the evolutionary relationships of insect endosymbiotic gamma-Proteobacteria as a test case.* Syst Biol, 2007. **56**(1): p. 1-16.
7. Creevey, C.J., et al., *Does a tree-like phylogeny only exist at the tips in the prokaryotes?* Proc Biol Sci, 2004. **271**(1557): p. 2551-8.
8. Fitzpatrick, D.A., C.J. Creevey, and J.O. McInerney, *Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales.* Mol Biol Evol, 2006. **23**(1): p. 74-85.
9. Lerat, E., V. Daubin, and N.A. Moran, *From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria.* PLoS Biol, 2003. **1**(1): p. E19.
10. Williams, K.P., B.W. Sobral, and A.W. Dickerman, *A robust species tree for the alphaproteobacteria.* J Bacteriol, 2007. **189**(13): p. 4578-86.
11. Keane, T.M., et al., *Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified.* BMC Evol Biol, 2006. **6**: p. 29.
12. Battistuzzi, F.U., A. Feijao, and S.B. Hedges, *A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land.* BMC Evol Biol, 2004. **4**: p. 44.
13. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
14. Strimmer, K. and A. von Haeseler, *Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment.* Proc Natl Acad Sci U S A, 1997. **94**(13): p. 6815-9.
15. Edgar, R.C., *MUSCLE: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, 2004. **5**: p. 113.
16. Abascal, F., R. Zardoya, and D. Posada, *ProtTest: selection of best-fit models of protein evolution.* Bioinformatics, 2005. **21**(9): p. 2104-5.
17. Guindon, S. and O. Gascuel, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.* Syst Biol, 2003. **52**(5): p. 696-704.
18. Ronquist, F. and J.P. Huelsenbeck, *MrBayes 3: Bayesian phylogenetic inference under mixed models.* Bioinformatics, 2003. **19**(12): p. 1572-4.
19. Felsenstein, J., *PHYLIP (Phylogeny Inference Package) version 3.6.* Department of Genome Sciences, University of Washington, Seattle, 2005.

20. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic Acids Res, 2000. **28**(1): p. 33-6.

21. Young, J.M., et al., *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): p. 89-103.

22. Degnan, J.H. and N.A. Rosenberg, *Discordance of species trees with their most likely gene trees.* PLoS Genet, 2006. **2**(5): p. e68.

23. Degnan, J.H. and L.A. Salter, *Gene tree distributions under the coalescent process.* Evolution, 2005. **59**(1): p. 24-37.

ANEXO 1

Genomas utilizados	GenBank	Tamaño en megabases
Agrobacterium radiobacter K84	CP000628	7.31
**Agrobacterium vitis* S4	CP000633	6.31
Sinorhizobium meliloti	AL591688	6.8
Rhodopseudomonas palustris HaA2	CP000250	5.33
Rhodopseudomonas palustris CGA009	BX571963	5.51
Rhodopseudomonas palustris BisB5	CP000283	4.89
Rhodopseudomonas palustris BisB18	CP000301	5.51
Rhodopseudomonas palustris BisA53	CP000463	5.51
Rhizobium leguminosarum bv viciae 3841	AM236080	7.79
Rhizobium etli CFN42	CP000133	6.53
Nitrobacter winogradskyi Nb-255	CP000115	3.4
Nitrobacter hamburgensis X14	CP000319	5.01
Mesorhizobium loti	BA000012	7.6
Mesorhizobium BNC1	CP000390	4.94
Caulobacter crescentus CB15	AE005673	4
**Brucella suis*1330	AE014291	3.31
**Brucella melitensis* biovar Abortus	AM040264	3.32
**Brucella melitensis* 16M	AE008917	3.29
Brucella abortus 9-941	AE017223	3.3
Bradyrhizobium japonicum	BA000040	9.1
Bartonella quintana strain Toulouse	BX897700	1.58
Bartonella henselae Houston-1	BX897699	1.93
Bartonella bacilliformis KC583	CP000524	1.4
**Agrobacterium tumefaciens* C58 UWash	AE007869	5.65
Agrobacterium tumefaciens C58 Cereon	AE007869	5.65

*Estas especies fueron excluidas porque son redundantes con otras especies del mismo género.

ANEXO 2

Artículos publicados o por publicarse.

BMC Evolutionary Biology

BioMed Central

Research article

Open Access

Factors affecting the concordance between orthologous gene trees and species tree in bacteria

Santiago Castillo-Ramírez* and Víctor González

Address: Programa de Genómica Evolutiva, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Apartado Postal 565-A, CP 62210, Cuernavaca, Morelos, México

Email: Santiago Castillo-Ramírez* - iago@ccg.unam.mx; Víctor González - vgonzal@ccg.unam.mx

* Corresponding author

Published: 30 October 2008

BMC Evolutionary Biology 2008, **8**:300 doi:10.1186/1471-2148-8-300

Received: 1 July 2008
Accepted: 30 October 2008

This article is available from: http://www.biomedcentral.com/1471-2148/8/300

Abstract

Background: As originally defined, orthologous genes implied a reflection of the history of the species. In recent years, many studies have examined the concordance between orthologous gene trees and species trees in bacteria. These studies have produced contradictory results that may have been influenced by orthologous gene misidentification and artefactual phylogenetic reconstructions. Here, using a method that allows the detection and exclusion of false positives during identification of orthologous genes, we address the question of whether putative orthologous genes within bacteria really reflect the history of the species.

Results: We identified a set of 370 orthologous genes from the bacterial order *Rhizobiales*. Although manifesting strong vertical signal, almost every orthologous gene had a distinct phylogeny, and the most common topology among the orthologous gene trees did not correspond with the best estimate of the species tree. However, each orthologous gene tree shared an average of 70% of its bipartitions with the best estimate of the species tree. Stochastic error related to gene size affected the concordance between the best estimated of the species tree and the orthologous gene trees, although this effect was weak and distributed unevenly among the functional categories. The nodes showing the greatest discordance were those defined by the shortest internal branches in the best estimated of the species tree. Moreover, a clear bias was evident with respect to the function of the orthologous genes, and the degree of divergence among the orthologous genes appeared to be related to their functional classification.

Conclusion: Orthologous genes do not reflect the history of the species when taken as individual markers, but they do when taken as a whole. Stochastic error affected the concordance of orthologous genes with the species tree, albeit weakly. We conclude that two important biological causes of discordance among orthologous genes are incomplete lineage sorting and functional restriction.

Background

Fitch coined the term orthologous genes to describe genes whose phylogenies represent the phylogeny of the species [1,2]. Classically, gene orthology is established by comparing the phylogenetic tree obtained from the gene in question with that for the reference species. As bacterial comparative genomics deal with large amounts of data, requiring extensive computational power and time, sophisticated phylogenetic analysis cannot be easily automated. Thus, most of the studies in this area have used sequence similarity approaches to infer orthology. The reciprocal best hits (RBH) and single gene families (SGF) approaches are the two most common bioinformatic techniques used to infer orthology in bacterial comparative genomics. However, both horizontal gene transfer (HGT), a very pervasive force among bacteria [3-6], and duplications with subsequent differential loss of orthologous genes (DSDL), may result in the misidentification of orthologous genes (false positives) whenever RBH or SGF are used. Moreover, even using *bona fide* orthologous genes and phylogenetically robust methods such as maximum likelihood, incorrect phylogenetic reconstructions may occur when inadequate substitution models are employed [7]. When phylogenetic inference is performed with proteins, inconsistencies may arise due to the use of an incorrect amino acid substitution matrix, or not taking into account for rate variations across sites or variation in the observed amino acid frequencies [8]. Even genome-scale analyses may be susceptible to systematic error when model selection is omitted or a poor model is chosen, particularly when divergence among genes is high. Furthermore, in the case of single markers, individual genes may be affected by stochastic error related to gene size.

Gammaproteobacteria and *Alphaproteobacteria* have been used as model organisms for examining whether a prokaryotic phylogenetic tree can be confidently inferred using many orthologous genes [4,5,9,10]. Phylogenetic concordance among virtually all (203 out of 205) of the selected gene families was found in the case of *Gammaproteobacteria* [10]. However, another study of the same data set determined that 10% of these families had been horizontally transferred and that too little phylogenetic signal was evident in the rest of the families [5]. More recently, it was found that only three out of 200 orthologous genes manifested the topology of the species tree, while 29% of the data set rejected the species tree [11]. In the case of *Alphaproteobacteria*, around 77% of the gene trees inferred from SGF manifested no significant differences with the proposed supertree, which was inferred from all the gene trees, and 76 gene trees were identical to this supertree [4]. In another study, although concatenated alignments indicated a robust tree for the *Alphaproteobacteria*, no two phylogenies obtained from individual families were alike [12]. This apparent incongruence among the trees derived for individual genes may be at least partly due to artefactual phylogenetic reconstruction. Notably, most of these studies did not undertake model selection for individual genes, but instead used a single matrix for all analyses. This may represent a significant flaw, as a recent study in *Proteobacteria* found that, depending on the genes studied, the use of different amino acid matrices is required [8]. However, it is also possible that false positives have caused distortions in some of the prior studies (i.e. the families that rejected the species tree could be subject to HGT and/or DSDL).

Here, we use a strict strategy to infer orthology. First, we establish a RBH approach that applies a higher threshold than regular RBH approaches; an E-value of 10e-12 is used, along with the requirement that the hits align across least 50% of their length. We then use confidence sets of gene trees and an observed sister group relationship to rule out false positives. In this study, we address the question of whether single bacterial orthologous genes, as defined by our strategy, reflect the history of the species. The number of genes in common among species and phylogenetic signal decrease as phylogenetic distance increases; thus, we avoid signal erosion and reduction in the numbers of genes by focusing on a group whose members are separated by only moderate phylogenetic distances. A previous genomic timescale study of prokaryotes estimated that *Caulobacter crescentus* diverged from some species belonging to the order *Rhizobiales* about 1.5 billion years ago, whereas *Alphaproteobacteria* and *Gammaproteobacteria* were estimated to diverge about 2 billion years ago [13]. Here, we use members of the *Rhizobiales* order to make reliable phylogenetic inferences and by applying model selection for each phylogeny we try to avoid artefactual reconstructions.

Our results indicate that orthologous genes manifest a great diversity of phylogenies, and this diversity implies different topologies and models of evolution, as well as an ample level of divergence. The concordance of the orthologous gene trees with the best estimate of the species tree is affected by stochastic error related to gene size, although weakly and the effect is not distributed evenly among functional categories. While the individual phylogenies inferred from orthologous genes are not found to reflect the exact history of the species, the majority of the bipartitions composing the individual phylogenies do reflect such history. The nodes presenting greatest discordance are those defined by the shortest internal branches in the best estimate of the species tree. We see a clear bias concerning the functional categories of the orthologous genes, and this influences their degree of divergence. These results indicate that both functional restriction and incomplete lineage sorting are important factors driving discordance.

Results

The initial set of potential orthologous genes and a probable species tree

The RBH method was used to define an initial set of potential orthologous genes (see methods), yielding 469 candidates. A multiple sequence alignment and phylogeny were constructed for each orthologous gene (see methods). We then used these potential orthologous genes to deduce a probable species tree that helped us refine the set of potential orthologous genes. A consensus tree (469CT) was produced (Figure 1a) using the 469 phylogenies. By concatenating all the individual alignments, a superalignment was created and Bayesian and maximum parsimony phylogenies were inferred. Both methods yielded the same topology; for convenience, the superalignment Bayesian phylogeny (469SBP; Figure 1b) was used for subsequent analyses. The topologies of the 469SBP and 469CT were almost identical, differing only in the position of *Bradyrhizobium japonicum*. The genera *Nitrobacter* and *B. japonicum* were grouped together under 469CT, excluding the genus *Rhodopseudomonas*, whereas *Nitrobacter* and *Rhodopseudomonas* clustered together under 469SBP, excluding *B. japonicum*. Under 469CT, the group comprised of *Nitrobacter* and *B. japonicum* had the smallest presence among single gene phylogenies, being contained in only 160 out of the 469 individual phylogenies.

Ruling out falsely positive orthologs

Even though our RBH approach was stringent, in that we applied BLAST searches with an E-value of 10e-12 and required proteins to align along at least 50% of their length, false positives may still result. In order reduce the risk of false positives, we inferred confidence sets for the 469 alignments. The 469SBP and 469CT topologies were tested for all alignments (see methods). The most accepted topology was 469SBP, which could not be

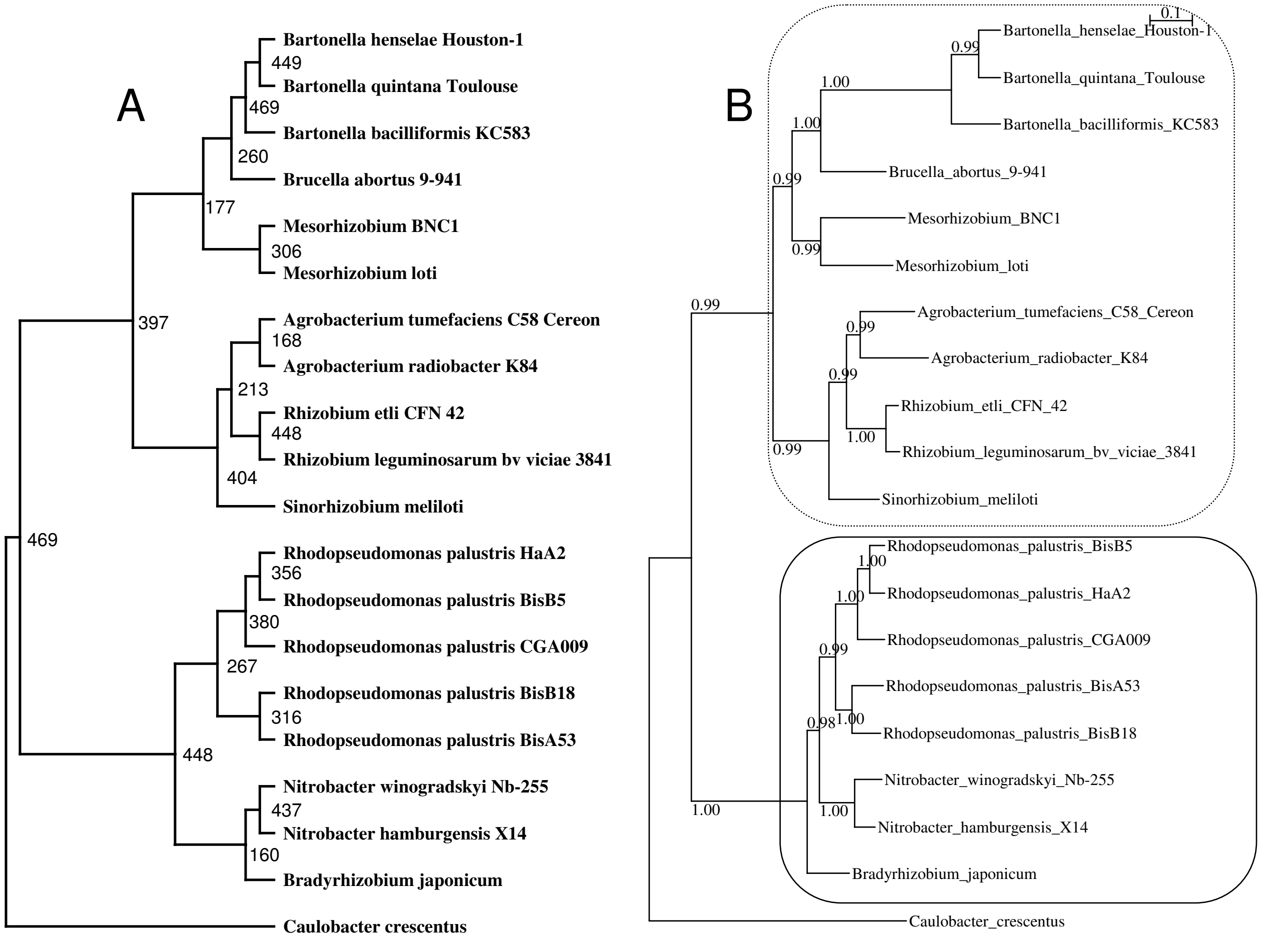

A
Bartonella henselae Houston-1
449
Bartonella quintana Toulouse
469
Bartonella bacilliformis KC583
260
Brucella abortus 9-941
177
Mesorhizobium BNC1
306
Mesorhizobium loti
397
Agrobacterium tumefaciens C58 Cereon
168
Agrobacterium radiobacter K84
213
Rhizobium etli CFN 42
448
Rhizobium leguminosarum bv viciae 3841
404
Sinorhizobium meliloti
469
Rhodopseudomonas palustris HaA2
356
Rhodopseudomonas palustris BisB5
380
Rhodopseudomonas palustris CGA009
267
Rhodopseudomonas palustris BisB18
316
Rhodopseudomonas palustris BisA53
448
Nitrobacter winogradskyi Nb-255
437
Nitrobacter hamburgensis X14
160
Bradyrhizobium japonicum
Caulobacter crescentus
B
0.1
Bartonella_henselae_Houston-1
0.99
Bartonella_quintana_Toulouse
1.00
Bartonella_bacilliformis_KC583
1.00
Brucella_abortus_9-941
0.99
Mesorhizobium_BNC1
0.99
Mesorhizobium_loti
0.99
Agrobacterium_tumefaciens_C58_Cereon
0.99
Agrobacterium_radiobacter_K84
0.99
Rhizobium_etli_CFN_42
1.00
Rhizobium_leguminosarum_bv_viciae_3841
0.99
Sinorhizobium_meliloti
Rhodopseudomonas_palustris_BisB5
1.00
Rhodopseudomonas_palustris_HaA2
1.00
Rhodopseudomonas_palustris_CGA009
0.99
Rhodopseudomonas_palustris_BisA53
0.98
1.00
Rhodopseudomonas_palustris_BisB18
Nitrobacter_winogradskyi_Nb-255
1.00
1.00
Nitrobacter_hamburgensis_X14
Bradyrhizobium_japonicum
Caulobacter_crescentus

Figure 1

The superalignment Bayesian phylogeny (SBP469) and the consensus tree (CT469) constructed from 469 potential orthologous genes. A: CT469, the numbers to the right of the internal branches indicate the number of orthologous genes that contain the group defined by that internal branch. B: SBP469, the numbers on the branches give the posterior probability of the group. The sister group relationship between groups 1 and 2 is denoted by the dashed line (group 1) and the thick line (group 2). The scale bar denotes the estimated number of amino acid substitutions per site.

rejected by 432 potential orthologous genes, whereas only 400 potential orthologous genes did not reject the 469CT topology. As the superalignment only accepted the 469SBP topology, we ruled out the 37 potential orthologous genes that rejected this topology. The 469SBP topology included two very well supported sister groups (Figure 1b, dashed and thick lines). The first group, which was supported by a posterior probability of 0.99, comprised *Sinorhizobium meliloti* and the genera *Rhizobium*, *Agrobacterium*, *Mesorhizobium*, *Bartonella*, and *Brucella abortus* 9–941. The second group, which had a posterior probability of 1.0, comprised the genera *Rhodopseudomonas* and *Nitrobacter*, and *B. japonicum*. This sister group relationship was used to screen the 432 potential orthologous genes that accepted the 469SBP topology, ruling out all potential orthologous genes that contradicted the sister group relationship. This filtering yielded a set of 370 potential orthologous genes, which was expected to not include false positives. New versions of the superalignment Bayesian phylogeny (370SBP) and consensus tree (370CT) were constructed using these 370 orthologous genes (Figure 2). Both analyses yielded the same topology as that found for 469SBP. Both 469SBP and 370SBP had similar branch lengths (see Figure 1b and Figure 2b); however, because 370SBP arguably contained no false positives, it represented the most accurate approximation of the species history. As many authors have used lower thresholds when identifying orthologous genes, we lowered the E-values to 10e-9 and 10e-6 and applied the two filters to see how many orthologous groups could be rescued under these E-values. With E-values of 10e-9 and 10e-6, we rescued 31 and 38 more groups, respectively, compared to

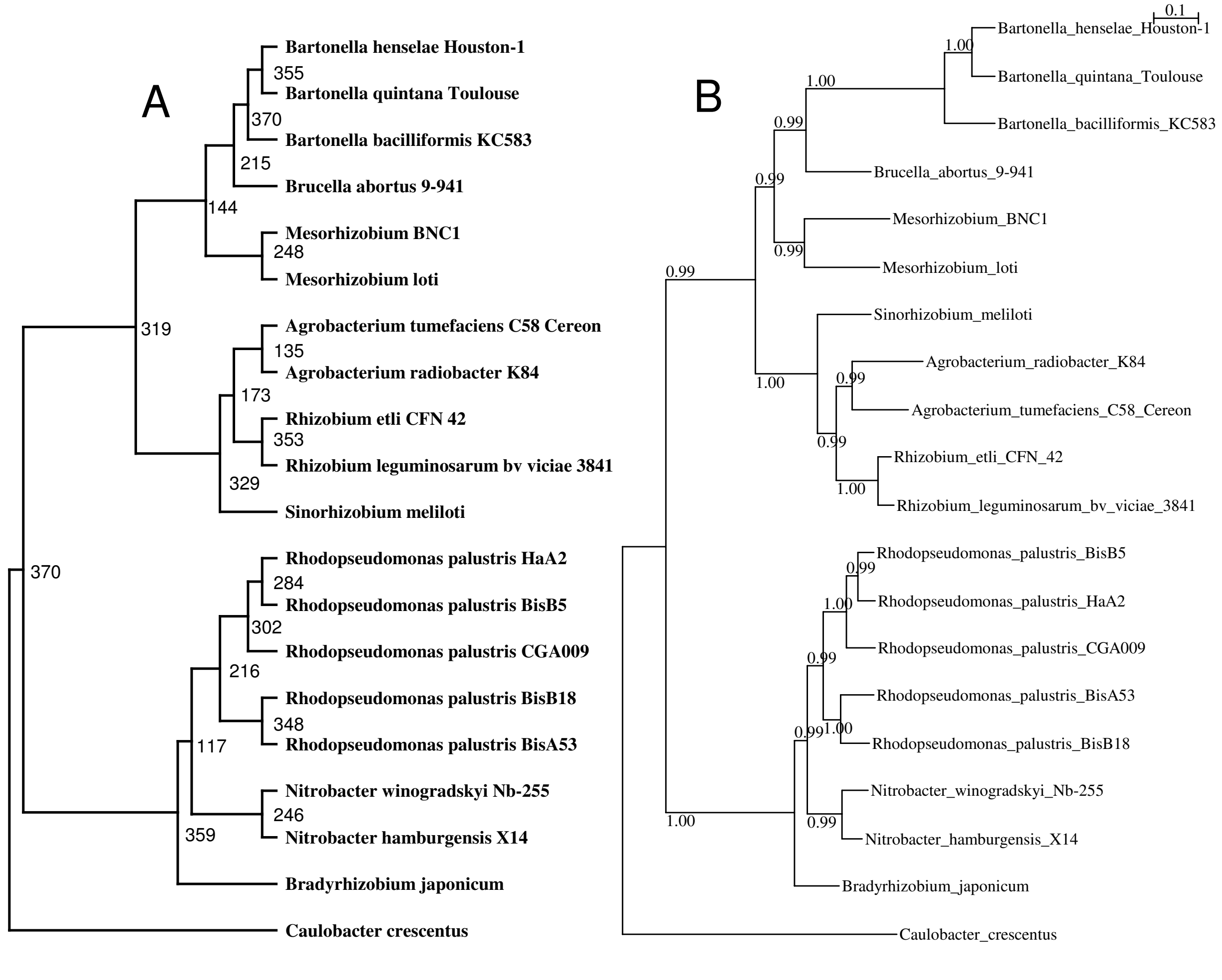

A
Bartonella henselae Houston-1
355
Bartonella quintana Toulouse
370
Bartonella bacilliformis KC583
215
Brucella abortus 9-941
144
Mesorhizobium BNC1
248
Mesorhizobium loti
319
Agrobacterium tumefaciens C58 Cereon
135
Agrobacterium radiobacter K84
173
Rhizobium etli CFN 42
353
Rhizobium leguminosarum bv viciae 3841
329
Sinorhizobium meliloti
370
Rhodopseudomonas palustris HaA2
284
Rhodopseudomonas palustris BisB5
302
Rhodopseudomonas palustris CGA009
216
Rhodopseudomonas palustris BisB18
348
Rhodopseudomonas palustris BisA53
117
Nitrobacter winogradskyi Nb-255
246
Nitrobacter hamburgensis X14
359
Bradyrhizobium japonicum
Caulobacter crescentus
B
0.1
Bartonella_henselae_Houston-1
1.00
Bartonella_quintana_Toulouse
1.00
Bartonella_bacilliformis_KC583
0.99
Brucella_abortus_9-941
0.99
Mesorhizobium_BNC1
0.99
Mesorhizobium_loti
0.99
Sinorhizobium_meliloti
Agrobacterium_radiobacter_K84
1.00
0.99
Agrobacterium_tumefaciens_C58_Cereon
0.99
Rhizobium_etli_CFN_42
1.00
Rhizobium_leguminosarum_bv_viciae_3841
Rhodopseudomonas_palustris_BisB5
0.99
Rhodopseudomonas_palustris_HaA2
1.00
Rhodopseudomonas_palustris_CGA009
0.99
Rhodopseudomonas_palustris_BisA53
0.99
1.00
Rhodopseudomonas_palustris_BisB18
Nitrobacter_winogradskyi_Nb-255
1.00
0.99
Nitrobacter_hamburgensis_X14
Bradyrhizobium_japonicum
Caulobacter_crescentus

Figure 2

The superalignment Bayesian phylogeny (SBP370) and the consensus tree (CT370), created from the 370 potential orthologous genes filtered from the larger data set. A: CT370, the numbers to the right of the internal branches indicate the number of orthologous genes that contain the group defined by that internal branch. B: SBP370, the numbers on the branches give the posterior probability of the group. The scale bar denotes the estimated number of amino acid substitutions per site.

the earlier analysis. This indicates that the majority of groups had E-values equal to or greater than 10e-12 (i.e. only 38 more groups were found when the E-values was lowered from 10e-12 to 10e-6). Because the difference between the use of E-values of 10e-6 and 10e-9 was only eight more groups, we further examined the former (38 rescued groups) using the filters described above. Of the 38 groups, 20 were ruled out by one or both of the filters. Thus, for the 38 groups that were picked up by an E-value of 10e-6 but not 10e-12, almost 50% were ruled out by the utilized filters. Notably, however, when both filters were applied, the percentage of rejection was almost equal for the data sets obtained using E-values of 10e-12 and 10e-6, with 370 out of 469 groups (79%) and 390 out of 507 groups (77%), respectively, passing both filters.

The identified orthologous genes had good phylogenetic content and substantial support

We used likelihood mapping analysis to analyze the phylogenetic content of the data set (see methods). Recognizing that a data set provides phylogenetic signal if it contains a high percentage of resolved quartets [14], we first determined the percentage of resolved quartets for each gene. The mean value of resolved quartets for all orthologous genes was 90.9% [standard error (SE), 1.32%; mode, 91.5%]. Even if all quartets are completely resolved, it is possible that the quartet-puzzling tree is not completely resolved when the quartets are not compatible with each other [14]. In our data set, only 82 orthologous groups presented a completely resolved puzzling tree; the groups yielding incompletely resolved puzzling trees comprised principally *B. japonicum* and the genus *Agrobacterium*. The superalignment had all quartets resolved and its puzzling tree was completely resolved. As a measure of support for our phylogenies, we calculated the median bootstrap value across the whole phylogeny, and then calculated the mean of the median values. The mean of the median values was 77 (SE = 4). These findings indicate that the identified orthologous genes had sufficient phylogenetic signal and substantial support.

Almost every orthologous gene had a unique topology, and the most common topology was not that of 370SBP

In order to evaluate the diversity of evolutionary histories among the orthologous genes, we determined the number of different topologies. Approximately 93% of the orthologous genes presented unique topologies, for a total of 346 different topologies. Only two orthologous genes, namely ATP-dependent Lon protease (COG0466) and DNA-directed RNA polymerase beta subunit (COG0085), yielded the SBP370 topology. Unexpectedly, the most frequent topology (shared by six orthologous genes) was that of 469CT.

Most bipartitions were in agreement with the 370SBP topology

In order to present a full account of phylogenic diversity, we examined the number of common bipartitions between the species tree and all the individual phylogenies. A bipartition represents the division of a phylogeny into two parts connected by a single internal branch; this divides the phylogeny into two groups but does not consider the relationships within each of the groups. The total number of different possible bipartitions for 20 taxa is 524,267; however, we only identified 254 different bipartitions in the individual phylogenies examined in the present study. The majority of bipartitions were in agreement with the 370SBP topology (71.5% of all observed bipartitions did not contradict this topology). Both 370CT and 370SBP yielded the same topology, thus they also shared the same bipartitions. Subsequently, 370CT reflected the frequencies of the bipartitions of 370SBP for the individual phylogenies. The frequencies of those bipartitions were not evenly distributed. There were only two cases where the nodes or bipartitions were supported by all of the orthologous gene trees. The separation of *Caulobacter crescentus* from the rest of the species represented one of these, while the other was the segregation of genus *Bartonella* from the other species (see Figure 2a). The two least frequently encountered bipartitions defined the genus *Agrobacterium* (supported by 135 phylogenies; see Figure 2a), and the group formed by *Rhodopseudomonas* and *Nitrobacter* but excluding *B. japonicum* (supported only 117 phylogenies). In addition, the branches that defined these two bipartitions/groups in 370SBP represented the second and third shortest branches across the whole phylogeny. Next, to estimate the similarities between each orthologous gene tree and the best estimate of the species tree, we calculated the percentage of common bipartitions between each orthologous gene tree and 370SBP (see methods). More than 90% of the 370 orthologous gene trees had more than 50% of their bipartitions in common with 370SBP. The mean percentage of common bipartitions among all orthologous genes was 71.76% and the mode was 76%. Thus on average, more than 70% of the bipartitions in each orthologous gene tree were also present in 370SBP.

The larger the gene size, the higher the percentage of bipartitions in common; as a branch in the species tree grew larger, its bipartition frequency increased

To assess whether the stochastic error related to gene length affected the percentage of bipartitions in common, we tested for correlation between gene size and the percentage of common bipartitions. We found a weak but significant correlation ($p < 0.00001$; coefficients of correlation and determination, 0.39 and 0.15, respectively). This suggests that longer genes shared a higher percentage of bipartitions in common with the species tree. We also determined the correlation between the number of phylogenies that supported a bipartition in 370CT and the length of that branch in 370SBP. The coefficient of correlation was 0.624 and the coefficient of determination was 0.384 ($p < 0.01$), suggesting that longer branches defined groups (bipartitions) among a greater number of orthologous gene trees.

Multiple best-fit protein models were selected

We then used the Akaike information criterion to allow each orthologous gene to select a model of protein evolution (see methods). The WAG matrix represented the most selected substitution model (selected by 58% of genes), followed by the JTT matrix (selected by around 18% of genes) (Figure 3a). Only four out of the eight selected models were chosen by more than 10 orthologous genes (Figure 3a). Although no single matrix was chosen for all genes, the preferred matrixes comprised a relatively small set. All of the orthologous genes had to be corrected for among-site rate variation. In addition, 68% also required correction concerning the frequencies of amino acids, and 40% were shown to have a proportion of invariable sites. To confirm that model selection improved our results, we examined the difference of the log likelihood values between the best and the worst models, according to the Akaike information criterion (where a high difference indicates an improvement). Approximately 56% and 85% of the genes showed differences higher than 1000 and 500, respectively, indicating that model selection improved our results (Figure 3b).

Orthologous genes were functionally biased

We used the COG database [15] to functionally categorize (see methods) the identified orthologous genes into the four broad categories of this database. The frequency distributions of the functional categories differed significantly between our data set and that of the COG database (chi-square test $p < 0.0005$), indicating that the identified orthologous genes were functionally biased. The most common category in our data set, comprising 34% of the identified genes, was that of "Information Storage and Processing;" in contrast, most common category throughout the COG database was the "Poorly Characterized" category, which comprised 40% of the database (Table 1). On the flip side, the least frequent category in the COG database was that of "Information Storage and Processing" (15% of genes), while that in our data set was the "Poorly Characterized" category (12% of genes) (Table 1). In order to analyze the congruence among these broad categories, superalignment Bayesian phylogenies and consensus trees were constructed for each category. The superalignment Bayesian phylogenies for all four categories indicated the 370SBP topology. The consensus trees obtained for the "Information Storage and Processing" and "Cellular Processes and Signaling" categories also indicated the 370SBP topology, whereas the consensus trees for the "Metabolism" and "Poorly Characterized" categories differed from one another and from the 370SBP topology. The consensus tree obtained for the "Metabolism" category revealed a topology identical to that of 469CT (Fig 1a), while that for the "Poorly Characterized" category manifested the same discordance and, in addition, the genus *Agrobacterium* did not form a monophyletic group. These two points of discrepancy contradicted the two least common bipartitions in 370CT (the ones defined by the shortest internal branches in 370SBP). The correlation between gene size and the percentage of common bipartitions differed among the functional categories (Table 1); the "Poorly Characterized" category had the strongest correlation (coefficient of correlation, 0.49), while "Metabolism" had the weakest (non-significant) correlation (0.15, $p = 0.072$) (Table 1).

There was a wide variation in total phylogeny length

To test for variation in the level of divergence within the set of orthologous genes, the total phylogeny length was determined for each individual phylogeny (see methods). We observed significant variation among the total lengths of the phylogenies (coefficient of variation, 53%), with a mean total length of 5.2 expected substitutions per site per phylogeny. Most of the phylogenies had between four and six expected substitutions per site per phylogeny (around 28%), followed by those having between two and four expected substitutions per site per phylogeny (almost 27%) (Figure 4, blue bars). When we tested whether the level of divergence was the same among the functional categories, we found significant differences among the categories (Kruskal-Wallis test, $p < 0.0005$). The "Poorly Characterized" category had the most diverged orthologous genes, with most genes (28%) having six to eight expected substitutions per site per phylogeny (Figure 4, purple bars). In contrast, the "Information Storage and Processing" and "Metabolism" categories had the least diverged orthologous genes, most of which fell into the range of between two and four expected substitutions per site per phylogeny (Figure 4, red and green bars, respectively). These observations suggest that the divergence of orthologous genes in this species appears to vary by functional class.

Discussion

In this study, our goal was to test whether orthologous genes reflect the history of the species. To answer this question, we selected a monophyletic group having moderate phylogenetic distances (allowing us to make a reliable phylogenetic inference). We obtained an initial data set of possible orthologous genes using the reciprocal RBH technique, and further used two filters to infer gene orthology, thereby avoiding the inadvertent inclusion of DSDL and/or HGT. These filters excluded more than 20%

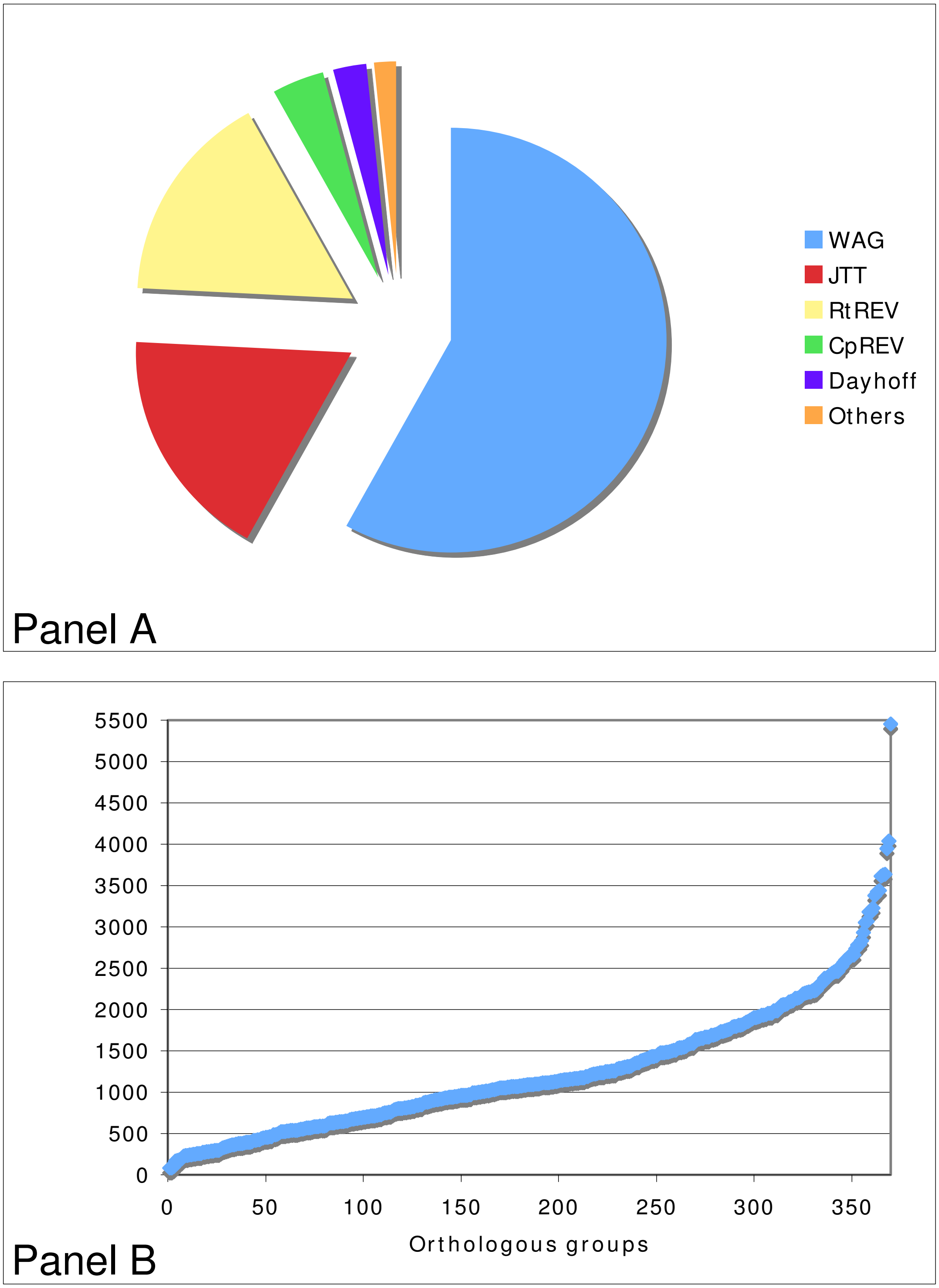


Figure 3

The models of evolution selected by the orthologous genes. The Akaike criterion of information was used to select the models of evolution. A: The different amino acid matrices selected. B: Difference between the best and worst models. The genes were ordered from lowest to highest with regard to the differences in the log likelihood values. Differences are positive because the worst models are more negative.

Table 1: Functional classification and percentage of genes in each functional category

Category	This data set	COG database	Stochastic error
Information storage and processing	34%	15%	0.36
Cellular processes and signaling	24%	18%	0.37
Metabolism	31%	28%	0.15*
Poorly characterized	12%	40%	0.49

Percentage of orthologous genes and COG database genes distributed across the four broad functional categories. The fourth column shows the correlation between gene size and the percentage of common bipartitions for each category. * This correlation was not significant.

of the initial data. The superalignment phylogenies and the consensus tree did not agree with one another when the initial data set was used. Once the false positives had been ruled out, however, both approaches produced the same tree. Thus, our results suggest that approaches using only computational definitions of orthology (e.g. RBH or SGF) can produce a considerable number of false positives, which contribute to disagreements among phylogenetic results.

As in other studies involving recently evolved groups [4,9,10], we found that orthologous genes had very good phylogenetic content. This set presented a strong vertical signal, indicated by the fact that both approaches used to infer the possible species trees revealed the same topology. Moreover, around 71% of the total bipartitions agreed with the inferred species trees. Therefore as a whole, our selected orthologous genes had a very strong vertical signal and manifested a tree-like organismal his-

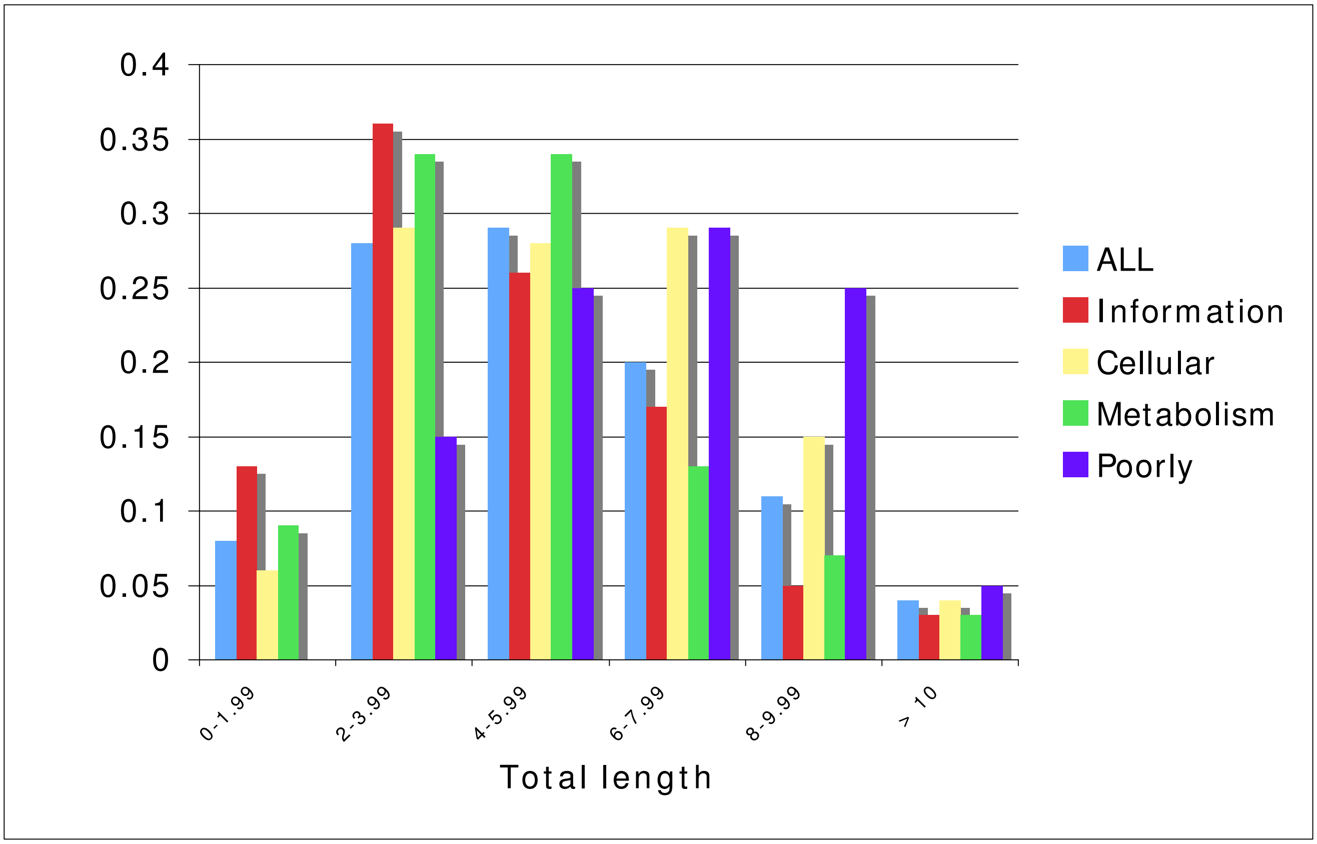


Figure 4

Total phylogeny lengths. As a measure of divergence we used the total phylogeny length, which is expressed as the estimated number of substitutions per site per phylogeny. This analysis was undertaken across the whole confidence set of orthologous genes (blue bars) as well as for the genes divided into four broad categories. Abbreviations are as follows: Information (red bars), the "Information Storage and Processing" category; Cellular (yellow bars), the "Cellular Processes and Signaling" category; Metabolism (green bars), the "Metabolism" category; Poorly (purple bars), the "Poorly Characterized" category.

tory. Other studies in *Alphaproteobacteria* and *Gammaproteobacteria* [4,9-12] reached the same conclusion when gene families were considered as a whole (e.g. superalignment and/or supertrees). In a recent study, a robust phylogeny for the *Alphaproteobacteria* was inferred [12], and the relationships revealed for the *Rhizobiales* group were equivalent to the species tree inferred in this study. We found a great diversity of topologies, many of which were well supported. Almost every orthologous gene revealed a distinct topology, yielding 346 different topologies. This is consistent with the findings of the previous study in *Alphaproteobacteria* [12], wherein none of the topologies from the individual genes were found to be equivalent. Furthermore, we found diversity not only in topology, but also in the models of protein evolution chosen by each of the orthologous genes. Eight amino acid substitution matrices were chosen, but only 4 had a frequency exceeding 10 genes. As in the previous study that identified a robust species tree for *Alphaproteobacteria* [12], the most frequent matrix identified among the orthologous genes was the WAG amino acid substitution matrix. The site rate variation and correction for amino acid frequency inequality parameters were strong performers in our study; all of the phylogenies described herein were based on models that accounted for site rate variation, and up to 68% of the phylogenies were corrected for inequalities in amino acid frequency.

The most frequently found topology differed from the species tree, although the only difference was the position of *B. japonicum*, which corresponded to one of the shortest branches in the best estimate of the species tree. Furthermore, only two orthologous genes yielded the species tree topology. These findings are in accordance with similar findings from other reports [11,12]. For instance, within *Gammaproteobacteria* only three out of 200 genes had the same topology as the reference tree [11]. These findings collectively suggest that most orthologous genes do not reflect the exact species tree when used as individual markers, and the most common topology can differ from the species tree. This is a significant point, because it suggests that even at moderate phylogenetic distances (where phylogenetic inference is reliable when adequately performed), neither a single orthologous gene nor the most common topology can be used to reconstruct the exact history of the species. However, even though only two of the orthologous gene trees manifested the species tree topology, all the orthologous gene trees together shared an average of 71% of their bipartitions with the species tree. Thus, the majority of bipartitions composing the orthologous gene trees in this study reflected a large part of the species history. Nevertheless, individual orthologous gene trees were not evenly distributed with regard to the species tree bipartitions. Only the genus *Bartonella* and the separation of the ingroup from outgroup occurred in all of the individual phylogenies. The two bipartitions that showed the smallest representation among the individual phylogenies were two of the three shortest internal branches in the species tree (see Figure 2b), and involved the placements of *B. japonicum* and the genus *Agrobacterium*.

Gene length also emerged as a factor influencing discordance in our study, both at the level of the species tree and for the single orthologous gene trees. Even though the correlation was weak, the phylogenies of longer genes had more bipartitions in common with the species tree. This agrees with a recent study analyzing *Alphaproteobacteria*, which concluded that part of the problem with inferring phylogenies from individual genes resulted from insufficient information content, due to the short length of the genes [12]. Notably, however, this correlation was not equal across all functional categories; the "Poorly Characterized" category presented the strongest correlation, while the "Metabolism" category did not show significant correlation. This suggests that, where possible, it is better to choose longer orthologous genes from the "Poorly Characterized" category. On the other hand, a stronger correlation was found between bipartition frequency among the individual phylogenies and the internal branches of the species tree (where the least common bipartitions were defined by the shortest internal branches). Therefore, as more changes accumulate in a branch that defines a group in the species tree, more of the individual orthologous gene trees will reflect this group. This implies that even for species trees, special attention should be paid to the shortest internal branches, which will tend to be more problematic.

There are several non-biological causes that could cause discordance, such as imperfect sequence alignment, stochastic error related to gene length (discussed above), and model violations. We feel that model violations are not the main source of incongruence in the present study, because each orthologous gene was allowed to indicate its own model of evolution and the phylogenies were constructed using models that accounted for site rate variation and (where necessary) corrected for amino acid frequency inequalities.

Incomplete lineage sorting has been recognized as a biological factor that can lead to discordance when phylogenies are inferred from genes [16,17], particularly where the internal branches of the species tree are short enough so that coalescence of gene lineages may occur more deeply than the speciation event. Degnan and Rosenberg showed that very short branches deep in a species tree comprising five or more species can lead to anomalous genes trees (AGT), i.e. gene trees that do not match the species tree [16]. Furthermore, the most probable gene tree can have a different topology from that of the species tree if multiple branch lengths are small enough in coalescent units [16]. Two trends lead us to believe that incomplete lineage sorting is one of the main causes of discordance among the orthologous genes examined in the present study. First, the two least common bipartitions from the individual orthologous gene trees involve two out of three of the shortest internal branches in the species tree (Figure 2). Second, the most common topology was not that of the species tree, but it only differed from the species tree in terms of the position of *B. japonicum*, which involves precisely one of the very short, deep, internal branches discussed above. Indeed, when we used the COAL [18] software to determine the probability of the genes trees that had the most common topology and the genes trees that had the species tree topology, given our best estimate of the species tree, although all the genes with the most common topology got the same very low probability, which was 0.00000000001, the probability got by the genes with the species tree topology was 0.00000000000. Thus, the most common topology appears to be an example of AGT.

It is common for orthologous genes to broadly indicate the history of the species, without reflecting it exactly. In the present case, this is not related to signal erosion because most of the orthologous genes studied herein had good phylogenetic content. Instead, we think that the type of function fulfilled by each gene influenced its ability to recover the true tree. We found that the orthologous genes recovered by our analysis were functionally biased, with genes of the "Information Storage and Processing" category representing 34% of the orthologous genes (as compared to 15% of the COG database), while the "Poorly Characterized" category represented only 12% of the orthologous genes (compared to 40% in the COG database). Furthermore, the level of divergence paralleled the functional bias, as the categories containing more orthologous genes were less diverged. The most diverged category was that of the "Poorly Characterized" genes, which contained a very few highly diverged orthologous genes and yielded a consensus tree that differed considerably from the species tree. To a certain extent, this aspect of functional restriction also relates to the discordance caused by incomplete lineage sorting. The "Metabolism" and "Poorly Characterized" categories were the most affected by incomplete lineage sorting, as their consensus trees differed precisely in those branches where incomplete lineage sorting was a factor. The "Poorly Characterized" category contained the most diverged orthologous genes, and was the most adversely affected by lineage sorting. It is reasonable to deduce that weak functional restrictions may have allowed this. Following the same logic, a category with highly conserved (i.e. functionally restricted) genes should be less affected by incomplete lineage sorting, as seen for the "Information Storage and Processing" category.

In conclusion, we observed that orthologous genes exhibited a great diversity of phylogenies, having different best-fit models of evolution, topologies, and degrees of divergence. Thus, almost no single orthologous gene by itself can reflect the exact history of the species. Notably, the most frequent topology did not match the species tree. Orthologous genes were affected by stochastic error relating to gene size, although this effect was relatively weak and was not evenly distributed across the functional categories. The most problematic clades were those defined by short internal branches, as these suffered from the effects of incomplete lineage sorting. The extent of these effects depended on the functional restrictions of the orthologous genes; for example, the "Information Storage and Processing" category appeared to be refractory to this process, whereas the "Poorly Characterized" category was more highly affected. When we used as many markers as possible, however, we could achieve a good reconstruction of the species history. For instance, when we employed superalignment, even the "Poorly Characterized" category indicated the topology of the species tree. Thus, when taken as a complete set, orthologous genes have a great capacity for depicting the history of a species.

Methods

Genomes used

We used the complete proteomes of 25 *Alphaproteobacteria* (see additional file 1), including 24 belonging to the *Rhizobiales* order and the genome of *Caulobacter crescentus*, which was used as outgroup. All of the genomes, except those of *Agrobacterium radiobacter* K84 and *Agrobacterium vitis* S4, were downloaded in February of 2007 from the NCBI ftp site. Those of *Agrobacterium radiobacter* K84 and *Agrobacterium vitis* S4 were downloaded from Agrobacterium.org http://depts.washington.edu/agro/.

Defining orthologous groups

Using the RBH approach to identify possible orthologous groups

As *B. quintana* strain Toulouse has the smallest proteome out of all the species considered herein, this strain was used as the reference genome to establish an RBH approach. Each of the 1142 proteins of *B. quintana* strain Toulouse were compared with the proteomes of the other strains, using BLAST [19] with an E-value cutoff of < 1.0e-12. We retained all cases where a protein of *B. quintana* strain Toulouse had a bidirectional best hit in each of the other proteomes, and the proteins aligned along at least 50% of their lengths.

The above analysis yielded 469 groups (potential orthologs). Each of these possible orthologous groups were aligned using MUSCLE [20] with the default parameters. The best model of amino acid substitution for each alignment was determined using ProtTest [21], and the most likely phylogeny was constructed using PHYML [22] with 100 non-parametric bootstrap replicates. The gamma shape parameter and the proportions of invariable sites were estimated by maximizing the likelihood of the phylogeny. Likelihood mapping analysis was carried out to determine the phylogenetic content for every individual alignment, using PUZZLE [14,23].

Excluding redundant species

In our preliminary analyses, we noted that the genera *Agrobacterium* and *Brucella* contained species that showed minimal divergence. As a result, many possible orthologous groups manifested identical protein sequences for certain species belonging to *Agrobacterium* and/or *Brucella*. In order to exclude redundant species, we used PUZZLE to establish maximum likelihood matrices for the 469 alignments, taking into account among-site rate variation [14,23]. We then took the mean of the maximum likelihood distance between any two species; if two species had a mean distance equal to or less than 0.05, one of these was excluded. Five species were removed (marked with asterisks in additional file 1). We then established new alignments, model selection, and phylogenies without the excluded species.

Ruling out false positives

Two filters were used to eliminate false positives. The first filter consisted of using confidence sets to assess whether the differences in topology between the probable species trees (see below) and individual gene trees exceeded those expected to occur by chance. We used expected likelihood weighting [23], which provides a simple and intuitive method for making multiple comparisons of models and constructing corresponding confidence sets. This test has the benefit of being less conservative than the SH test [23]. The topologies tested included the superalignment Bayesian topology and the consensus tree topology (see below). PUZZLE was used to carry out this test for each of the 469 alignments, as well as for the superalignment (see below). The 469SBP typology (see Figure 1b) contained a sister group relationship between the group comprising *Sinorhizobium meliloti*, *Brucella abortus* 9–941 and the genera *Rhizobium*, *Agrobacterium*, *Mesorhizobium*, and *Bartonella* and that comprising the genera *Rhodopseudomonas*, *Nitrobacter*, and *Bradyrhizobium japonicum*. The presence of this sister group relationship was used as the second filter; we used PAUP* 4.01 b10 [24] to see whether each of the 432 potential orthologous genes that passed the first filter had phylogenies manifesting the two sister groups. We then used likelihood mapping analysis (applied through PUZZLE) to determine the phylogenetic content for each of the remaining orthologous genes; the number of resolved quartets was counted for each gene, and then a mean and SE were calculated for the entire set.

Two approaches for establishing a probable species tree

Superalignment approach

A superalignment was created by concatenating the 469 individual alignments. Two phylogenies were derived. The first was undertaken with maximum parsimony, using PAUP* 4.01 b10 [24] with random addition of sequences and tree bisection reconnection. The second phylogeny was created using MrBayes v3.1.2 [25], allowing the MCMC sampler to explore all of the fixed-rated amino acid models included in MrBayes. The number of rate categories for gamma distributions was set to four, with an allowance for a proportion of sites to be invariable. Due to the computational burden, we performed a single run with four chains, for 500,000 generations. Trees were sampled every 500 generations, 25% of all generations were removed as burn-in, and a consensus was taken. Once the candidate orthologous genes had been filtered for removal of false positives, we generated a second Bayesian phylogeny from the remaining 370 genes, using the same specifications as above. Because we ran only one run, for each Bayesian phylogeny, we could not use the standard deviation of the split frequencies, instead we examined the log likelihood values. For both superalignments, these values stabilized very soon and started to fluctuate within a very narrow range. In additional file 2 we plotted the log likelihood values of the second phylogeny.

Consensus tree approach

A consensus tree was created from all 469 phylogenies using CONSENSE [26]. Once the candidate orthologous genes had been filtered for removal of false positives, we generated a second consensus tree from the remaining 370 genes.

Topologies and bipartitions

The number of different topologies for the confidence set of orthologous groups was deduced using the Robinson and Fould distance (RFd), as calculated through application of TREEDIST [26]. The RFd indicates the number of bipartitions that are unique to one of two phylogenies being compared; the RFd equals zero when the two phylogenies have the same topology. The number and proportion of total bipartitions were determined using an *ad hoc* perl script that is based on inputting the consensus file generated from CONSENSE [26].

Percentage of bipartitions in common between the 370SBP and each individual phylogeny

We calculated the RFd between each individual phylogeny and the species tree and used it to determine the percentage of shared bipartitions. Each phylogeny had 17 bipartitions, and two phylogenies were considered in each comparison, for a total of 34 bipartitions in each comparison. The RFd reflected the number of bipartitions that were unmatched within the data set. For example, an RFd of four indicated that 30 bipartitions were shared. In order to establish the percentage of common bipartitions for each phylogeny, the number of shared bipartitions was divided by two, because two phylogenies were being considered. In our example this would be 30/2, which equals 15. Thus, 15 out of 17 (88%) of the bipartitions would be common to the two phylogenies. Therefore, the formula for establishing the percentage of common bipartitions is as follows:

Percentage of common bipartitions = ((34-RFd)/2) × 100

Functional assignment

We used the COG database [15] to undertake functional annotation across the four broad categories of "Information Storage and Processing," "Cellular Processes and Signaling," "Metabolism," and "Poorly Characterized." A few orthologous genes that had not been functionally assigned within the COG database were placed in the "Poorly Characterized" category. We excluded all orthologous genes that belonged to two or more broad categories. We chose this method because broad classification is less prone to error.

Authors' contributions

SC-R conceived, designed, and performed the experiments. SC-R analyzed the data and wrote the manuscript. VG contributed materials and edited the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Genomes used.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-300-S1.doc]

Additional file 2

The log likelihood values of the 370SBP.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-300-S2.ppt]

Acknowledgements

We thank Luis Lozano, Bruno Contreras "El Tronco", Scott V. Edwards, and Pablo Vinuesa for their comments on the manuscript. SC-R acknowledges a PhD fellowship from CONACyT, and thanks Pablo Vinuesa for helpful discussions. Gratitude is also extended to Dr. Joao Setubal for allowing the use of the *A. vitis* S4 and *A. radiobacter* K84 genome sequences prior to publication. This work was supported by CONACYT grants U46333-Q and IN223005 from PAPIIT-UNAM. SC-R also thanks the Computational Unit of the Instituto de Biotecnologia, particularly to Jerome Verleyen, and the "Macroproyecto de Tecnologias de la Informacion y la Computacion de la UNAM" for the use of their computer facilities.

References

1. Fitch WM: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, **19(2):**99-113.
2. Fitch WM: **Homology a personal view on some of the problems.** *Trends Genet* 2000, **16(5):**227-31.
3. Bapteste E, *et al.*: **Do orthologous gene phylogenies really support tree-thinking?** *BMC Evol Biol* 2005, **5(1):**33.
4. Fitzpatrick DA, Creevey CJ, McInerney JO: **Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales.** *Mol Biol Evol* 2006, **23(1):**74-85.
5. Susko E, *et al.*: **Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the gamma-proteobacteria.** *Mol Biol Evol* 2006, **23(5):**1019-30.
6. Zhaxybayeva O, Lapierre P, Gogarten JP: **Genome mosaicism and organismal lineages.** *Trends Genet* 2004, **20(5):**254-60.
7. Kelchner SA, Thomas MA: **Model use in phylogenetics: nine key questions.** *Trends Ecol Evol* 2007, **22(2):**87-94.
8. Keane TM, *et al.*: **Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified.** *BMC Evol Biol* 2006, **6:**29.
9. Creevey CJ, *et al.*: **Does a tree-like phylogeny only exist at the tips in the prokaryotes?** *Proc Biol Sci* 2004, **271(1557):**2551-8.
10. Lerat E, Daubin V, Moran NA: **From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria.** *PLoS Biol* 2003, **1(1):**E19.
11. Comas I, Moya A, Gonzalez-Candelas F: **From phylogenetics to phylogenomics: the evolutionary relationships of insect endosymbiotic gamma-Proteobacteria as a test case.** *Syst Biol* 2007, **56(1):**1-16.
12. Williams KP, Sobral BW, Dickerman AW: **A robust species tree for the alphaproteobacteria.** *J Bacteriol* 2007, **189(13):**4578-86.
13. Battistuzzi FU, Feijao A, Hedges SB: **A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land.** *BMC Evol Biol* 2004, **4:**44.
14. Strimmer K, von Haeseler A: **Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment.** *Proc Natl Acad Sci USA* 1997, **94(13):**6815-9.
15. Tatusov RL, *et al.*: **The COG database: a tool for genome-scale analysis of protein functions and evolution.** *Nucleic Acids Res* 2000, **28(1):**33-6.
16. Degnan JH, Rosenberg NA: **Discordance of species trees with their most likely gene trees.** *PLoS Genet* 2006, **2(5):**e68.
17. Kubatko LS, Degnan JH: **Inconsistency of phylogenetic estimates from concatenated data under coalescence.** *Syst Biol* 2007, **56(1):**17-24.
18. Degnan JH, Salter LA: **Gene tree distributions under the coalescent process.** *Evolution* 2005, **59(1):**24-37.
19. Altschul SF, *et al.*: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17):**3389-402.
20. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5:**113.
21. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics* 2005, **21(9):**2104-5.
22. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52(5):**696-704.
23. Strimmer K, Rambaut A: **Inferring confidence sets of possibly misspecified gene trees.** *Proc Biol Sci* 2002, **269(1487):**137-42.
24. Swofford DL: **PAUP*: phylogenetic analysis using parsimony (*and other methods).** Sinauer Associates, Mass; 1998.
25. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19(12):**1572-4.
26. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6.** Department of Genome Sciences, University of Washington, Seattle; 2005.

Horizontal gene transfer and diverse functional constrains within a common replication-partitioning system in *Alphaproteobacteria*: the *repABC* operon

Santiago Castillo-Ramírez*, Jorge F. Vázquez-Castellanos, Víctor González, and Miguel A. Cevallos*
Address: Programa de Genómica Evolutiva, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Apartado Postal 565-A, CP 62210, Cuernavaca, Morelos, México.
*Corresponding authors: iago@ccg.unam.mx and mac@ccg.unam.mx.

Abstract

Background: The *repABC* plasmid family, which is extensively present within *Alphaproteobacteria*, and some secondary chromosomes of the *Rhizobiales* have the particular feature that all the elements involved in replication and partitioning reside within one transcriptional unit, the *repABC* operon. Given the functional interactions among the elements of the *repABC* operon, and the fact that they all reside in the same operon, a common evolutionary history would be expected if the entire operon had been horizontally transferred. Here, we tested whether there is a common evolutionary history within the *repABC* operon. We further examined different incompatibility groups in terms of their differentiation and degree of adaptation to their host.

Results: We did not find a single evolutionary history within the *repABC* operon. Each protein had a particular phylogeny, horizontal gene transfer events of the individual genes within the operon were detected, and different functional constraints were found within and between the Rep proteins. When different *repABC* operons coexisted in the same genome, they were well differentiated from one another. Finally, we found different levels of adaptation to the host genome within and between *repABC* operons coexisting in the same species.

Conclusions: Horizontal gene transfer with conservation of the *repABC* operon structure provides a highly dynamic operon in which each member of this operon has its own evolutionary dynamics. In addition, it seems that different incompatibility groups present in the same species have different degrees of adaptation to their host genomes, in proportion to the amount of time the incompatibility group has coexisted with the host genome.

Background

The *repABC* plasmids are a typical genome component of many *Alphaproteobacteria* species. In fact, more than 20 *Alphaproteobacteria* species have at least one *repABC* plasmid (see refs [1, 2] for recent reviews), these *repABC* plasmids may be the commonest plasmids in *Alphaproteobacteria* species. In some species these *repABC* plasmids constitute a significant amount of the bacterial genome; such is the case of *Rhizobium leguminosarum* 3841, in which *repABC* plasmids account for 35% of the genome [3]. This plasmid family includes several incompatibility groups, meaning that more than one type of *repABC* plasmid can reside in the same bacterial species [1, 2]. For instance, *Rhizobium etli* CFN42 has 6 plasmids, all of them *repABC* plasmids [4]. In contrast to other low copy-number plasmids, in which the elements involved in plasmid replication and segregation are located on different loci (each one under its own regulatory circuit), the *repABC* plasmids contain all the elements required for replication and partition within the *repABC* operon. In general, this transcriptional unit comprises three protein-encoding genes (*repA*, *repB,* and *repC*) and a gene encoding a small antisense RNA (ctRNA) [5], which is located within the *repB-repC* intergenic region. The proteins encoded in the *repABC* operon have an intricate relationship, with RepA and RepB interacting both with themselves and with each other. These proteins, in conjunction with the centromere-like sequence, *parS*, function as the plasmid's segregation machinery [1, 2, 6]. On one hand, RepA is a transcriptional repressor of the operon, while RepB acts as its co-repressor by contacting the operator sequence. The third protein-encoding gene of the operon, *repC*, is essential for plasmid replication; it encodes the initiator protein, RepC, which exerts its function by binding the origin of replication located within its own coding sequence [1, 2, 6]. Taking these observations into account, it is reasonable to hypothesize that the *repABC* operon is under concerted evolutionary pressures aimed at maintaining its functionality and avoiding incompatibility with other *repABC* operons. Remarkably, this operon is not only the replication system of *repABC* plasmids, but of some secondary chromosomes of some *Rhizobiales* species. For instance, the second chromosomes of *Agrobacterium* vitis S4 and *Agrobacterium tumefaciens* C58 have a *repABC* origin of replication [7].

At the structural level, the various *repABC* operons are only superficially homogeneous; they are highly diverse in DNA sequence, and some possess specific structural elements shared only by few members of the family. These distinctive elements fall into three types: (a) the number and class of regulatory elements involved in operon transcription; (b) the number and position of centromere-like sequences (*parS* sequences); and (c) the presence of peptide-encoding minigenes [1]. Several *Alphaproteobacteria* genomes possess *repAB* genes that are not in close association with the ctRNA or *repC* sequences. However, it has been shown that replication of some *Alphaproteobacteria* plasmids depends only on RepC and a ctRNA, without the involvement of the *repAB* genes. This suggests that fusion of different modules could participate in the generation of new *repABC* plasmids, indicating that the different elements may have experienced different evolutionary histories.

Plasmid stability requires an exquisite balance among all of the interacting molecules involved in plasmid replication and segregation. Perturbation of this balance, for example by the introduction of any replication or segregation element in excess, could lead to plasmid incompatibility. It has been shown that *repABC* plasmids contain at least four elements involved in plasmid incompatibility: the RepA and RepB proteins, the small antisense RNA, and the *parS* sequences [6, 8-10]. Phylogenies made with RepA, RepB, and RepC proteins have shown that different replicons residing in the same bacterial strain tend to belong to different clades [11]. Other study found that phylogenetic analyses of *repABC* gene lineages had a lack of evolutionary congruence with the species tree [7]. These observations suggest that divergent evolution followed by episodes of horizontal transfer have played a central role in originating new incompatibility groups. We might therefore expect that incompatibility groups residing in the same genome would be different enough so as to not interfere with each other.

In this study, we analyzed three aspects of *repABC* operons. First, because it is known that *repABC* operon has been horizontally transferred, through phylogenetic analyses, we examined horizontal gene transfer of entire operon versus horizontal transfer of individual genes within this operon. This is a key point, since a previous study has shown that some bacterial operons present horizontal gene transfer events that affect not the entire operons but single genes within the operons [12]. Second, we determined the degree of differentiation among *repABC* operons from different plasmids residing in the same strain (which implies different incompatibility groups). Third, we established the degree of evolutionary adaptedness among different *repABC* operons coexisting in a single species. In principle, because all the elements of the partition and replication systems are contained in the same operon and the encoded proteins interact, these elements might be expected to present almost the same history. Contrary to this, we found significantly different histories for the various elements of the *repABC* operon. Moreover, we detected different selective constraints among the elements composing the operon, and even within individual components. As expected, when different incompatibility groups coexisted in a species, these groups were clearly differentiated from one another. Finally, we found different levels of adaptation to the host genome within and between *repABC* operons coexisting in the same species.

Results

The collection of homologous *repABC* operons

To date, at least 81 *repABC* operons have been recognized across the class *Alphaproteobacteria* [1]. Because we wanted to utilize only homologous groups with the same domain structure, we established strict criteria for defining homologous *rep* genes and operons (see Methods). As a result of this, we analyzed only 49 operons herein (see Additional file 1). Twenty-one genomes had at least one *repABC* operon, and most of the operons were located on plasmids. A few genomes, such as those from genera *Brucella* and *Agrobacterium*, had *repABC* operons located on replicons that are considered secondary chromosomes (see Additional file 1). Two *Rhizobium* species, *R. etli* CFN42 and *R. leguminosarum* 3841, had the highest number of *repABC* operons, with seven operons each. All plasmids from these species had a single operon, with the exceptions of plasmid p42f from *R. etli* CFN42 and plasmid pRL11 from *R. leguminosarum* 3841, which each had two operons per plasmid. We also found six faulty operons that were missing one of the three protein-encoding genes; five out of six were composed of *repA* and *repB* genes, while the remaining one consisted of *repA* and *repC*. In four of six cases, the faulty operons coexisted with complete operons. In many species only one gene was present; by far the most widely distributed gene was *repA*, followed by *repC* (see Additional file 1).

There is no a single history for the *repABC* operon

Our first goal was to test whether the elements of the *repABC* operon have a common evolutionary history. A single history would be expected if the entire operon had been transferred; on the opposite, if the individual genes were transferred, several histories would be expected. Given that the partition and replication elements functionally interact with each other and compose a single transcriptional unit, we expected to find a single history *a priori*. To test this possibility, we constructed individual Bayesian phylogenies for each protein, and used the phylogenies to construct a strict consensus tree. We obtained phylogenies with strong support, but no two phylogenies gave the same topology (see Figure 1). For example, when we considered the phylogenies for RepA and RepC, only five nodes out of 40 achieved a posterior probability below 0.95 (see Figure 1). There was a large degree of conflict among the individual phylogenies, as demonstrated by the fact that the strict consensus tree had many polytomies and was poorly resolved (Figure 2). Only 25% of the nodes composing each phylogeny were shared among the three phylogenies. Since confidence sets of genes trees have been used to compare competing gene trees [13], we used this method to examine whether the differences among the phylogenies of RepA, RepB, and RepC were more than would be expected by chance (see Methods). The individual alignments of each protein rejected all but its own phylogeny, indicating that the phylogenies of the different proteins were significantly different from each other. Therefore, horizontal gene transfer has affected the individual genes within the operon. Actually each gene has had many unique horizontal gene transfer events since protein alignments rejected all but its own phylogeny. Here we will describe the positions of a couple plasmids in the rep phylogenies to make this clear. First example, the proteins coded by genes of the *repABC* operon located on plasmid pXAUT01 of *Xanthobacter autotrophicus* occupy drastically different positions in all 3 phylogenies (see Figure 1, green arrows). Actually, in each rep phylogeny pXAUT01 clusters with distinct groups, with very good support in every case. In other example, the horizontal gene transfer has affected either 2 of the genes or one gene, as 2 phylogenies agree while the third disagrees; for example, whereas the plasmids pSymA and pSMED02 cluster with pOANT01, in RepA and RepB phylogenies, plasmid pOANT01 does not cluster with the other 2 in RepC phylogeny (see Figure 1, red squares). The horizontal gene transfer events that have affected the *rep* genes are very particular, as they did not disrupt the operon structure. Gene displacement *in situ* is the most probable process behind this observation given that the operon is conserved in all the cases. As expected, the phylogenies for RepA and RepB, whose genes are next to each other, were more similar to one another than to RepC, as the Robison-Fould distance (a metric used to compare phylogenies, in which increasing distance indicates increasing disparity between phylogenies) between the phylogenies of RepA and RepB was smaller than that between RepC and either RepA or RepB (see Additional file 2). Since the evolutionary distance within the RepA, RepB, and RepC phylogenies is not that vast (see Figure 1), we checked if *in situ* gene displacement occurred by means of homologous recombination. To see if this might be the case here, we performed recombination analyses on the DNA alignments. In all three genes we found evidence of recombination, pairwise identity plots of the localized recombination events are presented in Additional file 3. We identified one event for *repA*, two for *repB*, and up to four for *repC* (see Additional file 3). The above results suggest that *in situ* gene displacement within the operon, through homologous recombination, has affected the *repABC* operon.

Different levels of functional restriction within and between Rep proteins

The most common method for modeling the variation of evolutionary rates among sites is the gamma distribution. Its shape parameter, α, determines the extent of rate variation among sites; a small α represents extreme rate variation, while a large α value represent a minor variation in rate [14]. Given that the main reason for the heterogeneity of evolutionary rates among sites seems to be differences in their selective constraints (due to the functional and/or structural requirements of the gene), we herein used the shape parameter α as a proxy for the functional restriction of each studied protein. In addition, we used the total length of each phylogeny as a means to examine the level of protein conservation. Among the three studied proteins, RepA showed the lowest total phylogenetic length and the highest among-site rate variation (reflected through the smallest shape parameter α), indicating that RepA was the most conserved protein, and that it experienced the highest level of functional restriction. The confidence intervals of the total length of the RepA phylogenies did not overlap with those of the two other phylogenies (see Table 1). Interestingly, the among-site rate variation was not significantly different between RepA and RepC, but the among-site rate variations of these two proteins were significantly different from that of RepB (see Table 1, shape parameter α column). Therefore, although RepA was the most conserved protein, RepA and RepC had similar levels of functional restriction.

To assess functional restriction inside the proteins, we next identified domains using Pfam [15], and assigned substitutions rates for individual sites for each protein using a discrete-gamma distribution (see Methods). We found that different domains had different substitution rates. For instance, in RepA protein, the ATPase domain almost did not have positions with highest substitution rates (see Figure 3, dotted lines, family MipZ), whereas the nucleotide-binding domain did have positions with the highest substitution rates (see Figure 3, domain CbiA). Similarly, most of the sites in the ParB-like nuclease domain of RepB (see Figure 3, dotted lines, family ParBc) had substitution rates that were smaller than those of the plasmid partition family domain (see Figure 3, family RepB). Only one domain was identified for RepC, but the substitution rates of its sites varied (Additional file 4). Notably, whereas RepA (the most conserved protein) was affected by a recombination event within its more variable domain (see Figure 3, Recombination, upper panel), RepB seems to have been affected by recombination throughout its sequence (see Figure 3, Recombination, lower panel). Thus, we detected different levels of functional restriction not only between the studied proteins, but also within them.

Well differentiated incompatibility groups

We used *Rhizobium etli* CFN42 and *Rhizobium leguminosarum* 3841 to compare and contrast incompatibility groups, because these strains each harbored six *repABC* compatible plasmids (i.e., six incompatibility groups). We made four DNA alignments, one for each *rep* gene and one for the intergenic region between the *repB* and *repC* genes, which encodes a small antisense RNA gene that acts as a strong incompatibility factor. To evaluate the degree of distinction among the *rep* genes and intergenic region of the different incompatibility groups, we determined maximum likelihood matrices and then calculated the average distance over all possible pairs of sequences (see Methods). The genes and intergenic region could be clearly differentiated across the different plasmids (see Table 2). In agreement with our protein phylogenies, the *repA* and *repC* genes presented shorter average distances and higher proportions of invariant sites compared to *repB*. Notably, the intergenic region comprised the shortest distance, but did not have any invariant position (see Table 2). Moreover, this locus had the highest among-site rate variation, as reflected in the smallest shape parameter α (see Table 2). This suggests that the intergenic region is under higher functional restriction compared to the *rep* genes; this finding is compatible with the presence of the small antisense RNA-encoding sequence in the intergenic region. Neither the intergenic region nor the *rep* genes showed any evidence of recombination. These results suggest that there was a high degree of differentiation among the examined incompatibility groups.

Codon Adaptation Index as a measure of evolutionary adaptedness

The Codon Adaptation Index (CAI) is a simple measure of synonymous codon usage bias. This index uses a reference set of highly expressed genes to assess the relative merits of every codon, and then determines a score for the gene or genes in question based on the use frequency of all codons in that gene [16]. The CAI can be used to evaluate the extent to which selection has been effective in molding the pattern of codon usage [16], and compare the codon usage of foreign genes versus that of highly expressed native genes. Here, we used the CAI to assess the adaptation of the *repA*, *repB*, and *repC* genes to their host genomes. We first calculated the relative synonymous codon usage values of highly expressed native genes (those encoding ribosomal proteins from each species), and then used CAI to compare the codon usage of the *repA*, *repB*, and *repC* genes to those of the reference genes (see Methods). CAI values can range from 0 (reflecting equal use of synonymous codons) to 1 (reflecting the strongest bias, codon usage is equal to that in the reference ribosomal protein-encoding genes). We found a clear trend in the CAI values within and between the studied *repABC* operons. In general, *repA* genes had the highest CAI values, followed by *repB* genes (see Figure 4). The *repABC* operons located on different plasmids had different CAI values, with those located on plasmids appearing to be the newest (e.g., p42a and p42d in *R. etli* CFN42, see Discussion) having the smallest CAI values (see Figure 4, red circles). Notably, in plasmids harboring two *repABC* operons, one always failed to meet the abovementioned pattern of CAI stratification. For example, plasmid pRL7 from *R. leguminosarum* 3841 contained the pRL7.1 and pRL7.2 operons, and the former had a higher CAI value for *repB* than *repA* (see Figure 4, green squares). Given that the degree of codon bias in unicellular organisms correlates with the level of gene expression, our results suggest that *repA* is more highly expressed than the other two genes, and *repB* is expressed at a higher level than *repC*. Furthermore, it seems that the different operons have different levels of expression.

Discussion

The *repABC* operon is not only important because it is the replication-partition system of *repABC* plasmids, a common component of *Alphaproteobacteria* species, but because it is also the replication-partition system of some secondary chromosomes in *Alphaproteobacteria* species. Our present analyses functioned at two levels: within the *repABC* operon and between *repABC* operons in those cases where several *repABC* operons coexisted in the same genome. We did not find a single history within the *repABC* operon; clearly, each protein had its own phylogeny. This is somewhat surprising, since *repA*, *repB*, and *repC* form an operon, and it would seem that they should have similar histories if the entire operon had been horizontally transferred. Instead, even RepA and RepB, which compose the partition system and physically interact, had different phylogenies. This contrast with a recent work in which relaxase sequences were used as tools for classification of conjugative systems. In that study it was found that relaxases and the IV coupling proteins (T4CP), which map next to each other and belong to a minimal gene set that allows plasmid to be conjugally transmitted, evolve congruently for long periods of time [17]. Thus, it seems that compared with some elements of the transfer machinery the *repABC* replication-partition system is highly diverse.

Quite notably, every single gene of this operon presented evidence of horizontal gene transfer. *In situ* gene displacement is a likely process behind this, since the structure of the *repABC* operon is completely conserved. We think *in situ* gene displacement could have occurred through homologous recombination, as we found homologous recombination events across the 3 *rep* genes. Although *in situ* gene displacement appears unlikely, there is evidence that shows that this process is not that scarce. Omelchenko *et al* found that within the bacterial operons they had analyzed *in situ* gene displacement was a frequent event [12]. A striking difference between *in situ* gene displacement and other types of horizontal gene transfer events is that the former leaves intact the operon structure, so that, the operon is completely functional.

The proteins differed not only at the topological level, but also at the level of functional restriction. RepA and RepC, which belong to different systems, were under similar levels of functional restriction, suggesting that key elements of the partitioning and replication systems are under similar functional restrictions. In contrast, RepB had a very different level of functional restriction. We also found different levels of functional restrictions within proteins. For example, the ATPase domain of RepA (Figure 3, family MipZ), which forms a complex with the chromosome partitioning protein and is indispensable for partitioning, presented the lowest substitution rates. As well the only recombination event presented in *repA* did not affect the ATPase domain but a relatively unconserved part of the gene. Therefore, it seems that the different proteins, and even the different parts of the proteins themselves, are under different functional and/or structural constraints. Of the three genes studied, *repA* was the most conserved and might have the highest expression level. This is not unexpected, as RepA is known to have several functions, and its expression is required in both the presence and absence of partition, suggesting the need for high-level translation in order to maintain sufficient RepA levels. In contrast, *repC*, which is a replication initiator protein, had the lowest CAI values, perhaps due to the higher levels of homologous recombination in this gene (see below). Horizontal gene transfer could be very important in allowing the variability of this operon. Indeed, if horizontal gene transfer had not affected the genes within the operon, these genes would have to have a single evolutionary history. Instead, we found that the reverse was true. The proteins encoded in those genes not only presented different phylogenies, they also had different functional restrictions, even within the proteins themselves, and the CAI values differed among the genes. Given the presence of differences at several levels, it is very logical to think that horizontal gene transfer has unconnected the various portions of the operon, allowing each part to have a particular evolutionary history. In this way, genes with very different functional restrictions could be located next to each other, as seen for *repB* and *repA*.

The existence of multiple *repABC* operons located on different replicons in the same genome implies the presence of different incompatibility groups. We herein showed that when multiple *repABC* operons coexisted in the same genome, they were well differentiated from one another. We did not find evidence of homologous recombination in these cases; this is not unexpected, since homologous recombination would homogenize the sequences, meaning that the different groups would no longer be compatible with each other. The intergenic region, which encodes a small antisense RNA (a very important determinant for incompatibility), was highly conserved and found to be under high functional restriction, yet it did not have any invariant sites. Although this sequence has changed only minimally due to functional restrictions, it has still accumulated sufficient changes to allow the coexistence of the different incompatibility groups. In agreement with our within-operon analysis, *repA* and *repC* were highly conserved, with *repC* being the most highly conserved between operons (it had the smallest average distance). As mentioned above, *repC* also had the most homologous recombination events. This suggests that homologous recombination might be reducing the divergence of *repC*, potentially also explaining the low CAI values for this gene (homologous recombination would be erasing any improvement in the CAI values). In a report on the genome sequence of *R. leguminosarum*, Young and coworkers suggested that a recent recombination event had taken place, and divergence of RepC was not critical for plasmid compatibility [3]. Here, one of the recombination events detected in *repC* involved the sequence from pRL8, which is a plasmid of *R. leguminosarum* 3841.

Different *repABC* operons had distinct levels of adaptation to their host genome, with no two *repABC* operons presenting the same CAI values. We think that amelioration might be playing a role in the adaptation of *repABC* operons to their hosts. Plasmids p42a and p42d were suggested to be newly acquired plasmids based on their lower GC values, poor conservation, and poor functional connectivity with the rest of the genome [4]. These two plasmids had the worst CAI values, implying that they are not well adapted to their host's genome. In contrast, the operon from p42f, which appeared to be the oldest plasmid harbored within *R. etli* CFN42, had the highest CAI values, suggesting that this operon is highly adapted. These findings indicate that the longer a *repABC* operon coexists with its host genome, the more adapted the operon becomes. This may result in more effective replication and partitioning processes. As well plasmids, which had the most adapted operons, presented essential genes as well; for instance plasmids pRL11, pRL12, and pRL10, which all have essential genes [3], had the operons with higher CAI values than the rest of plasmid of *R. leguminosarum* 3841.

In summary, we herein report finding different histories and functional constraints within the *repABC* operon. In addition, when multiple *repABC* operons were present in the same genome, they had different levels of adaptedness to the host genome, and this seems to be related to the length of time each operon had been associated with the host genome. Finally, horizontal gene transfer with conservation of the operon structure provides a highly dynamic operon in which each member could have its own evolutionary dynamics.

Methods

Detection of homologous genes and operons

We first identified the homologous of the RepA, RepB, and RepC proteins across the known *Alphaproteobacteria* genomes (see Additional file 5). The RepA, RepB, and RepC proteins from symbiotic plasmids of *R. etli* CFN42 and *S. meliloti* 1021 were used as seeds, and were queried against the proteomes encoded by the other genomes (Additional file 5), using BLAST [18] with an E-value cutoff of 1.0e-12. We retained all cases where a seed protein had a hit in any other proteome and the proteins aligned along at least 70% of their lengths. We then selected for dna sequences wherein *repA* was next to *repB*, and *repB* was next to *repC* (by definition, the only gene between *repA* and *repC* was *repB*), this was taken as a complete operon. The homologous protein groups contained only proteins whose genes formed complete operons. For each homologous protein group, we constructed an alignment with MUSCLE [19], and used this alignment to infer a phylogeny (see below). To generate the DNA alignments of *repA*, *repB*, and *repC*, we used their protein alignments as references, and performed nucleotide alignment using the "tranalign" program from The European Molecular Biology Open Software Suite (EMBOSS) [20]. The recombination analysis was carried out on these DNA alignments.

Other sets of DNA alignments were created for each of the operons contained in *R. etli* CFN42 and *R. leguminosarum* 3841. The intergenic region between *repB* and *repC* was also considered. We then used jModelTest [21] to carry out statistical selection of the best-fit models of nucleotide substitution for every DNA alignment. Finally, maximum likelihood distance matrices were inferred using the model specifications from jModelTest; this was done with PUZZLE [22].

Phylogenetic Analysis

Phylogenies were created using MrBayes v3.1.2 [23], allowing the MCMC sampler to explore all of the fixed-rated amino acid models included in MrBayes. The number of rate categories for gamma distributions was set to four, with a proportion of sites allowed to be invariable. We performed two runs with four chains each, for 5,000,000 generations. Trees were sampled every 1000 generations, 20% of all generations were removed as burn-in, and a consensus tree was taken. We also estimated the best amino acid models, including the amino acid matrices with the highest posterior probability, estimates of the proportion of invariable sites, and estimates of the gamma shape parameter.

A strict consensus tree was created from all three Bayesian phylogenies, using CONSENSE [24].

We established the similarities of the phylogenies using the Robinson and Fould distance (RFd), as calculated with TREEDIST [24].

We used confidence sets to assess whether the differences in topology between the individual Bayesian phylogenies exceeded those expected to occur by chance. We used expected likelihood weighting [13], which provides a simple and intuitive method for making multiple comparisons of models and constructing the corresponding confidence sets. This test has the benefit of being less conservative than the Shimodaira-Hasegawa test [13]. The topologies tested included those from the RepA, RepB, and RepC phylogenies. PUZZLE [22] was used to carry out this test for each protein alignment.

Recombination analysis

Although methods that use the substitution patterns or incompatibilities among sites seem be the most powerful strategy for identifying the presence of recombination events, no single method seems to perform optimally under all different scenarios [25]. Thus, the best strategy is often to use a combination of methods. Here, we used the RDP3 program [26], which implements a number of methods for identifying recombination events, including GENECONV [27], RDP [26], MaxChi [28], Chimera [28], SisCan [29], and Bootscanning [30]. We identified a recombination event as valid when at least three of the six methods indicated positive findings.

Functional regions and among-site rate variation in Rep proteins

We identified the various protein domains by applying the Pfam-A component of Pfam [15]. For this analysis, the RepA, RepB, and RepC proteins of symbiotic plasmid p42d from *R. etli* CFN42 were queried against Pfam-A. For every position of each protein alignment, a substitution rate was assigned using a discrete-gamma distribution. The discrete-gamma distribution used five rate classes and was implemented through PUZZLE.

Codon Adaptation Index as measure of evolutionary adaptedness

This analysis was done only for the *repA*, *repB*, and *repC* genes located on operons found within species *R. etli* CFN42 and *R. leguminosarum* 3841. We used the utility "cusp" from EMBOSS to calculate a codon usage table for the genes encoding the ribosomal proteins in each species. Using these tables as a reference, we applied the “cai” program of the EMBOSS suite to calculate Codon Adaptation Indices for the *repA*, *repB*, and *repC* genes.

Authors’ contributions

SC-R conceived and designed the experiments. SC-R and JFV-C performed the experiments. SC-R analyzed the data. SC-R and MAC discussed the results. MAC and SC-R wrote the manuscript. JFV-C and VG checked the manuscript. VG contributed materials. All authors read and approved the final manuscript.

Acknowledgments

We thank Luis Lozano for his comments on the manuscript. SC-R acknowledges a PhD fellowship from CONACyT, and thanks the Computational Unit of the Instituto de Biotecnología, particularly Jerome Verleyen, and the “Macroproyecto de Tecnologías de la Información y la Computación de la UNAM,” for the use of their computer facilities. This work was partially supported by PAPIIT Grant IN205808. SC-R would like to extend warm thanks to Valeria Lobos for all the non-academic support, which is the one that most counts.

References

1. Cevallos, M.A., R. Cervantes-Rivera, and R.M. Gutierrez-Rios, *The repABC plasmid family.* Plasmid, 2008. **60**(1): p. 19-37.
2. Pappas, K.M., *Cell-cell signaling and the Agrobacterium tumefaciens Ti plasmid copy number fluctuations.* Plasmid, 2008. **60**(2): p. 89-107.
3. Young, J.P., et al., *The genome of Rhizobium leguminosarum has recognizable core and accessory components.* Genome Biol, 2006. **7**(4): p. R34.
4. Gonzalez, V., 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): p. 3834-9.
5. Kumar, C.C. and R.P. Novick, *Plasmid pT181 replication is regulated by two countertranscripts.* Proc Natl Acad Sci U S A, 1985. **82**(3): p. 638-42.
6. MacLellan, S.R., et al., *The expression of a novel antisense gene mediates incompatibility within the large repABC family of alpha-proteobacterial plasmids.* Mol Microbiol, 2005. **55**(2): p. 611-23.
7. Slater, S.C., et al., *Genome sequences of three agrobacterium biovars help elucidate the evolution of multichromosome genomes in bacteria.* J Bacteriol, 2009. **191**(8): p. 2501-11.
8. Chai, Y. and S.C. Winans, *RepB protein of an Agrobacterium tumefaciens Ti plasmid binds to two adjacent sites between repA and repB for plasmid partitioning and autorepression.* Mol Microbiol, 2005. **58**(4): p. 1114-29.
9. Ramirez-Romero, M.A., et al., *Structural elements required for replication and incompatibility of the Rhizobium etli symbiotic plasmid.* J Bacteriol, 2000. **182**(11): p. 3117-24.
10. Venkova-Canova, T., et al., *Two discrete elements are required for the replication of a repABC plasmid: an antisense RNA and a stem-loop structure.* Mol Microbiol, 2004. **54**(5): p. 1431-44.
11. Cevallos, M.A., et al., *Rhizobium etli CFN42 contains at least three plasmids of the repABC family: a structural and evolutionary analysis.* Plasmid, 2002. **48**(2): p. 104-16.

12. Omelchenko, M.V., et al., *Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ.* Genome Biol, 2003. **4**(9): p. R55.
13. Strimmer, K. and A. Rambaut, *Inferring confidence sets of possibly misspecified gene trees.* Proc Biol Sci, 2002. **269**(1487): p. 137-42.
14. Yang, Z., *Among-site rate variation and its impact on phylogenetic analyses.* Trends Ecol Evol, 1996. **11**(9): p. 6.
15. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2008. **36**(Database issue): p. D281-8.
16. Sharp, P.M. and W.H. Li, *The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications.* Nucleic Acids Res, 1987. **15**(3): p. 1281-95.
17. Garcillan-Barcia, M.P., M.V. Francia, and F. de la Cruz, *The diversity of conjugative relaxases and its application in plasmid classification.* FEMS Microbiol Rev, 2009. **33**(3): p. 657-87.
18. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
19. Edgar, R.C., *MUSCLE: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, 2004. **5**: p. 113.
20. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: the European Molecular Biology Open Software Suite.* Trends Genet, 2000. **16**(6): p. 276-7.
21. Posada, D., *jModelTest: phylogenetic model averaging.* Mol Biol Evol, 2008. **25**(7): p. 1253-6.
22. Schmidt, H.A., et al., *TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.* Bioinformatics, 2002. **18**(3): p. 502-4.
23. Ronquist, F. and J.P. Huelsenbeck, *MrBayes 3: Bayesian phylogenetic inference under mixed models.* Bioinformatics, 2003. **19**(12): p. 1572-4.
24. Felsenstein, J., *PHYLIP (Phylogeny Inference Package) version 3.6.* Department of Genome Sciences, University of Washington, Seattle, 2005.
25. Posada, D., K.A. Crandall, and E.C. Holmes, *Recombination in evolutionary genomics.* Annu Rev Genet, 2002. **36**: p. 75-97.
26. Martin, D.P., C. Williamson, and D. Posada, *RDP2: recombination detection and analysis from sequence alignments.* Bioinformatics, 2005. **21**(2): p. 260-2.
27. Padidam, M., S. Sawyer, and C.M. Fauquet, *Possible emergence of new geminiviruses by frequent recombination.* Virology, 1999. **265**(2): p. 218-25.
28. Smith, J.M., *Analyzing the mosaic structure of genes.* J Mol Evol, 1992. **34**(2): p. 126-9.
29. Gibbs, M.J., J.S. Armstrong, and A.J. Gibbs, *Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences.* Bioinformatics, 2000. **16**(7): p. 573-82.
30. Salminen, M.O., et al., *Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning.* AIDS Res Hum Retroviruses, 1995. **11**(11): p. 1423-5.

Figure legends

Figure 1 - The individual Bayesian phylogenies

The Bayesian phylogenies for RepA, RepB, and RepC. The scale bar denotes the estimated number of amino acid substitution per site. The asterisks on the branches represent posterior probability values higher than 0.95, otherwise values are shown. All of the phylogenies were artificially rooted with the homologous gene on chromosome 2 of *Ochrobactrum anthropi*, to facilitate visual comparison between phylogenies.

Figure 2 - The strict consensus tree

A strict consensus tree was constructed using the Bayesian phylogenies of RepA, RepB, and RepC.

Figure 3 - Functional restrictions within the RepA and RepB proteins

Substitution rate variation among sites in the RepA and RepB proteins. For each protein, all sites were assigned to one of five gamma categories. The Pfam-A domains are shown for each protein, as well as the zones affected by recombination events.

Figure 4 - Codon Adaptation Index

CAI values are shown for each of the genes comprising the *repABC* operons found in *R. etli* CFN42 and *R. leguminosarum* 3841. Red circles indicate the putatively newest plasmids in *R. etli* CFN42. Green squares show the inconsistencies found herein.

Tables

Table 1 - Estimates of the best amino acid models for the individual Bayesian phylogenies

The amino acid matrix with the highest posterior probability, the estimated proportion of invariable sites, and the estimated gamma shape parameter for each Rep protein. Abbreviations. TL: total length of the phylogeny, PP: posterior probability. The values in parentheses is the 95% Cred. Interval.

Table 2 - Average between-locus distance for the different loci

The average distance over all possible sequence pairs for each locus, along with the specifications made by jModelTest regarding the substitution model.

+Average distance over all possible pairs of sequences.

All the loci but the “Intergenic region” selected the GTR model with correction for across site rate variation and invariant sites (GTR+I+G). The “Intergenic region” selected *TPM2 with correction for across site rate variation (TPM2+G).

*This model implies AC=AT; CG=GT; AG=CT;

Additional Files

Additional file 1 - Homologous genes of *repA, repB,* and *repC,* as well as complete and faulty *repABC* operons found across the studied *Alphaproteobacteria* genomes. For each gene it was registered whether it was located on a chromosome (C) or a plasmid (P).

Additional file 2 - In order to determine the similarity among the Rep phylogenies, Robison-Fould distances between Rep phylogenies were established.

Additional file 3 - Recombination events identified for *repA*, *repB*, and *repC*. Pairwise identity plots of the localized recombination events, showing major and minor parent sequences as well as the daughter sequence. Abbreviations are given in Additional file 6.

Additional file 4 - Functional restrictions within RepC. Substitution rate variation among sites in RepC. All sites were assigned to one of five gamma categories. Pfam-A domains are shown, as well as the zone affected by recombination events.

Additional file 5 - *Alphaproteobacteria* genomes used to search for *repABC* operons.

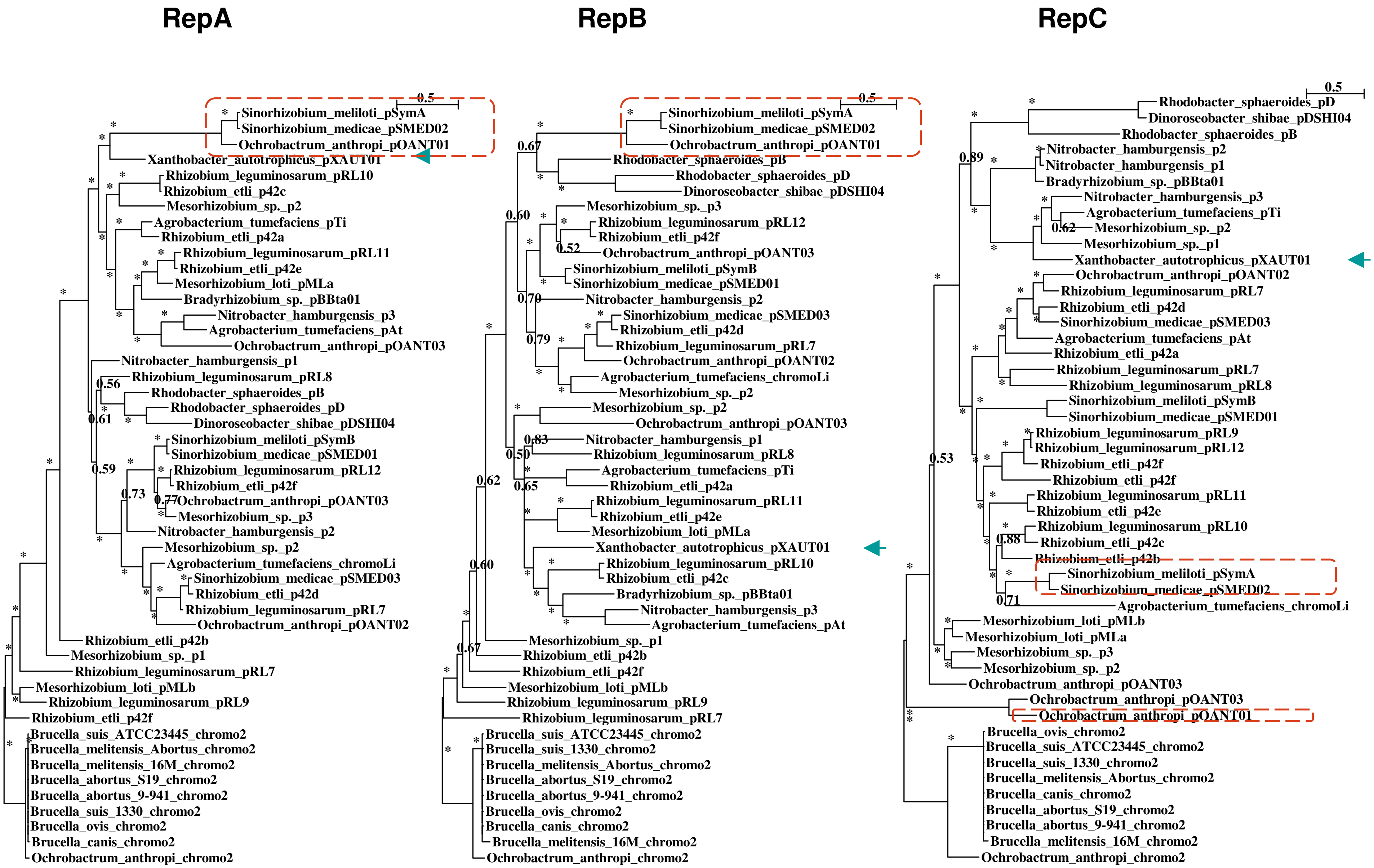
Additional file 6 – Abbreviations used in the pairwise identity plots.

Table 1. Amino acid models specifications for RepA, RepB, and RepC proteins.

Protein	*TL*	*Shape parameter α*	*P. Invariant sites*	*Model*
RepA	11.987 (11.128 12.906)	0.933 (0.776 1.102)	0.065 (0.022 0.107)	WAG (PP 1)
RepB	19.952 (18.617 21.323)	1.721 (1.487 1.983)	0.0698 (0.041 0.103)	WAG (PP 1)
RepC	17.678 (16.49 18.922)	1.122 (0.993 1.265)	0.068 (0.040 0.098)	JTT (PP 1)

Table 2. Average between-locus distance and model subtitution specifications

Locus	*Average distance+*	*P. Invariant sites*	*Shape parameter α*
repA	0.72530	0.194	1.176
repB	1.13479	0.089	1.661
Intergenic region	0.45827	0.0	0.47
repC	0.59197	0.186	1.108


RepA

RepB

RepC

0.5
Sinorhizobium_meliloti_pSymA
Sinorhizobium_medicae_pSMED02
Ochrobactrum_anthropi_pOANT01
Xanthobacter_autotrophicus_pXAUT01
Rhizobium_leguminosarum_pRL10
Rhizobium_etli_p42c
Mesorhizobium_sp._p2
Agrobacterium_tumefaciens_pTi
Rhizobium_etli_p42a
Rhizobium_leguminosarum_pRL11
Rhizobium_etli_p42e
Mesorhizobium_loti_pMLa
Bradyrhizobium_sp._pBBta01
Nitrobacter_hamburgensis_p3
Agrobacterium_tumefaciens_pAt
Ochrobactrum_anthropi_pOANT03
Nitrobacter_hamburgensis_p1
Rhizobium_leguminosarum_pRL8
0.56
Rhodobacter_sphaeroides_pB
Rhodobacter_sphaeroides_pD
0.61
Dinoroseobacter_shibae_pDSHI04
Sinorhizobium_meliloti_pSymB
Sinorhizobium_medicae_pSMED01
0.59
Rhizobium_leguminosarum_pRL12
Rhizobium_etli_p42f
0.73
0.77
Ochrobactrum_anthropi_pOANT03
Mesorhizobium_sp._p3
Nitrobacter_hamburgensis_p2
Mesorhizobium_sp._p2
Agrobacterium_tumefaciens_chromoLi
Sinorhizobium_medicae_pSMED03
Rhizobium_etli_p42d
Rhizobium_leguminosarum_pRL7
Ochrobactrum_anthropi_pOANT02
Rhizobium_etli_p42b
Mesorhizobium_sp._p1
Rhizobium_leguminosarum_pRL7
Mesorhizobium_loti_pMLb
Rhizobium_leguminosarum_pRL9
Rhizobium_etli_p42f
Brucella_suis_ATCC23445_chromo2
Brucella_melitensis_Abortus_chromo2
Brucella_melitensis_16M_chromo2
Brucella_abortus_S19_chromo2
Brucella_abortus_9-941_chromo2
Brucella_suis_1330_chromo2
Brucella_ovis_chromo2
Brucella_canis_chromo2
Ochrobactrum_anthropi_chromo2
0.5
Sinorhizobium_meliloti_pSymA
Sinorhizobium_medicae_pSMED02
0.67
Ochrobactrum_anthropi_pOANT01
Rhodobacter_sphaeroides_pB
Rhodobacter_sphaeroides_pD
Dinoroseobacter_shibae_pDSHI04
Mesorhizobium_sp._p3
0.60
Rhizobium_leguminosarum_pRL12
Rhizobium_etli_p42f
0.52
Ochrobactrum_anthropi_pOANT03
Sinorhizobium_meliloti_pSymB
Sinorhizobium_medicae_pSMED01
0.70
Nitrobacter_hamburgensis_p2
Sinorhizobium_medicae_pSMED03
Rhizobium_etli_p42d
0.79
Rhizobium_leguminosarum_pRL7
Ochrobactrum_anthropi_pOANT02
Agrobacterium_tumefaciens_chromoLi
Mesorhizobium_sp._p2
Mesorhizobium_sp._p2
Ochrobactrum_anthropi_pOANT03
0.83
Nitrobacter_hamburgensis_p1
0.50
Rhizobium_leguminosarum_pRL8
Agrobacterium_tumefaciens_pTi
0.62
0.65
Rhizobium_etli_p42a
Rhizobium_leguminosarum_pRL11
Rhizobium_etli_p42e
Mesorhizobium_loti_pMLa
Xanthobacter_autotrophicus_pXAUT01
0.60
Rhizobium_leguminosarum_pRL10
Rhizobium_etli_p42c
Bradyrhizobium_sp._pBBta01
Nitrobacter_hamburgensis_p3
Agrobacterium_tumefaciens_pAt
Mesorhizobium_sp._p1
0.67
Rhizobium_etli_p42b
Rhizobium_etli_p42f
Mesorhizobium_loti_pMLb
Rhizobium_leguminosarum_pRL9
Rhizobium_leguminosarum_pRL7
Brucella_suis_ATCC23445_chromo2
Brucella_suis_1330_chromo2
Brucella_melitensis_Abortus_chromo2
Brucella_abortus_S19_chromo2
Brucella_abortus_9-941_chromo2
Brucella_ovis_chromo2
Brucella_canis_chromo2
Brucella_melitensis_16M_chromo2
Ochrobactrum_anthropi_chromo2
0.5
Rhodobacter_sphaeroides_pD
Dinoroseobacter_shibae_pDSHI04
Rhodobacter_sphaeroides_pB
Nitrobacter_hamburgensis_p2
0.89
Nitrobacter_hamburgensis_p1
Bradyrhizobium_sp._pBBta01
Nitrobacter_hamburgensis_p3
Agrobacterium_tumefaciens_pTi
0.62
Mesorhizobium_sp._p2
Mesorhizobium_sp._p1
Xanthobacter_autotrophicus_pXAUT01
Ochrobactrum_anthropi_pOANT02
Rhizobium_leguminosarum_pRL7
Rhizobium_etli_p42d
Sinorhizobium_medicae_pSMED03
Agrobacterium_tumefaciens_pAt
Rhizobium_etli_p42a
Rhizobium_leguminosarum_pRL7
Rhizobium_leguminosarum_pRL8
Sinorhizobium_meliloti_pSymB
Sinorhizobium_medicae_pSMED01
Rhizobium_leguminosarum_pRL9
Rhizobium_leguminosarum_pRL12
0.53
Rhizobium_etli_p42f
Rhizobium_etli_p42f
Rhizobium_leguminosarum_pRL11
Rhizobium_etli_p42e
Rhizobium_leguminosarum_pRL10
0.88
Rhizobium_etli_p42c
Rhizobium_etli_p42b
Sinorhizobium_meliloti_pSymA
Sinorhizobium_medicae_pSMED02
0.71
Agrobacterium_tumefaciens_chromoLi
Mesorhizobium_loti_pMLb
Mesorhizobium_loti_pMLa
Mesorhizobium_sp._p3
Mesorhizobium_sp._p2
Ochrobactrum_anthropi_pOANT03
Ochrobactrum_anthropi_pOANT03
Ochrobactrum_anthropi_pOANT01
Brucella_ovis_chromo2
Brucella_suis_ATCC23445_chromo2
Brucella_suis_1330_chromo2
Brucella_melitensis_Abortus_chromo2
Brucella_canis_chromo2
Brucella_abortus_S19_chromo2
Brucella_abortus_9-941_chromo2
Brucella_melitensis_16M_chromo2
Ochrobactrum_anthropi_chromo2

Figure 1

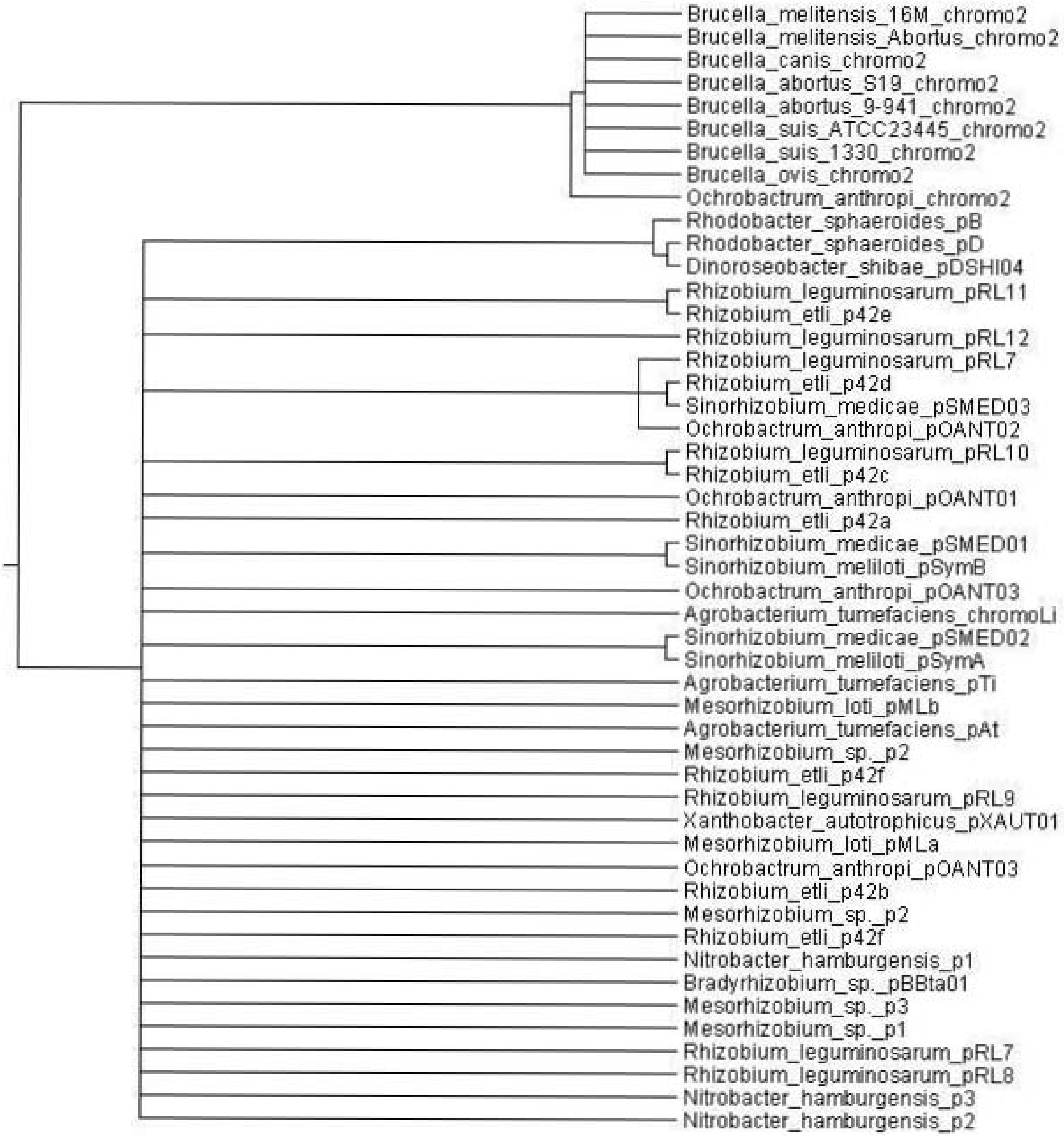

Brucella_melitensis_16M_chromo2
Brucella_melitensis_Abortus_chromo2
Brucella_canis_chromo2
Brucella_abortus_S19_chromo2
Brucella_abortus_9-941_chromo2
Brucella_suis_ATCC23445_chromo2
Brucella_suis_1330_chromo2
Brucella_ovis_chromo2
Ochrobactrum_anthropi_chromo2
Rhodobacter_sphaeroides_pB
Rhodobacter_sphaeroides_pD
Dinoroseobacter_shibae_pDSHI04
Rhizobium_leguminosarum_pRL11
Rhizobium_etli_p42e
Rhizobium_leguminosarum_pRL12
Rhizobium_leguminosarum_pRL7
Rhizobium_etli_p42d
Sinorhizobium_medicae_pSMED03
Ochrobactrum_anthropi_pOANT02
Rhizobium_leguminosarum_pRL10
Rhizobium_etli_p42c
Ochrobactrum_anthropi_pOANT01
Rhizobium_etli_p42a
Sinorhizobium_medicae_pSMED01
Sinorhizobium_meliloti_pSymB
Ochrobactrum_anthropi_pOANT03
Agrobacterium_tumefaciens_chromoLi
Sinorhizobium_medicae_pSMED02
Sinorhizobium_meliloti_pSymA
Agrobacterium_tumefaciens_pTi
Mesorhizobium_loti_pMLb
Agrobacterium_tumefaciens_pAt
Mesorhizobium_sp._p2
Rhizobium_etli_p42f
Rhizobium_leguminosarum_pRL9
Xanthobacter_autotrophicus_pXAUT01
Mesorhizobium_loti_pMLa
Ochrobactrum_anthropi_pOANT03
Rhizobium_etli_p42b
Mesorhizobium_sp._p2
Rhizobium_etli_p42f
Nitrobacter_hamburgensis_p1
Bradyrhizobium_sp._pBBta01
Mesorhizobium_sp._p3
Mesorhizobium_sp._p1
Rhizobium_leguminosarum_pRL7
Rhizobium_leguminosarum_pRL8
Nitrobacter_hamburgensis_p3
Nitrobacter_hamburgensis_p2

Figure 2

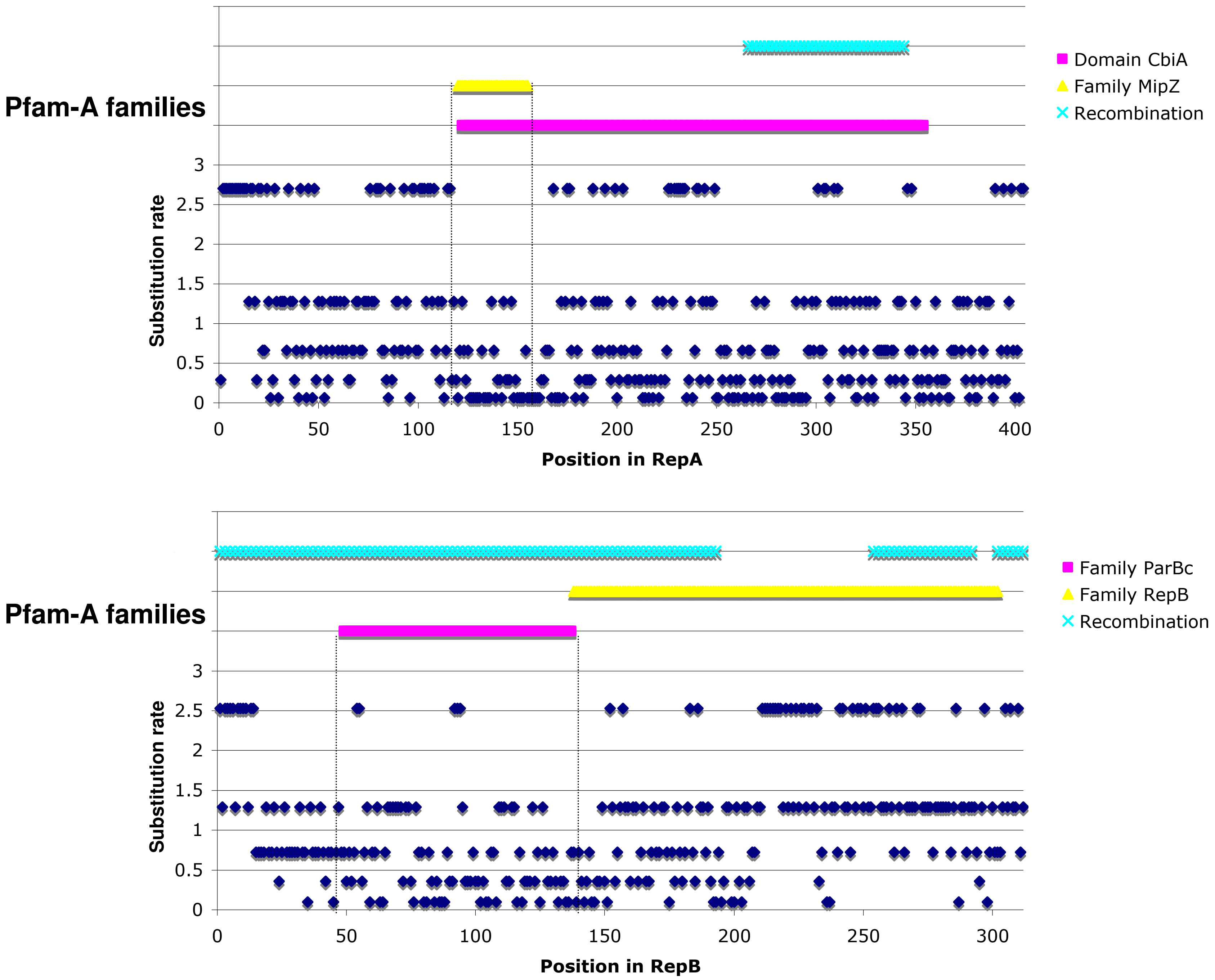

Pfam-A families
■ Domain CbiA
▲ Family MipZ
× Recombination
Substitution rate
3
2.5
2
1.5
1
0.5
0
0
50
100
150
200
250
300
350
400
Position in RepA
Pfam-A families
■ Family ParBc
▲ Family RepB
× Recombination
Substitution rate
3
2.5
2
1.5
1
0.5
0
0
50
100
150
200
250
300
Position in RepB

Figure 3

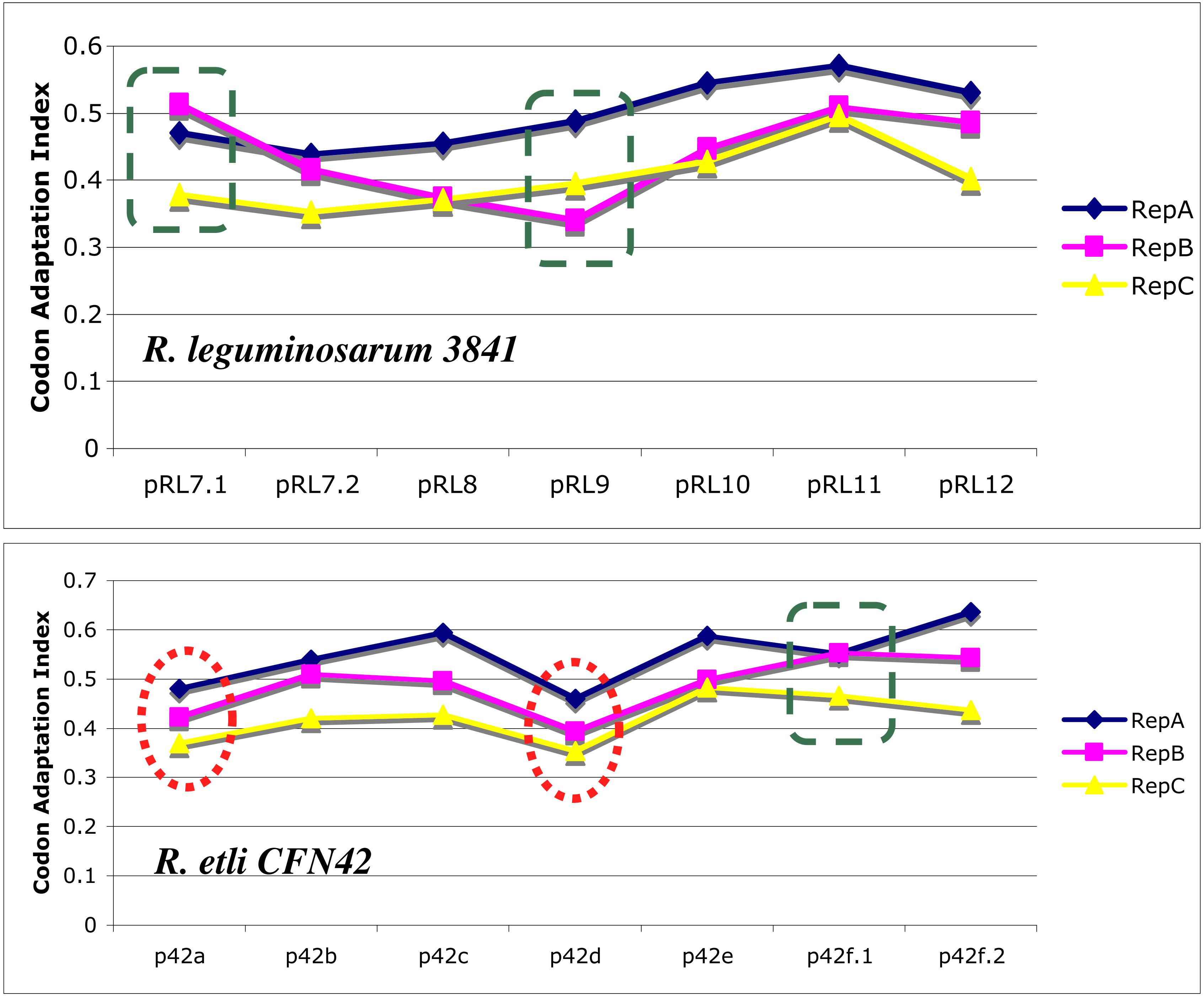

Codon Adaptation Index
0.6
0.5
0.4
0.3
0.2
0.1
0
R. leguminosarum 3841
pRL7.1
pRL7.2
pRL8
pRL9
pRL10
pRL11
pRL12
RepA
RepB
RepC
Codon Adaptation Index
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
R. etli CFN42
p42a
p42b
p42c
p42d
p42e
p42f.1
p42f.2
RepA
RepB
RepC

Figure 4

Additional files provided with this submission:

Additional file 1: additionalfile1.doc, 107K
http://www.biomedcentral.com/imedia/3536304302742797/supp1.doc
Additional file 2: additionalfile2.doc, 21K
http://www.biomedcentral.com/imedia/3247034732742797/supp2.doc
Additional file 3: additionalfile3.ppt, 334K
http://www.biomedcentral.com/imedia/1710426416274279/supp3.ppt
Additional file 4: additionalfile4.ppt, 156K
http://www.biomedcentral.com/imedia/1016089163274279/supp4.ppt
Additional file 5: additionalfile5.doc, 38K
http://www.biomedcentral.com/imedia/4392248242742798/supp5.doc
Additional file 6: additionalfile6.doc, 32K
http://www.biomedcentral.com/imedia/8392363672742798/supp6.doc

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 *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

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

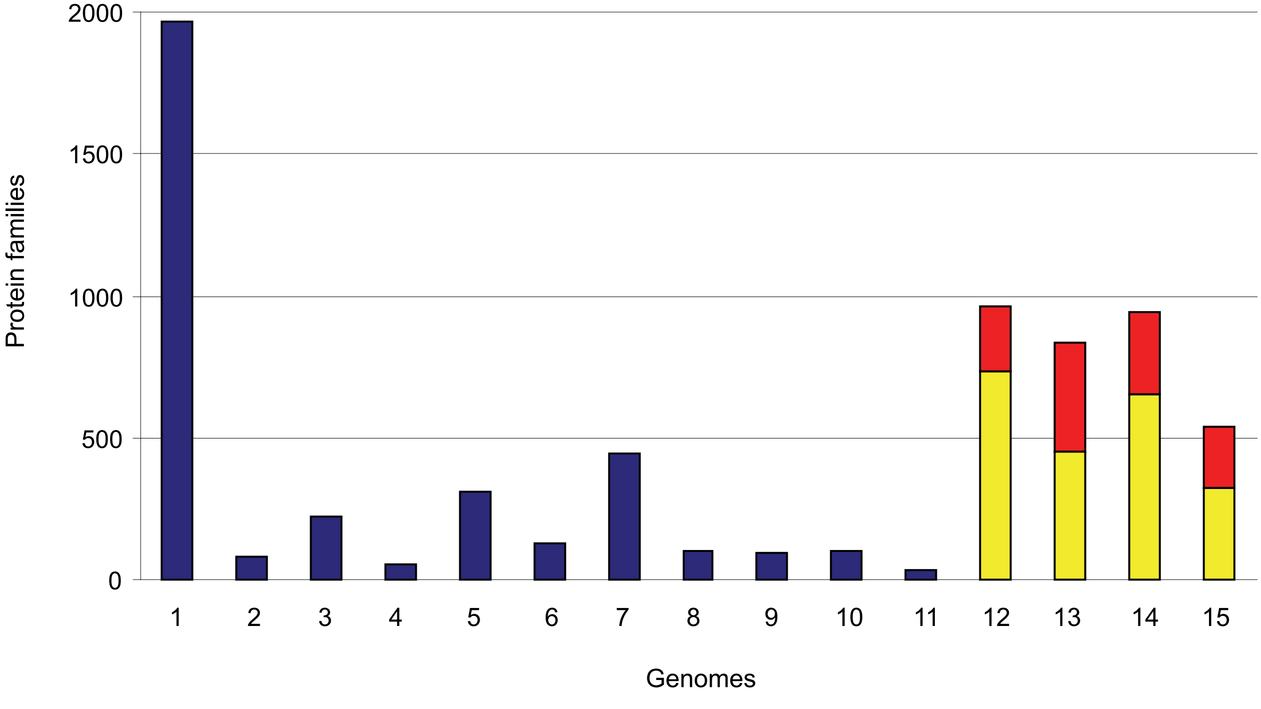


Figure 1. Distribution of protein families in the genomes of *S. meliloti* (S), *A. tumefaciens* (A), *R. leguminosarum* (L) and *R. etli* (E). - Bar number indicates the assignation of the protein families to the corresponding genome according to the following letters code: 1, SALE; 2, SAL; 3, ALE; 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

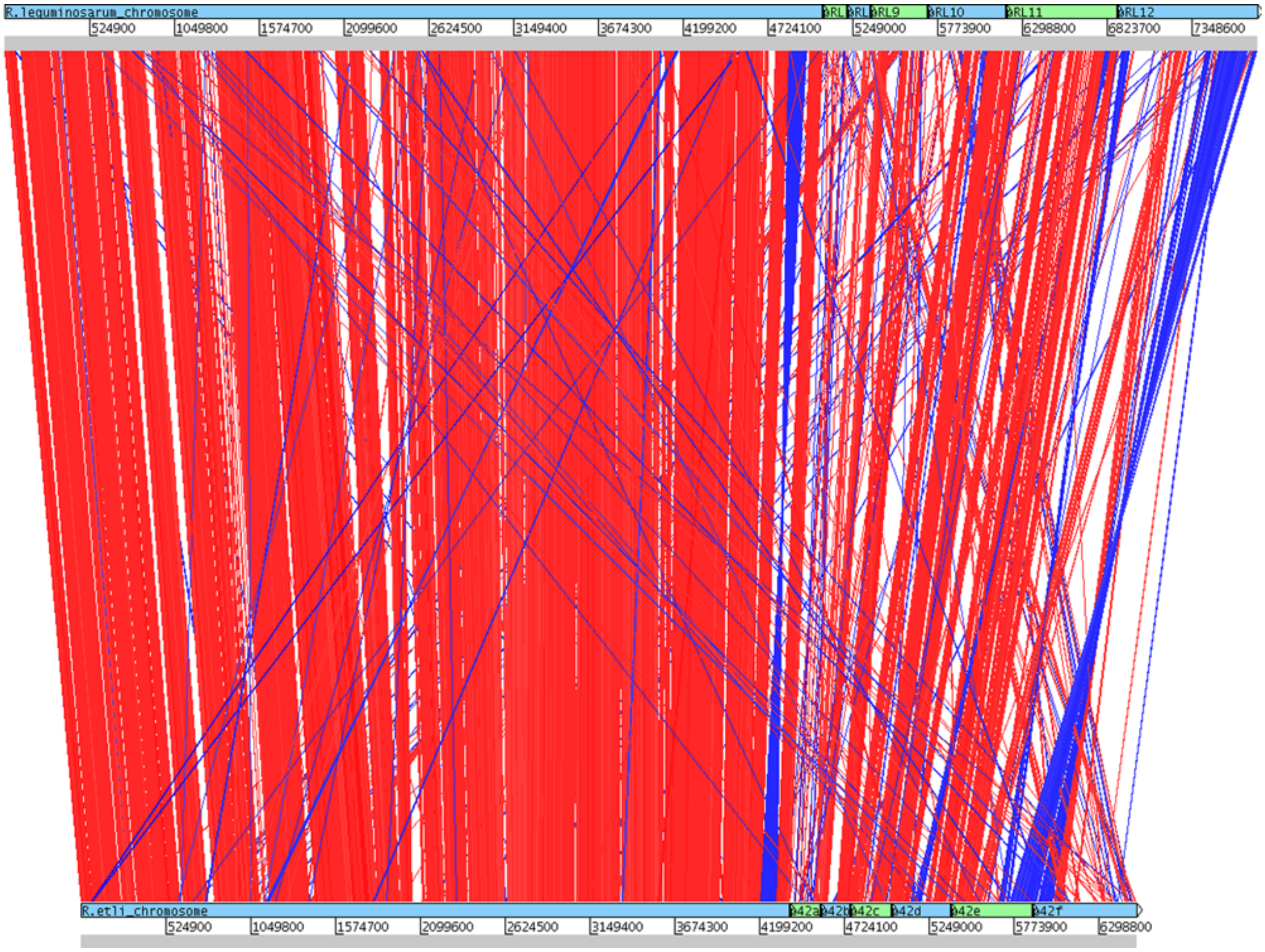


Figure 2. ACT View of Chromosome and Plasmids. The chromosomal and plasmid DNAs have been laid end-to-end and analysed using the Artemis comparison tool (ACT) [15]. Red bars represent close matches, whilst blue bars represent inverted close matches. The *R. leguminosarum* genome is at the top of the 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

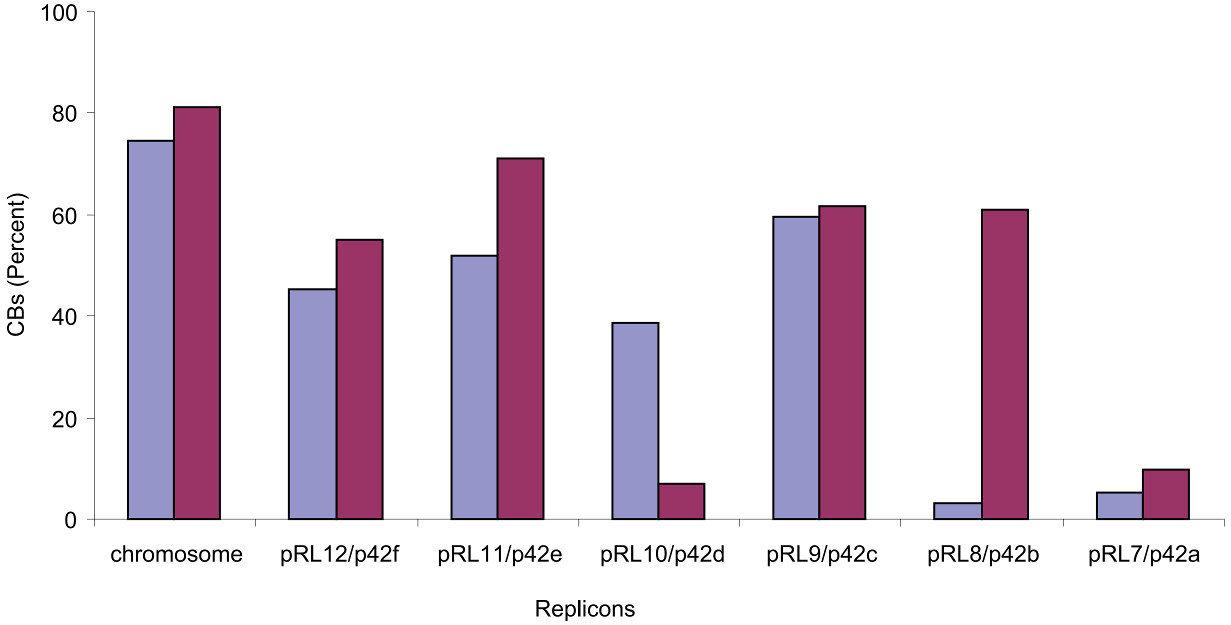


Figure 3. The proportion of synteny in the *R.etli* genome as compared to the *R. leguminosarum* genome. - The proportion of synteny is expressed as the percentage of the total DNA in CBs (Y axis) considering the total length of the pairs of replicons (X axis). Blue color *R. etli* CFN42; magenta, *R.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

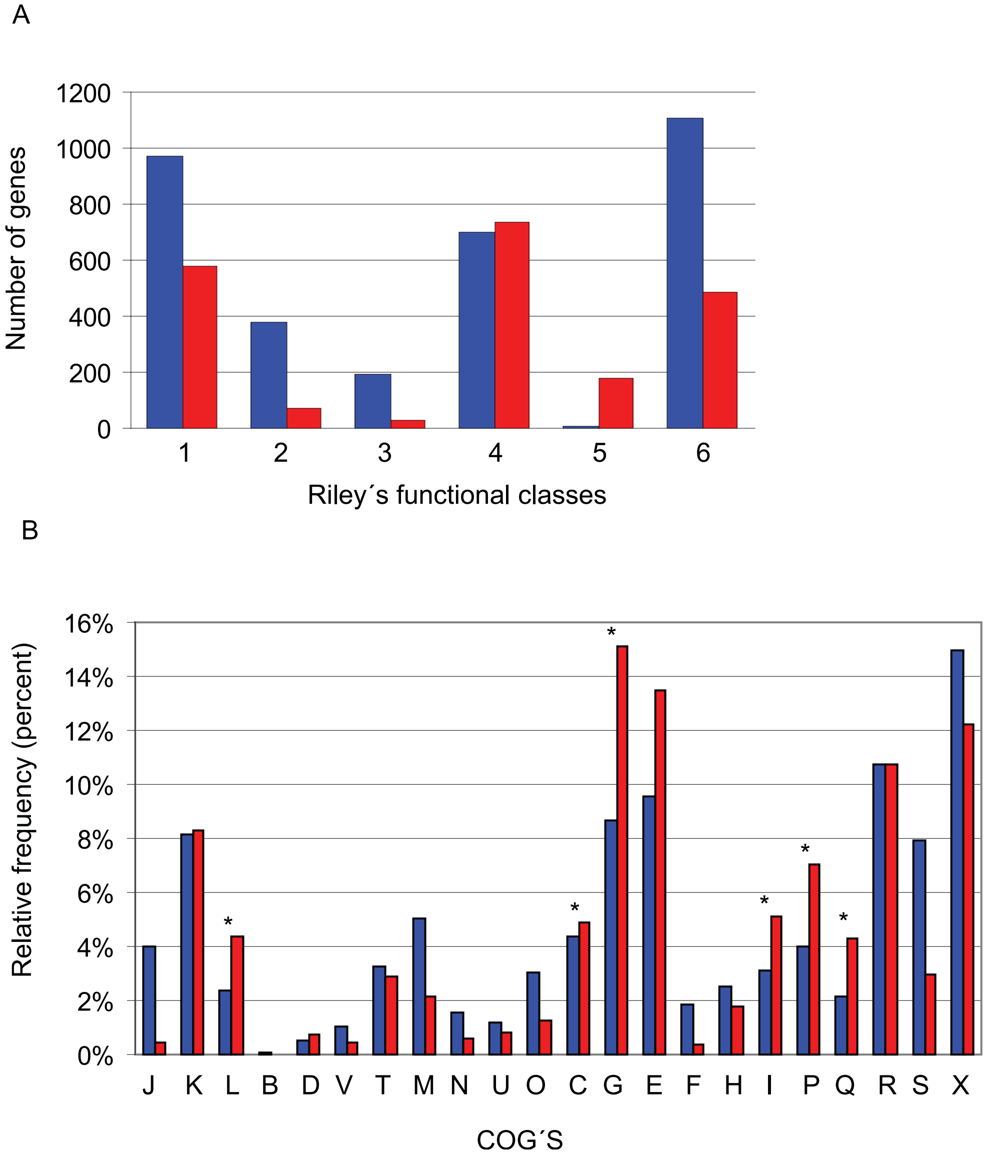


Figure 4. Functional bias in CHR-O (chromosomal only) and N-CHR (non-chromosomal) classes of homologues. - Figure 4a) Rileys categories: 1. small molecule metabolism. 2. Macromolecule metabolism. 3. Structural elements. 4. Cell process. 5. External origin. 6. Miscellaneous. 4b) COGs functional classification. 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

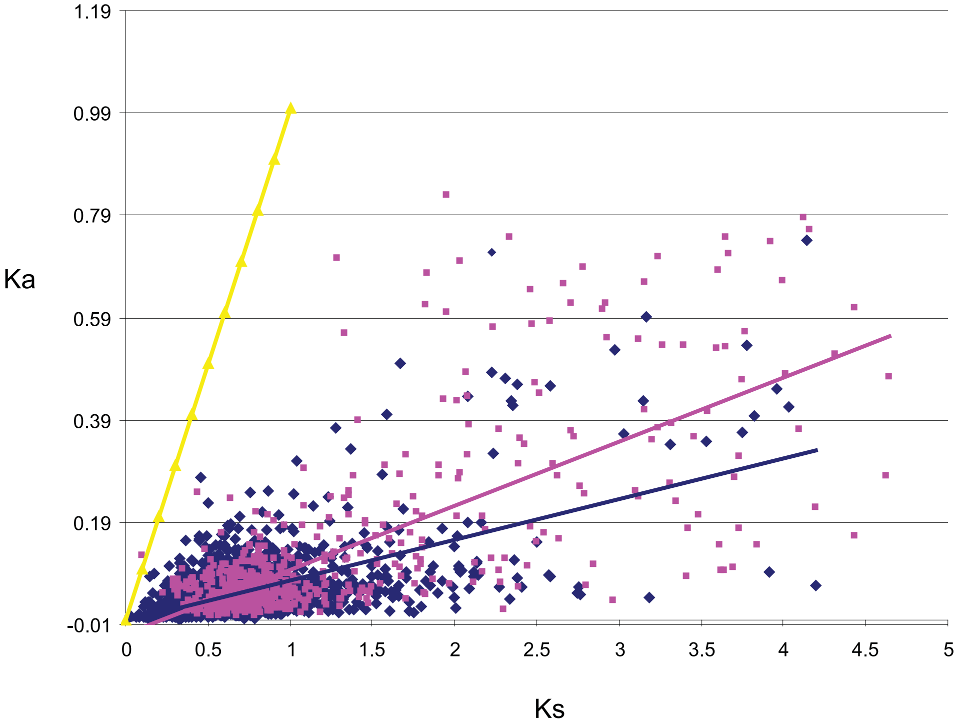


Figure 5. Rates of synonymous (Ks) and non-synonymous substitutions (Ka) in orthologous genes of *R. etli* and *R. leguminosarum*. - Neutrality line (Ka = Ks) is indicated in yellow. Linear regressions for CO class (blue color line and diamonds) and NC class (rose color line and diamonds) are indicated. 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 ≪ 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

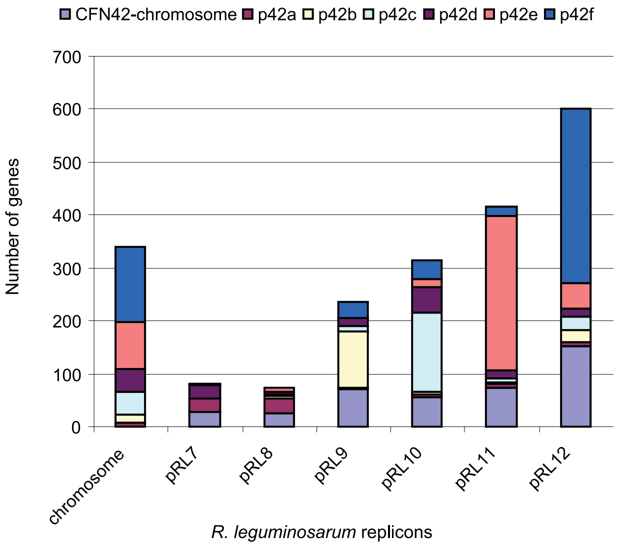


Figure 6. Composition of the *R. leguminosarum* and *R.etli* genomes according to N-CHR homologues. - The composition of the *R. leguminosarum* genome compared to the replicons of *R. etli* is shown. The replicon name is given at the base of the figure and color key to the right of the figure. Genes 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

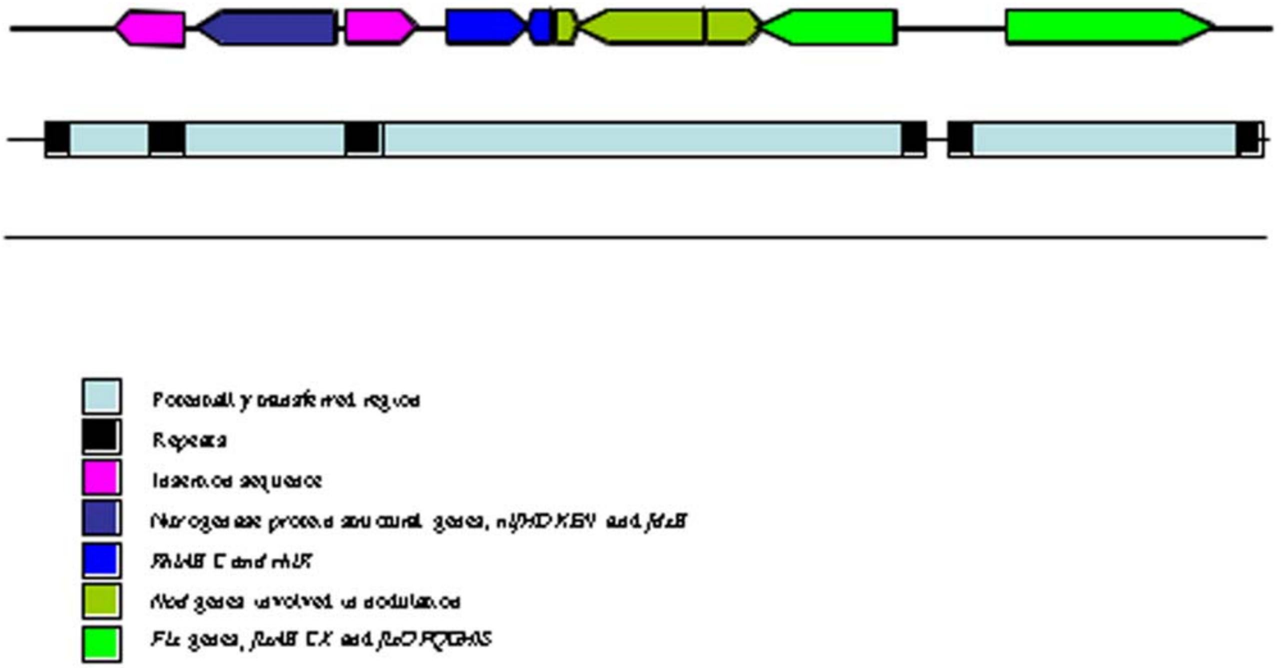

Potentially transferred region
Repeats
Insertion sequence
Nitrogenase protein structural genes, nifHDKEN and fdxB
RhiABC and rhiR
Nod genes involved in nodulation
Fix genes, fixABCX and fixOPQGHIS

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.

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.

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.

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.

BMC Genomics

BioMed Central

Research article

Open Access

Rapid evolutionary change of common bean (*Phaseolus vulgaris* L) plastome, and the genomic diversification of legume chloroplasts

Xianwu Guo*[1], Santiago Castillo-Ramírez[1], Víctor González[1], Patricia Bustos[1], José Luís Fernández-Vázquez[1], Rosa Isela Santamaría[1], Jesús Arellano[2], Miguel A Cevallos[1] and Guillermo Dávila[1]

Address: [1]Programa de Genómica Evolutiva, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Apartado Postal 565-A, C.P 62210, Cuernavaca, Morelos, México and [2]Programa de Genómica Funcional de Eucariotes, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Apartado Postal 565-A, C.P 62210, Cuernavaca, Morelos, México

Email: Xianwu Guo* - gxianwu@ccg.unam.mx; Santiago Castillo-Ramírez - iago@ccg.unam.mx; Víctor González - vgonzal@ccg.unam.mx; Patricia Bustos - paty@ccg.unam.mx; José Luís Fernández-Vázquez - jlfernan@ccg.unam.mx; Rosa Isela Santamaría - rosa@ccg.unam.mx; Jesús Arellano - jesus@ccg.unam.mx; Miguel A Cevallos - mac@ccg.unam.mx; Guillermo Dávila - davila@ccg.unam.mx

* Corresponding author

Published: 10 July 2007

BMC Genomics 2007, **8**:228 doi:10.1186/1471-2164-8-228

Received: 13 February 2007

Accepted: 10 July 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/228

Abstract

Background: Fabaceae (legumes) is one of the largest families of flowering plants, and some members are important crops. In contrast to what we know about their great diversity or economic importance, our knowledge at the genomic level of chloroplast genomes (cpDNAs or plastomes) for these crops is limited.

Results: We sequenced the complete genome of the common bean (*Phaseolus vulgaris* cv. Negro Jamapa) chloroplast. The plastome of *P. vulgaris* is a 150,285 bp circular molecule. It has gene content similar to that of other legume plastomes, but contains two pseudogenes, *rpl*33 and *rps*16. A distinct inversion occurred at the junction points of *trn*H-GUG/*rpl*14 and *rps*19/*rps*8, as in adzuki bean [1]. These two pseudogenes and the inversion were confirmed in 10 varieties representing the two domestication centers of the bean. Genomic comparative analysis indicated that inversions generally occur in legume plastomes and the magnitude and localization of insertions/deletions (indels) also vary. The analysis of repeat sequences demonstrated that patterns and sequences of tandem repeats had an important impact on sequence diversification between legume plastomes and tandem repeats did not belong to dispersed repeats. Interestingly, *P. vulgaris* plastome had higher evolutionary rates of change on both genomic and gene levels than *G. max*, which could be the consequence of pressure from both mutation and natural selection.

Conclusion: Legume chloroplast genomes are widely diversified in gene content, gene order, indel structure, abundance and localization of repetitive sequences, intracellular sequence exchange and evolutionary rates. The *P. vulgaris* plastome is a rapidly evolving genome.

Background

Chloroplasts are derived from an endosymbiotic cyanobacterium that invaded the eukaryotic cell a billion years ago. During the evolutionary process from endosymbiont to contemporary organelles, the cyanobacterium lost the bulk of its genome and retained the genes encoding the photosynthesis machinery and the components of several chemical pathways. During this process, it also acquired many host-derived properties and was thus transformed into a distinct organelle: the chloroplast.

Angiosperm chloroplast genomes present a similar gene content and gene order. They are circular molecules that can also be present in linear forms with multiple copies, ranging in size from 120 kb to 160 kb, but usually around 150 kb with about 90–110 unique genes [2]. A pair of large inverted repeats (IR) about 21–28 kb in length divides the genome into one large single-copy region (LSC) and one small single-copy region (SSC). rRNA genes are always located in IR regions.

Despite the overall conservation of plastomes, genomic diversification was also experienced in many respects. Many genes were lost phylogenetically, independently in parallel or uniquely lost in a particular species [3]. An extreme example is the cpDNA of the parasite plant *Epifagus virginiana*, which lost 13 tRNA genes and retained only 60 genes so that the genome was reduced to 70 kb [4]. It was found that several kinds of inversions interrupted the gene order of the plastome [5-11]. They are generally associated with specific lineages and thus could be a sign of important events in evolutionary diversification [12,13].

Sequence duplication is another feature of some land plant chloroplast genomes. For example, *Pelargonium* × *hortorum* contains some large duplicated fragments, including several genes, and numerous simple repeats as well as a tremendous extension of IR (75 kb) [14]. Definite evidence supporting transposition within plastid genomes is lacking, but intramolecular recombination mediated by short direct repeats has been reported [15].

The chloroplast genes have been extensively used to study the phylogenetic relationships at several taxonomic levels, especially in the analysis of basal clades, mainly because they have slower mutation rate in comparison with the nuclear genes [16]. The Fabaceae (legume) family is one of the largest and more diverse angiosperm families. It comprises about 20,000 species, which are distributed essentially in tropical regions. Chloroplast-derived markers have been used to study the evolutionary relationship between some legume plants (Fabaceae) [17-21]. However, to date, only the sequences of three legume chloroplast genomes have been reported: *Lotus japonicus*, *Glycine max*, [22,23] and *Medicago truncatula* (AC093544, unpublished). The common bean, *Phaseolus vulgaris*, is a major food crop, domesticated independently in two sites: Mesoamerica and South America[24]. The physical map of its chloroplast genome was published in 1983 [25] and some small pieces of the chloroplast genome were sequenced to study domestication [26] and phylogeny issues. Here we report the chloroplast sequence of *P. vulgaris* cv. Negro Jamapa. A comparative analysis of this sequence with other legume chloroplast genomes indicates that these genomes are highly diversified in sequence and organization. Moreover, we provide evidence that one plastome (*P. vulgaris*) evolved faster than another (*G. max*) at the genomic and gene levels, which could be the consequence of pressure coming from both mutation and natural selection.

Results

General features of the genome

The genome of *P. vulgaris* chloroplast is a circular molecule of 150,285 bp that contains an identical IR of 26,426 bp, separated by an LSC of 79,824 bp and an SSC of 17,610 bp (Fig. 1). The noncoding regions, including both introns and intergenic regions, comprises 40.4% of the genome. The overall A+T content for the genome is 64.6% in contrast to 68.7% for the noncoding regions. rRNA genes and tRNA genes have the lowest A+T composition with 45.1% and 47.6%, respectively. A total of 127 genes were assigned to the genome, 108 of which were unique and 19 were duplicated in IR regions. The unique genes included 75 coding-protein genes, 30 tRNA genes, and 4 rRNA genes. There were 17 genes containing one or two introns, six of which were tRNA genes.

Gene content

The gene content of chloroplast genomes of *P. vulgaris*, *G. max*, *L. japonicus*, and *M. truncatula*, the legume chloroplast genomes sequenced up to date, was similar. All lacked the *rpl*22 genes and *inf*A, which occurred in other flowering plants. A distinctive characteristic of the *P. vulgaris* chloroplast genome was the presence of two pseudogenes: *rps*16 and *rpl*33. *rps*16 is an intron-containing gene present as a functional gene in both *L. japonicus* and *G. max* but absent in *M. truncatula*. In *P. vulgaris*, *rps*16 has several features that define it as a pseudogene: firstly, it contains four stop-codons within the second exon; secondly, the gene lacks a functional motif located from the positions 16 to 47 of the amino acid sequence (comparing with the soybean sequence); finally, its initial amino acid is not ATG but ATA. The second pseudogene, *rpl*33, has three stop-codons within its CDS and possesses a GTC as the initial codon. To determine if the stop-codons in these pseudogenes were "corrected" during the RNA-editing process, we compared their sequence against an EST library of *P. vulgaris* cv Negro Jamapa [27]. A cDNA with a perfect match to *rpl*33 sequence was found, indicating that

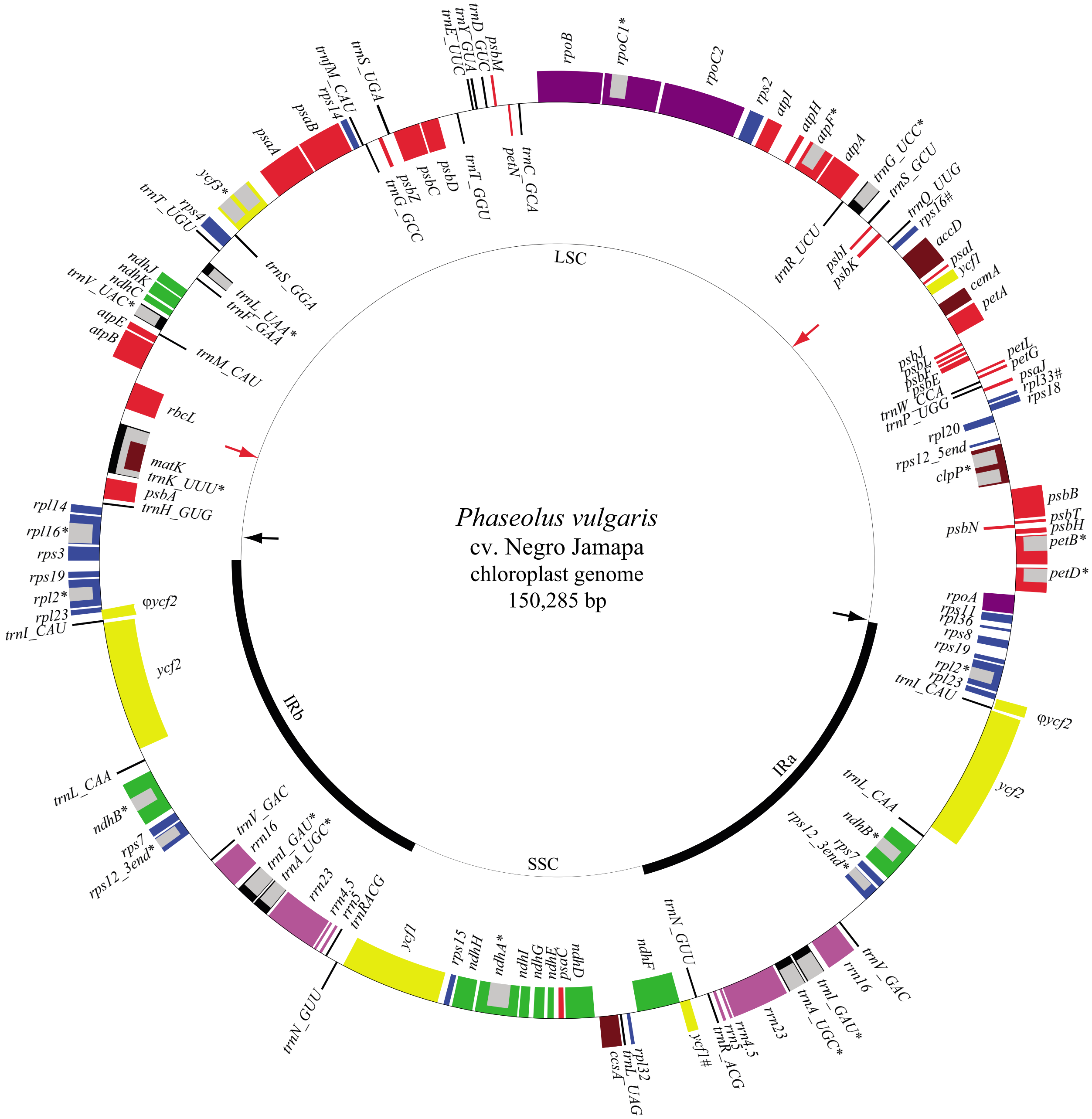


Figure 1

Schematic map of the *Phaseolus vulgaris* plastome. Genes on the outside of the map are transcribed in the clockwise direction and those on the inside are transcribed in the counterclockwise direction. Genes containing introns are indicated by an asterisk. Pseudogenes and incomplete genes are signified by #. Genes are color-coded by function, as shown: blue, ribosomal proteins; red, photosynthesis system; black, transfer RNAs; green, NADH dehydrogenases; yellow, *ycf*; purple, RNA polimerases; light purple, ribosomal RNAs; grey, intron; brown, others. The inner circle shows the quadripartite structure of the plastome. The arrows depict the boundaries of inversions: red arrow indicating the 51 kb-inversion; black arrow indicating the inversion between *trn*H-GUG/*rpl*14 and *rps*19/*rps*8.

this pseudogene was transcribed and that the stop codons were not edited in its mRNA. In contrast, the *rps*16 sequence was not represented in this library. To demonstrate that the presence of these pseudogenes is not a peculiarity of the bean cultivars that we used in this work, the regions containing *rps*16 or *rpl*33 from 10 other varieties of *P. vulgaris*, belonging to two different domestication centers, were amplified by PCR and the products were sequenced. They gave the same sequence, except for 1–3 SNPs (not shown), indicating that their presence is a common characteristic of the species. *P. vulgaris, G. max*, and *L. japonicus* chloroplast genomes contained 21 unique introns. However, *M. truncatula* lacked intron 1 of *clp*P and the intron present in the 3'-end of *rps*12.

Gene order

Each one of four-sequenced legume cpDNAs possessed its own genome structure (Fig. 2). In comparison with the *Arabidopsis* chloroplast genome (outgroup), *L. japonicus* chloroplast genome has almost the same gene order, except for a 51-kb inversion extending from *rbc*L to *rps*16 in the LSC region, which is present in most taxa of the Papilionoideae subfamily of Leguminosae [8,12,22]. In contrast to the plastome of *L. japonicus*, *G. max* cpDNA seems to have a second inversion embracing the region located between LSC and IRs, but is another isomer product of the flip-flop intramolecular recombination present in platomes [28]. *G. max* and *M. truncatula* shared the same gene order but the conspicuous difference between them was the absence of the IRb region in the latter. The *P. vulgaris* cpDNAcontained an inversion at the junction between *trn*H-GUG/*rpl*14 and *rps*19/*rps*8 which was absent in the three other legume chloroplast genomes. We confirmed the presence of this peculiar structure in 10 other *P. vulgaris* varieties originating from Mesoamerican and South American domestication centers, using a concatenated long PCR analysis. This genome inversion has also been reported in the adzuki bean (*Vigna angularis*) [1] and mung bean (*Vigna radiata*) [8]. These results indicate that the structure found in *L. japonicus* cpDNA was closer to the legume ancestral gene order.

IR region

The IR in *P. vulgaris* contained 19 complete genes and spanned 26,426 bp, longer than *G. max* (25,574 bp) and *L. japonicus* (25,156 bp). The *P. vulgaris* duplicated region included the whole *rps*19 gene and 572 bp of its downstream sequence, whereas in both *G. max* and *L. japonicus*, the IRs included only a partial fragment of the *rps*19 gene. Thus, the length increase of IR was principally attributed to the expansion of the IR region at the junction between IR/LSC.

The junction points of IR/LSC were located in 24 bp from the start base of *rps*3 CDS at one end and 53 bp from the start base of *rps*8 CDS at the other. This was exactly like the adzuki bean[1], indicating that this IR predated the speciation of these two bean species, but after the separation from soybean. The boundaries between SSC/IR are located within the *ycf*1 gene and for this reason, 505 bp of this gene's 5'-end is repeated. A similar repetition was found in *G. max* (478 bp) and *L. japonicus* (514 bp), which are shorter than the repeat in *Arabidopsis* (1027 bp).

Indel structure

A number of insertions/deletions (indels) present on cpDNA homologous regions shared by *M. truncatula*, *G. max*, *L. japonicus*, and *P. vulgaris* were detected by DNA alignments. In Figure 3, indels greater than 20 bp are shown. Indels in *P. vulgaris* were principally concentrated at the LSC region, only one was in IRs (24 bp); but deletion was more common than insertion in its cpDNA, which resulted in the reduction of the genome size. In contrast, *M. truncatula* had more and larger indels than other legume plants, and even lost one copy of IR. A large part of the indels was located at the intergenic regions or introns but some of them lay within genes, common in *ycf*1, *ycf*2, *psa*A, *rps*16, *rps*18, and *acc*D.

DNA repeat analysis

All repeated sequences of 20 bp or larger with 100% identity were examined in each of the four legume chloroplast genomes. *M. truncatula* had the largest number of repeats, as described by Saski [23], whereas *P. vulgaris* had the least. Repeats were generally located within the intergenic regions or within introns; however, some of them were present in genes, usually *ycf*1, *ycf*2, *psa*A, and *acc*D.

The biggest direct repeat found in *P. vulgaris* cpDNA was a 287-bp duplication of an internal fragment of *ycf*2 (ψ*ycf*2, Fig. 1). In *P. vulgaris* and *G. max*, this repeat had the same size, while in *L. japonicus* this segment was a little smaller, 265 bp. These two copies in *P. vulgaris* were identical, as well as in *G. max* and in *L. japonicus*, but in *M. truncatula*, it already diverged, sharing 56% of identity. Palindromic repeats were normally situated within intergenic regions and in proximity to the gene end. In *P. vulgaris*, an identical 20-bp-sized palindromic sequence was found within 70 bp from the ends of genes *trn*H-GUG, *ycf*3, and *ycf*1, indicating that they could have the same function.

Tandem-repeat analysis

The distribution of tandem repeats in the legumes cpDNAs is shown in Table 1. *Phaseolus* has five groups of tandem repeats, the smallest number of the sequenced legume cpDNAs. One repetitive unit of 16 bp was duplicated four times within the IR region and was located close to the boundaries of IR/LSC (coordinate positions: 80116–80179 and 149929 – 149992). The alignment of this region with the corresponding sequences of other leg-

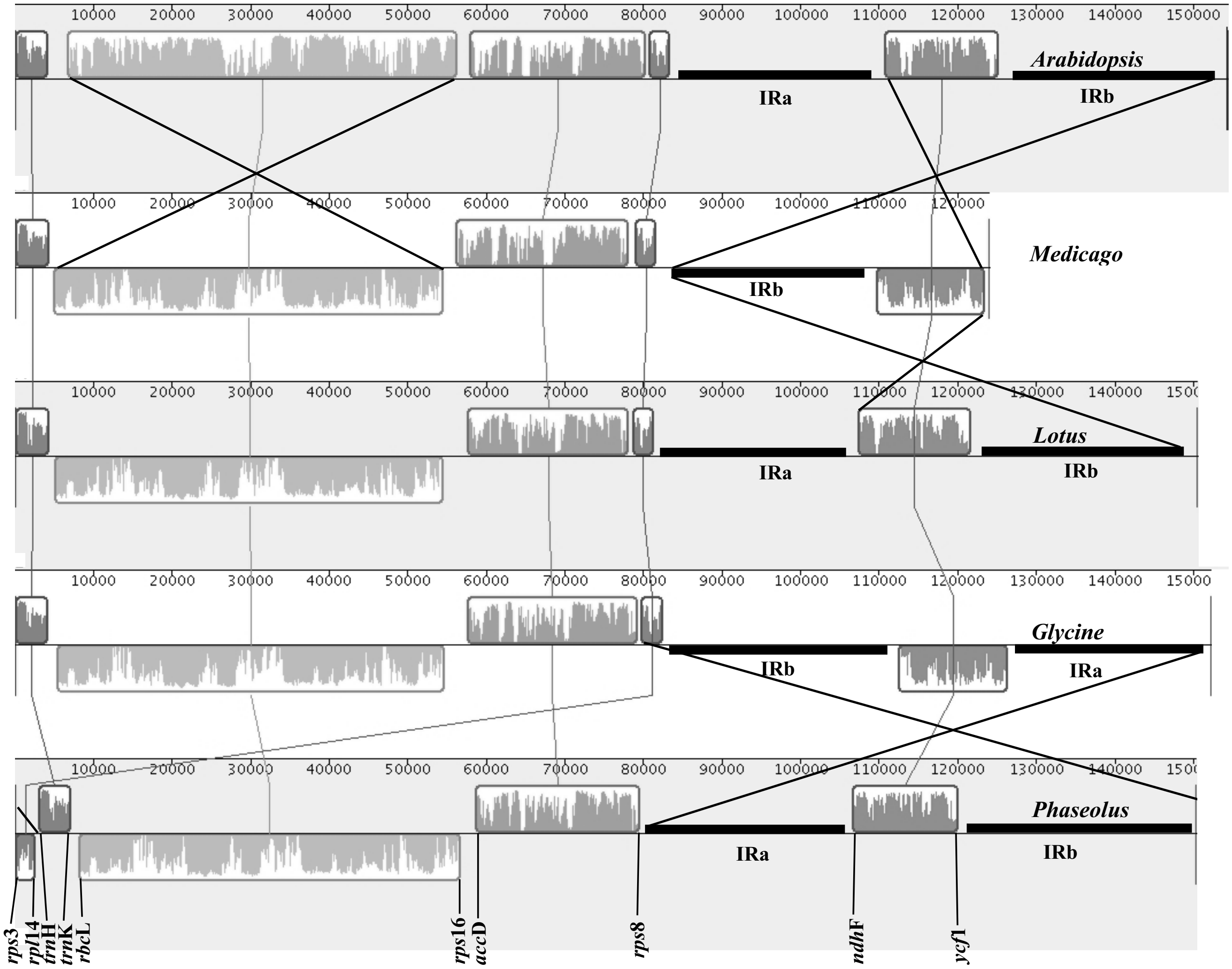


Figure 2

Gene order comparison of the legume plastome, with *Arabidopsis* as a reference, is principally produced by MAUVE. The boxes above the line represent the gene complex sequences in clockwise direction and the boxes below the line represent those sequences in the opposite direction. The gene names at the bottom indicate the genes that are located at the boundaries of the gene complex of the *P. vulgaris* plastome.

ume cpDNAs available from Genbank showed that adzuki bean possessed this duplicated tandem repeat, but with three repeated units each. *G. max* and *L. japonicus* lost this sequence. However, *M. truncatula* had only one 16-bp unit with 75% identity at this position.

M. truncatula had a similar number of reverse and palindrome repeats to other legume plastomes but had a higher proportion of tandem repeats (2% of its genome), compared to other legume cpDNAs. The majority of tandem repeats were located within coding regions of *acc*D, *ycf*1, and *ycf*2 genes and into intergenic regions between *clp*P/*rps*12-5'end and *ycf*1/*trn*N. For example, the *acc*D gene contained seven kinds of repeats in tandem from two to five copies. Of all tandem repeats found in *M. truncatula*, only one (coordinate number: 37267–37401) in *ycf*2 was, to a different extent, shared by all the legume plastomes. Consensus sequences of repetitive units of each tandem repeat present in *M. truncatula* cpDNA were obtained and searched in the other legume cpDNAs. The consensus sequences of repeats within *ycf*1, *ycf*2, *rps*18, and *psa*A were found in the other genomes but as single sequences (not repeated).

The largest tandem repeat in *M. truncatula*, spanning 286 bp, was situated at the end of *clp*P (coordinates 55590 and 55875), and it was exclusively found in cpDNA of this plant. It consisted of two identical tandem copies of 143

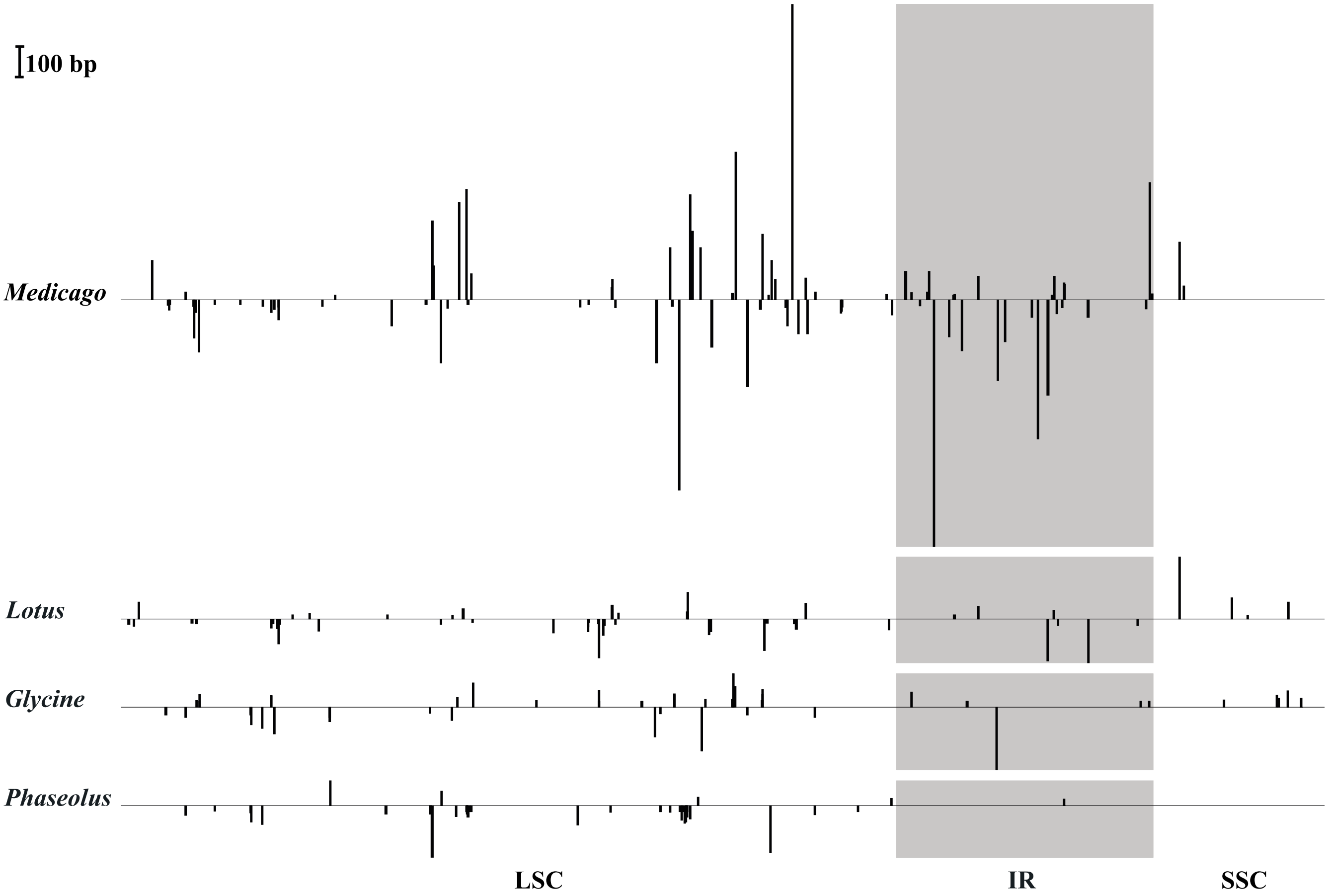


Figure 3
Indel profiles of legume plastomes. Indels were identified by the sequence alignments with Clustal-X [66]. The black bars above the horizontal axis indicate insertion and those below the axis show deletion. The height of the boxes represents the size of indel fragments. The sequence order is shown as in *P. vulgaris*. The shadow region represents one IR and another IR was removed from the figure.

bp, repeats A and B (Fig. 4). In fact, this segment was also composed of six copies of a smaller repeated unit of approximate 48 bp, of which some copies were altered by a few bases (a1, b1) or had some base insertions (a2, a3, b2, b3), but the backbone was conserved. This structure suggests that the 48 bp was first duplicated two consecutive times, and then each of these units underwent some degree of diversification to form the 143 bp. More recently, this last element was duplicated. Similar situations were found in the *acc*D gene and the intergenic region *ycf*1/*trn*N.

Phylogenetic analysis

Legume chloroplast phylogenies were established using a phylogenomic approach and the phylogenetic information of individual genes. In our analyses, we always used the *A. thaliana* chloroplast genome as the outgroup. From the phylogenomic perspective, we made two large alignments: one with all homologous regions of the five cpDNAs but excluding the paralogous regions, and the other, by pasting together the individual alignments of 102 individual genes. Both gave similar tree topologies, forming two subgroups with a bootstrap value of 100: *Phaseolus* with *Glycine* and *Medicago* with *Lotus* (Fig. 5a, b), which correspond to the previously well-established phylogeny [21]. It was apparent that, in the group of *Phaseolus* with *Glycine, Phaseolus* has accumulated more substitutions than *Glycine*, thus *Phaseolus* diversified much faster (2.3 times), while *M. truncatula* and *L. japonicus* has a similar substitution rate (Fig. 5a, b). To support the phylogeny obtained with genomes, we also did phylogenies with each of the 75 protein-encoding genes (*rps*12 is a divided gene: its -5' and 3'ends were considered here as two genes because they are encoded at different loci; *ycf*4 was not used due to the absence in *M. truncatula* and *L. japonicus* plastomes). Ribosomal RNA and transfer RNA genes were not included because of fewer base substitutions. 60 protein-coding genes produced phylogenies with bootstrap

Table 1: Distribution of tandem repeats (> 15 bp with 80% identity between copies) in four legume plastomes

	Initial position	Final position	Size (≥ 15 bp)	Copies	Identity (≥ 80%)	Position related genes
Phaseolus	66513	66572	15	4	80	*psa*J/*rpl*33
	65733	65783	17	3	92.2	*trn*W/*trn*P
	80116	80179	16	4	98	*rps*8/*rps*19, or *rps*3/*rps*19
	85700	85762	21	3	88.9	*ycf2*
	88119	88172	18	3	88.9	*ycf2*
Lotus	1694	1765	24	3	81.9	*psb*A/*trn*K
	14487	14543	19	3	84.2	*trn*L/*trn*T
	17838	17888	17	3	86.3	*ycf3*, intron
	24441	24492	26	2	100	*trn*G/*ycf9*
	47831	47878	16	3	96	*atp*H/*atp*F
	54191	54265	25	3	80	*psb*K/*trn*Q
	87031	87093	21	3	91	*ycf2*
	89444	89524	27	3	95	*ycf2*
	106513	106572	20	3	83.3	*trn*N/*ycf1*
	109580	109642	21	3	84.1	*ndh*F/*rpl*32
Glycine	28572	28640	23	3	81.2	*psb*D/*trn*T
	51493	51555	21	3	84.1	*atp*A/*trn*R
	51753	51818	22	3	86.4	*trn*R/*trn*G
	58325	58396	24	3	80.6	*acc*D/*psa*I
	64627	64674	24	2	96	*pet*G/*trn*W
	66304	66345	21	2	100	*rpl*33/*rps*18
	68386	58429	22	2	100	*clp*P/*rps*12_5'-end
	81892	81954	21	3	85.7	*rpl*16,*rps*3
	82665	82718	18	3	85.2	*rps*3,*rps*19
	83848	83901	18	3	85.2	*rpl*2, intron
	88334	88396	21	3	91	*ycf2*
	89622	89663	21	2	100	*ycf2*
	90774	90827	18	3	85.2	*ycf2*
	108203	108252	25	2	96	*trn*N/*ycf1*
	123651	123710	20	3	85	*trn*L/*rpl*32
	127141	127190	25	2	96	*ycf1*
Medicago	13248	13319	24	3	84.7	*rps*15/*ycf1*
	17087	17158	24	3	100	*ycf1*
	18922	19013	46	2	100	*ycf1*/*trn*N
	18847	19031	37	5	84.3	*ycf1*/*trn*N
	19100	19219	60	2	100	*ycf1*/*trn*N
	27448	27617	85	2	93	*rrn*16/*trn*V
	36490	36669	60	3	98	*ycf2*
	37267	37401	45	3	83.7	*ycf2*
	38869	38940	36	2	100	*ycf2*/*trn*I
	38954	38997	22	2	100	*ycf2*/*trn*I
	39247	39368	61	2	89	*trn*I/*rpl*23
	55590	55875	143	2	100	*clp*P/*rps*12_5'-end
	55807	55920	57	2	88.6	*clp*P/*rps*12_5'-end
	56146	56265	24	5	95	*clp*P/*rps*12_5'-end
	56392	56466	25	3	100	*clp*P/*rps*12_5'-end
	58382	58441	15	4	90	*rps*18
	58799	58867	23	3	81.2	*rps*18/*rpl*33
	65523	65586	32	2	96.9	*cem*A/*psa*I
	67538	67702	33	5	91.5	*acc*D
	67639	67818	60	3	98	*acc*D
	68026	68214	63	3	100	*acc*D
	68251	68322	24	3	88.1	*acc*D
	68311	68436	63	2	93	*acc*D
	68577	68624	24	2	86.7	*acc*D
	68907	68954	24	2	96	*acc*D
	69341	69422	41	2	96	*acc*D/*trn*Q
	91311	91394	28	3	81	*trn*C/*pet*N
	99689	99742	18	3	92.6	*psb*Z/*trn*G
	105175	105222	24	2	98	*psa*A

```
   a1:   1 CAAATAATGACATTCAAAAAAAAAGGAGTTAACTAATGTCATTATATGA 49
A  a2:  50 CA-TTAGTTAAATCC-AAAAAAAAGGAGTTAACTAATGTCATATAATGA 96
   a3:  97 CA-TTAGTTAAATCC-AAAAAAAAGCAGTTAACTAATGTCATTATATGA 143
   b1: 144 CAAATAATGACATTCAAAAAAAAAGGAGTTAACTAATGTCATTATATGA 192
B  b2: 193 CA-TTAGTTAAATCC-AAAAAAAAGGAGTTAACTAATGTCATATAATGA 239
   b3: 240 CA-TTAGTTAAATCC-AAAAAAAAGCAGTTAACTAATGTCATTATATGA 286
```

Figure 4
Largest tandem repeats in *Medicago* at the coordinate of 55590 and 55875. Repeats A and B are respectively composed of smaller tandem repeats, a1-3 and b1-3.

values higher than 50. These 60 phylogenies were classified into five topologies: three of them were obtained more frequently (Fig. 5c–e) and the other two topologies were only supported by single genes (not shown). The most frequent topology, representing 28 genes (47%), matched the topology obtained with phylogenomic analysis. Topologies D and E represent phylogenies of 18 (30%) and 12 (20%) genes, respectively. In all of these topologies *G. max* and *P. vulgaris* made a cluster, but *M. truncatula* or *L. japonicus* differed in the relation to *A. thaliana*, the outgroup. It is important to point out that phylogenies obtained with *mat*K and *rbc*L (topology D), two genes commonly used in plant phylogenic analysis, do not fit the genome-based topology, suggesting that care must be taken in interpreting data obtained with these gene-markers.

Relative evolutionary rate

The genome-based phylogenies indicate that legume chloroplast genomes change at different rates. To identify which genes and to what extent these genes contribute to the overall evolutionary rate, a relative rate test was performed. The relative rates between *Phaseolus* and *Glycine* and those between *Medicago* and *Lotus* in K, Ks, and Ka of all protein-coding genes were determined. Considering that the outgroup plastome could affect, to some extent, the analysis, each relative test employed one of three different genomes alternatively as an outgroup. The relative rate tests between *P. vulgaris* and *G. max* were evaluated using as a reference species, *A. thaliana, M. truncatula*, or *L. japonicus*. Similarly, the relative rate tests between *M. truncatula* and *L. japonicus* were calculated using *A. thaliana, P. vulgaris*, or *G. max* as reference group.

In the comparing *P. vulgaris* and *G. max*, we found a number of *P. vulgaris* genes with a strong tendency to evolve faster, despite the different reference species used (Fig. 6). All the genes with statistical significance ($p < 0.05$) K, Ka, and Ks values also produced the same results (Fig. 6, Tables 2 and 3). We therefore concluded that there was faster diversification of the *P. vulgaris* plastome than *G. max* at the genomic level. Comparing *M. truncatula-L. japonicus*,12 genes evolved at a significantly different rate (K), 10 of which accumulated more substitutions in *M. truncatula* (Fig. 6A, B, and 6C), and two of which had more substitutions in *L. japonicus*.

In both groups, P. vulgaris-G. max and M. truncatula-L. japonicus, all the pet, psa, psb, and atp genes showed no significant difference in substitution rates, and six genes (accD, ycf1, ycf2, clpP, ndhF, and rpoC2) evolved at different rates (Tables 2 and 3, Fig. 6). Some genes containing significant differences in the group P. vulgaris-G. max did not demonstrate significant differences in M. truncatula-L. japonicus. This result suggests that, in legume plastomes, some genes showed similar evolutionary tendency and others diversified faster in a particular plastome. accD and ycf2 presented different rates of both synonymous and nonsynonymous changes, implying that these genes have low functional compromise. Moreover, accD and ycf2 had a ω index (Ka/Ks) higher than 1, indicating that they are subjected to a strong diversifying process. The rest of the genes with significant change rates had a ω index lower than 1, showing that these genes are under purifying selection.

Discussion

Gene order and gene content of legume plastomes

In contrast to the genome organization in *A. thaliana*, most taxa of the subfamily Papilionoideae, including the four species of which plastomes are sequenced, present a 51-kb inversion within the LSC region [12]. Another inversion at the junction points of *trn*H-GUG/*rpl*14 and *rps*19/*rps*8 was only reported to occur in two genera, *Phaseolus* and *Vigna*[1,19,29], indicating that this chloroplast genome arrangement is characteristic of the *Phaseolus-Vigna* species complex. The chloroplast genome of *M. truncatula* lacks one IR, a feature shared with other legume

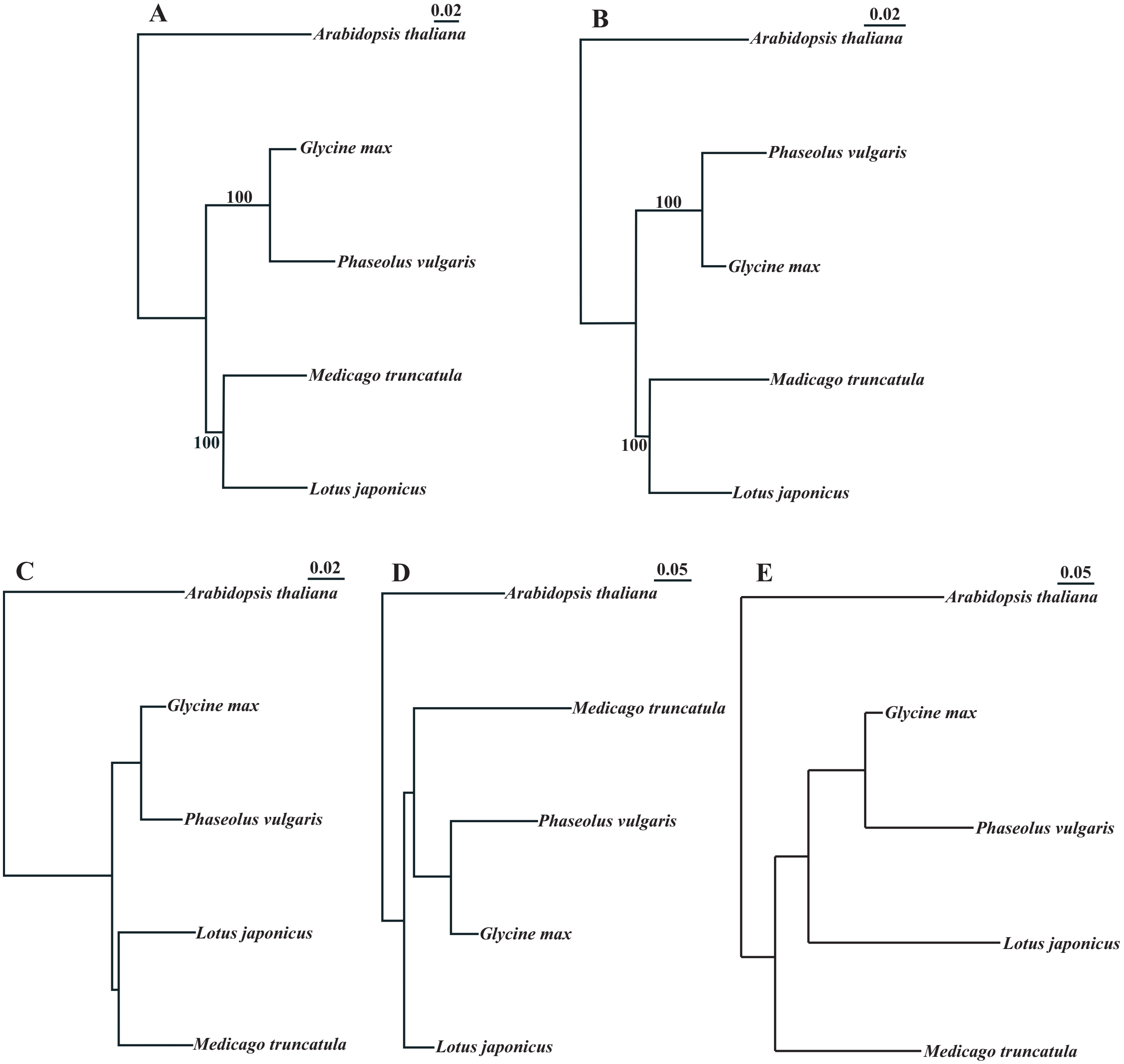


Figure 5
Diagrams of phylogenetic trees. Topology A was deduced from all the genome sequences and B was based on all the genes. C, D, E are different topologies of individual gene phylogenies.

tribes such as Carmichaelieae, Cicereae, Galegeae, Hedysareae, Trifolieae, and Vicieae and some genera of other groups [13]. Now, all these tribes form a new clade, IRLC (inverted-repeat-loss clade) [30]. Thus, the four-sequenced plastomes represent three types of plastome structure, suggesting that the cpDNA organization is very diverse in legume plants.

Legume cpDNAs do not contain *rpl*22 [31,32] and *inf*A [33] genes, indicating that they were phylogenetically lost from this lineage. A specific character of *P. vulgaris* cpDNA is the presence of the two pseudogenes *rps*16 and *rpl*33. The first is functional in *L. japonicus* and *G. max* but is lost in *M. truncatula* [23,32]. The cpDNAs of other land plants, *Selaginella uncinata*, *Psilotum nudum*, *Physcomitrella patens*, *E. virginiana*, and *Eucalyptus globules*, lost this gene inde-

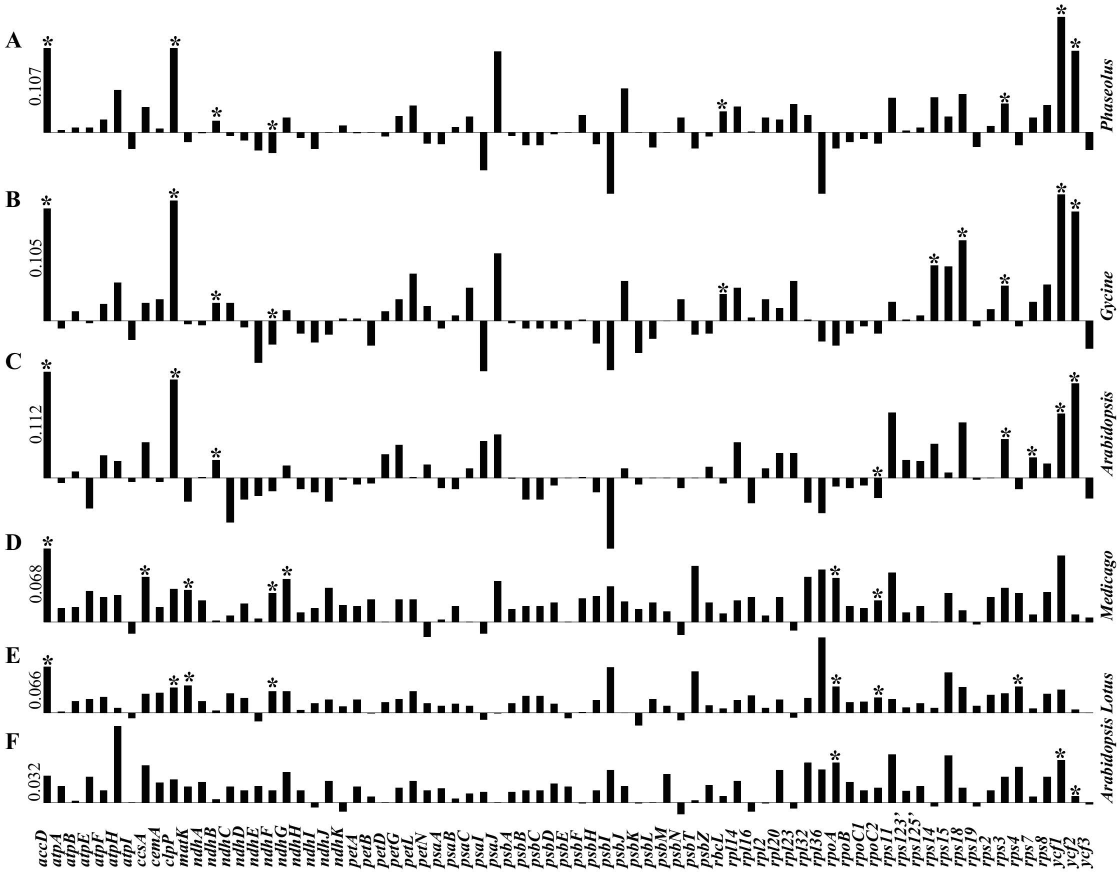


Figure 6

Diagrams of differences in evolutionary rates of "K", the number of nucleotide substitutions per site, of 75 protein-coding genes. Panels A, B, and C represent the variances in relative rates between *Medicago* and *Lotus* using the reference plastomes, respectively, as *Phaseolus, Glycine*, and *Arabidopsis*. Panels D, E, and F show those between *Phaseolus* and *Glycine* using the reference plastomes, respectively, as *Medicago*, *Lotus*, and *Arabidopsis*. The height of the black bar denotes the value of variances (the first bar showed the value, as a scale of this panel). The bars above the axis mean *Medicago* with higher substitution rates than *Lotus* in Panels A, B, and C or *Phaseolus* with higher substitution rates than *Glycine* in Panels D, E, and F and the bars below the axis represent the opposite case. The asterisk is a sign of significant difference (P < 0.05).

pendently [4,34,35]. *rpl*33 is a functional gene basically present in all land plant chloroplasts, except in *S. uncinata*. These data suggested that *P. vulgaris* cpDNA is still undergoing genome reduction.

The *acc*D gene encodes an acetyl-coenzyme A carboxylase subunit similar to prokaryotic *acc*D in structure[36], and is the most variable gene present in legume chloroplasts. Its size is widely different: 1299 bp in *G. max*; 1422 bp in *P. vulgaris*; 1506 bp in *L. japonicus*, and 2142 bp in *M. truncatula. Medicago* has the largest *acc*D of prokaryotic form, containing seven kinds of tandem repeats and one 43-bp-sized separate direct repeat situated between two conserved regions. We did a BLAST-search with the *acc*D gene against the EST bank of *M. truncatula*. One tentative consensus segment of 9334 bp (TC106672) was found to contain the identical sequences of chloroplast genes *trn*S-GCU, *trn*Q-UUG, *psb*I, *psb*K, *acc*D, *psa*I, *cem*A, and *pet*A, indicating that these genes are transcribed. Nevertheless, the large amount of tandem repeats present in the *M. truncatula acc*D gene calls into question its functionality.

Another landmark of the legume plastomes is the duplication of a portion of *ycf*2. The duplicated segment, named ψ*ycf*2, was first identified as a pseudogene in *Vigna angularis* [1]. It is present in the same relative position in the legume plastomes analyzed here. In *G. max*, *P. vulgaris* and *L. japonicus*, ψ*ycf*2 is identical to its copy within *ycf*2, but in *M. truncatula* they are very divergent (60 % of identity). This result indicates that the last common ancestor of these plants already had this duplication and gene conversion occurred in the plastomes containing IR.

Table 2: Synonymous (Ks) and Nonsynonymous (Ka) substitution rates of *P. vulgaris* and *G. max*.

	Arabidopsis as a reference		*Lotus* as a reference		*Medicago* as a reference	
	Ka	Ks	Ka	Ks	Ka	Ks
	Pha.#/Gly.	Pha./Gly.	Pha./Gly.	Pha./Gly.	Pha./Gly.	Pha./Gly.
*acc*D	---	---	0.1769/0.1234	0.1265/0.0572	0.2587/0.2096	0.2354/0.1512
*ccs*A	---	---	---	---	0.1061/0.0782	---
*clp*P	---	---	0.0603/0.0284	---	---	---
*mat*K	---	---	0.1692/0.1371	---	0.1633/0.1365	---
*ndh*F	---	---	0.1065/0.082	---	0.094/0.0723	---
*ndh*G	---	---	0.0609/0.0323	---	0.0659/0.0309	---
*psb*D	---	---	---	---	---	0.2203/0.1492
*rpo*A	0.1356/0.1026	---	0.0803/0.0552	---	0.0747/0.0493	---
*rpo*B	0.0685/0.0562	---	0.0535/0.0425	---	0.0472/0.0361	---
*rpo*C1	---	---	0.0497/0.0375	---	---	---
*rpo*C2	---	---	0.1123/0.0961	---	0.1089/0.093	---
*rps*15	0.1906/0.1207	---	0.1166/0.0613	---	---	---
*rps*2	---	---	0.0669/0.0442	---	---	---
*rps*4	---	---	0.0635/0.0389	---	---	---

# Pha. and Gly. represent respectively *Phaseolus* and *Glycine*.

Nature of tandem repeats

The sequence and distribution of repetitive elements are characteristic of each chloroplast genome, and they can be classified in two broad categories: large repeats and short dispersed repeats (SDRs). Both categories can be found in different proportions in chloroplast genomes. *Oenothera* and *Triticum* chloroplasts contain some dispersed repeats, but 20% of the *Chlamydomonas reinhardtii* plastome consists of repeated sequences, many of them are tandem repeats (TR) [37-39]. In legume plastomes, clear differences reside in the number, location, and sequence of TR. *M. truncatula* possess a plastome with greater number and larger TRs, and *P. vulgaris* has a plastome with fewer TRs.

Usually, TRs are classified as a subcategory of SDRs, but our analysis of the legume chloroplast genomes shows

Table 3: Synonymous (Ks) and Nonsynonymous (Ka) substitution rates of *M. truncatula* and *L. japonicus*.

	Arabidopsis as a refernce		*Glycine* as a reference		*Phaseolus* as a reference	
	Ka Med.#/Lot.	Ks Med./Lot.	Ka Med./Lot.	Ks Med./Lot.	Ka Med./Lot.	Ks Med./Lot.
*acc*D	0.2822/0.1869	0.2074/0.1222	0.2096/0.1234	0.1512/0.0572	0.2587/0.1769	0.2354/0.1265
*atp*A	---	---	0.0184/0.0092	0.2455/0.3494*	0.0226/0.0134	---
*atp*B	---	---	0.0232/0.0128	---	---	---
*atp*H	---	---	---	---	0.0218/0	---
*clp*P	0.1803/0.0668	---	0.1503/0.0284	---	0.1734/0.0603	---
*ndh*B	---	0.1297/0.0754	---	0.1189/0.0633	---	0.1126/0.0652
*ndh*E	---	---	---	0.186/0.4588*	---	---
*ndh*F	---	---	---	0.4389/0.5855*	---	0.4925/0.6659*
*pet*B	---	---	---	0.2429/0.3904*	---	---
*psa*B	---	0.3872/0.4844*	---	---	---	---
*rbc*L	---	---	---	---	---	0.5606/0.3989
*rpo*C2	---	0.3959/0.4735*	---	---	---	---
*rps*11	---	---	0.0596/0.0274	---	---	---
*rps*14	---	---	0.0706/0.0219	---	0.0704/0.0308	---
*rps*18	0.1263/0.0718	---	0.1107/0.0397	---	0.1152/0.0648	---
*rps*3	0.1048/0.069	---	0.0862/0.0442	---	0.1013/0.0602	---
*rps*7	0.0392/0.0055	---	0.0332/0.0055	---	0.0417/0.0137	---
*ycf*1	---	---	0.178/0.0946	0.2648/0.0946	0.2192/0.1205	0.3529/0.1409
*ycf*2	0.161/0.0674	0.1576/0.0681	0.1487/0.054	0.1481/0.0535	0.1556/0.0588	0.1511/0.058

* The star signal represents *Lotus* genes with higher substitution rates than *Medicago* genes.
#Med. and Lot. indicate respectively *Medicago* and *Lotus*.

that TRs have a different origin from the rest of the SDRs. The repetitive unit of an SDR family is dispersed throughout the genome and different members of an SDR family share high identity. In contrast, the repetitive unit of a TR is not dispersed, and the consensus sequence of each TR has low identity with the consensus sequences of other TRs, with the exception of some repeats with low complexity (*i. e.* ATATAT). In other words, each TR is specific to a site.

Multi-alignments among plastomes frequently show that a repetitive consensus unit of a TR can be found in other chloroplast genomes at similar positions without duplication, or the region containing corresponding sequences are completely deleted from a specific plastome. Moreover, some small insertions from 7 bp to 21 bp are the duplication events of one of the flanking sequences in a specific plastome to form a small TR (only two tandem units). On the other hand, more complicated TRs by consecutive duplication, as shown in Figure 4, also exist in other sites of the plastome. Taking together our observations, we conclude that TRs came from *in situ* sequences and do not share the same origin of dispersed repeats.

We propose that homology-facilitated illegitimate recombination is the mechanism that creates TRs. The reasons are: 1). TRs arise from *in situ* sequences, actually from 7 bp to 143 bp long in the present study; 2) About 4–17 bp initial bases of some larger insertions are the iteration of their flanking sequence; 3) There are many copies of the plastome in a cell, both in circle and in linear forms, which provide the opportunity of such recombination; 4) Homology-facilitated illegitimate recombination is corroborated by the gene transformation in the chloroplast of *Acinetobacter* sp. [40]. Recombination mediated by short direct repeats was reported in wheat chloroplast [15].

Intracellular sequence exchange

Recently, Kami reported the sequence from a nuclear BAC clone, 71F18, containing a chloroplast-derived DNA of *P. vulgaris* [41]. The sequence comparison between the *P. vulgaris* plastome and the BAC clone showed that two separate regions (*trn*G-*rps*14 in 914 bp, *trn*I-*ndh*B in 7901 bp) in the plastome were linked together in the nuclear genome, with the same similarity (99.01%) to their nuclear homologues. We noted that the nuclear homologues did not contain the insertion in comparison with its plastome sequence, but had 8 deletion segments ranging in size from 8 bp to 583 bp. We therefore postulate that the original fragment transferred from the plastome, likely spanned the whole fragment from *trn*I-GAU to *rps*14 (73 kb), and then some deletions occurred, including the deletion of 64 kb fragment from *trn*L to *psbZ*.

A BLAST-search of the *M. truncatula* plastome sequences with available nuclear genome sequences of this species found that 51% of the plastome is present in the nuclear genome with more than 99% identity. These identified chloroplast-derived segments of the *M. truncatula* nuclear genome can be as large as 25 Kb. One must take into account that we only had the opportunity to explore a partial nuclear genome that is available up to date in Genbank, suggesting that the whole plastome could be found in the nuclear genome if the complete nuclear genome becomes available. If so, it is similar to the case of the rice genome [42], but different from *A. thaliana*, in which the chloroplast-derived fragments found in the nuclear genome have a lesser degree of identity (commonly 92–98%) and the transferred fragments are smaller in size, generally less than 4 kb, indicating that cpDNA transfer occurs earlier in the *A. thaliana* genome. In the rice genome, cpDNAs are continuously transferred to the nuclear genome, which incessantly eliminates them, until an equilibrium is reached [42]. On the other hand, we did not find significant similarity between the plastome of *L. japonicus* and its nuclear genome. There are several hypotheses to explain the gene transfer from chloroplast to nuclear genomes [43]. The most common mechanism of transfer depends on chloroplast lysis, but it is still difficult to elucidate why the nuclear genome of *A. thaliana* did not integrate cpDNA with the same patterns as *M. truncatula* or *O. sativa*.

Rate of evolutionary change in legume plastomes

There are only a few reports that describe the evolutionary rate of the chloroplast genome [44-46]. In the present study, we demonstrate that one plastome (*P. vulgaris*) globally evolved faster than another plastome (*G. max*), which has not been observed before.

In regard to the evolutionary rate of legume plants, Lavin reported that *Phaseolus* and closely related genera have the fastest substitution rates at the *mat*K locus, within Leguminosae [21]. Delgado-Salinas recently suggested this accelerated substitution rate in *mat*K (within the intron of *trn*K) is related to the formation of the modern Trans-Mexico volcanic belt [47]. We present further evidence here that the *Phaseolus* plastome genomically diversified rapidly. Considering that all the genes in this genome were affected, we deduced that some factor likely impacted this plastome globally, leading to a higher rate of evolutionary change.

Evolutionary rate can be mainly affected by the following factors: generation time, population size, specific mutation rate, and natural selection [48]. The first three factors should influence all the genes of a genome as a whole, whereas the third is able to impinge on specific genes. Generation time is usually considered as an important cause for acting on the evolutionary rate, and has been applied in the elucidation of the discrepancy of evolutionary rates between rodents and other mammals [49], between the plastomes of *Phalaenopsis aphrodite* and grass crops [50], and between rice and maize [46]. However, it cannot be applied to explain the phenomenon in the present study because both *G. max* and *P. vulgaris* are annual crop plants, sharing the same generation time. Population sizes of *G. max* and *P. vulgaris* cultivars seem to be similar because they are important domesticated plants with a highly limited genetic diversity [51]. The divergent mutation rate could be one of the causes of the variance in the substitution rate between *Phaseolus* and *Glycine*. The reasons are: 1) overall Ks in *Phaseolus* is much higher than *Glycine* (see Additional File 1); 2) the sites of synonymous substitution are far from saturation in this plastome (< < 1); 3) and these two crop plants have the same generation time and similar reproductive mode (self-fertilization), which prevents genetic recombination from other plants; and 4) the chloroplast is rarely imported from other compartments of a cell as genetic elements. On the other hand, natural selection should be a factor for the relative rate of specific genes. The present research shows that almost all genes are under a purifying selection ($\omega < 1$). Therefore, we conclude that the different evolutionary rate between *Phaseolus* and *Glycine* is a consequence of the pressures of both mutation and natural selection.

The *M. truncatula* and *L. japonicus* plastomes evolved at a similar rate (K). However, the genes with significant differences showed a remarkably distinct rate: 10 *M. truncatula* genes evolved significantly faster than did their *L. japonicus* counterparts, but two genes, *rpo*C2 and *ndh*F, changed faster in *L. japonicus*. In this case, it seems that the particular reason that leads to faster evolution of some genes in one plastome must be natural selection.

Conclusion

Plastomes of leguminous plants have evolved specific genomic structures. They have undergone diversification in gene content, gene order, indel structure, abundance and localization of repetitive sequences, intracellular sequence exchange and evolutionary rates. In particular, the *P. vulgaris* plastome globally has evolved faster than that of *Glycine*.

Methods

Biological materials

The *P. vulgaris* cultivars used in this work were Negro Jamapa, Pinto V1-114, Kentucky wonder, Carioca, Olathe, Othello, MSU Fleet Wood, Jalo EEP558, and BAT93, derived from the mesoamerican domestication center and Cardinal and Red Kloud, derived from the Andean domestication center.

Chloroplast DNA extraction, DNA sequencing, and genome annotation

P. vulgaris cv. Negro Jamapa cpDNA was isolated from intact chloroplasts using the method reported by Jansen [52]. To construct the shotgun library, DNA was fragmented by nebulization. Fragments between 2 and 5 kb were recovered from 1% agarose gel, blunt-ended, and cloned in pZERO™-2 in its *Eco*RV site (Invitrogen). Recombinant clones were sequenced using the Dye-terminator cycle sequencing kit (Perkin Elmer Applied Biosystems, USA). Sequencing reactions were run in an ABI 3730 sequencer (Applied Biosystems). To seal small gaps, specific regions were amplified by polymerase chain reaction (PCR), and the obtained products were sequenced. Assemblages were obtained using the PHRED-PHRAP-CONSED software [53,54] with a final quality of < 1 error per 100,000 bases. Genome annotation was performed with the aid of the DOGMA program [55]. The start and stop codons and the boundaries between introns and exons for each protein-coding gene were determined by comparison with other published chloroplast genomes using BLASTX [56]. We also annotated the *M. truncatula* plastome because its annotation is not available from Genbank.

PCR amplification

Concatenated long PCR was adopted to confirm the gene order of the *P. vulgaris* chloroplast genome and to analyze the gene order of closely related bean varieties. Primers for amplifying the whole genome as overlapping segments are shown in Additional File 2. The pairs of primers for the amplification of pseudogenes, *rps*16 and *rpl*33, were: *rps*16F (5'-tgtagcgaatgaatcaatgc-3'), *rps*16R (5'-tgccttactcaatgtttgttc-3'); *rpl*33F (5'-aaattcggagtgaaactcg-3'), *rpl*33R (5'-tctcagtcgactcgctttt-3'). PCR assays were performed in a 25 μl reaction volume containing 250 ng template DNA, 1× reaction XL buffer II, 1.1 mM $Mg(OAc)_2$, 200 μM dNTPs, 5 pmol of each primer, and 1 unit of rTth DNA polymerase XL (Perkin Elmer). PCR amplifications were carried out in a 9700 thermocycler (Perkin Elmer) with the following conditions: an initial denaturation at 94°C for 1 min; 30 cycles of denaturation at 94°C for15 s, annealing and extension at 62°C for 3–15 min (depending on the fragment size needed to amplify); and a final extension at 72°C for 7 min.

Genome analysis

Gene order comparison between the chloroplast genomes of *P. vulgaris* (DQ886273), *A. thaliana* (AP000423),*G. max* (DQ317523),*L. japonicus* (AP002983), and *M. truncatula* (AC093544) was performed with MAUVE [57]. REPuter [58] was used to identify the number and location of direct, reverse, and palindromic repeats of genomes with minimum identical repeat size of 20 bp.

Meanwhile, Equicktandem and Etandem [59] were applied to find the distribution of tandem repeats.

Evolutionary analysis

Genes were defined as homologs with the criterion of E value, 1×10^{-12}, in a BLAST search, using as queries the *P. vulgaris* genes against other chloroplast genomes mentioned above [56]. Two big alignments were made. The first one was a multigenome alignment produced by MAUVE [57]. The second one was constructed by two steps: creating the homologous alignments of each of 74 individual protein-encoding genes that had at least one copy in each genome by MUSCLE [60] and then pasting all the individual gene alignments together to form a big one (concatenated alignment). Alignments were edited to exclude gap-containing columns.

A DNA substitution model was selected using Akaike information criterion with Modeltest, version 3.7 [61]. For the alignments described earlier, the General Time Reversible (GTR) model, including rate variation among sites (+G) and invariable sites (+I), was chosen as the best fit. One thousand replicates were generated with SEQBOOT. Phylogenies were constructed using PHYML [62] and DNAPARS and the consensus phylogenetic tree was obtained with CONSENSE. For each of the 74 individual gene alignments, a phylogeny was produced with PHYML, using a nonparametric bootstrap analysis of 100 replicates. TREEDIST was used to estimate how many different topologies there are, but only the topologies with nonparametric bootstrap values higher than 50 were considered. SEQBOOT, DNAPARS, CONSENSE, and TREEDIST were downloaded from the PHYLIP package version 3.61 [63].

The number of nucleotide substitutions per site "K" was calculated with MEGA3 [64]. The number of nucleotide substitutions per synonymous site "Ks" and the number of nucleotide substitutions per nonsynonymous site "Ka" were deduced with yn00 from PAML13.14 [65]. Based on these data, K, Ks, and Ka, a triplet relative rate test was employed to evaluate the evolutionary rate difference between *P. vulgaris* and *G. max* or that between *L. japonicus* and *M. truncatula*.

Abbreviations

IR, inverted repeat; SSC, small single copy; LSC, large single copy; *ycf*, hypothetical chloroplast reading frame; *rrn*, ribosomal RNA; cpDNA, chloroplast genomic DNA; CDS, coding sequences; EST, expressed sequence tags; SNPs, single nucleotide polymorphisms; K, the number of nucleotide substitutions per site; Ka, the number of nucleotide substitutions per nonsynonymous site; Ks, the number of nucleotide substitutions per synonymous site; ω, the index of Ka/Ks; SDRs, short dispersed repeats; TRs, tandem repeats;

Additional material

Additional file 1

Average synonymous (Ks) and nonsynonymous (Ka) substitution rates ofprotein-coding genes in the P. vulgaris *or* G. max *plastomes. The data show average synonymous (Ks) and nonsynonymous (Ka) substitution rates of 75 protein-coding genes derived from comparing* P. vulgaris *or* G. max *plastomes with the reference plastomes of* A. thaliana, L. japonicus *or* M. truncatula.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-228-S1.doc]

Additional file 2

Primers used for amplifying the complete plastome of the common bean. This file provides the sequences of primers used for amplifying the overlapped PCR products covering the complete plastome of the common bean.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-228-S2.doc]

Acknowledgements

The authors thank Oscar Brito, José Espíritu, Luis Lozano, Ismael Luis Hernández González, and Delfino García for technical and computational assistance. Financial support was by grants from Conacyt 46333-Q and UNAM PAPIIT IN223005.

References

1. Perry AS, Brennan S, Murphy DJ, Kavanagh TA, Wolfe KH: **Evolutionary re-organisation of a large operon in adzuki bean chloroplast DNA caused by inverted repeat movement.** *DNA Res* 2002, **9(5):**157-162.
2. Bendich AJ: **Circular chloroplast chromosomes: the grand illusion.** *Plant Cell* 2004, **16(7):**1661-1666.
3. Martin W, Stoebe B, Goremykin V, Hapsmann S, Hasegawa M, Kowallik KV: **Gene transfer to the nucleus and the evolution of chloroplasts.** *Nature* 1998, **393(6681):**162-165.
4. Wolfe KH, Morden CW, Palmer JD: **Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant.** *Proc Natl Acad Sci U S A* 1992, **89(22):**10648-10652.
5. Jansen RK, Palmer JD: **A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae).** *Proc Natl Acad Sci U S A* 1987, **84(16):**5818-5822.
6. Kim KJ, Choi KS, Jansen RK: **Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae).** *Mol Biol Evol* 2005, **22(9):**1783-1792.
7. Hachtel W Neuss A, Vomstei J: **A chloroplast DNA inversion marks an evolutionary split in the genus Oenothera.** *Evolution* 1991, **45:**1050-1052.
8. Palmer JD Osorio B, Thompson WF: **Evolutionary significance of inversions in legume chloroplast DNAs.** *Curr Genet* 1988, **14:**65-74.
9. Doyle JJ, Davis JI, Soreng RJ, Garvin D, Anderson MJ: **Chloroplast DNA inversions and the origin of the grass family (Poaceae).** *Proc Natl Acad Sci U S A* 1992, **89(16):**7722-7726.
10. Hoot SB, Palmer JD: **Structural rearrangements, including parallel inversions, within the chloroplast genome of Anemone and related genera.** *J Mol Evol* 1994, **38(3):**274-281.
11. Johansson JT: **There large inversions in the chloroplast genomes and one loss of the chloroplast gene rps16 suggest an early evolutionary split in the genus Adonis (Ranunculaceae) .** *Plant Syst Evol* 1999, **218:**318-318.
12. Doyle JJ, Doyle JL, Ballenger JA, Palmer JD: **The distribution and phylogenetic significance of a 50-kb chloroplast DNA inver-**

sion in the flowering plant family Leguminosae. *Mol Phylogenet Evol* 1996, **5(2):**429-438.

13. Lavin M Doyle JJ, Palmer JD: **Evolutionary significance of the loss of the chloroplst-DNA inverted repeat in the leguminosae subfamily Papilionoideae.** *Evolution* 1990, **44:**390-402.
14. Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK: **The Complete Chloroplast Genome Sequence of Pelargonium x hortorum: Organization and Evolution of the Largest and Most Highly Rearranged Chloroplast Genome of Land Plants.** *Mol Biol Evol* 2006, **23(11):**2175-2190.
15. Ogihara Y, Terachi T, Sasakuma T: **Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species.** *Proc Natl Acad Sci U S A* 1988, **85(22):**8573-8577.
16. Lynch M, Koskella B, Schaack S: **Mutation pressure and the evolution of organelle genomic architecture.** *Science* 2006, **311(5768):**1727-1730.
17. Wojciechowski MF, Lavin M, Sanderson MJ: **A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades within the family.** *Am J Botany* 2004, **91:**1846-1862.
18. Hu JM, Lavin M, Wojciechowski MF, Sanderson MJ: **Phylogenetic systematics of the tribe Millettieae (Leguminosae) based on chloroplast trnK/matK sequences and its implications for evolutionary patterns in Papilionoideae.** *Am J Bot* 2000, **87(3):**418-430.
19. Pardo C, Cubas P, Tahiri H: **Molecular phylogeny and systematics of Genista (Leguminosae) and related genera based on nucleotide sequences of nrDNA (ITS region) and cpDNA (trnL- trnF intergenic spacer).** *Plant Syst Evol* 2004, **244:**93-119.
20. Hilu KW, Borsch T, Müller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R, Sauquet H, Neinhuis C, Slotta TAB, Rohwer JG, Campbell CS, W. CL: **Angiosperm phylogeny based on matK sequence information.** *Am J Botany* 2003, **90:**1758-1776.
21. Lavin M, Herendeen PS, Wojciechowski MF: **Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary.** *Syst Biol* 2005, **54(4):**575-594.
22. Kato T, Kaneko T, Sato S, Nakamura Y, Tabata S: **Complete structure of the chloroplast genome of a legume, Lotus japonicus.** *DNA Res* 2000, **7(6):**323-330.
23. Saski C, Lee SB, Daniell H, Wood TC, Tomkins J, Kim HG, Jansen RK: **Complete chloroplast genome sequence of Gycine max and comparative analyses with other legume genomes.** *Plant Mol Biol* 2005, **59(2):**309-322.
24. Greps P Osborn, T. C., Rashka, K., Bliss, F., A.: **Phaseolin-Protein variability in wild forms and landraces of the common bean (Phaseolus vulgaris): evidence for multiple centers of domestication.** *Econ Bot* 1986, **40:**451-468.
25. Mubumbila M, Gordon KH, Crouse EJ, Burkard G, Weil JH: **Construction of the physical map of the chloroplast DNA of Phaseolus vulgaris and localization of ribosomal and transfer RNA genes.** *Gene* 1983, **21(3):**257-266.
26. Chacon SM, Pickersgill B, Debouck DG: **Domestication patterns in common bean (Phaseolus vulgaris L.) and the origin of the Mesoamerican and Andean cultivated races.** *Theor Appl Genet* 2005, **110(3):**432-444.
27. Ramirez M, Graham MA, Blanco-Lopez L, Silvente S, Medrano-Soto A, Blair MW, Hernandez G, Vance CP, Lara M: **Sequencing and analysis of common bean ESTs. Building a foundation for functional genomics.** *Plant Physiol* 2005, **137(4):**1211-1227.
28. Palmer JD: **Chloroplast DNA exist in two orientations.** *Nature* 1983, **301:**92-93.
29. Palmer JD, Thompson WF: **Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost.** *Cell* 1982, **29(2):**537-550.
30. Cronk Q, Ojeda I, Pennington RT: **Legume comparative genomics: progress in phylogenetics and phylogenomics.** *Curr Opin Plant Biol* 2006, **9(2):**99-103.
31. Gantt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD: **Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron.** *Embo J* 1991, **10(10):**3073-3078.
32. Doyle JJ Doyle JL, Palmer JD: **Multiple Independent Losses of 2 Genes and One Intron from Legume Chloroplast Genomes.** *Systematic Botany* 1995, **20(3):**272-294.
33. Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, Calie PJ, Jermiin LS, Wolfe KH: **Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus.** *Plant Cell* 2001, **13(3):**645-658.
34. Sugiura C, Kobayashi Y, Aoki S, Sugita C, Sugita M: **Complete chloroplast DNA sequence of the moss Physcomitrella patens: evidence for the loss and relocation of rpoA from the chloroplast to the nucleus.** *Nucleic Acids Res* 2003, **31(18):**5324-5331.
35. Steane DA: **Complete Nucleotide Sequence of the Chloroplast Genome from the Tasmanian Blue Gum, Eucalyptus globulus (Myrtaceae).** *DNA Res* 2005, **12(3):**215-220.
36. Lee SS, Jeong WJ, Bae JM, Bang JW, Liu JR, Harn CH: **Characterization of the plastid-encoded carboxyltransferase subunit (accD) gene of potato.** *Mol Cells* 2004, **17(3):**422-429.
37. Hupfer H, Swiatek M, Hornung S, Herrmann RG, Maier RM, Chiu WL, Sears B: **Complete nucleotide sequence of the Oenothera elata plastid chromosome, representing plastome I of the five distinguishable euoenothera plastomes.** *Mol Gen Genet* 2000, **263(4):**581-585.
38. Ogihara Y, Isono K, Kojima T, Endo A, Hanaoka M, Shiina T, Terachi T, Utsugi S, Murata M, Mori N, Takumi S, Ikeo K, Gojobori T, Murai R, Murai K, Matsuoka Y, Ohnishi Y, Tajiri H, Tsunewaki K: **Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA.** *Mol Genet Genomics* 2002, **266(5):**740-746.
39. Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB: **The Chlamydomonas reinhardtii plastid chromosome: islands of genes in a sea of repeats.** *Plant Cell* 2002, **14(11):**2659-2679.
40. 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.
41. Kami J, Poncet V, Geffroy V, Gepts P: **Development of four phylogenetically-arrayed BAC libraries and sequence of the APA locus in Phaseolus vulgaris.** *Theor Appl Genet* 2006, **112(6):**987-998.
42. Matsuo M, Ito Y, Yamauchi R, Obokata J: **The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux.** *Plant Cell* 2005, **17(3):**665-675.
43. Leister D: **Origin, evolution and genetic effects of nuclear insertions of organelle DNA.** *Trends Genet* 2005, **21(12):**655-663.
44. Wolfe KH, Li WH, Sharp PM: **Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs.** *Proc Natl Acad Sci U S A* 1987, **84(24):**9054-9058.
45. Muse SV, Gaut BS: **A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome.** *Mol Biol Evol* 1994, **11(5):**715-724.
46. Matsuoka Y, Yamazaki Y, Ogihara Y, Tsunewaki K: **Whole chloroplast genome comparison of rice, maize, and wheat: implications for chloroplast gene diversification and phylogeny of cereals.** *Mol Biol Evol* 2002, **19(12):**2084-2091.
47. Delgado-Salinas A Bibler R, Lavin M: **Phylogeny of the genus Phaseolus (Leguminosae): A recent diversification in an ancient landscape.** *Systematic Botany* 2006, **31(4):**779-791.
48. Ayala FJ: **Molecular clock mirages.** *Bioessays* 1999, **21(1):**71-75.
49. Yang Z, Nielsen R: **Synonymous and nonsynonymous rate variation in nuclear genes of mammals.** *J Mol Evol* 1998, **46(4):**409-418.
50. Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, Cheng CH, Lin CY, Liu SM, Chang CC, Chaw SM: **The chloroplast genome of Phalaenopsis aphrodite (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications.** *Mol Biol Evol* 2006, **23(2):**279-291.
51. Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB: **Impacts of genetic bottlenecks on soybean genome diversity.** *Proc Natl Acad Sci U S A* 2006, **103(45):**16666-16671.
52. Jansen RK, Raubeson LA, Boore JL, dePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ, Fourcade HM, Kuehl JV, McNeal JR, Leebens-Mack J, Cui L: **Methods for**

obtaining and analyzing whole chloroplast genome sequences. *Methods Enzymol* 2005, **395:**348-384.

53. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8(3):**175-185.
54. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing.** *Genome Res* 1998, **8(3):**195-202.
55. Wyman SK, Jansen RK, Boore JL: **Automatic annotation of organellar genomes with DOGMA.** *Bioinformatics* 2004, **20(17):**3252-3255.
56. 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.
57. Darling AC, Mau B, Blattner FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements.** *Genome Res* 2004, **14(7):**1394-1403.
58. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R: **REPuter: the manifold applications of repeat analysis on a genomic scale.** *Nucleic Acids Res* 2001, **29(22):**4633-4642.
59. Emboss: **Programs [http://bioweb.pasteur.fr/seqanal/EMBOSS].** .
60. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32(5):**1792-1797.
61. Posada D, Crandall KA: **MODELTEST: testing the model of DNA substitution.** *Bioinformatics* 1998, **14(9):**817-818.
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. PHYLIP: **[http://evolution.genetics.washington.edu/phylip.html].** .
64. Kumar S, Tamura K, Nei M: **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.** *Brief Bioinform* 2004, **5(2):**150-163.
65. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13(5):**555-556.
66. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25(24):**4876-4882.