Escherichia coli Plasmidomics

Val Fernández Lanza

Tesis Doctoral

Director:
Dr. Fernando de la Cruz Calahorra
Departamento de Biología Molecular
2015

Agradecimientos

Desde el primer día que ves una tesis te asalta un pensamiento "¿Qué pondría yo en mis agradecimientos?". Con los años te das cuenta que tus agradecimientos van creciendo, conoces nueva gente que te ayuda de una u otra forma y a la cual terminas debiéndole, al menos, un "gracias".

Gracias a Fernando, por esta magnífica oportunidad que me dio, por ser el "jefe" que necesitaba, por dejarme seguir mis propios caminos y por enseñarme algunos que yo era incapaz de ver. La puerta de ese despacho siempre ha estado abierta para debatir mis ideas a pesar de que al final nunca me leí el Lehninger.

Gracias a mis compañeros de "labo", mis compañeros de la "lobera"; Yera, Inma, Juancho, Alej, Getino, Maris, David, Eric, Sandra, Andrés y muy especialmente a Jorge y a Lillo que nunca me han fallado y que siempre están ahí. A los post-Docs; a Mapi porque siempre esta pendiente de nosotros; enseñando, ayudando y cuidando. A "sATHAnasia" por ser la mejor y más divertida griega del mundo y a Ruli, por esa primera entrevista de trabajo que me abrió las puertas del laboratorio. A l@s tecnic@s, Ana, Raquel, Sheila, Carlos, Mati y Sandra que sin ellos el laboratorio no funcionaría. A María, mi pupila, co-autora y revisora. Gran parte de esta tesis se debe a su trabajo incansable.

Gracias al resto de los "jefes", a Gabi, Elena e Iñaki con quien he tenido el placer de trabajar, aprender y discutir (esto último sobretodo con Iñaki).

Gracias a los bioquímicos; a Manu, Gabi, Juan y Bolado por los partidillos y el "frikismo". A los "piero": Ana, Javi, Adán, Lorena e Iñaki porque en ese labo me siento como en casa. A "risoto"; por las tardes en su casa, el padel, la piscina y

sus sonrisa constante. A las chicas de Matxalen; Coral y Anabel por los buenos ratos. A la gente de inmuno; Fernanda, Aramburu y Jorge; especialmente a Marcos, el más grande de Gordaliza del Pino y parte del mundo. Y por supuesto a Maigui, sin ella esto no hubiese ocurrido.

Gracias a mis dos chicas especiales: Esther, mi pequeña ninja, con ese carácter que diga lo que te diga tu sabes que siempre te está diciendo lo que piensa. A Paula, sabes que siempre serás una de mis personas favoritas y que siempre te estaré agradecido.

Gracias a Nuria y el moro. Gracias y mil gracias Nuria por convencerme de hacer el master de bioinformática. Gracias a los dos por demostrar que los kilómetros nunca son un obstáculo para la amistad.

Gracias a los externos, a Clara y Beri que de vez en cuando me aguantan la chapa científica y ni siquiera protestan. A Roberto e Irene que afortunadamente el destino ha hecho que nos volvamos a reunir en Madrid.

Y hablando de Madrid, a todos los que me han dado la bienvenida en mi nueva casa; a la gente del Ramón y Cajal: Anasofi, Conchi, Merche, Aida, Marta e Irene que además me ha ayudado con sus incisivas críticas a mi redacción. A Fernando y Teresa; que me han acogido como nunca podría haber imaginado y que me enseñan la parte de la ciencia que esta fuera del teclado.

A “mañocao”, Ra y Fer, Sara y Ramón, Santi y Ros: que os habéis convertido en mi pequeña familia en Madrid.

Y ya termino con lo más importante, mi familia: A mis padres que siempre me han apoyado, que siempre han creído en mi y que me han inculcado la necesidad de saber. A mi hermano, sé que siempre ha estado orgulloso de mí. A Ana por traer la alegría a casa. Gracias papá por tus grandes lecciones, el

pensamiento crítico y tu optimismo infinito: “si ellos pueden tu puedes”. Soy lo que soy gracias a vosotros.

Gracias a Mila y Santiago por abrirme las puertas de su casa.

Gracias a ti, mi amiga, mi compañera y mi mujer. Porque tu siempre estás a mi lado, porque siempre me empujas para que siga. Porque me haces ser mejor y querer ser mejor. Por hacerme feliz. A ti Raquel por aparecer en mi vida.

Prólogo **8**
Objetivos *9*

Introducción **10**
Escherichia coli: "Ese gran conocido". *11*
Evolución y genómica de *Escherichia coli.* 12
Patogenicidad de *Escherichia coli*. 19
E. coli O25b:H4-ST131: Pandémico, multirresistente y comensal. 28
Plásmidos. *31*
Clasificación. 32
Estructura y funcionalidad. 39
Plásmidos de *E. coli*. 43
Secuenciación de nueva generación y Bioinformática. *46*
PLAsmid Constellation NETworks (PLACNET) 61

Publicaciones **64**
Blanco J, et al. 2013. *Four Main Virotypes among Extended-Spectrum-β-Lactamase-Producing Isolates of Escherichia coli O25b:H4-B2-ST131: Bacterial, Epidemiological, and Clinical Characteristics.* 66
Lanza VF, et al. 2014. *Plasmid Flux in Escherichia coli ST131 Sublineages, Analyzed by Plasmid Constellation Network (PLACNET), a New Method for Plasmid Reconstruction from Whole Genome Sequences.* 76
De Been M, et al. 2014. *Dissemination of Cephalosporin Resistance Genes between Escherichia coli Strains from Farm Animals and Humans by Specific Plasmid Lineages.* 128

Discusión **150**
Escherichia coli plasmidomics 151
PLACNET 152
Diseminación de plásmidos resistentes a antibióticos. 155
E. coli O25b:H4-B2 ST131 y su plasmidoma. 158
Conclusiones *161*

Bibliografía. **164**

Otras publicaciones **186**

Prólogo

En esta tesis hemos acuñado el nuevo término "Plasmidomics" como un estudio integral de la biología de los plásmidos: la genética, la evolución, la interacción con el hospedador o el impacto en la población bacteriana. Queremos establecer el estudio de los plásmidos desde un punto de vista tanto global, analizando el impacto que tienen los plásmidos en las poblaciones bacterianas, como un punto de vista más individual: cómo evoluciona un plásmido y cómo es la co-evolución de un plásmido y su hospedador. Para ello nos hemos centrado en *Escherichia coli*, un comensal de la flora animal y humana, y patógeno oportunista que ha sido estudiado intensamente desde los inicios de la microbiología. En esta tesis presentamos tres trabajos que pretenden analizar qué implicación tienen los plásmidos en la biología de *Escherichia coli*. Los dos primeros artículos (Blanco et al. 2013; Lanza et al. 2014) tratan del conjunto de plásmidos (plasmidoma) presentes en el clon epidémico *Escherichia coli* ST131. Para ello hemos desarrollado un nuevo método (PLAsmid Constellation NETworks, PLACNET) que permite identificar y caracterizar los plásmidos a partir de secuenciaciones de nueva generación. Este nuevo método nos da una profundidad en el análisis que hasta ahora no existía. Este estudio pone de manifiesto cómo la variabilidad de los plásmidos es muy superior a la de su hospedador, indicando que su evolución es más rápida que la de los cromosomas y reforzando la idea de que los plásmidos son los principales vectores de transmisión de información entre bacterias (Halary et al. 2010). Incluso en un conjunto de cepas que representan un clon epidémico encontramos una heterogeneidad plasmídica superior a la esperada. Esta heterogeneidad se ha podido determinar mediante PLACNET debido a que los métodos clásicos de tipado no tienen la suficiente resolución para observarla.

En el tercer trabajo (de Been et al. 2014), hemos estudiado el plasmidoma de un conjunto de cepas de *Escherichia coli* productores de beta-lactamasas de espectro extendido (BLEE). Para ello, se analizó el contenido

plasmídico de 32 cepas aisladas de pollos, carne de pollo, cerdos y humanos en el grupo del Dr. Rob Willems (Centre Medical de Utretch). El estudio está orientado a investigar cómo se produce la diseminación de la resistencia a antibióticos a través de la cadena alimentaria. El objetivo fue analizar si existe una dispersión que implique a las estabulaciones animales, la manipulación del producto y a los consumidores. Además, el trabajo incorporó una serie de muestras de cerdos y los granjeros encargados de su estabulación para estudiar la posible transmisión directa. La comparación de filogenia de los plásmidos y los hospedadores nos ha permitido establecer la transmisión de los determinantes de resistencia, diferenciando entre transmisiones clonales y transmisiones plasmídicas. Nuevamente hemos necesitado aplicar PLACNET para reconstruir los plásmidos y realizar los estudios filogenéticos. Este trabajo ha demostrado que la transmisión de los determinantes de resistencia se debe a una expansión plasmídica y no a una expansión clonal, como parecían indicar los métodos clásicos de tipado.

Objetivos

1. Desarrollar un software que nos permita la identificación y definición de plásmidos a partir de los datos de Secuenciación de Nueva Generación.
2. Validación del software frente a cepas contenidas en bases de datos públicas, secuenciación alternativa y experimentos de biología molecular.
3. Análisis del plasmidoma en el clon epidémico *Escherichia coli* B2 ST131.
4. Influencia del plasmidoma de *Escherichia coli* B2 ST131 en su evolución, expansión y virulencia.
5. Análisis del plasmidoma de cepas de *Escherichia coli* productoras de beta-lactamasas de espectro extendido con distintos orígenes de la cadena alimentaria.
6. Transferencia de los genes $bla_{CTX\text{-}M\text{-}1}$ y $bla_{CMY\text{-}2}$ en aislados de *Escherichia coli* pertenecientes a aislados de la cadena alimentaria.

Introducción

Escherichia coli: "Ese gran conocido".

Escherichia coli (E. coli) es un comensal habitual de la flora intestinal animal y por lo tanto de la humana (Tenaillon et al. 2010; Chaudhuri and Henderson 2012). Sin género de dudas, *E. coli* es el organismo procariota más estudiado del mundo. Una simple búsqueda bibliográfica de la presencia del término *"Escherichia coli"* o *"E. coli"* en el título de los artículos devuelve una cantidad de 104.387 publicaciones (Octubre de 2014), lo cual es un claro indicativo del interés que suscita entre los microbiólogos y entre la biología en general.

E. coli fue descrita por primera vez en 1885 por el físico alemán Theodor Escherich. Aunque inicialmente fue bautizada como *Bacterium coli commune,* en 1919 Castellani y Chalmers la renombraron como *Escherichia coli* en referencia a su descubridor. Paralelamente, 1897 Kiyoshi Shiga describió el *Bacterium dysentericus* que nuevamente fue renombrado por Castellani y Chalmers como *Shigella* en el 1919. Los géneros *Escherichia* y *Shigella* están fuertemente relacionadas. Muchos autores sugieren que en realidad son el mismo género, ya que las pruebas filogenéticas demuestran que algunos grupos de *E. coli* están tan distantes entre sí como lo están de *Shigella* (Chaudhuri and Henderson 2012).

Atendiendo a su descripción clásica, *E. coli* es un "*Bacilo Gramnegativo, anaerobio facultativo, móvil por flagelos perítricos, no forma esporas, es capaz de fermentar la glucosa y la lactosa y su prueba de IMVIC (Indol, Rojo de metilo, Voges-Proskauer, Citrato) es* ++--". En la actualidad esta descripción está obsoleta ya que diversas publicaciones muestran que existen cepas de *E. coli* que son citrato positivas (Ishiguro 1978). Taxonómicamente se clasifica como *Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriacea; Escherichia*. Además, *E. coli* se puede clasificar en distintos grupos monofiléticos

(filogrupos o subgrupos). Aunque el número de estos grupos puede variar dependiendo de los estudios, se acepta que los principales son A, B1, B2, D y E (Chaudhuri and Henderson 2012). Los nuevos grupos C, F e I son de reciente definición y aún no se han caracterizado en profundidad. Actualmente, somos capaces de establecer nuevos esquemas de clasificación que permiten indagar en la naturaleza y origen de estas cepas (Clermont et al. 2013). Existen evidencias que indican que los filogrupos presentan preferencias por distintos nichos ecológicos. En humanos los filogrupos A (40.5%) y B2 (25.5%) son los más frecuentes y B1 y D (17%) son menos prevalentes, mientras que en animales domésticos y salvajes las cepas del grupo B1 son predominantes (41%), seguidas por A (22%), B2 (21%) and D (16%) (Tenaillon et al. 2010).

Evolución y genómica de *Escherichia coli*.

Existen muchas cuestiones que rodean a la naturaleza de *E. coli*. ¿Cómo un comensal se convierte en un patógeno?, ¿qué presión o fuerza evolutiva provoca este cambio?, ¿cuáles son los mecanismos más importantes en su evolución?, ¿cuál es su estructura poblacional?…

En la era pre-molecular, para clasificar las cepas de *E. coli* se usaban serotipos basados en el antígeno somático (O), el antígeno flagelar (H) y en menor medida el antígeno capsular (K) (Chaudhuri and Henderson 2012). Aún, hoy en día, es habitual usar esta metodología para clasificar las cepas. En la actualidad los principales sistemas de clasificación se basan en métodos moleculares. Por una parte tenemos el esquema de Clermont que, basándose en la presencia/ausencia de una serie de genes, es capaz de diferenciar entre los principales filogrupos (Clermont et al. 2000). Una actualización de este método añade la detección de otros filogrupos (Clermont et al. 2013). La presencia o ausencia de los genes se determina mediante PCR y por lo tanto supone un método sencillo, rápido y económico en comparación con otros. Sin embargo, en ocasiones la determinación del filogrupo no es lo

suficientemente informativo ya que con él es imposible rastrear brotes epidémicos o expansiones clonales.

En 1998 se desarrolló una nueva técnica basándose en la secuencia de nucleótidos de una serie de genes y se llamó tipificación multilocus de secuencias o MLST por sus siglas en inglés (*Multilocus Sequence Typing*). El MLST se desarrolló por primera vez para tipificar *Neisseria meningitidis* basándose en las secuencias de 11 genes (Maiden et al. 1998). El MLST consiste en numerar cada alelo de cada gen y asignar un ST (*Sequence Type*) a cada combinación de alelos. Existen multitud de esquemas de MLST para una gran diversidad de especies. Los repositorios más importantes son posiblemente PubMLST (http://pubmlst.org/) y MLST.net (http://www.mlst.net/), donde podemos encontrar información de todos los esquemas MLST existentes: genes que componen cada esquema, cebadores de amplificación para cada gen, alelos existentes, etc. En el caso de *E. coli* existen tres esquemas diferentes para MLST. El más utilizado es el propuesto por el grupo de Mark Achtman (Wirth et al. 2006) (http://mlst.warwick.ac.uk/mlst/), seguido de los esquemas del Instituto Pasteur (http://www.pasteur.fr/recherche/genopole/PF8/mlst/EColi.html) y Thomas Whittam (http://www.shigatox.net/). A fecha actual (Enero 2015) no existe ningún estudio que relacione los tres esquemas. Los estudios con MLST permiten obtener información poblacional y trazar posibles expansiones clonales. Aunque durante mucho tiempo se admitió que dos cepas que compartían el mismo ST eran clones, actualmente esa afirmación no es aceptada. El esquema de Achtman se compone de 7 genes (*adk, fumC, gyrB, icd, mdh, purA y recA*). Para que un gen esté presente en un esquema de MLST debe cumplir varias condiciones: 1) que sea un gen "*housekeeping*", es decir que esté presente en todas las cepas; 2) que varíe lo suficiente como para dar una variabilidad genética; y 3) que no se encuentre en zonas recombinantes. Aunque en un principio el MLST se desarrolló simplemente para el tipado de las cepas, el uso de la secuencia de estos genes para extrapolar la filogenia de las muestras es un recurso ampliamente utilizado.

Lo común es generar una única secuencia de DNA uniendo la secuencia de los 7 genes, crear un alineamiento múltiple con otras secuencias de cepas conocidas y generar un árbol filogenético. En el caso de *Escherichia coli* y el esquema de Achtman existe una gran correlación entre la filogenia producida por los genes MLST con las filogenias basadas en la secuencia completa de los genomas (Tenaillon et al. 2010; Chaudhuri and Henderson 2012). Sin embargo existen discrepancias entre los arboles filogenéticos producidos por los distintos esquemas de MLST debido a que la tasa de recombinación de *E. coli* es mayor de la observada inicialmente (Chaudhuri and Henderson 2012). En la actualidad, con la proliferación de las tecnologías de secuenciación masiva, se puede realizar una clasificación más exacta de las cepas usando los genomas completos de las bacterias (ver el capítulo Secuenciación de Nueva Generación y Bioinformática).

Los estudios genéticos van poniendo un orden en la evolución de *E. coli*. Como hemos mencionado antes, existen diversos grupos filogenéticos (A, B1, B2, D, E y F (Tenaillon et al. 2010)), a estos grupos incluyen tanto a comensales, patógenos o ambientales (Luo et al. 2011). A estos grupos establecidos se añaden *E. coli* crípticos que se posicionan entre los *E. coli* conocidos y otras especies como *E. fergusonii* o *E. albertii,* lo que hace aún más difusa la línea que separa a la especie *E coli* del resto de especies del género (Walk et al. 2009). Si esta diversidad es característica de *Escherichia* o en realidad es la diversidad intrínseca en cualquier género bacteriano es

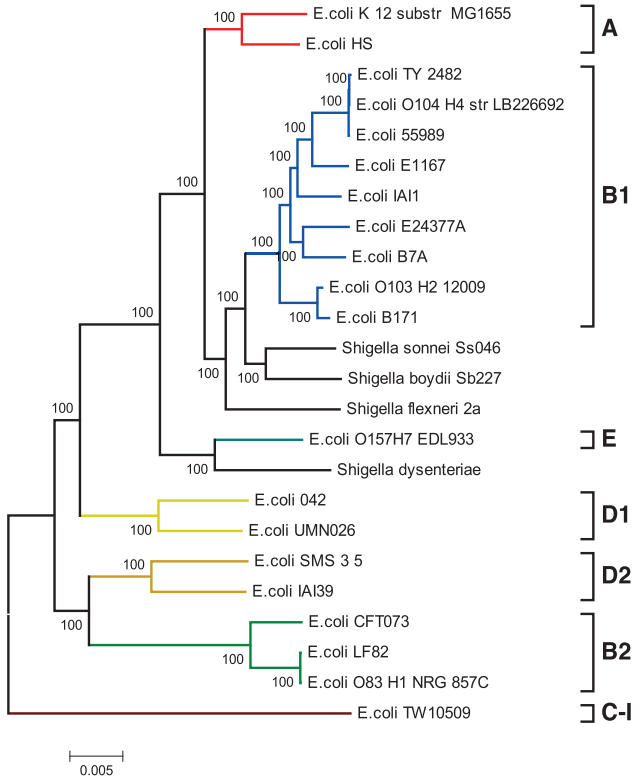


Figura 1. Filogenia de *E. coli* realizada a partir de los genomas completos de algunas de las cepas más relevantes que definen los filogrupos de *E. coli*. Figura extraída de (Chaudhuri and Henderson 2012)

difícil de determinar.

E. coli también se clasifica en función de su potencial patógeno en cepas comensales y cepas patógenas o virulentas. El conjunto de genes que determina la diferencia entre una cepa comensal y una cepa patógena se definen como factores de virulencia e incluye una gran variedad de proteínas, (ver Patogenicidad de *Escherichia coli*)

No parece existir una correlación entre filogrupos y patogenicidad (Escobar-Páramo et al. 2004). Podemos encontrar cepas patógenas en cualquiera de los filogrupos. Por ejemplo, *E. coli* O25b:H4-ST131 pertenece al grupo B2, O157:H7 EDL933 pertenece al grupo E y O104:H4 2009EL-2050 al grupo B1 (Escobar-Páramo et al. 2004; Bohlin et al. 2014). Tampoco parece existir una exclusividad de factores de virulencia asociados a algún filogrupo. Todo esto indica que la adquisición horizontal de factores de virulencia es lo que determina si una cepa es virulenta o comensal (Leimbach et al. 2013).

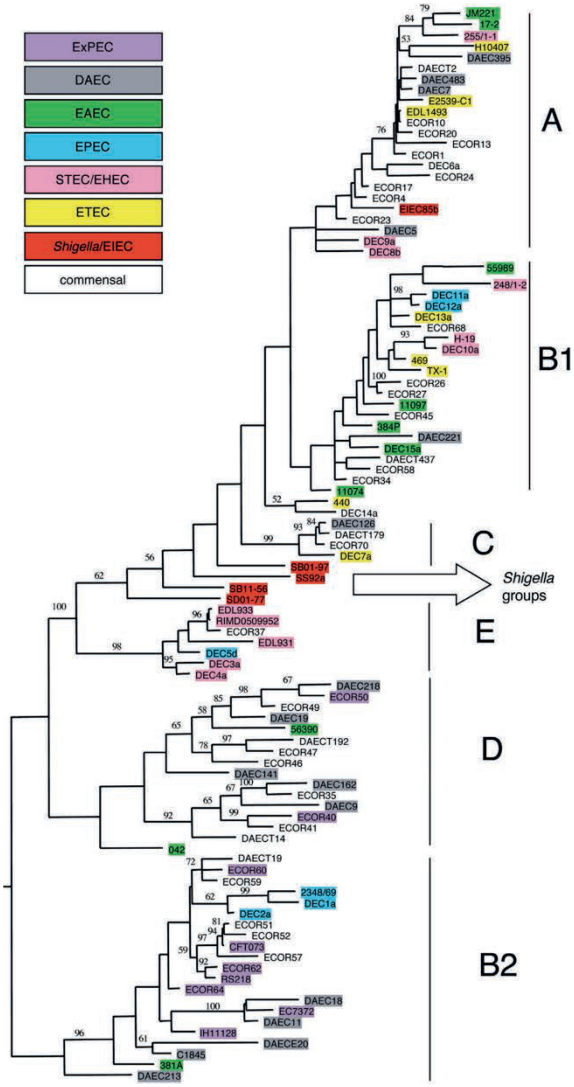


Figura 2. Filogenia de *E. coli* basado en la concatenación y alineamiento de los genes: *trpA*, *trpB*, *pabB*, *putP*, *icd* y *polB*. En colores quedan resaltados los distintos patotipos así como las cepas de *Shigella*. Figura extraída de (Escobar-Páramo et al. 2004)

Este no es un caso exclusivo de *E. coli*, existen multitud de ejemplos de bacterias comensales que pueden llegar a ser patógenos mediante la adquisición de determinados factores de virulencia (Pallen and Wren 2007).

Genéricamente podemos dividir un genoma en dos partes: el núcleo del genoma, o *"core-genome"*, y el pangenoma. El *core-genome* lo componen todos aquellos genes que son comunes a la especie, es decir que

están presentes en todas las cepas. El pangenoma por su parte, está representado por todos aquellos genes que están presentes en alguna o muchas cepas de una especie pero no pertenecen al *core-genome*. Idealmente el *core-genome*, son todos aquellos genes que hacen las funciones básicas e imprescindibles de la bacteria. En la realidad no es exactamente así, ya que no se puede construir una bacteria únicamente con su *core-genome*. Por ejemplo, el *core-genome* de *E. coli* se estima en 1.729 genes (Bohlin et al. 2014), mientras que el genoma mínimo más pequeño que se ha conseguido construir es de ~3.700 genes (Pósfai et al. 2006). Teóricamente se pueden obtener genomas más reducidos, suficientes para mantener un funcionamiento normal en un entorno específico (Gil et al. 2004; Zhang et al. 2010). De estos datos se deduce que en el pangenoma están incluidos genes con funciones redundantes, pero esenciales para la bacteria. Es decir, que un determinado número de funciones son imprescindibles para la bacteria, pero pueden ser realizadas por distintos conjuntos de genes.

Existen tres fuerzas que moldean un genoma: la ganancia, la pérdida y el intercambio de genes. La ganancia de genes en bacterias puede producirse por 4 vías: duplicación génica, conjugación, transformación y transducción. La duplicación génica es un proceso básico en la evolución de los organismos. Por un error en la replicación de ADN o una recombinación asimétrica, se producen dos copias del mismo gen. Éste puede evolucionar y convertirse en un gen parálogo. Existen ejemplos donde la virulencia de una cepa está influenciada por la duplicación de un gen, como es el caso de *M. Tuberculosis*, donde las duplicaciones de los genes involucrados en la secreción de la proteína ESAT6 han incrementado su virulencia en comparación con otras cepas (Pallen and Wren 2007).

Los otros tres procesos de adquisición de genes se consideran transferencia horizontal. Ésta es la forma en la que las bacterias comparten información, más allá de la herencia genética o evolución vertical. Generalmente son estos procesos los que modifican el fenotipo de la bacteria

más rápidamente y permiten la colonización de nuevos nichos (Wirth et al. 2006). La conjugación es un mecanismo por el cual una bacteria llamada donadora transmite un plásmido a una bacteria receptora o recipiente. Este proceso, a diferencia de los fenómenos de transformación y transducción, requiere un contacto físico entre las dos células. El proceso de conjugación y los plásmidos serán explicados en detalle en una sección posterior (ver Plásmidos).

En los eventos de transformación y transducción esta también implicado el proceso de recombinación. Tanto la recombinación homóloga, como la no homóloga juegan un papel fundamental en la evolución de las bacterias (Didelot, Méric, et al. 2012). Se define la tasa de recombinación/mutación de una especie como el ratio entre los eventos de recombinación homóloga y los eventos de mutación (mutaciones puntuales o pequeñas inserciones/deleciones) (Vos and Didelot 2009). Esta medida se suele realizar sobre los genes MLST ya que es el conjunto de datos más amplio que normalmente está disponible. *E. coli* tiene una tasa de recombinación/mutación (r/m) de 0.7, que en comparación con otras bacterias estudiadas es muy baja, por ejemplo, *Salmonella enterica* tiene un r/m de 30.2 ó *Haemophilus influenzae* un 3.7 (Vos and Didelot 2009). Esto significa que existen más eventos de mutación que de recombinación, lo que es congruente con la incapacidad de *E. coli* para adquirir ADN externo espontáneamente vía transformación (fenómenos de competencia natural) . Esto no significa que no existan procesos de transformación, sino que ocurren a frecuencias bajas en comparación con los procesos de conjugación o de transducción (Sinha and Redfield 2012). Respecto a estos datos existe mucha controversia. Se debe diferenciar entre la recombinación del *core-genome* y la adquisición, vía recombinación, de nuevos elementos en el cromosoma. No sólo la transformación está implicada en la recombinación, sino que está descrito que los plásmidos pueden integrarse en el cromosoma y posteriormente transferir grandes segmentos de ADN mediante conjugación (Ochman et al. 2000; Schubert et al. 2009). Además algunos bacteriófagos

pueden integrar hasta 100kb de ADN extra que posteriormente puede transferirse por transducción. Como ya hemos mencionado, las tasas de recombinación-mutación (*r/m*) se suelen medir estudiando los genes de MLST, que por definición pertenece al *core-genome*. De esta manera en *E. coli* se obtienen tasas bajas de *r/m* lo que implica que las recombinaciones homólogas en el *core-genome* son poco frecuentes. Sin embargo, muchos otros estudios señalan que los eventos de recombinación son mucho más frecuentes que esta estimación, pero que principalmente ocurren en genes accesorios (Mau et al. 2006; Schubert et al. 2009; Didelot, Méric, et al. 2012; McNally et al. 2013).

Atendiendo a los elementos que se pueden adquirir mediante transferencia horizontal tenemos: plásmidos, transposones, bacteriófagos e integrones (De la Cruz and Davies 2000). Esta clasificación no es absoluta. Los integrones, por ejemplo, suelen estar contenidos en transposones; la mayoría de los plásmidos tienen transposones; un bacteriófago puede contener un integrón y/o un transposón, etc… Las combinaciones son numerosas y dan lugar a un enorme abanico de posibilidades. En el caso de los cromosomas, se clasifica como isla genómicas, o GI por sus siglas en inglés (Genomic Islands), a un conjunto de genes agrupados en una región del cromosoma que han sido adquiridos horizontalmente y, generalmente, estan constituidos por un combinación de bacteriófagos, transposones o integrones. Si una GI contiene factores de virulencia, se suele clasificar como isla de patogenicidad o PAI por sus siglas en inglés (Pathogenicity Island). Las PAIs suelen cumplir los siguientes criterios: contener genes de virulencia y estar presentes en cepas patógenas, pero no ser comunes en cepas no-patógenas de la misma especie o cercanas; tienen tamaños entre 10kb y 200kb, son relativamente inestables, suelen tener un contenido en GC distinto al del *core-genome*, suelen estar asociadas con genes tARN, frecuentemente contienen genes de movilización como secuencias de inserción (IS), transposones, integrasas y bacteriófagos, suelen estar flanqueadas por secuencias repetidas y generalmente presentan un

mosaicismo compuesto por pequeños segmentos de ADN posiblemente adquiridos durante diferentes eventos de transferencia horizontal (Lloyd et al. 2009; Leimbach et al. 2013). Inicialmente el término PAI fue acuñado para cepas *E. coli* uropatógenas , pero ahora el uso es generalizado.

A pesar de la ganancia de genes por transferencia horizontal, los genomas tienden a mantener el mismo tamaño, por lo que la ganancia de genes debe complementarse con la pérdida de otros (Mira et al. 2001; Pallen and Wren 2007). Las bacterias parecen guiarse por la máxima de "úsalo o piérdelo". Así, si una bacteria permanece mucho tiempo en un mismo nicho y este es estable, perderá todos aquellos genes que no suponen una ventaja evolutiva en dicho nicho. Esta pérdida de genes es una ventaja adaptativa *per se,* ya que la simple presencia de los genes supone un pequeño coste biológico (Pallen and Wren 2007). Los cambios genéticos son el evento más básico de la evolución: mutación, presión, selección. Una mutación puntual puede provocar un beneficio biológico que seleccione a una población por encima del resto. Como se comenta más adelante en este capítulo, una mutación puntual puede provocar la resistencia a un conjunto de antibióticos como es el caso de el gen *gyrA* y los antibióticos basados en las quinolonas (Vila et al. 1994). Esto puede favorecer una expansión clonal bajo determinadas circunstancias de presión selectiva con estos antibióticos.

Patogenicidad de *Escherichia coli*.

En un principio se pensó que *E. coli* era un organismos estrictamente comensal y que por lo tanto no generaba infecciones lo que le diferenciaba de *Shigella*, estrictamente patógena. No fue hasta la década de los 40 en que se empezaron a describir los primeros brotes de diarreas ocasionados por *E. coli*. Ahora sabemos que *E. coli* es capaz de producir distintos tipos de infecciones, principalmente: diarreas entéricas, infecciones del tracto urinario (ITUs) y sepsis/meningitis. En base a la patología que producen, se ha confeccionado una clasificación de ocho patotipos distintos. Inicialmente se dividen en dos grupos: extra-intestinales (ExPEC) y diarreicas. Dentro de estos

dos grupos tenemos: entero-patogénicos (EPEC), entero-hemorragicos (EHEC), entero-toxigénicos (ETEC), entero-invasivos (EIEC), entero-agregativos (EAEC), difusamente adherentes (DAEC), uropatogénicos (UPEC) y causantes de meningitis neonatal (NMEC) (Walk et al. 2009). Además de estos patotipos existe un grupo más, referente a animales (APEC) que causa infecciones extra-intestinales (infecciones respiratorias, pericarditis y septicemias) en aves.

ETEC y las DAEC colonizan el intestino delgado y causan diarrea. Las EHEC y las EIEC causan enfermedades en el intestino grueso. Las EAEC pueden colonizar tanto en intestino delgado como el grueso. Las UPEC anidan en el tracto urinario, pudiendo ascender hasta la vejiga y causar cistitis, que en último lugar genera pielonefritis. Tantos las UPEC como las NMEC son capaces de pasar a torrente sanguíneo causando bacteremias. Además, NMEC pueden traspasar la barrera hematoencefálica y producir meningitis (ver **Figura 1**).

El nicho original de *E. coli* es la mucosa del colon. A pesar de los esfuerzos y recursos invertidos en el estudio de *E. coli* aún se desconoce la función simbiótica que pueda realizar esta bacteria en el intestino humano. Los estudios de metagenómica de la flora intestinal revelan que las proteobacterias representan tan sólo entre el 1% y el 0.1% de la microbiota intestinal total (Eckburg et al. 2005; Tap et al. 2009).

A pesar de este dato *E. coli* es el patógeno con mayor incidencia en bacteriemias (Gradel et al. 2014; Laupland and Church 2014), el patógeno con mayor frecuencia en infecciones del tracto urinario (ITU) (Foxman 2003; Foxman 2010) y diarreas (Croxen et al. 2013), y se estima que causa la muerte a cerca de 2 millones de personas al año en todo el mundo (Tenaillon et al. 2010). Esto indica que a pesar de no abundar en el intestino, su eficacia para ocasionar determinadas infecciones, como las ITUs, es mayor que en especies más abundantes en el intestino (Eckburg et al. 2005; Foxman 2010).

Esto marca el carácter versátil de esta especie capaz de ocupar diversos nichos ecológicos, pero ninguno de forma dominante (tracto

digestivo de humanos y otros animales, comida o medio ambiente (Kaper et al. 2004; Wirth et al. 2006; Tenaillon et al. 2010; Chaudhuri and Henderson 2012)) .

En general la mayoría de las infecciones de *E. coli* son fácilmente tratables y no suponen una elevada tasas de mortalidad. Sin embargo su enorme incidencia en la población mundial la convierte en una de las principales causas de mortalidad en países sub-desarrollados (Croxen et al. 2013). En estos países son las diarreas el principal problema que produce *E. coli*. La inaccesibilidad a tratamientos médicos básicos y a condiciones de higiene y salubridad, agravan el pronóstico de la enfermedad. Por el contrario, en los países desarrollados las diarreas son fácilmente tratables y

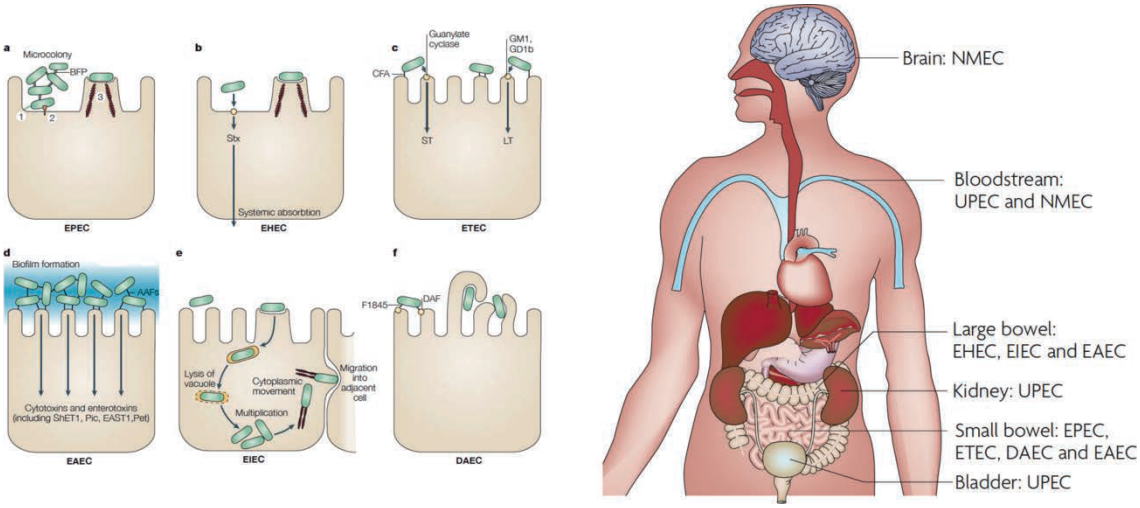


Figura 3. a| Las EPEC se adhieren a los eritrocitos del intestino delgado pero destruyen las microvellosidades induciendo las características lesiones de adherencia y esfacelamiento (A/E). La desestructuración del citoesqueleto vienen acompañados de una respuesta inflamatoria y diarrea. b| EHEC También inducen lesiones por (A/E) pero en este caso en el colon. La característica distintiva de las EHEC es la producción de la toxina Stx (Shigella toxin). La absorción de esta toxina durante largos periodos puede incluso producir la muerte. c| El caso de las ETEC es parecido, se adhieren también al intestino delgado e inducen diarrea mediante la secreción de las enterotoxinas termo-lábil (LT) y/o termo-estable (ST). d| las EAEC se pueden adherir tanto a las células epiteliales del intestino delgado como del intestino grueso formando un fino biofilm y secretan enterotoxinas y citotoxinas. e| Las EIEC invaden las células epiteliales del colon y producen la lisis del fagosoma. Una vez lisado el fagosoma son libres de moverse dentro de la célula y entre las células adyacentes. f| Finalmente las DAEC producen una señal en los enterocitos del intestino delgado provocando la proyección de una estructura celular que rodea la bacteria. (AAF) Fimbria agregativa adherente. (BFP) Pilus de racimo. (CFA) Antígeno del factor de colonización. (DAF) Factor de la aceleración del deterioro. (LT) Toxina termo-lábil. (ST) Toxina termo-estable. (ShET1) Enterotoxina Shigella 1. (EAST1) E. coli entero-agregativo ST1 Modificada de (Croxen and Finlay 2010) y (Kaper et al. 2004).

son limitados los casos de muerte por este tipo de infecciones. Sin embargo la incidencia de infecciones urinarias es muy alta, sobre todo en la población femenina. Algunos estudios indican que al menos la mitad de las mujeres sufrirán una infección urinaria una vez en su vida, y que una de cada tres necesitará tratamiento con antibióticos antes de los 24 años (Foxman 2003).

Estudios *in vitro* han revelado algunos de los mecanismos que usan las bacterias en los procesos de colonización y/o infección. Estos resultados podrían no ser extrapolables al intestino humano, por lo que aún existe una gran laguna en el conocimiento de los procesos infecciosos de *E.coli*.

El proceso de infección de *E. coli* es similar al que emplean otros patógenos de la mucosa intestinal, pudiendo resumirse en cuatro pasos básicos: colonizar de la mucosa, sortear las defensas del hospedador, proliferar y dañar al hospedador. Independientemente del patotipo de *E. coli,* todos ellos tienen funciones esenciales para poder colonizar y/o infectar tejidos fuera de su nicho natural. La función más básica en este sentido es la adherencia a las células del intestino, uretra u otros tejidos. Las proteínas que realizan esta función se denominan "adhesinas". Dentro de la adhesinas existen diversas familias siendo las más abundantes las pilinas (ver **Tabla 1**). Existe una cierta asociación entre algunos patotipos y adhesinas, aunque no siempre están presentes en todas las cepas del patotipo, por ejemplo, *Afa/Dr* esta asociado con UPECs pero no todas las UPECs tienen este operón (Kaper et al. 2004).

Muchas de estas adhesinas son las causantes de la respuesta del sistema inmune del hospedador que en casos excepcionales pueden llegar a causar shock séptico e incluso la muerte. Al ser un elemento común en todos los *E. coli* las fimbrias se usan para la clasificación de distintos grupos. En el caso del lipopolisacárido (LPS) (antígeno O) define los serogrupos, mientras que si se usa el antígeno O junto con el antígeno H (antígeno flagelar) se definen serotipos. Como las adhesinas son el principal componente implicado en los primeros estadíos de la infección/colonización se catalogan como factores de virulencia. Aunque las adhesinas no son propiamente un elemento agresivo para las células del hospedador, son necesarias para la colonización de los tejidos ya sea en procesos infectivos o en la colonización del intestino de forma comensal. Esto significa que la presencia de las adhesinas no hace a la bacteria propiamente patógena sino que es un elemento más que contribuye a su patogenicidad.

Tipo de Adhesina	Patotipos asociados	Descripción
Intimina a	EPEC, EHEC	Adhesina, induce la respuesta de TH 1; 10 variantes descritas
Adhesinas Afa/Dr	DAEC, UPEC	Adhesina, se une al factor DAF
Fimbria P (Pap)	UPEC	Adhesina; induces la expresión de citoquinas
CFAs	ETEC	Adhesina, >20 factores distintos designados como: CFA, CS ó PCF
Fimbria Tipo1	All	UPEC Adhesina; se une a la proteína UPK1B (*uroplakin*)
Fimbria F1C	UPEC	Adhesina
Fimbria S	UPEC, MNEC	Adhesina
Bundle-forming pilus (BFP)	EPEC	Pilus Tipo IV
Fimbria AAF	EAEC	Adhesina; >4 subtipos
Paa	EPEC, EHEC	Adhesina
ToxB	EHEC	Adhesina
Efa-1/LifA	EHEC	Adhesina
Long polar fimbriae (LPF)	EHEC, EPEC	Adhesina
Saa	EHEC	Adhesina
OmpA	MNEC, EHEC	Adhesina
Curli	Varios	Adhesina; Se una a la fibronectina

Tabla 1. Adhesinas comunes de *E. coli*. Adaptada de (Kaper et al. 2004)

Las toxinas, sin embargo, presentan una actividad agresiva frente a las células del hospedador y representan un auténtico factor de virulencia (**Tabla 2**). Generalmente las toxinas provocan alteraciones en las células del hospedador que pueden variar entre la secreción de metabolitos esenciales; principalmente iones, la modificación del citoesqueleto o promover la muerte celular ya sea por apoptosis o por necrosis. Las razones evolutivas de la existencia de las toxinas son discutibles, pero la funcionalidad es clara. En la mayoría de los casos la funcionalidad es la obtención de metabolitos necesarios que en determinados ambientes pueden resultar escasos (Kaper et al. 2004; Croxen and Finlay 2010). En otros casos se neutraliza la acción del sistema inmune, como por ejemplo las toxinas RTX de las cepas UPEC que atacan a los leucocitos y los lisan (Kaper et al. 2004). Muchas toxinas son efectores que necesitan sistemas de secreción independientes para ser segregadas/actuar sobre las células del hospedador como pueden ser los sistemas de secreción tipo I, II ó III (Kaper et al. 2004). No existe ninguna evidencia de que alguna toxina use el sistema de secreción tipo IV en *E. coli* aunque existan ejemplos del uso de este sistema, en otras bacterias, para secretar efectores. Otras toxinas son autotransportadores y por lo tanto no dependen de otros sistemas de secreción para actuar sobre las células del

hospedador. Podemos encontrar toxinas codificadas por genes localizados tanto en el cromosoma (en islas de patogenicidad, transposones o ICEs[1]) como en plásmidos, lo que indica que la transferencia horizontal juega un papel fundamental en la patogénesis de *E. coli,* pudiendo una cepa inocua convertirse en un patógeno oportunista al adquirir horizontalmente los factores de virulencia necesarios (Leimbach et al. 2013).

Otro grupo de factores de virulencia importante son aquellos que facilitan la adaptación de la cepa a nuevos nichos. En este grupo se encuentran los sistemas de importación de hierro. El hierro es un oligoelemento esencial para muchos procesos biológicos tanto de procariotas como de eucariotas. La concentración de hierro libre en los mamíferos, principal hospedador de *E. coli*, es muy baja, se estima que en la sangre es de 10^{-25}M y en otros fluidos como la orina es aún menor, mientras que la concentración de hierro en el citosol de *E. coli* es de 10^{-6}M Así que para que *E. coli* pueda colonizar y/o invadir un tejido como el epitelio de la uretra (por ejemplo en una cistitis) o en la sangre (en una bacteriemia) necesita un sistema que le proporcione el hierro necesario (Wiles et al. 2008). La limitación del hierro libre es un mecanismo de defensa básico en los animales contra las bacterias patógenas.

Los organismos eucariotas también usan sideróforos para captar el hierro del medio. Las transferrinas que transportan el hierro en el plasma, son proteínas altamente conservadas en los mamíferos, aves, peces o anfibios y tienen una alta afinidad por el hierro libre (Constante de afinidad: Kd ~10^{-20}) (Fischbach et al. 2006). En general los sistemas de captación de hierro se componen de dos factores: el sideróforo y el transportador. Los sideróforos son pequeñas moléculas que se secretan al medio y que tienen una afinidad muy grande por el Fe^{+3} que se encuentra libre en el medio. La afinidad media de los receptores de hierro de los animales es de $Kd \sim 10^{-20}$ y la de los sideróforos es $Kd \sim 10^{-49}$ lo que muestra la gran eficiencia de estos sistemas y

1 ICE "Integrative Conjugative Element" son elementos genéticos insertados en el cromosoma que contienen un sistema de conjugación y se pueden mover independientemente sin necesidad de otros genes del hospedador.

la capacidad de adaptación de estas bacterias a ambientes hostiles (Fischbach et al. 2006).

Factor	Patotipo	Clase de Toxina	Objetivo	Actividad/Effecto
Enterotoxina termo lábil (LT)	ETEC	Effector tipo II, Sub-unidad AB	G_S	Activa la adenil-ciclasa causando la secreción de iones
Toxina *Shiga*	EHEC	Sub-unidad AB	rARN	Daña el rARN provocando la inhibición de la síntesis de proteínas. Induce apoptosis.
CDT (Toxina de distensión-ciclo-letal)	Varios	Sub-unidad ABC	ADN	Actividad DNasa I. Bloquea la mitosis en la fase G2/M.
Enterotoxina Shigella 1 (ShET1)	EAEC EIEC	Sub-unidad AB	-	Secreción de iones.
Ureasa	EHEC	Sub-unidad ABC	Urea	Disocia la urea en NH_3 y CO_2
EspC	EPEC	Autotransportador	Desconocido	Serin-proteasa. Secreción de iones.
EspP	EHEC	Autotransportador	Desconocido	Serin-proteasa. Rompe el factor V de coagulación.
Hemoglutinina termo-sensible (Tsh)	ExPEC APEC	Autotransportador	Hemoglobina	Degrada la hemoglobina y libera el hierro.
Pet	EAEC	Autotransportador	Espectrina	Serin-proteasa; Secreción de iones; Citotóxico
Pic	UPEC EAEC EIEC	Autotransportador	Desconocido	Proteasa, mucinasa
Sat	UPEC	Autotransportador	Desconocido	Vacuolación.
SepA	EIEC	Autotransportador	Desconocido	Serin-proteasa.
SigA	EIEC	Autotransportador	Desconocido	Secrecion de iones.
Inhibidor del ciclo celular (Cif)	EPEC EHEC	Efector tipo III	Desconocido	Bloquea la mitosis en la fase G2/M mediante la inactivación de Cdk1.
EspF	EPEC EHEC	Efector tipo III	Desconocido	Induce apoptosis.
EspH	EPEC EHEC	Efector tipo III	Desconocido	Modula la formación de pedestales y filopodias.
Map	EPEC EHEC	Efector tipo III	Mitocondria	Altera el potencial de membrana de la mitocondria.
Tir	EPEC EHEC	Efector tipo III	Nck	Promueva la nucleación de las proteínas del citoesqueleto. Perdida de las micro-vellosidades. Actividad GAP.
IpaA	EIEC	Efector tipo III	Vinculina	Despolimerización de la actina.
IpaB	EIEC	Efector tipo III	Caspasa 1	Apoptosis. Liberación de IL-1. Inserciones de membrana.
IpaH	EIEC	Efector tipo III	Nucleo	Modula la inflamación.
IpgD	EIEC	Efector tipo III	PtdIns $(4,5)P_2$	Inositol 4-fosfatasa. Formación de burbujas en la membrana
VirA	EIEC	Efector tipo III	Tubulina	Destabilización de los microtubulos. Erizado de la membrana
StcE	EHEC	Efector tipo II	Inhibidor C1-esterasa (C1-INH)	Disocia el complejo C1-INH. Altera las cascadas complementarias.
HlyA	UPEC	Toxina RTX	Eritrocitos Leucocitos	Lisa las células.
Ehx	EHEC	Toxina RTX	Eritrocitos Leucocitos	Lisa las células.
Factor necrotizante citotoxico (CNF-1,-2)	MNEC UPEC NTEC		Cdc42, RhoA Rac	Alteración del citoesqueleto. Necrosis.
LifA/Efa	EPEC EHEC		Linfocitos	Inhibe la activación de los linfocitos. Promueva adhesión.
Enterotoxina *Shiga* 2 (ShET2)	EIEC ETEC		Desconocido	Secreción de iones.
Enterotoxina Termo-estable (STa)	ETEC	Toxinas termo-estables	Guanilato ciclasa	Activa la guanilato-ciclasa y provoca la secreción de iones.
Enterotoxina Termo-estable (STb)	ETEC	Toxinas termo-estables	Desconocido	Aumenta la concentración intracelular de calcio provocando la secreción de iones.
Toxina enteroagragativa Termo-estable (EAST)	Varios	Toxinas termo-estables	Guanilato ciclasa	Activa la guanilato-ciclasa y provoca la secreción de iones.

Tabla 2. Resumen de las principales toxinas producidas por *E. coli*. Adaptada de (Kaper et al. 2004)

Una vez que el sideróforo se ha asociado con el F^{+3} el trasportador introduce el complejo en el citosol de la bacteria y se libera el hierro del sideróforo. *E. coli* tiene un sistema propio de sideróforo-transportador (Enterobactina *Ent-Fep*) pero éste resulta ineficiente para la invasión/colonización en los procesos infectivos. El hospedador tiene su propio sistema para actuar sobre las enterobactinas, la lipocalina asociada a gelatinasa de neutrófilos (ngal) o lipocalina-2. Esta proteína actúa como agente bacteriostático uniéndose específicamente y secuestrando las enterobactinas del sistema *Ent-Fep* (Wiles et al. 2008). Es la adquisición de otros sistemas lo que proporciona la capacidad a *E. coli* para infectar ya que algunos de ellos, como el sistema *IroABCDEN* es capaz de escapar de la lipocalina-2.

Factor de Virulencia	Patotipo	Descripción
IbeABC	MNEC	Promueve la invasión
AslA	MNEC	Promueve la invasión
Dispersinas	EAEC	Promueve la colonización, contribuye a la penetración en la mucosa
Antigeno K	MNEC	Factor antifagocítico; Existen más de 80 tipos distintos
Aerobactinas (IutA)	EIEC	Adquisición de Hierro, sideróforo
Yersiniabactinas (FyuA)	Varios	Adquisición de Hierro, sideróforo
IreA	UPEC	Adquisición de Hierro, receptor de sideróforos
IroN	UPEC	Adquisición de Hierro, receptor de sideróforos
SitA	UPEC	Adquisición de Hierro, receptor de sideróforos
Chu (Shu)	EIEC, UPEC, MNEC	Adquisición de Hierro, Trasportador del grupo *hemo*
Flagelina	Todos	Mobilidad, más de 50 serotipos distintos (H). Induce la expresión de citokinas.
Lipopolisacárido	Todos	Más de 180 tipos distintos de antígenos O. Induce la expresión de citokinas

Tabla 3. Otros factores de virulencia de *E. coli*. Adaptada de (Kaper et al. 2004)

La mayoría de los sistemas de captación de hierro, se transmiten horizontalmente y es común encontrarlos codificados en plásmidos. Otros, como por ejemplo *IroABCDEN*, pueden estar integrados en el cromosoma (Sorsa et al. 2003). Además, se puede encontrar una redundancia de estos sistemas, por ejemplo el plásmido pEcoS88 tiene hasta 3 sistemas de importación de hierro: *IroN, SitA e IutA* (Peigne et al. 2009).

Aunque la mayoría de las infecciones causadas por *E. coli* son fácilmente curables en ocasiones se requiere el uso de antibióticos durante el tratamiento. Las infecciones pueden complicarse gravemente con la presencia de cepas resistentes a antibióticos. Recientemente el centro de control de

enfermedades y prevención de Estados Unidos (CDC)[2] resaltaba en un informe (http://www.cdc.gov/drugresistance/threat-report-2013/), el problema a nivel mundial de la evolución y dispersión de cepas resistentes a antibióticos. En este informe se indica una larga lista con los patógenos más peligrosos que han adquirido resistencia o multi-resistencia a antibióticos. En ella se encuentran *Enterobacteriaceae* resistentes a carbapenémicos, entre los que se incluye *E. coli*. Se pueden encontrar gran variedad de resistencias a antibióticos en *E. coli*, no sólo a carbapenémicos. Es difícil determinar si *E. coli* es especialmente propenso a adquirir resistencias o si lo que vemos es un sesgo en las bases de datos debido a que algunas cepas de *E. coli* han sido especialmente estudiadas en los últimos años. Lo que sí sabemos es que es capaz de adquirir un gran abanico de resistencias a todo tipo de antibióticos: macrólicos, beta-lactámicos, carbapenémicos, quinolonas, etc. Además pueden ser adquiridos horizontalmente (generalmente en plásmidos) o como ya hemos comentado anteriormente, por mutaciones puntuales en los genes diana de algunos antibióticos, como por ejemplo el caso de los genes *gyrA* involucrados en la resistencia a quinolonas (Vila et al. 1994). Las resistencia a antibióticos, así como a otros biocidas y/o metales, juega un papel fundamental en la evolución bacteriana (De la Cruz and Davies 2000). Los determinantes de resistencia permiten a las bacterias expandirse en sus nichos ecológicos e incluso invadir nuevos nichos que estén bajo presión de antibioticos. Tienen especial relevancia para la salud aquellos ambientes en los cuales la presión es mucho mayor, como por ejemplo los hospitales, centros de la tercera edad e incluso en las granjas, donde el uso de antibióticos está generalizado (Trevisi et al. 2014) .

En el año 2011 *E. coli* tomó especial relevancia debido al brote de la cepa enteroagregativa y enterohemorrágica O104:H4 focalizado en el centro de Alemania, que finalmente se saldó con la muerte de 39 personas y secuelas importantes en la gran mayoría de los pacientes supervivientes (Buchholz et al. 2011; Frank et al. 2011; Grad et al. 2013). Esta cepa

2 CDC. Centers for Desease Control and Prevention. 2013 Threat Report 2013 Antimicrobial Resistance. CDC website.

combinaba una serie de características que la hacían especialmente peligrosa por su virulencia. Por una parte, la adquisición de los factores de virulencia asociados con cepas Shiga-tóxicas (*stx2 , iha, lpfO26 y lpfO113*) y por otra los factores de adherencia propios de cepas enteroagregativas (*aggA, aggR, set1, pic, aap*) generó una virulencia más activa de lo normal. La combinación de una mayor adherencia puede potenciar la absorción de la toxina *stx2* provocando el síndrome urémico hemolítico que provoca infecciones complicadas en su tratamiento. Además la presencia de un gen de resistencia a beta-lactámicos de espectro extendido (ESBL), en un plásmido provocó que en muchos casos el tratamiento con antibióticos no fuese efectivo agravando la situación (Grad et al. 2013). Éste fue un claro ejemplo de la peligrosidad de las resistencia a antibióticos y cómo la suma de factores específicos puede hacer de *E. coli* un patógeno mortal. Finalmente, el brote se controló efectivamente y son pocos los casos de infecciones que se reportan actualmente debidos a esta cepa.

Otro ejemplo bien distinto es el de *E. coli* O25b:H4-ST131, objeto de estudio en esta tesis y que merece un capítulo.

E. coli O25b:H4-ST131: Pandémico, multirresistente y comensal.

En el año 2008 varios grupos detectaron en distintos países la aparición de una cepa causante de infecciones urinarias que presentaba un fenotipo de multirresistencia (Coque et al. 2008; Lau et al. 2008; Nicolas-Chanoine et al. 2008). El estudio genético mostró que estas cepas pertenecían a la secuencia tipo ST131 y serotipo O25b:H4. Posteriores estudios han demostrado una amplia diseminación a nivel mundial hasta ahora nunca descrita (Rogers et al. 2011; Novais et al. 2012; Petty et al. 2014). Las razones de su exitosa diseminación no son claras, pese a que el fenotipo de multirresistencia fue inicialmente apuntado como causa principal (Singer 2015). Si bien muchos de los aislados estudiados son resistentes a beta-lactámicos y fluoroquinolonas, no son mayoría los portadores de beta-

lactamasas pero sí los resistentes a fluoroquinolonas (J.R. Johnson et al. 2010). La resistencia a fluoroquinolonas está asociada a mutaciones específicas en los genes *gyrA* y *parC* (Johnson et al. 2013), mientras que la resistencia a beta-lactámicos, se asocian generalmente a la presencia de genes codificantes de beta-lactamasas de espectro extendido (BLEEs) (Totsika et al. 2011; Johnson et al. 2013; Petty et al. 2014), principalmente localizados en plásmidos. Algunos estudios retrospectivos han encontrado muestras de estos clones desde 1991 (Novais et al. 2012), lo que indica que ST131 es un comensal común de nuestra flora intestinal en contraposición a casos como el del brote alemán de O104:H4. Aunque se considera a ST131 un clon, en realidad contiene una amplia diversidad de cepas y/o virotipos que han sido identificados por distintos grupos y técnicas, alguno de los cuales forma parte de esta tesis (Novais et al. 2012; Blanco et al. 2013; Petty et al. 2014).

Al analizar su contenido en factores de virulencia, como razón para su rápida diseminación, destaca la presencia de la adhesina de tipo I, codificada por el gen *fimH*, que pertenece a un operón de 7 genes *(fimAICDFGH)*. La proteína *FimH* es determinante para la adhesión a los eritrocitos, leucocitos y células epiteliales, mediante la detección de manosa. Existen estudios que demuestran que *fimH* es esencial para la colonización del epitelio uretral y por lo tanto para la infección urinaria (Krekeler et al. 2012). Se puede clasificar a ST131 según el alelo de *fimH* que posea (J.R. Johnson et al. 2010). Los estudios epidémicos muestran que el alelo *fimH30* es predominante en las colecciones descritas por (Petty et al. 2014) y que aglutina a los virotipos A, B y C establecidos por Blanco *et al.* Además se asocia el alelo *fimH30* con la resistencia a fluoroquinolonas (Dahbi et al. 2014). Hay que aclarar que la fimbria de tipo I (cualquiera de sus alelos) no interfiere en la resistencia a fluoroquinolonas, en este caso lo que se observa es una asociación entre un alelo especifico de *fimH* con los alelos resistentes a fluoroquinolonas de *gyrA* y *parC*. De esta forma, la asociación entre estos alelos parece que se ha producido por deriva genética, y que existe una expansión clonal de estos alelos, aunque no queda claro cuál de ellos es el responsable de la

expansión. Por una parte si el alelo *fimH30* es más efectivo en la adhesión al epitelio esto podría provocar una mejora en la colonización y por tanto explicar la expansión, pero no existe ningún estudio que concluya esta hipótesis. Se ha descrito previamente que distintos alelos de *fimH* se asocian a distintos fenotipos de virulencia, modificando la capacidad de colonización (Hamrick et al. 2000; Schembri et al. 2001; Weissman et al. 2007; Schwartz et al. 2013).

Sin embargo, los estudios realizados *in vivo* no clarifican si este alelo de *fimH* tiene un fenotipo más exitoso en colonización (Totsika et al. 2011; Mora et al. 2014). Por otra parte, la resistencia a fluoroquinolonas podría explicar la expansión. Esta hipótesis es controvertida por tres razones principales: (1) la procedencia habitual de las muestras son hospitales, que como se ha mencionado anteriormente, se encuentran sometidos a una gran presión selectiva. (2) Por otro lado, el tratamiento más común en las infecciones urinarias son las fluoroquinolonas (ciprofloxacina, norfloxacina y levofloxacina); y (3) la mayoría de los urocultivos que se reportan proceden de infecciones recurrentes, y que por lo tanto se han recibido tratamiento antibiótico previamente. Esta cadena de acontecimientos puede generar un sesgo en los estudios que no ha sido evaluado, ya que podríamos estar sobrerrepresentando la muestra con cepas resistentes (Singer 2015). No tenemos información de la mayoría de infecciones urinarias que generalmente se tratan en atención primaria, y si el tratamiento inicial es exitoso no se reporta ningún urocultivo. Nos falta información acerca de las frecuencias reales de ST131 en la flora comensal que nos ayude a esclarecer cómo es la expansión clonal de esta cepa. Lo que parece claro es que a la luz de los datos, la mayoría de las infecciones complicadas están causadas por clones de ST131 con el alelo *fimH30*, pero si esto es extrapolable a la frecuencia de colonización o no, aún no lo sabemos. Además ST131 puede contener los genes típicos de cepas ExPEC como son *papA, papAH, papC, sfa/focDE, afa/draBC, iutA o kpsMII* (Johnson et al. 2007; Mora et al. 2014) que están asociados con la colonización del tracto digestivo (Singer 2015).

Plásmidos.

Existen muchas definiciones para los plásmidos, las más generalistas pueden resumirse en *"elementos extracromosomales, estables y auto-replicativos"*. Incluso las definiciones más sencillas para los plásmidos tienen sus excepciones. ¿Son todos los plásmidos extracromosomales? Podríamos considerar que un plásmidos es plásmido solo cuando no está integrado en el cromosoma, sin embargo los plásmidos integrados en el cromosoma o ICEs (*Integrative Congugative Elements*) son más comunes que los plásmidos independientes (Guglielmini et al. 2011), y además su integración/escisión del cromosoma parece ser un estado transitorio, por lo que podríamos encontrarlos en ambos estados en una misma población. ¿Son todos los plásmidos estables? Si entendemos "estables" como que son persistentes en la población, no necesariamente, muchos plásmidos se pierden por segregación debido a que son una carga metabólica que no ofrece ninguna ventaja evolutiva en ese momento y por lo tanto no son estables (Smillie et al. 2010). Si entendemos "estables" como, que son genéticamente estables y que no varían su contenido genético, tampoco es válido para todos los plásmidos. Existen plásmidos muy adaptados al hospedador que se convierten en pseudo-cromosomas y que se pueden encontrar en la mayoría de las cepas asociadas con esa especie (por ejemplo el segundo cromosoma de *Vibrio cholerae*) pero muchos otros plásmidos varían su contenido genético mucho más rápidamente de lo que lo hacen los cromosomas dando lugar a una variabilidad muy superior a la de las bacterias (Smillie et al. 2010; Petersen 2011). Si esto puede considerarse estable o no es otra discusión. ¿Son todos los plásmidos auto-replicativos? Tampoco, existen plásmidos que necesitan de la maquinaria del hospedador (más allá de la polimerasa o los tARNs) para poder replicarse como por ejemplo algunos plásmidos de *E. coli* de replicación por circulo rodante (Bruand and Ehrlich 2000).

A pesar de todo, la definición más básica sí que nos da una idea intuitiva de lo que es un plásmido. Un elemento ajeno al cromosoma, que de alguna forma es autónomo, que se replica, que evoluciona paralelamente al cromosoma y que en ocasiones es capaz de dispersarse entre las bacterias. Salvando todas esas excepciones que podemos aplicar a su definición, un plásmido es: *"un conjunto de módulos genéticos funcionales que están organizados en un estructura estable y auto-replicativa y que frecuentemente no contienen genes involucrados en las funciones esenciales de la célula"* (Frost et al. 2005).

Clasificación.

Los plásmidos son junto a los bacteriófagos y a los transposones conjugativos los principales actores en la transferencia horizontal. Los plásmidos son capaces de transmitir genes involucrados en virulencia, resistencia a antibióticos, detoxificación o interacciones ecológicas. Existen varios esquemas para clasificar los plásmidos. A diferencia de las bacterias, la clasificación filogenética de los plásmidos con las herramientas clásicas, como los arboles filogenéticos, tienen serios inconvenientes. Si bien la asociación entre los genes 16S rARN de las bacterias y su evolución es claro, en el caso de los plásmidos no existe ningún elemento común por el cual se puedan clasificar filogenéticamente. Por lo tanto su clasificación filogenética se restringe a la información obtenida a partir de módulos conservados como pueden ser la movilización o la replicación, pero que no son comunes a todos. Esta información suele quedar restringida a plásmidos cercanos, de modo que rápidamente se pierde la señal genética que nos permite asociar una historia evolutiva. Esto se debe a que los plásmidos no derivan de un único elemento, sino que cada familia de plásmidos es el producto de eventos evolutivos independientes, que además están muy condicionados por la transferencia horizontal.

La forma más extendida y aceptada de catalogar los plásmidos es el sistema de incompatibilidad. Dos plásmidos son del mismo grupo de

incompatibilidad si no pueden coexistir en un hospedador (Petersen 2011). Antes de la *"era molecular"* de la biología, los experimentos se basaban en el cultivo de las cepas conteniendo dos plásmidos para probar su compatibilidad, o no. Estos experimentos son tediosos y requieren de mucho tiempo, además de tener altas tasas de falsos positivos y falsos negativos. En la actualidad se extrapolan los grupos de incompatibilidad a partir de las secuencias de sus proteínas de replicación o de sus orígenes de replicación (llamados *oriV*). De esta forma se trata de clasificar los replicones mediante técnicas moleculares, generalmente por PCR. El esquema más extendido para tipificar plásmidos es el desarrollado por Carattoli *et al.* (Carattoli et al. 2005; García-Fernández et al. 2009; Villa et al. 2010) con el que se pueden clasificar los principales replicones de gammaproteobacterias. El problema de este esquema es que es útil en proteobacterias pero no clasifica otros plásmidos presentes en otras especies. Existen esquemas similares para *Firmicutes,* desarrollados por múltiples laboratorios (Jensen et al. 2010; Rosvoll et al. 2010; Lozano et al. 2012; Freitas et al. 2013; Wardal et al. 2013; Sadowy and Luczkiewicz 2014; Wardal and Markowska 2014). El tipado a través de los replicones tiene el inconveniente que en ocasiones algunos plásmidos tienen multi-replicones, y por lo tanto su clasificación puede resultar ambigua. Cierto es que generalmente ambos replicones pertenecen *sensu estricto* a los mismo grupos de incompatibilidad, pero a distinto sub-grupo, por ejemplo, el caso de los replicones de IncF, con subgrupos IncFIA, IncFIB o IncFII que pueden coexistir en el mismo plásmido (Villa et al. 2010). Otro de los problemas de estos esquemas basados es que se necesita conocer específicamente los replicones, por lo que el método no es capaz de detectar replicones previamente no descritos. Recientemente han propuesto esquemas de clasificación similares al MLST de bacterias. Este sistema, *plasmid multi-locus sequence type* (pMLST), está diseñado para el tipado de plásmidos de las familias IncF, IncI1, IncN, IncHI2 e IncHI1. De forma análoga al MLST, la presencia/ausencia de genes esenciales para el plásmido, se establece una clasificación para cada uno de ellos (Jolley and Maiden 2010; Carattoli et al. 2014).

Otro de los esquemas de clasificación propuestos para los plásmidos es el basado en el tipado a través de su relaxasa. Éste método ha sido desarrollado por nuestro laboratorio durante los últimos años (Francia et al. 2004; Garcillán-Barcia et al. 2009; Smillie et al. 2010; Alvarado et al. 2012). La relaxasa es una de las principales proteínas en el proceso de conjugación y por lo tanto está presente en todos los sistemas de conjugación, característica clave a la hora de hacer una clasificación.

Como se ha comentado previamente, la conjugación bacteriana es un proceso en el cual un plásmido es transferido de una cepa (donador) a otra cepa (receptor). En este proceso están implicados dos sistemas, el primero es el sistema de secreción, compuesto por una decena de proteínas. Este sistema es el encargado de formar el canal por el cual va a producirse la conjugación que se denomina pilus conjugativo. El otro sistema implicado es el de movilización, que en general esta compuesto por tres proteínas, la proteína acopladora, la proteína accesoria, y la relaxasa. Algunos sistema carecen de proteína accesoria, bien porque la función no es necesaria o porque la función puede ser complementada por otra proteína. Lo mismo ocurre con la proteína acopladora, aunque es menos frecuente (Smillie et al. 2010). En relación a su capacidad para conjugar, los plásmidos pueden clasificarse en 3 tipos: no movilizable, movilizables y conjugativos. Se considera que un plásmido no es movilizable cuando carece de relaxasa. Se considera movilizable si tiene relaxasa y un origen de transferencia (*oriT*) funcional. Se considera que un plásmido es conjugativo si además de relaxasa tiene integrado un sistema de secreción que permita la conjugación (Smillie et al. 2010). En este caso se dice que el plásmido tiene un sistema de conjugación completo y por lo tanto es autónomo para conjugar. En el caso de los plásmidos movilizables deben usar un sistema de secreción auxiliar que no está codificado en el mismo plásmido. Éste puede

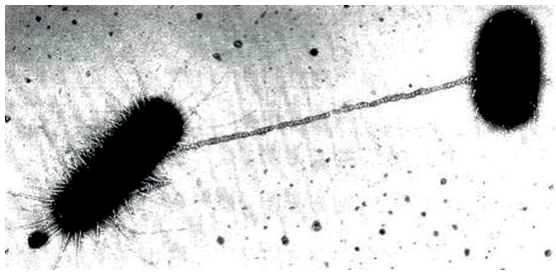


Figura 4. Formación del pilus conjugativo de un plásmido F entre dos *E. coli*

ubicarse en otro plásmido coexistente o en el cromosoma. La mitad de los plásmidos secuenciados son no movilizables, de la otra mitad, aproximadamente el 50% son conjugativos y el resto solo son movilizables (Smillie et al. 2010). Entre los plásmidos no movilizables hay que hacer una diferencia. Algunos plásmidos muy pequeños (< 2000 pb) sólo codifican lo necesario para replicarse, y por lo tanto no contienen ningún elemento que les permita diseminarse. Sin embargo este tipo de plásmidos, aunque no movilizables por conjugación, deben usar otro métodos para diseminarse, como por ejemplo la transducción (integrándose previamente en un fago), ya que su dispersión es evidente y están en alta frecuencia en especies como *E. coli* (Frost et al. 2005; Brolund et al. 2013).

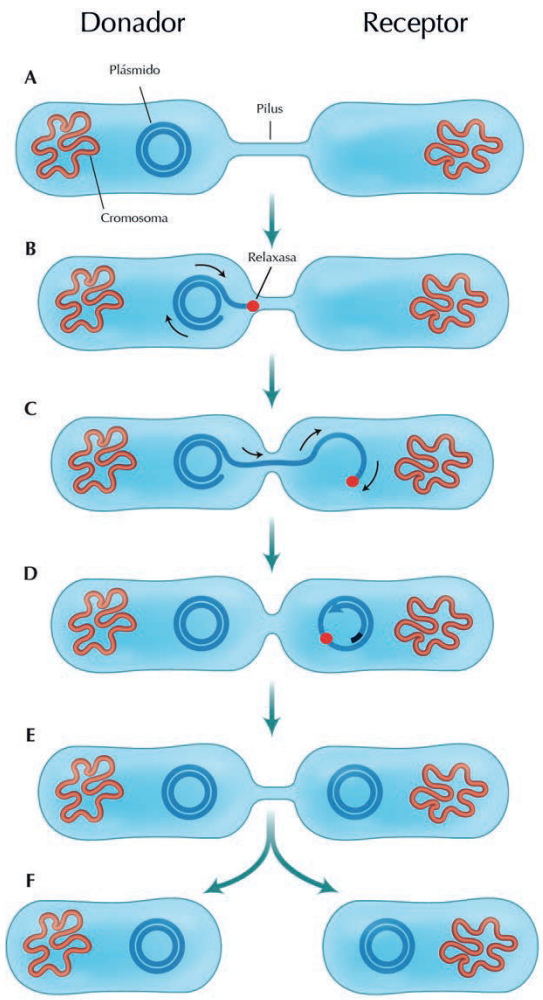


Figura 5. Conjugación Bacteriana: A|Formación del pilus conjugativo entre la cepa donadora y la receptora. **B|**La relaxasa corta por en el sitio especifico y se une a la hebra simple de DNA. **C|**La relaxasa guía al ADN a través del pilus conjugativo usando el sistema de secreción. **D|**La relaxasa une los extremos del ADN y se separa del ADN. En este punto se inicia la replicación de la hebra complementaria. **E|**Se disocia el pilus conjugativo. **F|**Como resultado de la conjugación tenemos la célula donadora inicial y la célula receptora con el plásmido transferido. A esta célula se le denomina transconjugante.

Por otra parte existen plásmidos, generalmente de gran tamaño (>200kb), que están fuertemente adaptados a su hospedador y que llegan a contener genes esenciales para la bacteria.

Se desconoce si estos plásmidos fueron adquiridos horizontalmente y posteriormente han perdido el sistema de conjugación, o si por el contrario son replicones que han ido creciendo durante la evolución hasta convertirse en elementos esenciales para la bacteria. Raramente contienen sistemas de conjugación funcionales que consigan transferir el plásmido completamente (Smillie et al. 2010). En algunas especies existen plásmidos de gran tamaño

muy adaptados al hospedador que se denominan megaplásmidos (Frost et al. 2005; Smillie et al. 2010). Incluso algunos cromosomas en realidad son megaplásmidos, ya que sus sistemas de replicación están más relacionados con los plásmidos que con los cromosomas, como es el caso de especies pertenecientes a los géneros *Burkholderia, Vibrio o Rhizobium* (Smillie et al. 2010).

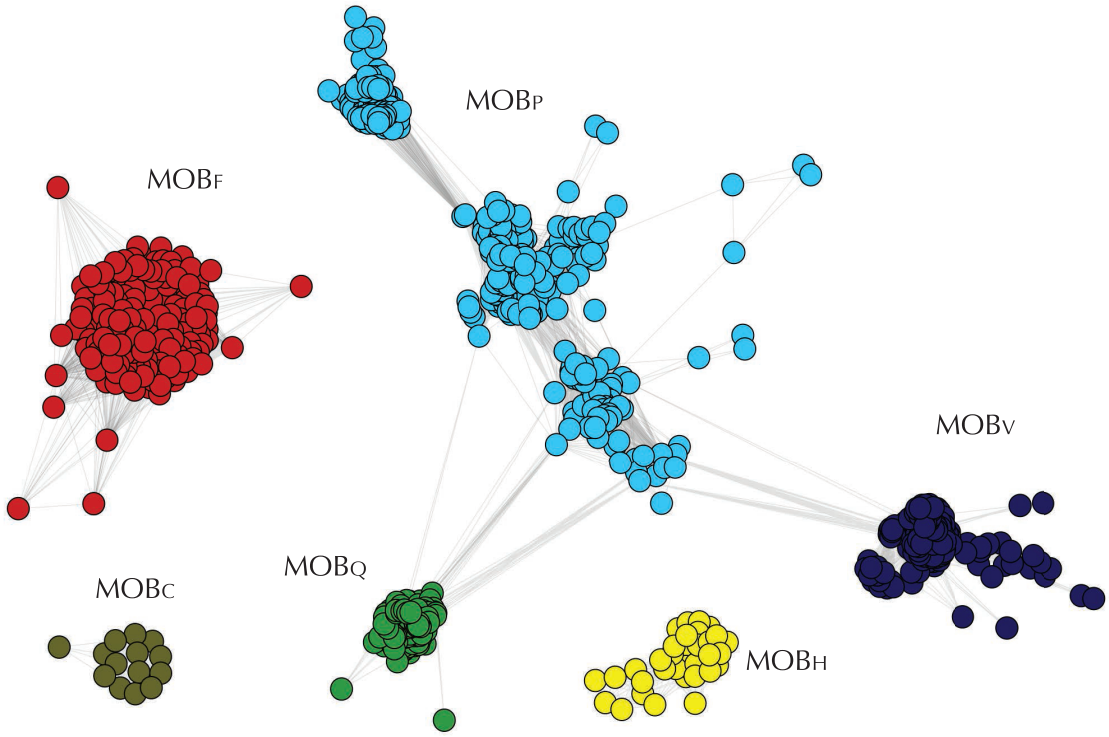


Figura 6. Red de homología de los seis grupos de relaxasas presentes en plásmidos. Esta red representa la relación filogenética entre las familias de relaxasas. Cada nodo representa un relaxasa y quedan conectadas entre sí si existe una relación de homología entre ellos. Los nodos quedan agrupados debido a que las conexiones entre ellos son más densas que con el exterior. La red se construye mediante la realización de un BLAST (Altschul 1997) de todos contra todos, con un E-value menor de 1e-6. Las relaxasas se recompilaron basándose en los criterios establecidos en (Smillie et al. 2010). Como se observa en la figura algunas de las familias de relaxasas están relacionadas filogenéticamente como son MOB_Q, MOB_V y MOB_P lo que indicaría un posible origen común. Sin embargo las familiar MOB_C, MOB_F y MOB_H parecen no tener una relación directa lo que indica que la diversificación entre estas familias es mucho mas antigua (Guglielmini et al. 2013).

Las relaxasas se pueden dividir en seis grandes familias denominadas MOB (MOB_F, MOB_Q, MOB_P, MOB_V, MOB_C, MOB_H) (ver **Figura 6**., además cada familia se puede subdividir en grupos filogenéticos, como pueden ser MOB_{F1} ó MOB_{F2} (Garcillán-Barcia et al. 2009)(ver **Figura 7**). Este esquema, que en principio se especificó para plásmidos, posteriormente se amplió con otras dos familias de relaxasas, en este caso asociadas con ICEs: MOB_B y MOB_T (Guglielmini et al. 2011).

Existe una asociación entre las relaxasas y los grupos de incompatibilidad clásicos (ver Figura 7), esto ocurre debido a que las relaxasas contienen una información evolutiva más profunda que la de los replicones. En otras palabras, mientras que los replicones no tienen una traza evolutiva visible, las relaxasas contienen información filogenética suficiente para establecer una jerarquía entre los plásmidos acerca de cuáles están más cercanos evolutivamente. Además son muy pocos los casos de plásmidos con más de una relaxasa, por lo que casi no existen tipados ambiguos como en el caso de los replicones (Smillie et al. 2010; Alvarado et al. 2012). El punto débil del esquema es que sólo la mitad de los plásmidos son movilizables y por lo tanto sólo la mitad de los plásmidos son potencialmente tipables.

El sistema de clasificación de plásmidos a través de su relaxasa se denomina DPTM (Degenerate Primer MOB Typing) y consiste en una serie de oligonucleótidos degenerados diseñados para amplificar las regiones conservadas de cada familia de relaxasas (Alvarado et al. 2012).

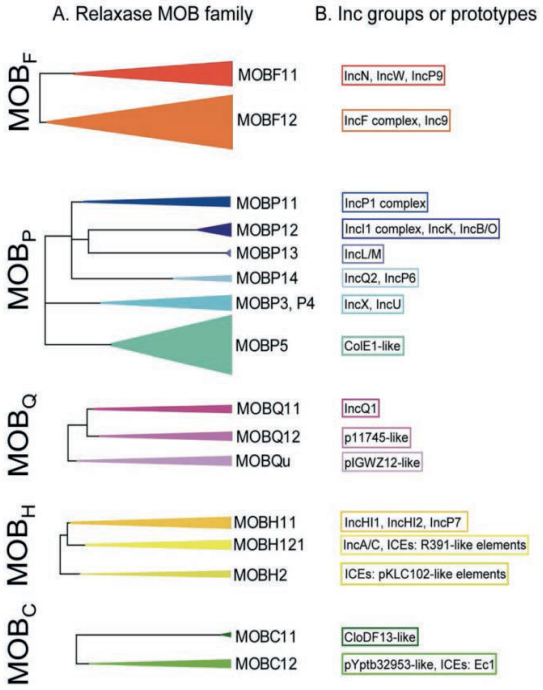


Figura 7. Filogenia resumida de las familias de relaxasas y su asociación con los principales grupos de incompatibilidad. Extraída de (Alvarado et al. 2012)

De igual forma que en el esquema de tipado por replicones, el proceso experimental consiste en realizar una serie de PCRs con cada conjunto de oligonucleótidos y mediante presencia/ausencia de los amplicones determinar el grupo al que pertenece el plásmido en cuestión.

Desde el punto de vista del análisis de plásmidos, sus clasificaciones tienen serios inconvenientes. Como hemos mencionado anteriormente a diferencia de las bacterias los plásmidos no tienen ninguna proteína común

que permita la realización de un análisis filogenético. Sólo aquellos plásmidos que tengan módulos en común son susceptibles de compararse en un estudio filogenético. El problema es que pueden ocurrir recombinaciones en las que alguno de estos módulos se pierda y por lo tanto ya no tengan esos elementos en común. Por ejemplo, si un plásmido pierde su relaxasa por una inserción o recombinación, al no tener este plásmido relaxasa no podemos tomar a la relaxasa como elemento común entre todos ya que esto dejaría a este plásmido fuera del análisis. Esto se puede solventar usando otro tipo de aproximaciones basadas en el contenido genético. En el caso de la filogenia son las mutaciones comunes o distintas las que agrupan o separan a las muestras del estudio. En una aproximación basada en el contenido genético lo que se busca son genes (de manera cuantitativa) que tienen en común dos muestras (Snel et al. 1999). En el caso de los plásmidos para cada uno se construye un perfil de presencia/ausencia de genes. Para ello, lo primero es construir una base de datos de referencia con todas las familias de proteínas presentes en el conjunto de los plásmidos, una vez que tenemos todas las familias, asignamos de forma binaria (0 o 1) la presencia de cada una de las familias en el plásmido. Posteriormente se calcula la distancia que existe entre los perfiles de cada uno de los plásmidos. Existen multitud de algoritmos que permiten calcular este tipo de distancias pero los más usados son *Jaccard, Manhattan o Euclideana*. Cada uno de ellos tiene sus pros y sus contras y su efectividad depende del conjunto de datos a analizar (Tekaia and Yeramian 2005; Snipen and Ussery 2010; Zhou et al. 2013). Con todas las distancias calculadas, se forma una matriz de distancias y se construye un dendrograma que igualmente puede estar basado en varios métodos como pueden ser UPGMA, Neighbord Joining o cualquier método basado en matrices de distancia.

Así podemos construir un dendrograma que nos da información jerárquica a cerca de las similitudes de los plásmidos sin la necesidad de que compartan todos ellos algún gen en común, cosas que sólo ocurre con plásmidos muy parecidos. Hay que tener siempre en cuenta que éste no es un

método filogenético por lo que cualquier conclusión acerca de la relaciones de evolución entre unos plásmidos y otros pueden llevar a equívoco (Snipen and Ussery 2010).

Cuando los plásmidos son relativamente cercanos suelen compartir un conjunto de genes que se suele denominar *"backbone"*. El backbone de un plásmido está formado por aquellos genes que codifican las funciones básicas del plásmido (replicación, persistencia y movilización). El backbone no siempre está totalmente conservado y es la distancia evolutiva entre dos plásmidos la que hará que el backbone sea mayor o menor, además de las posibles interferencias mediante recombinaciones (adquisición o pérdida de genes del backbone).

En general, las distancias entre dos plásmidos tienen una alta correlación con la filogenia de las relaxasas, por lo que para plásmidos conjugativos la relaxasa es un marcador extraordinario que nos indica las relaciones evolutivas del plásmido (Smillie et al. 2010; Petersen 2011).

Estructura y funcionalidad.

Los plásmidos tienen estructuras modulares que se pueden clasificar en cuatro tipos distintos: módulo de replicación, módulo de movilización/propagación, módulo de estabilidad y mantenimiento, y "el cargo". La replicación es el proceso esencial en la biología del plásmido. Aunque ya hemos comentado antes que no todos los plásmidos tienen una replicación autónoma, la gran mayoría son autosuficientes para la replicación. El módulo de replicación suele estar compuesto por una o más proteínas y por un origen de replicación denominado *oriV*. Además, el sistema de replicación controla el número de copias del plásmido dentro de la célula, esencial para la simbiosis entre la bacteria y el plásmido. Un número de copias excesivo afectaría negativamente a la capacidad adaptativa del hospedador al generar una carga metabólica extra, mientras que un número de copias bajo puede provocar la pérdida accidental del plásmido durante la división celular. Algunos sistemas de replicación también están

involucrados en la segregación del plásmido durante la división celular (Cervantes-Rivera et al. 2011). En algunos plásmidos, como los de la familia ColE-1, la replicación está controlada por ARNs antisentido y no por proteínas (Waters and Storz 2009). Otros, como ya hemos comentado, pueden tener más de un sistema de replicación que hace que puedan coexistir con otro plásmidos de su misma *especie*[3] como por ejemplo el plásmido pJIE186-2 (Zong 2013). No todos los plásmidos son compatibles con cualquier bacteria, sino que sólo son viables en algunos hospedadores de la misma especie. Según *"Biotechnology for Food and Agriculture"* se especifican como plásmidos de amplio rango de hospedador (*Broad Host Range, BHR)*[4] aquellos plásmidos capaces de replicar en diferentes especies; mientras que se denominan plásmidos de reducido rango de hospedador (*Narrow Host Range, NHR*)[5] aquellos plásmidos que sólo replican en una especie o un número reducido de ellas.

El módulo de movilización está compuesto por los genes de movilización (o genes *tra*) y, en el caso de que el plásmido sea conjugativo, por el sistema de secreción. En los plásmidos conjugativos de bacterias gram-negativas el sistema de secreción más habitual es el sistema de tipo IV (Type 4 Secretion System, T4SS, en inglés). Los sistemas de secreción tipo II y III, que son similares al sistema de secreción tipo IV, codifican una ATPasa de la familia *TadA* que está asociada con la generación del pilus y con la exportación/importación de ADN (Frost et al. 2005). En algunos casos (como veremos más adelante en los artículos que conforman esta tesis) algunos plásmidos tienen truncados los sistemas de movilización. Por ejemplo, el

3 Aunque no existen "especies" en los plásmidos tal y como lo entendemos para bacterias, se sobrentiende como especie a las familias de plásmidos con similares características, como son los plásmidos derivados del plásmido F de Escherichia coli o del plásmido de resistencia R100.

4 "broad-host-range plasmid." Glossary of Biotechnology for Food and Agriculture. UN Food and Agriculture Organization. 23 January. 2015http://www.expertglossary.com/definition/broad-host-range-plasmid.

5 "narrow-host-range plasmid." Glossary of Biotechnology for Food and Agriculture. UN Food and Agriculture Organization. 23 January. 2015http://www.expertglossary.com/definition/narrow-host-range-plasmid.

plásmido pEK499 analizado en (Woodford et al. 2009) tiene truncado parte del sistema de conjugación.

El módulo de mantenimiento es el encargado de todas aquellas funciones que influyen en la persistencia del plásmido en la población. Esto incluye los sistemas de partición que optimizan la segregación del plásmido durante la división celular. En ocasiones estos sistemas también influyen en la compatibilidad de los plásmidos. Ésta es una de las razones por las cuales el tipado clásico de la incompatibilidad tiene esas tasas de error. Dos plásmidos pueden no ser compatibles debido a sus sistemas de partición y no a su sistema de replicación como sería de esperar (Bouet et al. 2007). Los plásmidos con múltiples copias tienen el problema de generar multímeros debido a que los eventos de recombinación entre las distintas copias pueden ocurrir con relativa facilidad (Summers et al. 1993; Field and Summers 2011). Es por esto que algunos plásmidos contienen sistemas que resuelven los dímeros para incrementar su estabilidad (Thomas 2002).

Otros sistemas de mantenimiento son los encargados de eliminar a las poblaciones libres de plásmidos (Thomas 2002; Fernández-López et al. 2006) asegurando la persistencia en la población del plásmido. Entre estos sistemas se encuentran los sistemas toxina/antitoxina. Estos sistemas están compuestos por una toxina muy estable y una antitoxina inestable, de tal forma que si durante el proceso de división celular se perdiese el plásmido, al estar presente la toxina pero no la antitoxina (que se degrada rápidamente) la bacteria muere o impide su división (efecto bacteriostático) (Hayes 2003; Gerdes et al. 2005; Hayes and Van Melderen 2011). A nivel poblacional significa que sólo las bacterias portadoras del plásmido son viables, aunque esto afecta a la capacidad adaptativa de la bacteria, ya que su crecimiento se verá afectado por la eficiencia de los sistemas de partición que tenga el plásmido (Hayes 2003; Gerdes et al. 2005). Estos sistemas tienen transcendencia en la circulación de los plásmidos entre bacterias. Teóricamente, un plásmido con un sistema toxina/antitoxina solo podría verse sustituido por otro plásmido que contenga el mismo sistema toxina/antitoxina.

En la realidad los sistemas no son infalibles y en ocasiones se pueden perder plásmidos con sistemas toxina/antitoxina sin que exista ningún plásmido que lo sustituya (Hayes and Van Melderen 2011). Algunos poseen varios sistemas toxina/antitoxina que podrían ser un indicio de su capacidad de dispersión y de su promiscuidad (Rankin et al. 2012).

Existen además un conjunto de proteínas encargadas de la estabilización del plásmido después de la conjugación (Fernández-López et al. 2006). Este conjunto de proteínas incluye a sistemas anti-restricción que anulan los sistemas de defensa del hospedador (sistemas restricción/modificación) y sistemas de protección de ADN de cadena sencilla que protegen de las endonucleasas del hospedador (proteínas pertenecientes a la familia SSB (Meyer and Laine 1990; Lohman and Ferrari 1994)).

No todos los plásmidos tienen módulo de mantenimiento. En algunos casos las funciones de mantenimiento las realizan los sistemas de replicación. Existen plásmidos con un número de copias alto (>10) o muy alto (>100) en los que su permanencia en la cepa se basa en las pocas probabilidades de que durante la división celular alguna de sus copias no esté en uno de las células hija (Summers et al. 1993; Field and Summers 2011). Esto es frecuente en plásmidos pequeños (<2000 pb) que tienen un número de copias alto y que se encuentran frecuentemente en cepas de *E. coli* (Brolund et al. 2013).

El último módulo, el cargo, lleva genes con función adaptativa. En esta clasificación entran los factores de virulencia, genes metabólicos, genes de resistencia a antibióticos, etc. Algunos cargos son fundamentales para el fenotipo que adquiere la bacteria al obtener un plásmido, como por ejemplo en el caso de la resistencia a antibióticos. Así como en el caso de la patogenicidad donde un solo gen no convierte una bacteria comensal en una patógena (salvo contadas excepciones), en el caso de la resistencia a antibióticos el efecto puede ser directo. Por ejemplo, la adquisición de un plásmido codificante de una beta-lactamasa de tipo CTX-M puede ser

suficiente para que la bacteria sea resistente a cefalosporinas de tercera generación.

Los plásmidos tienen una tasa de recombinación aparentemente mayor que los cromosomas, lo que parece incrementar la adquisición de elementos móviles como transposones o integrones (Fernández-López et al. 2006). Este fenómeno parece estar motivado por la presencia de secuencias de inserción en proporciones mayores a lo que se puede encontrar en los cromosomas (Mahillon and Chandler 1998). Nuevamente hay que diferenciar entre el tipo de plásmidos. Los plásmidos endógenos, muy adaptados a la bacteria, suelen tener las mismas características que el cromosoma (contenido G+C, frecuencia de trinucleótidos, número de secuencias de inserción, etc…). Son los plásmidos movilizables, como por ejemplo los plásmidos de *Enterobacterias* MOB_F ó MOB_H (IncF, IncA/C, IncH etc…) los que tienen mayor contenido de secuencias de inserción. El hecho de ser plásmidos conjugativos que se diseminan entre las poblaciones hace que aumenten la posibilidad de adquirir elementos nuevos por el simple hecho de estar en presencia de una mayor diversidad de genes, en contraposición de lo que ocurre con los plásmidos endógenos que no suelen ser movilizables (Mahillon and Chandler 1998; Touchon and Rocha 2007). Algunos estudios indican que la mayor transferencia de ADN entre especies es por el intercambio de plásmidos, superando a la producida por fagos (transducción) (Halary et al. 2010). Esto significa que la mayoría de las recombinaciones en el cromosoma vienen precedidas por la adquisición de un plásmido. Es difícil observar plásmidos con segmentos de ADN atípicos que sin embargo en los cromosomas se sabe que son adquiridos de forma horizontal. Por ejemplo las islas de patogenicidad de *Escherichia coli* no se suelen observar en plásmidos y sin embargo se sabe que se transmiten horizontalmente. Esta transmisión puede ocurrir de distintas formas. Uno es que el elemento primero se inserte en un plásmido y posteriormente se produzca la recombinación de forma inversa. Este evento de doble recombinación (cromosoma -> plásmido, plásmido -> cromosoma) está ampliamente descrito en la bibliografía

(Thomas and Nielsen 2005). Otro mecanismo involucra la integración inicial de los plásmidos en el cromosoma, para posteriormente transferir gran parte de él, donde se encuentra incluida la región adquirida (por ejemplo una PAI). Finalmente, este fragmento recombinaría nuevamente en el cromosoma de la cepa receptora (Frost et al. 2005; Johnson and Nolan 2009). En último lugar, existen eventos de movilización de regiones del ADN que no necesitan de una recombinación previa y que son promovidos por algunos plásmidos conjugativos (Meyer 2009).

Plásmidos de *E. coli*.

E. coli es capaz de albergar una gran variedad de plásmidos, hasta 39 grupos de incompatibilidad distintos, la mayoría asociados con virulencia y/o resistencia a antibióticos (Johnson and Nolan 2009). Esto genera una gran plasticidad en los genomas de *E. coli*, y explica por qué no existe una asociación entre patotipo y filogrupos. En cuanto a la variedad de tamaños en los plásmidos de *E. coli* podemos encontrar plásmidos desde 1 kb hasta más de 200 kb. En cuanto a la prevalencia de los plásmidos, entre los plásmidos de gran tamaño sobresalen los plásmidos de los grupos IncF (MOB_{F1}), al igual que en el resto de *Enterobacterias* (Garcillán-Barcia et al. 2009). Los plásmidos MOB_F que incluyen a grupos de incompatibilidad como IncF, IncN, IncW o IncP9 han jugado un papel fundamental en la evolución de *E. coli* (Fernández-López et al. 2006; Garcillán-Barcia et al. 2009; Johnson and Nolan 2009; Smillie et al. 2010).

Muchos de los patotipos de *E. coli* tienen asociados diversos plásmidos fundamentales en su virulencia y/o resistencia (Johnson and Nolan 2009). Las cepas ETEC tienen asociado un plásmido que codifica factores de colonización y toxinas (por ejemplo pH10407_95 (Evans et al. 1975)); las cepas EAEC contienen frecuentemente un plásmido que interviene en la adherencia y en la secreción de toxinas (pAA (Nataro et al. 1987) y pO42 (GenBank: AB255435.1)). Las cepas EIEC codifican la maquinaria de invasión en el plásmido pINV (Parsot 2005). Un ejemplo de este plásmido es el

p53638_226 (GenBank: CP001064.1). Las toxinas de las cepas EHEC suelen estar contenidas en plásmidos como en el caso del pO157 (Burland et al. 1998; Makino et al. 1998). Las cepas EPEC contienen un plásmido de adherencia o *"EAF"* el primero que se secuenció fue pB171 (Tobe et al. 1999). Las cepas ExPEC suelen tener plásmidos de virulencia, se suelen distinguir en dos tipos: plásmidos *vir,* como el pVir86 (T.J. Johnson et al. 2010) y plásmidos ColV, como el como el pEcoS88 (Peigne et al. 2009). Además de todas estas asociaciones podemos encontrar gran variedad de combinaciones, cepas con plásmidos de virulencia y de resistencia al mismo tiempo. En la última década se están reportando muchos plásmidos que contiene tanto factores de virulencia como genes de resistencia a antibióticos (Venturini et al. 2010; Villa et al. 2010; Kunne et al. 2012).

En *E. coli* podemos encontrar también otros tipos de plásmidos. Son comunes los plásmidos del tipo ColE1 que codifican colicinas. Las colicinas forma parte de un sistema similar a un toxina/antitoxina pero que en este caso no solo afecta al hospedador sino que la toxina es secretada al medio, eliminando a las *E. coli* sensibles que se encuentran en el mismo nicho (Cascales et al. 2007). Esto es una ventaja evolutiva directa tanto para el hospedador, por eliminar a posibles cepas que puedan competir por el mismo nicho, como para el plásmido ya que se fomenta que todas las cepas presentes tengan el plásmido o sino serán sensibles a la colicina. Existen varios tipos de colicinas codificadas en distintos plásmidos, muchos de ellos movilizables lo que involucra que su dispersión puede jugar un papel fundamental en la estructura poblacional de *E. coli* en algunos nichos como puede ser el intestino humano.

Existen tambien plásmidos sin función evidente en su módulo cargo, o bien que carecen de este módulo. No portan genes de resistencia, ni factores de virulencia conocidos. Estos plásmidos contienen unos pocos genes de función desconocida, que son anotados como *"proteínas hipotéticas"* (Brolund et al. 2013). Algunos son conjugativos y pueden funcionar como posibles transmisores de elementos móviles . Otros, son pequeños plásmidos,

normalmente llamados crípticos ya que no tienen un módulo cargo y por lo tanto no se entiende que posible ventaja evolutiva dan a la bacteria. Serían plásmidos egoístas o parásitos (Burian et al. 1997).

Secuenciación de nueva generación y Bioinformática.

Sin género de dudas las nuevas tecnologías de secuenciación han ocasionado una revolución en la biología actual. El acceso de todo tipo de grupos de investigación a la posibilidad de secuenciar organismos completos con un coste muy bajo (actualmente menos de 300€/cepa para un organismo como *Escherichia coli*) ha impulsado los estudios genómicos a un nuevo nivel. Centrándonos en la microbiología y dejando a un lado la revolución que ha supuesto en campos como el cáncer donde la aportación de estas nuevas tecnologías resulta fundamental, la secuenciación masiva está abriendo una visión mucho más extensa acerca de la evolución, la dispersión o la epidemiología de todo tipo de bacterias (Chan et al. 2012; Didelot, Bowden, et al. 2012; Olsen et al. 2012). El incremento de información en las bases de datos es un claro ejemplo del crecimiento exponencial que está teniendo el uso de la secuenciación masiva en la biología (ver **Figura 6**).

En el campo de la secuenciación existe un punto de inflexión con la aparición de lo que se denominó: secuenciadores de segunda generación. Fue un cambio radical en cuanto al concepto de secuenciación. Mientras que en la secuenciación convencional se obtenía una única secuencia de aproximadamente 1.000 pb; en la secuenciación de nueva generación se obtienen millones de fragmentos de entre 32 a 300 pb, en función de la tecnología.

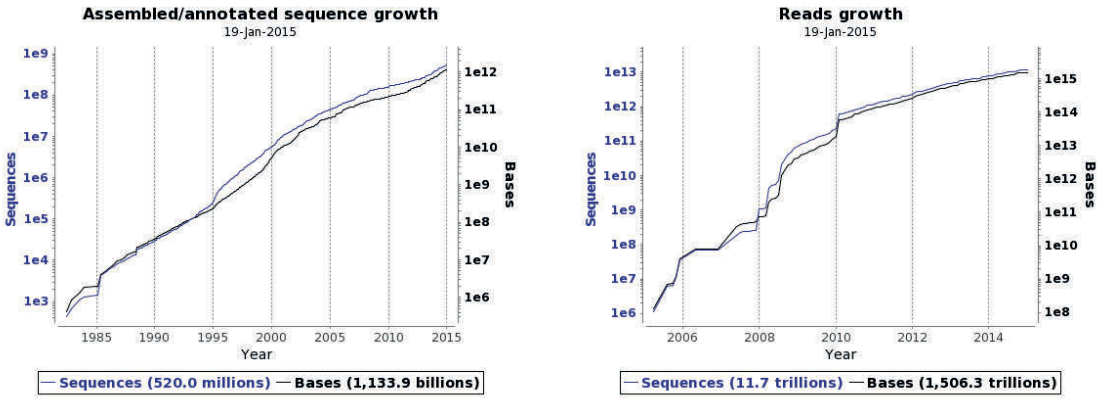


Figura 8. Estadísticas de la base de datos *European Nucleotide Archive*. Crecimiento de la base de datos de secuencias ensambladas y anotadas desde sus inicios (izquierda). Progreso de la base de datos de secuencias procedentes de plataformas de nueva secuenciación (derecha). Observar que el eje de ordenadas se encuentra en escala logarítmica, lo que implica que el crecimiento de ambas bases de datos es exponencial

Entre los años 2005-2007 aparecieron en el mercado los primeros secuenciadores de segunda generación como el 454 de Life Sciences que fue posteriormente adquirida por Roche, el Genome Analyzer de Solexa (posteriormente Illumina) y los secuenciadores SOLiD de Applied Biosystem y Heliscope de Helicos (MacLean et al. 2009). Independientemente de la tecnología, el concepto de secuenciación masiva es el mismo: dado un organismo, por ejemplo una bacteria, se fracciona su genoma y se secuencian paralelamente en millones de segmentos cortos de ADN. Estos fragmentos se denominan lecturas o *reads* en inglés.

En la actualidad existen tres tecnologías dominantes: Illumina, claro dominador del mercado con su serie de secuenciadores HiSeq, NextSeq y MiSeq; IonTorrent de la compañía Life Technologies, sucesor de los secuenciadores 454 de Roche por sus características; y Pacific Biosciences con su secuenciador PacBio RSII, una nueva tecnología que permite obtener fragmentos mucho mayores pero con mayor tasa de error (Van Dijk et al. 2014). Los secuenciadores SOLiD y Helicos llevan años en clara desventaja y es difícil encontrarlos en funcionamiento en los grandes centros de secuenciación. Recientemente Roche anunció el final del soporte oficial para los secuenciadores GS 454 FLX+ y GS Junior programado para mediados del

2016[6]. En la Tabla 4 se muestra un resumen del actual estado de las tecnologías de secuenciación.

Compañía	Modelo	Tipo de Carrera	Tiempo de Carrera	Longitud de Read (pb)	Capacidad
Roche	FLX+	Single end	23h	700	700Mb
	Junior+	Single end	18h	700	70Mb
Life Techonologies	PGM	Single end	4h-7h	200-400	600Mb-1Gb
	Proton	Single end	4h	125	8-10Gb
	SOLiD 5500W	Single & Pair end	10 días	2x 50	320Gb
Illumina	HiSeq2500	Single & Pair end	60h - 6 días	2x250 2x125	300Gb – 1Tb
	HiSeq X Ten	Single & Pair end	<3 días	2x150	1.8Tb
	NextSeq 1500	Single & Pair end	<3 días	2x150	100-120Gb
	MiSeq	Single & Pair end	<3 días	2x300	15Gb
Pacific Bioscience	RSII	Single end	2 días	50% reads > 10kb	5Gb(16 SMRT)
Helicos	Heliscope	Single end	10 días	30	15Gb

Tabla 4. Tabla comparativa de las distintas tecnologías de secuenciación disponibles en la actualidad (Enero 2015). Los datos están recogidos según las especificaciones de cada casa comercial. Los datos pueden variar en función de la configuración del equipo y de los reactivos usados.

Existen dos metodologías de secuenciación: *single-end* y *pair-end*. Como ya hemos mencionado la secuenciación masiva lee millones de pequeños fragmentos de ADN de forma paralela e independiente. Estos fragmentos de ADN no tienen por qué ser de la misma medida que el tamaño de las lecturas que se obtienen sino que en general son más grandes. En una secuenciación s*ingle-end* solo se secuencia un extremo del fragmento, mientras que en una secuenciación *pair-end* se secuencian los dos extremos. Al preparar las muestras para la secuenciación se pueden configurar el tamaño de los fragmentos de ADN que se van a secuenciar y por lo tanto las lecturas de ambos extremos quedan asociadas con una distancia. A esta distancia se le denomina *"tamaño del inserto"*. No todas las plataformas son aptas para este tipo de secuenciación, que es realmente útil en procesos como la secuenciación *de novo*.

6 Fuente: https://www.genomeweb.com/sequencing/roche-shutting-down-454-sequencing-business

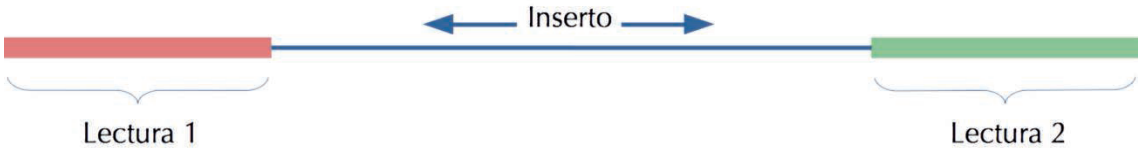


Figura 9. Esquema resumen de la secuenciación *pair-end.* Una secuenciación *pair-end* se compone de dos ciclos. En el primer ciclo se secuencia uno de los extremos (Lectura 1), cuando se alcanza la longitud deseada se detiene el ciclo. A continuación se reinicia la secuenciación pero en este caso por el extremo opuesto (Lectura2) y en dirección opuesta.

En el campo de la microbiología el salto cuantitativo y cualitativo ha sido muy significativo. Las capacidades de secuenciación superan por mucho las necesidades para secuenciar un microorganismo lo que nos permite secuenciar paralelamente cientos de muestras (Chan et al. 2012). Se está pasando de realizar estudios poblacionales o epidemiológicos basados en el tipado de las muestras por MLST o por PFGE (Pulsed Field Gel Electroforesis[7]) a realizar los estudios mediante secuenciación. La mejora en la calidad y profundidad de los datos es considerable. Tanto el MLST como el PFGE tienen limitaciones claras. El MLST por su parte no ofrece ningún tipo de información acerca del genoma accesorio, el cual es esencial en muchos casos para entender el fenotipo de la muestra. Por su parte en el PFGE se puede llegar a inferir cambios importantes en el genoma accesorio como puede ser la adquisición o pérdida de un plásmido pero no ofrece ningún tipo de información filogenética. La secuenciación de las cepas resuelve ambos problemas. La información filogenética es muy superior a la que se puede obtener mediante MLST, ya que en vez de basar la filogenia de la cepa en unos pocos genes (varía según la especie pero nunca superior a 10 genes) se puede reconstruir basándose en el genoma completo. Aunque los estudios han avanzado mucho, existen limitaciones inherentes a la tecnología y a la naturaleza de las bacterias.

El flujo de trabajo normal en el análisis de los resultados de secuenciación se compone de 4 pasos (Edwards and Holt 2013; Kisand and Lettieri 2013):

7 El PFGE es una técnica de tipado ampliamente usada en microbiología clínica. Se basa en la fragmentación del DNA mediante encimas de restricción que posteriormente se corre un gel de electroforesis aplicando una corriente eléctrica pulsada. El gel resultante es un patrón de bandas por el cual se puede extrapolar la similitud entre muestras, permitiendo la agrupación y clasificación .

1. Análisis de la calidad de las secuencias.
2. Ensamblaje.
3. Anotación funcional.
4. Análisis mediante genómica comparada.

Los secuenciadores se rigen por unos estándares establecidos. Aunque algunos de ellos tienen formatos propios de salida para la secuenciación, todos ellos se convierten al estándar mundial que es el formato FASTQ. El formato FASTQ fue inicialmente desarrollado por el Wellcome Trust Sanger Institute aunque con el tiempo se ha convertido en un estándar mundial. Un archivo FASTQ contiene no sólo la secuencia generada sino un valor de calidad por cada base secuenciada o QC (*Quality Control*). Durante el proceso de secuenciación uno de los pasos es la determinación del nucleótido que se está secuenciando. A este proceso se le denomina *base calling*. Todos los secuenciadores asocian una valor de fiabilidad a este *base calling* que indica con que probabilidad ese nucleótido está correctamente asignado. Esta calidad se representa con un valor en la escala PHRED (Ewing and Green 1998). Durante el proceso de *Análisis de la calidad de las secuencias* se eliminan aquellas con calidades malas. El umbral de la calidad suele venir condicionado por la plataforma de secuenciación utilizada, buscando siempre un equilibrio entre la profundidad de secuenciación y la calidad de las lecturas. Además se eliminan los fragmentos de las secuencias que son utilizadas durante el proceso de elaboración de las librerías (adaptadores, etiquetas, etc…). Existen también herramientas para la corrección de errores en las lecturas que mejoran levemente los análisis posteriores (Yang et al. 2013).

Una vez que las lecturas están correctamente recortadas y filtradas (proceso que en general se suele realizar en los centros de secuenciación y no por el usuario final) se procede al ensamblaje o ensamblaje *de novo*. El resultado del ensamblaje son una serie de secuencias ADN que se denominan *contigs*. Idealmente un solo *contig* debería representar todo el cromosoma, pero en la práctica los genomas tienen repeticiones que los

programas de ensamblaje no consiguen resolver. Durante años se han desarrollado algoritmos para el ensamblaje de datos obtenidos mediante secuenciación tradicional pero que son ineficaces con los datos de nueva generación (MacLean et al. 2009). Estos programas no están diseñados para manejar este tipo de datos; la distribución de los errores, la cantidad de secuencias o el tamaño de las lecturas hacen que estos programas estén obsoletos. Por ello una nueva generación de ensambladores ha ido apareciendo junto con las nuevas plataformas de secuenciación. La principal innovación ha sido la implementación de métodos basados en grafos de *de Bruijn*. A diferencia de los algoritmos clásicos de ensamblaje (denominados OverLaping Consensus, OLC) también basados en grafos, en los grafos de *de Bruijn* los nodos representan fragmentos de una longitud fija denominados *k-mer* y los enlaces representan el número de veces que dichos nodos solapan entre ellos en el conjunto de secuencias. En los algoritmos clásicos cada lectura es un nodo y su enlace con otro nodo depende de si existe un solapamiento o no. Aunque a primera vista no es clara la diferencia, los recursos computacionales son muy distintos. En los métodos basados en *de Bruijn* se almacenan todos los posible *k-mer* que existen en las secuencias, muchos de ellos son repetidos y por lo tanto sólo se almacenan en memoria una vez, mientras que en los métodos clásicos cada nodo se guarda por separado. Además cada vez que un nodo es añadido en la red de solapamiento debe compararse con todos los nodos existentes para establecer con cuales es solapante y enlazarlos. En los métodos basados en *de Bruijn* sólo se enlazan entre si los *k-mer* resultantes de cada lectura, osea los *k-mer* adyacentes que resultan de cada lectura. Por lo tanto no es necesario realizar una búsqueda sobre toda la red resisando cuales son los *k-mer* solapantes lo que hace que los recursos computacionales sean mucho menores. Una vez que los grafos quedan establecidos cada programa utiliza sus propios algoritmos para encontrar los caminos óptimos que reproducen mejor el genoma. A pesar de que los métodos basados en *de Bruijn* han demostrado su eficacia existen diversos programas para ensamblar basados en técnicas de solapamiento, por ejemplo SHARCGS (Dohm et al. 2007), SSAKE (Warren et

al. 2007), VCAKE (Jeck et al. 2007), ABySS (Simpson et al. 2009) o Newbler que fue diseñado por Roche para sus secuenciadores. Pero los grandes dominadores de los ensambladores están basados en los grafos de *de Bruijn*, algunos de los más usados son: Velvet (Zerbino and Birney 2008), MIRA (Chevreux et al. 1999), SPADES (Bankevich et al. 2012), ALLPATHS (Butler et al. 2008) ó SOAPdenovo (Luo et al. 2012). Existen varias comparativas al respecto (Hernandez et al. 2008; Narzisi and Mishra 2011; Salzberg et al. 2012; Rahman and Pachter 2013) pero los resultados suelen estar muy condicionados por el origen de los datos. Por ejemplo, la calidad de las secuencias en ocasiones resulta esencial ya que los programas basados en *de Bruijn* son mucho más sensibles a los errores de secuenciación. Un error en la secuencia genera un *k-mer* distinto al real, mientras que en el solapamiento (OLC) se pueden permitir cierto grado de libertad y un único error no sería determinante en la asignación de la unión entre dos nodos. Por esta razón algunos programas como SPADES incluyen correctores de errores para las lecturas (Bankevich et al. 2012).

A pesar de todos los avances en la algoritmia de los ensambladores existen limitaciones que van más allá de los programas usados. En el fondo son situaciones en las que debemos resolver un sistema de ecuaciones en donde existe más de una solución. Se podría llegar a realizar una aproximación heurística por la cual se escoge una solución válida, pero eso no significa ni que sea la mejor solución ni que sea la real, por lo que los programas ante esta adversidad no continúan con el ensamblaje y dejan los *contigs* aislados. Por esta razón los resultados que obtenemos al ensamblar una bacteria es un conjunto de *contigs* y muy raramente es un único *contig* representando al cromosoma (u otros *contigs* representando a los plásmidos en el caso de que los tuviese).

El principal problema en el ensamblaje de secuencias *de novo* son las repeticiones. A este respecto es muy recomendable la lectura de (Treangen and Salzberg 2012). Para solucionar los problemas que ocasionan las repeticiones es indispensable la información que se obtiene al realizar

secuenciaciones con *pair-end*. Como ya hemos descrito antes, en una secuenciación *pair-end* se puede inferir que dos lecturas están a una distancia determinada porque ambas pertenecen al mismo fragmento de ADN.

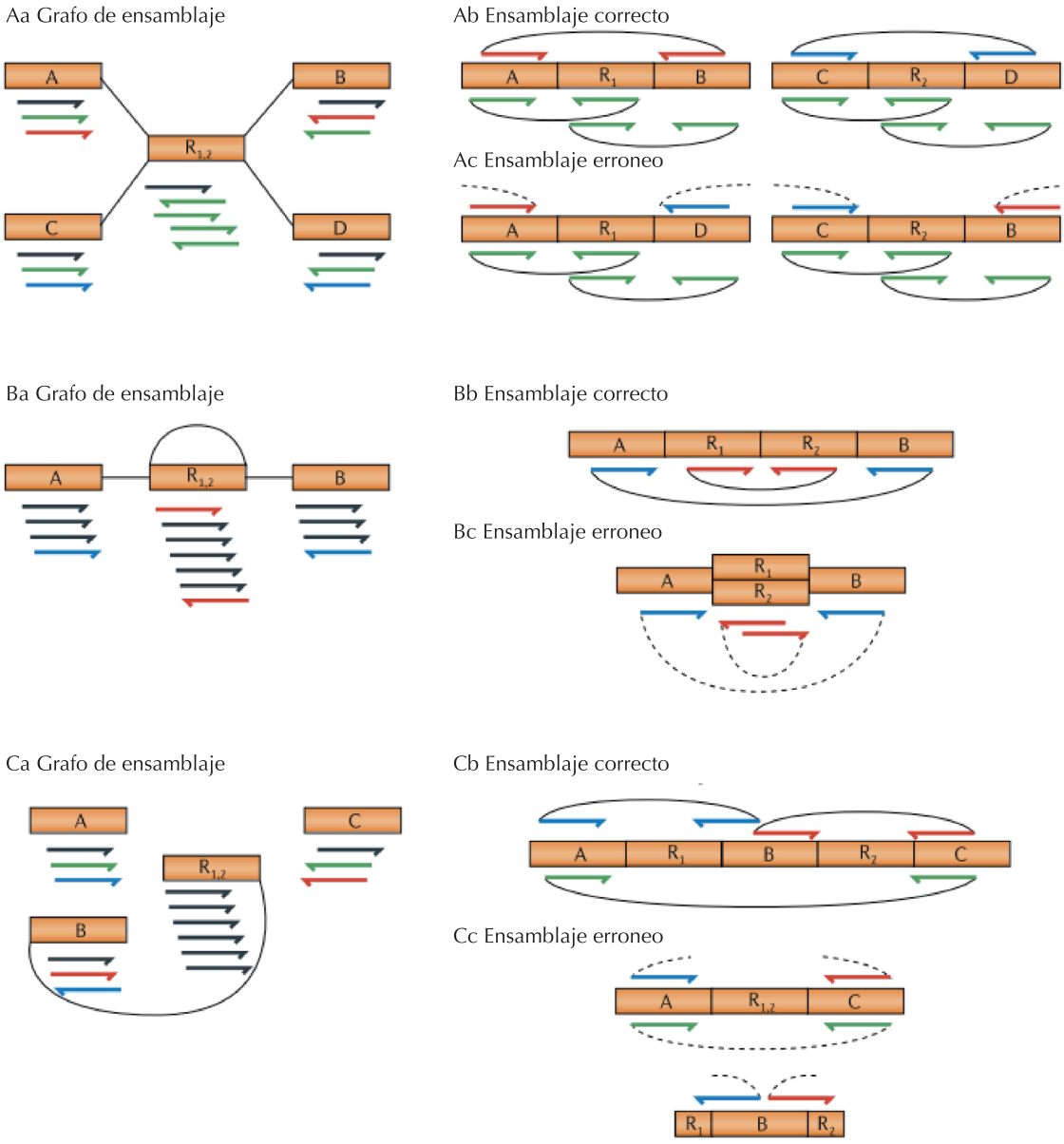


Figura 10. Errores de ensamblaje provocados por las repeticiones. A| Error de ensamblaje por una mala organización de los *contigs*. Aa| Un ejemplo de un grafo de ensamblaje que involucra a 6 *contigs* 2 de ellos idénticos (R1 y R2) . Las flechas que se muestran debajo de cada *contig* representan los reads que alinean con ellos. Ab| Situación real de cada *contig*. Ac| Dos ensamblajes erróneos formando quimeras causadas por las repeticiones. B| Error de ensamblaje por simplificación de dos repeticiones en un solo contig. Ba| El grafo de ensamblaje muestra 4 *contigs* 2 de ellos son idénticos. Bb| El ensamblaje correcto en donde las lecturas se encuentran en su orientación correcta y con las repeticiones ordenadas consecutivamente. Bc| Ensamblaje erróneo en donde las repeticiones se han reducido a un solo *contig* (con mayor cobertura, pero uno solo). C| Error de ensamblaje debido a repeticiones insertadas en distintos puntos. Ca| Grafo de ensamblaje consistente en 5 *contigs*, 2 de los cuales son idénticos. Cb| Ensamblaje correcto en donde se observa el orden correcto de los *contigs*. Cc| Ensamblaje erróneo. Las dos copias idénticas aparecen resumidas en un solo *contig* flanqueadas por los dos extremos, mientras que el *contig* central (B) aparece aislado y con parte de los extremos de las repeticiones flanqueándolo. Figura adaptada de (Treangen and Salzberg 2012).

Esto es muy importante porque añade información al grafo de ensamblaje, no solo se pueden conectar dos nodos por su solapamiento o porque tienen *k-mer* comunes sino que si son lecturas del mismo fragmento se sabe que están cercanos, a una distancia aproximada del tamaño del inserto y por lo tanto relacionados.

Como se muestra en la Figura 10 los problemas ocasionados por las repeticiones son variados. El más habitual es el ejemplo de secuencias repetidas dispersas por el genoma. Este tipo de repeticiones son muy comunes, elementos como transposones o secuencias de inserción están totalmente conservadas y dan repeticiones perfectas a lo largo del genoma. Imaginemos la situación mostrada en la Figura 10C si no existe ninguna lectura que salte la repetición el algoritmo es incapaz de saber si después de fragmento A+R1 le sigue la región B ó C y por lo tanto lo común es que la salida del programa sean los *contigs* A+R1, B y C de manera aislada o incluso A, B, C y R en *contigs* sueltos. Este tipo de problemas no es dependiente del algoritmo de ensamblaje sino del problema matemático que representa. La única forma de resolverlo y si disponemos de información extra que nos indique cual es el orden de los *contigs*. Para ellos existen varias soluciones. La primera es generar librerías de *pair-end* con insertos muy grandes. Las librerías convencionales de *pair-end* (por ejemplo de Illumina) tiene un tamaño máximo que viene limitado por el tamaño del fragmento de ADN que es capaz de secuenciar. En el caso de Illumina, no se aconsejan realizar

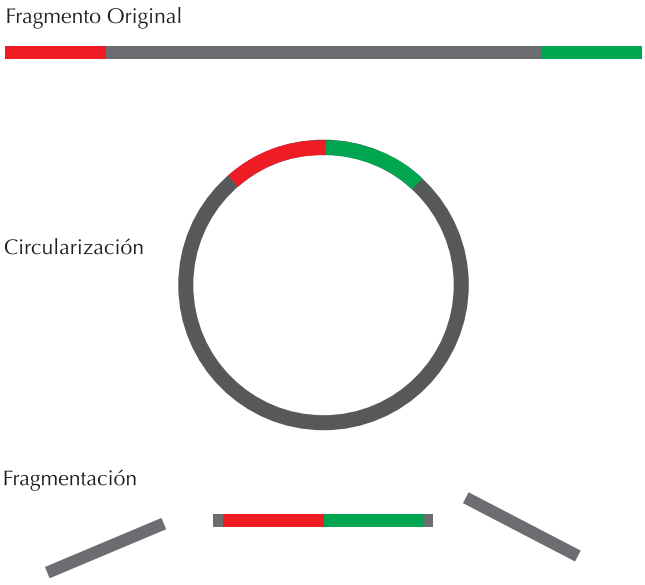


Figura 11. Esquema de preparación de una librería *mate-pair*. En inicio se fragmenta el genoma en fragmentos grandes, mayores de 1kb e incluso 1Mb en función del destino de la muestra. Posteriormente se circulariza el DNA, generalmente biotinizando los extremos. Finalmente se vuelve a fragmentar, pero en este caso en tamaños acordes con una librería *pair-end* (300-500pb) y se purifica la muestra para enriquecer los fragmentos que contienen los

librerías con insertos mayores de 500pb, de esta forma la repetición más larga que se puede saltar mediante la información del *pair-end* es de 500pb, insuficiente en la mayoría de los casos.

Para salvar este impedimento se realizan librerías *mate-pair*. Estas librerías se basan en la circularización de segmentos grandes de ADN (1kb a 1Mb), posteriormente se fragmenta el ADN circularizado en fragmentos del tamaño indicado para secuencias (500pb por ejemplo). Finalmente se enriquecen los fragmentos que contienen la unión de los extremos del ADN inicial (ver **Figura 11**).

La forma más habitual para el enriquecimiento es marcar los extremos del ADN inicial con biotina, así en el proceso de enriquecimiento seleccionamos la biotina que está contenida en la unión del ADN al circularizar y aumentamos la probabilidad de que los fragmentos seleccionados contengan ambos extremos de ADN.

De esta forma lo que se está secuenciando son los extremos de un ADN de un tamaño muy superior al que se puede alcanzar por *pair-end*. El *mate-pair* también tiene sus limitaciones y aunque mejora los ensamblajes no es la solución definitiva. El primer problema es que el enriquecimiento no es perfecto. Debemos pensar que para que el *mate-pair* sea óptimo debemos secuenciar cada uno de los extremos del ADN original, esto significa que a la hora de enriquecer el elemento seleccionado, por ejemplo mediante la captura biotina, deberíamos tener suficiente información en ambos extremos para que cada una de las lecturas pertenezca únicamente a un extremo. En la práctica este sistema tiene sus fallos y en ocasiones sólo se secuencia un extremo o alguna de las lecturas es una mezcla de ambos extremos por lo que introduce ruido en el proceso de ensamblaje.

Para obtener ensamblajes óptimos se recomienda combinar más de un tipo de librerías, por ejemplo *pair-end* junto con *mate-pair* o un *pair-end* de Illumina junto con un set de datos de 454. En ocasiones, al usar librerias *pair-end* o *mate-pair* se da la circunstancia de que una de las lecturas (R1) solape

con un *contig* (C1) y la otra lectura (R2) solape con otro *contig* (R2). Esta información se usa para relacionar ambos *contigs* asumiendo que en la realidad están contiguos en el genoma, pero que por alguna razón no se puede determinar la secuencia que existe entre ellos. En esos casos se inserta un conjunto de bases ambiguas marcadas con "N" entre ambos *contigs* y se fusionan en uno solo. Al suceso de que dos *contigs* queden conectados entre sí por una lectura se le denomina *scaffold*.

La que parece como la solución definitiva, es el uso de los secuenciadores más modernos (secuenciadores de tercera generación) que obtienen lecturas superiores a las 5kb, como por ejemplo el secuenciador PacBio RSII de Pacific Biosciences (ver **Tabla 4**). Estos datos en combinación con una secuenciación común (Illumina, 454 etc…) ofrecen los mejores resultados, llegando en muchos casos a ofrecer genomas cerrados. Es necesario el uso de una secuenciación común para poder corregir la tasa de errores cometidos por el PacBio que aún sigue siendo demasiado elevada para conseguir ensamblajes completos. Los secuenciadores PacBio han sido utilizados en la caracterización de cepas relevantes como la O104:H4 causante del brote urémico-hemolítico de Alemania (Rasko et al. 2011). Existen varios los artículos de revisión que hacen mención de la mejoría que supone el uso de estos secuenciadores (Chan et al. 2012; Koren et al. 2013; Huddleston et al. 2014) pero su uso aún es minoritario debido al alto coste y a los pocos equipos disponibles en el mundo.

Los secuenciadores PacBio son el exponente de los llamados secuenciadores de tercera generación. Sin embargo no es el único secuenciador de este tipo. La empresa Oxford Nanopore lleva años desarrollando un dispositivo muy prometedor que ofrece lecturas largas (desde 1kb hasta 10kb) con un coste muy bajo. En estos momentos (Enero 2015) se encuentra activo el programa de pruebas del MinION, el primer secuenciador de Oxford Nanopore en ver la luz (Mikheyev and Tin 2014). Se trata de un dispositivo compacto y portable. Este dispositivo del tamaño de un pequeño disco duro externo que se puede conectar a cualquier ordenador a

través de un puerto USB ha levantado durante años una gran expectativa. La promesa de una secuenciación a partir ADN sin amplificar, lecturas de gran tamaño, sin desviación por el contenido GC del ADN y la posibilidad de escalar distintos dispositivos para aumentar las capacidades del sistema, han suscitado un gran interés. De mismo modo que la secuenciación mediante PacBio RSII, la tasa de error por base es significativa y por lo tanto no es apto para algunas aplicaciones, como la detección de mutaciones puntuales. Sin embargo, la longitud de sus lecturas en combinación con datos de secuenciación Illumina supone un gran avance para completar genomas. Este secuenciador tiene unas ventajas añadidas, dado que será capaz no solo secuenciar ADN, sino también ARN, proteínas e incluso la detección de compuestos químicos (nanoporetech.com/technology/analytes-and-applications-ADN-ARN-proteins/analytes-and-applications-ADN-ARN-proteins).

El resultado del ensamblaje oscila entre las decenas de *contigs* (un ensamblaje muy bueno) hasta miles de *contigs* que hace que sea imposible realizar un análisis correcto. El resultado habitual se encuentra en torno a un centenar de *contigs*. Existen varios parámetros para medir la calidad de un ensamblaje. El más usado es *N50*. El valor *N50* de un ensamblaje indica que *"El 50% de las secuencias ensambladas se encuentra en contigs superiores a este valor"* (Salzberg et al. 2012). Es una medida que indica la calidad del ensamblaje aunque no es determinante y existen otros parámetros como la cantidad de *contigs* largos o el número total de bases ensambladas, que se deben observar para determinar si un ensamblaje es óptimo o no.

El proceso de anotación funcional consiste en la predicción de los genes y su caracterización. Para la predicción de los genes se pueden usar diversos programas como GeneMark (Besemer et al. 2001) o Glimmer (Delcher et al. 1999) recomendados en el manual del NCBI (www.ncbi.nlm.nih.gov/books/NBK174280/). Una vez determinados los CDS[8] de cada gen se procede a su anotación que generalmente se realiza mediante BLAST

8 *Coding Sequence* (CDS). Es el segmento de DNA que codifica para una proteína

usando una base de datos no redundante como pueden ser RefSeq o Uniref100. Existen alternativas como RATT (Otto et al. 2011) en donde se transfiere la anotación de una referencia (una bacteria cercana filogenéticamente a la nuestra y que esta previamente anotada) a la bacteria que se está analizando. Existen aplicaciones que integran la predicción de los genes y su anotación como por ejemplo BG7 (Pareja-Tobes et al. 2012), DIYA (Stewart et al. 2009), PROKKA (Seemann 2014) o EuGene-PP (Sallet et al. 2014). Finalmente, existen servidores que implementan todas las herramientas necesarias para el análisis de secuenciación como por ejemplo RAST (Aziz et al. 2008; Overbeek et al. 2014), ISGA (Hemmerich et al. 2010), BASys (Van Domselaar et al. 2005), WeGAS (Lee et al. 2009) o KOBAS (Wu et al. 2006; Xie et al. 2011).

En cualquier caso, llegados a un punto, lo que tenemos es el genoma representado por un conjunto de *contigs*. El problema es que no podemos diferenciar a cada uno de los elementos del genoma: el cromosoma y los plásmidos. El objetivo principal de esta tesis es desarrollar herramientas que nos permitan el estudio de los plásmidos contenidos en las bacterias.

Una vez que el ensamblaje y la anotación esta finalizada se procede a la ultima fase del análisis. Dependiendo de los objetivos que se establezcan al inicio del estudio su usan distintas aproximaciones. Por ejemplo, en los estudios evolutivos y filogenéticos en donde el objetivo es conocer las relaciones evolutivas de las muestras bien entre ellas y/o con bacterias conocidas se usan herramientas para poder identificar las mutaciones existentes. Nuevamente existen dos aproximaciones distintas, una de ellas es usar directamente las lecturas de la secuenciación y “mapear” una referencia para identificar que mutaciones existen entre la referencia y las lecturas, y otra distinta es usar los ensamblajes para observar modificaciones genomicas. La técnica de mapear frente a una referencia es usada ampliamente en eucariotas para la identificación de mutaciones en enfermedades como el cáncer (Schweiger et al. 2011). Existe gran variedad de utilidades tanto para el alineamiento de las lecturas frente a la referencia (Bowtie (Langmead and

Salzberg 2012), BWA (Li and Durbin 2009), BLAT (Kent 2002) o SOAP2 (R. Li, Yu, et al. 2009)) como para la identificación de mutaciones (VarScan2 (Koboldt et al. 2012), Pindel (Ye et al. 2009), GATK (McKenna et al. 2010), SOAPsnp (R. Li, Li, et al. 2009), SAMtools (H. Li et al. 2009)) incluso algunas son especificas para organismos procariotas como la utilidad Breseq (Barrick et al. 2014; Deatherage et al. 2014; Deatherage and Barrick 2014). Estos métodos permiten la identificación de mutaciones puntuales (SNP) y pequeñas inserciones y deleciones (Indels), pero sin embargo no son útiles frente a grandes re-arreglos como son la adquisición o perdida de islas genómicas, recombinaciones del genoma o la adquisición/perdida de plásmidos.

Para visualizar grandes modificaciones genéticas es necesario usar datos a partir de los ensamblajes. Se pueden realizar comparaciones entre dos cepas (aproximación *one-to-one*), un conjunto de cepas frente a una referencia (aproximación *one-to-many*) o alineamientos múltiples sin usar referencia (aproximación *many-to-many*) . La mayoría de las utilidades que realizan one-to-many realizan también one-to-many. Destro de estas herramientas podemos encontrar el paquete Mummer (Kurtz et al. 2004) que es la base de muchas otras utilidades como ACT (Carver et al. 2005), EasyFig (Sullivan et al. 2011) o Abacas (Assefa et al. 2009). Otras utilidades basadas en referencias usan BLAST (Altschul 1997) como base para sus comparativas. En este conjunto de utilidades destaca BRIG (Alikhan et al. 2011) por su versatilidad y la calidad de las figuras finales (Edwards and Holt 2013). En todas las aproximaciones que usen una referencia siempre encontraremos la misma limitación. Solo se puede observar aquello que se tiene en común con la referencia o la ausencia de elementos que tiene la referencia, pero nunca los elementos que tiene los genomas que no tiene la referencia.

Para salvar esta limitación existen métodos que no usan referencia y que son capaces de comparar un conjunto de genomas entre sí realizando un alineamiento multiple. Las dos herramientas mas usadas son Mauve (Darling et al. 2004) y Mugsy (Angiuoli and Salzberg 2011). Estas herramientas son

capaces de detectar los re-arreglos del genoma, grandes recombinaciones, inserciones o deleciones. Mauve incluso es capaz de establecer el numero de ortólogos que tienen en común las cepas o detectar los SNPs o pequeñas mutaciones a nivel genético.

Existen otros métodos que no se basan en alineamientos que permiten identificar los ortólogos comunes (core-genome) de un conjunto de muestras o incluso la identificación de los eventos de recombinación que hayan podido producirse. OrthoMCL (Li et al. 2003) es la herramienta más usada para la caracterización de ortólogos (1465 citaciones bibliográficas), pero existen más como por ejemplo BLASTO (Zhou and Landweber 2007), BratNextGen (Marttinen et al. 2012) o INPARANOID (Remm et al. 2001). Además se pueden implementar métodos caseros basados en programas de clustering de secuencias como CD-HIT (Li and Godzik 2006) o kClust (Hauser et al. 2013). Solo es necesario filtrar los resultados para obtener aquellos genes o proteínas (en función de si trabajamos con el genoma o con el proteoma) que son comunes en todas las muestras.

La determinación del core-genome es muy útil en los estudios filogenéticos. A pesar de que se pueden calcular arboles filogenéticos a partir de procedimientos de mapeo y búsqueda de mutaciones, la realidad es que estos métodos son menos precisos ya que requieren regiones altamente conservadas, mientras que el core-genome puede admitir mayor grado de libertad y por lo tanto la filogenia está basada en más información.

A pesar de que existen herramientas concretas para determinar el core-genome existe un gran vacío en utilidades que nos permitan analizar los genes accesorios. Existen herramientas con algunos objetivos específicos como identificar secuencias de inserción (ISfinder (Siguier et al. 2006)), fagos (PHAST (Zhou et al. 2011), Phage_Finder (Fouts 2006)) o Islas genómicas (CompBio (Tu and Ding 2003), INDeGenIUS (Shrivastava et al. 2010), IslandPath (Hsiao et al. 2003), GIDetector (Che et al. 2010), SIGI-HMM (Waack et al. 2006), IVOMS (Vernikos and Parkhill 2006)).

Existen diversas herramientas para caracterizar cepas a partir de datos de secuenciación. Se pueden realizar análisis de MLST *in silico*, predicción de genes de resistencia a antibióticos, factores de virulencia e incluso contenido plasmídico (Edwards and Holt 2013). Por ejemplo el software SRST2 (Inouye et al. 2014) contiene un conjunto de herramientas que aúna las principales bases de datos para el tipado completo de una cepa (MLST, resistencias a antibióticos, factores de virulencia y plásmidos). El tipado de los plásmidos está basado en la herramienta PlasmidFinder (Carattoli et al. 2014) que usa los replicones caracterizados para *Enterobacteriaceas* desarrollados por el grupo de Carattoli *et al.* (Ver el apartado **Clasificación** del capitulo **Plásmidos**) y los esquemas de pMLST depositados en la base de datos pública www.pubmlst.org/plasmids.

Aunque el tipado de los plásmidos pueda resultar útil, en muchos casos es insuficiente para realizar estudios de diseminación apropiados. Con las frecuencias de aparición que tienen algunos plásmidos en algunas especies, por ejemplo plásmidos IncF en *Escherichia coli* (Villa et al. 2010; Carattoli et al. 2014), identificar estos plásmidos no es significativo aunque tengan el mismo conjunto de replicones. Es necesaria una caracterización completa para poder establecer si un plásmido está diseminado o simplemente son plásmidos de la misma familia.

PLAsmid Constellation NETworks (PLACNET)

PLACNET ha sido desarrollado para resolver el problema de discriminar los *contigs* que conforman el cromosoma de aquellos que pertenecen a cada uno de los plásmidos que contiene la bacteria. Ya hemos repasado la importancia que tienen los plásmidos en la evolución, la transferencia de material genético o en el fenotipo de las bacterias. Por estas razones la caracterización de los plásmidos es fundamental no solo para poder realizar un análisis completo de una bacteria sino para realizar un estudio comparativo sobre un conjunto de muestras. Resulta fundamental

para realizar estudios de diseminación de plásmidos como el propuesto en el artículo (de Been et al. 2014), que forma parte de esta tesis .

PLACNET está basado en 2 premisas:

1. Todos los *contigs* de un plásmidos tienen similitud con un mismo conjunto de plásmidos.
2. Los *contigs* de un plásmido tienden a tener más *scaffold links* entre ellos que con el resto de elementos.

Siguiendo estas dos premisas se calculan dos conjuntos de datos. Primero todas las referencias para cada uno de los *contigs*, para ello se hace un BLAST de todos los *contigs* frente a una base de datos con todos los cromosomas y plásmidos cerrados. Segundo se calculan todos los posibles *scaffolds* entre *contigs* alineando todas las lecturas frente a los *contigs* y buscando aquellas lectura que tengan cada extremo en *contigs* distintos (para más detalles leer el capítulo *Material and Methods* de (Lanza et al. 2014)). Esta información se implementa en un grafo en donde quedan conectados los *contigs*, que representan los nodos entre ellos a través de sus *scaffolds* y a los cromosomas o plásmidos de referencia que cumplan con los umbrales de inclusión.

Idealmente, en el grafo, cada elemento (cromosoma y plásmidos) queda definido por un agrupamiento de nodos que representan los *contigs* y las referencias. En la realidad las repeticiones que existen en los elementos y que son comunes a cromosomas y plásmidos interconectan los agrupamientos llegando incluso a hacer que sea imposible definir donde empieza un elemento y donde acaba otro. Para resolver este problema la red se poda manualmente identificando estas repeticiones y eliminándolas o duplicándolas, en función del tamaño que tienen, maximizando la separación de los agrupamientos. Esta es quizá la mayor limitación que tiene el método ya que, de momento, es un proceso que se elabora manualmente. Otra de sus debilidades es la dependencia de las referencias disponibles. Cuantas menos referencias similares al plásmido que se quiere definir, peor se definirá. El

método también es sensible ante secuencias nunca observadas en determinados contextos genéticos. Por ejemplo un transposón nunca observado en una familia de plásmidos puede no quedar agrupado por las referencias junto con el resto de los contigs del plásmidos. En estos casos los *scaffold links* juegan un papel fundamental a la hora de la definición del plásmido.

Atendiendo al funcionamiento de PLACNET se deduce que el método es más robusto identificando plásmidos epidémicos que plásmidos aislados. Si tenemos un plásmidos diseminado en un conjunto de bacterias, en los grafos este plásmido generará siempre un agrupamiento muy similar, ya que siempre rescatará las mismas referencias, principal fuente de información para el método y por lo tanto su definición siempre será muy parecida, por lo que a la hora del análisis su secuencia es muy parecida.

Dada la importancia que tienen los plásmidos dentro de la microbiología no se pueden subestimar los plásmidos en los análisis genómicos. Hasta ahora no existían herramientas que permitiesen realizar estos estudios con la profundidad suficiente. Ahora abrimos una nueva vía de estudio que nos permitirá investigar el impacto de la diseminación y adquisición de plásmidos en comunidades bacterianas, la dinámica de diseminación de los plásmidos en brotes epidémicos o la frecuencia de los plásmidos en comunidades complejas, como por ejemplo la microbiota humana, en donde podríamos observar cuales son las especies que funcionan como reservorios naturales de los plásmidos y entre cuales son posibles los intercambios.

Publicaciones

JCM
Journals.ASM.org

Four Main Virotypes among Extended-Spectrum-β-Lactamase-Producing Isolates of *Escherichia coli* O25b:H4-B2-ST131: Bacterial, Epidemiological, and Clinical Characteristics

Jorge Blanco,[a] Azucena Mora,[a] Rosalia Mamani,[a] Cecilia López,[a] Miguel Blanco,[a] Ghizlane Dahbi,[a] Alexandra Herrera,[a] Juan Marzoa,[a] Val Fernández,[b,c] Fernando de la Cruz,[b,c] Luis Martínez-Martínez,[b,d] María Pilar Alonso,[e] Marie-Hélène Nicolas-Chanoine,[f,g] James R. Johnson,[h] Brian Johnston,[h] Lorena López-Cerero,[i] Álvaro Pascual,[i,j] Jesús Rodríguez-Baño,[i,k] the Spanish Group for Nosocomial Infections (GEIH)

Laboratorio de Referencia de *E. coli* (LREC), Departamento de Microbioloxía e Parasitoloxía, Facultade de Veterinaria, Universidade de Santiago de Compostela (USC), Lugo, Spain[a]; Departamento de Biología Molecular, Universidad de Cantabria, Santander, Spain[b]; Instituto de Biomedicina y Biotecnología de Cantabria, Santander, Spain[c]; Servicio de Microbiología, Hospital Universitario Marqués de Valdecilla, IFIMAV, Santander, Spain[d]; Unidade de Microbioloxía Clínica, Hospital Universitario Lucus Augusti, Lugo, Spain[e]; Service de Microbiologie, Hôpital AP-HP Beaujon, Clichy, France[f]; Faculté de Médecine D. Diderot, Université Paris 7, Paris, France[g]; Veterans Affairs Medical Center and University of Minnesota, Minneapolis, Minnesota, USA[h]; Unidad Clínica de Enfermedades Infecciosas y Microbiología, Hospital Universitario Virgen Macarena, Seville, Spain[i]; Departamento de Microbiología, Facultad de Medicina, Seville, Spain[j]; Departamento de Medicina, Universidad de Sevilla, Seville, Spain[k]

A total of 1,021 extended-spectrum-β-lactamase-producing *Escherichia coli* (ESBLEC) isolates obtained in 2006 during a Spanish national survey conducted in 44 hospitals were analyzed for the presence of the O25b:H4-B2-ST131 (sequence type 131) clonal group. Overall, 195 (19%) O25b-ST131 isolates were detected, with prevalence rates ranging from 0% to 52% per hospital. Molecular characterization of 130 representative O25b-ST131 isolates showed that 96 (74%) were positive for CTX-M-15, 15 (12%) for CTX-M-14, 9 (7%) for SHV-12, 6 (5%) for CTX-M-9, 5 (4%) for CTX-M-32, and 1 (0.7%) each for CTX-M-3 and the new ESBL enzyme CTX-M-103. The 130 O25b-ST131 isolates exhibited relatively high virulence scores (mean, 14.4 virulence genes). Although the virulence profiles of the O25b-ST131 isolates were fairly homogeneous, they could be classified into four main virotypes based on the presence or absence of four distinctive virulence genes: virotypes A (22%) (*afa* FM955459 positive, *iroN* negative, *ibeA* negative, *sat* positive or negative), B (31%) (*afa* FM955459 negative, *iroN* positive, *ibeA* negative, *sat* positive or negative), C (32%) (*afa* FM955459 negative, *iroN* negative, *ibeA* negative, sat positive), and D (13%) (*afa* FM955459 negative, *iroN* positive or negative, *ibeA* positive, *sat* positive or negative). The four virotypes were also identified in other countries, with virotype C being overrepresented internationally. Correspondingly, an analysis of XbaI macrorestriction profiles revealed four major clusters, which were largely virotype specific. Certain epidemiological and clinical features corresponded with the virotype. Statistically significant virotype-specific associations included, for virotype B, older age and a lower frequency of infection (versus colonization), for virotype C, a higher frequency of infection, and for virotype D, younger age and community-acquired infections. In isolates of the O25b:H4-B2-ST131 clonal group, these findings uniquely define four main virotypes, which are internationally distributed, correspond with pulsed-field gel electrophoresis (PFGE) profiles, and exhibit distinctive clinical-epidemiological associations.

In recent years, extended-spectrum β-lactamase (ESBL) production in *Enterobacteriaceae*, particularly *Escherichia coli*, has significantly increased in many countries, including Spain (1). In 2000, the first nationwide study of ESBL-producing *E. coli* (ESBLEC) in Spain was developed (GEIH-BLEE-2000) (2). The overall prevalence of ESBL production was 0.5%, with CTX-M-9, SHV-12, and CTX-M-14 predominating. No CTX-M-15-producing *E. coli* isolates were detected. In contrast, in 2006, a similarly designed nationwide study (GEIH-BLEE-2006) showed that in just 6 years, the prevalence of ESBLEC in Spain had increased 8-fold, to 4% (3, 4). In the 2006 study, it was found that CTX-M-15 had joined CTX-M-14, SHV-12, and CTX-M-9 as a prevalent ESBL type. Thus, the predominant ESBL type had changed quickly, primarily due to the introduction and dissemination of a single clonal group characterized by CTX-M-15, serotype O25b:H4, phylogenetic group B2, and sequence type 131 (ST131), i.e., the international O25b:H4-B2-ST131 clonal group (5–9).

Unlike most other antimicrobial-resistant *E. coli* isolates, most of which derive from non-B2 phylogenetic groups (i.e., A, B1, and D) but are similar to other group B2 clonal groups, O25b:H4-B2-ST131 *E. coli* isolates typically exhibit multiple virulence factors, including adhesins, toxins, siderophores, and group 2 capsules. Thus, this clonal group combines both resistance and virulence genes, which in classical extraintestinal pathogenic *E. coli* (ExPEC) isolates have been mutually exclusive (6, 10–14).

In the report of the GEIH-BLEE-2006 study isolates (4), only the 37 CTX-M-15-positive isolates were screened for O25b-ST131 status, and all but five corresponded to this clonal group; the vir-

Received 14 June 2013 Returned for modification 24 July 2013
Accepted 29 July 2013
Published ahead of print 7 August 2013
Address correspondence to Jorge Blanco, jorge.blanco@usc.es.
J.B., A.M., and R.M. contributed equally to this work.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JCM.01555-13.

doi:10.1128/JCM.01555-13

ulence genotypes of the O25b-ST131 isolates were not assessed. Here, we screened all 1,021 ESBLEC isolates from the GEIH-BLEE-2006 study for O25b-ST131 status and investigated a subset of the detected O25b-ST131 isolates for virulence genotypes and pulsed-field gel electrophoresis (PFGE) profiles. The main objective was to characterize the virulence profile diversity of the ESBL-producing *E. coli* human O25b-ST131 isolates from Spain, thereby identifying prominent virotypes. Secondarily, we screened for these virotypes among ST131 isolates from other countries and sought to find associations between individual virotypes and epidemiological and clinical features.

MATERIALS AND METHODS

Bacterial isolates. Forty-four hospitals representing all Spanish regions participated in the GEIH-BLEE-2006 project. During the study period (1 February to 30 March 2006), 1,021 ESBL isolates were obtained from clinical samples (4). Species identification was performed with the API 20E system (bioMérieux, Marcy l'Etoile, France). ESBL production was confirmed by broth microdilution according to the CLSI guidelines (15).

For comparison, 52 international O25b:H4-B2-ST131 isolates (50 of human and 2 of avian origin), representing eight countries and three continents and taken from the reference collections of Nicolas-Chanoine et al. (6) and Johnson et al. (16), were analyzed to detect globally spread clonal variants in Spain. These isolates included the most prevalent ST131-associated XbaI pulsotypes (800, 812, 905, and 968) and others (699, 788, 797, 806, 903, 904, 909, 910, 911, 913, 915, 916, 917, 919, 979, 981, 1160, 1201, and 1202), as previously described by Johnson et al. (16).

Identification of O25b:H4-B2-ST131 isolates: serotyping, phylogenetic grouping, and multilocus sequence typing. The 1,021 Spanish study isolates were screened using a triplex PCR that was previously developed for the presumptive identification of O25b-ST131 isolates (5). Presumptive O25b-ST131 isolates were confirmed as O25:H4 by serotyping, using type-specific O and H antisera. The major *E. coli* phylogenetic group was determined by triplex PCR (17). Multilocus sequencing typing (MLST) relied on a sequence analysis of seven housekeeping genes (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*) according to the protocol and primers that are specified on the *E. coli* MLST website (http://mlst.ucc.ie/mlst/dbs/Ecoli) (18).

Antibiotic susceptibility testing and molecular characterization of resistance mechanisms. MICs were determined by the MicroScan WalkAway automated system (Siemens, Spain), used according to the manufacturer's instructions, and were interpreted as specified by the CLSI (15). ESBL production was screened for by using cephalosporin resistance and the double-disk synergy test. ESBL genotype was determined by PCR using published TEM-, SHV-, CTX-M-1-, and CTX-M-9-group-specific primers (19). The resistance genes bla_{OXA-1} and *aac(6′)-Ib-cr* were detected by PCR screening and bidirectional sequencing of amplicons (10). The genetic environment of $bla_{CTX-M-15}$ was investigated by a specific PCR for upstream insertion elements (IS*Ecp1* and IS*26*) in isolates of clone O25b-ST131 (20).

Virulence genotyping. The detection of 40 ExPEC-associated virulence genes was done by multiplex PCR (21, 22). The genes we sought to detect included *fimH*, $fimAv_{MT78}$, F10 *papA*, *papA*, *papC*, *papEF* (positive isolates were tested for the *papG I*, *papG II*, and *papG III* alleles), *sfa/focDE* (positive isolates were tested for *sfaS* and *focG*), *afa/draBC*, *afa* operon FM955459, *iha*, *bmaE*, *gafD*, *sat*, *cdtB*, *cnf1*, *hlyA*, *iucD*, *iutA*, *iroN*, *fyuA*, *chuA*, *kpsM II* (with its *neuC-K1*, *K2*, and *K5* variants), *kpsMT III*, *cvaC*, *iss*, *traT*, *ibeA*, *malX*, *usp*, *tsh*, and *ompT* (Table 1). Isolates were classified as ExPEC if they carried ≥2 of *papEF* (P fimbriae), *sfa/focDE* (S/F1C fimbriae), *afa/draBC* (Afa/Dr adhesins), *iutA* (aerobactin receptor), and *kpsM II* (group 2 capsule synthesis) (23). The virulence score was the number of virulence genes that were detected in an isolate. PCR screening of the *fimB* insertion sequence (IS*3*-like transposase) was performed as described elsewhere (24). Investigation of the *fim* operon sequence in the *E. coli* EC958 ST131 strain revealed a 1,895-bp insertion in the *fimB* gene, which encodes the FimB recombinase that switches on the expression of type 1 fimbriae (24).

***fimH* subtyping.** The $fimH_{TR}$ allele was identified based on sequence variation in the *E. coli* type-1 fimbrial adhesion gene (positions 64 to 552). Amplification and sequencing were performed as previously described (25).

Pulsed-field gel electrophoresis. XbaI PFGE analysis was performed as previously described (21). Profiles were analyzed with the BioNumerics fingerprinting software (Applied Maths, St-Martens-Latem, Belgium). Cluster analysis of the Dice similarity indices based on the unweighted-pair group method using average linkages (UPGMA) was done to generate a dendrogram describing the relationship among the PFGE profiles.

Epidemiological and clinical features. Epidemiological and clinical features were collected by using a structured questionnaire based on the following data: age, sex, health care environment, underlying conditions, invasive procedures performed during the preceding week, antimicrobial use during the preceding month, whether the isolate represented colonization or infection (and, if infection, the type of infection), and outcome. Isolates were classified as nosocomially acquired (NA) if obtained 48 h after hospital admission, as health care-associated (HCA) if the patient had been admitted to an acute or long-term care center or had received hemodialysis, specialized home care, or day hospital care during the preceding 3 months, and as community-acquired (CA) if none of these applied. The ethics committee of each participating center approved the study.

Statistical analysis. Comparisons of proportions and scores were tested using Fisher's exact test, chi-square test, and the Mann-Whitney U test, as appropriate. The criterion for statistical significance was set at a P value of <0.05.

Nucleotide sequence accession number. The sequence for the CTX-M-103 gene was deposited in EMBL under accession number HG423149.

RESULTS

Prevalence and distribution of the O25b:H4-B2-ST131 clonal group. According to PCR-based detection, the *E. coli* O25b-ST131 clonal group accounted for 195 (19%) of the 1,021 ESBLEC isolates from the Spanish GEIH-BLEE-2006 project. The O25b-ST131 isolates were widely distributed across Spain and were recovered from 30 of the 44 participating hospitals, including in 13 of the 17 autonomous communities. By hospital, the prevalence of ST131 among the local ESBL study isolates ranged from 0% to 52% (see Table S1 in the supplemental material), with the highest values observed in Madrid and the Canary Islands (Fig. 1).

ESBL enzymes produced by O25b:H4-B2-ST131 isolates. From the 195 total O25b-ST131 isolates, 130 (a maximum of 10 per hospital) were selected randomly for further characterization. Of these, 96 (74%) were positive for CTX-M-15, 15 (12%) for CTX-M-14, 9 (7%) for SHV-12, 6 (5%) for CTX-M-9, and 5 (4%) for CTX-M-32, with the remaining two isolates being positive, respectively, for CTX-M-3 and a new ESBL enzyme, CTX-M-103, which was first detected in this study. CTX-M-103 (EMBL accession no. HG423149) has a single amino acid change compared to CTX-M-15 and CTX-M-3, and it may be grouped in cluster CTX-M-1 (26). Three O25b-ST131 isolates from three different hospitals in the Galicia autonomous community exhibited both CTX-M-14 and CTX-M-15 (Fig. 1; see also Table S1 in the supplemental material).

Virulence genes and virotypes of O25b:H4-B2-ST131 isolates. Of the 40 studied ExPEC-associated virulence genes, 13 were detected by PCR in a majority of the 130 selected O25b-ST131 isolates (Table 1). These 13 genes included *fimH* (in 99%), F10 *papA* (77%), *iha* (75%), *sat* (76%), *iucD* (93%), *iutA* (94%), *fyuA*

TABLE 1 Comparison of virulence gene prevalences of 130 representatives of O25b:H4-B2-ST131 ESBLEC isolates in relation to their virotypes

Virulence gene	Isolates in virotype:[a] A	B	C	D	O25b-ST131 isolates	P^b
Total no.	29	40	41	17	130	
Adhesins (no. [%])						
fimH	29 (100)	40 (100)	40 (98)	17 (100)	129 (99)	
IS*L3*-like in *fimB*	**29 (100)**	**40 (100)**	**41 (100)**	0	110 (85)	<0.001
fimAv$_{MT78}$	0	0	0	1 (6)	1 (1)	
F10 *papA*	**27 (93)**	**31 (78)**	**41 (100)**	0	100 (77)	<0.001
papA	0	0	0	**12 (71)**	12 (9)	<0.001
papC	0	0	2 (5)	**9 (53)**	11 (8)	<0.001
papEF	0	0	0	**12 (71)**	12 (9)	<0.001
papG I	0	0	0	0	0	
papG II	0	0	0	0	0	
papG III	0	0	0	**12 (71)**	12 (9)	<0.001
sfa/focDE	0	0	0	0	0	
sfaS	0	0	0	0	0	
focG	0	0	0	0	0	
afa/draBC	**29 (100)**	0	1 (2)	0	30 (23)	<0.001
afa FM955459	**29 (100)**	0	0	0	29 (22)	<0.001
iha	**27 (93)**	**29 (73)**	**41 (100)**	0	98 (75)	<0.001
bmaE	0	0	0	0	0	
gafD	0	0	0	0	0	
Toxins (no. [%])						
sat	**28 (97)**	**30 (75)**	**41 (100)**	0	99 (76)	<0.001
cdtB	0	0	0	**5 (29)**	5 (4)	<0.001
cnf1	0	0	0	**10 (59)**	10 (8)	<0.001
hlyA	0	0	0	**10 (59)**	10 (8)	<0.001
Siderophores (no. [%])						
iucD	27 (93)	39 (98)	41 (100)	13 (76)	121 (93)	
iutA	28 (97)	40 (100)	41 (100)	13 (76)	122 (94)	
iroN	0	**40 (100)**	0	13 (76)	53 (41)	<0.001
fyuA	29 (100)	40 (100)	41 (100)	17 (100)	130 (100)	
chuA	29 (100)	40 (100)	41 (100)	17 (100)	130 (100)	
Capsula (no. [%])						
kpsM II	29 (100)	31 (78)	28 (68)	17 (100)	106 (82)	
kpsM II-K2	**29 (100)**	0	0	0	29 (22)	<0.001
kpsM II-K5	0	**31 (78)**	**28 (68)**	**15 (88)**	75 (58)	<0.001
neuC-K1	0	0	0	2 (12)	2 (2)	
kpsM III	0	0	0	0	0	
Miscellaneous (no. [%])						
cvaC	0	1 (3)	0	**9 (53)**	10 (1)	<0.001
iss	0	**39 (98)**	0	**16 (94)**	55 (42)	<0.001
traT	21 (72)	21 (53)	36 (88)	16 (94)	96 (74)	
ibeA	0	0	0	**17 (100)**	17 (13)	<0.001
malX (PAI)	29 (100)	40 (100)	41 (100)	17 (100)	130 (100)	
usp	29 (100)	40 (100)	41 (100)	17 (100)	130 (100)	
tsh	0	0	0	0	0	
ompT	29 (100)	40 (100)	41 (100)	17 (100)	130 (100)	
ExPEC status (no. [%])	29 (100)	31 (78)	28 (68)	17 (100)	106 (82)	
Range of virulence genes (no.)	10–15	11–17	12–14	9–20	9–20	NA
Mean no. of virulence genes	14.4	14.5	13.6	16.8	14.4	NA

[a] Significant differences are indicated in bold.
[b] *P* values (by Fisher's exact test) are shown where there was a *P* value of <0.05. NA, not applicable.

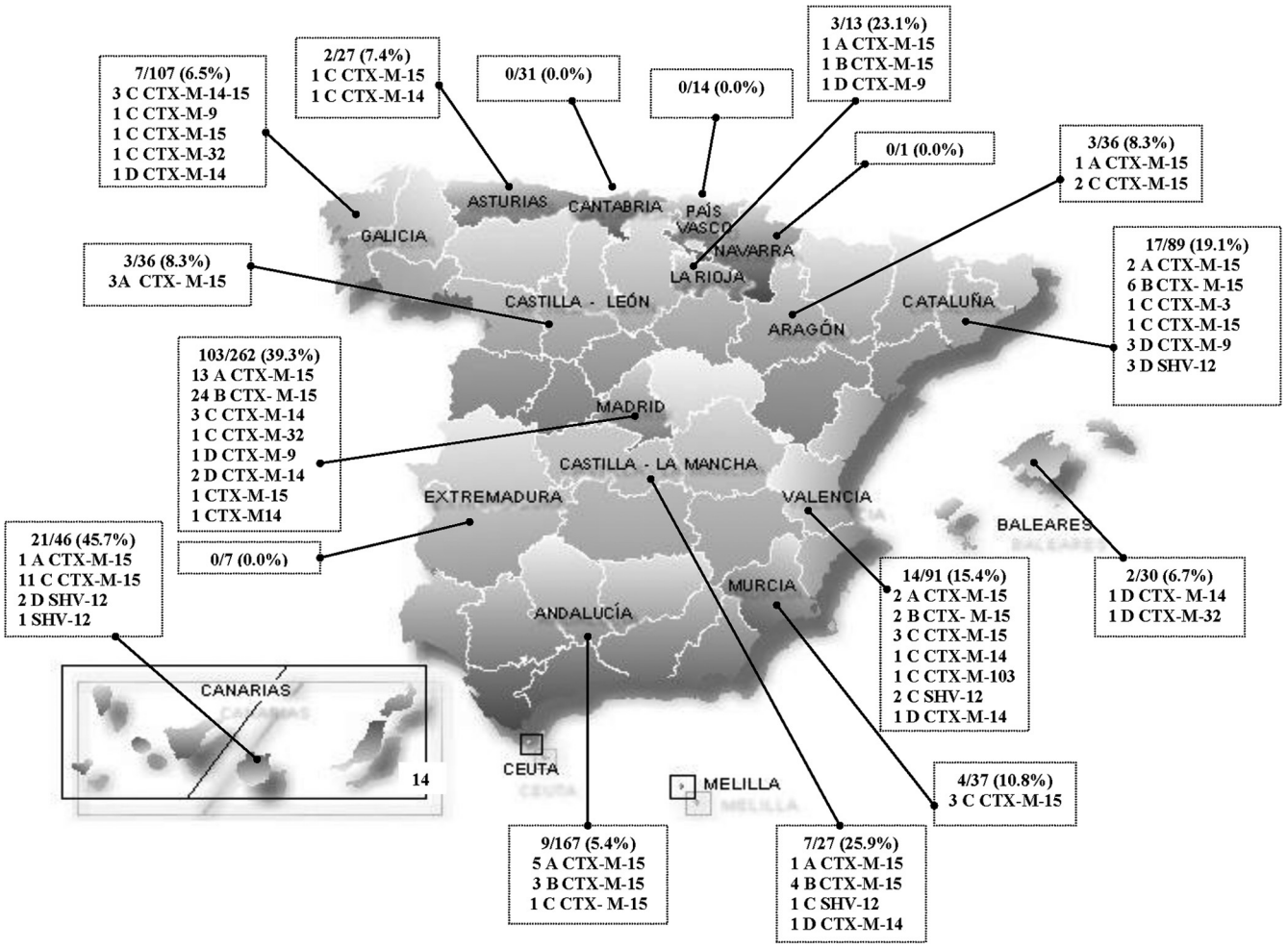


FIG 1 Distribution of O25b-ST131 isolates in Spain. The O25b-ST131 isolates were widely distributed across Spain, being recovered from 30 of the 44 participant hospitals, including 13 of the 17 autonomous communities of Spain.

(100%), *chuA* (100%), *kpsM II* (82%), *traT* (74%), *malX* (100%), *usp* (100%), and *ompT* (100%). In contrast, nine virulence genes were not detected in any O25b-ST131 isolate; these included *papG I*, *papG II*, *sfa/focDE*, *sfaS*, *focG*, *bmaE*, *gafD*, *kpsM III*, and *tsh*. Overall, the 130 O25b-ST131 isolates exhibited relatively high aggregate virulence scores (mean, 14.4; range, 9 to 20), and 106 (82%) satisfied the molecular criteria for ExPEC.

Although the virulence profiles were fairly homogeneous overall, four discrete virotypes (labeled arbitrarily as A, B, C, and D) were resolved based on the presence or absence of four distinctive virulence genes, including *afa* FM955459 (specific for an ST131 clone encoding an Afa/Dr adhesin), *iroN* (catecholate siderophore receptor), *ibeA* (invasion of brain endothelium), and *sat* (secreted autotransporter toxin). The patterns were as follows: virotype A (*afa* FM955459 positive, *iroN* negative, *ibeA* negative, *sat* positive or negative), virotype B (*afa* FM955459 negative, *iroN* positive, *ibeA* negative, *sat* positive or negative), virotype C (*afa* FM955459 negative, *iroN* negative, *ibeA* negative, *sat* positive), and virotype D (*afa* FM955459 negative, *iroN* positive or negative, *ibeA* positive, *sat* positive or negative). The 130 isolates were distributed among the four virotypes as follows: A, 29 (22%), B, 40 (31%), C, 41 (32%), and D, 17 (13%), with three isolates remaining unclassified as to virotype since they lacked all four virotype-defining genes.

Interestingly, the 17 virotype D isolates exhibited significantly higher virulence scores than did those of other virotypes (mean, 16.8 for D versus 14.4 for A, 14.5 for B, and 13.6 for C) (Table 1).

All four virotypes identified for the Spanish O25b:H4-B2-ST131 isolates were also represented in 51 of the 52 studied international O25b:H4-B2-ST131 isolates. Virotype C was by far the most prevalent (65%), followed by virotypes A (17%), B (10%), and D (6%).

Insertion of IS*L3*-like transposase gene in *fimB* in relation to virotype. The 130 selected Spanish O25b-ST131 isolates were also analyzed for an insertion in *fimB*, similar to the IS*L3* transposase gene observed in *E. coli* strain Nissle 1917 (GenBank accession no. AF188737). The *fimB* insertion was detected in all 113 isolates of virotypes A, B, and C but in none of the 17 virotype D isolates ($P < 0.001$ for all comparisons versus virotype D) (Table 1).

***fimH* alleles and O25b:H4-B2-ST131 virotypes.** *fimH* sequence analysis was done for 40 of the Spanish O25b-ST131 isolates, including 10 each for the four virotypes, which were selected to include a representation of all virulence gene profiles and PFGE groups encountered within a particular virotype. Two different *fimH* alleles were identified among these 40 isolates. All representatives of virotypes A, B, and C contained *fimH30*, whereas all

TABLE 2 ESBL enzymes, genes associated with CTX-M-15 plasmids, and antimicrobial resistance of 130 representatives of O25b:H4-B2-ST131 ESBLEC isolates in relation to their virotypes

Characteristic	No. (%) of isolates by virotype[a]:				No. (%) of O25b-ST131 isolates
	A	B	C	D	
Total no.	29	40	41	17	130
ESBL enzymes					
CTX-M-3	0	0	1 (2)	0	1 (0.8)
CTX-M-9	0	0	1 (2)	**5 (29)**	6 (5)
CTX-M-14	0	0	**8 (20)**	**6 (35)**	15 (12)
CTX-M-15	**29 (100)**	**40 (100)**	**26 (63)**	0	96 (74)
CTX-M-32	0	0	4 (10)	1 (6)	5 (4)
CTX-M-103	0	0	1 (2)	0	1(0.8)
SHV-12	0	0	3 (7)	**5 (29)**	9 (7)
Genes associated with CTX-M-15					
bla_{OXA-1}	**20 (69)**	**35 (88)**	**21 (51)**	0	85 (65)
aac(6′)-Ib-cr	**29 (100)**	**35 (88)**	**21 (51)**	0	85 (65)
IS*26* element	**29 (100)**	0	0	0	30 (23)
Antimicrobial resistance					
Ciprofloxacin	**29 (100)**	**40 (100)**	**40 (98)**	5 (29)	117 (90)
Gentamicin	0	**27 (68)**	**16 (39)**	1 (6)	46(35)
Tobramycin	**29 (100)**	**33 (83)**	**28 (68)**	3 (18)	95 (73)
Trimethoprim-sulfamethoxazole	**29 (100)**	31 (78)	20 (50)	10 (59)	92 (71)
Amoxicillin-clavulanic acid	**10 (34)**	6 (15)	9 (22)	0	26 (20)
Fosfomycin	**5 (17)**	3 (8)	2 (5)	0	11 (9)

[a] Significant differences are indicated in bold. *P* values are shown in Results.

representatives of virotype D contained *fimH22* ($P < 0.001$ for all comparisons versus virotype D).

Virotype versus ESBL variants, CTX-M-15-associated genes, and antimicrobial resistance. Among the 130 selected O25b:H4-B2-ST131 isolates, all 69 isolates from virotypes A and B and 63% (26/41) of the virotype C isolates produced CTX-M-15; however, no virotype D isolates produced CTX-M-15 ($P < 0.001$ for all comparisons versus virotype D). In contrast, virotype D isolates produced CTX-M-9 (5 isolates), CTX-M-14 (6 isolates), SHV-12 (5 isolates), and CTX-M-32 (1 isolate) (Table 2).

As expected, the bla_{OXA-1} and *aac(6′)-Ib-cr* genes were associated with CTX-M-15-producing isolates. These genes were detected in 69% and 100% of virotype A isolates, respectively, 88% (both genes) of virotype B isolates, and 51% (both genes) of virotype C isolates but in 0% of virotype D isolates ($P < 0.001$ for all comparisons versus virotype D) (Table 2).

Only the 29 virotype A isolates showed an IS*26* element within the terminal inverted repeat of the IS*Ecp1*-like element upstream of $bla_{CTX-M-15}$, separating the $bla_{CTX-M-15}$ allele from its usual promoter, as found in the epidemic ST131 "strain A" that is prevalent in the United Kingdom (5, 20) ($P < 0.001$ for comparisons with virotypes B, C, and D) (Table 2).

Different patterns of associated antimicrobial resistances were detected in relation to virotype. Specifically, ciprofloxacin and tobramycin resistance were significantly associated with virotypes A, B, and C ($P < 0.001$ for all comparisons versus virotype D). In contrast, virotypes B and C were significantly associated with gentamicin resistance ($P < 0.05$ for all comparisons versus virotypes A and D), and virotype A was significantly associated with trimethoprim-sulfamethoxazole resistance ($P < 0.05$ for all comparisons versus virotypes B, C, and D) (Table 2).

PFGE profiles of O25b:H4-B2-ST131 isolates in relation to virotype. In the PFGE-based dendrogram, 116 of the 130 Spanish O25b-ST131 isolates were distributed among 4 large virotype-specific clusters, defined at similarity levels of approximately 72% (11 virotype D isolates), 76% (28 virotype A isolates), 77% (33 virotype C isolates), and 74% (39 virotype B isolates) (see Fig. S1 in the supplemental material).

A similar clustering of PFGE profiles by virotype was detected when the 130 Spanish study isolates were compared with the 52 international isolates (see Fig. S2 in the supplemental material). Four main clusters, each largely virotype specific, comprised 162 of the 182 total isolates. Each of these clusters, defined at similarity levels of approximately 80% (47 isolates, 40 of virotype B), 78% (60 isolates, 58 of virotype C), 82% (37 isolates, 36 of virotype A), and 67% (18 isolates, 16 of virotype D), included isolates from multiple countries.

In the present study, among 13 international isolates representing the top four pulsotypes described by Johnson et al. (16), we found that virotype A corresponded with pulsotype 812, virotype B with pulsotype 905, and virotype C with all four pulsotypes (968, 800, 905, and 812), whereas none of the virotype D isolates belonged to these top four pulsotypes.

Epidemiological and clinical associations of ESBLEC O25b-ST131. Associated clinical data were compared for O25b-ST131 and non-O25b-ST131 Spanish ESBLEC isolates (Table 3). In univariate analyses, compared with the non-O25b-ST131 isolates, the O25b-ST131 isolates were significantly associated with older age, nursing home residents, a health care-related origin, asymptomatic bacteriuria, and bacteremia. However, after adjustment for acquisition type and age by logistic regression analysis, the association of O25b-ST131 isolates with bacteremia was not statistically

TABLE 3 Epidemiological and clinical data of patients with colonization or infection due to ESBLEC O25b:H4-B2-ST131 compared to those with non-ST131 infection

Patient variable	Value for isolate[a]: ST131	Non-ST131	*P*[e]
Total no.	190	818	
Male gender (no. [%])	70 (37)	328 (40)	
Median age (IQR) (yr)	**75 (68–84)**	69 (50–79)	0.01[b]
No. (%) of pediatric patients (age ≤ 14 yr)	4 (2)	**53 (6)**	0.01[c]
Acquisition type (no. [%])			
Nosocomial	66 (35)	238 (29)	
Health care related	**47 (25)**	115 (14)	<0.001
Community	77 (41)	**465 (57)**	<0.001
Nursing home resident	**37 (19)**	39 (5)	<0.001
Infection site (no. [%])[d]	136 (72)	**651 (80)**	0.02
Urinary tract	102 (75)	485 (75)	
Soft tissue	16 (12)	71 (11)	
Digestive tract	6 (4)	37 (6)	
Other types	20 (15)	116 (18)	
Bacteremia (primary or secondary)	**22 (16)**	56 (9)	0.01
Crude 30-day mortality (no. [%])[d]	5 (4)	12 (2)	

[a] Data from 5 patients with ST131 and from 8 patients with non-ST131 isolates were unavailable. *P* values were calculated by chi-square test except where otherwise noted.
[b] Calculated by Mann-Whitney U test.
[c] Calculated by Fisher's exact test.
[d] Only patients with infection were considered.
[e] *P* values are shown where $P < 0.05$.

TABLE 4 Association of virotype with epidemiological and clinical data among 130 O25b:H4-B2-ST131 ESBLEC isolates

Patient variable	Isolates by virotype[a]: A	B	C	D
Total no.	29	40	41	17
Male gender (no. [%])	10 (35)	16 (40)	22 (54)	7 (41)
Age ≤ 14 yr (no. [%])	1 (3)	0	1 (2)	2 (12)
Median age (IQR) (yr)	77 (63–82)	75 (64–83)[a]	64 (56–79)	60 (34–78)[b]
Acquisition type (no. [%])				
Nosocomial	9 (31)	18 (45)	18 (44)	4 (24)
Health care related	9 (31)	9 (23)	7 (17)	1 (6)
Community	11 (38)	13 (32)	16 (39)	12 (71)[c]
Nursing home resident	7 (24)	10 (25)	3 (7)	1 (6)
Infection type	22 (85)	21 (57)[d]	36 (92)[e]	13 (77)
Urinary tract	20 (69)	13 (33)[f]	25 (61)	13 (77)
Respiratory tract	0	4 (10)[g]	1 (2)	0
Digestive tract	0	0	2 (5)	0
Primary bacteremia	0	2 (5)	3 (7)	0
Other types	2 (7)	2 (5)	5 (12)	0
Bacteremia (primary or secondary)	5 (17)	8 (20)	8 (15)	0
Crude 30-day mortality (no. [%])	0	6 (15)	4 (10)	1 (6)
Infection-related deaths (no. [%])	0	3 (8)	1 (2)	0

[a] $P = 0.05$ (calculated by Mann-Whitney U test). All comparisons other than those indicated by a footnote have a *P* value of >0.05.
[b] $P = 0.03$ (calculated by Mann-Whitney U test).
[c] $P = 0.03$ (calculated by Fisher's exact test).
[d] $P = 0.009$.
[e] $P = 0.01$.
[f] $P = 0.0001$.
[g] $P = 0.03$ (calculated by Fisher's exact test).

significant (odds ratio [OR], 1.58; 95% confidence interval [CI], 0.91 to 2.75) ($P = 0.10$).

Associations of the four virotypes with demographic data, acquisition type, and type of infection also were explored (Table 4). Virotype B was significantly associated with older patients and a lower likelihood of symptomatic infection, specifically urinary tract infection, but a higher likelihood of respiratory tract infection. In contrast, virotype C was significantly associated with a higher likelihood of symptomatic infection, whereas virotype D was significantly associated with younger patients and community acquisition. Additionally, virotypes A and B were more highly associated with nursing home residents (17/69 [25%]) than were virotypes C and D (4/61 [7%]) ($P = 0.005$).

Triplex PCR for virotyping. A triplex PCR based on the detection of *sat*, *iroN*, and *ibeA* was designed for virotyping of the O25b:H4-B2-ST131 isolates. When applied to the present 130 Spanish and 52 international O25b-ST131 isolates, this assay exhibited 100% specificity and sensitivity (not shown). This assay can be combined with a previously described triplex PCR (based on detection of the *afa* operon FM955459, the O25b *rfb* allele, and the 3′ end of $bla_{CTX\text{-}M\text{-}15}$) to establish the virotypes of the O25b-ST131 isolates (Table 5; see also Fig. S3 in the supplemental material).

DISCUSSION

The prevalence and epidemiology of ESBLEC are changing rapidly. In recent years, ESBL production in *E. coli* has increased significantly, due primarily to the spread of CTX-M types. In fact, the prevalence of ESBLEC strains in Spain increased 8-fold between 2000 and 2006, from 0.5% to 4% (4). The emergence and dissemination of ESBLEC have two possible explanations, that it occurs through dissemination of mobile genetic elements between non-clonally related strains or through clonal spread. The two mechanisms might occur simultaneously, thereby contributing to the rapid dissemination of ESBLEC strains. Until a few years ago, most ESBLEC strains were clonally unrelated; however, recently, the O25b:H4-B2-ST131 intercontinental *E. coli* clonal group that produces CTX-M-15 with high virulence potential has been reported worldwide, representing a major public health problem (6, 10).

In the present study, 195 (19%) of the 1,021 total ESBLEC study isolates were of the O25b-ST131 group. The O25b-ST131 isolates were widely distributed across Spain, as they were at 30 of the 44 participating centers and accounted for up to 52% of ESBLEC strains per hospital. This contrasts with the 9% prevalence of ST131 among the 92 analyzed ESBLEC isolates from a similar study from 2004 done in 11 Spanish hospitals (7 of which were included in the present study) (8). More recent studies found even higher prevalences of O25b-ST131 among ESBLECs in various Spanish cities, including Lugo in 2008 (23%) (5) and in 2012 (61%) (27), Barcelona (32% in 2008) (28), Seville (13% in 2010) (29), and Madrid (21% in 2008) (30). Other countries have also had high reported prevalences of ST131 among the ESBLEC strains, including Denmark (38%) (31), Japan (41%) (32), the United States (47%) (33), and Canada (78%) (34).

The Spanish O25b-ST131 isolates in this study carried not only CTX-M-15 but also CTX-M-14, SHV-12, CTX-M-9, CTX-M-32,

TABLE 5 Primers included in the two triplex PCR assays used for specific identification of clonal group O25b:H4-B2-ST131 and virotyping

Gene	Primer	Oligonucleotide sequence (5′ to 3′)	Fragment size (bp)	Annealing temp (°C)	Reference
afa FM955459[a]	AFA-O25F	GAGTCACGGCAGTCGCGGCGG	207	55	5
	AFA-O25R	TTCACCGGCGACCAGCCATCTCC			
rfbO25b[a]	rfb.1bis	ATACCGACGACGCCGATCTG	300	55	45
	rfbO25b.r	TGCTATTCATTATGCGCAGC			
$bla_{\text{CTX-M-15}}$[a]	CTX-M-F1	ATAAAACCGGCAGCGGTG	483	55	19
	CTX-M-F2	GAATTTTGACGATCGGGG			
iroN[b]	IRON-F	AAGTCAAAGCAGGGGTTGCCCG	667	60	46
	IRON-R	GACGCCGACATTAAGACGCAG			
sat[b]	SAT-F	ACTGGCGGACTCATGCTGT	387	60	47
	SAT-R	AACCCTGTAAGAAGACTGAGC			
ibeA[b]	IBEA 10F	AGGCAGGTGTGCGCCGCGTAC	170	60	22
	IBEA 10R	TGGTGCTCCGGCAAACCATGC			

[a] Triplex PCR used for specific identification of O25b-ST131 isolates (5).
[b] Triplex PCR used for virotyping of O25b-ST131 isolates (developed in this study).

CTX-M-3, and the new ESBL enzyme CTX-M-103 first described here. Although in most countries, ST131 is associated mainly with CTX-M-15 (6, 33–35), exceptions do occur, such as in Japan, where O25b-ST131 is frequently associated with CTX-M-14 (36), or in Ireland, where it is associated with CTX-M-3 (37). In the present study, we showed that association between *E. coli* O25b-ST131 and CTX-M types other than CTX-M-15 was mainly identified in the isolates with virotype D.

ExPEC isolates have specialized virulence factors enabling them to colonize host surfaces, injure host tissues, and avoid host defense systems. Thirteen genes (*fimH*, F10 *papA*, *iha*, *sat*, *iucD*, *iutA*, *fyuA*, *chuA*, *kpsM II*, *traT*, *malX*, *usp*, and *ompT*) were detected in most of the 130 Spanish O25b-ST131 ESBLEC isolates in this study. In two recent studies, Coelho et al. (28) in Spain and Johnson et al. (38) in the United States found that these virulence genes were significantly more prevalent among ST131 than non-ST131 isolates. Therefore, they might be important in the worldwide dissemination of *E. coli* O25b-ST131. Moreover, the present 130 Spanish O25b-ST131 isolates exhibited high virulence scores, and most qualified molecularly as being ExPEC (23).

After analyzing all the virulence gene profiles together with the PFGE pulsotypes, we observed that the virotypes we established corresponded well with the PFGE clusters. Although ST131 isolates share a large set of putative virulence factors, we selected the four virulence genes (*afa* FM955459, *iroN*, *ibeA*, and *sat*) that clearly define the four virotypes within the clonal group O25b-ST131. Notably, the four virotypes, most prominently virotype C, were also present in other countries (France, Portugal, Switzerland, United States, Canada, Korea, and Lebanon). The XbaI PFGE dendrogram revealed four major clusters that corresponded closely with the virotypes, suggesting a clonal basis for the virotypes. Isolates of virotype D exhibited significantly higher virulence scores, which might explain the association of this virotype with younger patients and community-acquired infections.

Like Banerjee et al. (39), we found that patients with O25b-ST131 isolates were older and more frequently had health care-associated acquisition (particularly through nursing home residency) than those with non-O25b-ST131 ESBLEC isolates. Although the O25b-ST131 isolates were associated with bacteremia in univariate analyses, this association was not significant when age and acquisition type were considered in multivariate analysis. Chung et al. (40) found that patients with clone ST131 were more likely to have secondary bacteremia than those with non-ST131 isolates. As has been found in other studies (39–41), we did not find that O25b-ST131 isolates were associated with increased mortality. Further studies are needed to more fully assess the clinical implications of the comparatively high virulence scores and the group B2 background of the O25b-ST131 ESBLEC isolates. Of note, virotype D was linked with community-acquired infections and virotypes A and B with nursing home residents.

In a recent PFGE analysis of 579 ST131 isolates from diverse years, hosts, and locales, Johnson et al. (16) found that the four most prevalent pulsotypes (among 170 total) accounted for 46% of the population and tended to occur in more recent years, which is consistent with recent emergence and expansion and implies greater fitness of these pulsotypes. In the present study, among 13 international isolates representing the top four pulsotypes of Johnson et al. (16), we found virotypes A, B, and C.

Seven distinct *fimH*-based putative clonal lineages (H15, H22, H27, H30, H35, H41, and H49) were found among 352 historical and recent ST131 isolates obtained primarily from the United States (42). The H22 subclone was dominant (73%) among the historical isolates (those from 1967 to 1999), whereas the H30 subclone was the most prevalent (75%) among the current clinical isolates (those from 2000 to 2011) and accounted for nearly all fluoroquinolone-resistant isolates. Interestingly, in the present study, the H30 subclone accounted for all analyzed Spanish O25b-ST131 isolates of virotypes A, B, and C, whereas the H22 subclone accounted for all analyzed virotype D isolates. Consistent with the association of the H30 ST131 subclone with fluoroquinolone resistance, we found the prevalence of ciprofloxacin resistance to be significantly higher among the (H30-derived) virotype A, B, and C isolates (100%, 100%, and 98%, respectively) than among the (non-H30) virotype D isolates (29%) ($P < 0.001$).

Diverse explanations have been proposed for the rapid and successful dissemination of *E. coli* O25b-ST131 among humans, including possible transmission through animal contact and food consumption (43). Mora et al. (44) reported an association of *E. coli* O25b-ST131 with clinical disease in poultry and identified this clonal group in retail chicken meat in Spain. In that study, the 19 avian O25b-ST131 isolates (of which 7 produced CTX-M-9) were of virotype D, defined according to the criteria specified here. However, in the present study, only 2 of the 17 Spanish human isolates of virotype D (isolates 16.30 and 16.31; *fimH*, *iucD*, *iroN*, *kpsMII-K1*, *iss*, *traT*, *ibeA*, *malX*, *usp*, and SHV-12) and the 2 avian virotype D isolates from the United States (isolates CD285 and CD287; *fimH*, *iucD*, *iroN*, *kpsMII-K1*, *iss*, *traT*, *ibeA*, *malX*, *usp*, and *tsh*) showed virulence gene profiles that were highly similar to those of the Spanish avian isolates (44).

In conclusion, we found a high prevalence of *E. coli* O25b:H4-B2-ST131 among (mostly CTX-M-15-producing) ESBLEC isolates from hospitals across Spain in 2006. We newly defined four main virulence gene profile variants within ST131, which we termed virotypes A to D, and showed these to be prevalent also among isolates from other countries. The four ST131 virotypes, which appeared to represent clonal variants within ST131, exhibited distinctive distributions and clinical and epidemiological associations. Future studies of the conventional and molecular epidemiology of ST131 should address relevant within-ST divisions, including *fimH*-based subclones and virotypes. Such studies should improve our understanding of the basis for the worldwide dissemination and emergence of the *E. coli* ST131 clonal group.

ACKNOWLEDGMENTS

We thank Veronika Tchesnokova and Mariya Billig (University of Washington School of Medicine, Seattle, WA) for their help during the *fimH* subtyping. We thank Monserrat Lamela for skillful technical assistance.

A.M. acknowledges the Ramón y Cajal program from the Spanish Ministerio de Economía y Competitividad, Gobierno de España. R.M. acknowledges the grant of the Agencia Española de Cooperación Internacional (AECI) (Ministerio de Asuntos Exteriores y de Cooperación). This work was partially supported by the Red Española de Investigación en Patología Infecciosa (REIPI) (no. RD06/0008/1018-1016, RD12/0015) and grant no. PI09/01273, 070190, 10/02021, 10/01955, 10/00795, and PI11/01117 (Instituto de Salud Carlos III, Fondo de Investigación Sanitaria, Ministerio de Economía y Competitividad, Gobierno de España), CN2012/303 09TAL007261PR and 10MRU261023PR (Consellería de Cultura, Educación e Ordenación Universitaria, Xunta de Galicia and the European Regional Development Fund [ERDF]), 0048/2008 and CTS-5259 (Junta de Andalucía), BFU2011-26608 (Spanish Ministry of Education), 282004/FP7-HEALTH.2011.2.3.1-2 (European VII Framework Program), and FEDER-INNTERCONECTA-COLIVAC (CDTI, Ministerio de Economía y Competitividad, Gobierno de España; Consellería de Economía e Industria, Xunta de Galicia; ERDF). This material also is based partly upon work supported by the Office of Research and Development, Medical Research Service, Department of Veterans Affairs, grant no. 1 I01 CX000192 01.

The Spanish GEIH-BLEE 2006 study group includes C. Martínez Peinado (Villajoyosa), J. F. Ordás (Cangas de Nancea), E. Garduño (Badajoz), M. A. Domínguez (Barcelona), F. Navarro (Barcelona), G. Prats (Barcelona), F. Marco (Barcelona), E. Ojeda (Burgos), P. Marín (Cádiz), R. Carranza (Alcazar de San Juan), F. Rodríguez (Cordova), C. García Tejero (Figueras), F. Artiles (Gran Canaria), B. Palop (Granada), I. Cuesta (Jaén), M. Cartelle (A Coruña), M. D. Rodríguez (Ferrol), I. Fernández (León), E. Ugalde (Logroño), R. Cantón (Madrid), E. Cercenado (Madrid), F. Chaves (Madrid), J. J. Picazo (Madrid), A. Delgado (Alcorcón), C. Guerrero (Murcia), B. Fernández (Orense), A. Fleites (Oviedo), A. Oliver (Palma de Mallorca), J. J. García (Pamplona), M. García (Pontevedra), J. Elías (Salamanca), J. Calvo (Santander), M. Treviño (Santiago de Compostela), M. Ruiz (Seville), M. A. Díaz and J. R. Hernández-Bello (Seville), M. Lara (Tenerife), L. Torres (Teruel), E. García (Toledo), D. Navarro (Valencia), M. Gobernado (Valencia), A. Tenorio (Valladolid), I. Otero (Vigo), L. Michaus (Vitoria), and J. Castillo (Zaragoza).

REFERENCES

1. **Livermore DM, Canton R, Gniadkowski M, Nordmann P, Rossolini GM, Arlet G, Ayala J, Coque TM, Kern-Zdanowicz I, Luzzaro F, Poirel L, Woodford N.** 2007. CTX-M: changing the face of ESBLs in Europe. J. Antimicrob. Chemother. **59:**165–174.
2. **Hernández JR, Martínez-Martínez L, Cantón R, Coque TM, Pascual A, Spanish Group for Nosocomial Infections (GEIH).** 2005. Nationwide study of *Escherichia coli* and *Klebsiella pneumoniae* producing extended-spectrum beta-lactamases in Spain. Antimicrob. Agents. Chemother. **49:**2122–2125.
3. **Díaz MA, Hernández JR, Martínez-Martínez L, Rodríguez-Baño J, Pascual A, Grupo de Estudio de Infección Hospitalaria (GEIH).** 2009. Extended-spectrum beta-lactamase-producing *Escherichia coli* and *Klebsiella pneumoniae* in Spanish hospitals: 2nd multicenter study (GEIH-BLEE project, 2006). Enferm. Infecc. Microbiol. Clin. **27:**503–510. (Article in Spanish.)
4. **Díaz MA, Hernández-Bello JR, Rodríguez-Baño J, Martínez-Martínez L, Calvo J, Blanco J, Pascual A, Spanish Group for Nosocomial Infections (GEIH).** 2010. Diversity of *Escherichia coli* strains producing extended-spectrum β-lactamases in Spain: second nationwide study. J. Clin. Microbiol. **48:**2840–2845.
5. **Blanco M, Alonso MP, Nicolas-Chanoine MH, Dahbi G, Mora A, Blanco JE, López C, Cortés P, Llagostera M, Leflon-Guibout V, Puentes B, Mamani R, Herrera A, Coira MA, García-Garrote F, Pita JM, Blanco J.** 2009. Molecular epidemiology of *Escherichia coli* producing extended-spectrum β-lactamases in Lugo (Spain): dissemination of clone O25b:H4-ST131 producing CTX-M-15. J. Antimicrob. Chemother. **63:**1135–1141.
6. **Nicolas-Chanoine MH, Blanco J, Leflon-Guibout V, Demarty R, Alonso MP, Caniça MM, Park YJ, Lavigne JP, Pitout J, Johnson JR.** 2008. Intercontinental emergence of *Escherichia coli* clone O25:H4-ST131 producing CTX-M-15. J. Antimicrob. Chemother. **61:**273–281.
7. **Blanco J, Mora A, Mamani R, López C, Blanco M, Dahbi G, Herrera A, Blanco JE, Alonso MP, García-Garrote F, Chaves F, Orellana MÁ, Martínez-Martínez L, Calvo J, Prats G, Larrosa MN, González-López JJ, López-Cerero L, Rodríguez-Baño J, Pascual A.** 2011. National survey of *Escherichia coli* causing extraintestinal infections reveals the spread of drug-resistant clonal groups O25b:H4-B2-ST131, O15:H1-D-ST393 and CGA-D-ST69 with high virulence gene content in Spain. J. Antimicrob. Chemother. **66:**2011–2021.
8. **Oteo J, Diestra K, Juan C, Bautista V, Novais A, Pérez-Vázquez M, Moyá B, Miró E, Coque TM, Oliver A, Cantón R, Navarro F, Campos J, Spanish Network in Infectious Pathology Project (REIPI).** 2009. Extended-spectrum beta-lactamase-producing *Escherichia coli* in Spain belong to a large variety of multilocus sequence typing types, including ST10 complex/A, ST23 complex/A and ST131/B2. Int. J. Antimicrob. Agents **34:**173–176.
9. **Oteo J, Navarro C, Cercenado E, Delgado-Iribarren A, Wilhelmi I, Orden B, García C, Miguelañez S, Pérez-Vázquez M, García-Cobos S, Aracil B, Bautista V, Campos J.** 2006. Spread of *Escherichia coli* strains with high-level cefotaxime and ceftazidime resistance between the community, long-term care facilities, and hospital institutions. J. Clin. Microbiol. **44:**2359–2366.
10. **Coque TM, Novais A, Carattoli A, Poirel L, Pitout J, Peixe L, Baquero F, Cantón R, Nordmann P.** 2008. Dissemination of clonally related *Escherichia coli* strains expressing extended-spectrum beta-lactamase CTX-M-15. Emerg. Infect. Dis. **14:**195–200.
11. **Peirano G, Pitout JD.** 2010. Molecular epidemiology of *Escherichia coli* producing CTX-M beta-lactamases: the worldwide emergence of clone ST131 O25:H4. Int. J. Antimicrob. Agents **35:**316–321.
12. **Rogers BA, Sidjabat HE, Paterson DL.** 2011. *Escherichia coli* O25b-ST131: a pandemic, multiresistant, community-associated strain. J. Antimicrob. Chemother. **66:**1–14.
13. **Horcajada JP, Soto S, Gajewski A, Smithson A, Jiménez de Anta MT, Mensa J, Vila J, Johnson JR.** 2005. Quinolone-resistant uropathogenic

Escherichia coli strains from phylogenetic group B2 have fewer virulence factors than their susceptible counterparts. J. Clin. Microbiol. **43:**2962–2964.
14. **Da Silva GJ, Mendonça N.** 2012. Association between antimicrobial resistance and virulence in *Escherichia coli*. Virulence **3:**18–28.
15. **Clinical and Laboratory Standards Institute.** 2010. Performance standards for antimicrobial susceptibility testing; 20th informational supplement. CLSI M100-S19. Clinical and Laboratory Standards Institute, Wayne, PA.
16. **Johnson JR, Nicolas-Chanoine MH, DebRoy C, Castanheira M, Robicsek A, Hansen G, Weissman S, Urban C, Platell J, Trott D, Zhanel G, Clabots C, Johnston BD, Kuskowski MA, MASTER Investigators.** 2012. Comparison of *Escherichia coli* ST131 pulsotypes, by epidemiologic traits, 1967–2009. Emerg. Infect. Dis. **18:**598–607.
17. **Clermont O, Bonacorsi S, Bingen E.** 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. Appl. Environ. Microbiol. **66:**4555–4558.
18. **Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, Achtman M.** 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol. Microbiol. **60:**1136–1151.
19. **Leflon-Guibout V, Jurand C, Bonacorsi S, Espinasse F, Guelfi MC, Duportail F, Heym B, Bingen E, Nicolas-Chanoine MH.** 2004. Emergence and spread of three clonally related virulent isolates of CTX-M-15-producing *Escherichia coli* with variable resistance to aminoglycosides and tetracycline in a French geriatric hospital. Antimicrob. Agents Chemother. **48:**3736–3742.
20. **Woodford N, Ward ME, Kaufmann ME, Turton J, Fagan EJ, James D, Johnson AP, Pike R, Warner M, Cheasty T, Pearson A, Harry S, Leach JB, Loughrey A, Lowes JA, Warren RE, Livermore DM.** 2004. Community and hospital spread of *Escherichia coli* producing CTX-M extended-spectrum beta-lactamases in the UK. J. Antimicrob. Chemother. **54:**735–743.
21. **Mora A, López C, Dabhi G, Blanco M, Blanco JE, Alonso MP, Herrera A, Mamani R, Bonacorsi S, Moulin-Schouleur M, Blanco J.** 2009. Extraintestinal pathogenic *Escherichia coli* O1:K1:H7/NM from human and avian origin: detection of clonal groups B2 ST95 and D ST59 with different host distribution. BMC Microbiol. **9:**132. doi:10.1186/1471-2180-9-132.
22. **Johnson JR, Stell AL.** 2000. Extended virulence genotypes of *Escherichia coli* strains from patients with urosepsis in relation to phylogeny and host compromise. J. Infect. Dis. **181:**261–272.
23. **Johnson JR, Murray AC, Gajewski A, Sullivan M, Snippes P, Kuskowski MA, Smith KE.** 2003. Isolation and molecular characterization of nalidixic acid-resistant extraintestinal pathogenic *Escherichia coli* from retail chicken products. Antimicrob. Agents Chemother. **47:**2161–2168.
24. **Totsika M, Beatson SA, Sarkar S, Phan MD, Petty NK, Bachmann N, Szubert M, Sidjabat HE, Paterson DL, Upton M, Schembri MA.** 2011. Insights into a multidrug resistant *Escherichia coli* pathogen of the globally disseminated ST131 lineage: genome analysis and virulence mechanisms. PLoS One **6:**e26578. doi:10.1371/journal.pone.0026578.
25. **Weissman SJ, Johnson JR, Tchesnokova V, Billig M, Dykhuizen D, Riddell K, Rogers P, Qin X, Butler-Wu S, Cookson BT, Fang FC, Scholes D, Chattopadhyay S, Sokurenko E.** 2012. High-resolution two-locus clonal typing of extraintestinal pathogenic *Escherichia coli*. Appl. Environ. Microbiol. **78:**1353–1360.
26. **Poirel L, Gniadkowski M, Nordmann P.** 2002. Biochemical analysis of the ceftazidime-hydrolysing extended-spectrum beta-lactamase CTX-M-15 and of its structurally related beta-lactamase CTX-M-3. J. Antimicrob. Chemother. **50:**1031–1034.
27. **Dahbi G, Mora A, López C, Alonso MP, Mamani R, Marzoa J, Coira A, García-Garrote F, Pita JM, Velasco D, Herrera A, Viso S, Blanco JE, Blanco M, Blanco J.** 2013. Emergence of new variants of ST131 clonal goup in extraintestinal pathogenic *Escherichia coli* producing extended-spectrum β-lactamases. Int. J. Antimicrob. Agents. doi:10.1016/j.ijantimicag.2013.06.017.
28. **Coelho A, Mora A, Mamani R, López C, González-López JJ, Larrosa MN, Quintero-Zarate JN, Dahbi G, Herrera A, Blanco JE, Blanco M, Alonso MP, Prats G, Blanco J.** 2011. Spread of *Escherichia coli* O25b:H4-B2-ST131 producing CTX-M-15 and SHV-12 with high virulence gene content in Barcelona (Spain). J. Antimicrob. Chemother. **66:**517–526.
29. **López-Cerero L, Bellido MD, Serrano L, Liró J, Cisneros JM, Rodríguez-Baño J, Pascual A.** 2012. *Escherichia coli* O25b:H4/ST131 are prevalent in Spain and are often not associated with ESBL or quinolone resistance. Enferm. Infecc. Microbiol. Clin. **31:**385–388.
30. **Novais Â, Viana D, Baquero F, Martínez-Botas J, Cantón R, Coque TM.** 2012. Contribution of IncFII and broad-host IncA/C and IncN plasmids to the local expansion and diversification of phylogroup B2 *Escherichia coli* ST131 clones carrying $bla_{CTX-M-15}$ and *qnrS1* genes. Antimicrob. Agents Chemother. **56:**2763–2766.
31. **Olesen B, Hansen DS, Nilsson F, Frimodt-Møller J, Leihof RF, Struve C, Scheutz F, Johnston B, Krogfelt KA, Johnson JR.** 2013. Prevalence and characteristics of the epidemic multiresistant *Escherichia coli* ST131 clonal group among extended-spectrum beta-lactamase-producing *E. coli* isolates in Copenhagen, Denmark. J. Clin. Microbiol. **51:**1779–1785.
32. **Matsumura Y, Yamamoto M, Higuchi T, Komori T, Tsuboi F, Hayashi A, Sugimoto Y, Hotta G, Matsushima A, Nagao M, Takakura S, Ichiyama S.** 2012. Prevalence of plasmid-mediated AmpC β-lactamase-producing *Escherichia coli* and spread of the ST131 clone among extended-spectrum β-lactamase-producing *E. coli* in Japan. Int. J. Antimicrob. Agents. **40:**158–162.
33. **Johnson JR, Urban C, Weissman SJ, Jorgensen JH, Lewis JS, Jr, Hansen G, Edelstein PH, Robicsek A, Cleary T, Adachi J, Paterson D, Quinn J, Hanson ND, Johnston BD, Clabots C, Kuskowski MA, AMERECUS Investigators.** 2012. Molecular epidemiological analysis of *Escherichia coli* sequence type ST131 (O25:H4) and $bla_{CTX-M-15}$ among extended-spectrum-β-lactamase-producing *E. coli* from the United States, 2000 to 2009. Antimicrob. Agents Chemother. **56:**2364–2370.
34. **Peirano G, van der Bij AK, Gregson DB, Pitout JD.** 2010. Molecular epidemiology over an 11-year period (2000 to 2010) of extended-spectrum β-lactamase-producing *Escherichia coli* causing bacteremia in a centralized Canadian region. J. Clin. Microbiol. **50:**294–299.
35. **Brisse S, Diancourt L, Laouénan C, Vigan M, Caro V, Arlet G, Drieux L, Leflon-Guibout V, Mentré F, Jarlier V, Nicolas-Chanoine MH, Coli β Study Group.** 2012. Phylogenetic distribution of CTX-M- and non-extended-spectrum-β-lactamase-producing *Escherichia coli* isolates: group B2 isolates, except clone ST131, rarely produce CTX-M enzymes. J. Clin. Microbiol. **50:**2974–2981.
36. **Suzuki S, Shibata N, Yamane K, Wachino J, Ito K, Arakawa Y.** 2009. Change in the prevalence of extended-spectrum-beta-lactamase-producing *Escherichia coli* in Japan by clonal spread. J. Antimicrob. Chemother. **63:**72–79.
37. **Clermont O, Dhanji H, Upton M, Gibreel T, Fox A, Boyd D, Mulvey MR, Nordmann P, Ruppé E, Sarthou JL, Frank T, Vimont S, Arlet G, Branger C, Woodford N, Denamur E.** 2009. Rapid detection of the O25b-ST131 clone of *Escherichia coli* encompassing the CTX-M-15-producing strains. J. Antimicrob. Chemother. **64:**274–277.
38. **Johnson JR, Johnston B, Clabots C, Kuskowski MA, Castanheira M.** 2010. *Escherichia coli* sequence type ST131 as the major cause of serious multidrug-resistant *E. coli* infections in the United States. Clin. Infect. Dis. **51:**286–294.
39. **Banerjee R, Johnston B, Lohse C, Porter SB, Clabots C, Johnson JR.** 2013. *Escherichia coli* sequence type 131 is a dominant, antimicrobial-resistant clonal group associated with healthcare and elderly hosts. Infect. Control Hosp. Epidemiol. **34:**361–369.
40. **Chung HC, Lai CH, Lin JN, Huang CK, Liang SH, Chen WF, Shih YC, Lin HH, Wang JL.** 2012. Bacteremia caused by extended-spectrum-β-lactamase-producing *Escherichia coli* sequence type ST131 and non-ST131 clones: comparison of demographic data, clinical features, and mortality. Antimicrob. Agents Chemother. **56:**618–622.
41. **Rodríguez-Baño J, Mingorance J, Fernández-Romero N, Serrano L, López-Cerero L, Pascual A, ESBL-REIPI Group.** 2013. Outcome of bacteraemia due to extended-spectrum β-lactamase-producing *Escherichia coli*: impact of microbiological determinants. J. Infect. **67:**27–34.
42. **Johnson JR, Tchesnokova V, Johnston B, Clabots C, Roberts PL, Billig M, Riddell K, Rogers P, Qin X, Butler-Wu S, Price LB, Aziz M, Nicolas-Chanoine MH, Debroy C, Robicsek A, Hansen G, Urban C, Platell J, Trott DJ, Zhanel G, Weissman SJ, Cookson BT, Fang FC, Limaye AP, Scholes D, Chattopadhyay S, Hooper DC, Sokurenko EV.** 2013. Abrupt emergence of a single dominant multidrug-resistant strain of *Escherichia coli*. J. Infect. Dis. **207:**919–928.
43. **Platell JL, Johnson JR, Cobbold RN, Trott DJ.** 2011. Multidrug-resistant extraintestinal pathogenic *Escherichia coli* of sequence type ST131 in animals and foods. Vet. Microbiol. **153:**99–108.
44. **Mora A, Herrera A, Mamani R, López C, Alonso MP, Blanco JE, Blanco M, Dahbi G, García-Garrote F, Pita JM, Coira A, Bernárdez MI, Blanco**

J. 2010. Recent emergence of clonal group O25b:K1:H4-B2-ST131 *ibeA* strains among *Escherichia coli* poultry isolates, including CTX-M-9-producing strains, and comparison with clinical human isolates. Appl. Environ. Microbiol. **76:**6991–6997.

45. **Clermont O, Lavollay M, Vimont S, Deschamps C, Forestier C, Branger C, Denamur E, Arlet G.** 2008. The CTX-M-15-producing *Escherichia coli* diffusing clone belongs to a highly virulent B2 phylogenetic subgroup. J. Antimicrob. Chemother. **61:**1024–1028.
46. **Johnson JR, Russo TA, Tarr PI, Carlino U, Bilge SS, Vary JC, Jr, Stell AL.** 2000. Molecular epidemiological and phylogenetic associations of two novel putative virulence genes, *iha* and *iroN* (*E. coli*), among *Escherichia coli* isolates from patients with urosepsis. Infect. Immun. **68:** 3040–3047.
47. **Johnson JR, Gajewski A, Lesse AJ, Russo TA.** 2003. Extraintestinal pathogenic *Escherichia coli* as a cause of invasive nonurinary infections. J. Clin. Microbiol. **41:**5798–5802.

OPEN ACCESS Freely available online

PLOS | GENETICS

Plasmid Flux in *Escherichia coli* ST131 Sublineages, Analyzed by Plasmid Constellation Network (PLACNET), a New Method for Plasmid Reconstruction from Whole Genome Sequences

Val F. Lanza[1☯], María de Toro[1☯], M. Pilar Garcillán-Barcia[1], Azucena Mora[2], Jorge Blanco[2], Teresa M. Coque[3,4,5], Fernando de la Cruz[1]*

1 Departamento de Biología Molecular (Universidad de Cantabria) and Instituto de Biomedicina y Biotecnología de Cantabria IBBTEC (UC-SODERCAN-CSIC), Santander, Spain, **2** Laboratorio de Referencia de *E. coli* (LREC), Departamento de Microbiología y Parasitología, Facultad de Veterinaria, Universidad de Santiago de Compostela, Lugo, Spain, **3** Departamento de Microbiología, Hospital Universitario Ramón y Cajal, Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain, **4** Unidad de Resistencia a Antibióticos y Virulencia Bacteriana asociada al Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain, **5** Centros de Investigación Biomédica en Red de Epidemiología y Salud Pública, (CIBER-ESP), Madrid, Spain

Abstract

Bacterial whole genome sequence (WGS) methods are rapidly overtaking classical sequence analysis. Many bacterial sequencing projects focus on mobilome changes, since macroevolutionary events, such as the acquisition or loss of mobile genetic elements, mainly plasmids, play essential roles in adaptive evolution. Existing WGS analysis protocols do not assort contigs between plasmids and the main chromosome, thus hampering full analysis of plasmid sequences. We developed a method (called plasmid constellation networks or PLACNET) that identifies, visualizes and analyzes plasmids in WGS projects by creating a network of contig interactions, thus allowing comprehensive plasmid analysis within WGS datasets. The workflow of the method is based on three types of data: assembly information (including scaffold links and coverage), comparison to reference sequences and plasmid-diagnostic sequence features. The resulting network is pruned by expert analysis, to eliminate confounding data, and implemented in a Cytoscape-based graphic representation. To demonstrate PLACNET sensitivity and efficacy, the plasmidome of the *Escherichia coli* lineage ST131 was analyzed. ST131 is a globally spread clonal group of extraintestinal pathogenic *E. coli* (ExPEC), comprising different sublineages with ability to acquire and spread antibiotic resistance and virulence genes via plasmids. Results show that plasmids flux in the evolution of this lineage, which is wide open for plasmid exchange. MOB_{F12}/IncF plasmids were pervasive, adding just by themselves more than 350 protein families to the ST131 pangenome. Nearly 50% of the most frequent γ–proteobacterial plasmid groups were found to be present in our limited sample of ten analyzed ST131 genomes, which represent the main ST131 sublineages.

Citation: Lanza VF, de Toro M, Garcillán-Barcia MP, Mora A, Blanco J, et al. (2014) Plasmid Flux in *Escherichia coli* ST131 Sublineages, Analyzed by Plasmid Constellation Network (PLACNET), a New Method for Plasmid Reconstruction from Whole Genome Sequences. PLoS Genet 10(12): e1004766. doi:10.1371/journal.pgen.1004766

Editor: Paul M. Richardson, MicroTrek Incorporated, United States of America

Received May 24, 2014; **Accepted** September 19, 2014; **Published** December 18, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files except for the raw data, which are available using the accession numbers provided in the manuscript and at http://www.ebi.ac.uk/ena/ (Acc. No. PRJEB6262).

Funding: Work was financed by the Spanish Ministry of Economy and Competitivity (BFU2011-26608 to FdlC, FIS-PI09/01273 and AGL2013-47852-R to JB and FIS-PI12-01581 and CB06/02/0053 to TMC), by the European Seventh Framework Program (612146/FP7-ICT-2013-10 to FdlC and 282004/FP7-HEALTH-2011-2.3.1-2 to FdlC and TMC); by Red Española de Investigación en Patología-a Infecciosa (REIPI RD06/0008/1018-1016) to JB, by Consellería de Cultura, Educación e Ordenación Universitaria, Xunta de Galicia and European Regional Development Fund, ERDF (CN2012/303 and EM2014/001) to JB and by the regional government of Madrid (PROMPT-S2010/BMD2414) to TMC. We are also grateful to the Spanish Network for the Study of Plasmids and Extrachromosomal Elements (REDEEX) for funding cooperation among Spanish microbiologists working on the biology of MGEs (Spanish Ministry of Science and Innovation BFU2011-14145-E). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* Email: delacruz@unican.es

☯ These authors contributed equally to this work.

Introduction

Clinical microbiology is being transformed by whole genome sequencing (WGS) [1]. A case in point is *Escherichia coli*: there were 1,618 *E. coli* projects submitted to NCBI compared to just 68 complete genomes by year 2013. Within the realms of clinical and environmental microbiology, plasmid analysis is increasingly used to track the dissemination of genes encoding virulence, resistance to antibiotics, heavy metals and biocides [2–4] and, to a lesser extent, to analyze differences in the adaptive evolution of certain clonal backgrounds [5,6]. Hybridization with specific probes [7], amplification of plasmid replication initiator proteins (RIP) [8–10], and relaxases (REL) [11] allow preliminary identification of plasmid families. In addition, plasmid MLST (pMLST) is used for

Author Summary

Plasmids are difficult to analyze in WGS datasets, due to the fragmented nature of the obtained sequences. We developed a method, called PLACNET, which greatly facilitates this analysis. As an example, we analyzed the plasmidome of *E. coli* ST131, an ExPEC clonal group involved in human urinary tract infections and septicemia. Relevant variation within this clone (e.g., antibiotic resistance and virulence) is frequently caused by the acquisition and loss of plasmids and other mobile genetic elements. Nevertheless, our knowledge of the ST131 plasmidome is limited to a few antibiotic resistance plasmids and to identification of replicons from known plasmid groups. PLACNET analysis extends the number of sequenced plasmids in ST131, which can be used for comparative genomics, from 11 to 50. The ST131 plasmidome is seemingly huge, encompassing roughly 50% of the main plasmid groups of γ–proteobacteria. MOB_{F12}/IncF plasmids are apparently the most active players in the dissemination of relevant genetic information.

epidemiological surveillance, but is restricted to individual plasmids of a few plasmid families of Enterobacteriaceae (http://pubmlst.org/plasmid/). This precludes the detection of plasmid mutations or rearrangements, as well as the identification of conjugative plasmids not represented in the pMLST database and of most mobilizable plasmids [11]. Finished plasmid/genome sequencing provides accurate and non-biased information, but is still expensive and thus seldom used specifically for plasmid analysis. Draft WGS dramatically cut down cost and analysis time. Although it allowed rapid and cheap data acquisition, WGS datasets typically result in more than a hundred contigs for a given genome, due to the short read lengths generally obtained. Genome fragmentation makes it difficult to distinguish between physical units, that is, between chromosome and plasmid sequences, as well as between different plasmids that usually coexist in bacterial cells. Several strategies can be followed to analyze WGS genome sequences, the workflow described by [12] being a typical example. There are also applications to identify plasmids in WGS sequences, such as PlasmidFinder (http://cge.cbs.dtu.dk/services/PlasmidFinder/), which identifies plasmids according to PCR-based replicon typing (PBRT) [8–10] and the subtyping scheme included in the pMLST web page (http://pubmlst.org). PlasmidFinder is limited by its inability to reconstruct the sequences of entire plasmids, underscoring the urgent need for improvement over existing tools.

E. coli ST131 is a successful high-risk clonal complex of pandemic distribution, able to cause extraintestinal infections in humans [13-18]. The increasing recovery of ST131 isolates from hospitalized and non-hospitalized individuals and, more recently, from companion and foodborne animals [17,19–25], sewage and main rivers of large European cities [26,27] highlights the rapid spread and local adaptation to different habitats of this lineage. ST131 is characterized by high metabolic potential [28] and a variable number of virulence factors, including adhesins, siderophores, toxins, polysaccharide coats (capsules and lipopolysaccharides), protectins and invasins [19,29,30], mostly acquired by recombination and by the interplay of mobile genetic elements (MGEs) [16]. Such traits, which are common among different lineages of the *E. coli* B2 phylogroup [31,32], enable strains to colonize mucosal surfaces, invade tissues, foil defence mechanisms and yield injurious inflammatory responses in the host. *E. coli* populations identified as ST131 by the widely used 'Achtman scheme' of multilocus sequence typing (MLST) [33] (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli), split in diverse clusters or subclones on the basis of genomic profile, serotype, content of virulence factors, antibiotic susceptibility pattern and the presence of certain *fimH* alleles [21,29,34–36]. The most prevalent ST131 clonal sublineage (*H*30) is characterized by the presence of a *fimH*30 allele, serotype O25:H4 and a specifically conserved *gyrA/parC* allele combination that confers fluoroquinolone resistance (FQ-R). Most human infections caused by ST131 are due to isolates of the *H*30 sublineage [13,16,37–39], many of them carrying the $bla_{\text{CTX-M-15}}$ gene which is responsible for resistance to third generation of cephalosporins. Some authors suggested differences between CTX-M-15 and non-CTX-M-15 producers, referred to as *H*30-R and *H*30-Rx sublineages, respectively [13,35,37,38]. Currently, diverse O25b:H4 ST131 variants (e.g. *fimH*22, *fimH*30) or O16:H5 (e.g. *fimH*41) seem also to be widely spread [13-16,40]. Full genome sequencing of several ST131 *E. coli* genomes, most of them *H*30-Rx variants [16,41–44], revealed further differences among strains, mainly chromosomal SNPs, indels and plasmid variations [16,43,44]. Heterogeneity of MGEs has been reported in other relevant *E. coli* clones, mainly Shiga-toxin producing *E. coli* (STEC) as O157:H7, O104:H4 or O26:H11 [5,6,45–47], often associated with ecological diversification of *E. coli* populations that can influence host-pathogen interactions [48,49]. Recently, International and European organisations including European Food Safety Agency, EFSA; European Centre for Disease Control, ECDC; Food Drug Administration, FDA; Centre for Diseases Control, CDC) and national food safety authorities underscored the need to identify clonal variants with enhanced transmissibility or pathogenicity as well as to infer the evolutionary history of pathogens of interest in Public Health (http://www.efsa.europa.eu/en/events/event/140616.htm). Because relevant adaptive traits are plasmid located, there is an urgent need to consider MGEs in population genetic studies.

In this work we describe PLACNET, a method to reconstruct plasmids from WGS datasets, and its application to the comprehensive analysis of bacterial plasmidomes. As a specific example, we describe the ST131 plasmidome and discuss its possible impact in the diversification of this clinically important lineage. PLACNET allows the identification of plasmids currently circulating among *E. coli* and other enterobacterial species that may be underestimated, thus providing a useful tool to approach comprehensive plasmid population genetic studies.

Results

Phylogeny of *E. coli* ST131 genomes

We analyzed ten *E. coli* genomes, classified as ST131 according to the Achtman scheme (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli), which branch in three main clusters identified as ST43, ST9 and ST506 (Fig. 1) according to the cgMLST Pasteur Institute scheme (http://www.pasteur.fr/recherche/genopole/PF8/mlst/EColi.html). The use of these two schemes is widely accepted in epidemiology [50] and increasingly used for *E. coli* typing. The ST43 branch contains isolates of the *H*30 lineage, which split in three subclusters (four strains of virotype C, two of virotype A, one of virotype B). The ST9 branch corresponds to isolates of the *H*22/*H*324 sublineage (virotype D). The most distal branch to the main cluster is represented by the commensal strain SE15, a member of sublineage *H*41 identified as ST506 [16]. It does not contain any marker used for the virotype subtyping method described by Blanco et al (*afa, sat, ibeA, iroN*) [36,51,52].

	JJ1886	BWH24	SE15	FV9873	E35BA	E2022	E61BA	BIDMC38	HVH177	BIDMC20B
JJ1886	0	308	9994	671	639	285	4621	672	4591	307
BWH24	308	0	10024	658	619	133	4647	657	4617	7
SE15	9994	10024	0	10127	9967	10001	7990	10128	7958	10023
FV9873	671	658	10127	0	719	631	4782	57	4754	657
E35BA	639	619	9967	719	0	594	4595	724	4569	620
E2022	285	133	10001	631	594	0	4620	638	4594	132
E61BA	4621	4647	7990	4782	4595	4620	0	4785	206	4646
BIDMC38	672	657	10128	57	724	638	4785	0	4755	656
HVH177	4591	4617	7958	4754	4569	4594	206	4755	0	4616
BIDMC20B	307	7	10023	657	620	132	4646	656	4616	0

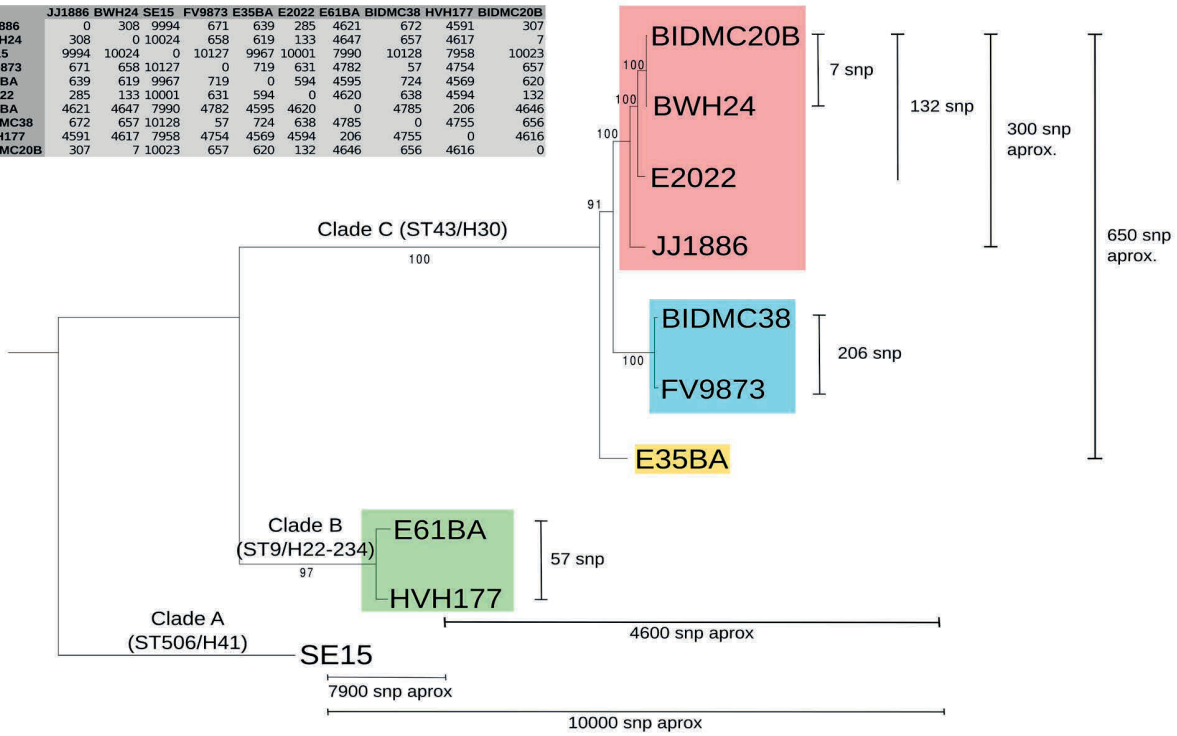


Fig. 1. Phylogenetic tree of ST131 *E. coli*. The tree is based on a 3,629,034 bp core genome (3,734 orthologous genes: 90% identity and 90% coverage) and 100 bootstrapping replicates. ST131 clades are named according to [16] and further subdivided and colored according to virotypes [36]: virotype A (blue), virotype B (yellow), virotype C (pink), virotype D (green). Virotype classification is based on the presence/absence of four putative virulence factors: *afaFM955459* (encoding an Afa/Dr adhesin), *sat* (secreted autotransporter toxin, present in PAI-CFT073-pheV), *ibeA* (invasion of brain endothelium) and *iroN* (salmochelin siderophore receptor). The commensal ST131 strain SE15 was used to root the tree (virotype non typable; serotype O150 in the original publication [91] but lying within the H41 cluster in the phylogenomic study of [16]). Given SNP numbers are approximate averages of individual comparisons.
doi:10.1371/journal.pgen.1004766.g001

Thus, the sample analyzed in this work includes representatives of all ST131 branches described to date [13,16]. The core genome of the 10 strains encompasses 3.6 Mb (Fig. 1 inset). As can be seen, the phylogenetic tree of ST131 genomes can be rooted at the commensal strain SE15. It should be noted, however, that SE15 is not necessarily the ancestor of the pathogenic lineages, as inferred by recent evidence [16]. The divergence of SE15 from the other ST131 strains is of about 3,000 SNP/Mb, a measure of the depth of the ST131 phylogenetic branch (<0.3% divergence in the core genome). There are only 650 SNPs among the genomes of cluster C lineage (i.e., <200 SNP/Mb), indicating their close phylogenetic relationship. There are <300 SNPs within a given virotype. The average distance between clades A and B is of about 4,600 SNPs (i.e., 1,300 SNP/Mb).

Plasmid reconstruction in *E. coli* ST131 genomes

The PLACNET protocol was used as explained in Materials and Methods. We proceeded with plasmid reconstruction, as exemplified in Fig. 2 for the reconstruction of the E61BA genome (ST9/H324/virotype D). When we applied the rules for reference homology, scaffold links and plasmid protein tagging, the E61BA network shown as "original network" was produced. Obviously, this network was not neat enough to allow plasmid reconstruction. Expert pruning of the network consisted on several steps. First, contigs smaller than 200 bp were eliminated. Second, hubs were identified (see arrows in the original network of Fig. 2), duplicated and assigned to separate disjoint connected components. Scaffold links and coverage information, as well as score values of conflict edges, were used to decide on valid component assignment. Inspection of the coding potential of hubs usually showed them to correspond to ISs, transposons or other known repeated elements (as shown in S9 and S10 Figs.). As a result, a pruned network was reconstructed as shown in Fig. 2. Differential coloring of disjoint connected components in the pruned network thus displayed the final network of plasmids (as contig constellations). In PLACNET Cytoscape representation, most plasmids can be identified by their RIP and/or REL proteins. Thus, the reconstructed E61BA genome contains seven plasmids: a 134 kb MOB_{F12}/IncF plasmid (pE61BA-1), a 37.7 kb MOB_{P6}/IncI2 plasmid (pE61BA-7), a 24.5 kb MOB_{C12} plasmid (pE61BA-2), a 18 kb MOB_{P11}/IncP1 plasmid (pE61BA-4), two MOB_{P5}/ColE1-like plasmids of 6.6 and 6.9 kb (pE61BA-5 and pE61BA-6, respectively) and one MOB_{Q12} 5.0 kb plasmid (pE61BA-3). Only plasmid pE61BA-2 could be closed, the remaining contained at least two contigs. Thus, their reported sizes are minimum sizes, since they might include small repeated sequences that were taken out of the analysis during network pruning. Two contigs remained as "not assigned" to any physical unit in this particular genome because they did not show any reference or scaffold link that bind them to other contigs: a 2,953 bp contig (containing a putative DNA primase and a lytic transglycosylase) and a 1,301 bp contig (containing two conjugation-related genes: *trbI* and a partial *traB* gene).

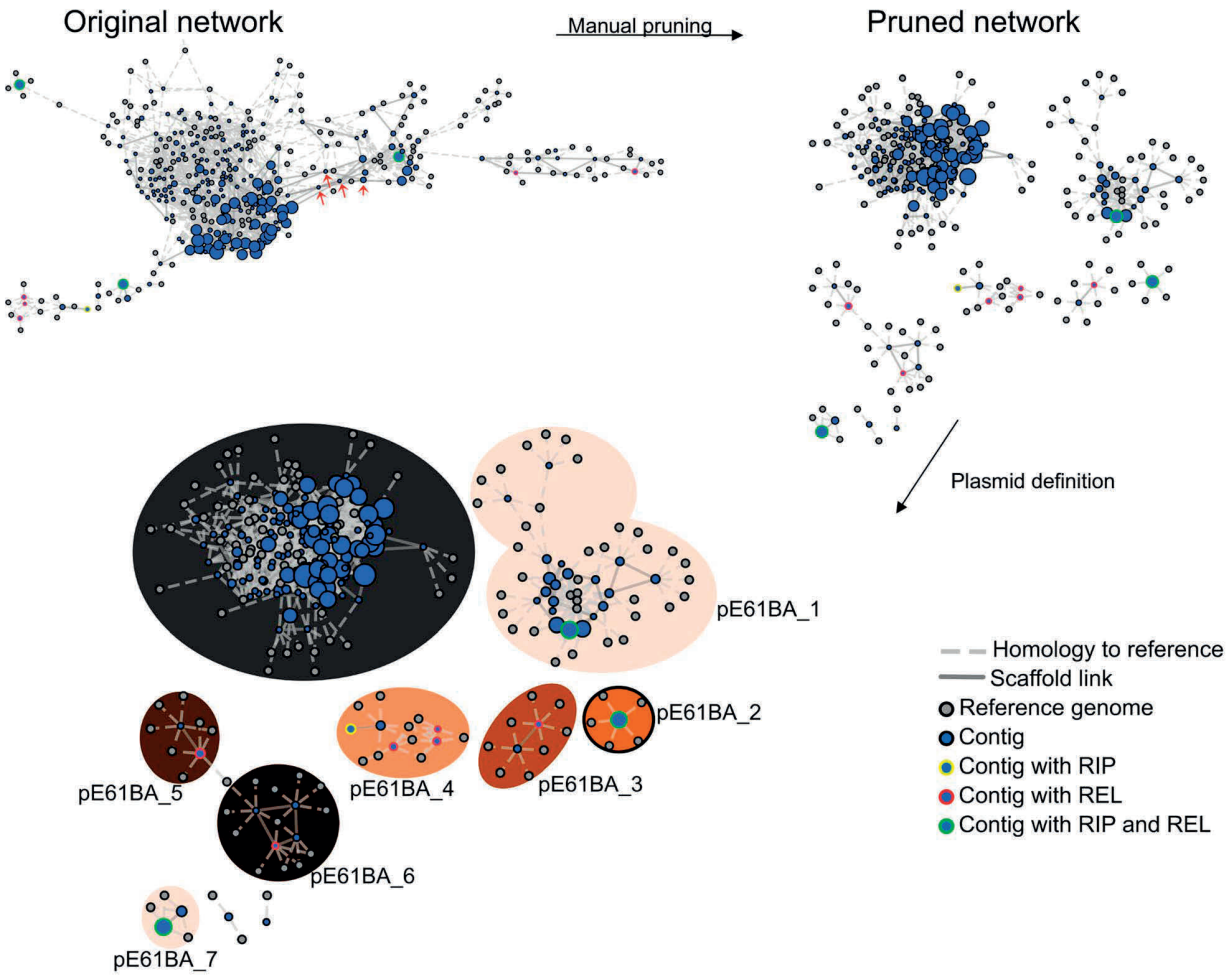


Fig. 2. PLACNET plasmid reconstruction of ST131 genome E61BA (ST9/H324/virotype D). The network contains nodes of two different colors (blue for contigs, grey for reference genomes). The size of reference nodes is always the same. The size of the contig nodes is proportional to the contig length. Besides, outlines are yellow for contigs containing RIP proteins, red for relaxases and green for both proteins. Edges are either solid (scaffold links) of dotted (homologous references). The length of the edges is arbitrarily selected by Cytoscape algorithm. In the upper left, the network output (original network) is shown, which resulted from automatic reference search, scaffold links and protein tagging rules. The original network was converted to a pruned network by eliminating contigs smaller than 200 bp and duplicating specific hubs (red arrows). Two contigs could not be assigned for lack of scaffold links: a 2,953 bp contig (putative DNA primase + lytic transglycosylase) and a 1,301 bp contig (TrbI + TraB-partial). Closed plasmids (e.g., pE61BA_2, size: 24,447 bp) are shown with a black outline in the final PLACNET network.
doi:10.1371/journal.pgen.1004766.g002

The same procedure was applied to the three other strains sequenced for this work as well as to the four genomes obtained from public DBs as Illumina reads. The plasmid content of the four strains sequenced in this work was confirmed by the analysis of S1-digested genomic DNA profiles by PFGE. This analysis fully confirmed the presence of plasmids of similar size to those identified by PLACNET (S2 Text and S4 Table), In the case of strain E35BA, in which PLACNET identified two IncF plasmids that could not be separated (totaling 211 kb), S1-PFGE identified two plasmids of 140 kb and 75 kb. As a result of PLACNET analysis, we obtained the plasmid constellation networks shown in S1 to S8 Figs. A summary of the results, i.e., the reconstructed plasmids, is shown in Table 1, which includes also the plasmids of the ST131 reference strains JJ1886 and SE15. As can be seen, the number of plasmids in the ST131 genomes is variable, even from strains belonging to the same ST131 sublineage, ranging from just one plasmid in HVH177 (clade B/ST9/*fimH22*) or SE15 (clade A/ST506/*fimH41*) to seven plasmids in E61BA (clade B/ST9/*fimH22*), to give an average of 4 plasmids per genome. There is not a single plasmid group that appears specific of a particular sublineage. S1 Table contains the complete list of contigs assigned to each plasmid or chromosome.

Overall plasmid diversity is visualized in plasmid dendrograms. Overall, the ten ST131 genomes analyzed contain 39 plasmids (including one potential ICE), which can be assorted by their relative sizes and MOB groups [53], as shown in Table 1. The most conspicuous group was that of MOB_{F12}/IncF plasmids (11 plasmids), present in all ten sequenced ST131 genomes. Other relevant plasmid backbones belong to the MOB_P (RIP groups IncI1/K, IncI2, IncX1, IncX4, ColE1), MOB_Q (Qu, Q12) and MOB_C (C12) REL families. The non-F plasmids comprise a total of 20 plasmids belonging to eight plasmid groups. Two plasmids were phage-like and belong to the Rep-3 RIP family. Finally, 5 plasmids corresponded to the no-MOB category. The E35BA genome (ST43/H30 virotype B) showed a MOB_{P11} relaxase within a 234 kb chromosomal contig, implying the

Table 1. Summary of plasmid content.

Genome	*fimH* allele	Strain virotype	MOB_{F12}/IncF	Phage-related/ RepFIB	MOB_{P12}/ IncI	MOB_{P6}/ IncI2	MOB_{F11}/IncN	MOB_{P3}/IncX	MOB_{P11}/IncP1	MOB_{C12}[d]	MOB_{P5}/ ColE1-like	MOB_{Qu}[d]	MOB_{Q12}[d]	no-MOB[d]
FV9873	*H*30	A	pFV9873_5 (91.4 Kb; ΔTraI)					pFV9873_4 (33.3 Kb)				pFV9873_1 (4.1 Kb)	pFV9873_6 (5.2 Kb)	pFV9873_2 (2.2 Kb); pFV9873_3 (4.6 Kb)
BIDMC38	*H*30	A	pBIDMC38_5 (123 Kb)								pBIDMC38_1 (11.8 Kb)	pBIDMC38_4 (4.2 Kb)	pBIDMC38_2 (5.3 Kb)	pBIDMC38_3 (1.6 Kb)
E35BA	*H*30	B	p35BA_2+3 (211 Kb)						IME_E35BA (14.2 Kb)			pE35BA_1 (4.1 Kb)		
E2022	*H*30	C	pE2022_2 (103 Kb)		pE2022_1 (98.3 Kb)			pE2022_3 (35.0 Kb)				pE2022_4 (4.1 Kb)		pE2022_5 (2.2 Kb)
BIDMC20B	*H*30	C	pBIDMC20B_1 (128 Kb, ΔTraI[b])	pBIDMC20B_2 (109 Kb)										
BWH24	*H*30	C	pBWH24_1 (123 Kb, ΔTraI[b])	pBWH24_2 (109 Kb)		pBWH24_3 (60.3 Kb)								
JJ1886	*H*30	C	pJJ1886-5 (110 Kb ΔTraI)						pJJ1886-4 (55.9 Kb)		pJJ1886-3 (5.6 Kb)		pJJ1886-2 (5.2 Kb)	pJJ1886-1 (1.6 Kb)
E61BA	*H*324	D	pE61BA_1 (137 Kb)			pE61BA_7 (37.9 Kb)			pE61BA_4 (18.3 Kb)	pE61BA_2 (24.5 Kb)	pE61BA_5 (6.5 Kb); pE61BA_6 (6.9 Kb)		pE61BA_3 (5.5 Kb)	
HVH177	*H*22	D	pHVH177_1 (78.6 Kb)											
SE15	*H*41	Commensal	pECSF1 (122 Kb)											
Other ST131 plasmids[a]			pEK516 (64.5 Kb, ΔTra); pEK499 (117 Kb, ΔTra); pJIE186-2[c] (138 Kb); pGUE-NDM (87.0 Kb)		pEK204 (93.7 Kb)		pKC394 (53.2 Kb); pKC396 (44.2 Kb); pNDM-ECS01 (41.2 Kb); pECN580 (64.9 Kb)	pJIE143 (34.3 Kb)						

[a]Plasmid references: pEK516 ([93]; EU935738); pEK499 ([93]; EU935739); pJIE186-2 ([56]; NC_020271); pGUE-NDM (in [119]; JQ364967); pEK204 ([93]; EU935740); pKC394 ([120]; HM138652); pKC396 ([120], HM138653); pNDM-ECS01 (Unpublished; KJ413946); pECN580 ([121]; KF914891); pJIE143 ([94]; JN194214).

[b]pBIDMC20B_1 and pBWH24_1 plasmids lacked the REL domain of the TrwC protein.

[c]Plasmid pJIE186-2 was isolated from strain JIE186 [94], although GenBank acc. n° NC_020271 specifies it is located at EC958 strain. Strain JIE186 also contains plasmid pJIE186-1, not included in this study as it is not available at public DBs.

[d]No correlation with RIP typing methods.

doi:10.1371/journal.pgen.1004766.t001

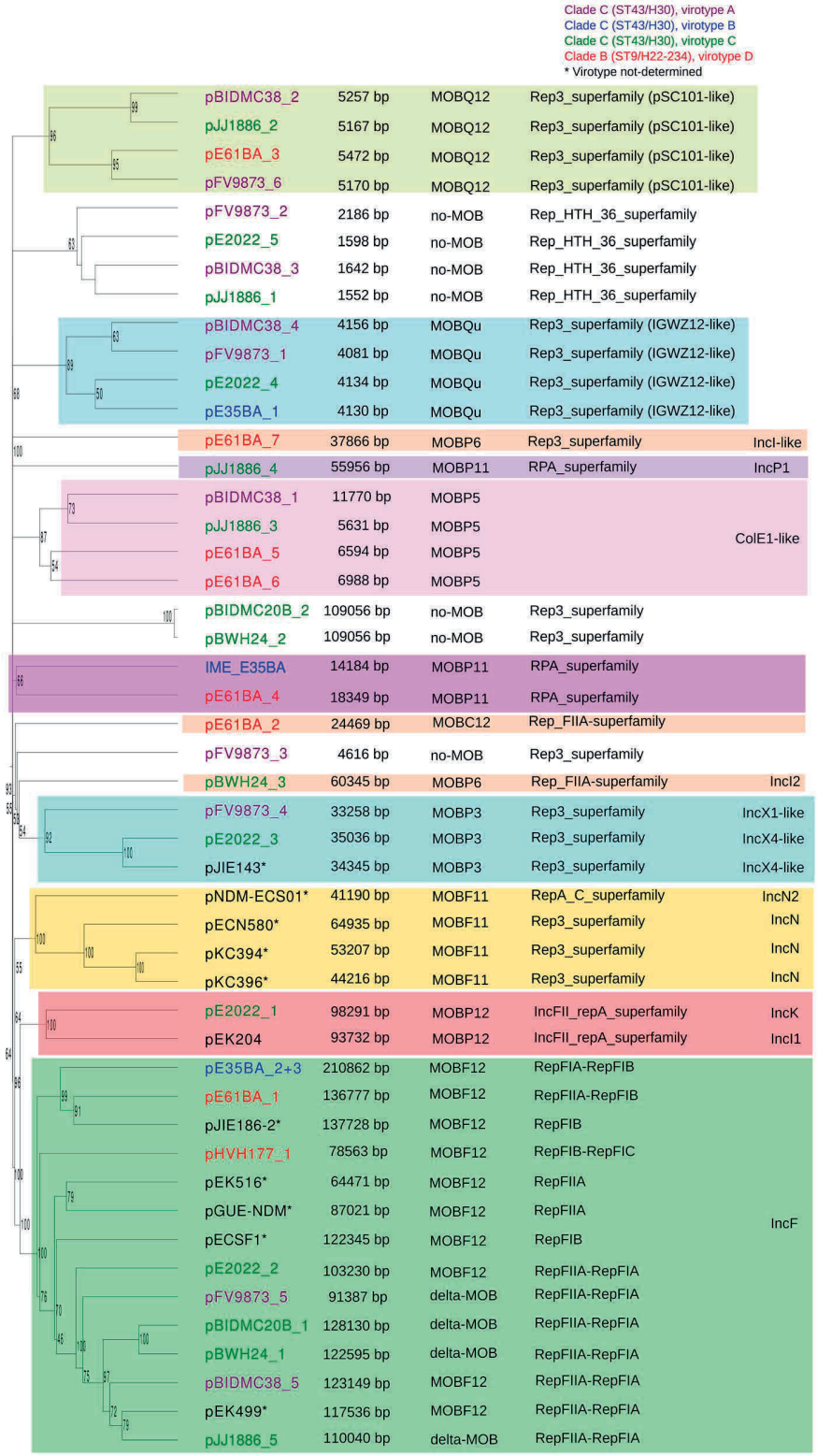
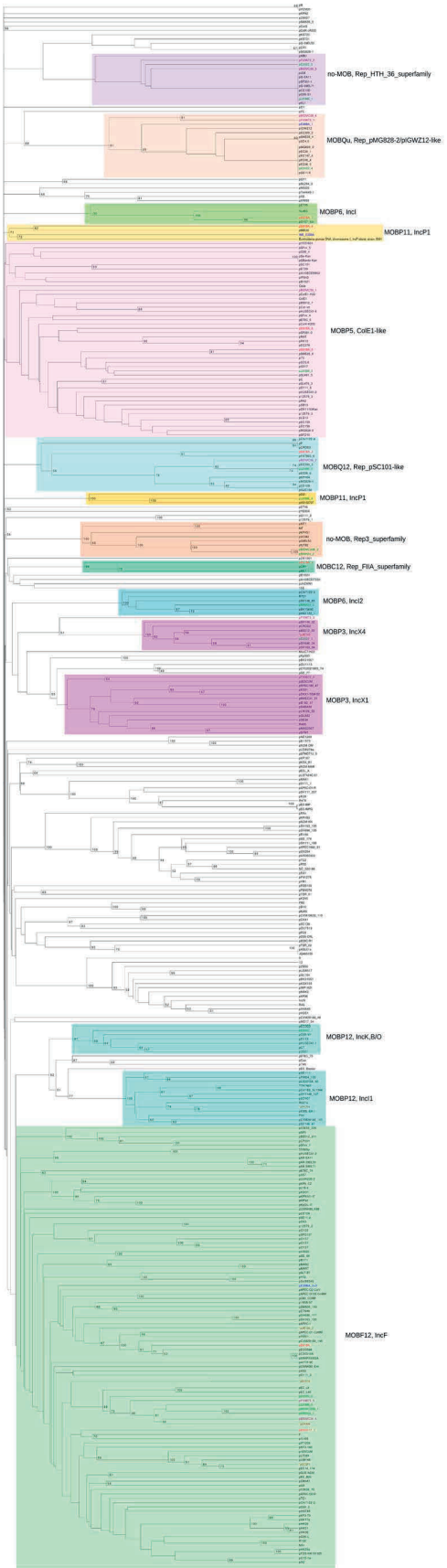


Fig. 3. Hierarchical clustering dendrogram of ST131 plasmids. The UPGMA dendrogram was based on protein cluster analysis using 60% sequence identity and 80% coverage. Plasmid names are colored according to their clade, taking into account ST, *fimH* allele and virotype, following the color code shown at the upper right. The five plasmid names in black correspond to previously sequenced plasmids from ST131 strains. Different color backgrounds are shown to emphasize branches of related plasmids. To the right of the dendrogram, four columns show, respectively, plasmid size, MOB type, RIP type and Inc type.
doi:10.1371/journal.pgen.1004766.g003



Clade C (ST43/H30), virotype A
Clade C (ST43/H30), virotype B
Clade C (ST43/H30), virotype C
Clade B (ST9/H22-234), virotype D
* Virotype not determined

MOBF12, IncF

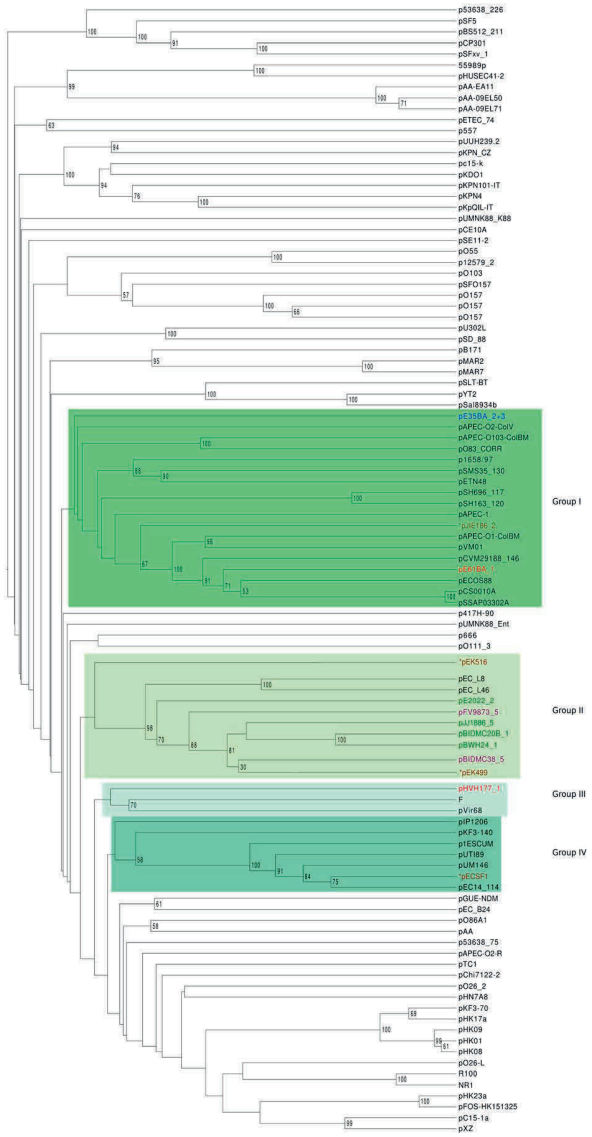


Fig. 4. Hierarchical clustering dendrogram of ST131 plasmids and relevant references. The left dendrogram shows the complete tree, with references. Dendrogram construction and color codes are as in Fig. 4. The right dendrogram expands the MOB_{F12}/IncF branch, with new background colors highlighting plasmid groups within this branch that are mentioned in the text.
doi:10.1371/journal.pgen.1004766.g004

presence of an ICE (integrative and conjugative element). Detailed inspection of this contig identified a 14.2 kb IME (integrative and mobilizable element) (see below and Suppl. Mat.). Once plasmids were identified by PLACNET and contigs assorted, the next step in plasmid analysis consisted in the construction of a dendrogram that summarizes plasmid gene content, allowing a visualization of relatedness between individual plasmids. The dendrogram of the 44 plasmids found in ST131 genomes (39 described in this report plus five plasmids already published) is shown in Fig. 3. The figure shows how the plasmids divide into branches that coincide with backbone MOB groups. There are 14 plasmid groups, according to the dendrogram, shown in the figure by different color backgrounds. Since each dendrogram group links related plasmids, they can be now analyzed individually, by comparing them either among themselves (Fig. 3) or with selected references (Fig. 4).

Analysis of IncF plasmids. Representatives of the largest plasmid group, formed by 15 MOB_{F12}/IncF plasmids, were found in each of the analyzed ST131 strains (Table 1). Included in this set are four plasmids lacking MOB and *tra* regions but containing RIP and other backbone genes related to IncF plasmids. The dendrogram of Fig. 3 clearly indicates that these plasmids belong to the IncF plasmid family, underscoring the usefulness of this step in plasmid reconstruction. Judging from their position in the dendrogram, it seems IncF plasmid genes are scrambled as if the precise constitution of each individual IncF plasmid could not be predicted at all for isolates of each specific ST131 cluster or phylotype. This is even clearer in Fig. 4. In this figure, ST131 plasmids are represented together with the reference sequences that were used for PLACNET reconstruction and analysis. Besides, BRIG comparison of IncF plasmids (Fig. 5) reveals high heterogeneity between them, with not a single completely conserved gene (confirmed by the fact that not a single plasmid gene was found to belong to the ST131 core genome). KClust software [54] at 30% identity and 50% coverage was used to group all proteins coded by IncF plasmids into 354 reference clusters. Manual curation was used to classify these 354 clusters in three groups (Fig. 5): (i) plasmid backbone (i.e., conjugation, RIP and maintenance genes) and metabolic protein genes, (ii) antibiotic resistance and virulence genes, and (iii) other protein genes such as ISs, transposases or hypothetical proteins. Conjugation proteins represent 30 of the 53 backbone proteins and constitute the most conserved set. As mentioned above, 4/14 plasmids do not keep a complete backbone. Table 2 contains the functional annotation of the IncF plasmids. As shown, 11/15 of the MOB_{F12}/IncF plasmids contain antibiotic resistance genes (conferring resistance to beta-lactams, in all cases, but also to sulfonamides, aminoglycosides, trimethoprim, chloramphenicol, tetracycline and macrolides, in some of them). In addition, nine of the ten antibiotic resistance-plasmids confer a multidrug-resistance (MDR) phenotype (equal or more than four antibiotic families). Besides, 10/15 MOB_{F12}/IncF plasmids presented putativevirulence genes (Table 2). As previously noted, there was an apparent trade-off between antibiotic resistance and virulence, genes coding for these adaptive traits being located in different plasmids [2]. Finally, a DNA modification gene (adenine-specific DNA methylase) was conserved in all IncF plasmids except in pHVH177_1.

The ST131 IncF plasmids belong to four different branches of the dendrogram, as shown in Fig. 4 inset. Group I includes four plasmids similar to the well-known virulence plasmids pAPEC-ColV like (also called pS88-like), which are commonly detected among avian pathogenic *E. coli* (APEC) [2,55]. A suitable reference is the ST131 plasmid pJIE186-2, coming from a ST131 strain previously recovered in Australia in 2006 [56]. As shown in S11A Fig., group I IncF plasmids share two large homologous regions: a 70 kb region containing virulence genes *iss, iroBCDEN, iucABCD, iutA, cvaBC* and *sitC* and the cassette *ompT-hlyF-mig14*, eventually also linked to *estABCDE* [55] and a 40 kb region containing the *tra* region and other backbone genes. Group II contains 10 MDR plasmids, 8 of which are ST131 plasmids with characteristic F2:A1:B- replicons and multiple antibiotic resistancegenes. A suitable reference is the ST131 plasmid pJJ1886-5, coming from a ST43/*fimH*30 lineage from USA. As shown in S11B Fig., group II IncF plasmids share most of their genomes. It should be noted that three of these plasmids (pFV9873_5, pEK499 and pEK516) have extensive deletions within their *tra* regions, as seen in the figure. Groups III and IV are just represented by one plasmid each. None of them contains antibiotic resistance genes and they are poor in virulence genes. While group III plasmid pHVH177_1 is not similar to any reference outside the backbone genes (S11C Fig.), the group IV plasmid pECSF1 is extensively similar to various large *E.coli* plasmids, (S11D Fig.). A more comprehensive comparison of F plasmids recovered from ST131 with previously described F-like plasmids is given in Suppl. Mat.

Analysis of other ST131 plasmid groups. Besides MOB_{F12}/IncF plasmids, 28 other plasmids were represented in ST131 isolates (Table 1). Ten were large, presumably conjugative plasmids (>18 kb), while 17 were small plasmids (<12 kb) and there was one IME.

Among the large plasmids, a most remarkable branch is composed by two almost identical 109 kb plasmids (pBIDMC20B_2 and pBWH24_2) present in two ST43/*H*30 isolates of virotype C, for which only RepFIB (Rep3-superfamily) and the maintenance protein ParB could be identified as plasmid backbone genes. No conjugative genes were detected. On the other hand, they code for an integrase protein and several phage-typical proteins. The plasmids are highly similar to pECOH89, recently recovered from a CTX-M-15 producer *E. coli* isolate from Germany [57]. Closest reference hits were the adherent invasive *E. coli* (AIEC) plasmid pLF82, isolated from a patient with Crohn's disease [58], the STEC plasmid p09EL50 [5], the *Salmonella* plasmid pHCM2 [59] and the *Salmonella* bacteriophage SSU5 [60]. These are all cryptic plasmids isolated from pathogenic enterobacteria that have been barely analyzed and thus are poorly annotated. The possibility arises of these elements being similar to lysogenic phages that are stably maintained as plasmids, analogous to phage P1 [57]. S12 Fig. compares this branch of related plasmids, using plasmid pECOH89 as a reference. As can be seen, both ST131 plasmids share most of their sequences with this 111 kb plasmid, including several phage-like protein genes. Significantly, none of the cryptic plasmids described in this study or those mentioned in the references, except pECOH89, harbor a resistance gene.

The 98.3 kb MOB_{P12}/IncK plasmid pE2022_1 is most similar to pCT [61]. pE2022_1 contains a $bla_{CTX\text{-}M\text{-}14}$ gene identical to that in pCT, a plasmid carrying $bla_{CTX\text{-}M\text{-}14}$ that is globally spread among humans and animals and particularly prevalent in clinical

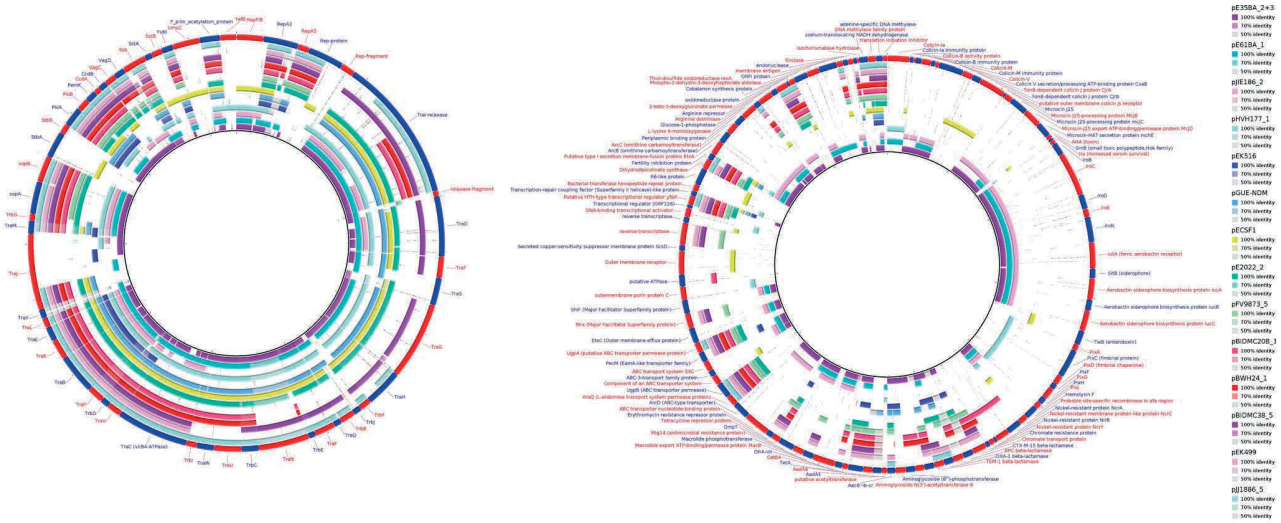


Fig. 5. MOB_{F12}/IncF plasmid analysis. Protein cluster analysis was performed with kClust software (parameters: 30% identity, 50% coverage) on the set of 14 plasmids shown in Table 4. Plasmid pGUE-NDM [119] was excluded from this comparison since it is only distantly related to the others (see dendrogram in Fig. 5). A total of 354 protein clusters were obtained and annotated versus the NCBI protein database (Blastp). Manual inspection was carried out to classify the reference proteins of each cluster into one of these three groups (comparative analysis shown with BRIG): (i) Backbone and metabolic proteins (panel A); (ii) Virulence and Antibiotic resistance proteins (panel B); and (iii) ISs and hypothetical proteins (not shown).
doi:10.1371/journal.pgen.1004766.g005

isolates form Spain [61,62]. The backbone of plasmid pE2022_1 is homologous to that of the reference IncI1 plasmid pEK204. These plasmids are described in S13. Fig. Despite their different Inc names, IncI1, IncK and IncB/O have similar backbones, belonging to different branches of the IncI complex (an analogous case to IncF plasmids).

MOB_{P6} is a large plasmid family, as can be observed in the phylogenetic tree of the MOB_{P6} relaxase family (S14A Fig.). The two ST131 MOB_{P6}/IncI2 plasmids are rather different, as judged by the distant positions of their REL. Plasmid pBWH24_3 (60.3 kb) is similar to the IncI2 prototype R721 [63], but most similar to the APEC plasmid pChi7122_3 [64]. In turn, pE61BA_7 (37.9 kb) is most similar to *Salmonella agona* plasmid SL483 and the enterohemorrhagic *E. coli* (EHEC) plasmid pO157_Sal [65]. They were recovered from isolates identified as *H*30_virotype C from the USA and *H*324_virotype D from Spain, respectively.

Another ST131 important group is MOB_{P3}/IncX, composed by three ST131 plasmids (pFV9873_4 and pE2022_3, as well as the reference pJIE143). Plasmids pFV9873_4 and pE2022_3, obtained from different *H*30 subgroups, are rather different between them and belong to different plasmid groups (IncX1 and IncX4), as shown their relatively distant positions in the MOB_{P3} phylogenetic tree of S15A Fig., even if showing similar sizes of about 34 kb. Their coding capacity is shown in the BRIG representations shown in S15B and S15C Fig. The IncX1 plasmid pFV9873_4 is most similar to the EC plasmid p2ESCUM [66], although genetic similarity is constrained to their backbone genes, occupying about half of the reference plasmid sequences. Conversely, the IncX4 plasmid pE2022_3 is most similar to pSH696_34 and the ST131 reference plasmid pJIE143 all over its sequence length (S15C Fig.).

Two plasmids and the IME belong to the MOB_{P11}/IncP1 family, as shown in the MOB_{P11} relaxase phylogenetic tree of S16A Fig. Plasmids pJJ1886_4 and pE61BA_4 showed widely different sizes (56 and 18 kb, respectively). Plasmid pE61BA_4 is only distantly related in its backbone genes to environmental plasmid pMBUI2, isolated from an uncultured bacterium [67]. Plasmid pJJ1886_4, on the other hand, is similar to the *E. coli* plasmid pHS102707 (GeneBank Acc. N° KF701335). These two plasmids thus represent new additions to the ST131 plasmidome (see S16B Fig.). The IME_E35BA is a 14.2 kb insertion within a 234 kb chromosomal contig. S16C Fig. shows some detail on the genetic structure of the IME and its insertion site in the ST131 core genome.

The last potentially conjugative plasmid is the 24.5 kb MOB_{C12} plasmid pE61BA_2, also not closely related to any reference, as seen in the REL phylogenetic tree of S17A Fig. This plasmid resulted in a single contig, so it could be closed. The closest homolog is the *Yersinia pestis* plasmid pCRY, with which it shares all backbone genes. Its RIP belongs to the Rep_FIIA superfamily, although this plasmid group is not represented in the classical PBRT method [8]. It does not contain any gene with known adaptive function, except a protease and a putative secreted thermonuclease. As in the case above, this plasmid represents a new addition to the ST131 plasmidome (S17B Fig.), its relaxase being only 70% identical to its closest homolog, the pCRY relaxase.

Among the small plasmids, the first group is composed by four MOB_{P5}/ColE1-like plasmids (11.8, 6.9, 6.6 and 5.6 kb). Three of the plasmids are relatively different, as judged by the MOB_{P5} phylogenetic tree of S18A Fig. The large plasmid (pBIDMC38_1) contains a type II restriction-modification system (Cfr10I) and is almost identical to the ST131 reference pJJ1886_3 (S18B Fig.). The two MOB_{P5}/ColE1 plasmids of strain E61BA (plasmids pE61BA_5 and pE61BA_6) contain colicin ColE and ColK genes, respectively (S18C and S18D Fig.). Colicins are considered both as virulence factors as well as traits that influence bacterial fitness and survival in the presence of competitors [68].

Besides, there were four almost identical MOB_{Qu} plasmids of around 4.1 kb (S19A Fig.), which populate all *H*30 subgroups (two of virotype A, one of virotype B and one of virotype C). Nothing remarkable could be distinguished it their genetic constitution, besides a common MOB region and a pIGWZ12 -like Rep protein

Table 2. Resistance genes and virulence determinants in MOB_{F12} plasmids.

Plasmid	Strain Virotype	Size (Kb)	RIP (FAB formula)[a]	Antibiotic resistance genes[b]	Virulence genes[c]
pFV9873_5	A	91.4	RepFIIA-RepFIA (F2:A1:B-)	*bla*$_{CTX-M-15}$, *bla*$_{OXA-1}$, *sul1*, *aadA5*, *aac*(6′)-Ib-cr, *dfr*A17, *cat*B4, *tet*(A), *mphA*	None-detected
pBIDMC38_5	A	123	RepFIIA-RepFIA (F2:A1:B-)	*bla*$_{TEM-1}$, *sul1*, *aadA5*, *tet*(A), *mphA*	*traT*, *finO*
pE2022_2	C	103	RepFIIA-RepFIA (F2:A1:B-)	*bla*$_{CTX-M-15}$, *bla*$_{TEM-1}$, *bla*$_{OXA-1}$, *aac*(6′)-Ib-cr-like[d], *cat*B4, *tet*(A)	*traT*, *finO*
pBIDMC20B_1	C	128	RepFIIA-RepFIA (F2:A1:B-)	*bla*$_{KPC-3}$, *sul1*, *sul2*, *strA*, *strB*, *aadA5*, *dfr*A17, *tet*(A)	*finO*
pBWH24_1	C	123	RepFIIA-RepFIA (F2:A1:B-)	*bla*$_{KPC-3}$, *sul1*, *sul2*, *strA*, *strB*, *aadA5*, *dfr*A17, *tet*(A)	*finO*
pJJ1886-5	C	110	RepFIIA-RepFIA (F2:A1:B-)	*bla*$_{TEM-1}$, *bla*$_{OXA-1}$, *aac*(6′)-Ib-cr	None-detected
pEK499	-	117	RepFIIA-RepFIA (F2:A1:B-)	*bla*$_{CTX-M-15}$, *bla*$_{TEM-1}$, *bla*$_{OXA-1}$, *sul1*, *aadA5*, *aac*(6′)-Ib-cr, *dfr*A17[e], *tet*(A), *mphA*	None-detected
pEK516	-	64.5	RepFIIA (F2:A-:B-)	*bla*$_{CTX-M-15}$, *bla*$_{TEM-1}$, *bla*$_{OXA-1}$, *aac*(6′)-Ib-cr, *aac*(3′)-IIa, *cat*B4, *tet*(A)	None-detected
pGUE-NDM	-	87.0	RepFIIA (F2:A-:B-)	*bla*$_{NDM-1}$, *bla*$_{OXA-1}$, *ble*$_{MBL}$, *sul1*, *aac*(6′)-Ib-cr, *aac*(3′)-II, *aad*A2, *dfr*A12, Δ*cat*B4	*traT*
pE61BA_1	D	137	RepFIIA-RepFIB (F2:A-:B1)	*bla*$_{TEM-1}$, *tet*(A)	*ompT*, *iss*, *iroBCDEN*, *iucABCD*, *cvaBC*, *traT*, *etsABC*, *mig14-hlyF-finO*
pECSF1	Commensal	122	RepFIIA-RepFIB (F29:A-:B10)	None-detected	*traT*, *finO*
pE35BA_2+3	B	211	RepFIA-RepFIB (F-:A2:B1)	*bla*$_{CTX-M-15}$, *bla*$_{TEM-1}$, *bla*$_{OXA-1}$, *sul2*, *strA*, *strB*, *aac*(6′)-Ib-cr, *aac*(3′)-IIa, *dfr*A14, *cat*B4, *tet*(A)	*ompT*, *iss*, *iroBCDEN*, *iucABCD*, *iutA*, *cvaBC*, *sitC*, *traT*, *etsABC*, *mig14-hlyF-finO*
pJIE186-2	-	138	RepFIB (F-:A-:B1)	None-detected	*ompT*, *iss*, *iroBCDEN*, *iucABCD*, *iutA*, *cvaBC*, *etsABC*, *mig14-hlyF-finO*
pHVH177_1	D	78.6	RepFIB (F-:A-:B31)	None-detected	*traT*, *etsABC*, *mig14-hlyF-finO*

[a]FAB formula according to http://pubmlst.org/plasmid/classification scheme [9].
[b]According to the ARG-annot database (>90% amino acid identity) [http://en.mediterranee-infection.com].
[c]According to our in-house database (>90% amino acid identity).
[d]*aac(6′)-Ib-cr*-like presents the Glu72Gly additional mutation.
[e]In the original paper [93] *dfrA7* is reported, instead of *dfrA17*. However, inspection of its amino acid sequence indicates it is a DfrA17 protein.
doi:10.1371/journal.pgen.1004766.t002

(S19B Fig.). Four very similar MOB_{Q12} plasmids (around 5.2 kb) are also represented in *H*30 (two of virotype A, one of virotype C) and H22 (one of virotype D). They contain RIP and REL proteins but, as in the case of the MOB_{Qu} plasmids, no phenotype could be pointed out (S20 Fig.). MOB_{Qu} and MOB_{Q12} plasmids have received little attention because they are cryptic and remain unnoticed in most typing schemes. The present ST131 plasmidome analysis suggests they can be surprisingly abundant in *E. coli*. Finally, there were five no-MOB cryptic plasmids (four of them were 1.5 kb long and the other 5.0 kb). They all contain distinguishable Rep proteins (Rep_HTH_36_superfamily), without assignment in the PBRT method. Four of them are almost identical among themselves (S21 Fig.), while the fifth (pFV9873_3) was unique and unrelated to any reference. A detailed analysis of MOBF11/IncN plasmid family is detailed in S22 Fig.

Discussion

There are two aspects of this work that will focus the discussion. On one side, the applicability, usefulness and limitations of PLACNET will be discussed. On the other, the plasmidome of *E. coli* ST131 genomes that were reconstructed by PLACNET will be analyzed as an example of the applicability of the method. Analysis of the individual reconstructed plasmids, meant for plasmid specialists, is expanded in S1 Text.

Bacterial genomes and plasmid reconstructions

Most bacterial genomes contain more than one physical unit of DNA. Besides the main chromosome, some bacteria contain additional chromosomes and most contain plasmids. We propose that PLACNET can be used as a new method to analyze bacterial genomes. It allows the assignment of chromosomes and plasmids as separate physical units within a genome. Visual representation of the network in Cytoscape, in which plasmids appear as constellations in a starry sky, allows user-friendly apprehension of that genome constitution. We applied PLACNET in this work to analyze the plasmidome of *E. coli* ST131 genomes, but it has been shown to work also for a series of prototypic bacterial genera with different GC content and genome architecture, such as *Salmonella*, *Klebsiella*, *Agrobacterium*, *Staphylococcus* or *Bacillus*. As an example, the PLACNET representation of the genome of *Staphylococcus aureus* strain 118 (ST772) (GenBank acc number AJGE00000000) is shown in S23 Fig. PLACNET scope of

application also includes multi-chromosome bacteria like *Vibrio* or *Brucella*, where it correctly predicts both chromosomes present in these species. One *Vibrio cholerae* Pacini 1854 genome (Bioproject ID: PRJEB2215) is shown in S24 Fig. as an example. Once contigs belonging to each plasmid are defined, classical plasmid analysis ensues, as explained in the Results section. Contigs selected as part of a single plasmid are taken together and its overall proteome used to build a clustering dendrogram with reference plasmids present in the network. The dendrogram tree gathers plasmids according to the number of homologous proteins they share, providing an indication on prototype plasmids closely related with the query plasmid. There are two issues in PLACNET analysis that require additional work and for which additional improvement can be expected:

Unassigned contigs and reference sets. After plasmid reconstruction, occasionally, one or a few contigs may remain unassigned. In the set of ST131 genomes analyzed in this work, there were only two unassigned contigs (>200 bp), both in E61BA (Fig. 2). The fact that only two unassigned contigs appeared in the analysis of eight *E. coli* genomes suggests that this is not a quantitatively important problem. As could be expected, unassigned contigs are more frequent in genomes for which there are fewer references available. The lack of a suitable reference set results in poor quality clustering and an increased fraction of contigs without references. It is obvious from the preceding discussion that bacterial taxons for which not enough references exist will be more problematic for plasmid reconstruction. Thus, any such project should start by the generation of a sufficiently ample set of plasmid references. In this respect, *E. coli* constitutes probably the best choice, due to the large reference set available.

Repeat sequences and difficult plasmids. Usually PLACNET works well because contigs belonging to individual plasmids pair with different selected references and thus cluster in disjoint connected components in the Cytoscape representation after a single pruning step. The pruning step consists in identifying the bridging contigs (network hubs) as repeat sequences (RS). Two sets of evidence were used: (i) homology to known ISs or transposons, and (ii) existence of three or more scaffold links. Contigs fulfilling these two criteria were assumed to be in fact repeated in the connected network. Thus, the pruning operation consisted of duplicating the alluded nodes and splitting their scaffold links. In the tested set of *E. coli* genomes that were used to validate PLACNET (the set of eight ST131 genomes analyzed here, the 32 genome set analyzed by de Been et al. (2014) < submitted to Plos Genet together with this work >, plus another set of 10 other ESBL genomes obtained from clinical strains of bioprojects PRJNA186205 and PRJNA202876), there was only one case in which RS pruning operation was not sufficient to obtain disjoint components. It was the case of genome E35BA, where two coexisting MOB_{F12}/IncF plasmids (pE35BA_2 and pE35BA _3) could be inferred, but repeated pruning did not result in disjoint components. The evidence for the existence of two plasmids was the finding of two sets of contigs containing REL and other plasmid backbone genes. PLACNET failed in discriminating both plasmids probably because network links to references were interlocking, since several PLACNET-selected references established best hits to different components of each set. Besides, the assembly program could not distinguish among parts of both sequences and considered them as RSs. Closely related plasmids that coexist in a given cell poses the most serious problem we encountered in the application of PLACNET.

The ST131 plasmidome

HGT plays a critical role in shaping bacterial lineages, especially those of multi-environment opportunistic pathogens. Comprehensive characterization of plasmidomes has been impeded by methodological limitations, although they are essential for multilevel population genetics analysis, an approach necessary to explain selection and diversification of bacterial populations and to understand the reservoir dynamics of antibiotic resistance and virulence genes [69]. The application of PLACNET to ST131 genomes allowed the detection of emerging plasmid variants, important for the evolutionary history of this ExPEC lineage, which constitutes an outstanding example of a "high risk clonal complex", a concept increasingly important in Public Health [69].

Plasmidome description. We describe a remarkable heterogeneity of plasmids among the *E. coli* ST131 genomes analyzed, with the identification of 39 plasmids to add to the 11 plasmids in the ST131 lineage already sequenced (Table 1). Interestingly, these plasmids encompass 8 out of 17 main MOB plasmid groups found in the whole class of γ-proteobacteria [11,53], namely F12 (IncF), P3 (IncX), P5 (ColE1), P6 (IncI2), P11 (IncP), P12 (IncI/K/BO), Q12 (Rep_pSC101-like), Qu (Rep_pMG828-2/IGWZ12-like), several of them undetectable by PBRT. Besides conjugative or mobilizable plasmids, there were other plasmids, lacking REL, but identifiable by their RIP proteins. An in-depth analysis of each plasmid group identified in this study has been diverted to a Supplementary Discussion (though exciting for clinical epidemiology or plasmid biology, it is outside the mainstream goal of this work). Our findings substantially enlarge the repertoire of plasmids identified among *E. coli* ST131 isolates, which now reflect a genome widely open for plasmid infection. It is of note that this scenario has also been described for *E. coli* clones of different pathovars [47,70–73]. Comparative genomics of the few *E. coli* lineages comprehensively analyzed to date suggests that this species is a generalist able to colonize and infect humans. It also suggests that phages and plasmids make an important contribution to specialization by accessorizing the genome with new adaptive traits and tools that modify genome structure and, eventually, by modifying transcriptional regulation [70,71,74,75]. It should also be emphasized that almost all available studies on ST131, included this one, focused on strains involved in the spread of antibiotic resistance genes, which constitute, undoubtedly, a biased fraction of the ST131 plasmidome and thus preclude an accurate view of its evolutionary history [76] (see also below).

Plasmids and *E. coli* diversification. Specific ExPEC lineages have scarcely been analyzed in the context of multilevel population genetics with the exception of punctual cases involving clonally unrelated isolates [70]. A recent phylogenomic analysis of 95 ST131 isolates from different geographical areas identified the same three clusters studied in the present work [16]. This analysis concluded that point-mutations and recombination events associated with diverse MGEs, including prophages and genomic islands, determined the diversification of this ExPEC lineage. However, the diversity of plasmids was only inferred by searching for incompatibility regions based on PBRT schemes. The role of plasmids in genomic versatility were not further analyzed [16]. Even though our study analyzed a smaller number of isolates, some observations can be drawn about the role of plasmids in the diversification of the ST131 lineage.

The rate of mutagenesis of *E. coli* has been roughly estimated in one mutation per genome per year [77]. Although this number is no doubt controversial, such study and those addressing the role of recombination in the ST131 lineage add context to understand its evolution as represented in Fig. 1. Compared to the limited

sequence divergence among the ST131 core genomes (the ST43/*H*30 branch includes just about 600 SNPs), plasmids represent a very active fraction of ST131 adaptive evolution, as can be concluded from the analysis of Table 1. Such plasmid variability suggests that independent plasmid acquisitions and losses frequently occur within and between ST131 sublineages. Within the *H*30 cluster, the presence of antibiotic resistance plasmids is remarkable, specially the identification of structurally similar F2:A1:B- plasmids carrying genes conferring antibiotic resistance, since early acquisition of $bla_{\text{CTX-M-15}}$, mainly associated with F2 plasmids, is considered a key event in the selection of the ST131 cluster C/*H*30 subclone [13,16,17]. The modular structure of F2 plasmids, containing multiple copies of ISs, facilitates gene rearrangements and the interchange of antibiotic resistance platforms linked to resistance to first line antibiotics between plasmids of the same and different families. This notion, inferred from our present analysis, has already been proposed [17,78–80] and is of great concern nowadays because of the increasing risk of encountering *E. coli* isolates carrying bla_{KPC} of bla_{NDM} genes predominant in *Klebsiella* (http://www.cdc.gov/drugresistance/threat-report-2013/) [81]. Beyond F2 plasmids, the presence of other plasmid groups (N, I1/K/BO, I2, A/C, X) carrying antibiotic resistance genes is observed at variable rates in this and other studies, clearly influenced by local ecology. Most of these antibiotic resistant non-F plasmids occur only rarely in *E. coli* isolates [82]. This could be due to an intrinsic lack of fitness of these plasmids in *E. coli* under natural conditions. Alternatively, they could represent cryptic indigenous plasmids now identifiable because of the acquisition of antibiotic-resistance cassettes. Nevertheless, the acquisition of mosaic regions carrying multiple antibiotic resistance genes by broad host plasmids (e.g. N, I2, A/C) increases the risk to spread resistance to first line antibiotics to different bacterial species in and outside hospitals [79,83]. Besides antibiotic resistance plasmids, an outstanding finding of this work was the frequent detection other plasmid groups, generally considered cryptic, that are clearly underrepresented in previous ST131 studies, as ColE1, MOB_{Q}, and phage-like Rep3 plasmids. All of them are highly heterogeneous plasmid groups able to acquire adaptive traits or contribute to the mobilization of other plasmids (*See supplementary dataset for details*).

Members of the ST131 plasmidome such as MOB_{F12}/IncF, MOB_{H}/IncA/C and Rep3/phage-like plasmids can also shape the *E. coli* chromosome by facilitating mobilization in trans of genetic islands or integrating new genetic material [32,70,84]. Interestingly, recombination of large chromosomal regions occurring at the sites of insertion of either prophages or transferable genomic islands seems to have contributed to the split of the ST131 lineage in different clusters [16]. Although experimental studies on ST131 did not yet associate plasmids with genome structure, the hypothesis is plausible taking into account the frequency at which such events occur in other B2 *E. coli* populations.

On a more general note, our study identifies antibiotic resistance plasmids, which are favored in high density environments, such as the human gastrointestinal tract, and under antibiotic selective pressure, that are predominant in hospitals [85] together with phages (or phage-like cryptic plasmids), apparently predominant in low density environments [86,87], and other cryptic plasmids (frequently very small and devoid of any possible adaptive gene). This mixed constitution, which is difficult to understand on purely selective grounds, highlights potential roles of plasmids in the context of multilevel selection, a recurrent issue in evolutionary biology. The plasmid flux in ST131 strains occurs while disseminating genes coding for resistance to extended spectrum beta-lactamases (ESBL). The results presented here find a complement in the study of de Been *et al*. (accompanying paper), where the authors document the dissemination of ESBL-carrying epidemic plasmids from animal to human clonally unrelated *E. coli* lineages. Thus, many plasmids appear in a clonal lineage, and many lineages can be infected by a single predominant plasmid. These conceptual notions have relevance in Public Health as they deal with the hierarchical units of selection that contribute to increase the population size of antibiotic resistance genes in human and animal pathogens [69,88,89].

In summary, our study reveals the utility of PLACNET in multilevel population genetics analysis, critical to understand the evolutionary processes and dynamics of both bacterial and plasmid lineages. Its application to *E. coli* ST131 allowed us to infer the roles of plasmids in the dissemination of globally spread antibiotic resistance and virulencegenes, some of them being underrepresented in Genbank. It is probable that these plasmids are critically relevant to understand the adaptive evolution of *E. coli* populations and their bacterial exchange communities. Armed with this new tool for plasmid analysis, future scrutiny of a larger number of significant strains will allow us to understand the interplay among different plasmid associations that often appear in bacterial pathogens.

Conclusion

The evolutionary processes of main bacterial pathogens are often discussed in the context of lineage-associated acquisition of a specific virulence gene set. The present study demonstrates how *E. coli* ST131 strains, even when they are practically identical in their core genomes, contain a striking variety of different plasmids. Many of them remain unnoticed, since they are apparently cryptic. Prevalent plasmids, such as IncFs, undergo frequent recombination, continuously resulting in novel gene repertoires. Our results shed light on the role of plasmids in *E. coli* ST131 evolution. Horizontal transmission of plasmids that carry not only antibiotic resistance and virulence genes, but also other poorly analyzed functions (metabolic genes, colicins and as yet cryptic functions) is common in the ST131 plasmidome and results in frequent and rapid adaptive changes. Arrival to these conclusions has been made possible by the application of PLACNET, a plasmid reconstruction method for WGS datasets.

Materials and Methods

Epidemiological background of bacteria and plasmids

Comprehensive plasmidome analysis was carried out for 10 *E. coli* ST131 genomes, representing main ST131 sublineages described to date [13,16]. They include strains coming from Spain (three *fimH*30, one *fimH*324), USA (three *fimH*30), Australia (one *fimH*30), Denmark (one *fimH*22) and Japan (one *fimH*41). The *fimH*30 strains from Spain were CTX-M-15 producers and belonged to the *H*30-Rx sublineage (additionally, one strain was also CTX-M-14), while those collected in the USA were KPC-2 producers. The four strains from Spain represent predominant ST131 variants on the basis of PFGE patterns and the presence/absence of four putative virulence markers (*afaFM955459*, encoding an Afa/Dr adhesion; *sat*, secreted autotransporter toxin; *ibeA*, invasion of brain endothelium; and *iroN*, salmochelin siderophore receptor) [36,51,52] and sequenced for this work.

The ST131 isolates studied represent epidemic variants exhibiting particular combination of putative virulence traits and were previously designed as distinct "virotypes" by capital letters A to D [36]. It should be noted that no correlation exists between

Table 3. Human *E. coli* ST131 genomes analyzed in this work.

Strain	Accession[a]	Location	Collection date	Isolation source	Plasmid name (Accession number)[b]	Reference
HVH-177	PRJNA186205	Denmark	2003	Blood	pHVH177_1	PRJNA186413
BIDMC20B	PRJNA202031	USA	-	Urine	pBIDMC20B_1and _2	PRJNA202876
BWH24	PRJNA201983	USA	-	-	pBWH24_1 to _3	PRJNA202876
BIDMC38	PRJNA202050	USA	2012	-	pBIDMC38_1 to _5	PRJNA202876
FV9873	PRJEB6262	Spain	2007	Urine	pFV9873_1 to _6	This study
E35BA	PRJEB6262	Spain	2008	Urine	pE35BA_1 to _3, IME_E35BA	This study
E2022	PRJEB6262	Spain	2006	Urine	pE2022_1 to _5	This study
E61BA	PRJEB6262	Spain	2008	Abscess	pE61BA_1 to _7	This study
SE15	AP009378	Japan	-	Feces	pECSF1 (AP009379)	[91]
JJ1886	NC_022648.1	USA	-	Urine	pJJ1886_1 to _5 (NC_022649; NC_022650; NC_022651; NC_022661; NC_022662)	[90]

[a]PRJNA and PRJEB6262 accession numbers correspond to SRA datasets. AP009378 and NC_022648.1 correspond to finished genomes.
[b]Plasmids derived from this study are named according to Table 1.
doi:10.1371/journal.pgen.1004766.t003

these "virotype" designations and "ST131 strain designation" in other studies that also used capital letters to distinguish among ST131 clonal variants [16]. Other genome datasets were taken either from Bioproject NCBI database (https://www.ncbi.nlm.nih.gov/bioproject/), or from NCBI genomes database (*E. coli* JJ1886 [90] and *E. coli* SE15 [91]). In addition, fully sequenced plasmids pEK499, pEK204, pEK516, pJIE186-2 and pJIE143, previously found in other ST131 isolates [56,92–94], were used for plasmid comparisons. Information about all genomes is detailed in Table 3.

DNA sequencing

Total DNA from *E. coli* ST131 strains FV9873, E35BA, E2022 and E61BA was extracted with QIAmp DNA Mini Kit (Qiagen). DNA concentration was measured with Nanodrop 2000 (Thermo Scientific) and Qubit 2.0 Fluorometer (Life Technologies). 1.0 μg DNA was sonicated (20 cycles of 30 s at 4°C, low intensity) with Bioruptor Next Generation (Diagenode). Sample quality was checked in a Bioanalyzer 2100 (Agilent Technologies). DNA samples were preconditioned for sequencing by using the TruSeq DNA Sample Preparation Kit (Illumina) and quantified with Step One Plus Real-Time PCR System (Applied Biosystems). Flow-cells were prepared with TruSeq PE Cluster Kit v5-CS-GA (Illumina). Sequencing was carried out using a standard 2×71 base protocol (300-400 bp insert size) in a Genome Analyzer IIx (Illumina, San Diego, CA) at the sequencing facility of the University of Cantabria. The main statistics of the eight sequence datasets analyzed are shown in Table 4.

Phylogenetic analysis of the ST131 core genome

The ST131 core genome was defined as the collection of genes present in the ten ST131 genomes analyzed, with more than 90% similarity and 90% coverage. CD-HIT-EST [95] was used to cluster genes. A homemade Perl script was created to parse the cluster and define the core genome set. All core genes were concatenated and aligned with progressive Mauve [96]. A tabular list of SNPs was extracted from the Mauve alignment by applying the SNP export tool of Mauve GUI. The tabular list of polymorphic sites was parsed by a homemade script. A given position was counted as an SNP if it varied between two given sequences. The number of SNPs was added for each pair of strains to give the final SNP count. Polymorphic sites with gaps were removed from the SNP count matrix. The Mauve alignment was curated by trimAl [97]. RAxML [98] was used to build the core genome phylogenetic tree, using 100 replicates for bootstrap determination.

Plasmid Constellation Networks (PLACNET)

PLACNET was developed to associate contigs with specific physical DNA units in WGS experiments. Networks are powerful models that allow visualization and analysis of sequence information. PLACNET networks are composed of two types of nodes (contigs and reference genomes) and two types of edges (similarity to reference sequences and scaffold links). Commonly, network layout algorithms simulate repulsion forces between nodes and attraction forces by the edges that link two nodes. Thus, node distribution in the network will depend on the intensity of forces that define the edges. In such network model, a plasmid will be represented by a connected component (a set of linked nodes) or, in other words, a constellation of contigs. Different physical units (plasmids and chromosomes) should be represented by disjoint connected components (separate constellations). The workflow (Fig. 6) involves the following steps:

Assembly. Velvet assembly software [99] and its script VelvetOptimiser.pl were used to determine the best assembly and to scan the optimum parameters. Velvet provides also coverage information for each contig, which adds useful information for network interpretation.

Scaffold links. Although assembly programs perform scaffolding between contigs, when the assignation is ambiguous, contigs remain unbound. A method based in the mapping tool Bowtie 2 [100] was used to find all possible scaffold links. All reads were mapped using the contigs as references, using default parameters of Bowtie 2, with the option to report all possible hits. The output file was converted to SAM format [101] to give the potential adjacency information for each contig. We considered as potential PLACNET scaffold links those which comply with two rules: (i) the contigs were paired at their extremities, themselves defined as twice the read length, and (ii) the number of pair-end reads linking those two contigs had to be higher than one third of the mean of the total pair-end reads that scaffold all

Table 4. Genomes assembled in this study.

Strain	ID	N° Libraries	Read length	N°Contigs	Total bp	N50	Kmer
HVH177	SRS399685	2	101	81	5035548	242711	83
BIDMC20B	SRS420795	3	101	115	5311918	209342	91
BWH24	SRS420803	2	101	114	5369063	192138	83
BIDMC38	SRS420798	4	101	135	5226831	190988	91
FV9873	ERS450218	3	71	262	5160060	153685	55
E35BA	ERS450219	3	71	419	5243070	159702	57
E2022	ERS450220	2	71	346	5296607	159635	53
E61BA	ERS450221	3	71	246	5168482	198396	57

doi:10.1371/journal.pgen.1004766.t004

contigs. This procedure was implemented as an in-house Perl script.

Reference search. At least for bacteria widely covered by sequencing projects, most contigs in any new sequence are similar to one or more previously published sequences (reference sequences). Our initial hypothesis was that, for any physical DNA unit, its contigs will "BLAST" a related set of reference sequences. Thus, in the PLACNET network representation, they will cluster around the homologous references. The more DNA databases grow, the closer the references will be to the query sequence and the better PLACNET will work. A homemade BLAST [102] database was constructed from the NCBI genomes database by joining all sequences contained in [ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/] and [ftp://ftp.ncbi.nlm.nih.gov/genomes/Plasmids/]. The version used in this work was from March 7th 2013 and contains 6,432 genomes (plasmids and chromosomes). Megablast search of all contigs was carried out against the homemade BLAST-genome database with the objective of selecting a few best matches for network construction. Due to the different length of each contig, fixed thresholds by e-value or score cannot be chosen. Since the score is not a normalized parameter, and varies depending on sequence length, hits were selected by applying a dynamic threshold, based on the number of homologous sequences and the score of each hit. If the threshold is defined as $85\%+2n$ of the mean of the n previous sequence scores, and n is the ranking position of sequence i retrieved by megablast with score S_i, then T_{n+1} is the threshold for sequence $n+1$:

$$T_{n+1} = \frac{\sum_{i=1}^{n} S_i}{n} \cdot (0.85 + 2n)$$

All reference sequences above the threshold were taken as nodes in the PLACNET representation.

Protein prediction of replication initiator proteins (RIP) and relaxases (REL). Some genes are indicative of a plasmid sequence. Among them we selected REL, key proteins in the conjugation process [53,103] and RIPs, key proteins in the replication of most plasmids [104]. Although not all plasmids contain a RIP and/or REL, their presence in a contig is diagnostic of a plasmid (or ICE) sequence. Some plasmids have more than one RIP (i.e. IncF family plasmids) [9,105,106] but plasmids rarely have more than one REL [107]. ORF prediction was carried out by GeneMark [108], which optimizes predictions based on GC content of DNA. The heuristic prediction implemented in this software is especially useful to predict ORFs in plasmid-containing genomes because it takes each contig individually and selects the best prediction model case by case.

To implement specific search protocols for REL and RIPs, three homemade databases (DB) were developed. A REL database (REL-DB) was constructed according to [53,109]. Similarly, RIP-DB was constructed from all RIPs annotated in UniProt database. RIP sequences were clustered by CD-HIT [95], using 40% identity as a threshold. Next, a Hidden Markov Model was built from each cluster by hmmer3 [110]. Finally, the HMM profiles were used in a search against UniProt. The HMM search and initial dataset were joined in one database (RIP-DB). An additional step was necessary to classify RIPs according to the widely used plasmid classification protocol PBRT [8]. A homemade nucleotide database (INC-DB) was created to identify PBRT types in silico using blastn. Finally, the relevant ORFs (REL and RIPs) identified to specific contigs by using REL-DB and RIP-DB

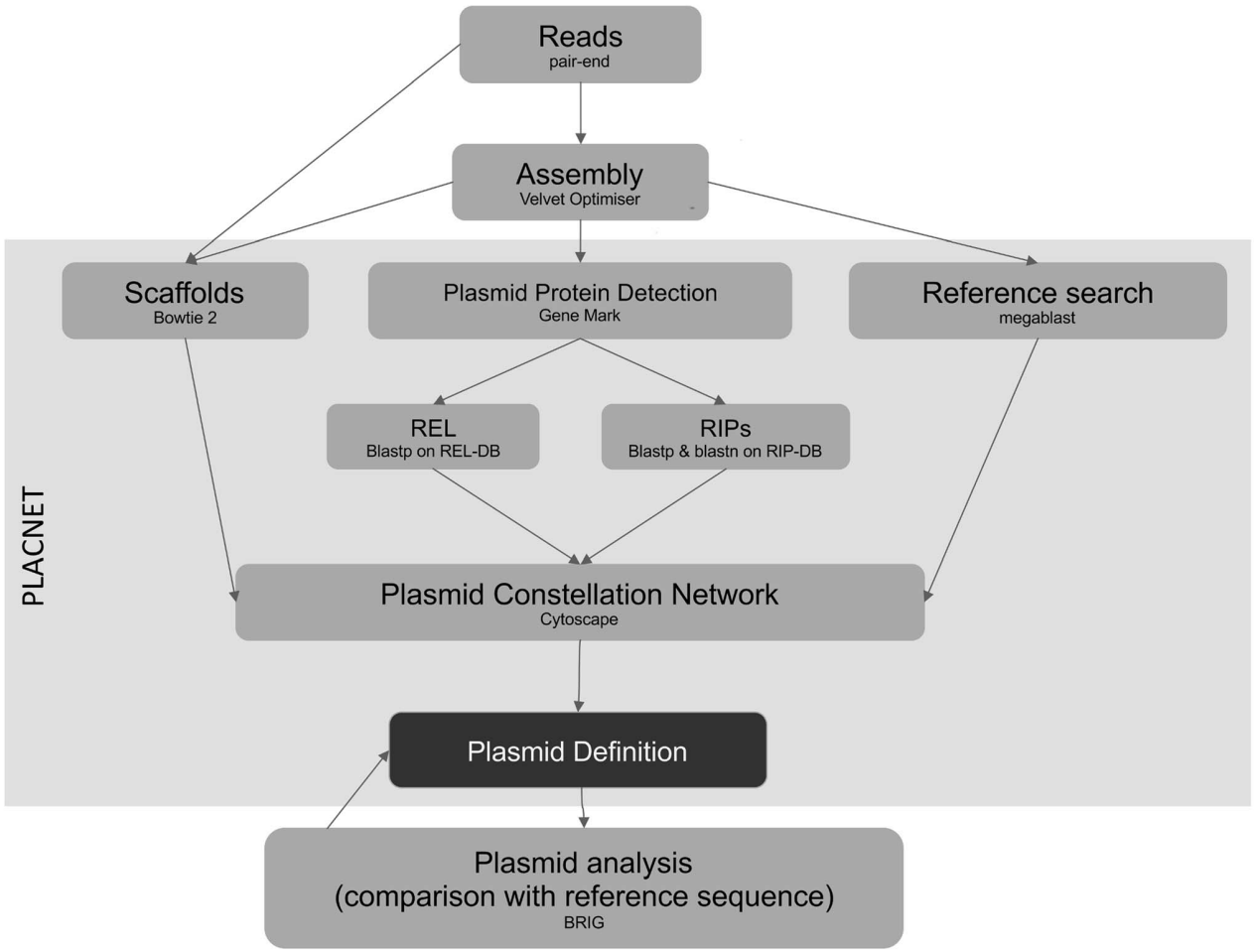


Fig. 6. PLACNET flow diagram. The diagram represents the PLACNET workflow to analyze an Illumina bacterial genome dataset. It can be separated in two sub-process: network delineation and plasmid analysis. Network delineation consists on contig assembly, determination of scaffold interactions, reference search of homologous genomes and plasmid protein prediction. Plasmid analysis basically consists in the construction of a dendrogram of plasmid protein profiles, which identifies the most relevant reference sequences, followed by plasmid cluster analysis, which compares query plasmids with its closest references. Plasmid analysis is a feedback process that helps to resolve uncertainties and results in a final definition of plasmid and chromosome content.
doi:10.1371/journal.pgen.1004766.g006

were incorporated to the network as tags. All relevant steps in network construction were implemented as a Perl script available at the following web page: http://placnet.sourceforge.net/.

Plasmid constellations. As explained above, each plasmid is represented in PLACNET by a connected component (a constellation). Thus, different physical units (plasmids and chromosomes) should be represented by disjoint (unlinked) connected components. Cytoscape software [111] was used to visualize and analyze plasmid constellations, which incorporate all the information (similarity to reference sequences, scaffold links, and protein tags) in a single network. Node attributes such as contig size, coverage and reference description, are added to the network. At this stage, network pruning is needed to resolve individual plasmids as disjoint components. When a genome has a number of repeated sequences (e.g., insertion sequences (ISs) or transposons), or two very similar plasmids, the assembly process outputs those sequences as contigs with multiple scaffold links. In the network context, they represent hubs, that is, nodes with a high number of connections. This makes the network very dense and complicates the analysis of network connected components. In PLACNET, contigs smaller than 200 bp were directly eliminated from the analysis. Hubs were examined by *blastx* against protein databases (i.e. UniProtKN or NCBI nr). If there was identity to any transposase gene, the hub was duplicated, and scaffold links were partitioned among them, to maximize the number of disjoint components. Contigs that remain unbound are classified as "unassigned sequence" in the contig assignation table.

Plasmid definition, dendrograms and cluster analysis. The final steps in plasmid reconstruction involve the definition and verification of each plasmid. This is an iterative process, as shown in Fig. 6. First, each contig was assigned to a putative plasmid (or chromosome) based on visualization of disjoint connected components in the Cytoscape representation. Assignments take into consideration additional types of evidence like the presence of REL and/or RIPs, size of the putative plasmid compared to reference plasmids and coverage of each contig (contigs belonging to the same plasmid must have similar coverage). Taking into account the information provided by related genomes within the same sequencing project can also be helpful (same sequences cluster around the same references). In

this respect, PLACNET is more robust in multi-strain collections. The performance of PLACNET was validated by testing a number of previously sequenced and annotated *E. coli* genomes (S2 Table). The ART software (Huang et al., 2012) was used to simulate pair-end Illumina reads from those genomes, which were then analyzed by PLACNET as explained above. Results are shown in S2 Text, S3 Table and S25-S34 Figs.

After PLACNET has defined the plasmids carried in the relevant genomes, the next step in plasmidome analysis is to build a dendrogram that produces a hierarchical clustering of plasmid proteomes similar to those described in [112–114]. CD-HIT (thresholds: 70% identity and 80% coverage) was used for clustering references and query plasmids. Based on the output file, a presence/absence table (present or absence of each protein cluster in each plasmid) was built. Each table row represents a plasmid protein profile. Raup-Crick distance method, implemented in *vegan* package for R software [115], was used to calculate the distance matrix of plasmid protein profiles. The Ape package [116] was used to calculate the dendrogram bootstrapping confidence value. Finally, a hierarchical clustering dendrogram was built using the UPGMA algorithm.

Putative plasmids and references belonging to the same dendrogram branch were compared using BRIG [117] or Abacas [118]. While BRIG is not sensitive to contig arrangement, Abacas can be used to order contigs according to a given reference. With these tools, the curator is able to visualize the correspondence between reconstructed plasmids and references, or can go back to dendrograms or PLACNET in the search for missing or extra contigs. This iterative mode of analysis is represented in Fig. 6 by the backward arrow linking plasmid cluster analysis with plasmid definition.

Plasmids were mainly classified according to their REL in MOB families, as described by [53]. Classical Inc families are also given when typing them by in silico PBRT was possible. Plasmids that could not be classified one way or the other were termed no-MOB by exclusion.

Supporting Information

S1 Fig PLACNET reconstruction for FV9873 genome (*E. coli* ST131/*H*30/virotype A). A total of 262 contigs were classified in chromosome and six plasmids. The black line surrounding the pFV9873_1 plasmid (4.1 bp) indicates that it is a closed plasmid. There are not unassigned contigs.
(PDF)

S2 Fig PLACNET reconstruction for E35BA genome (*E.coli* ST131/*H*30/virotype B). One 14.2 kb MOBP11 Integrative Mobilizable Element (IME) was detected in the chromosome. The pE35BA_1 plasmid (4.1 kb) is closed. There was a conflict of separation between two IncF plasmids (pE35BA_2 and pE35BA_3, total size: 211 kb). The annotation of this particular genome is limited by the quality of the assembly (many small and non-scaffolded contigs). This is the network with the highest number of contigs in our study (total: 419). No contigs remained unassigned.
(PDF)

S3 Fig PLACNET reconstruction for E2022 genome (*E. coli* ST131/*H*30/virotype C). A total of five plasmids, two of them as closed plasmids, and the chromosome were obtained in a 346-contig network. No contigs remained unassigned.
(PDF)

S4 Fig PLACNET reconstruction for E61BA genome (*E. coli* ST131/*H*324/virotype D). Seven different plasmids were obtained in the analysis. The pE61BA_2 plasmid (24.5 kb), containing a single contig, was closed. Two contigs remained unassigned, as no scaffold links were detected for them. One of them (2,953 bp) encodes for a putative DNA primase and a lytic transglycosilase, while another (1,301 bp) encodes for TrbI and TraB partial proteins.
(PDF)

S5 Fig PLACNET reconstruction for BIDMC20B genome (*E. coli* ST131/*H*30/virotype C). A total of 115 contigs were fully assigned to the chromosome and two plasmids. Plasmid pBIDMC20B_2 (109 kb) appeared as a single contig that could be closed.
(PDF)

S6 Fig PLACNET reconstruction for BIDMC38 genome (*E. coli* ST131/*H*30/virotype A). Five plasmids were detected. Three small plasmids (1.6, 4.2 and 5.3 kb) are closed. No contigs remained unassigned.
(PDF)

S7 Fig PLACNET reconstruction for BWH24 genome (*E. coli* ST131/*H*30/virotype C). Three plasmids were detected, only one of them as a closed plasmid (pBWH24_2, 109 kb). No contigs remained unassigned.
(PDF)

S8 Fig PLACNET reconstruction for HVH177 genome (*E. coli* ST131/*H*324/virotype D). Only one plasmid (pHVH177_1, 78.6 kb) composed by three contigs was detected in the HVH177 genome. No contigs remained unassigned.
(PDF)

S9 Fig Contig coverage in E61BA genome (*E. coli* ST131/*H*324/virotype D). Contigs belonging to each plasmid are shown in different colors, according to the code below the histogram. Plasmid copy numbers are inferred from their contig coverage. Average coverage of chromosomal contigs was 56. All contigs with >2X average are named according to their predicted gene products. When there is more than one contig, annotations are separated by colons, if adequate.
(PDF)

S10 Fig Contig coverage in E35BA genome (*E.coli* ST131/*H*30/virotype B). Contigs belonging to each plasmid are shown in different colors, according to the code below the histogram. Plasmid copy numbers are inferred from their contig coverage. Average coverage of chromosomal contigs was 55. All contigs with >2X average were named according to their predicted gene products. If more than one contig coincided within the same coverage section, annotations of the individual contigs were separated by colons. As is shown in the histogram, contigs corresponging to the MOBF12/IncF plasmids pE35BA_2+3 (in green color) show similar coverage than the chromosome, which indicates that both IncF plasmids have the same copy number than the chromosome.
(PDF)

S11 Fig BRIG comparative analysis of MOBF12/IncF plasmids. These plasmids were subdivided in four groups according to Fig. 5 inset. **S11A**: Group I. Plasmid pJIE186-2 is used as the reference for the BRIG comparison. **S11B**: Group II. Plasmid pJJ1886-5 is the inner reference. **S11C**: Plasmid F is used as a reference. **S11D**: Plasmid pECSF1 is used as a reference.
(PDF)

S12 Fig Comparative analysis of phage-related/RepFIB plasmids. **S12A**: BRIG representation using the 111kb plasmid

pECOH89 as reference. **S12B**: Phylogenetic analysis of RepFIB family of RIP proteins. RaxML software (v.7.2.8) was used to infer the Maximum Likelihood tree and MEGA5.2.2 to represent the result. Bootstrap values for 100 replicates are indicated. The tree was rooted with the RepFIB protein of the IncN plasmid N3.
(PDF)

S13 Fig BRIG comparative analysis of MOBP12/IncI-complex. **S13A**: The IncI1 plasmid pEK204 is used as inner ring in the BRIG analysis. **S13B**: Plasmid pCT [62] was used as reference.
(PDF)

S14 Fig Comparative analysis of MOBP6/IncI2 plasmids. **S14A**: Phylogenetic tree of MOBP6 REL proteins, calculated as in S12B Fig. The tree was rooted with MOBP6 REL of Plasmid2 from *Nitrosomonas eutropha* C91. **S14B**: BRIG comparative analysis of pBWH24_3 plasmid, using pChi7122_3 as reference. **S14C**: BRIG comparative analysis of pE61BA_7 plasmid, using pO157_Sal as inner reference.
(PDF)

S15 Fig Comparative analysis of MOBP3/IncX plasmids. **S15A**: Phylogenetic tree of MOBP3 REL proteins, calculated as in S12B Fig. The tree was rooted with VirD2_pSD25 (MOBP2 subfamily). ST131 plasmids are shown in red. IncX subgroups are indicated in different color backgrounds. **S15B**: BRIG comparative analysis of IncX1 plasmids using p2ESCUM as a reference. **S15C**: BRIG comparative analysis of IncX4-like plasmids using pSH696_34 as reference.
(PDF)

S16 Fig Comparative analysis of MOBP11/IncP plasmids. **S16A**: Phylogenetic tree of MOBP11 REL proteins, calculated as in S12B Fig. The tree was rooted with NikB_R64 (MOBP12 subfamily). ST131 plasmids are colored in red. Two clearly separated groups are colored. **S16B**: BRIG comparative analysis of IncP1 plasmids using pJJ1886_4 as a reference. **S16C**: Comparison of JJ1886 and E35BA genomes, showing the genetic map of the inserted IME_E35BA, and its homology to *Bukholderia glumae* IncP island. The figure was drawn with EasyFig [75]. Specific genes are specifically colored according to the code in the lower part of the figure.
(PDF)

S17 Fig Comparative analysis of MOBC12 plasmid. **S17A**: Phylogenetic tree of MOBC12 REL proteins, calculated as in S12B Fig. The tree was rooted with MobC_CloDF13 (MOBC11 subfamily). **S17B**: BRIG comparative analysis of MOBC12 plasmids using pCRY as a reference.
(PDF)

S18 Fig Comparative analysis of MOBP5/ColE1-like plasmids. **S18A**: Phylogenetic tree of MOBP5 REL proteins, calculated as in Fig SF12B. **S18B, C and D**: BRIG comparative analysis of MOBP5 plasmids using ColE1 (SF18B), pJJ1886_3 (SF18C) and pColK-K235 (SF18D) as references.
(PDF)

S19 Fig Comparative analysis of MOBQu plasmids. **S19A**: Phylogenetic tree of MOBQu REL proteins, calculated as in Fig SF12B. The tree was rooted with the MOBQu2 subfamily. ST131 plasmids are colored in red. Different color backgrounds are used to represent MOBQu1, where ST131 MOBQu plasmids are located, and MOBQu2 branches. **S19B**: BRIG comparative analysis of MOBQu plasmids using pSE11-6 as a reference.
(PDF)

S20 Fig Comparative analysis of MOBQ12 plasmids. **S20A**: Phylogenetic tree of MOBQ12 REL proteins, calculated as in Fig SF12B. The tree was rooted with MobA_RSF1010 (MOBQ11 subfamily). **S20B**: BRIG comparative analysis of MOBQ12 plasmids using pCE10B as a reference.
(PDF)

S21 Fig Comparative analysis of small no-MOB plasmids. BRIG comparative analysis of no-MOB plasmids using pCE10D as a reference.
(PDF)

S22 Fig Comparative analysis of MOBF11/IncN plasmids. **S22A**: Phylogenetic tree of MOBF11 REL proteins, calculated as in Fig SF12B. The tree was rooted with R388 (MOBF11/IncW). IncN1 and IncN2 subgroups are indicated with different background colors. **S22B**: BRIG comparative analysis of IncN1 plasmids, using R46 as a reference. **S22C**: BRIG comparative analysis of IncN2 plasmids using p271A as a reference.
(PDF)

S23 Fig PLACNET reconstruction of the genome of *Staphylococcus aureus* strain 118 (ST772) (ID: PRJNA82607). Assembly data: Number of libraries: 1; read length: 75 bp; number of contigs: 73; total bp: 2,798,022 bp; N50: 224673 and Kmer: 73. One 12,819 bp plasmid was identified and reconstructed. No REL or RIP proteins were detected.
(PDF)

S24 Fig PLACNET reconstruction for a genome of *Vibrio cholerae* Pacini 1854 (ID: PRJEB2215). Assembly data: Number of libraries: 1; read length: 75 bp; number of contigs: 149; total bp: 4,022,287 bp; N50: 180089 and Kmer: 65. The two *V. cholerae* chromosomes were fully reconstructed. Chromosome I (2,992,142 bp in our study) was reported to be about 3 million bp and encodes most essential functions [95]. As shown in Fig S24, it harbors a MOBH12 relaxase, identical to that of the integrative and conjugative element (ICE) of the SXT/R391 family [96]. Chromosome II is smaller (1,034,286 bp in our study). Finally, a 26.3 kb plasmid containing a RIP protein (87% identity with *E.coli* pABU plasmid [97]) could be reconstructed.
(PDF)

S25 Fig Cytoscape representation of the reconstructed *E. coli* JJ1886 genome. The network was constructed and codes used (in this and following figures) as explained in Fig. 6. The pruned network (Step 1) was obtained after deleting 19 contigs smaller than 200 bp.
(PDF)

S26 Fig Definition (Step 2) of *E. coli* JJ1886 plasmids p1 to p4. These plasmids contain a RIP and/or REL protein and appeared as single contigs. The inset Table shows some properties of relevant nodes. The nodes are represented in the network surrounded by circles of the same color than the background color in the Table.
(PDF)

S27 Fig Resolution of hubs and definition of plasmid p5 (Step 3). Hub nodes that were duplicated are shown in the inset Table and indicated by red arrows in the Cytoscape network. After hub duplication, the IncF plasmid p5 is shown to contain the 12 contigs surrounded by a red circle.
(PDF)

S28 Fig BRIG comparison of the reconstructed plasmid p5 with the reference plasmid pJJ1886_5. The reference plasmid (inner ring) is compared to the p5 reconstructed plasmid (purple ring).

Outer black and white ring sectors represent pJJ1886_5 gene annotations.
(PDF)

S29 Fig Inverse BRIG comparison of reconstructed plasmid p5 vs pJJ1886_5 reference plasmid. The reconstructed plasmid is placed here as the reference inner ring (thin black circle line) to which the reference plasmid (purple ring) and the reconstructed p5 contigs (outer blue and red ring) are compared.
(PDF)

S30 Fig Cytoscape representation of the reconstructed *E. coli* SE15 genome. The network was constructed and codes used as explained in Fig. 6. The pruned network (Step 1) was obtained after deleting 16 contigs smaller than 200 bp.
(PDF)

S31 Fig Plasmid definition (PLACNET steps 2 and 3) of *E. coli* SE15 genome. Three particular nodes in the pruned network (surrounded by the red circle) were scrutinized due to their loose connection to the chromosome. As shown in the inset Table (red background files), a blastx comparison indicates they correspond to "typical" *E. coli* chromosomal segments, so were finally assigned to the chromosome. Three other nodes (surrounded by a green circle in the left Cytoscape representation) corresponded to hubs (green background in the inset) and were thus duplicated. The reconstructed plasmid (p1) in the final network is surrounded by a purple ring.
(PDF)

S32 Fig BRIG comparison of SE15 *E. coli* genome reconstructed IncF plasmid (p1) with the reference plasmid pECSF1. The reference plasmid (inner ring) is compared to the IncF reconstructed plasmid (purple ring). Outer black and white ring sectors represent pECSF1 gene annotations. Three regions (5,709 bp in total) were missing from the p1 reconstruction.
(PDF)

S33 Fig Cytoscape representation of the reconstructed genome of *E. coli* strain MG1655 containing plasmids pEC958 and R46. The network was constructed and codes used as explained in Fig. 6. The pruned network was obtained after deleting 25 contigs smaller than 200 bp and duplicating 2 hubs (surrounded by a red circle). Plasmid p1 is the reconstructed R46 while p2 is the reconstructed pEC958. Nodes surrounded by a blued circle, and described by the blue background files in the inset Table, could not be assigned. See text for further details.
(PDF)

S34 Fig Final Cytoscape representation of reconstructed H0407 *E. coli* genome. The network was constructed as explained in Fig. 6. The pruned network was obtained after deleting 44 contigs smaller than 200 bp. Plasmids p1 and p2, represented by single contigs, are surrounded by red and blue circles, respectively. A single contig (surrounded in an intense blue circle) remained isolated from other genetic units. Nevertheless, blastx analysis demonstrate it correspond to chromosomal background, as shown in the inset Table. Nodes described by the grey background files in the inset Table correspond to one node assigned to the chromosome, plus seven hubs, which were duplicated. The green circle surrounds 52 contigs, adding 119,735 bp, which represent plasmids p3 and p4, the two IncF plasmids that PLACNET was not able to resolve (see text for further details). The red arrow indicates a node containing REL, RIP and backbone genes in common to both IncF plasmids.
(PDF)

S35 Fig Hierarchical clustering dendrogram of ST131 plasmids and relevant references. The left dendrogram shows the complete tree, with references. Dendrogram construction and color codes are as in Fig. 4. The right dendrogram expands the MOB_{F12}/IncF branch, with new background colors highlighting plasmid groups within this branch that are mentioned in the text.
(TIF)

S1 Table Complete list of contigs assigned to each plasmid or chromosome.
(XLS)

S2 Table Ten *E. coli* genomes analyzed as examples of PLACNET performance.
(PDF)

S3 Table PLACNET performance on a set of ten *E. coli* genomes.
(PDF)

S4 Table Molecular size of the plasmids estimated by S1-PFGE and PLACNET reconstruction in the four strains sequenced in this study.
(PDF)

S1 Text The plasmidome of *E. coli* ST131.
(PDF)

S2 Text PLACNET validation by reconstruction of finalized genomes.
(PDF)

Acknowledgments

The authors thank Carlos Revilla for expert technical assistance.

Author Contributions

Conceived and designed the experiments: VFL FdlC. Performed the experiments: VFL MdT AM MPGB. Analyzed the data: MdT VFL TMC FdlC MPGB. Contributed reagents/materials/analysis tools: JB. Wrote the paper: MdT VFL AM JB TMC FdlC MPGB.

References

1. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW (2012) Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet 13: 601–612.
2. Johnson TJ, Nolan LK (2009) Pathogenomics of the virulence plasmids of Escherichia coli. Microbiol Mol Biol Rev 73: 750–774.
3. Carattoli A (2011) Plasmids in Gram negatives: molecular typing of resistance plasmids. Int J Med Microbiol 301: 654–658.
4. Carattoli A (2013) Plasmids and the spread of resistance. Int J Med Microbiol 303: 298–304.
5. Ahmed SA, Awosika J, Baldwin C, Bishop-Lilly KA, Biswas B, et al. (2012) Genomic comparison of Escherichia coli O104:H4 isolates from 2009 and 2011 reveals plasmid, and prophage heterogeneity, including shiga toxin encoding phage stx2. PLoS One 7: e48228.
6. Boerlin P, Chen S, Colbourne JK, Johnson R, De Grandis S, et al. (1998) Evolution of enterohemorrhagic Escherichia coli hemolysin plasmids and the locus for enterocyte effacement in shiga toxin-producing E. coli. Infect Immun 66: 2553–2561.
7. Couturier M, Bex F, Bergquist PL, Maas WK (1988) Identification and classification of bacterial plasmids. Microbiol Rev 52: 375–395.
8. Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL, et al. (2005) Identification of plasmids by PCR-based replicon typing. J Microbiol Methods 63: 219–228.
9. Villa L, Garcia-Fernandez A, Fortini D, Carattoli A (2010) Replicon sequence typing of IncF plasmids carrying virulence and resistance determinants. J Antimicrob Chemother 65: 2518–2529.
10. Garcia-Fernandez A, Fortini D, Veldman K, Mevius D, Carattoli A (2009) Characterization of plasmids harbouring qnrS1, qnrB2 and qnrB19 genes in Salmonella. J Antimicrob Chemother 63: 274–281.

11. Alvarado A, Garcillan-Barcia MP, de la Cruz F (2012) A Degenerate Primer MOB Typing (DPMT) Method to Classify Gamma-Proteobacterial Plasmids in Clinical and Environmental Settings. PLoS One 7: e40438.
12. Edwards DJ, Holt KE (2013) Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microb Inform Exp 3: 2.
13. Price LB, Johnson JR, Aziz M, Clabots C, Johnston B, et al. (2013) The epidemic of extended-spectrum-beta-lactamase-producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-Rx. MBio 4: e00377–00313.
14. Johnson JR, Clermont O, Johnston B, Clabots C, Tchesnokova V, et al. (2014) Rapid and specific detection, molecular epidemiology, and experimental virulence of the O16 subgroup within Escherichia coli sequence type 131. J Clin Microbiol.
15. Blanc V, Leflon-Guibout V, Blanco J, Haenni M, Madec JY, et al. (2014) Prevalence of day-care centre children (France) with faecal CTX-M-producing Escherichia coli comprising O25b:H4 and O16:H5 ST131 strains. J Antimicrob Chemother.
16. Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, et al. (2014) Global dissemination of a multidrug resistant Escherichia coli clone. Proc Natl Acad Sci U S A.
17. Coque TM, Novais A, Carattoli A, Poirel L, Pitout J, et al. (2008) Dissemination of clonally related Escherichia coli strains expressing extended-spectrum beta-lactamase CTX-M-15. Emerg Infect Dis 14: 195–200.
18. Nicolas-Chanoine MH, Bertrand X, Madec JY (2014) Escherichia coli ST131, an Intriguing Clonal Group. Clin Microbiol Rev 27: 543–574.
19. Nicolas-Chanoine MH, Blanco J, Leflon-Guibout V, Demarty R, Alonso MP, et al. (2008) Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. J Antimicrob Chemother 61: 273–281.
20. Leflon-Guibout V, Blanco J, Amaqdouf K, Mora A, Guize L, et al. (2008) Absence of CTX-M enzymes but high prevalence of clones, including clone ST131, among fecal Escherichia coli isolates from healthy subjects living in the area of Paris, France. J Clin Microbiol 46: 3900–3905.
21. Platell JL, Johnson JR, Cobbold RN, Trott DJ (2011) Multidrug-resistant extraintestinal pathogenic Escherichia coli of sequence type ST131 in animals and foods. Vet Microbiol 153: 99–108.
22. Albrechtova K, Dolejska M, Cizek A, Tausova D, Klimes J, et al. (2012) Dogs of nomadic pastoralists in northern Kenya are reservoirs of plasmid-mediated cephalosporin- and quinolone-resistant Escherichia coli, including pandemic clone B2-O25-ST131. Antimicrob Agents Chemother 56: 4013–4017.
23. Hernandez J, Bonnedahl J, Eliasson I, Wallensten A, Comstedt P, et al. (2010) Globally disseminated human pathogenic Escherichia coli of O25b-ST131 clone, harbouring blaCTX-M-15, found in Glaucous-winged gull at remote Commander Islands, Russia. Environ Microbiol Rep 2: 329–332.
24. Pallecchi L, Bartoloni A, Fiorelli C, Mantella A, Di Maggio T, et al. (2007) Rapid dissemination and diversity of CTX-M extended-spectrum beta-lactamase genes in commensal Escherichia coli isolates from healthy children from low-resource settings in Latin America. Antimicrob Agents Chemother 51: 2720–2725.
25. Mora A, Herrera A, Mamani R, Lopez C, Alonso MP, et al. (2010) Recent emergence of clonal group O25b:K1:H4-B2-ST131 ibeA strains among Escherichia coli poultry isolates, including CTX-M-9-producing strains, and comparison with clinical human isolates. Appl Environ Microbiol 76: 6991–6997.
26. Dhanji H, Murphy NM, Akhigbe C, Doumith M, Hope R, et al. (2011) Isolation of fluoroquinolone-resistant O25b:H4-ST131 Escherichia coli with CTX-M-14 extended-spectrum beta-lactamase from UK river water. J Antimicrob Chemother 66: 512–516.
27. Colomer-Lluch M, Mora A, Lopez C, Mamani R, Dahbi G, et al. (2013) Detection of quinolone-resistant Escherichia coli isolates belonging to clonal groups O25b:H4-B2-ST131 and O25b:H4-D-ST69 in raw sewage and river water in Barcelona, Spain. J Antimicrob Chemother 68: 758–765.
28. Gibreel TM, Dodgson AR, Cheesbrough J, Bolton FJ, Fox AJ, et al. (2012) High metabolic potential may contribute to the success of ST131 uropathogenic Escherichia coli. J Clin Microbiol 50: 3202–3207.
29. Novais A, Pires J, Ferreira H, Costa L, Montenegro C, et al. (2012) Characterization of globally spread Escherichia coli ST131 isolates (1991 to 2010). Antimicrob Agents Chemother 56: 3973–3976.
30. Blanco J, Mora A, Mamani R, Lopez C, Blanco M, et al. (2011) National survey of Escherichia coli causing extraintestinal infections reveals the spread of drug-resistant clonal groups O25b:H4-B2-ST131, O15:H1-D-ST393 and CGA-D-ST69 with high virulence gene content in Spain. J Antimicrob Chemother 66: 2011–2021.
31. Johnson JR, Kuskowski MA, Gajewski A, Sahm DF, Karlowsky JA (2004) Virulence characteristics and phylogenetic background of multidrug-resistant and antimicrobial-susceptible clinical isolates of Escherichia coli from across the United States, 2000-2001. J Infect Dis 190: 1739–1744.
32. Schubert S, Darlu P, Clermont O, Wieser A, Magistro G, et al. (2009) Role of intraspecies recombination in the spread of pathogenicity islands within the Escherichia coli species. PLoS Pathog 5: e1000257.
33. Wirth T, Falush D, Lan R, Colles F, Mensa P, et al. (2006) Sex and virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol 60: 1136–1151.
34. Johnson JR, Nicolas-Chanoine MH, DebRoy C, Castanheira M, Robicsek A, et al. (2012) Comparison of Escherichia coli ST131 pulsotypes, by epidemiologic traits, 1967–2009. Emerg Infect Dis 18: 598–607.
35. Johnson JR, Tchesnokova V, Johnston B, Clabots C, Roberts PL, et al. (2013) Abrupt emergence of a single dominant multidrug-resistant strain of Escherichia coli. J Infect Dis 207: 919–928.
36. Blanco J, Mora A, Mamani R, Lopez C, Blanco M, et al. (2013) Four main virotypes among extended-spectrum-beta-lactamase-producing isolates of Escherichia coli O25b:H4-B2-ST131: bacterial, epidemiological, and clinical characteristics. J Clin Microbiol 51: 3358–3367.
37. Peirano G, Pitout JD (2014) Fluoroquinolone resistant Escherichia coli ST131 causing bloodstream infections in a centralized Canadian region: the rapid emergence of H30-Rx sublineage. Antimicrob Agents Chemother.
38. Colpan A, Johnston B, Porter S, Clabots C, Anway R, et al. (2013) Escherichia coli sequence type 131 (ST131) subclone H30 as an emergent multidrug-resistant pathogen among US veterans. Clin Infect Dis 57: 1256–1265.
39. Banerjee R, Robicsek A, Kuskowski MA, Porter S, Johnston BD, et al. (2013) Molecular epidemiology of Escherichia coli sequence type 131 and Its H30 and H30-Rx subclones among extended-spectrum-beta-lactamase-positive and -negative E. coli clinical isolates from the Chicago Region, 2007 to 2010. Antimicrob Agents Chemother 57: 6385–6388.
40. Dahbi G, Mora A, Lopez C, Alonso MP, Mamani R, et al. (2013) Emergence of new variants of ST131 clonal group among extraintestinal pathogenic Escherichia coli producing extended-spectrum beta-lactamases. Int J Antimicrob Agents 42: 347–351.
41. Avasthi TS, Kumar N, Baddam R, Hussain A, Nandanwar N, et al. (2011) Genome of multidrug-resistant uropathogenic Escherichia coli strain NA114 from India. J Bacteriol 193: 4272–4273.
42. Totsika M, Beatson SA, Sarkar S, Phan MD, Petty NK, et al. (2011) Insights into a Multidrug Resistant Escherichia coli Pathogen of the Globally Disseminated ST131 Lineage: Genome Analysis and Virulence Mechanisms. PLoS One 6: e26578.
43. Clark G, Paszkiewicz K, Hale J, Weston V, Constantinidou C, et al. (2012) Genomic analysis uncovers a phenotypically diverse but genetically homogeneous Escherichia coli ST131 clone circulating in unrelated urinary tract infections. J Antimicrob Chemother 67: 868–877.
44. Lavigne JP, Vergunst AC, Goret L, Sotto A, Combescure C, et al. (2012) Virulence potential and genomic mapping of the worldwide clone Escherichia coli ST131. PLoS One 7: e34294.
45. Kunne C, Billion A, Mshana SE, Schmiedel J, Domann E, et al. (2012) Complete sequences of plasmids from the hemolytic-uremic syndrome-associated Escherichia coli strain HUSEC41. J Bacteriol 194: 532–533.
46. Grad YH, Godfrey P, Cerquiera GC, Mariani-Kurkdjian P, Gouali M, et al. (2013) Comparative genomics of recent Shiga toxin-producing Escherichia coli O104:H4: short-term evolution of an emerging pathogen. MBio 4: e00452–00412.
47. Brunder W, Schmidt H, Frosch M, Karch H (1999) The large plasmids of Shiga-toxin-producing Escherichia coli (STEC) are highly variable genetic elements. Microbiology 145 (Pt 5): 1005–1014.
48. Pallen MJ, Wren BW (2007) Bacterial pathogenomics. Nature 449: 835–842.
49. Keen EC (2012) Paradigms of pathogenesis: targeting the mobile genetic elements of disease. Front Cell Infect Microbiol 2: 161.
50. Woodford N, Turton JF, Livermore DM (2011) Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. FEMS Microbiol Rev 35: 736–755.
51. Blanco M, Alonso MP, Nicolas-Chanoine MH, Dahbi G, Mora A, et al. (2009) Molecular epidemiology of Escherichia coli producing extended-spectrum {beta}-lactamases in Lugo (Spain): dissemination of clone O25b:H4-ST131 producing CTX-M-15. J Antimicrob Chemother 63: 1135–1141.
52. Coelho A, Mora A, Mamani R, Lopez C, Gonzalez-Lopez JJ, et al. (2011) Spread of Escherichia coli O25b:H4-B2-ST131 producing CTX-M-15 and SHV-12 with high virulence gene content in Barcelona (Spain). J Antimicrob Chemother 66: 517–526.
53. Garcillan-Barcia MP, Alvarado A, de la Cruz F (2011) Identification of bacterial plasmids based on mobility and plasmid population biology. FEMS Microbiol Rev 35: 936–956.
54. Hauser M, Mayer CE, Soding J (2013) kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics 14: 248.
55. Peigne C, Bidet P, Mahjoub-Messai F, Plainvert C, Barbe V, et al. (2009) The plasmid of Escherichia coli strain S88 (O45:K1:H7) that causes neonatal meningitis is closely related to avian pathogenic E. coli plasmids and is associated with high-level bacteremia in a neonatal rat meningitis model. Infect Immun 77: 2272–2284.
56. Zong Z (2013) Complete sequence of pJIE186-2, a plasmid carrying multiple virulence factors from a sequence type 131 Escherichia coli O25 strain. Antimicrob Agents Chemother 57: 597–600.
57. Falgenhauer L, Yao Y, Fritzenwanker M, Schmiedel J, Imirzalioglu C, et al. (2014) Complete Genome Sequence of Phage-Like Plasmid pECOH89, Encoding CTX-M-15. Genome Announc 2.
58. Miquel S, Peyretaillade E, Claret L, de Vallee A, Dossat C, et al. (2010) Complete genome sequence of Crohn's disease-associated adherent-invasive E. coli strain LF82. PLoS One 5.
59. Kidgell C, Pickard D, Wain J, James K, Diem Nga LT, et al. (2002) Characterisation and distribution of a cryptic Salmonella typhi plasmid pHCM2. Plasmid 47: 159–171.

60. Kim M, Kim S, Ryu S (2012) Complete genome sequence of bacteriophage SSU5 specific for Salmonella enterica serovar Typhimurium rough strains. J Virol 86: 10894.
61. Cottell JL, Webber MA, Coldham NG, Taylor DL, Cerdeno-Tarraga AM, et al. (2011) Complete sequence and molecular epidemiology of IncK epidemic plasmid encoding blaCTX-M-14. Emerg Infect Dis 17: 645–652.
62. Valverde A, Canton R, Garcillan-Barcia MP, Novais A, Galan JC, et al. (2009) Spread of bla(CTX-M-14) is driven mainly by IncK plasmids disseminated among Escherichia coli phylogroups A, B1, and D in Spain. Antimicrob Agents Chemother 53: 5204–5212.
63. Kim SR, Komano T (1992) Nucleotide sequence of the R721 shufflon. J Bacteriol 174: 7053–7058.
64. Mellata M, Maddux JT, Nam T, Thomson N, Hauser H, et al. (2012) New insights into the bacterial fitness-associated mechanisms revealed by the characterization of large plasmids of an avian pathogenic E. coli. PLoS One 7: e29481.
65. Wang P, Xiong Y, Lan R, Ye C, Wang H, et al. (2011) pO157_Sal, a novel conjugative plasmid detected in outbreak isolates of Escherichia coli O157:H7. J Clin Microbiol 49: 1594–1597.
66. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, et al. (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet 5: e1000344.
67. Brown CJ, Sen D, Yano H, Bauer ML, Rogers LM, et al. (2013) Diverse broad-host-range plasmids from freshwater carry few accessory genes. Appl Environ Microbiol 79: 7684–7695.
68. Smajs D, Micenkova L, Smarda J, Vrba M, Sevcikova A, et al. (2010) Bacteriocin synthesis in uropathogenic and commensal Escherichia coli: colicin E1 is a potential virulence factor. BMC Microbiol 10: 288.
69. Baquero F, Coque TM (2011) Multilevel population genetics in antibiotic resistance. FEMS Microbiol Rev 35: 705–706.
70. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, et al. (2008) The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol 190: 6881–6893.
71. Hazen TH, Sahl JW, Fraser CM, Donnenberg MS, Scheutz F, et al. (2013) Refining the pathovar paradigm via phylogenomics of the attaching and effacing Escherichia coli. Proc Natl Acad Sci U S A 110: 12810–12815.
72. Skippington E, Ragan MA (2011) Lateral genetic transfer and the construction of genetic exchange communities. FEMS Microbiol Rev.
73. Ogura Y, Ooka T, Iguchi A, Toh H, Asadulghani M, et al. (2009) Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli. Proc Natl Acad Sci U S A 106: 17939–17944.
74. Croucher NJ, Harris SR, Grad YH, Hanage WP (2013) Bacterial genomes in epidemiology—present and future. Philos Trans R Soc Lond B Biol Sci 368: 20120202.
75. Johnson TJ, Logue CM, Johnson JR, Kuskowski MA, Sherwood JS, et al. (2012) Associations Between Multidrug Resistance, Plasmid Content, and Virulence Potential Among Extraintestinal Pathogenic and Commensal Escherichia coli from Humans and Poultry. Foodborne Pathog Dis 9: 37–46.
76. Polz MF, Alm EJ, Hanage WP (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet 29: 170–175.
77. Reeves PR, Liu B, Zhou Z, Li D, Guo D, et al. (2011) Rates of mutation and host transmission for an Escherichia coli clone over 3 years. PLoS One 6: e26907.
78. Sandegren L, Linkevicius M, Lytsy B, Melhus A, Andersson DI (2012) Transfer of an Escherichia coli ST131 multiresistance cassette has created a Klebsiella pneumoniae-specific plasmid associated with a major nosocomial outbreak. J Antimicrob Chemother 67: 74–83.
79. Chen L, Hu H, Chavda KD, Zhao S, Liu R, et al. (2014) Complete Sequence of a KPC-Producing IncN Multidrug-Resistant Plasmid from an Epidemic Escherichia coli Sequence Type 131 Strain in China. Antimicrob Agents Chemother 58: 2422–2425.
80. Partridge SR, Zong Z, Iredell JR (2011) Recombination in IS26 and Tn2 in the evolution of multiresistance regions carrying blaCTX-M-15 on conjugative IncF plasmids from Escherichia coli. Antimicrob Agents Chemother 55: 4971–4978.
81. O'Hara JA, Hu F, Ahn C, Nelson J, Rivera JI, et al. (2014) Molecular Epidemiology of KPC-Producing Escherichia coli: Occurrence of ST131-fimH30 Subclone Harboring pKpQIL-like IncFIIk Plasmid. Antimicrob Agents Chemother.
82. Johnson TJ, Wannemuehler YM, Johnson SJ, Logue CM, White DG, et al. (2007) Plasmid replicon typing of commensal and pathogenic Escherichia coli isolates. Appl Environ Microbiol 73: 1976–1983.
83. Chen L, Chavda KD, Al Laham N, Melano RG, Jacobs MR, et al. (2013) Complete nucleotide sequence of a blaKPC-harboring IncI2 plasmid and its dissemination in New Jersey and New York hospitals. Antimicrob Agents Chemother 57: 5019–5025.
84. Johnson TJ, Lang KS (2012) IncA/C plasmids: An emerging threat to human and animal health? Mob Genet Elements 2: 55–58.
85. Kim YA, Qureshi ZA, Adams-Haduch JM, Park YS, Shutt KA, et al. (2012) Features of infections due to Klebsiella pneumoniae carbapenemase-producing Escherichia coli: emergence of sequence type 131. Clin Infect Dis 55: 224–231.
86. Muniesa M, Colomer-Lluch M, Jofre J (2013) Potential impact of environmental bacteriophages in spreading antibiotic resistance genes. Future Microbiol 8: 739–751.
87. Hoyland-Kroghsbo NM, Maerkedahl RB, Svenningsen SL (2013) A quorum-sensing-induced bacteriophage defense mechanism. MBio 4: e00362–00312.
88. Schimke RT (1984) Gene amplification in cultured animal cells. Cell 37: 705–713.
89. Baquero F, Tedim AP, Coque TM (2013) Antibiotic resistance shaping multi-level population biology of bacteria. Front Microbiol 4: 15.
90. Andersen PS, Stegger M, Aziz M, Contente-Cuomo T, Gibbons HS, et al. (2013) Complete Genome Sequence of the Epidemic and Highly Virulent CTX-M-15-Producing H30-Rx Subclone of Escherichia coli ST131. Genome Announc 1.
91. Toh H, Oshima K, Toyoda A, Ogura Y, Ooka T, et al. (2010) Complete genome sequence of the wild-type commensal Escherichia coli strain SE15, belonging to phylogenetic group B2. J Bacteriol 192: 1165–1166.
92. Boyd DA, Tyler S, Christianson S, McGeer A, Muller MP, et al. (2004) Complete nucleotide sequence of a 92-kilobase plasmid harboring the CTX-M-15 extended-spectrum beta-lactamase involved in an outbreak in long-term-care facilities in Toronto, Canada. Antimicrob Agents Chemother 48: 3758–3764.
93. Woodford N, Carattoli A, Karisik E, Underwood A, Ellington MJ, et al. (2009) Complete nucleotide sequences of plasmids pEK204, pEK499, and pEK516, encoding CTX-M enzymes in three major Escherichia coli lineages from the United Kingdom, all belonging to the international O25:H4-ST131 clone. Antimicrob Agents Chemother 53: 4472–4482.
94. Partridge SR, Ellem JA, Tetu SG, Zong Z, Paulsen IT, et al. (2011) Complete sequence of pJIE143, a pir-type plasmid carrying ISEcp1-blaCTX-M-15 from an Escherichia coli ST131 isolate. Antimicrob Agents Chemother 55: 5933–5935.
95. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659.
96. Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5: e11147.
97. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.
98. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.
99. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821–829.
100. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359.
101. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.
102. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10: 421.
103. Garcillan-Barcia MP, de la Cruz F (2013) Ordering the bestiary of genetic elements transmissible by conjugation. Mob Genet Elements 3: e24263.
104. del Solar G, Giraldo R, Ruiz-Echevarria MJ, Espinosa M, Diaz-Orejas R (1998) Replication and control of circular bacterial plasmids. Microbiol Mol Biol Rev 62: 434–464.
105. Zheng J, Peng D, Ruan L, Sun M (2013) Evolution and dynamics of megaplasmids with genome sizes larger than 100 kb in the Bacillus cereus group. BMC Evol Biol 13: 262.
106. Osborn AM, da Silva Tatley FM, Steyn LM, Pickup RW, Saunders JR (2000) Mosaic plasmids and mosaic replicons: evolutionary lessons from the analysis of genetic diversity in IncFII-related replicons. Microbiology 146 (Pt 9): 2267–2275.
107. Smillie C, Garcillan-Barcia MP, Francia MV, Rocha EP, de la Cruz F (2010) Mobility of plasmids. Microbiol Mol Biol Rev 74: 434–452.
108. Besemer J, Lomsadze A, Borodovsky M (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29: 2607–2618.
109. Guglielmini J, Quintais L, Garcillan-Barcia MP, de la Cruz F, Rocha EP (2011) The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. PLoS Genet 7: e1002222.
110. Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7: e1002195.
111. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431–432.
112. Zhou Y, Call DR, Broschat SL (2013) Using protein clusters from whole proteomes to construct and augment a dendrogram. Adv Bioinformatics 2013: 191586.
113. Tekaia F, Lazcano A, Dujon B (1999) The genomic tree as revealed from whole proteome comparisons. Genome Res 9: 550–557.
114. Tekaia F, Yeramian E (2005) Genome trees from conservation profiles. PLoS Comput Biol 1: e75.
115. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, et al. (2013) vegan: Community Ecology Package.

116. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20: 289–290.
117. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA (2011) BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12: 402.
118. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M (2009) ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25: 1968–1969.
119. Bonnin RA, Poirel L, Carattoli A, Nordmann P (2012) Characterization of an IncFII plasmid encoding NDM-1 from Escherichia coli ST131. PLoS One 7: e34752.
120. Cullik A, Pfeifer Y, Prager R, von Baum H, Witte W (2010) A novel IS26 structure surrounds blaCTX-M genes in different plasmids from German clinical Escherichia coli isolates. J Med Microbiol 59: 580–587.
121. Chen L, Chavda KD, Melano RG, Jacobs MR, Koll B, et al. (2014) Comparative Genomic Analysis of KPC-Encoding pKpQIL-Like Plasmids and Their Distribution in New Jersey and New York Hospitals. Antimicrob Agents Chemother 58: 2871–2877.

Text S1: The plasmidome of *E. coli* ST131

This work describes the remarkable heterogeneity of plasmids found in ten *E. coli* ST131 sequenced genomes. We find members of 8 out of the 17 main MOB plasmid groups found in γ-proteobacteria [11,53]. Specifically, we find members of Inc groups F12 (IncF), P3 (IncX), P5 (ColE1), P6 (IncI2), P11 (IncP), P12 (IncI/K/BO), Q12 (Rep_pSC101-like), Qu (Rep_pMG828-2/IGWZ12-like) plus others lacking relaxases that were identifiable as phage-related/RepFIB and no-MOB small plasmids. To date, IncF [15,17,29,56,80,92,93,119, Suzuki et al., 2009, Matsumura et al., 2013], IncFIIK [81], IncI1 [93, Curiao et al., 2011]; IncN [79,120, Suzuki et al., 2009, Zong et al., 2011, Novais et al., 2012] and, sporadically, IncA/C [Naseer et al., 2010, Novais et al., 2012], IncX [94], IncU [Matsumura et al., 2013], IncY [Dhanji et al., 2011], IncK [Calhau et al., 2013], Inclγ [Calhau et al., 2013], IncB/O [Matsumura et al., 2013] and ColE [Matsumura et al., 2012] plasmids were identified in *E. coli* ST131 by PBRT. They were associated with the spread of class A [CTXM (-1,-2, -3, -9, -14, -14, -15, -27, -32, -61), TEM (-4, -24, -116), SHV (-2, -5, -7, -12), NDM (-1), KPC (-2, -3, -4)], class B (VIM-1), class C (CMY-2) and class D (OXA-1) beta lactamases [12,13,23-29] [26,81,83,121, Rogers et al., 2011, Ma et al., 2013, Matsumura et al., 2013, Accogli et al., 2014, Cai et al., 2014] and genes coding for resistance to different antibiotic families (tetracyclines, macrolides, aminoglycosides, trimethoprim and sulfonamides), which are located in composite genomic islands, mostly derived from IncF plasmids (see main text). Despite extensive plasmid analysis of ST131 isolates in different studies, the ST131 plasmidome was not described comprehensively, as most information comes uniquely from identification of antibiotic resistant plasmid replicons by PBRT. As of April 2014, ten plasmids were sequenced from ST131 strains, as shown in Table 3. They include four IncFII, four IncN, one IncI1 and one IncX4 plasmids. Here we provide an individualized discussion of the 8 plasmid groups (39 plasmids) found in our ST131 survey, together with the ten previously sequenced plasmids. This information underscores the relevance of the ST131 plasmidome in the adaptation of this *E. coli* clonal group by providing emergent functions (e.g., antibiotic resistance, virulence, colonization enhancement, ability to outcompete other microbial species) and thus, the usefulness of PLACNET in bacterial population studies, as explained in the main body text of the manuscript.

MOB_{F12}/IncF plasmids (Suppl. Fig. S11). Recent plasmid history suggests that IncF plasmids represent the most abundant family among *E. coli* metapopulations. They are drivers of *E. coli* evolution because of their ability to acquire multiple adaptive traits coding for antibiotic resistance and virulence [2,3,9,82,

Hanni et al., 1982, Womble et al., 1988]. Figure 5 (inset) shows that IncF plasmids identified in the ST131 genomes correspond to four dendrogram groups. The two main groups were Group I (FIB), frequent among avian pathogenic *E. coli* (APEC) and extranintestinal pathogenic *E. coli* (ExPEC), and linked to putative virulence traits and Group II (FII/FIA), widely disseminated among ExPEC and extensively associated with antibiotic resistance (see Table 4). Besides, there were two outlier groups, Group III and Group IV.

Group I (FIB) plasmids mostly comprises virulence plasmids (such as pAPEC-like ones) and four ST131 plasmids containing numerous virulence genes. They are compared in Fig S11A, using ST131 plasmid pJIE186_2 as a reference. Three FIB (ΔFIA:B1) plasmids, pE61BA_1 and pE35BA2+3, were identified in two strains corresponding to ST9/H22/virotype D and ST43/H30/virotype B, respectively. To date, only one report [56] documented FIB plasmid carriage in *E. coli* ST131, despite the identification of putative virulence markers associated with ColV plasmids often appeared in surveys of ST131 *E. coli*[29,36, Kudinha et al., 2013]. Fig S11A emphasizes that all ST131 FIB plasmids (and relatives) share a 80kb genetic island (coordinates 93 kb to 21 kb in the map of Fig S11A) that includes a conserved ColV region comprising *iss, iroBCDEN, iucABCD, iutA, cvaBC* and *sitC* and an *ompT-hlyF-mig14* cassette[55]. While some traits of the conserved region can also be located in the chromosome (linked to PAIs), the cassette *ompT-hlyF-mig14* has only been detected in plasmids[2, Billard-Pomares et al., 2011].

Group II (FII/FIA/FIB) plasmids are compared in Fig S11B, using plasmid pJJ1886_5 as a reference. The seven plasmids analyzed show extensive sequence similarity over at least 100 kb, which includes not only backbone genes but also metabolic, transport and genes associated with mobile genetic elements (MGE). Interestingly, five of the seven plasmids show extensive deletions within their Tra region (see Fig S11B from coord 40 kb to 70 kb). A wide diversity of IncFII/FIA/FIB plasmids carrying *bla* genes have been described among ST131 clonal variants[17,92,93, Fiett et al., 2014], some plasmid types being apparently overrepresented at a global [29] or local level [Doumith et al., 2012, Matsumura et al., 2013] as F2:A1:B- (similar to pEK399) and F2:A-:B- (similar to pC15-1a, pHK01, pEK516), or F29:B10 (identified in ST131 from South Europe). The diversity of replicons, RFLP patterns and sequence mosaicism of *E. coli* ST131 IncF plasmids, even within situations that resemble plasmid outbreak scenarios [17,80,92,119, Smet et al., 2010] suggest that multiple DNA rearrangements among IncF2 plasmids of Enterobacteriaceae are extremely frequent under selection, as described earlier[Hanni et al., 1982, Womble et al., 1988]. More recently, it has been shown that homologous recombination mediated by

IS*26* and Tn*2* between particular ST131 IncF2 plasmids occurs often and partly explains the mentioned diversity[80, Smet et al., 2010]. Recombination of F plasmids among them or with other plasmid groups in enterobacterial populations involved in multispecies outbreaks of $bla_{CTX\text{-}M\text{-}15}$ and $bla_{KPC\text{-}2}$, confirms rearrangements among large composite regions containing antibiotic resistance genes, transposons and ISs and/or transfer regions of F plasmids is extraordinarily common[Sandegren et al., 2012]. The overrepresentation of F2:A1:B- (pEK499 derivatives) [93, Dhanji et al., 2011]and F2:A-:B- plasmids (e.g. pC15-1a, pHK01, pEK516), the last group only inferred in this study in an H30 strain of "virotype B"[17,29,92, Ho et al., 2012], can reflect the expansion of different *fimH30* ST131 sublineages (here designated as virotypes A, B and C) in different areas.

Group III contains only plasmid pHVH177_1, without antibiotic resistance genes and low content of virulence genes. The BRIG representation (Fig S11C), which compares pHVH177_1 with its closest relatives in the dendrogram of Figure 5, shows conservation of just the MOB_{F12} backbone genes (about 32 kb).

Group IV is represented by plasmid pECSF1, from the commensal ST131 strain SE15. No resistance genes and low content of virulence genes were detected. The most similar plasmids (Fig S11D) are the 122 kb plasmid p1ESCUM[66], the 115 kb plasmid pUM146 (GenBank acc NC_017630) and the 114 kb plasmid pEC14_114[DebRoy et al., 2010]. Besides backbone genes, these plasmids share more than 60 additional kb that includes a number of potential transposable elements. Other close relatives are the 168 kb plasmid pIP1206 [Perichon et al., 2008] and the *K. pneumoniae* MDR plasmid pKF3-140[Zhao et al., 2010].

All these data taken together underscore the complex evolution of MOB_{F12}/IncF plasmids, facilitated by the recombinogenic potential of transposable elements[92, Hanni et al., 1982, Womble et al., 1988, Partridge et al., 2004, Johnson et al., 2010].

Phage-related/ RepFIB plasmids (Suppl Fig S12). The ST131 *E.coli* plasmids pBIDMC20B_2 and pBWH24_2, found in two ST43/H30/virotype C isolates, are closely related among them and to the antibiotic resistance plasmid pECOH89 [57] and the STEC plasmid p09EL50[5]. They all share the same RepFIB replication and ParB proteins as backbone remnants, and an extensive set of phage-related proteins (Fig S12A). This group also comprises other phage-related plasmids such as plasmid pLF82 [58] from the prototype strain of adherent invasive *E. coli* (AIEC), the *Salmonella* plasmid pHCM2[59], the

Salmonella bacteriophage SSU5 [60] and the *Yersinia pestis* pMT1 and MT plasmids[Hu et al., 1999]. An extensive set of phage-related genes are shared by all seven plasmids. As shown in Fig S12A, only pECOH89 plasmid harbors the $bla_{CTX\text{-}M\text{-}15}$ antibiotic resistance gene. As shown in the phylogenetic tree of Fig S12B, the RIP protein RepFIB (Rep3_superfamily, pfam 01051) from ST131 plasmids is identical to those of plasmids pECOH89 and p09EL50 (here named as Group C). They are only 40% identical to Group B proteins, represented by plasmids pLF82, pHCM2, pMT1 and MT. Group A in the figure is represented by the well-known IncF (FIB) plasmids discussed in the previous section. At this point it must be stressed that, although a majority of plasmids carrying RepFIB belong to the IncF group, RepFIB also appears in more than 20 other RIP groups such as IncN, IncP or IncI groups[Gibbs et al., 1993]. Further work is required to rigorously classify RepFIB proteins (all belonging to the Rep3-superfamily) according to plasmid RIP groups. Moreover, the biology of RepFIB/phage-related plasmids remains largely unexplored, although they might be more extended than previously thought[57].

MOB_{P12}/IncI-complex (Suppl Fig S13). The single MOB_{P12}/IncI representative in our collection is plasmid pE2022_1, most similar to the IncI1/ST16 plasmid pEK204, carrying $bla_{CTX\text{-}M\text{-}15}$, already isolated from a ST131 strain[93]. Plasmids pEK204 and pE2022_1 are not phylogenetically close among themselves. Rather, they are similar to widespread IncI1 ($bla_{CTX\text{-}M\text{-}15}$) plasmids from Europe and IncK ($bla_{CTX\text{-}M\text{-}14}$) plasmids from Asia, respectively[16,93, Zang et al., 2013, Onnberg et al., 2014].

The IncI1 plasmid pEK204 shares its backbone with the eight IncI1 (also called IncIα) plasmids used as references in Fig S13A. While pEK204 carries ($bla_{CTX\text{-}M\text{-}15}$ and bla_{TEM}), these antibiotic resistance genes are absent in its closest relatives (see coord. 6 kb and 14 kb in S13A). Plasmids in this cluster include the IncIγ prototype, the 93.2 kb plasmid R621a from *S. enterica* [Takahashi et al., 2011].

The ST131 plasmid pE2022_1 (98.3 kb) is similar to the IncK prototype 93.6 kb plasmid pCT [61]. They share more than 80 kb including Tra regions and the $bla_{CTX\text{-}M\text{-}14}$ gene coding for resistance to beta-lactam antibiotics. IncK pCT-like plasmids harboring $bla_{CTX\text{-}M\text{-}14}$ are globally spread among animals and humans, being prevalent in UK and Spain among other countries [63,64, Dhanji et al., 2011]. Although widely spread, they have been only sporadically described among ST131 isolates, since IncF plasmids are the main vectors carrying $bla_{CTX\text{-}M\text{-}15}$ and $bla_{CTX\text{-}M\text{-}14}$ genes in ST131 to date.

MOB_{P6}/IncI2 plasmids (Suppl Fig S14). The two IncI2 plasmids found in our ST131 genomes (pBWH24_3 and pE61BA_7) belong to different clusters of a large plasmid family that, however, contains few

sequenced plasmids. Fig S14A shows the phylogenetic tree of MOB_{P6} REL proteins, underscoring the distant positions of the ST131 plasmid REL proteins. Plasmid pBWH24_3 (60.3 kb) is similar to the *E.coli* IncI2 plasmid prototype R721 (75.6 kb; [63]) and the *E.coli* APEC plasmid pChi7122_3 (56.7 kb, GenBank acc FR851304) over most of their length (Fig S14B). Plasmid R721 is the representative of a number of enterobacterial plasmids commonly isolated from *K. pneumoniae* and associated with spread of antibiotic resistance, mainly bla_{CMY-2} and bla_{KPC-2} [83]. Plasmids pBWH24_3 and pChi7122_3, on the other hand, lack any antibiotic resistance gene. These two plasmids are highly similar, the last one having a role in acid resistance and biofilm formation [64]. Besides, it is of note that IncI2 plasmids have a high number of putative integration sites, which would facilitate the acquisition of different antibiotic resistance genes, resulting in large composite multidrug platforms[83].

Plasmid pE61BA_7 (37.9 kb) is most similar (Fig S14C) to the *S. enterica* plasmid SL483 (37.9kb; [Fricke et al., 2011] and *E. coli* plasmid pO157_Sal (37.8 kb; [65]). Plasmid pE61BA_7 belongs to a different IncI2 subcluster than pBWH24_3, as shown in Fig S14A. This group is composed of cryptic plasmids, represented by SL483 from *Salmonella agona* and pO157_Sal from *E. coli* O157:H5, and seems to be widely spread among enterobacterial isolates from animals [65]. IncI2 plasmids have been identified among CTX-M-15 producers of the ST131/H30 sublineage collected in Canada, Australia and New Zealand [16] but they were not fully characterized in any of those studies.

MOB_{P3} /IncX plasmids (Suppl Fig S15). IncX plasmids are prevalent in *E. coli* and belong to a large plasmid family with four recognized subgroups X1 to X4 [Johnson et al., 2012]. The two IncX plasmids identified in this survey belong to the IncX1 (pFV9873_4) and IncX4 (pE2022_3) subclusters (Fig S15A). To these, the previously reported IncX4 plasmid pJIE143 should be added. As shown in Fig S15B, the IncX1 plasmid pFV9873_4 (33.3 kb) was most similar to the 33.8 kb plasmid p2ESCUM [66] over its entire length, and no so much to other reference hits. p2ESCUM was originally isolated from the emblematic strain UMN026, representative of a widely disseminated urinary pathogenic ST69 *E. coli* (UPEC) clone [Lescat et al., 2009]. The IncX4 plasmid pE2022_3 (35.0 kb) shares extensive homology with the *S. enterica* plasmid pSH696_34 (33.8 kb, [Gokulan et al., 2013]), which is used as a reference in Fig S15C. Compared to these two plasmids, the ST131 reference plasmid pJIE143 (34.3 kb;[94]), widely disseminated among *E. coli* from foodborne animals, lacks two important backbone genes, the conjugative coupling protein gene *traG* (around coord 19 kb) and the RIP encoding region (coord 9 kb to 10 kb). Plasmid pJIE143 codes for a completely different Pir protein (13% amino acid identity) although

also belonging to the Rep3 family. Therefore, it is probable that pSH696_34 and pE2022_3 belong to a different incompatibility group than pJIE143. IncX4 plasmids can be transferred at different temperatures and are frequent vehicles of antibiotic resistance, mainly *qnr*. Difficulties for detecting this plasmid group using PBRT might have underestimated its prevalence.

MOB_{P11}/IncP1 plasmids (Suppl Fig S16). Members of the IncP1 family are frequently isolated in the environment and often carry modules containing diverse accessory genes can be transferred between unrelated bacteria [Schluter et al., 2007, Venturini et al., 2013]. IncP1 plasmids from *E. coli* remain largely unexplored although detection of IncP sequences from animal [Dotto et al., 2014] and human *E. coli* isolates, included ST131 [16], suggest that IncP1 plasmids can be easily acquired by human associated Proteobacteria. As can be observed in the phylogenetic tree of Fig S16A, plasmids pJJ1886_4 and pE61BA_4 represent different new branches of the MOB_{P11} family and thus new additions to the ST131 plasmidome. The pJJ1886_4 REL protein is most similar to that of the 40.6 kb plasmid pDS1. However, proteome analysis resulting in the dendrogram of Fig. 5, shows plasmid pHS102707 (*E. coli*, 69.5 kb, Genbank acc NC_023907) as its closest homolog. This is confirmed in Fig S16B, where the extensive homology of pJJ1886_4 and pHS102707 is depicted. The second branch of the phylogenetic tree contains the ST131 plasmid pE61BA_4 (18.3 kb). The 37.9 kb plasmid pMBUI2 [67], from an uncultured bacterium, appears as its closest homolog at the level of REL protein. Plasmid pE61BA_4 seems a crippled plasmid, lacking most of its backbone, with just four genes shared with pMBUI2 (Suppl. Table S1), suggesting that it is the first representative of a new plasmid group. Finally, the same phylogenetic branch contains IME_E35BA REL protein. To analyze this integrative and mobilizable element (IME), we compared the relevant regions of E35BA and JJ1886 genomes using EasyFig [Sullivan et al., 2011]. Fig S16C shows the IME_E35BA (14.2 kb) inserted between a hypothetical protein and a GMP synthase encoding genes in *E.coli* JJ1886. IME_E35BA is similar to the *Burkholderia glumae* IncP island (14.7 kb; [Yoshii et al., 2012]). Both share a *tra* conjugative region, *repA*, *alpA* (encoding a DNA binding protein) and *int-P4* (coding for a P4 phage integrase family protein). They differ in their *repC* (resolvase) genes. Besides, IME_E35BA does not harbor the kasugamycin-2'-N-acetyltransferase gene [aac(2')-Iia] present in the *Bukholderia glumae* strain that confers resistance to kasugamycin, an aminoglycoside antibiotic widely used in agriculture [Yoshii et al., 2012].

MOB_{C12} plasmids (Suppl Fig S17). The MOB_C family constitutes a wide plasmid group, present in γ-proteobacteria, Firmicutes and Tenericutes. MOB_C relaxases corresponding to γ-proteobacterial

plasmids are clustered in the MOB_{C1} group, subdivided as well in two branches: MOB_{C11} and MOB_{C12}[11, Garcillan-Barcia et al., 2009]. Plasmid pE61BA_2 (24.5kb), located in a ST131 ST9/H234 *E.coli* strain, is not close to any reference from Figure 5, but clusters in the MOB_{C12} group according to its REL (Fig S17A). pE61BA_2 REL protein is only 70% identical to the closest hits: the 22 kb *Y. pestis* cryptic plasmid pCRY [Song et al., 2004] and the 42 kb *K.pneumoniae* multidrug resistant (MDR) plasmid pMET1 [Soler Bistue et al., 2008]. As shown in Fig S17B, pE61BA_2 shares mobilization (*mobBC –virB*) and replication (*repA*) genes with pCRY and pMET1, but lacks the micrococcal nuclease-like and the Mpr coding genes. Since PBRT does not detect MOB_C plasmids, they are probably underrepresented in the plasmidome of Enterobacteriaceae.

MOB_{P5}/ColE plasmids (Suppl Fig S18). The MOB_{P5} REL phylogenetic tree shows three different branches, constituting the MOB_{P51}, MOB_{P52} and MOB_{P53} subfamilies [11]. The four ST131 MOB_{P5} plasmids are located in the MOB_{P51} subfamily (Fig S18A). Plasmid pE61BA_5 (6.6 kb) is similar to the prototype MOB_{P5} plasmid ColE1 (6.6 kb) [Tomizawa et al., 1977] over their entire length (including MOB proteins and colicin-E1), as shown in Fig S18B. The ST131 plasmids pBIDMC38_1 (11.8 kb) and pJJ1886_3 (5.6 kb) are very similar among them, including their MOB and RIP proteins (Fig S18C). They are unique in carrying a large (2 kb) gene encoding an intriguing EamA-like transporter protein (pfam00892). Finally, plasmid pE61BA_6 (6.9 kb) is similar to enterohemorrhagic *E. coli* (EHEC) O157 plasmid pColD-157 (6.7 kb, [Hofinger et al., 1998]) and colicin K plasmid pColK-K235 (8.3 kb, [Rijavec et al., 2007]). As Fig S18D shows, pE61BA_6 plasmid shares MbkABCD mobilization and colicin K encoding regions with the pColK-K235 reference plasmid.

Although widely used as cloning vectors for genetic engineering, ColE plasmids represent a largely unexplored family in epidemiological studies, since they not usually carry genes coding for antibiotic resistance. On the other hand, they frequently, but not always, carry colicins [68]. The MOB_{P5}/ColE1-like plasmids identified in this study exhibited a wide size range (5.6-10.5 kb) and variable presence of colicins, two of them carrying either ColE1 or ColK, both in the same strain (Suppl figures S18A-D). These two colicins have traditionally been associated with UPEC and the B2 phylogenetic group of *E. coli* [68, Rijavec et al., 2007]. Although the carriage of ColE-1 like plasmids, inferred by the presence of either RIP or REL sequences, seems to be frequent among non-outbreak O25b and non-O25b ST131 isolates [Matsumura et al., 2013], the lack of full plasmid sequences and functional information preclude any conclusion about a role in the pathogenesis of B2 and ST131 clones. Besides a possible involvement of

colicin production in virulence, perhaps as a competition weapon to avoid the presence of other *E. coli* pathogenic strains [Budic et al., 2011], ColE1 plasmids can promote rearrangements and contribute to the mobilization of other plasmids [Rijavec et al., 2007]. As an example, the large plasmid (pBIDMC38_1) is almost identical to the ST131 reference pJJ1886_3, with an additional type II restriction-modification system (Cfr10I) (Fig. S18). In spite of the above, ColE plasmids carrying genes encoding antibiotic resistance are increasingly being reported although by lack of phylogenetic analysis, they cannot be accurately classified within the MOB_{P5} group [10, Dotto et al., 2014].

MOB_{Qu} plasmids (Suppl Fig S19). MOB_{Qu} and MOB_{Q12} plasmid clusters were previously identified by our group [11,53]. They have scarcely been documented among *E. coli* to date, probably because they do not carry antimicrobial resistance genes. This report is the first to document their presence in the ST131 lineage. Nevertheless, they seem to be common among UPEC, as shown in this study. The four MOB_{Qu} plasmids identified in this study were almost identical among them and with pSE11_6 [Oshima et al., 2008] in a DNA stretch of 4 kb, as shown in Fig S19B. Apparently they only code for RIP and MOB proteins. The role of these cryptic plasmids in the adaptation of ST131 is intriguing. Although not explored at all, a role in the adaptation by providing useful mobilization tools cannot be excluded.

MOB_{Q12} plasmids (Suppl Fig S20). As seen in the REL phylogenetic tree of fig S20A, the four ST131 MOB_{Q12} plasmids have almost identical relaxases. Three of the four MOB_{Q12} plasmids (of about 5.2 kb) (pBIDMC38_2, pFV9873_6 and pJJ1886_2) were almost identical among them and to the UPEC plasmid pCE10B [Lu et al., 2011], as shown in Fig S20B. They constitute a cluster of plasmids within the MOB_{Q12} family previously defined by the authors [11,53]. The fourth plasmid (pE61BA_3; 5.5 kb) has a unique 1.6 kb DNA segment that codes for a 266 amino acid protein identical to a colicin from *S. enterica* (NCBI protein acc. WP_024155916). These plasmids frequently coexist with IncF plasmids and form cointegrates with them (our unpublished data).

Small cryptic no-MOB plasmids (Suppl Fig S21). Small non-mobilizable cryptic plasmids remain scarcely documented among Enterobacteriaceae. Four highly similar plasmids of about 1.6 to 2.2 kb were found in our study. They are also similar to the 1.5 kb *E. coli* plasmid pCE10D [Lu et al., 2011], as shown in Fig S21. As can be expected, they have minimal coding capacity. Their only annotated gene codes for a RIP protein of the Rep_HTH_36_superfamily (pfam13730). Similar plasmids have also been detected in various ExPEC, including ST131 [90], Shiga-toxin producing *E. coli* (STEC) [Lu et al., 2011, Brolund et al.,

2013] and *Klebsiella* [Liu et al., 2012] isolates. A fifth, 5kb no-MOB plasmid (pFV9873_3) was unique and unrelated to any reference plasmid (see Figure 5). The adaptive functions of these plasmids, if any, are unknown.

Other ST131 plasmid groups: MOB_{F11}/IncN plasmids (Suppl. Fig. S22). We carried out extensive literature analysis in the search of plasmid groups present in ST131 but not detected in our survey. There were two plasmid groups, IncN and IncA/C, for which no plasmid was identified in our study although they are being increasingly detected among ST131 isolates. Both groups have been associated with $bla_{CTX\text{-}M}$ [120, Novais et al., 2012] and more recently with bla_{KPC} [89,81,121, Matsumura et al., 2013] and bla_{NDM} genes (GenBank acc nº KJ413946). These plasmids have spread worldwide and might have been acquired by ST131 from other enterobacterial species, e.g., during nosocomial polyclonal outbreaks [85].

Four complete sequences of MOB_{F11}/IncN plasmids originating from ST131 strains are available: the IncN1 plasmids pECN580 [79], pKC394 and pKC396 [120] (GenBank acc nº HM138652 and HM138653, respectively) and the IncN2 plasmid pNDM-ECS01 (GenBank acc nº KJ413946). Suppl Fig S22A shows the REL phylogenetic tree of the MOB_{F11}/IncN plasmid branch, which displays the IncN1 and IncN2 subclusters. BRIG comparative analysis of the three ST131 IncN1 plasmids versus reference plasmids (Suppl Fig S22B) reveals a highly conserved 34kb backbone region (coordinates 0kb to 34kb). As previously reported [11], IncN1 and IncN2 plasmid backbones, including REL proteins, are homologous but their RIP proteins are different [Poirel et al., 2011, Partridge et al, 2012]. Accordingly, the IncN2 pNDM-ECS01 plasmid RIP has no similarity to the RIPs from pECN580, pKC394 or pKC396 IncN1 plasmids. Comparative analysis of the pNDM-ECS01 plasmid (Suppl Fig S22C), harboring the $bla_{NDM\text{-}1}$ gene, versus the reference plasmids p271A [Poirel et al, 2011], pTR3 and pTR4 [Chen et al., 2012] and pJIE137 [Partridge et al., 2012] shows that all are almost identical except pJIE137, which harbors the $bla_{CTX\text{-}M\text{-}62}$ gene. Despite of the low prevalence of IncN1 and IncN2 plasmids in ST131 *E.coli* strains, these plasmids seem to be good vehicles for the transmission of several antibiotic resistance genes in Enterobacteriaceae by maintaining conserved backbones but variable regions that harbor genes encoding CTX-M, KPC or NDM beta-lactamases.

Supplemental references

Suzuki S, Shibata N, Yamane K, Wachino J, Ito K, et al. (2009) Change in the prevalence of extended-spectrum-beta-lactamase-producing Escherichia coli in Japan by clonal spread. J Antimicrob Chemother 63: 72-79.

Matsumura Y, Yamamoto M, Nagao M, Ito Y, Takakura S, et al. (2013) Association of fluoroquinolone resistance, virulence genes, and IncF plasmids with extended-spectrum-beta-lactamase-producing Escherichia coli sequence type 131 (ST131) and ST405 clonal groups. Antimicrob Agents Chemother 57: 4736-4742.

Curiao T, Canton R, Garcillan-Barcia MP, de la Cruz F, Baquero F, et al. (2011) Association of composite IS26-sul3 elements with highly transmissible IncI1 plasmids in extended-spectrum-beta-lactamase-producing Escherichia coli clones from humans. Antimicrob Agents Chemother 55: 2451-2457.

Zong Z, Yu R, Wang X, Lu X (2011) blaCTX-M-65 is carried by a Tn1722-like element on an IncN conjugative plasmid of ST131 Escherichia coli. J Med Microbiol 60: 435-441.

Novais A, Viana D, Baquero F, Martinez-Botas J, Canton R, et al. (2012) Contribution of IncFII and broad-host IncA/C and IncN plasmids to the local expansion and diversification of phylogroup B2 Escherichia coli ST131 clones carrying blaCTX-M-15 and qnrS1 genes. Antimicrob Agents Chemother 56: 2763-2766.

Naseer U, Haldorsen B, Simonsen GS, Sundsfjord A (2010) Sporadic occurrence of CMY-2-producing multidrug-resistant Escherichia coli of ST-complexes 38 and 448, and ST131 in Norway. Clin Microbiol Infect 16: 171-178.

Dhanji H, Doumith M, Rooney PJ, O'Leary MC, Loughrey AC, et al. (2011) Molecular epidemiology of fluoroquinolone-resistant ST131 Escherichia coli producing CTX-M extended-spectrum beta-lactamases in nursing homes in Belfast, UK. J Antimicrob Chemother 66: 297-303.

Calhau V, Ribeiro G, Mendonca N, Da Silva GJ (2013) Prevalent combination of virulence and plasmidic-encoded resistance in ST 131 Escherichia coli strains. Virulence 4: 726-729.

Cai JC, Zhang R, Hu YY, Zhou HW, Chen GX (2014) Emergence of Escherichia coli sequence type 131 isolates producing KPC-2 carbapenemase in China. Antimicrob Agents Chemother 58: 1146-1152.

Rogers BA, Sidjabat HE, Paterson DL (2011) Escherichia coli O25b-ST131: a pandemic, multiresistant, community-associated strain. J Antimicrob Chemother 66: 1-14.

Ma L, Siu LK, Lin JC, Wu TL, Fung CP, et al. (2013) Updated molecular epidemiology of carbapenem-non-susceptible Escherichia coli in Taiwan: first identification of KPC-2 or NDM-1-producing E. coli in Taiwan. BMC Infect Dis 13: 599.

Accogli M, Giani T, Monaco M, Giufre M, Garcia-Fernandez A, et al. (2014) Emergence of Escherichia coli ST131 sub-clone H30 producing VIM-1 and KPC-3 carbapenemases, Italy. J Antimicrob Chemother.

Hanni C, Meyer J, Iida S, Arber W (1982) Occurrence and properties of composite transposon Tn2672: evolution of multiple drug resistance transposons. J Bacteriol 150: 1266-1273.

Womble DD, Rownd RH (1988) Genetic and physical map of plasmid NR1: comparison with other IncFII antibiotic resistance plasmids. Microbiol Rev 52: 433-451.

Kudinha T, Johnson JR, Andrew SD, Kong F, Anderson P, et al. (2013) Distribution of phylogenetic groups, sequence type ST131, and virulence-associated traits among Escherichia coli isolates from men with pyelonephritis or cystitis and healthy controls. Clin Microbiol Infect 19: E173-180.

Billard-Pomares T, Tenaillon O, Le Nagard H, Rouy Z, Cruveiller S, et al. (2011) Complete nucleotide sequence of plasmid pTN48, encoding the CTX-M-14 extended-spectrum beta-lactamase from an Escherichia coli O102-ST405 strain. Antimicrob Agents Chemother 55: 1270-1273.

Fiett J, Baraniak A, Izdebski R, Sitkiewicz I, Zabicka D, et al. (2014) The first NDM metallo-beta-lactamase-producing Enterobacteriaceae isolate in Poland: evolution of IncFII-type plasmids carrying the bla(NDM-1) gene. Antimicrob Agents Chemother 58: 1203-1207.

Doumith M, Dhanji H, Ellington MJ, Hawkey P, Woodford N (2012) Characterization of plasmids encoding extended-spectrum beta-lactamases and their addiction systems circulating among Escherichia coli clinical isolates in the UK. J Antimicrob Chemother 67: 878-885.

Smet A, Van Nieuwerburgh F, Vandekerckhove TT, Martel A, Deforce D, et al. (2010) Complete nucleotide sequence of CTX-M-15-plasmids from clinical Escherichia coli isolates: insertional events of transposons and insertion sequences. PLoS One 5: e11202.

Sandegren L, Linkevicius M, Lytsy B, Melhus A, Andersson DI (2012) Transfer of an Escherichia coli ST131 multiresistance cassette has created a Klebsiella pneumoniae-specific plasmid associated with a major nosocomial outbreak. J Antimicrob Chemother 67: 74-83.

Ho PL, Yeung MK, Lo WU, Tse H, Li Z, et al. (2012) Predominance of pHK01-like incompatibility group FII plasmids encoding CTX-M-14 among extended-spectrum beta-lactamase-producing Escherichia coli in Hong Kong, 1996-2008. Diagn Microbiol Infect Dis 73: 182-186.

DebRoy C, Sidhu MS, Sarker U, Jayarao BM, Stell AL, et al. (2010) Complete sequence of pEC14_114, a highly conserved IncFIB/FIIA plasmid associated with uropathogenic Escherichia coli cystitis strains. Plasmid 63: 53-60.

Perichon B, Bogaerts P, Lambert T, Frangeul L, Courvalin P, et al. (2008) Sequence of conjugative plasmid pIP1206 mediating resistance to aminoglycosides by 16S rRNA methylation and to hydrophilic fluoroquinolones by efflux. Antimicrob Agents Chemother 52: 2581-2592.

Zhao F, Bai J, Wu J, Liu J, Zhou M, et al. (2010) Sequencing and genetic variation of multidrug resistance plasmids in Klebsiella pneumoniae. PLoS One 5: e10141.

Partridge SR, Hall RM (2004) Complex multiple antibiotic and mercury resistance region derived from the r-det of NR1 (R100). Antimicrob Agents Chemother 48: 4250-4255.

Johnson TJ, Jordan D, Kariyawasam S, Stell AL, Bell NP, et al. (2010) Sequence analysis and characterization of a transferable hybrid plasmid encoding multidrug resistance and enabling zoonotic potential for extraintestinal Escherichia coli. Infect Immun 78: 1931-1942.

Hu P, Elliott J, McCready P, Skowronski E, Garnes J, et al. (1998) Structural organization of virulence-associated plasmids of Yersinia pestis. J Bacteriol 180: 5192-5202.

Gibbs MD, Spiers AJ, Bergquist PL (1993) RepFIB: a basic replicon of large plasmids. Plasmid 29: 165-179.

Zhang L, Lu X, Zong Z (2013) The emergence of blaCTX-M-15-carrying Escherichia coli of ST131 and new sequence types in Western China. Ann Clin Microbiol Antimicrob 12: 35.

Onnberg A, Soderquist B, Persson K, Molling P (2014) Characterization of CTX-M-producing Escherichia coli by repetitive sequence-based PCR and real-time PCR-based replicon typing of CTX-M-15 plasmids. APMIS.

Takahashi H, Shao M, Furuya N, Komano T (2011) The genome sequence of the incompatibility group Igamma plasmid R621a: evolution of IncI plasmids. Plasmid 66: 112-121.

Fricke WF, Mammel MK, McDermott PF, Tartera C, White DG, et al. (2011) Comparative genomics of 28 Salmonella enterica isolates: evidence for CRISPR-mediated adaptive sublineage evolution. J Bacteriol 193: 3556-3568.

Johnson TJ, Bielak EM, Fortini D, Hansen LH, Hasman H, et al. (2012) Expansion of the IncX plasmid family for improved identification and typing of novel plasmids in drug-resistant Enterobacteriaceae. Plasmid 68: 43-50.
Lescat M, Calteau A, Hoede C, Barbe V, Touchon M, et al. (2009) A module located at a chromosomal integration hot spot is responsible for the multidrug resistance of a reference strain from Escherichia coli clonal group A. Antimicrob Agents Chemother 53: 2283-2288.
Gokulan K, Khare S, Rooney AW, Han J, Lynne AM, et al. (2013) Impact of plasmids, including those encodingVirB4/D4 type IV secretion systems, on Salmonella enterica serovar Heidelberg virulence in macrophages and epithelial cells. PLoS One 8: e77866.
Schluter A, Szczepanowski R, Puhler A, Top EM (2007) Genomics of IncP-1 antibiotic resistance plasmids isolated from wastewater treatment plants provides evidence for a widely accessible drug resistance gene pool. FEMS Microbiol Rev 31: 449-477.
Venturini C, Hassan KA, Roy Chowdhury P, Paulsen IT, Walker MJ, et al. (2013) Sequences of two related multiple antibiotic resistance virulence plasmids sharing a unique IS26-related molecular signature isolated from different Escherichia coli pathotypes from different hosts. PLoS One 8: e78862.
Dotto G, Giacomelli M, Grilli G, Ferrazzi V, Carattoli A, et al. (2014) High prevalence of oqxAB in Escherichia coli isolates from domestic and wild lagomorphs in Italy. Microb Drug Resist 20: 118-123.
Sullivan MJ, Petty NK, Beatson SA (2011) Easyfig: a genome comparison visualizer. Bioinformatics 27: 1009-1010.
Yoshii A, Moriyama H, Fukuhara T (2012) The novel kasugamycin 2'-N-acetyltransferase gene aac(2')-IIa, carried by the IncP island, confers kasugamycin resistance to rice-pathogenic bacteria. Appl Environ Microbiol 78: 5555-5564.
Garcillan-Barcia MP, Francia MV, de la Cruz F (2009) The diversity of conjugative relaxases and its application in plasmid classification. FEMS Microbiol Rev 33: 657-687.
Song Y, Tong Z, Wang J, Wang L, Guo Z, et al. (2004) Complete genome sequence of Yersinia pestis strain 91001, an isolate avirulent to humans. DNA Res 11: 179-197.
Soler Bistue AJ, Birshan D, Tomaras AP, Dandekar M, Tran T, et al. (2008) Klebsiella pneumoniae multiresistance plasmid pMET1: similarity with the Yersinia pestis plasmid pCRY and integrative conjugative elements. PLoS ONE 3: e1800.
Tomizawa JI, Ohmori H, Bird RE (1977) Origin of replication of colicin E1 plasmid DNA. Proc Natl Acad Sci U S A 74: 1865-1869.
Hofinger C, Karch H, Schmidt H (1998) Structure and function of plasmid pColD157 of enterohemorrhagic Escherichia coli O157 and its distribution among strains from patients with diarrhea and hemolytic-uremic syndrome. J Clin Microbiol 36: 24-29.
Rijavec M, Budic M, Mrak P, Muller-Premru M, Podlesek Z, et al. (2007) Prevalence of ColE1-like plasmids and colicin K production among uropathogenic Escherichia coli strains and quantification of inhibitory activity of colicin K. Appl Environ Microbiol 73: 1029-1032.
Budic M, Rijavec M, Petkovsek Z, Zgur-Bertok D (2011) Escherichia coli bacteriocins: antimicrobial efficacy and prevalence among isolates from patients with bacteraemia. PLoS One 6: e28769.
Oshima K, Toh H, Ogura Y, Sasamoto H, Morita H, et al. (2008) Complete genome sequence and comparative analysis of the wild-type commensal Escherichia coli strain SE11 isolated from a healthy adult. DNA Res 15: 375-386.
Lu S, Zhang X, Zhu Y, Kim KS, Yang J, et al. (2011) Complete genome sequence of the neonatal-meningitis-associated Escherichia coli strain CE10. J Bacteriol 193: 7005.

Brolund A, Franzen O, Melefors O, Tegmark-Wisell K, Sandegren L (2013) Plasmidome-analysis of ESBL-producing escherichia coli using conventional typing and high-throughput sequencing. PLoS One 8: e65793.

Liu P, Li P, Jiang X, Bi D, Xie Y, et al. (2012) Complete genome sequence of Klebsiella pneumoniae subsp. pneumoniae HS11286, a multidrug-resistant strain isolated from human sputum. J Bacteriol 194: 1841-1842.

Poirel L, Bonnin RA, Nordmann P (2011) Analysis of the resistome of a multidrug-resistant NDM-1-producing Escherichia coli strain by high-throughput genome sequencing. Antimicrob Agents Chemother 55: 4224-4229.

Partridge SR, Paulsen IT, Iredell JR (2012) pJIE137 carrying blaCTX-M-62 is closely related to p271A carrying blaNDM-1. Antimicrob Agents Chemother 56: 2166-2168.

Chen YT, Lin AC, Siu LK, Koh TH (2012) Sequence of closely related plasmids encoding bla(NDM-1) in two unrelated Klebsiella pneumoniae isolates in Singapore. PLoS One 7: e48737.

Text S2: PLACNET validation by reconstruction of finalized genomes

Methods

The objective of this report is the validation of PLACNET by plasmid reconstruction in genomes that were published as complete, finished genomes. The ten analyzed genomes are shown in **Table S2**. Genome sequences were downloaded from NCBI in FASTA format. As a consequence of using FASTA, all sequences were considered as linear DNA sequences. The ART software (Huang et al., 2012) was used to simulate pair-end Illumina reads. Main parameters used were: read length, 101 bp; insert size, 400 bp; standard deviation of fragment size for pair end, 10 bp. For each plasmid, the relative coverage was estimated according to its reported or estimated copy number, as shown in **Table S2**. Resulting reads were analyzed using the workflow shown in **Figure 6** of the main text, strictly following the same steps as in plasmid reconstruction of real data. For the analysis of each individual genome, the query genome sequences (including plasmids) were subtracted from the local megablast NCBI database. For each reconstructed plasmid, we define the ***error rate*** as the % of total DNA sequence of the finalized plasmid that is lost (or misplaced) in the plasmid reconstruction.

Besides, DNA from strains FV9873, E35BA, E2022 and E61BA was extracted as explained in Materials and Methods of Valverde et al., 2009. DNA samples were digested with S1-nuclease and visualized after Pulsed-Field Gel Elecrophoresis (PFGE) using exactly the conditions described by Valverde et al., 2009.

Exemplar reconstruction of a multi-plasmid genome: *E. coli* ST131 ExPEC strain JJ1886

Figure S25 shows the original and pruned Cytoscape networks of the JJ1886 genome, constructed as explained in Materials and Methods. PLACNET-based plasmid reconstruction is summarized in **Figures S26** and **S27**. **Figure S26** shows plasmid definition in the pruned network. Four single contigs (surrounded by colored circles in the figure) represent plasmids p1 to p4, since they contain a Replication Initiator Protein (RIP) and/or a relaxase protein (REL). These are unambiguous assignations, defined in the inset table (Step 2). One additional potential plasmid is shown by a chromosome-connected network with two replication proteins. It is defined as p5 by the red circle in **Figure S27**. p5 is a 16 contig plasmid, adding a total of 104,416 bp. It is an IncF plasmid, according to its two RIP proteins (yellow-tagged contigs containing IncFrepb and IncFIA in **Figure S27**). Four small contigs (shown in the Blastn descriptions in the inset Table Step 3), corresponding to known hubs, were duplicated. The resulting summary of plasmid definitions is shown in **Table S3**. As shown in the Table, plasmid reconstructions give almost zero error rate for plasmids p1 to p4. The 5.6 kb DNA sequences missing in the PLACNET reconstruction of plasmid p5 correspond to insertion sequences (IS66, ISEC23 and IS26) that were repeated two, two and six times in the original sequence but were considered only once in the reconstructed plasmid. The IS elements appeared after the assembly as either isolated contigs or linked

to one of the unique plasmid sequences. Although these IS elements show additional scaffold links, IS multiplication within the plasmid was not attempted. Nevertheless, as seen in **Figure S28**, PLACNET reconstruction includes essentially all sequences in plasmid p5. Conversely, only two very small contigs (nodes 104 and 120) corresponding to IS911 transposases, are not correctly assigned to plasmid p5, as shown in **Figure S29**. These nodes were hubs, (wrongly) duplicated during network pruning, due to their link to references connecting the chromosome with the plasmid.

Exemplar reconstruction of single-plasmid genomes: *E. coli* ST131 commensal strain SE15 and UPEC strain UTI89

The reconstruction of strain SE15 genome is summarized in **Figures S30** to **S33**. **Figure S30** shows the transition from the original to the pruned SE15 network, underscoring the presence of one RIP and one REL protein (color-tagged contigs). **Figure S31** helps interpreting the network. Three nodes, loosely bound to the chromosome, correspond to chromosomal sequences, as inferred from nearest BLAST hits (see the Blastn descriptions of the red background nodes in Inset Table Step 2). Three other "hub" nodes, which were duplicated, are also indicated with a green circle in the network and described in Inset table Step 2. The IncF plasmid was reconstructed from 15 contigs, as shown in **Figure S32**. As shown in the figure, there are only three differences (totaling 5,709 bp) between plasmid pECSF1 and the reconstructed plasmid. They correspond to three copies of IS66 (two complete copies plus one incomplete copy of the 2.436 bp element), which is also contained in the main chromosome. No node was wrongly assigned to the plasmid (data not shown).

The reconstruction of strain UTI89 genome was straight forward. Since it presented no special problems, the detailed steps of the reconstruction are not shown. The results are summarized in **Table S3**. As can be seen, the IncF plasmid is almost perfectly reconstructed, with just 0.5% error rate.

Additional plasmid reconstructions, underscoring main PLACNET reconstruction problems

Seven other *E. coli* genomes were reconstructed to validate PLACNET as well as to outline potential problems in plasmid reconstruction. A summary of the results is shown in **Table S3**. In general, reconstruction of small plasmids is almost perfect most of the times, with error rates <1%. Genomes containing large plasmids are also straightforwardly reconstructed if the corresponding genome contains a single large plasmid, even in the presence of co-resident small plasmids, as shown by the reconstruction of the genome of strain SMS-3-5 (**Table S3**) and that of the previously discussed strain JJ1186. In most cases, the error rate for large plasmids is <5%. Genomes with one large plasmid (with or without small plasmids) were the most commonly encountered situation in the set of 68 sequenced *E. coli* genomes as of August 2014.

Nevertheless, more complicated situations were sometimes found. The next two genomes in **Table S3** (corresponding to strain MG1655 carrying either the recombinant plasmid pEC_L46 or the two separate plasmids pEC958 and R46) illustrate PLACNET discrimination in the reconstruction of a plasmid

cointegrate with respect to the same strain containing the plasmids that originated the cointegrate as two separate genetic entities. In this particular case, the cointegrate is nicely reconstructed with 3% error rate. On the other hand, the reconstruction of the 2-plasmid strain is more problematic. The IncN plasmid cannot be fully reconstructed (resulting in a 19% error) by the lack of assignment of a five contig set (containing sequences of the In1 integron) with the IncN backbone. This problem is represented in the Cytoscape network shown in **Figure S33**. The In1 element (5 contig element circled in blue) should belong to one or the other plasmid in the strain, but this cannot be decided by the lack of scaffold links. In this particular case (as in others), relying solely on reference links does not help, since the In1 integron is mobile and thus could be linked to different backbones. This result underscores the crucial importance for the dataset to provide sufficient scaffold links.

The three next genomes (corresponding to pathogenic *E. coli* strains SS17, RM12581 and 11368), are examples of strains containing two or more large plasmids in genome datasets that were highly fragmented in the assembled datasets (500 - 700 contigs; see **Table S2**). In these situations, PLACNET analysis becomes more complicated, since the number of contigs and reference/scaffold links to analyze grow disproportionately with the number of contigs. In such cases, we found it convenient to eliminate contigs >200 bp, which was the default threshold. By taking out additional (small) contigs, we could reconstruct the plasmids, but sometimes incurring in a penalty in the error rate of assignment (up to 15%). Results are shown in **Table S3**. For strain SS17, it was sufficient to eliminate contigs <350 bp. Using this modification, the reconstructed plasmids showed error rates of 1-3%. In the cases of RM12581 and 11368, all contigs <500 bp had to be eliminated for proper plasmid reconstruction, resulting in error rates up to 15%. This result underscores another property of PLACNET: plasmid definition improves by increasing contig size elimination in the pruning step, but this comes at a price in terms of a larger error rate in the reconstructed plasmid.

The last genome in **Table S3** reports the only case we analyzed in which PLACNET failed to separate individual plasmids in an *E. coli* genome. It corresponds to the ETEC strain H0407. The strain contains two IncF plasmids of 66,681 bp and 94,797 bp. These two plasmids show extensive homologous regions, in which DNA identity is >90% (incidentally, they can be compatible, belonging to IncFI and IncFII incompatibility groups, for instance, due to a few point mutations in their replication region). Thus, it is impossible for the Illumina assembling programs to distinguish among many DNA segments that are almost identical in both plasmids. Thus, the Cytoscape representation of these plasmids is a densely connected network, as shown in **Figure S34**. This is because the non-segregated sequences produce scaffold links with contigs belonging to both plasmids. As an example, there is just a single contig containing the REL and RIP proteins of both plasmids (red arrow in **Figure S34**).

Confirmation of plasmid sizes by S1-PFGE.

The use of S1-PFGE allows the visualization of plasmids in a DNA sample as well as the estimation of the molecular weight (Barton et al., 1995). Table S4 shows a summary of the results obtained by S1-PFGE on the four strains (FV9873, E35BA, E2022 and E61BA) that were sequenced for this work. As can be seen in the table, there is a good correlation between the number and molecular size of the plasmids as estimated by S1-PFGE when compared to PLACNET reconstructions. The only significant difference is the

identification of two molecular species of aprox. 140 and 75 kb in the case of S1-PFGE that could not be differentiated by PLACNET reconstruction of strain E35BA. Nevertheless, PLACNET suggested the existence of two plasmid species based on the existance of two different relaxases in this genome. It shoud be also noted that S1-PFGE does not allow the visualization of plasmids of less than 15 kb.

Conclusions

PLACNET correctly identifies plasmids in Illumina-based WGS datasets. With practically no exception, PLACNET identifies and assembles plasmid backbones either in a single contig or as a connected component of several contigs. Plasmids containing multiple contigs may not be covered to 100% of their sequences, sometimes carrying a small error rate in the contigs assigned to each plasmid. The error is very low (<1%) for small plasmids, but can be as high as 20% for genomes containing several large plasmids. There are two main sources of error:

1. Repeated sequences originating from mobile genetic elements (MGEs). They are usually spotted as hubs, because they show multiple scaffold links, and their contigs are hits to known MGEs. Two alternative strategies are used by PLACNET to deal with repeated sequences. They can be eliminated, mainly if the contigs are small, or duplicated. As a result, the network connectivity diminishes and allows the identification of disjoint connected networks, as explained in the text.

2. Plasmids with extensive regions of high homology. This is in fact equivalent to the presence massive repeated sequences. In these cases, the relevant plasmids cannot be separated. They are identified by their signature sequences (REL and/or RIP).

Sources of error are more relevant if:

- A particular genome assembly results in highly fragmented genomes (400 contigs or more). This can be minimized by combining paired-end with mate-pair sequencing.

- The analyzed genome contains many repeated sequences. This is an intrinsic source of error that can only be minimized by improving the quality of the assembly process.

- There is a lack of a sufficiently wide reference dataset. This was clearly not a problem with *E. coli*, but certainly is for other genomes for which no many plasmid sequences are available. Future work will deal with this problem, when we report on plasmid reconstruction for genomes of other bacteria.

Besides, the existence and molecular size of the plasmids reconstructed by PLACNET from the four strains sequenced in our laboratory were confirmed by S1-PFGE. In all cases, but the two IncF plasmids of strain E35BA, which could be not separated by PLACNET, all plasmids were correctly identified and their molecular sizes were calculated with acceptable error rate (less than 3%).

Supplementary reference

Huang, W., Li, L., Myers, J.R., and Marth, G.T. (2012). ART: a next-generation sequencing read simulator. Bioinformatics (Oxford, England) *28*, 593-594.

Valverde, A., Cantón, R., Garcillán-Barcia, M.P., Novais, A., Galán, J.C., Alvarado, A., de la Cruz, F., Baquero, F., Coque, T.M. (2009). Spread of *bla*$_{CTX-M-14}$ is driven mainly by IncK plasmids disseminated among *Escherichia coli* phylogroups A, B1, and D in Spain. Antimicrob Agents Chemother *53(12),* 5201-5212.

Barton, B. M., Harding, G.P., Zuccarelli, A.J. (1995). A general method for detecting and sizing large plasmids. Anal Biochem, *226*, 235-240.

Table S2. Ten *E. coli* genomes analyzed as examples of PLACNET performance [1]

Genome[2]	Plasmid	Accession no.	Estimated copy number	Coverage	Nº contigs	Total bp	N50 (bp)	Kmer	Longest contig (bp)
JJ1886 (ST131, ExPEC)	pJJ1886_1	NC_022661	20	2000x	150	5,211,142	237,023	79	710,527
	pJJ1886_2	NC_022649	10	1000x					
	pJJ1886_3	NC_022662	10	1000x					
	pJJ1886_4	NC_022650	6	600x					
	pJJ1886_5	NC_022651	1	100x					
	chromosome	NC_022648	1	100x					
SE15 (ST131, ExPEC)	pECSF1	NC_013655	1	100x	91	4,785,093	367,892	79	659,604
	chromosome	NC_013654	1	100x					
UTI89 (ExPEC)	pUTI89	NC_007941	1	100x	124	5,129,511	240,702	79	724,339
	chromosome	NC_007946	1	100x					
SMS-3-5 (Environmental)	pSMS35_3	NC_010487	50	5000x	188	5,169,770	200,074	83	377,480
	pSMS35_4	NC_010486	50	5000x					
	pSMS35_8	NC_010485	20	2000x					
	pSMS35_130	NC_010488	1	100x					
	chromosome	NC_010498	1	100x					
MG1655 + pEC_L46	pEC_L46	NC_014385	1	100x	146	4,703,719	176,611	83	327,117
	chromosome	NC_000913	1	100x					
MG1655 + pEC958 + R46	pEC958	HG941719	1	100x	171	4,761,832	176,611	83	327,117
	R46	NC_003292	1	100x					
	chromosome	NC_000913.3	1	100x					
O157:H7 str. SS17 (EHEC)	pO157	CP008807	1	100x	556	5,481,778	179,816	83	386,831
	pSS17	CP008806	2	100x					
	chromosome	CP008805	1	100x					
O145:H28 str. RM12581 (STEC)	pRM12581	CP007137	1	100x	646	5,497,629	137,834	79	264,117
	PO145-12581	CP007138	1	100x					
	chromosome	CP007138	1	100x					
O26:H11 str. 11368 (EHEC)	pO26_1	NC_013369	1	100x	707	5,616,433	104,130	81	250,500
	pO26_2	NC_013362	1	100x					
	pO26_3	NC_013363	10	1000x					
	pO26_4	NC_014543	10	1000x					
	chromosome	NC_013361	1	100x					
H0407 (ETEC)	p52	NC_017721	10	1000x	336	5,205,705	87,499	83	248,902
	p58	NC_017723	10	1000x					
	p666	NC_017722	1	100x					
	p948	NC_017724	1	100x					
	chromosome	NC_017633	1	100x					

[1] The Table shows, for each genome, the accession number for each plasmid and chromosome, the estimated relative copy number of each plasmid, and some details of the simulated assembly of Illumina reads, as explained in the Method section.

[2] ExPEC: Extraintestinal pathogenic *Escherichia coli*, EHEC: Enterohemorrhagic *Escherichia coli*; STEC: Shiga toxin-producing *Escherichia coli;* ETEC: Enterotoxigenic *Escherichia coli.*

Table S3. PLACNET performance on a set of ten *E. coli* genomes

Genome[1]	Plasmid	Accession no.	MOB group	Inc group	PLACNET size (bp)	NCBI size (bp)	Delta-size (% error)
JJ1886 (ST131, ExPEC)	pJJ1886_1	NC_022661	-	-	1,552	1,552	0 bp (0 %)
	pJJ1886_2	NC_022649	MOB_{Q12}	-	5,167	5,167	0 bp (0 %)
	pJJ1886_3	NC_022662	MOB_{P51}	ColE1-like	5,601	5,631	30 bp (0.5 %)
	pJJ1886_4	NC_022650	MOB_{P11}	IncP1	55,955	55,956	1 bp (0.002 %)
	pJJ1886_5	NC_022651	MOB_{F12}	IncF	104,426	110,040	5,614 bp bp (5 %)
	chromosome	NC_022648	-	-	5,038,008	5,129,938	91,930 bp (2 %)
UTI89 (ExPEC)	pUTI89	NC_007941	MOB_{F12}	IncF	113,653	114,230	577 bp (0.5 %)
	chromosome	NC_007946	-	-	5,033,656	5,065,741	32,085 bp (0.6 %)
SE15 (ST131, ExPEC)	pECSF1	NC_013655	MOB_{F12}	IncF	116,636	122,345	5,709 bp (5 %)
	chromosome	NC_013654	-	-	4,666,391	4,717,338	50,947 bp (1 %)
SMS-3-5 (Environmental)	pSMS35_3	NC_010487	-	-	3,496	3,565	69 bp (2 %)
	pSMS35_4	NC_010486	MOB_{Qu}	-	3,981	4,074	93 bp (2 %)
	pSMS35_8	NC_010485	MOB_{P51}	ColE1-like	8,909	8,909	0 bp (0 %)
	pSMS35_130	NC_010488	MOB_{F12}	IncF	122,954	130,440	7,486 bp (6 %)
	chromosome	NC_010498	-	-	5,039,290	5,068,389	29,099 bp (1 %)
MG1655 + pEC_L46	pEC_L46	NC_014385	MOB_{F11} &MOB_{F12}	IncN & IncF	141,163	144,871	3,708 bp (3 %)
	chromosome	NC_000913	-	-	4,582,408	4,641,652	59,244 bp (1 %)
MG1655 + pEC958 + R46	pEC958	HG941719	MOB_{F12}	IncF	129,928	135,602	5,674 bp (4 %)
	R46	NC_003292	MOB_{F11}	IncN	41,266	50,969	9,703 bp (19 %)
	chromosome	NC_000913	-	-	4,581,208	4,641,652	60,444 bp (1 %)
O157:H7 str. SS17 (EHEC)[2]	pO157	CP008806	$MOB_{P12\text{-}like}$ & ΔTraI-MOB_{F12}	IncF	92,025	94,645	2,620 bp (3 %)
	pSS17	CP008807	MOB_{P6}	IncI2	37,122	37,447	325 bp (0.9 %)
	chromosome	CP008805	-	-	5,293,573	5,523,849	230,276 bp (4 %)
O145:H28 str. RM12581 (STEC)[3]	pRM12581	CP007137	-	IncA/C	61,449	64,562	3,113 bp (5 %)
	PO145-12581	CP007138	MOB_{P12}	IncB/O & IncF	74,498	87,120	12,622 bp (15 %)
	chromosome	CP007138	-	-	5,264,714	5,585,611	320,897 bp (6 %)
O26:H11 str. 11368 (EHEC)[3]	pO26_1	NC_013369	MOB_{P12}	IncB/O & IncF	74,108	85,167	11,059 bp (13 %)
	pO26_2	NC_013362	MOB_{F12}	IncF	63,357	63,365	8 bp (0.01 %)
	pO26_3	NC_013363	MOB_{P51}	ColE1-like	5,031	5,686	655 bp (12 %)
	pO26_4	NC_014543	-	-	3,690	4,073	383 bp (9 %)
	chromosome	NC_013361	-	-	5,353,548	5,697,240	343,692 bp (6 %)
H0407 (ETEC)	p52	NC_017721	MOB_{P51}	ColE1-like	5,175	5,175	0 bp (0 %)
	p58	NC_017723	MOB_{P51}	ColE1-like	5,799	5,800	1 bp (0.02 %)
	p666	NC_017722	MOB_{F12}	IncF	118,831	66,681	-
	p948	NC_017724	-	IncF		94,797	-
	chromosome	NC_017633	-	-	5,070,175	5,153,435	83,260 bp (1.6 %)

[1]ExPEC: Extraintestinal pathogenic Escherichia coli, EHEC: Enterohemorrhagic Escherichia coli; STEC: Shiga toxin-producing Escherichia coli; ETEC: Enterotoxigenic Escherichia coli. [2]In these genomes contigs <350 bp were removed. [3]In this genome contigs <500 bp were removed.

Figure S1

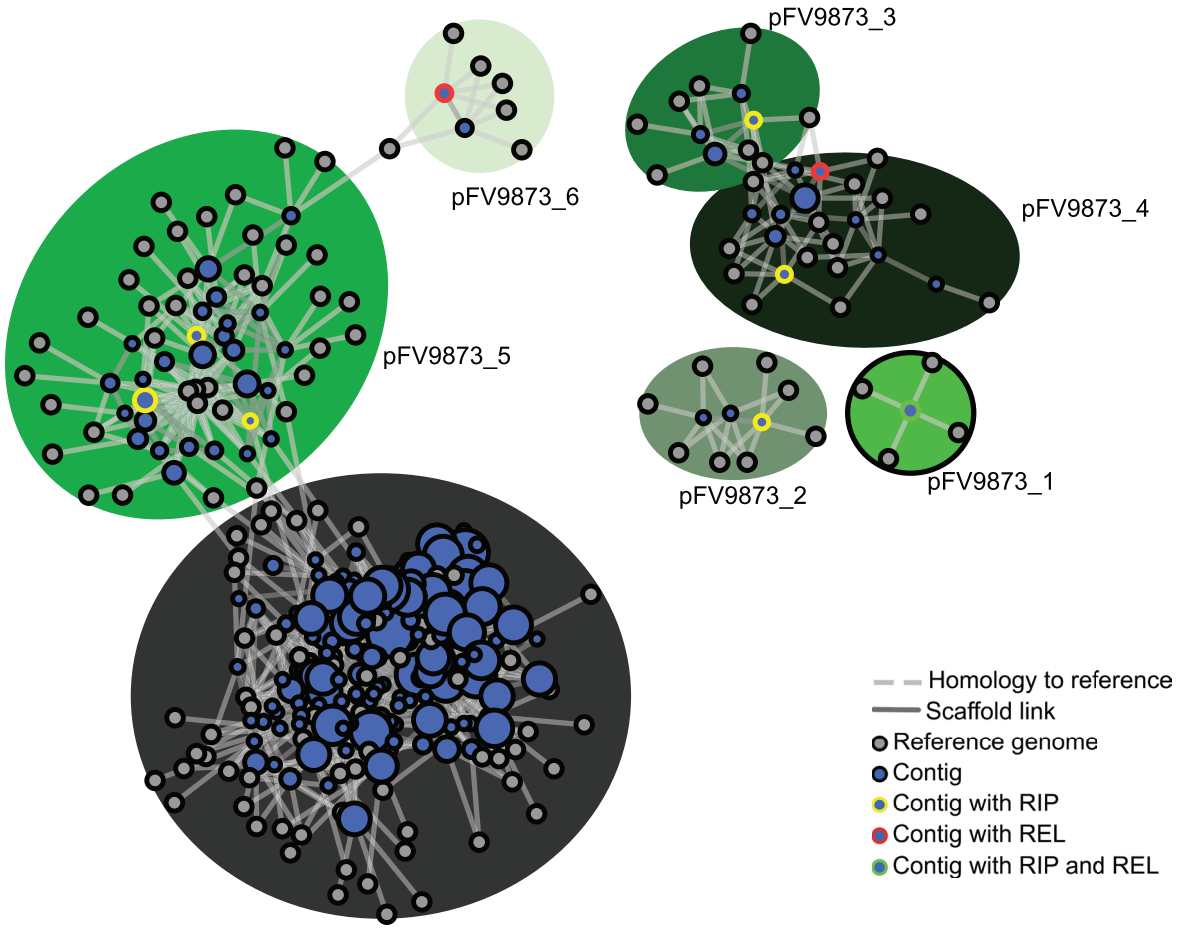


Figure S2

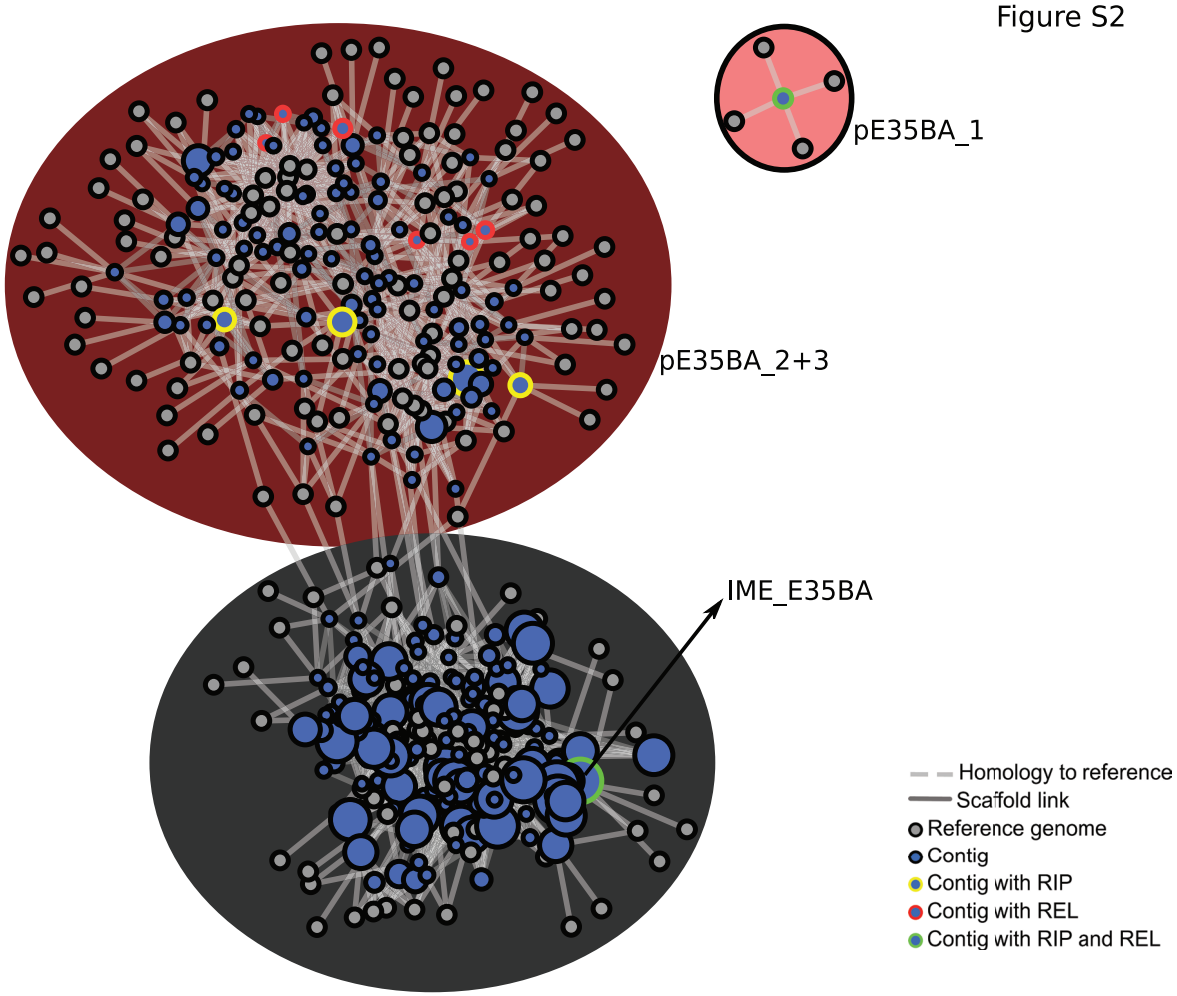


Figure S3

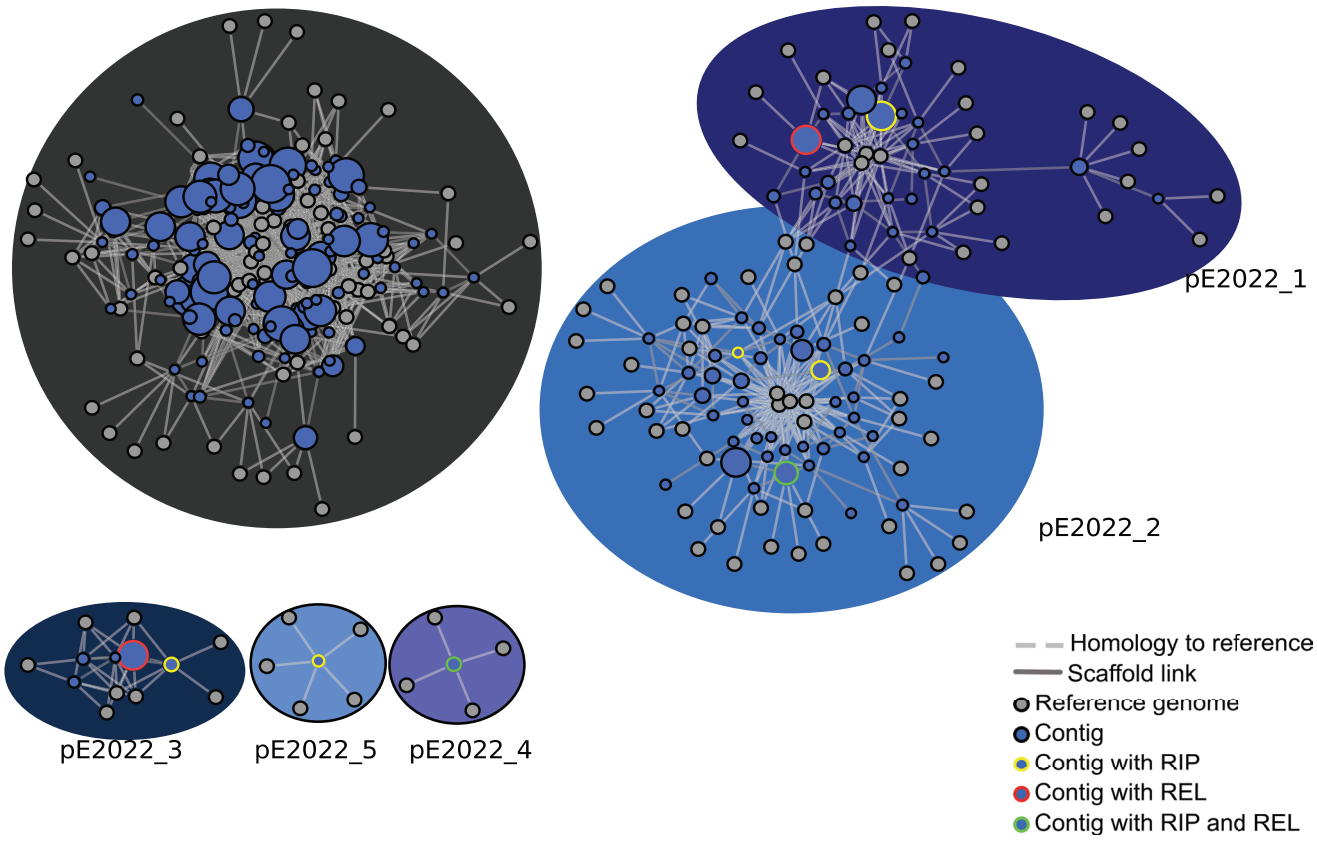


Figure S4

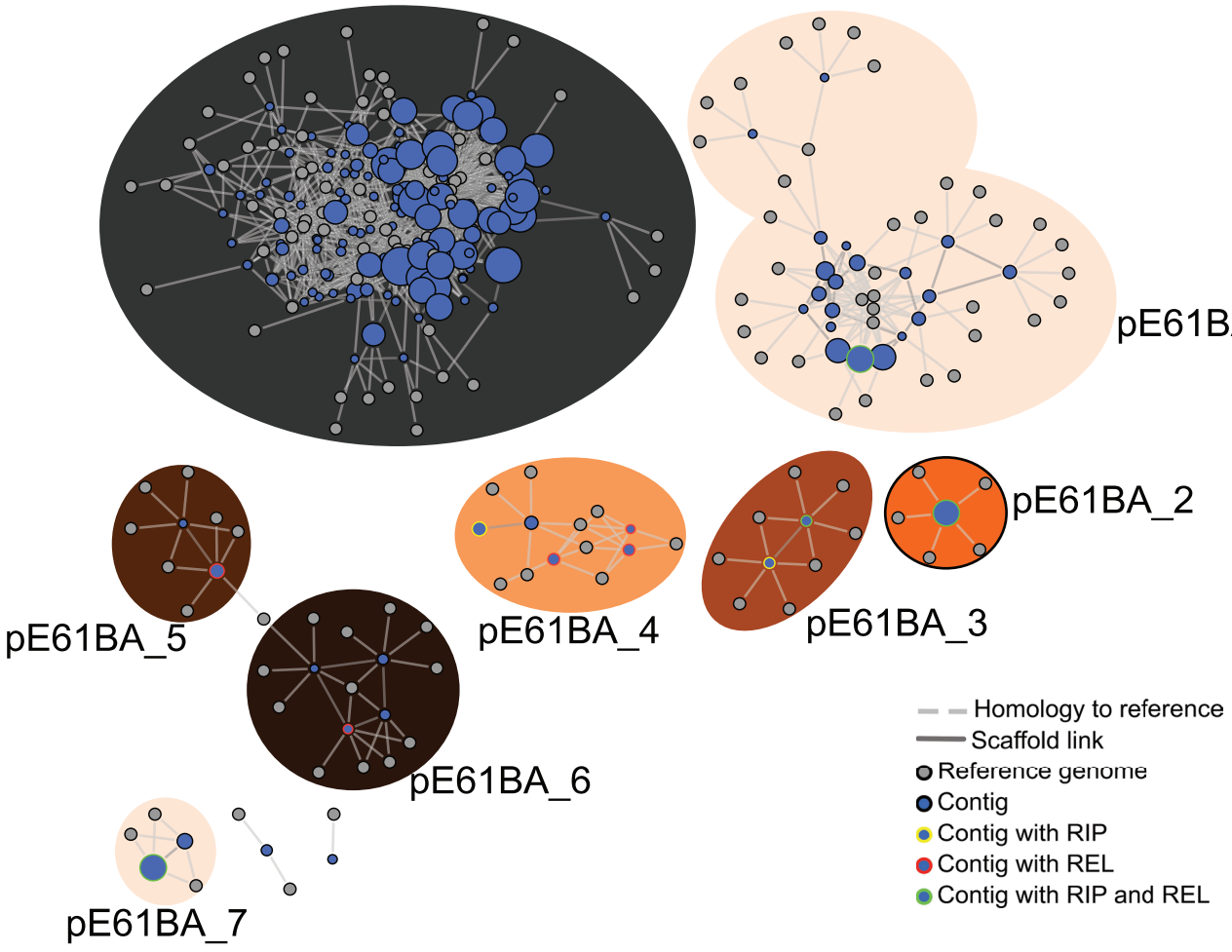


Figure S5

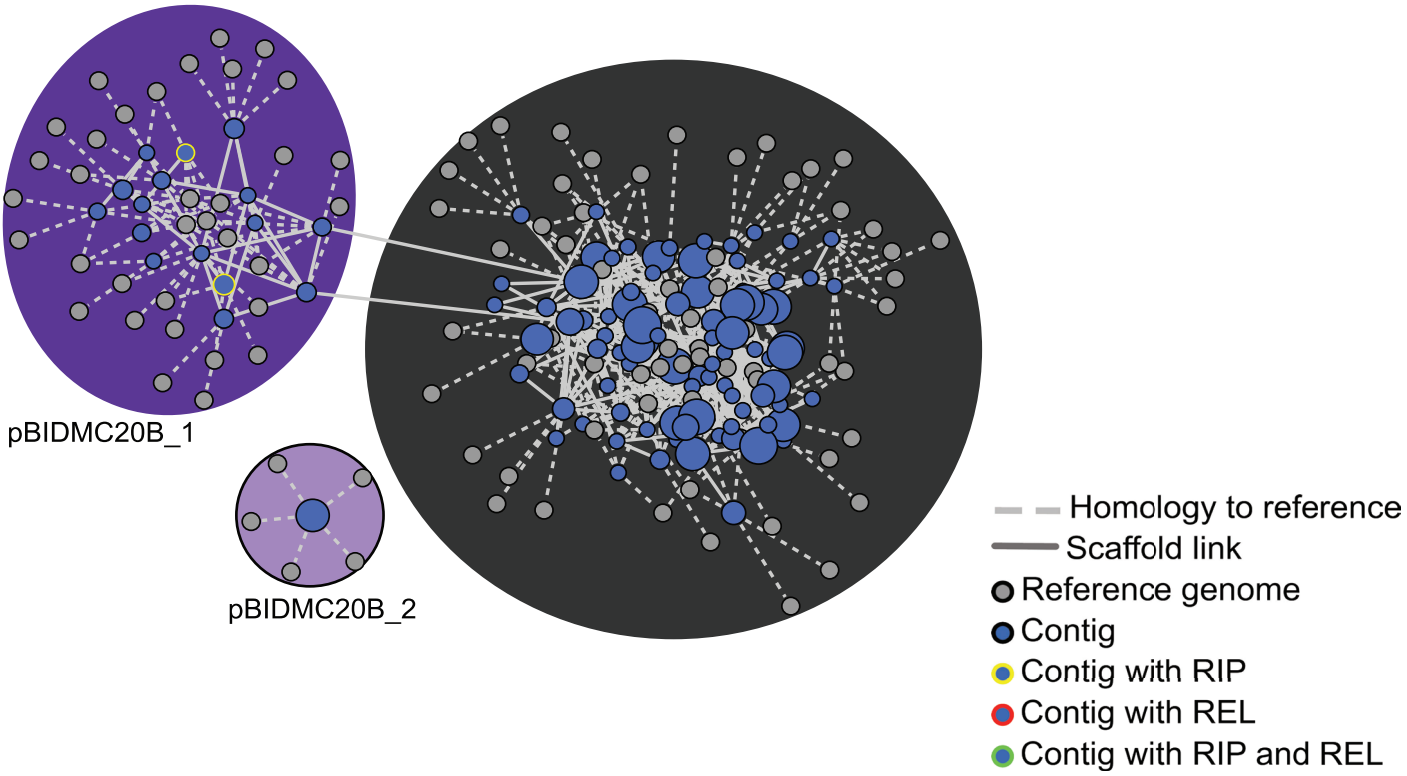


Figure S6

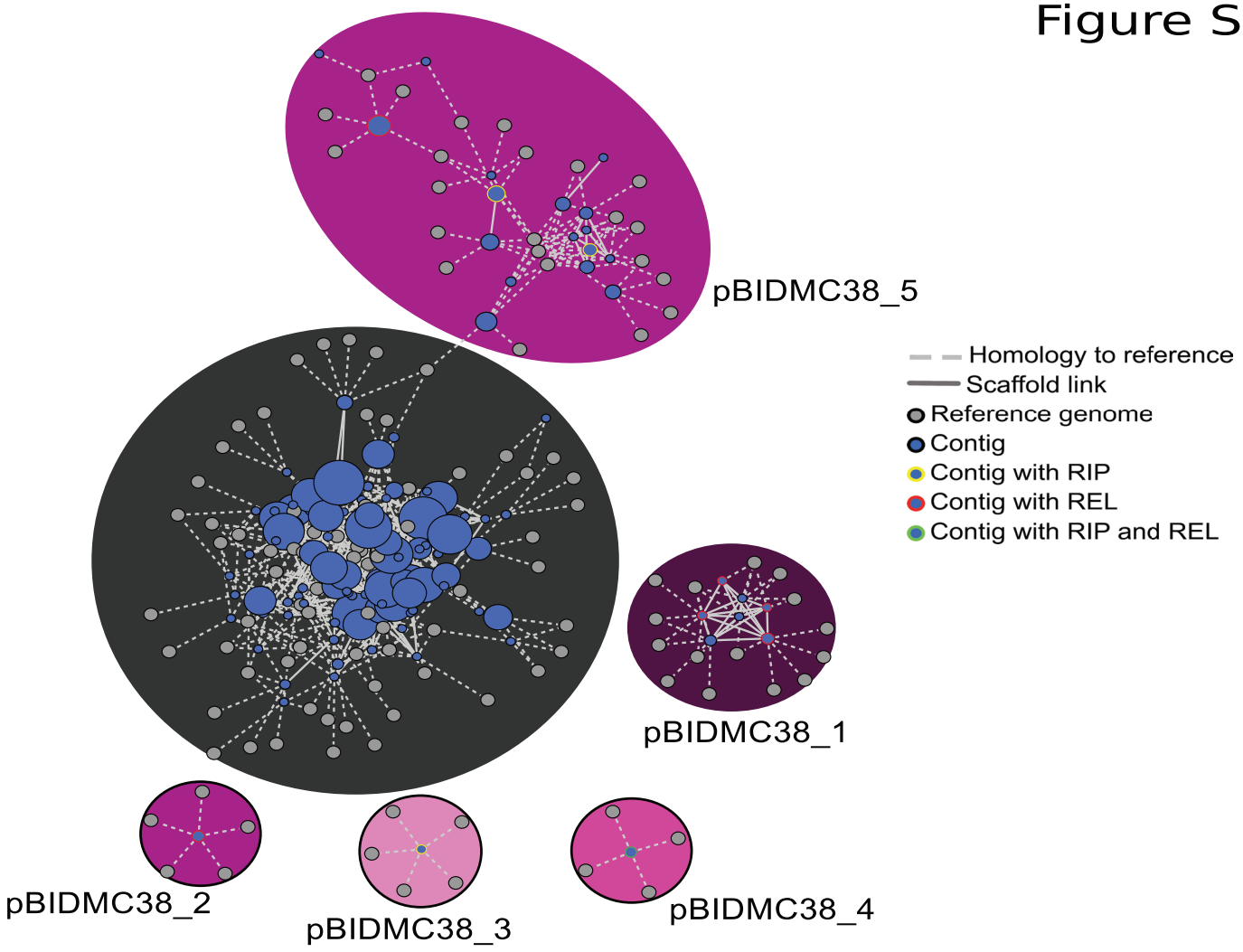


Figure S7

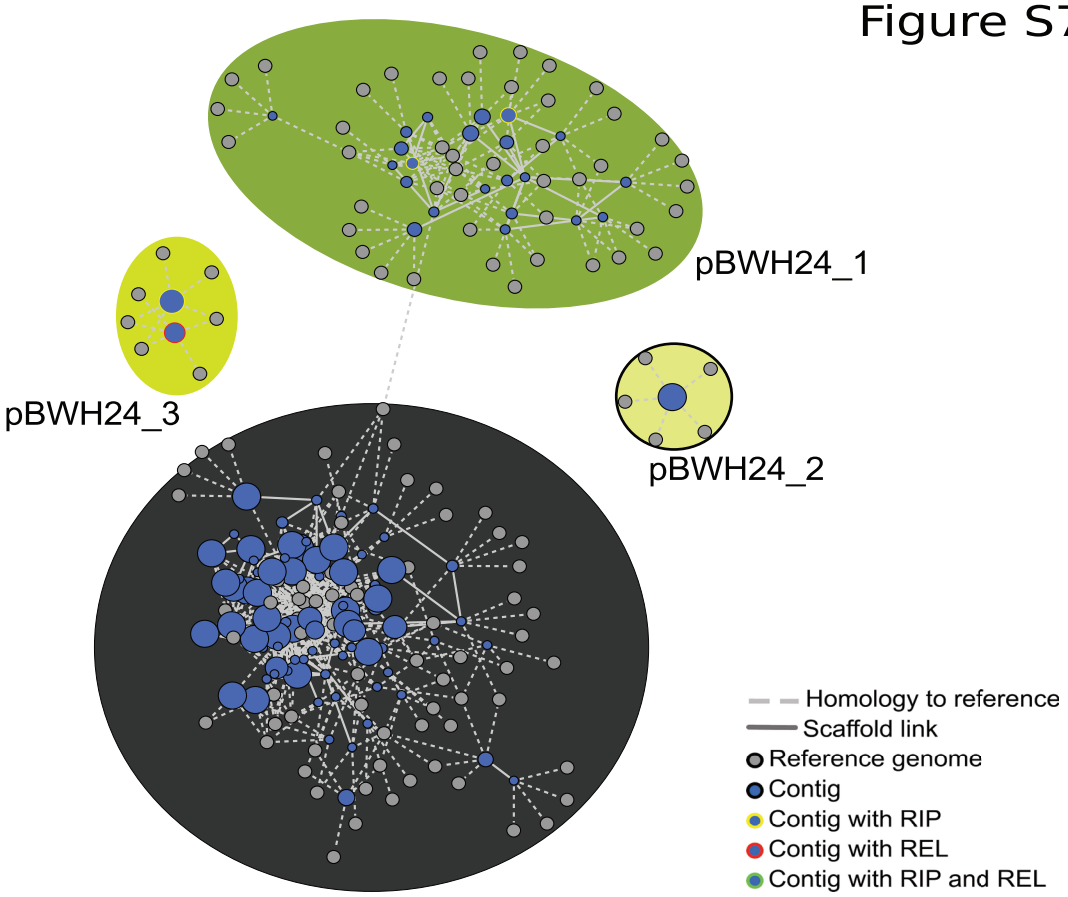


Figure S8

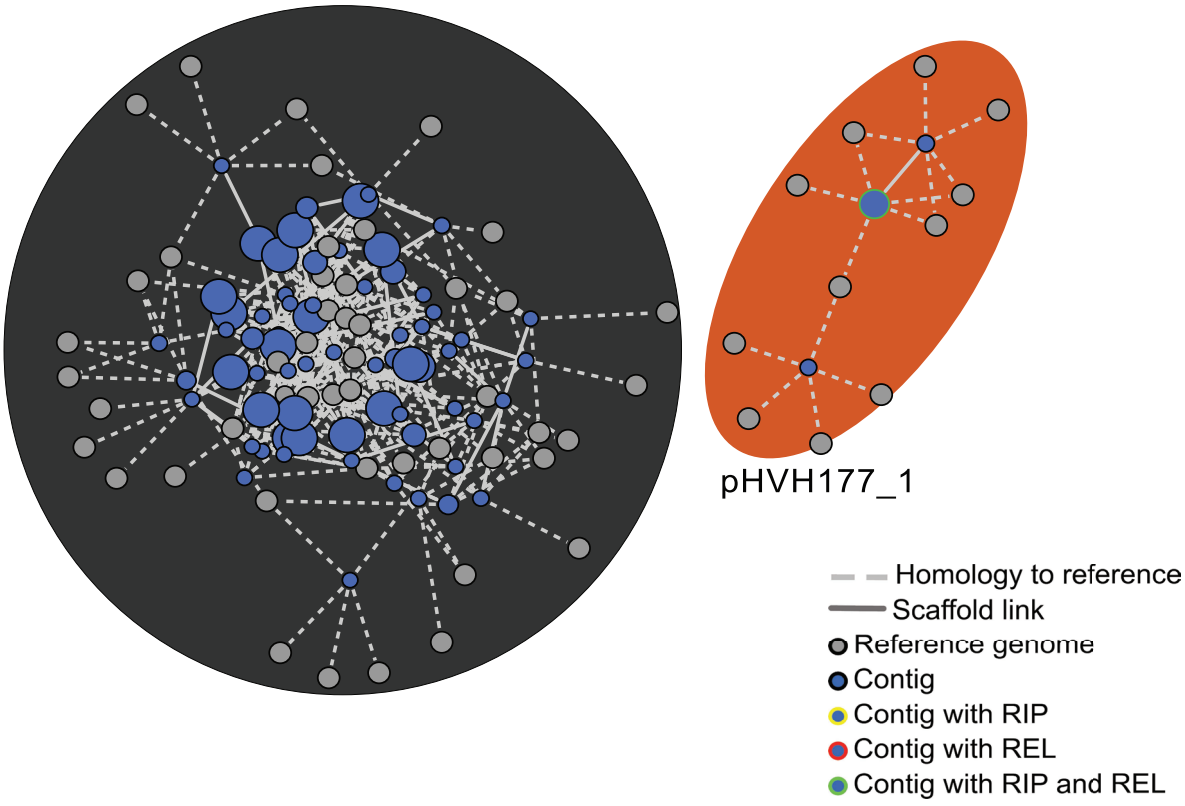


Figure S9

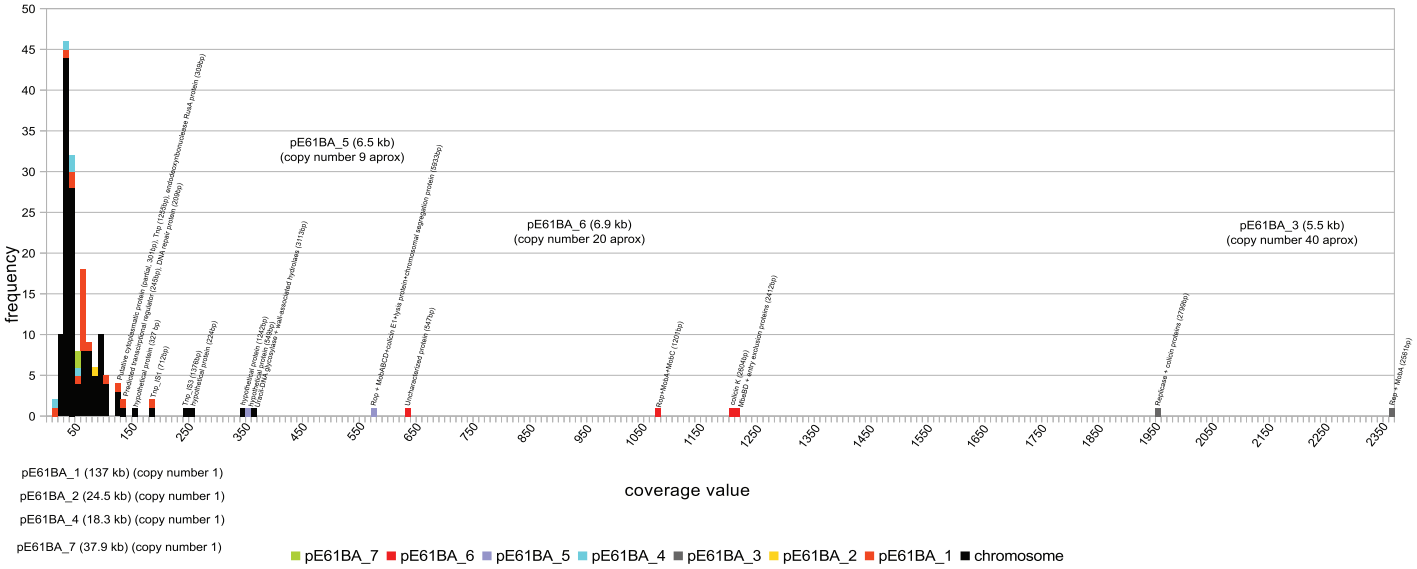


Figure S10

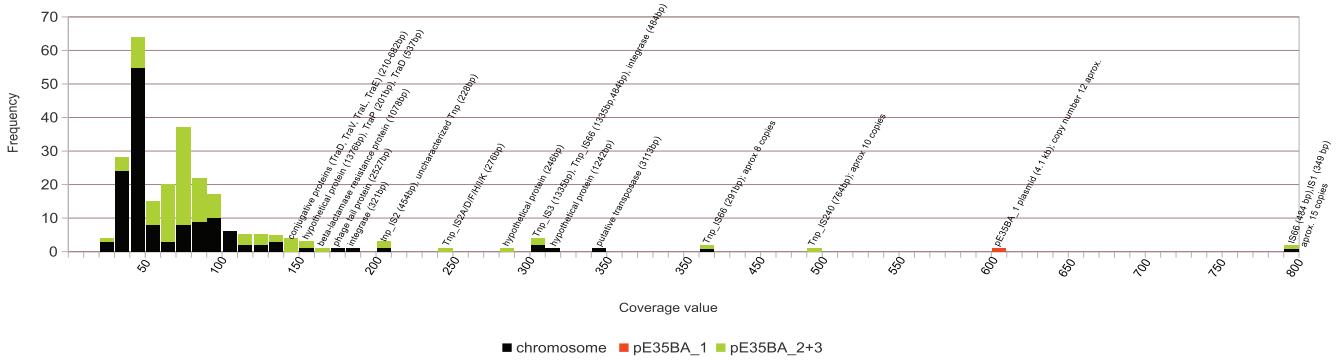


Figure S11A

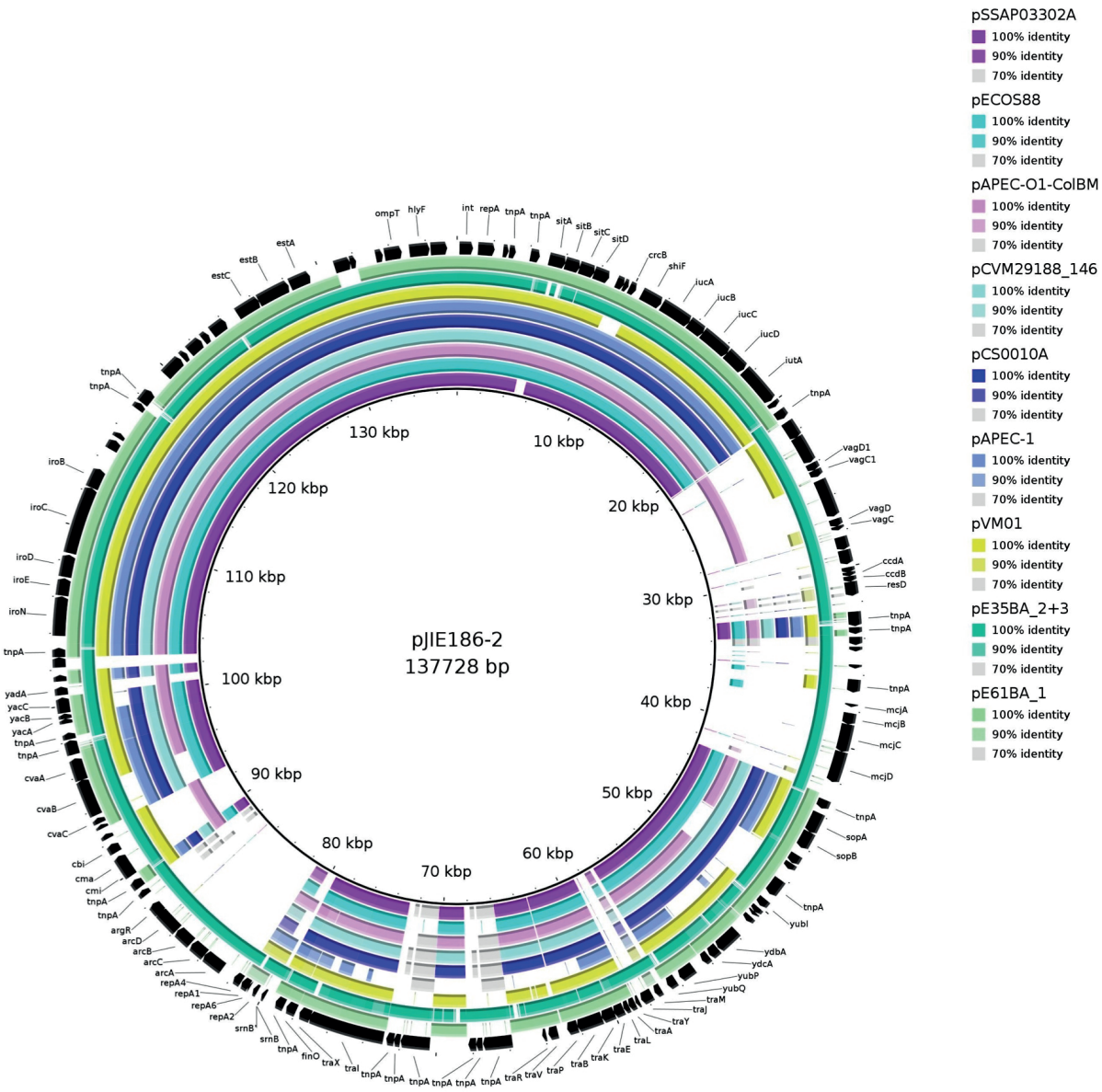

pJIE186-2
137728 bp
10 kbp
20 kbp
30 kbp
40 kbp
50 kbp
60 kbp
70 kbp
80 kbp
90 kbp
100 kbp
110 kbp
120 kbp
130 kbp
pSSAP03302A
100% identity
90% identity
70% identity
pECOS88
100% identity
90% identity
70% identity
pAPEC-O1-ColBM
100% identity
90% identity
70% identity
pCVM29188_146
100% identity
90% identity
70% identity
pCS0010A
100% identity
90% identity
70% identity
pAPEC-1
100% identity
90% identity
70% identity
pVM01
100% identity
90% identity
70% identity
pE35BA_2+3
100% identity
90% identity
70% identity
pE61BA_1
100% identity
90% identity
70% identity

Figure S11B

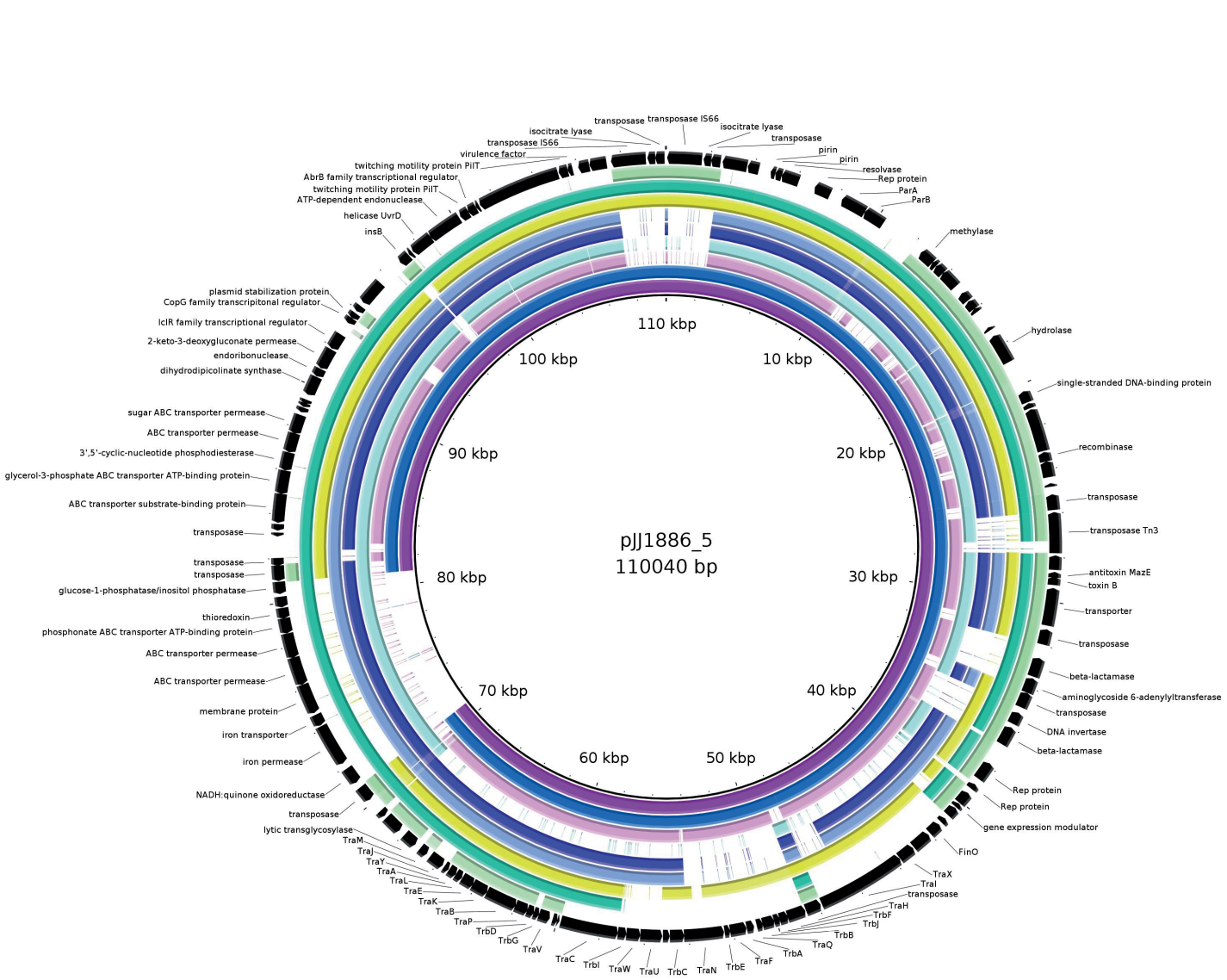

pJJ1886_5
110040 bp
10 kbp
20 kbp
30 kbp
40 kbp
50 kbp
60 kbp
70 kbp
80 kbp
90 kbp
100 kbp
110 kbp
transposase
transposase IS66
isocitrate lyase
isocitrate lyase
transposase
pirin
pirin
resolvase
Rep protein
ParA
ParB
methylase
hydrolase
single-stranded DNA-binding protein
recombinase
transposase
transposase Tn3
antitoxin MazE
toxin B
transporter
transposase
beta-lactamase
aminoglycoside 6-adenylyltransferase
transposase
DNA invertase
beta-lactamase
Rep protein
Rep protein
gene expression modulator
FinO
TraX
TraI
transposase
TraH
TrbF
TrbJ
TrbB
TraQ
TrbA
TraF
TrbE
TraN
TrbC
TraU
TraW
TrbI
TraC
TraV
TrbG
TrbD
TraP
TraB
TraK
TraE
TraL
TraA
TraY
TraJ
TraM
lytic transglycosylase
transposase
NADH:quinone oxidoreductase
iron permease
iron transporter
membrane protein
ABC transporter permease
ABC transporter permease
phosphonate ABC transporter ATP-binding protein
thioredoxin
glucose-1-phosphatase/inositol phosphatase
transposase
transposase
transposase
ABC transporter substrate-binding protein
glycerol-3-phosphate ABC transporter ATP-binding protein
3',5'-cyclic-nucleotide phosphodiesterase
ABC transporter permease
sugar ABC transporter permease
dihydrodipicolinate synthase
endoribonuclease
2-keto-3-deoxygluconate permease
IclR family transcriptional regulator
CopG family transcriptonal regulator
plasmid stabilization protein
insB
helicase UvrD
ATP-dependent endonuclease
twitching motility protein PilT
AbrB family transcriptional regulator
twitching motility protein PilT
virulence factor
transposase IS66
isocitrate lyase
pEC-L46
100% identity
90% identity
70% identity
pEC-L8
100% identity
90% identity
70% identity
pE2022_2
100% identity
90% identity
70% identity
pFV9873_5
100% identity
90% identity
70% identity
pBIDMC20B_1
100% identity
90% identity
70% identity
pBWH24_1
100% identity
90% identity
70% identity
pBIDMC38_5
100% identity
90% identity
70% identity
pEK499
100% identity
90% identity
70% identity
pEK516
100% identity
90% identity
70% identity

Figure S11C

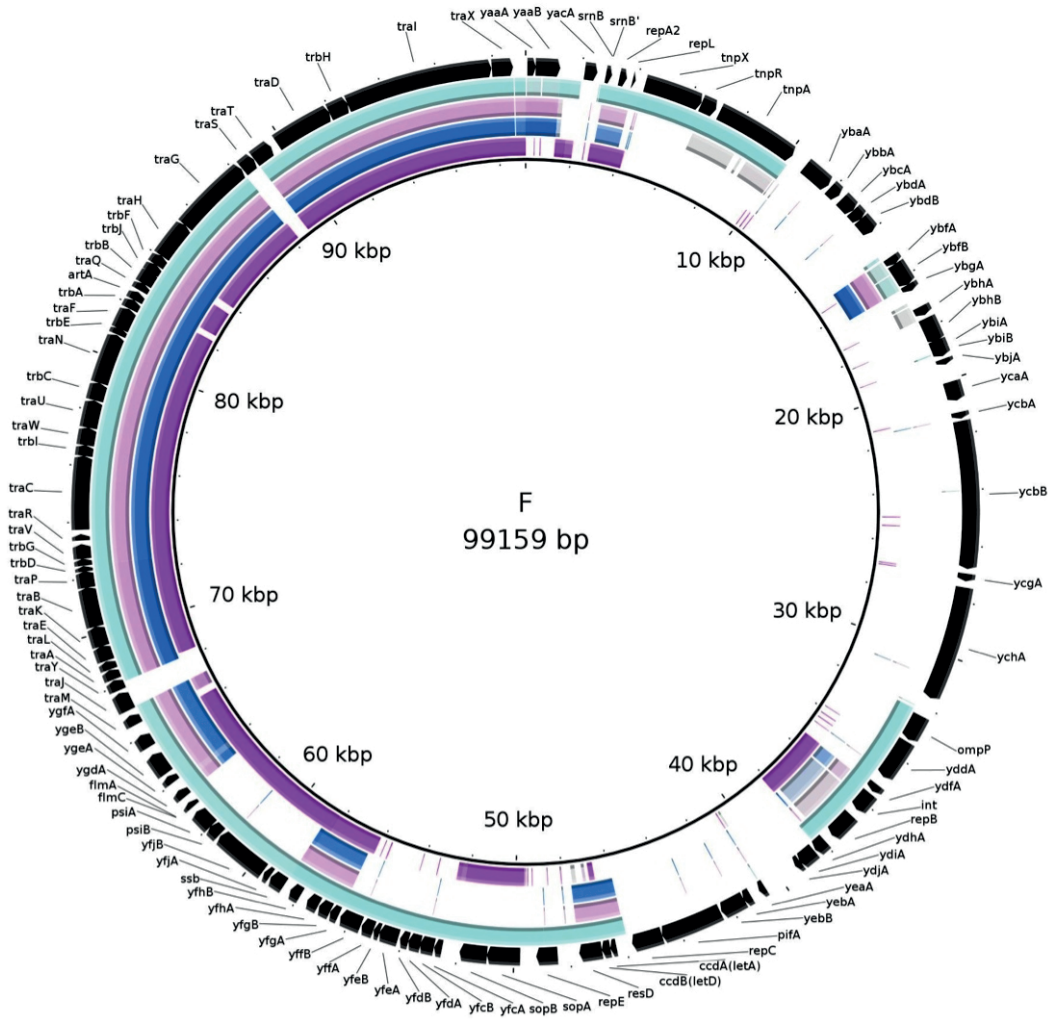

pHVH177_1
100% identity
90% identity
70% identity
pMAR2
100% identity
90% identity
70% identity
pMAR7
100% identity
90% identity
70% identity
pVir68
100% identity
90% identity
70% identity
F
99159 bp
10 kbp
20 kbp
30 kbp
40 kbp
50 kbp
60 kbp
70 kbp
80 kbp
90 kbp
traX yaaA yaaB yacA srnB srnB' repA2 repL tnpX tnpR tnpA ybaA ybbA ybcA ybdA ybdB ybfA ybfB ybgA ybhA ybhB ybiA ybiB ybjA ycaA ycbA ycbB ycgA ychA ompP yddA ydfA int repB ydhA ydiA ydjA yeaA yebA yebB pifA repC ccdA(letA) ccdB(letD) resD repE sopA sopB yfcA yfcB yfdA yfdB yfeA yfeB yffA yffB yfgA yfgB yfhA yfhB ssb yfjA yfjB psiB psiA flmC flmA ygdA ygeA ygeB ygfA traM traJ traY traA traL traE traK traB traP trbD trbG traV traR traC trbI traW traU trbC traN trbE traF trbA artA traQ trbB trbJ trbF traH traG traS traT traD trbH traI

Figure S11D

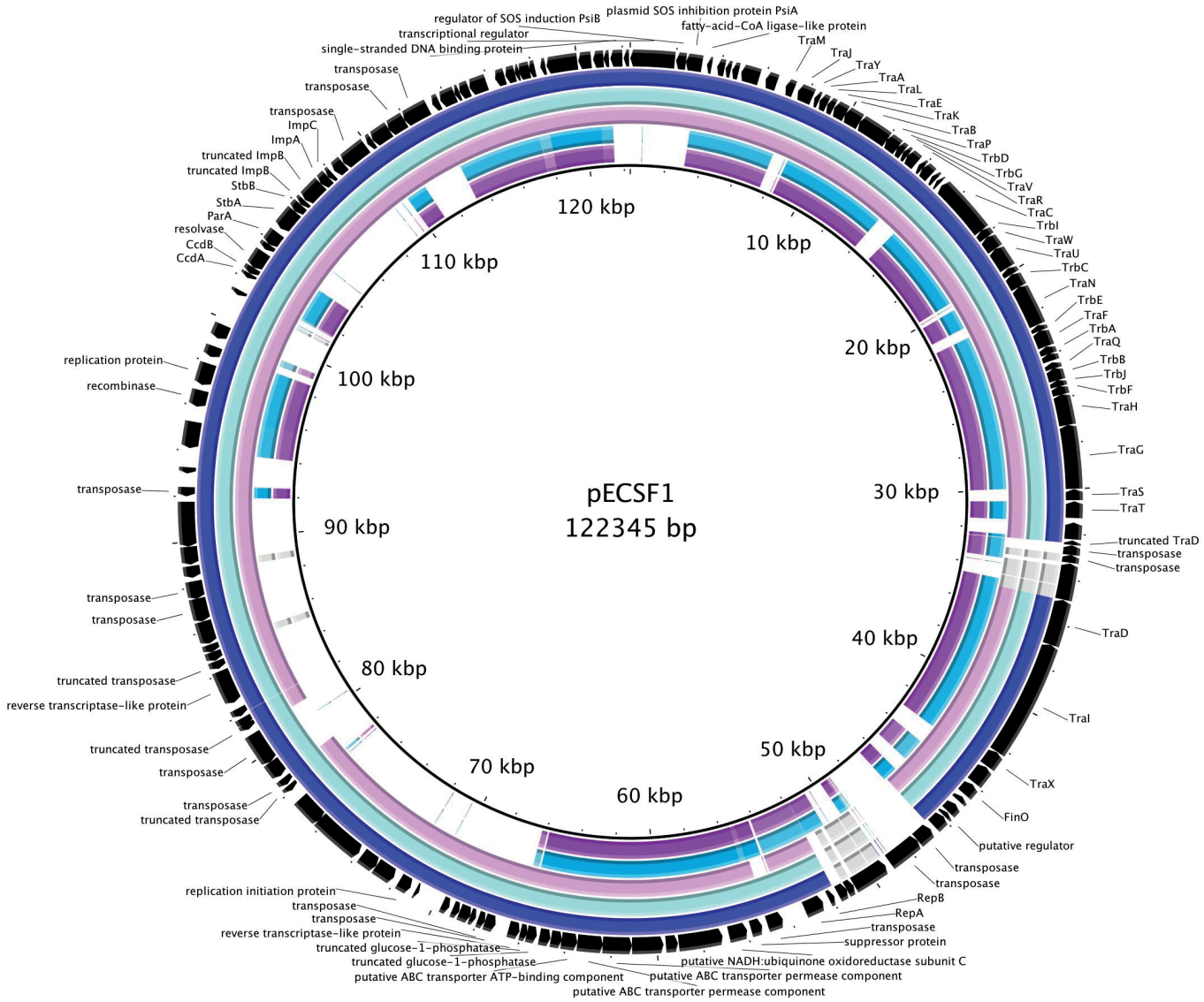

pIP1206
100% identity
90% identity
70% identity
pKF3-140
100% identity
90% identity
70% identity
p1ESCUM
100% identity
90% identity
70% identity
pUM146
100% identity
90% identity
70% identity
pEC14_114
100% identity
90% identity
70% identity
pECSF1
122345 bp
10 kbp
20 kbp
30 kbp
40 kbp
50 kbp
60 kbp
70 kbp
80 kbp
90 kbp
100 kbp
110 kbp
120 kbp
regulator of SOS induction PsiB
transcriptional regulator
single-stranded DNA binding protein
plasmid SOS inhibition protein PsiA
fatty-acid-CoA ligase-like protein
TraM
TraJ
TraY
TraA
TraL
TraE
TraK
TraB
TraP
TrbD
TrbG
TraV
TraR
TraC
TrbI
TraW
TraU
TrbC
TraN
TrbE
TraF
TrbA
TraQ
TrbB
TrbJ
TrbF
TraH
TraG
TraS
TraT
truncated TraD
transposase
transposase
TraD
TraI
TraX
FinO
putative regulator
transposase
transposase
RepB
RepA
transposase
suppressor protein
putative NADH:ubiquinone oxidoreductase subunit C
putative ABC transporter permease component
putative ABC transporter permease component
putative ABC transporter ATP-binding component
truncated glucose-1-phosphatase
truncated glucose-1-phosphatase
reverse transcriptase-like protein
transposase
transposase
replication initiation protein
transposase
truncated transposase
transposase
truncated transposase
reverse transcriptase-like protein
truncated transposase
transposase
transposase
transposase
replication protein
recombinase
transposase
resolvase
ParA
StbA
StbB
truncated ImpB
truncated ImpB
ImpA
ImpC
transposase
transposase
CcdA
CcdB

Figure S12A

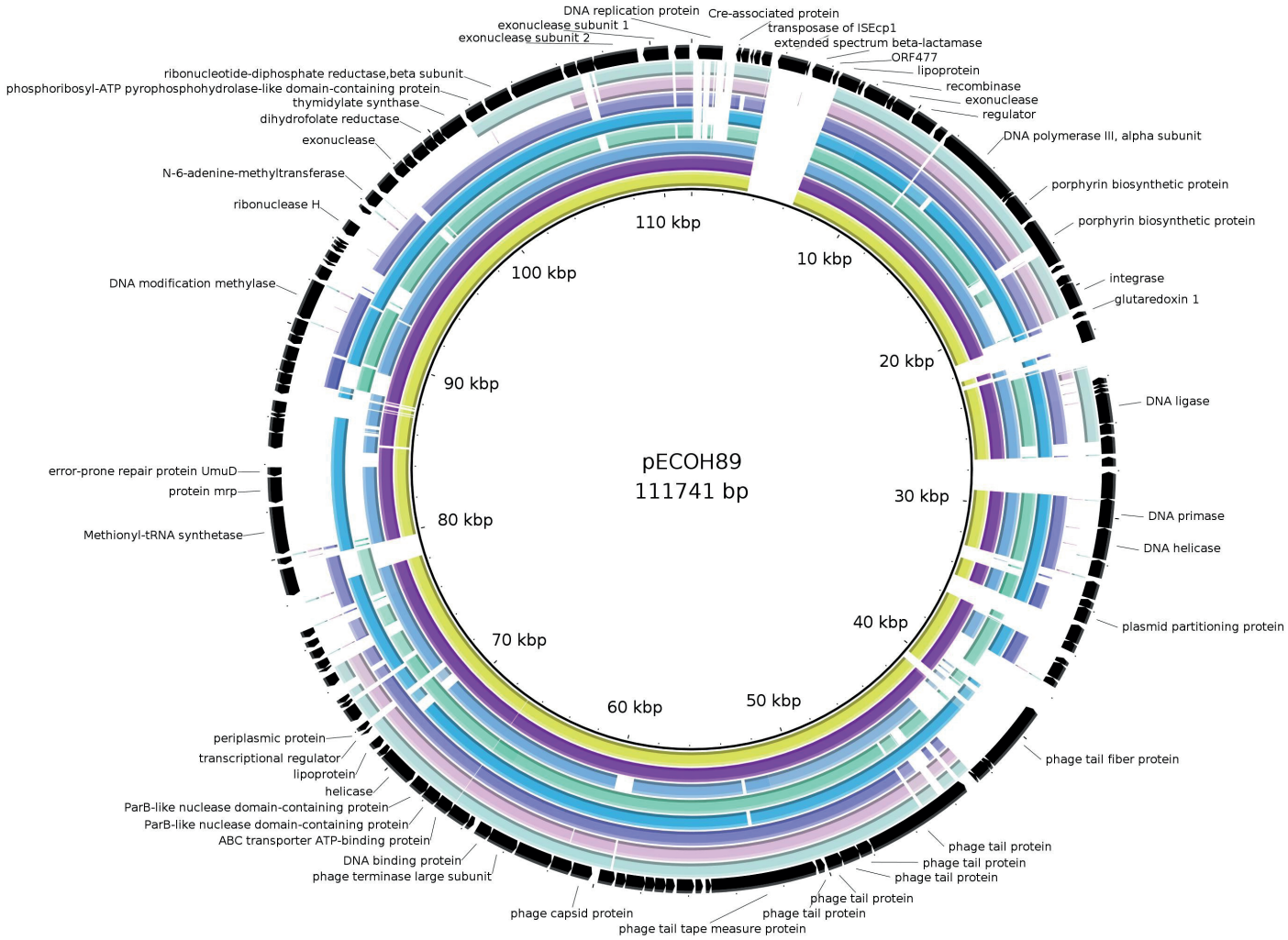
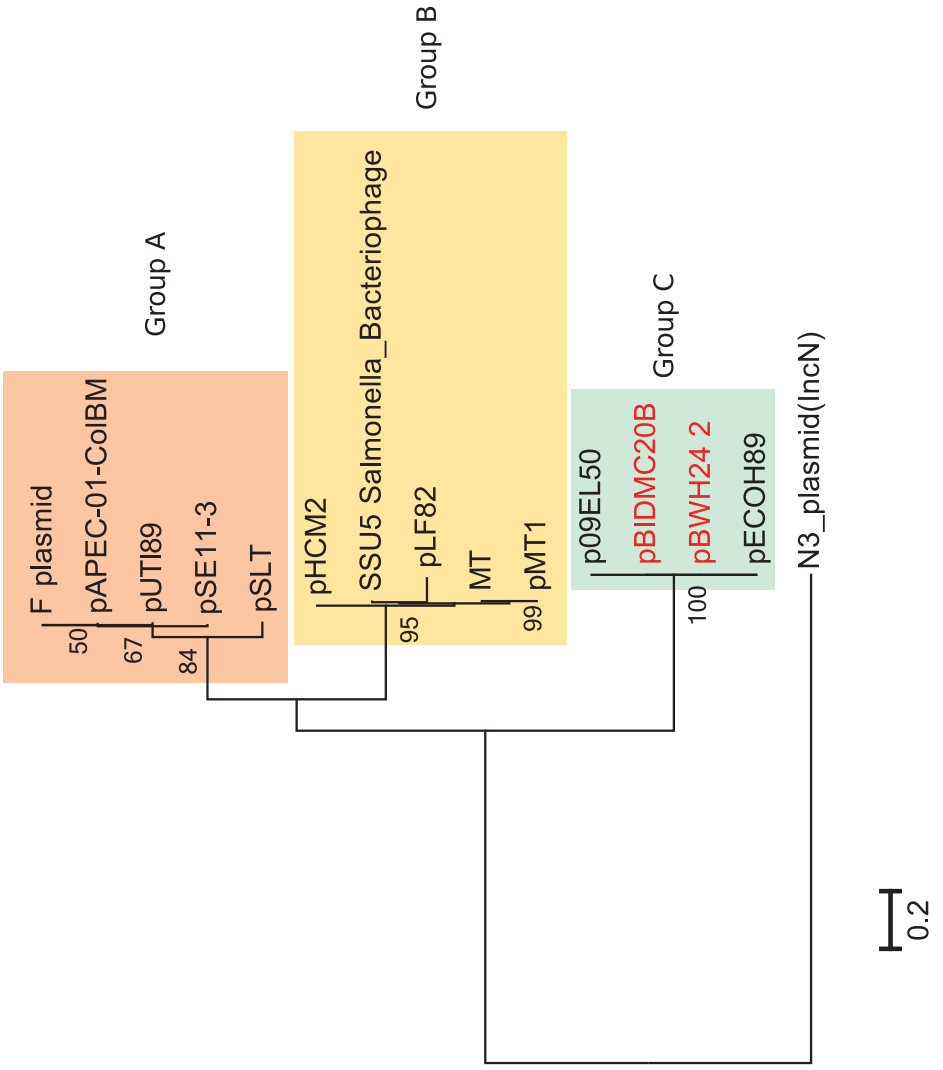

pECOH89
111741 bp
10 kbp
20 kbp
30 kbp
40 kbp
50 kbp
60 kbp
70 kbp
80 kbp
90 kbp
100 kbp
110 kbp
DNA replication protein
exonuclease subunit 1
exonuclease subunit 2
Cre-associated protein
transposase of ISEcp1
extended spectrum beta-lactamase
ORF477
lipoprotein
recombinase
exonuclease
regulator
DNA polymerase III, alpha subunit
porphyrin biosynthetic protein
porphyrin biosynthetic protein
integrase
glutaredoxin 1
DNA ligase
DNA primase
DNA helicase
plasmid partitioning protein
phage tail fiber protein
phage tail protein
phage tail protein
phage tail protein
phage tail protein
phage tail protein
phage tail tape measure protein
phage capsid protein
DNA binding protein
phage terminase large subunit
ABC transporter ATP-binding protein
ParB-like nuclease domain-containing protein
ParB-like nuclease domain-containing protein
helicase
lipoprotein
transcriptional regulator
periplasmic protein
Methionyl-tRNA synthetase
protein mrp
error-prone repair protein UmuD
DNA modification methylase
ribonuclease H
N-6-adenine-methyltransferase
exonuclease
dihydrofolate reductase
thymidylate synthase
phosphoribosyl-ATP pyrophosphohydrolase-like domain-containing protein
ribonucleotide-diphosphate reductase,beta subunit
pBIDMC20B_2
100% identity
70% identity
50% identity
pBWH24_2
100% identity
70% identity
50% identity
p09EL50
100% identity
70% identity
50% identity
SSU5
100% identity
70% identity
50% identity
pLF82
100% identity
70% identity
50% identity
pHCM2
100% identity
70% identity
50% identity
pMT1
100% identity
70% identity
50% identity
MT
100% identity
70% identity
50% identity

Figure S12B


Group A
Group B
Group C
F plasmid
pAPEC-01-ColBM
pUTI89
pSE11-3
pSLT
pHCM2
SSU5 Salmonella_Bacteriophage
pLF82
MT
pMT1
p09EL50
pBIDMC20B
pBWH24 2
pECOH89
N3_plasmid(IncN)
50
67
84
95
99
100
0.2

OPEN ACCESS Freely available online

PLOS | GENETICS

Dissemination of Cephalosporin Resistance Genes between *Escherichia coli* Strains from Farm Animals and Humans by Specific Plasmid Lineages

Mark de Been[1☯*], Val F. Lanza[2☯], María de Toro[2], Jelle Scharringa[1], Wietske Dohmen[3], Yu Du[4], Juan Hu[4], Ying Lei[4], Ning Li[5], Ave Tooming-Klunderud[6], Dick J. J. Heederik[3], Ad C. Fluit[1], Marc J. M. Bonten[1], Rob J. L. Willems[1], Fernando de la Cruz[2*], Willem van Schaik[1]

1 Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, The Netherlands, **2** Instituto de Biomedicina y Biotecnología de Cantabria, Universidad de Cantabria-Sodercan-CSIC, Santander, Spain, **3** Institute for Risk Assessment Sciences, Division of Environmental Epidemiology, Utrecht University, Utrecht, The Netherlands, **4** BGI-Shenzhen, Shenzhen, China, **5** BGI-Europe, Copenhagen, Denmark, **6** Norwegian High-Throughput Sequencing Centre, Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Oslo, Norway

Abstract

Third-generation cephalosporins are a class of β-lactam antibiotics that are often used for the treatment of human infections caused by Gram-negative bacteria, especially *Escherichia coli*. Worryingly, the incidence of human infections caused by third-generation cephalosporin-resistant *E. coli* is increasing worldwide. Recent studies have suggested that these *E. coli* strains, and their antibiotic resistance genes, can spread from food-producing animals, via the food-chain, to humans. However, these studies used traditional typing methods, which may not have provided sufficient resolution to reliably assess the relatedness of these strains. We therefore used whole-genome sequencing (WGS) to study the relatedness of cephalosporin-resistant *E. coli* from humans, chicken meat, poultry and pigs. One strain collection included pairs of human and poultry-associated strains that had previously been considered to be identical based on Multi-Locus Sequence Typing, plasmid typing and antibiotic resistance gene sequencing. The second collection included isolates from farmers and their pigs. WGS analysis revealed considerable heterogeneity between human and poultry-associated isolates. The most closely related pairs of strains from both sources carried 1263 Single-Nucleotide Polymorphisms (SNPs) per Mbp core genome. In contrast, epidemiologically linked strains from humans and pigs differed by only 1.8 SNPs per Mbp core genome. WGS-based plasmid reconstructions revealed three distinct plasmid lineages (IncI1- and IncK-type) that carried cephalosporin resistance genes of the Extended-Spectrum Beta-Lactamase (ESBL)- and AmpC-types. The plasmid backbones within each lineage were virtually identical and were shared by genetically unrelated human and animal isolates. Plasmid reconstructions from short-read sequencing data were validated by long-read DNA sequencing for two strains. Our findings failed to demonstrate evidence for recent clonal transmission of cephalosporin-resistant *E. coli* strains from poultry to humans, as has been suggested based on traditional, low-resolution typing methods. Instead, our data suggest that cephalosporin resistance genes are mainly disseminated in animals and humans via distinct plasmids.

Citation: de Been M, Lanza VF, de Toro M, Scharringa J, Dohmen W, et al. (2014) Dissemination of Cephalosporin Resistance Genes between *Escherichia coli* Strains from Farm Animals and Humans by Specific Plasmid Lineages. PLoS Genet 10(12): e1004776. doi:10.1371/journal.pgen.1004776

Editor: Paul M. Richardson, MicroTrek Incorporated, United States of America

Received May 20, 2014; **Accepted** September 24, 2014; **Published** December 18, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files. All sequence data have been deposited at DDBJ/EMBL/GenBank. Accession numbers for the Illumina sequence data are listed in Table 1. Pacific Biosciences sequence data have been deposited with accession numbers PRJNA260957 for strain 53C and PRJNA260958 for strain FAP1.

Funding: This work was supported by The European Union Seventh Framework (http://ec.europa.eu/research/fp7/) Programmes "Evolution and Transfer of Antibiotic Resistance" (EvoTAR; FP7-HEALTH-2011-single-stage; grant number 282004; to MdB, VFL, MdT, RJLW, FdlC, and WvS), and "Plaswires" (FP7 ICT 2009 4; grant number 248919; to VFL, MdT, and FdlC), the Spanish Ministry of Education (http://www.mecd.gob.es/portada-mecd/) (BFU2011 26608; to VFL, MdT, and FdlC), and The Netherlands Organisation for Research and Development ZonMw (http://www.nwo.nl/) (Contract number 50-51700-98-053; to WD). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* Email: M.deBeen-2@umcutrecht.nl (MdB); fernando.cruz@unican.es (FdlC)

☯ These authors contributed equally to this work.

Introduction

Antibiotic resistance among opportunistic pathogens is rapidly rising globally, hampering treatment of infections and increasing morbidity, mortality and health care costs [1,2]. Of particular concern is the increased incidence of infections caused by *Escherichia coli* isolates producing extended-spectrum β-lactamases (ESBLs), which has rendered the use of third generation cephalosporins increasingly ineffective against this pathogen [3].

During the 1990s, the most commonly encountered ESBL genes were bla_{TEM} and bla_{SHV}, and their spread occurred mainly through cross-transmission in hospitals. However, the epidemiology of ESBL-producing *E. coli* has changed. Nowadays, the most prevalent ESBL gene type is $bla_{\mathrm{CTX\text{-}M}}$ [4] and infections with ESBL-producing *E. coli* also occur in the community [5,6]. The

Author Summary

The rapid global rise of infections caused by *Escherichia coli* that are resistant to clinically relevant antimicrobials, including third-generation cephalosporins, is cause for concern. The intestinal tract of livestock, in particular poultry, is an important reservoir for drug resistant *E. coli*, but it is unknown to what extent these bacteria can spread to humans. Food is thought to be an important source because drug-resistant *E. coli* have been detected in animals raised for meat consumption and in meat products. Previous studies that used traditional, low-resolution, genetic typing methods found that drug resistant *E. coli* present in humans and poultry were indistinguishable from each other, suggesting dissemination of these bacteria through the food-chain to humans. However, by applying high-resolution, whole-genome sequencing methods, we did not find evidence for such transmission of bacteria through the food-chain. Instead, by employing a novel approach for the reconstruction of mobile genetic elements from whole-genome sequence data, we discovered that genetically unrelated *E. coli* isolates from both humans and animal sources carried nearly identical plasmids that encode third-generation cephalosporin resistance determinants. Our data suggest that cephalosporin resistance is mainly disseminated via the transfer of mobile genetic elements between animals and humans.

intestinal tracts of mammals and birds are important reservoirs for ESBL-producing *E. coli* [7], but it is unclear to what extent these bacteria can spread to humans. Food may be an important source, since ESBL genes have been detected in food-producing animals, especially poultry [8,9], and on retail meat [10]. The presence of ESBL-producing bacteria in food has been attributed to widespread use of antimicrobials, including third generation cephalosporins, in industrial farming practices [11].

In The Netherlands, antibiotic use and prevalence of antibiotic resistance in humans are among the lowest in Europe [12], whereas antibiotic use in food-producing animals ranks among the highest in Europe [13]. These circumstances render The Netherlands particularly suitable to study the transfer of third-generation cephalosporin-resistant bacteria through the food-chain. Recent studies performed in The Netherlands suggested clonal transfer of ESBL-producing *E. coli* from poultry to humans [14–16]. However, these interpretations were based on typing methods that target a limited number of genes, and which may not have provided sufficient resolution to accurately monitor the epidemiology of pathogens [17]. In this study, we have therefore sequenced 28 ESBL-producing and four ESBL-negative *E. coli* strains that had previously been collected from humans, poultry, retail chicken meat and pigs and tested whether previous claims on the relationship between strains from different reservoirs could be confirmed at the whole-genome sequence level. Furthermore, we investigated the relatedness of cephalosporin resistance gene-carrying plasmids, which were derived from different backgrounds and reservoirs, at the genomic level.

Results

Sequencing of ESBL-producing *E. coli*

We assessed the relatedness of ESBL-producing *E. coli* from humans, animals and food by using Whole-Genome Sequencing (WGS). The genomes of 32, mostly ESBL-producing, *E. coli* strains isolated in The Netherlands in the period 2006–2011 were sequenced (Table 1). One set of isolates (n = 24) included five pairs of human and poultry-associated strains that had previously been found indistinguishable based on Multi Locus Sequence Typing (MLST), plasmid typing (pMLST) and ESBL gene sequencing [15,18]. This set also included 11 human and poultry-associated isolates that carried an AmpC-type β-lactamase gene on an IncK plasmid [18]. The second set of isolates contained eight ESBL-producing strains that were isolated from pigs (n = 4) and their farmers (n = 4) (Table 1).

Illumina sequencing yielded draft genomes with an average assembly size of 5.2 Mbp (±0.17 Mbp), consisting of an average number of 133 scaffolds (±41) of size ≥500 bp and a mean N50 of 153 kbp (±47.9 kbp) (S1 Table). WGS-based MLST and ESBL gene analysis provided good agreement with previous typing data. Previously obtained MLST profiles and WGS-based MLST profiles were in complete agreement with each other. Although ESBL genes had previously been detected by both microarray-based methods and Sanger sequencing [15], the previously typed ESBL genes of four (out of 28) strains were absent from their assembled genomes. In three of these cases (strains 681, 320 and 38.34), we detected a $bla_{\text{TEM-1}}$ or $bla_{\text{TEM-20}}$ gene in the assembled genome, whereas a $bla_{\text{TEM-52}}$ gene should have been found according to the typing data. Mapping the Illumina reads of these strains against their own assemblies showed that the assembled bla_{TEM} genes contained several ambiguous positions pointing to the presence of more than one type of bla_{TEM} gene (most likely a combination of $bla_{\text{TEM-1}}$ and $bla_{\text{TEM-52}}$) in these strains (S2 Table). In comparison, no ambiguous positions were found in the assembled bla_{TEM} genes of other strains using the same mapping approach. In addition, the relative coverage of the assembled bla_{TEM} genes of strains 681, 320 and 38.34 was higher than that of the assembled bla_{TEM} genes of other strains (S2 Table). These findings suggested that strains 681, 320 and 38.34 contain multiple nearly identical bla_{TEM} genes (i.e. $bla_{\text{TEM-1}}$ and $bla_{\text{TEM-52}}$) that hampered the correct assembly of these genes. The fourth inconsistency between WGS and typing data was the absence of $bla_{\text{CTX-M-1}}$ from the assembly of strain 435. Mapping the reads of strain 435 against the $bla_{\text{CTX-M-1}}$ gene sequence did suggest the presence of this gene in the WGS data, but with a depth of around $1/10^{\text{th}}$ the average genomic sequencing depth. Possible explanations include a relatively poor isolation efficiency of the $bla_{\text{CTX-M-1}}$-carrying plasmid and/or the loss of this plasmid from the bacterial cells during culturing in the absence of antibiotics. The previous AmpC typing data [18] and our WGS data were in complete agreement.

Phylogeny and epidemiology of ESBL-producing *E. coli*

To assess the phylogenetic context of the sequenced strains within the genus *Escherichia* and *Shigella*, we used publicly available genome sequences of *Escherichia* (n = 126) and *Shigella* (n = 12) strains. Based on COG assignments, we identified 215 core proteins in the 170 analysed genomes, from which a concatenated core genome alignment of 170461 bp was built. A phylogenetic tree based on the 18169 variable positions in this alignment confirmed previous clustering based around phylogroups A, B1, B2, D, E and F (Fig. 1) [19]. The sequenced strains clustered together in accordance with their ST. Strains did not cluster based on isolation source, year, plasmid or ESBL gene. The ESBL-producing strains were spread throughout the tree, indicating that acquisition of ESBLs arises in different *E. coli* genetic backgrounds and has occurred multiple times during evolution (Fig. 1).

There were four clusters of ESBL-producing strains isolated from humans and animals/meat (clusters I–IV, Fig. 1). Cluster I

Table 1. *E. coli* strains sequenced in this study.

Strain	Source	Place of isolation	Date of isolation	MLST	ESBL	Inc-type of ESBL-carrying plasmid	AmpC	Inc-type of AmpC-carrying plasmid	GenBank BioProject
148	Human (blood)	Utrecht	14/02/2009	10	CTX-M-1	I1 (CC7, ST7)	n.d.	n.d.	PRJNA224190
320	Human (urine)	Utrecht	03/03/2009	10	TEM-52*	I1 (CC5, ST36)	n.d.	n.d.	PRJNA224195
681	Human (urine)	Delft	17/02/2009	10	TEM-52*	I1 (CC5, ST36)	n.d.	n.d.	PRJNA224196
38.27	Chicken (caecum)	Putten	14/06/2006	10	CTX-M-1	I1 (CC7, ST7)	n.d.	n.d.	PRJNA224199
38.34	Chicken (caecum)	Nunspeet[1]	20/06/2006	10	TEM-52*	I1 (CC5, ST10)	n.d.	n.d.	PRJNA224200
53A	Chicken meat	Utrecht[2]	10/05/2010	10	CTX-M-1	n.d.	n.d.	n.d.	PRJNA224201
85B	Chicken meat	Utrecht[3]	07/06/2010	10	TEM-52	n.d.	n.d.	n.d.	PRJNA224202
1240	Human (urine)	Schiedam	16/03/2009	58	CTX-M-1	n.d.	n.d.	n.d.	PRJNA224151
1350	Human (urine)	Leeuwarden	13/02/2009	58	CTX-M-1	I1 (CC7, ST7)	n.d.	n.d.	PRJNA224152
1365	Human (urine)	Leeuwarden	03/02/2009	58	CTX-M-1	I1 (CC7, ST7)	n.d.	n.d.	PRJNA224154
38.16	Chicken (caecum)	Nunspeet[1]	31/05/2006	58	CTX-M-1	I1 (CC7, ST7)	n.d.	n.d.	PRJNA224188
897	Human (pulmonary)	Terneuzen	22/02/2009	117	CTX-M-1	I1 (CC7, ST7)	n.d.	n.d.	PRJNA224139
1047	Human (faeces)	Velp	02/02/2009	117	CTX-M-1	I1 (CC7, ST7)	CMY-2	K	PRJNA224146
38.52	Chicken (caecum)	Nunspeet[1]	13/07/2006	117	CTX-M-1	I1 (CC7, ST7)	n.d.	n.d.	PRJNA224147
53C	Chicken meat	Utrecht[2]	10/05/2010	117	CTX-M-1	n.d.	CMY-2	K	PRJNA224234
435	Human (faeces)	Deventer	19/03/2009	68	CTX-M-1*[†]	n.d.	CMY-2[‡]	K	PRJNA224205
328	Human (urine)	Utrecht	04/03/2009	69	*negative*	n.d.	CMY-2[‡]	K	PRJNA224204
597	Human (urine)	Groningen	13/03/2009	95	*negative*	n.d.	CMY-2	K	PRJNA224228
668	Human (urine)	Delft	06/02/2009	648	CTX-M-15	n.d.	CMY-2[‡]	K	PRJNA224230
606	Human (pulmonary)	Groningen	18/02/2009	unknown	*negative*	n.d.	CMY-2[‡]	K	PRJNA224229
1A	Chicken meat	Utrecht[4]	2010	23	SHV-12	n.d.	CMY-2	K	PRJNA224231
27A	Chicken meat	Utrecht	26/04/2010	23	TEM-52	n.d.	CMY-2	K	PRJNA224233
9B	Chicken meat	Utrecht[4]	12/04/2010	93	SHV-12	n.d.	CMY-2	K	PRJNA224232
87A	Chicken meat	Utrecht[3]	07/06/2010	115	*negative*	n.d.	CMY-2	K	PRJNA224235
FAH1	Human (faeces)	farm A	18/04/2011	n.d.	CTX-M-1	n.d.	n.d.	n.d.	PRJNA224240
FAH2	Human (faeces)	farm A	19/04/2011	n.d.	CTX-M-1	n.d.	n.d.	n.d.	PRJNA224238
FAP1	Pig (faeces)	farm A	04/04/2011	n.d.	CTX-M-1	n.d.	n.d.	n.d.	PRJNA224241
FAP2	Pig (faeces)	farm A	04/04/2011	n.d.	CTX-M-1	n.d.	n.d.	n.d.	PRJNA224242
FBH1	Human (faeces)	farm B	24/05/2011	n.d.	CTX-M-1	n.d.	n.d.	n.d.	PRJNA224243
FBP1	Pig (faeces)	farm B	2011	n.d.	CTX-M-1	n.d.	n.d.	n.d.	PRJNA224244

Table 1. Cont.

Strain	Source	Place of isolation	Date of isolation	MLST	ESBL	Inc-type of ESBL-carrying plasmid	AmpC	Inc-type of AmpC-carrying plasmid	GenBank BioProject
FCH1	Human (faeces)	farm C	28/06/2011	n.d.	CTX-M-1	n.d.	n.d.	n.d.	PRJNA224245
FCP1	Pig (faeces)	farm C	28/06/2011	n.d.	CTX-M-1	n.d.	n.d.	n.d.	PRJNA224246

For all strains that were ESBL- and IncI1 plasmid-positive or AmpC- and IncK plasmid-positive, the Inc group of the plasmid that carried the ESBL/AmpC gene had previously been determined using a transformation-based approach [15,18]. Identical numbers behind places of isolation indicate the exact same locations: i.e. the same retail store for chicken meat isolates, the same slaughterhouse for chicken isolates and the same farm (referred to as farms A, B, and C) for pig and pig farmer isolates.

* These ESBL genes were not found in the genome assemblies: bla_{TEM-20} or bla_{TEM-1} were found instead of bla_{TEM-52}.

†This ESBL gene had previously only been typed using microarrays (no sequencing). The gene was found to belong to the CTX-M-1 group.

‡These CMY genes were divided over two contigs that were connected on a scaffold. BLAST runs of the partial CMY-2 sequences against GenBank's nr database gave best hits with bla_{CMY-2} and mapping of raw Illumina reads against bla_{CMY-2} indicated that the full bla_{CMY-2} gene was present in the corresponding strain.

n.d.: not determined.

doi:10.1371/journal.pgen.1004776.t001

contained human and pig isolates from two pig farms, with strains from farm A being particularly closely related. The other three clusters contained the five pairs of human and chicken isolates that had previously been considered indistinguishable based on traditional typing methods [15].

Among the five pairs of human and chicken isolates, the most closely related pairs were in cluster IV. The COG-based core genome alignment showed 171 SNPs between these strains, corresponding to 1003 SNPs/Mbp. To better elucidate the minimum number of SNPs between human and chicken isolates, we performed a core genome analysis using OrthoMCL [20] on the strains in cluster IV. For comparison, ten clonal O104:H4 strains from the 2011 German EHEC outbreak [21] and the four strains from pig farm A (cluster I) were included in this analysis (Fig. 1). We identified 3574 core proteins in this dataset translating to a concatenated nucleotide alignment of 3.34 Mbp. Within cluster IV there were 4216 SNPs between the most closely related isolates, corresponding to 1263 SNPs/Mbp. In contrast, only 0–6 SNPs (0–1.8 SNPs/Mbp) were found between any two strains in the German EHEC outbreak and only 6 SNPs were found between farmer isolate FAH2 and any of its two related pig isolates, suggesting recent clonal transmission of *E. coli* between pig and human in farm A (Fig. 2).

Given an estimated *E. coli* mutation rate of 2.3×10^{-7} to 3.0×10^{-6} substitutions per site per year [21,22] and an average *E. coli* genome size of 5.2 Mbp, the number of SNPs (1263/Mbp) between the two most closely related human and chicken isolates largely exceeded the number of 3–41 SNPs that is expected to arise in 2.6 years (the difference in isolation dates between both strains, Table 1). Even if 10% of the detected SNPs were due to recombination, which is considerably more than the reported upper limit for recombinant DNA (~3.5%) in *E. coli* [19], the number of SNPs due to mutation would exceed the expected maximum number of SNPs in case of recent clonal transmission. As the genetic distance between all other pairs of human and poultry isolates was even larger, our findings do not support a scenario of recent clonal transmission of ESBL-producing *E. coli* strains between humans and poultry.

Reconstruction of plasmids from WGS data

To investigate the possibility of horizontal spread of ESBLs via plasmids, we employed a Plasmid Constellation Networks (PLACNET) approach to reconstruct plasmids from WGS data [23]. Application of this approach resulted in the reconstruction of 147 plasmids (average of 4.6±2.1 plasmids per strain), with plasmid sizes ranging from 1.1 kbp to 290.4 kbp (Table 2). The plasmid sizes showed a trimodal distribution (Fig. 3) that was similar to the distribution previously reported for plasmids from a wide range of bacterial taxa [24]. The median size of large (conjugative) plasmids was 93.6 kbp (n = 91). Small plasmids could be further subdivided into two groups: one with a median size of 5.9 kbp (n = 41), predominated by mobilizable plasmids (i.e. containing MOB genes) and one with a median size of 1.7 kbp (n = 15), predominated by non-mobilizable plasmids. Based on the classification of their MOB genes [25] and using a hierarchical clustering analysis of gene content (Fig. 4), reconstructed plasmids belonged to a limited number of plasmid families, of which the most abundant ones were IncF-MOB_{F12} (n = 38; average size of 107.4±57.7 kbp) and IncI1-MOB_{P12} (n = 26; average size of 95.7±20.0 kbp). Other abundant families included MOBP5 (n = 25), IncK (n = 12) and MOBQ (n = 11). Finally, there were 18, mostly small-sized, plasmids (median size of 1.6 kbp; range of 1.1–106.3 kbp) that were scattered throughout the dendrogram and could not be clearly subdivided into any family. A comparison

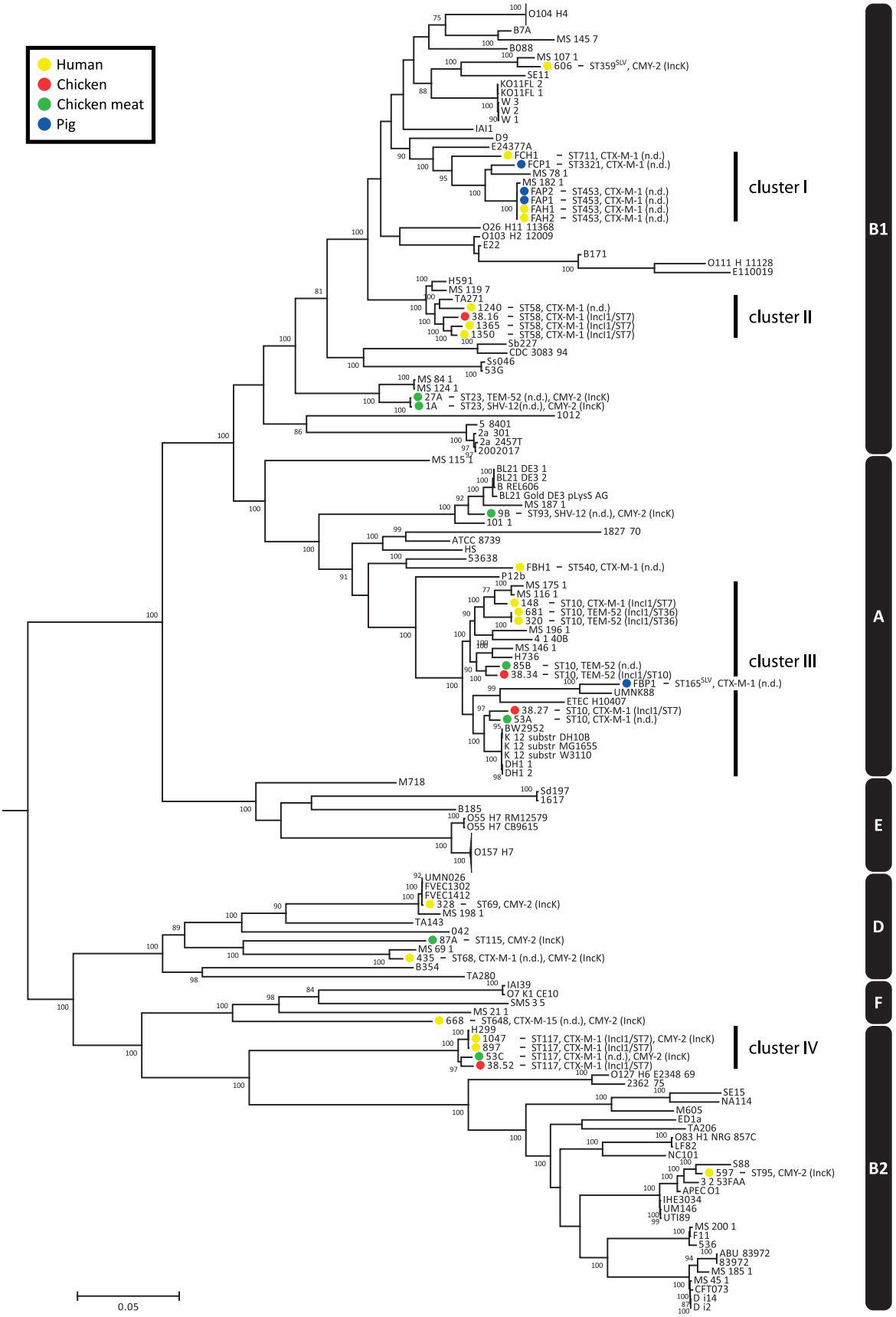

Human
Chicken
Chicken meat
Pig
O104 H4
B7A
MS 145 7
B088
MS 107 1
606 – ST359SLV, CMY-2 (IncK)
SE11
KO11FL 2
KO11FL 1
W 3
W 2
W 1
IAI1
D9
E24377A
FCH1 – ST711, CTX-M-1 (n.d.)
FCP1 – ST3321, CTX-M-1 (n.d.)
MS 78 1
MS 182 1
FAP2 – ST453, CTX-M-1 (n.d.)
FAP1 – ST453, CTX-M-1 (n.d.)
FAH1 – ST453, CTX-M-1 (n.d.)
FAH2 – ST453, CTX-M-1 (n.d.)
cluster I
O26 H11 11368
O103 H2 12009
E22
B171
O111 H 11128
E110019
H591
MS 119 7
TA271
1240 – ST58, CTX-M-1 (n.d.)
38.16 – ST58, CTX-M-1 (IncI1/ST7)
1365 – ST58, CTX-M-1 (IncI1/ST7)
1350 – ST58, CTX-M-1 (IncI1/ST7)
cluster II
Sb227
CDC 3083 94
Ss046
53G
MS 84 1
MS 124 1
27A – ST23, TEM-52 (n.d.), CMY-2 (IncK)
1A – ST23, SHV-12(n.d.), CMY-2 (IncK)
1012
5 8401
2a 301
2a 2457T
2002017
MS 115 1
BL21 DE3 1
BL21 DE3 2
B REL606
BL21 Gold DE3 pLysS AG
MS 187 1
9B – ST93, SHV-12 (n.d.), CMY-2 (IncK)
101 1
1827 70
ATCC 8739
HS
53638
FBH1 – ST540, CTX-M-1 (n.d.)
P12b
MS 175 1
MS 116 1
148 – ST10, CTX-M-1 (IncI1/ST7)
681 – ST10, TEM-52 (IncI1/ST36)
320 – ST10, TEM-52 (IncI1/ST36)
MS 196 1
4 1 40B
MS 146 1
H736
85B – ST10, TEM-52 (n.d.)
38.34 – ST10, TEM-52 (IncI1/ST10)
FBP1 – ST165SLV, CTX-M-1 (n.d.)
UMNK88
ETEC H10407
38.27 – ST10, CTX-M-1 (IncI1/ST7)
53A – ST10, CTX-M-1 (n.d.)
BW2952
K 12 substr DH10B
K 12 substr MG1655
K 12 substr W3110
DH1 1
DH1 2
cluster III
M718
Sd197
1617
B185
O55 H7 RM12579
O55 H7 CB9615
O157 H7
UMN026
FVEC1302
FVEC1412
328 – ST69, CMY-2 (IncK)
MS 198 1
TA143
042
87A – ST115, CMY-2 (IncK)
MS 69 1
435 – ST68, CTX-M-1 (n.d.), CMY-2 (IncK)
B354
TA280
IAI39
O7 K1 CE10
SMS 3 5
MS 21 1
668 – ST648, CTX-M-15 (n.d.), CMY-2 (IncK)
H299
1047 – ST117, CTX-M-1 (IncI1/ST7), CMY-2 (IncK)
897 – ST117, CTX-M-1 (IncI1/ST7)
53C – ST117, CTX-M-1 (n.d.), CMY-2 (IncK)
38.52 – ST117, CTX-M-1 (IncI1/ST7)
cluster IV
O127 H6 E2348 69
2362 75
SE15
NA114
M605
ED1a
TA206
O83 H1 NRG 857C
LF82
NC101
S88
597 – ST95, CMY-2 (IncK)
3 2 53FAA
APEC O1
IHE3034
UM146
UTI89
MS 200 1
F11
536
ABU 83972
83972
MS 185 1
MS 45 1
CFT073
D i14
D i2
B1
A
E
D
F
B2
0.05

Fig. 1. Phylogeny of *Escherichia* and *Shigella* species, including ESBL- and AmpC-positive strains sequenced for the purpose of this study. The tree was built using 18169 variable positions present in 215 core genes. The strains sequenced in this study are indicated in coloured bullets according to isolation source. Typing characteristics (Table 1) are given behind strain names. In case the MLST had not been determined before, it was determined using MLST v1.6 [55]. Clusters I–IV (see main text) are indicated behind the tree. Phylogroups are also indicated behind the tree (white text on black bars). The O104:H4 and O157:H7 branches are collapsed and represent 11 and 20 strains, respectively. Bootstrap support was implemented by running 100 bootstrap replicates. Values <75% are not displayed. SLV indicates Single Locus Variants of corresponding MLSTs.
doi:10.1371/journal.pgen.1004776.g001

between previous typing data and the PLACNET reconstructions showed that both data types were in excellent agreement with each other. First of all, the 11 strains that were previously found to contain an IncI1 plasmid were also found to contain such a plasmid using PLACNET. The sizes of these 11 reconstructed plasmids (average size of 92.7 kbp±5.7 kbp) were also in agreement with their previously estimated sizes on the basis of gel electrophoresis (average size of 97.7 kbp±3.8 kbp) [15]. Furthermore, the reconstructed plasmids for ten of these 11 strains had exactly the same ST as was previously found using pMLST. The only inconsistency was found for strain 38.34, which should contain an IncI1/ST10 plasmid according to pMLST, whereas we reconstructed an IncI1/ST36 plasmid. However, IncI1/ST10 and IncI1/ST36 are single locus variants that differ by only one SNP (http://pubmlst.org/plasmid/), indicating that this inconsistency was not a result of PLACNET, but was likely due to typing errors. Of the 11 strains that had previously been found to contain an IncK plasmid, ten were also found to contain

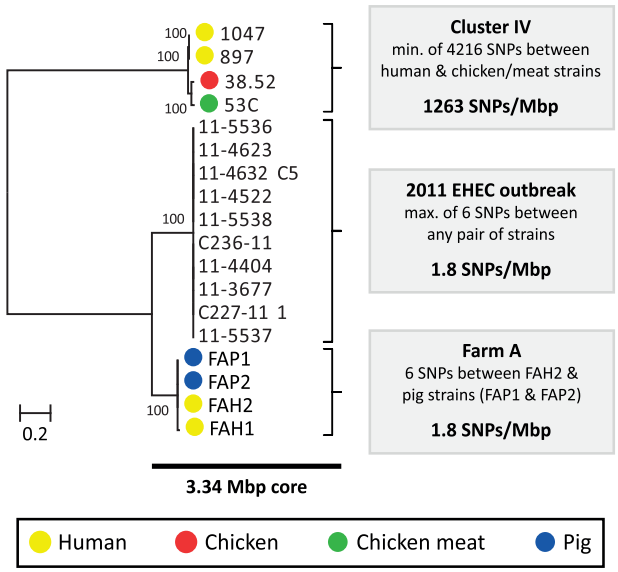


Fig. 2. Phylogeny and SNP analysis of closely related ESBL-producing *E. coli* strains from human and poultry. A high resolution core genome analysis was performed for a subset of strains which, based on an initial phylogenetic analysis (Cluster IV, Fig. 1), included the most closely related pairs of human and poultry-associated ESBL-producing strains within our dataset. Strains within Cluster IV had previously been found to be identical with respect to MLST, ESBL gene and ESBL-carrying plasmid (Table 1). For comparative purposes, clonally related *E. coli* strains from the 2011 German EHEC outbreak [21] and four potentially clonally related strains isolated from a single pig farm (Farm A) were included in this analysis. A phylogenetic tree, built from the 107919 variable positions present in the resulting 3.34 Mbp core genome alignment is shown to the left. Bootstrap support was implemented by running 1000 bootstrap replicates. Coloured bullets refer to the isolation source. The number of SNPs found in each of the three clusters (Cluster IV, EHEC outbreak, Farm A) is shown to the right.
doi:10.1371/journal.pgen.1004776.g002

such a plasmid using PLACNET, the only exception being strain 1047.

We also examined to what extent we were able to correctly connect ESBL and AmpC genes to reconstructed plasmids. Of the 28 previously typed ESBL genes, 24 were correctly identified in their genomes (Table 1) and among these, 15 were connected to a reconstructed plasmid. Four of the remaining nine unconnected ESBL genes ($bla_{\text{CTX-M-1}}$ in strains 1350, 1365, 1047 and 38.52) should have been connected to an IncI1 plasmid according to previous typing data (Tables 1–2). The reason that these ESBL genes remained unassigned was because they were located on relatively small scaffolds (average size of 6.6 kbp) that did not contain enough genetic information to unequivocally match them to a single plasmid using our reference database. For the 15 cases where we were able to connect an ESBL gene to a reconstructed plasmid, typing data indicating where the ESBL gene should be located was available for four cases (strains 148, 897, 38.16 and 38.27) and for all these cases we had connected the ESBL gene ($bla_{\text{CTX-M-1}}$) to the correct plasmid (IncI1/ST7) (Tables 1–2). Of the 11 AmpC ($bla_{\text{CMY-2}}$) genes, ten were connected to their correct plasmid (IncK). The only exception was found again for strain 1047 for which we could not reconstruct an IncK plasmid (Table 2). The above findings show that PLACNET worked efficiently to assemble plasmids from WGS data, although the assignment of small scaffolds to plasmids can be problematic, as is illustrated above by the ESBL genes that were not linked to a specific reconstructed plasmid (see also discussion below and in [23]).

Identification of distinct ESBL-associated plasmid lineages

Fifteen ESBL genes were connected to a reconstructed plasmid, of which 13 were connected to an IncI1 plasmid. Frequently (eight out of 13), these IncI1 plasmids were also unequivocally linked to other antibiotic resistance genes, such as *sul*, *dfrA*, *aadA* or *tet*. We also found IncK plasmids that were commonly (ten out of 12 plasmids) associated with the AmpC β-lactamase-encoding gene $bla_{\text{CMY-2}}$ (Fig. 4).

As IncI1 and IncK were the only plasmid families that included reconstructed ESBL-/AmpC-containing plasmids in strains from both humans and animals/meat, we further investigated their potential role in the transfer of resistance genes through the food-chain. To this aim we built a gene content-based dendrogram that also included closely related and publicly available plasmid sequences. In the resulting dendrogram, all reconstructed ESBL-containing IncI1 plasmids, except the $bla_{\text{SHV-12}}$-carrying plasmids p1A_2 and p9B_1, clustered into one specific branch that did not contain any other previously sequenced plasmid (Fig. 5). This branch also contained 12 of the 13 reconstructed IncI1 plasmids that did not include an ESBL gene. Similarly, all of the reconstructed IncK plasmids, except p87A_5, clustered into one specific branch that did not include any previously sequenced plasmid. These findings suggest the existence of IncI1 and IncK plasmids with a genetic profile distinct from previously characterised plasmids. We did not find any single gene that unequivocally explained the formation of the IncK branch, pointing to a

Table 2. Plasmids reconstructed using PLACNET.

Strain	IncI1	IncK	IncF	IncN	IncHI1	IncX	IncI2	MobC12	MobQu	MobQ1	MobP5	MobV2	Other	Nr. of plasmids	Total plasmid seq. (kbp) and scaffolds
148	p2: 98 (7), CTX-M-1	-	p3: 32 (2)	-	-	p1: 48 (3)	-	-	-	-	-	-	-	4	254 (18)
			p4: 76 (6)												
320	p2: 87 (5)*	-	p1: 132 (7)	-	-	-	-	p4: 26 (1)	-	-	p3: 7 (1)	-	-	4	253 (14)
681	p2: 87 (4)*	-	p1: 127 (4)	-	-	-	-	-	-	-	p3: 7 (1)	-	-	3	221 (9)
38.27	p1: 91 (8), CTX-M-1	-	p3: 96 (1)	-	p4: 185 (7)	-	-	-	-	-	-	-	p2: 3 (3)	4	375 (19)
38.34	p2: 84 (6)*	-	p1: 290 (41)	-	-	-	-	-	-	-	-	-	-	2	375 (47)
53A	p1: 142 (9), CTX-M-1	-	-	-	-	-	-	-	-	-	-	-	-	1	142 (9)
85B	p2: 82 (4), TEM-52	-	p3: 109 (14)	-	-	-	-	-	-	-	p1: 7 (1)	-	-	3	198 (19)
1240	p6: 63 (7)	p5: 77 (6)	p3: 154 (24)	-	-	-	-	-	-	p2: 5 (1)	-	-	p1: 2 (1)	7	449 (58)
			p4: 57 (18), CTX-M-1												
			p7: 91 (1)												
1350	p4: 89 (6)*	-	p5: 172 (47)	-	-	p2: 49 (1)	-	p3: 29 (5)	-	-	-	-	p1: 3 (1)	5	341 (60)
1365	p4: 93 (5)*	-	p3: 132 (7)	-	-	-	-	-	-	-	p2: 5 (1)	-	p1: 2 (1)	4	231 (14)
38.16	p1: 101 (7), CTX-M-1	-	p3: 149 (13)	-	-	p2: 39 (1)	-	-	-	-	-	-	-	3	289 (21)
897	p1: 101 (9), CTX-M-1	-	p5: 62 (3)	-	-	-	-	-	-	-	p2: 7 (1)	-	p4: 1 (2)	6	265 (23)
			p6: 90 (7)								p3: 3 (1)				
1047	p3: 93 (7)*	-*	p4: 155 (12)	-	-	-	-	-	-	-	p1: 10 (2)	-	p2: 1 (1)	4	259 (22)
38.52	p2: 93 (11)*	-	p1: 181 (40)	-	-	-	-	-	-	-	-	-	-	2	274 (51)
53C	p2: 83 (8), CTX-M-1	p3: 70 (5), CMY-2	p4: 128 (10)	-	-	-	p1: 57 (8)	-	-	-	-	-	-	4	338 (31)
435	-	p1: 96 (20), CMY-2	p2: 87 (2)	-	-	-	-	-	-	-	p3: 9 (1)	-	p4: 2 (1)	4	195 (24)
328	-	p1: 78 (7), CMY-2	p8: 119 (12)	-	-	p2: 33 (3)	-	-	p7: 4 (1)	-	p5: 6 (1)	p4: 3 (1)	p3: 2 (1)	9	346 (28)
			p9: 96 (1)								p6: 5 (1)				
597	-	p2: 86 (5), CMY-2	p1: 65 (5)	-	-	-	-	-	-	p3: 5 (1)	-	-	-	3	156 (11)

Table 2. Cont.

Strain	IncI1	IncK	IncF	IncN	IncHI1	IncX	IncI2	MobC12	MobQu	MobQ1	MobP5	MobV2	Other	Nr. of plasmids	Total plasmid seq. (kbp) and scaffolds
668	-	p1: 86 (9), CMY-2	-	p2: 58 (14), CTX-M-15	-	-	-	-	p3: 4 (1)	p6: 5 (1)	p5: 3 (1)	-	p4: 2 (1)	6	159 (27)
606	-	p3: 74 (7), CMY-2	p1: 16 (3)	p2: 55 (6)	-	-	-	-	-	-	p6: 7 (1)	-	p7: 2 (1)	7	521 (58)
			p4: 276 (36)												
			p5: 92 (4)												
1A	p2: 96 (10), SHV-12	p3: 79 (6), CMY-2	p1: 102 (7)	-	-	-	-	-	-	-	p4: 6 (1)	-	p6: 2 (1)	6	372 (26)
			p5: 87 (1)												
27A	p10: 78 (11)	p8: 88 (10), CMY-2	p9: 154 (20)	-	-	-	-	-	-	p4: 6 (1)	p3: 3 (1)	-	p1: 1 (1)	10	345 (48)
											p6: 3 (1)		p2: 1 (1)		
											p7: 7 (1)		p5: 4 (1)		
9B	p1: 105 (9), SHV	p2: 92 (13), CMY-2	p3: 46 (1)	-	-	-	-	-	p8: 6 (1)	-	p5: 5 (1)	p9: 4 (2)	p4: 2 (1)	9	274 (30)
											p6: 6 (1)				
											p7: 8 (1)				
87A	-	p5: 88 (29), CMY-2	-	-	-	p4: 47 (1)	-	-	p2: 7 (1)						
p3: 5 (1)	-	p1: 9 (1)	-	-	5	156 (33)									
FAH1	p2: 80 (20), CTX-M-1	p1: 74 (5)	p3: 123 (11)	-	-	-	-	-	p4: 7 (1)	-	-	-	-	4	284 (37)
FAH2	p2: 166 (20), CTX-M-1	-	p1: 57 (1)	-	-	-	-	-	-	-	-	-	-	2	223 (21)
FAP1	p2: 90 (6)	-	p1: 57 (1)	-	-	-	p3: 58 (7)	-	-	-	-	-	p5: 46 (2)	5	319 (26)
			p4: 68 (10)												
FAP2	p4: 89 (5)	-	p2: 64 (8)	-	-	-	p3: 60 (7)	-	-	p1: 6 (1)	-	-	-	5	286 (32)
			p5: 67 (11)												
FBH1	p2: 82 (8)	-	p1: 76 (9)	-	-	-	-	-	-	-	p4: 3 (1)	-	p5: 1 (1)	5	225 (20)
			p3: 61 (1)												
FBP1	p2: 103 (12)	-	p1: 132 (8)	-	-	-	-	-	-	-	-	p3: 2 (2)	p4: 106 (1)	5	345 (25)
												p6: 2 (2)			
FCH1	p1: 107 (10), CTX-M-1	-	-	-	-	-	-	-	-	-	p2: 7 (1)	-	-	3	119 (12)

Table 2. Cont.

Strain	IncI1	IncK	IncF	IncN	IncHI1	IncX	IncI2	MobC12	MobQu	MobQ1	MobP5	MobV2	Other	Nr. of plasmids	Total plasmid seq. (kbp) and scaffolds
											p3: 5 (1)				
FCP1	p1: 101 (10), CTX-M-1	-	-	-	-	-	-	-	-	-	p2: 3 (1)	-	-	3	109 (12)
											p3: 5 (1)				

For each strain, the type and number of reconstructed plasmids are indicated in columns 2–14. Plasmid numbering (e.g. p1) corresponds with the numbering in Figs. 4–6, and is followed by the plasmid size (in kbp), the number of assigned scaffolds (between brackets), and assigned bla_{ESBL} or bla_{CMY-2} genes, if applicable. The same format is used for the last summarising column.

* Inconsistency between plasmid typing data and WGS data/PLACNET (typing data is between brackets below): 320 (bla_{TEM-52} on IncI1); bla_{TEM-52} not found in assembly and thus not linked to an IncI1 plasmid; 681 (bla_{TEM-52} on IncI1): bla_{TEM-52} not found in assembly and thus not linked to an IncI1 plasmid; 38.34 (bla_{TEM-52} on IncI1); bla_{TEM-52} not found in assembly and thus not linked to an IncI1 plasmid; 1350 ($bla_{CTX-M-1}$ on IncI1); $bla_{CTX-M-1}$ found in assembly, but not linked to an IncI1 plasmid; 1365 ($bla_{CTX-M-1}$ on IncI1): $bla_{CTX-M-1}$ found in assembly, but not linked to an IncI1 plasmid; 1047 ($bla_{CTX-M-1}$ on IncI1): $bla_{CTX-M-1}$ found in assembly, but not linked to an IncI1 plasmid; 1047 (bla_{CMY-2} on IncK): bla_{CMY-2} found in assembly, but no IncK plasmid reconstructed; 38.52 ($bla_{CTX-M-1}$ on IncI1): $bla_{CTX-M-1}$ found in assembly, but not linked to an IncI1 plasmid.

doi:10.1371/journal.pgen.1004776.t002

delicate configuration of genes that gives these plasmids their unique genetic profile. However, for the IncI1 branch, we found a characteristic shufflon-related gene (UniProt P10487) that was present in all 26 reconstructed IncI1 plasmids, but which was absent from related IncI1 plasmids (Fig. 5).

To further characterise the IncI1 and IncK resistance plasmids, phylogenetic trees were built from the sequences of the reconstructed plasmids and their closest plasmid relatives. For the IncI1 phylogenetic reconstruction, the 23 plasmids belonging to the specific IncI1 branch as well as 27 related plasmids were included. An OrthoMCL analysis of these plasmids resulted in 8 core proteins (S3 Table), corresponding to a concatenated nucleotide alignment of 8.6 kbp, including 763 variable positions. In the phylogenetic tree built from these variable positions the reconstructed IncI1 plasmids were assigned to four distinct branches (Fig. 6A), each of which also contained previously characterised plasmids. However, the reconstructed plasmids within each branch were always more similar to each other than to any of these previously characterised plasmids. Two of the four branches, corresponding to IncI1/ST3 and IncI1/ST7, contained reconstructed ESBL-harbouring plasmids from both humans and animals or meat. Further rounds of OrthoMCL analyses showed that the reconstructed plasmids within each of these two sets were highly similar to each other: a maximum of only four SNPs (all attributable to p53C_2) was found in the 40 kbp plasmid core of the IncI1/ST3 subset, whereas no SNPs were found in the almost 50 kbp plasmid core of the IncI1/ST7 subset (Fig. 6A). Similarly, a subset of the bla_{CMY-2}-carrying IncK plasmids contained a plasmid core of almost 37 kbp with a maximum of 27 SNPs (Fig. 6B), which were mostly attributable to p435_1. Leaving out p435_1 from the comparisons revealed a maximum of only seven SNPs. These data strongly support the existence of ESBL-associated IncI1 and AmpC-associated IncK plasmids that have spread through phylogenetically distinct *E. coli* populations, possibly contributing to the dissemination of ESBLs and AmpC-type β-lactamases through the food-chain.

Validation of plasmid reconstructions by PacBio sequencing of strains 53C and FAP1

To validate the conclusions drawn from the PLACNET reconstructions, we sequenced two strains (53C and FAP1) using long-read DNA sequencing technology (Pacific Biosciences). Strain 53C was selected because it has both an IncI1 and an IncK plasmid, carrying $bla_{CTX-M-1}$ and bla_{CMY-2}, respectively. Strain FAP1 was selected because it contained an IncI1 plasmid of the same lineage as the one in strain 53C (Fig. 6A). The total amount of reconstructed plasmid sequence for strains 53C and FAP1 was 338 kbp and 319 kbp, respectively (Table 2). Genomes were assembled to an average depth of 66.7- and 77.0-fold, respectively, resulting in 11 contigs for strain 53C and five contigs for strain FAP1 (S4 Table). Inspection of the contig sequences showed the presence of four large plasmids in both strains. These were assigned to Inc groups F, I1 (carrying $bla_{CTX-M-1}$), I2, and K (carrying bla_{CMY-2}) in strain 53C and F, I1 (carrying $bla_{CTX-M-1}$), and I2, in strain FAP1. A single plasmid in strain FAP1 could not be assigned to an Inc group. Except for the IncI1 plasmid of FAP1, all plasmid contigs could be circularized (S4 Table). The plasmid content was in agreement with our reconstructions, except for two inconsistencies in strain FAP1: (i) PLACNET did not assign a $bla_{CTX-M-1}$ gene to its IncI1 plasmid, and (ii) PLACNET reconstructed two IncF plasmids. Blast analysis of both reconstructed IncF plasmids against the FAP1 long-read assembly suggested that they should indeed have been merged into one single plasmid. The reason for this incorrect prediction by

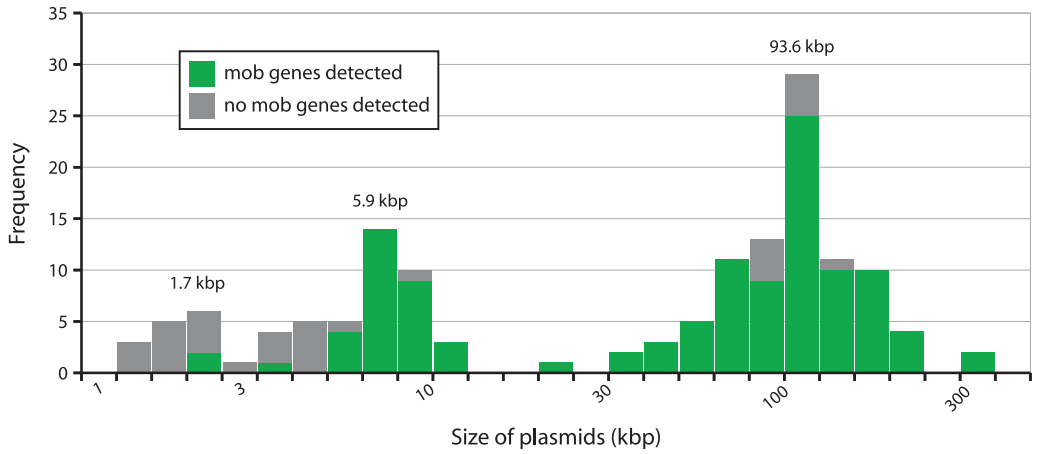


Fig. 3. Distribution of plasmid sizes in the collection of 32 sequenced *E. coli* strains. The histogram shows the total number of reconstructed plasmids corresponding to each size class (in a logarithmic scale). The plasmid size distribution shows a trimodal abundance curve. Numbers above the three peaks refer to the median size for each class. Plasmids in which a relaxase gene was detected are shown in green and those in which it was not detected are shown in grey.
doi:10.1371/journal.pgen.1004776.g003

PLACNET is unclear, but in the constellation network the two plasmids were relatively far away from each other, suggesting that the IncF plasmid in FAP1 is a fusion between previously observed IncF plasmids present in the reference database. These data show that caution must be taken in case PLACNET predicts multiple plasmids of the same Inc group in one strain. For the remaining plasmids, blast analysis showed that the precision rate of PLACNET was high, ranging from 97–100% (Table 3). Also in terms of sensitivity, PLACNET performed well being able to recover 72.1–99.7% of the plasmids (Table 3). The plasmid regions that were not reconstructed by PLACNET mostly aligned with small scaffolds (average size of 2.0±1.9 kbp, n = 33) in the assemblies built from Illumina short-read data, which indicates that these regions are difficult to assemble. Notably, these small scaffolds encoded many mobile element-, phage-, transposon- and integrase-associated proteins (29.7% of all predicted proteins in these scaffolds) as compared to the correctly assigned scaffolds, where only 6.7% of the proteins had these predicted roles. These observations are in line with results obtained from the PLACNET validation analyses described in [23] and show that PLACNET efficiently reconstructs plasmids from WGS data. Finally, the PLACNET-based prediction that both IncI1 plasmids from strains 53C and FAP1 are highly similar (Fig. 6A) was confirmed by aligning the two complete IncI1 plasmid sequences assembled from the long-read sequencing data. Filtering out repetitively aligning regions resulted in a pairwise alignment of 94.8 kbp containing only 4 SNPs. These data further substantiate our conclusions regarding highly successful plasmid lineages disseminating cephalosporin resistance.

Discussion

We assessed the epidemiology of ESBL-producing *E. coli* from humans, animals and food using WGS. Our findings strongly suggest the existence of highly successful ESBL-carrying plasmids facilitating transmission of ESBL genes between different reservoirs. This has important implications for our understanding of the dynamics of the spread of ESBL genes and for evaluating control measures.

Several strains that were sequenced in this study and which originated from humans and poultry had previously been considered indistinguishable based on MLST, plasmid and ESBL gene typing, suggesting clonal transfer of these strains through the food-chain, to humans [15]. The claim that ESBL-producing *E. coli* strains from humans and poultry are frequently identical was also made in other studies that made use of traditional sequence-based typing methods [14,16]. However, as has been demonstrated for different bacterial pathogens and in varied contexts, especially bacterial outbreak investigations, WGS provides superior resolution over traditional typing methods in terms of ruling in and out epidemiological connections between strains [26–28]. Similarly, we demonstrate that conclusions on the clonal spread of ESBL-producing *E. coli* through the food-chain cannot realistically be drawn on the basis of traditional sequence-based typing methods, due to their insufficient discriminative power. More specifically, we found that none of the five pairs of human and poultry-associated isolates, previously typed as indistinguishable, were particularly closely related. The most similar pair of isolates differed by 1263 SNPs/Mbp compared to a difference of 1.8 SNPs/Mbp for known/expected clonally related isolates. Hence, inferences from classical typing-based studies regarding the extent of transfer of ESBL-producing *E. coli* strains from animals via food to humans and the burden of disease and mortality due to the use of third-generation cephalosporins in food production must be considered as highly speculative [11].

In fact, our findings strongly suggest that distinct plasmids disproportionately contribute to the spread of antibiotic resistance between different reservoirs. We have demonstrated the existence of highly similar cephalosporin resistance-encoding IncI1/ST3 (40.0 kbp core, 0–4 SNPs), IncI1/ST7 (49.7 kbp core, 0 SNPs), and IncK (36.9 kbp core, 0–27 SNPs) plasmids in different reservoirs. Reconstructed $bla_{\text{CTX-M-1}}$-carrying IncI1/ST3 plasmids were found in one human and two poultry isolates, $bla_{\text{CTX-M-1}}$-carrying IncI1/ST7 plasmids were found in three human, two poultry, and one pig isolate; and $bla_{\text{CMY-2}}$-carrying IncK plasmids were found in five human and four poultry isolates. The isolates carrying these plasmids belonged to evolutionarily distinct backgrounds (IncI1 in phylogroups A, B1 and B2; IncK in phylogroups A, B1, B2, D and F), suggesting that these plasmids efficiently spread through *E. coli* populations and play an important role in the dissemination of ESBL and AmpC-type β-lactamases between different reservoirs.

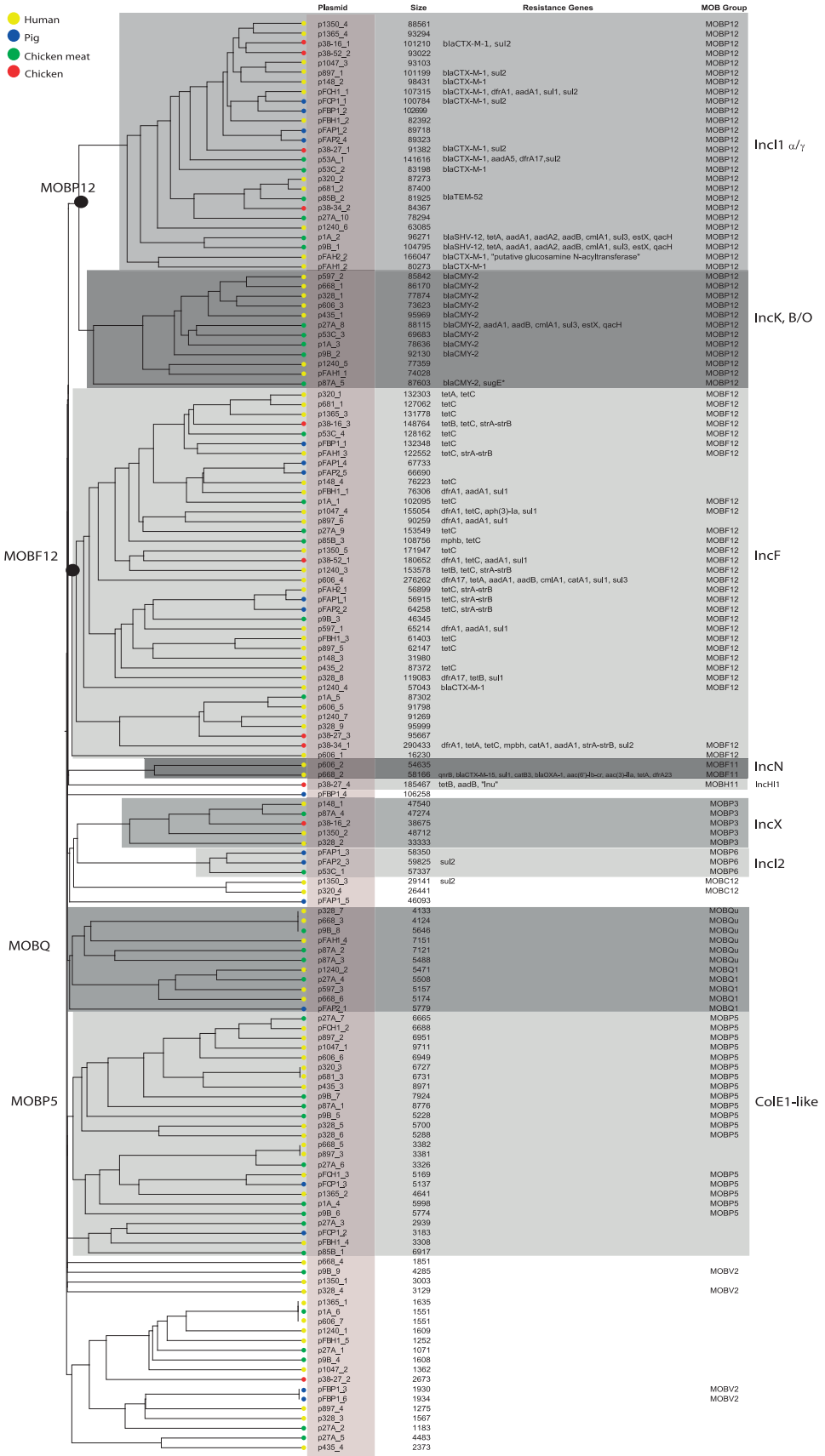

Human
Pig
Chicken meat
Chicken
Plasmid
Size
Resistance Genes
MOB Group
MOBP12
MOBF12
MOBQ
MOBP5
IncI1 α/γ
IncK, B/O
IncF
IncN
IncHI1
IncX
IncI2
ColE1-like

Fig. 4. Hierarchical clustering dendrogram of reconstructed plasmids contained in the collection of 32 sequenced *E. coli* strains. The dendrogram was constructed as explained in Methods. Reconstructed plasmids are indicated with colored bullets according to isolation source. The dendrogram construction automatically grouped plasmids into Inc families, which are shown by background colours. Mob types are also indicated. Additional columns show plasmid sizes, resistance genes and MOB subfamilies.
doi:10.1371/journal.pgen.1004776.g004

Based on their genetic content, the IncI1 and IncK plasmids in our dataset clustered into specific sub-branches that did not contain any previously characterised plasmid. However, phylogenetic analyses revealed that these sub-branches could be split into evolutionarily distinct plasmids, some of them being distantly related to previously sequenced plasmids. These findings suggest that evolutionarily distinct plasmids have been accumulating genes from the same genetic reservoir, resulting in plasmids with a similar genetic inventory. The reconstructed IncI1 plasmids all harboured a characteristic shufflon-related gene that was absent from previously characterised IncI1 plasmids. Shufflons are site-specific recombination systems that produce variable C-terminal extensions of the PilV adhesin, resulting in variations of recipient ability in IncI1 plasmid mating [29]. Whether this shufflon explains the promiscuous nature of ESBL-carrying IncI1 plasmids remains to be determined.

One important question is to what extent the IncI1 and IncK resistance plasmids found in this study have spread beyond The Netherlands. Given the trees in Fig. 6, it is clear that currently available plasmid sequences in public databases do not contain any plasmids that are particularly closely related to our reconstructed IncI1 and IncK plasmids. The pMLST repository (http://pubmlst.org/plasmid/) shows that $bla_{\text{CTX-M-1}}$-carrying IncI1/ST3 plasmids have been isolated from six different European countries, whereas $bla_{\text{CTX-M-1}}$-carrying IncI1/ST7 plasmids have until now been isolated only from The Netherlands and Germany. The location of $bla_{\text{CMY-2}}$ on an IncK plasmid, as found here, has only been occasionally reported before, in The Netherlands [30,31], but also in Sweden [32] and Canada [33]. Future sequencing projects are needed to determine whether the previously identified plasmids isolated outside The Netherlands are closely related to those described here.

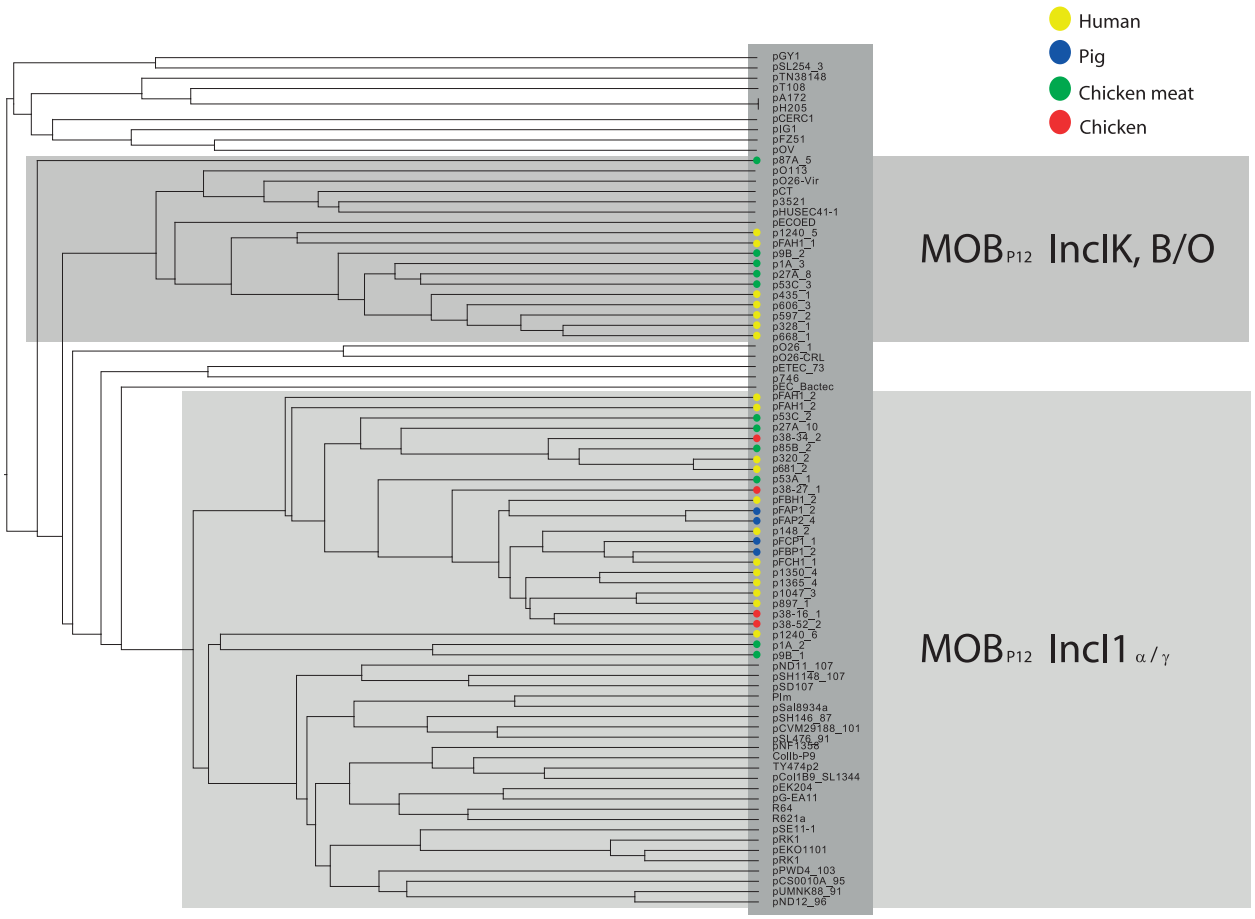


Fig. 5. Hierarchical clustering dendrogram of reconstructed IncI1 and IncK plasmids contained in the collection of 32 sequenced *E. coli* strains together with relevant and similar reference plasmids. The dendrogram was constructed as explained in Methods. Reconstructed plasmids are indicated with colored bullets according to isolation source. All other (reference) plasmids were taken from public sequence repositories.
doi:10.1371/journal.pgen.1004776.g005

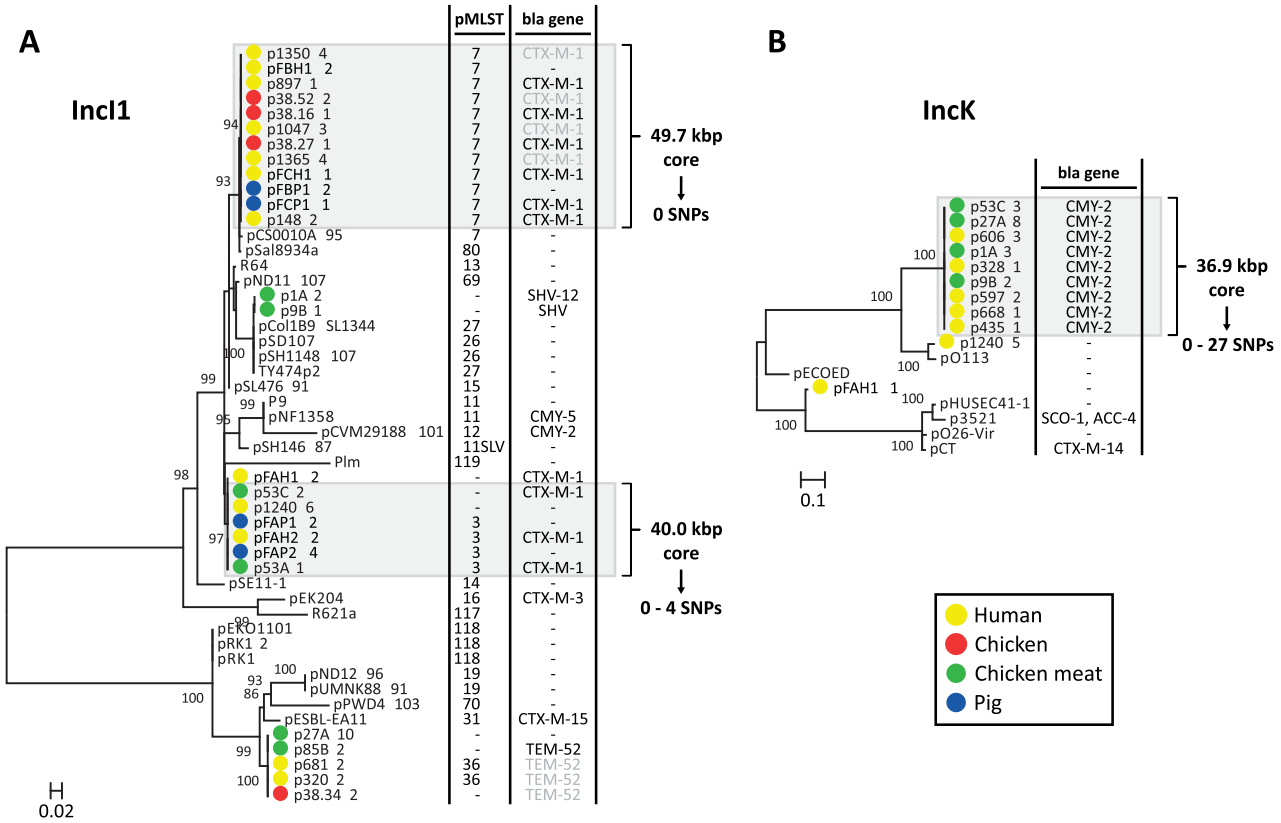


Fig. 6. Phylogeny of reconstructed IncI1 and IncK plasmids and their closest relatives. Phylogenetic tree of IncI1 plasmids built from 763 variable positions present in an 8.6 kbp alignment, representing 8 core proteins (S3 Table) (A). Phylogenetic tree of IncK plasmids built from 2724 variable positions present in a 19.9 kbp alignment, representing 27 core proteins (S3 Table) (B). Bootstrap support was implemented by running 1000 bootstrap replicates. Reconstructed plasmids are indicated with coloured bullets according to isolation source (plasmid names include associated strain names, followed by a unique plasmid identifier). All other plasmids were taken from public sequence repositories. Plasmid STs (for IncI1 only) and encoded β-lactamases (with the exception of TEM-1, which does not provide resistance to third generation cephalosporins) are indicated to the right of the trees. 11SLV indicates a single locus variant of ST11. pMLST negative means that the reconstructed plasmid lacks one or more pMLST loci. *Bla* genes in light grey were not connected to the reconstructed plasmid, but should be located on this plasmid according to typing data. Light grey panels indicate potential epidemic $bla_{CTX\text{-}M\text{-}1}$- and $bla_{CMY\text{-}2}$-carrying plasmids. Core genome analysis of these plasmid subsets revealed virtually identical backbones of up to 50 kbp.
doi:10.1371/journal.pgen.1004776.g006

We found that none of the human *E. coli* strains in our dataset were closely related to strains from poultry. In contrast, nine out of 17 human isolates (53%) contained a $bla_{CTX\text{-}M\text{-}1}$ or a $bla_{CMY\text{-}2}$ gene located on plasmids that were highly similar to those found in poultry. These data cannot be interpreted to mean that clonal transfer of antibiotic resistant *E. coli* strains between poultry and humans does not occur, but rather that such transfer occurs less frequently than the transfer of resistance plasmids between both reservoirs. One drawback of our study is that we have used a relatively small sample size (32 strains). Future studies, using larger sample sizes, are needed in order to make more accurate estimates of the relative (and absolute) contributions of clonal versus plasmid transfer towards the spread of antibiotic resistance and the associated health-care burden. In addition, our study focuses on IncI1 and IncK plasmids. Future studies are needed that also focus on other plasmid families, such as IncF plasmids, which are commonly detected in *E. coli* from human infections and are associated with the dissemination of many virulence and antibiotic resistance determinants [34,35].

Conjugal transfer of plasmids carrying antibiotic resistance genes has been shown to frequently occur among *Enterobacteriaceae* in different environments, including milk, meat, and feces, even in the absence of antibiotic pressure [36,37]. Moreover, it has been shown that *bla*-carrying plasmids are readily transferred from invading *Enterobacteriaceae* to *Enterobacteriaceae* that are indigenous to the animal and human intestine and that the invading clone itself generally does not persist in the intestine [38,39]. Nonetheless, it is difficult to infer to what extent the reservoir of *bla*-type resistance genes in poultry contributes to the carriage of such genes by human *E. coli* strains. If successful plasmids are largely responsible for the rising prevalence of ESBL- and AmpC-producing *E. coli* in healthy humans, their emergence in poultry and humans may simply be a reflection of selection of strains carrying these plasmids due to antibiotic usage in human and veterinary medicine.

A better understanding of the dynamics of ESBLs and other resistance genes in different hosts is needed to design effective control measures, both in the community and within health care settings. Our findings strongly suggest the occurrence of clonal transfer of ESBL-producing *E. coli* between pigs and pig farmers, which may well occur through direct contact or through aerosols. Whether such events represent a public health threat remains to be

Table 3. PLACNET precision and sensitivity rates for seven reconstructed plasmids.

PLACNET plasmid (kbp)	Corresponding PacBio plasmid (kbp)	PLACNET precision (%)	PLACNET sensitivity (%)*
p53C_1 (57.3)	IncI2(56.9)	97.0	96.4
p53C_2 (83.2)	IncI1 (109.7)	99.1	72.1
p53C_3 (69.7)	IncK (86.0)	100	76.7
p53C_4 (128.2)	IncF (134.8)	97.6	90.6
pFAP1_2 (89.7)	IncI1 (129.4)	100	82.1
pFAP1_3 (58.4)	IncI2 (62.4)	100	95.7
pFAP1_5 (46.1)	unclassified plasmid (46.2)	100	99.7

* Note that the region of the PacBio plasmid that was recovered can exceed the reconstructed plasmid size because of repetitive elements, which are collapsed in Illumina assemblies, but are uncollapsed in PacBio assemblies. For each reconstructed plasmid the PLACNET precision rate was calculated by the formula [(assembly size of all correctly assigned scaffolds/assembly size of all correctly+incorrectly assigned scaffolds)×100%]. Sensitivity reflects the percentage of each plasmid sequence (assembled using PacBio data) that was correctly reconstructed in the PLACNET analysis. The sensitivity rate was calculated by the formula [(total nr. of non-overlapping aligning residues/size of the plasmid that was assembled using PacBio data)×100%].

doi:10.1371/journal.pgen.1004776.t003

determined. The occurrence of transmission of ESBL-producing *E. coli* from poultry through the food-chain is less evident. The occurrence of highly-related plasmids that carry ESBL- and AmpC-type resistance genes among genotypically distinct *E. coli* strains from different sources is cause for concern because this suggests that plasmids can spread with relative ease between the different reservoirs and the spread of these plasmids may be exceedingly difficult to control. Clearly, there still remains an urgent need to minimize the use of third-generation cephalosporins in animal husbandry as this is an important selective pressure for the occurrence of ESBL- and AmpC-producing *E. coli* in animals raised for food production.

Materials and Methods

Isolates and molecular analyses

The genomes of 32, mostly ESBL-producing, *E. coli* strains isolated from different reservoirs in The Netherlands in the period 2006–2011, were sequenced. One set of isolates (n = 24) has been studied previously using classical typing methods [15,18]. This set contained strains from human clinical infections (n = 13) which had been obtained from geographically dispersed laboratories in The Netherlands, servicing secondary and tertiary care hospitals, general practitioners and long-term care facilities. Additional isolates were from chickens raised on production farms (n = 4) and chicken retail meat (n = 7) (Table 1). All 24 isolates were previously genotyped by MLST [40] (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli) and plasmid characterization was previously performed using PCR-based replicon typing [31,41] and additional pMLST for IncI1 plasmids [42,43] (http://pubmlst.org/plasmid/). Detection of ESBL genes had been performed for all 24 strains using microarray analysis and gene sequencing [44]. In addition, detection of AmpC-type β-lactamase-encoding genes had been performed for 11 strains, using gene sequencing [18]. The association between ESBL/AmpC genes and plasmids was previously determined by both Southern blot hybridization and transformation [31]. Four non-ESBL-producing isolates were included as controls and were analysed for the carriage of plasmids that can incorporate ESBL genes via horizontal gene transfer. The second set of isolates contained eight ESBL-producing strains that had been isolated from three different pig farms in The Netherlands in 2011 (Table 1). These farm strains were part of a larger cohort that will be described in detail elsewhere (Dohmen *et al.*, unpublished data). For one farm (farm A), four strains were collected, two from different fecal pools of six unique pigs and two from the feces of different farmers. For each of the other two farms (farms B and C), one strain was collected from a fecal pool of six pigs and one from the feces of a farmer. Detection of the ESBL ($bla_{\text{CTX-M-1}}$) gene was performed using a CTX-M-1 group-specific PCR and additional gene sequencing (Dohmen *et al.*, unpublished data).

Genome sequencing and assembly

Genomic DNA was isolated from cell pellets using the Ultraclean Microbial DNA isolation kit (Mo Bio Laboratories, Inc., Carlsbad, CA, USA) according to the manufacturer's instructions. Strains were sequenced using Illumina HiSeq 2000 sequencing technology (Illumina, Inc., San Diego, CA, USA) generating 90 bp paired-end reads from a library with an average insert size of 500 bp and a total amount of quality-filtered raw sequence of over 600 Mbp per strain. Quality filtering included the removal of duplicate reads and reads that contained ≥15 bp overlap with the adapter sequences. The corresponding paired-end reads were also removed in these cases. Reads were assembled *de novo* using SOAPdenovo v1.05 [45]. For each Illumina dataset, a range of different k-mer lengths (21–63 bp) was empirically tested to obtain the assembly with the lowest number of scaffolds of size ≥500 bp. In cases where more than one assembly contained the lowest number of scaffolds, the parameters of choice to pick the best assembly were: the lowest number of contigs of size ≥200 bp, the highest N50 for the scaffolds, and the highest N50 for the contigs, in order of priority. Assembly statistics are reported in S1 Table. Two strains (53C and FAP1, Table 1) were also sequenced on a Pacific Biosciences RS II instrument (Pacific Biosciences, Inc., Menlo Park, CA, USA). Libraries were prepared using the PacBio 20 kbp library preparation protocol. Size selection (5 kbp cut-off) of the final libraries was performed using a BluePippin instrument (Sage Science, Inc., Beverly, MA, USA). Sequencing was performed using P4-C2 chemistry. Three and five SMRT cells were used for sequencing strains FAP1 and 53C, respectively, generating 159191 and 95263 reads and a total of 997.1 and 471.5 Mbp, respectively. Reads were assembled using HGAP v3 (Pacific Biosciences, SMRT Analysis Software v2.2.0). Minimus2, part of the AMOS package [46], was used to circularize contigs. The SMRT Analysis Software was used to map reads back to the

contigs and correct sequences after circularization. Assembly statistics are reported in S4 Table.

Sequence data analysis

Publicly available sequence data were retrieved from GenBank (ftp://ftp.ncbi.nih.gov/genomes/Bacteria and ftp://ftp.ncbi.nih.gov/genomes/Bacteria_DRAFT). Whole genome sequence data for 126 *Escherichia* and 12 *Shigella* species were downloaded in June 2012, whereas sequence data for 4188 completely sequenced plasmids, 797 of them from *Enterobacteriaceae*, were downloaded in June 2013. The strains that were sequenced in this study were annotated with RAST v4.0 [47] using default settings. Predicted proteins were assigned to Clusters of Orthologous Groups (COG) [48] as described previously [49]. On the basis of COG assignments, a core proteome was defined by (i) extracting, per analysed genome, all proteins with one or more COGs assignments and which represented the only protein in that given COG or combination of COGs and by (ii) selecting from those proteins the ones that occurred in all genomes analysed. Alternatively, in the cases where smaller genomic datasets were analysed (see Results), core proteomes were determined by first subjecting all associated protein sequences to an all-vs-all blastp similarity search (defaults settings, except for: -F 'm S'; -e 1×10^{-5}; -z [the total number of proteins used in the analysis]). Groups of orthologous proteins were determined from the blastp output using OrthoMCL v2.0.2 [20]. Orthologous groups with exactly one representative protein from each input genome were considered to be part of the core proteome. Core genome alignments were built as follows: for each group of orthologous proteins, the corresponding nucleotide sequences were extracted and aligned using Muscle v3.7 [50], after which gaps were stripped from each alignment using trimAl v1.2 [51]. The resulting alignments were concatenated to yield a core genome alignment. Phylogenies were reconstructed by building maximum likelihood phylogenetic trees from the variable positions in core genome alignments using RAxML v7.2.8 [52] under the GTRCAT model. Confidence was inferred by running 100 or 1000 bootstrap replicates under the same model. Trees were mid-point rooted and visualised in MEGA v5.05 [53]. Bowtie2 [54] was used for mapping Illumina reads against scaffolded assemblies and gene sequences. MLST profiling of sequenced bacteria was performed using MLST v1.6 [55]. Pairwise large-scale nucleotide alignments were built using NUCmer v3.23 (with –mum option), which is part of the MUMmer package [56].

Plasmid reconstructions from WGS data

Plasmid reconstructions were based on the Plasmid Constellation Networks (PLACNET) method of genome representation [23]. In short, for all genomes, a PLACNET representation that clusters all plasmid-associated contigs was built using (i) contig similarities with reference genomes, (ii) all possible contig linkages, and (iii) plasmid-specific relaxase and replication initiator genes. This information was implemented in a network, where genomic contigs, together with reference plasmid and genome sequences are shown as nodes. The nodes are linked by edges of homology and scaffolding information. As a result, contigs fall into clusters, the largest one being the chromosome and additional ones being plasmids. Manual curation of the resulting networks helped solving most of the remaining ambiguities. Reference data from GenBank contained 4188 plasmids and 2728 chromosomes. Contig similarity analysis was performed using megablast against these reference data. Contig homology edges were defined by the five best blast hits (e-value $<1\times10^{-20}$). Scaffolds were determined by mapping all reads against contigs using Bowtie2 [54], and allocating as scaffold links all discordant paired-end reads that matched two different contigs. To provide additional evidence for the plasmid origin of a cluster, a blastp search against in-house databases containing plasmid-specific relaxases and replication initiator proteins was performed. Contigs encoding these proteins were tagged in the PLACNET. Plasmid Neighbour-Joining dendrograms were built based on previously described methodologies [57] using CD-HIT [58] to construct protein profiles and the Jaccard formula to calculate distance metrics between profiles. PLACNET results were validated as follows: for the two strains 53C and FAP1, the scaffolds assigned to the reconstructed plasmids were queried using megablast against the assemblies resulting from Pacific Biosciences (PacBio) sequencing. The best blast hit (e-value $\leq1\times10^{-10}$) was inspected to assess whether the scaffolds had been assigned to the correct plasmid. For each reconstructed plasmid the PLACNET precision rate was calculated by the formula [(assembly size of all correctly assigned scaffolds/assembly size of all correctly+incorrectly assigned scaffolds)$\times$100%]. To assess PLACNET sensitivity (the percentage of each plasmid sequence size that was recovered) blast hits against the corresponding PacBio plasmids were collected (e-value $\leq$ 1×10^{-10}, minimum of 250 identical residues). The sensitivity rate was calculated by the formula [(total nr. of non-overlapping aligning residues found by blast/size of the PacBio plasmid)$\times$100%].

Accession numbers

All sequence data have been deposited at DDBJ/EMBL/GenBank. Accession numbers for the Illumina sequence data are listed in Table 1. Pacific Biosciences sequence data have been deposited with accession numbers PRJNA260957 for strain 53C and PRJNA260958 for strain FAP1.

Supporting Information

S1 Table Assembly statistics of genome sequences of strains sequenced with Illumina technology.
(DOCX)

S2 Table Assembled bla_{TEM} genes and ambiguous nucleotide positions.
(DOCX)

S3 Table Core proteomes of IncI1 and IncK plasmids.
(DOCX)

S4 Table Assembly statistics of genome sequences of strains 53C and FAP1 (sequenced with Pacific Biosciences long-read technology).
(DOCX)

Acknowledgments

The Centre for Molecular and Biomolecular Informatics (CMBI, Radboud University Nijmegen, Nijmegen, The Netherlands) is acknowledged for providing computing space. The Pacific Biosciences sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no).

Author Contributions

Conceived and designed the experiments: MJMB RJLW FdlC WvS. Performed the experiments: JS YD JH YL NL ATK. Analyzed the data: MdB VFL MdT. Contributed reagents/materials/analysis tools: JS WD DJJH ACF. Wrote the paper: MdB VFL MdT WD DJJH ACF MJMB RJLW FdlC WvS.

References

1. De Kraker MEA, Davey PG, Grundmann H, on behalf of the BURDEN study group (2011) Mortality and Hospital Stay Associated with Resistant *Staphylococcus aureus* and *Escherichia coli* Bacteremia: Estimating the Burden of Antibiotic Resistance in Europe. PLoS Med 8: e1001104. doi:10.1371/journal.pmed.1001104.
2. Mauldin PD, Salgado CD, Hansen IS, Durup DT, Bosso JA (2010) Attributable Hospital Cost and Length of Stay Associated with Health Care-Associated Infections Caused by Antibiotic-Resistant Gram-Negative Bacteria. Antimicrob Agents Chemother 54: 109–115. doi:10.1128/AAC.01041-09.
3. Coque TM, Baquero F, Cantón R (2008) Increasing prevalence of ESBL-producing *Enterobacteriaceae* in Europe. Euro Surveill 13: pii: 19044
4. Hawkey PM, Jones AM (2009) The changing epidemiology of resistance. J Antimicrob Chemother 64: i3–i10. doi:10.1093/jac/dkp256.
5. Dubois V, Barbeyrac BD, Rogues A-M, Arpin C, Coulange L, et al. (2010) CTX-M-producing *Escherichia coli* in a maternity ward: a likely community importation and evidence of mother-to-neonate transmission. J Antimicrob Chemother 65: 1368–1371. doi:10.1093/jac/dkq153.
6. Valverde A, Coque TM, Sánchez-Moreno MP, Rollán A, Baquero F, et al. (2004) Dramatic Increase in Prevalence of Fecal Carriage of Extended-Spectrum β-Lactamase-Producing *Enterobacteriaceae* during Nonoutbreak Situations in Spain. J Clin Microbiol 42: 4769–4775. doi:10.1128/JCM.42.10.4769-4775.2004.
7. Kaper JB, Nataro JP, Mobley HLT (2004) Pathogenic *Escherichia coli*. Nat Rev Microbiol 2: 123–140. doi:10.1038/nrmicro818.
8. Smet A, Martel A, Persoons D, Dewulf J, Heyndrickx M, et al. (2008) Diversity of Extended-Spectrum β-Lactamases and Class C β-Lactamases among Cloacal *Escherichia coli* Isolates in Belgian Broiler Farms. Antimicrob Agents Chemother 52: 1238–1243. doi:10.1128/AAC.01285-07.
9. Machado E, Coque TM, Cantón R, Sousa JC, Peixe L (2008) Antibiotic resistance integrons and extended-spectrum β-lactamases among *Enterobacteriaceae* isolates recovered from chickens and swine in Portugal. J Antimicrob Chemother 62: 296–302. doi:10.1093/jac/dkn179.
10. Doi Y, Paterson DL, Egea P, Pascual A, López-Cerero L, et al. (2010) Extended-spectrum and CMY-type β-lactamase-producing *Escherichia coli* in clinical samples and retail meat from Pittsburgh, USA and Seville, Spain. Clin Microbiol Infect 16: 33–38. doi:10.1111/j.1469-0691.2009.03001.x.
11. Collignon P, Aarestrup FM, Irwin R, McEwen S (2013) Human deaths and third-generation cephalosporin use in poultry, Europe. Emerg Infect Dis 19: 1339–1340. doi:10.3201/eid.1908.120681.
12. Van de Sande-Bruinsma N, Grundmann H, Verloo D, Tiemersma E, Monen J, et al. (2008) Antimicrobial drug use and resistance in Europe. Emerg Infect Dis 14: 1722–1730. doi:10.3201/eid1411.070467.
13. Grave K, Greko C, Kvaale MK, Torren-Edo J, Mackay D, et al. (2012) Sales of veterinary antibacterial agents in nine European countries during 2005–09: trends and patterns. J Antimicrob Chemother 67: 3001–3008. doi:10.1093/jac/dks298.
14. Overdevest I, Willemsen I, Rijnsburger M, Eustace A, Xu L, et al. (2011) Extended-spectrum β-lactamase genes of *Escherichia coli* in chicken meat and humans, The Netherlands. Emerg Infect Dis 17: 1216–1222. doi:10.3201/eid1707.110209.
15. Leverstein-van Hall MA, Dierikx CM, Cohen Stuart J, Voets GM, van den Munckhof MP, et al. (2011) Dutch patients, retail chicken meat and poultry share the same ESBL genes, plasmids and strains. Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis 17: 873–880. doi:10.1111/j.1469-0691.2011.03497.x.
16. Kluytmans JAJW, Overdevest ITMA, Willemsen I, Kluytmans-van den Bergh MFQ, van der Zwaluw K, et al. (2013) Extended-spectrum β-lactamase-producing *Escherichia coli* from retail chicken meat and humans: comparison of strains, plasmids, resistance genes, and virulence factors. Clin Infect Dis Off Publ Infect Dis Soc Am 56: 478–487. doi:10.1093/cid/cis929.
17. Sabat AJ, Budimir A, Nashev D, Sá-Leão R, van Dijl J m, et al. (2013) Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull 18: 20380.
18. Voets GM, Fluit AC, Scharringa J, Schapendonk C, van den Munckhof T, et al. (2013) Identical plasmid AmpC beta-lactamase genes and plasmid types in *E. coli* isolates from patients and poultry meat in the Netherlands. Int J Food Microbiol 167: 359–362. doi:10.1016/j.ijfoodmicro.2013.10.001.
19. McNally A, Cheng L, Harris SR, Corander J (2013) The evolutionary path to extraintestinal pathogenic, drug-resistant *Escherichia coli* is marked by drastic reduction in detectable recombination within the core genome. Genome Biol Evol 5: 699–710. doi:10.1093/gbe/evt038.
20. Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189. doi:10.1101/gr.1224503.
21. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, et al. (2012) Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. Proc Natl Acad Sci U S A 109: 3065–3070. doi:10.1073/pnas.1121491109.
22. Reeves PR, Liu B, Zhou Z, Li D, Guo D, et al. (2011) Rates of Mutation and Host Transmission for an *Escherichia coli* Clone over 3 Years. PLoS ONE 6: e26907. doi:10.1371/journal.pone.0026907.
23. Lanza VF, de Toro M, Garcillán-Barcia MP, Mora A, Blanco J, Coque TM, de la Cruz F. Turbulent plasmid flux in *Escherichia coli* ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. PLoS Gen 10: e1004766
24. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F (2010) Mobility of plasmids. Microbiol Mol Biol Rev MMBR 74: 434–452. doi:10.1128/MMBR.00020-10.
25. Garcillán-Barcia MP, Francia MV, de la Cruz F (2009) The diversity of conjugative relaxases and its application in plasmid classification. FEMS Microbiol Rev 33: 657–687.
26. Den Bakker HC, Allard MW, Bopp D, Brown EW, Fontana J, et al. (2014) Rapid Whole-Genome Sequencing for Surveillance of *Salmonella enterica* Serovar Enteritidis. Emerg Infect Dis 20: 1306–1314. doi:10.3201/eid2008.131399.
27. Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, et al. (2012) Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med 366: 2267–2275. doi:10.1056/NEJMoa1109910.
28. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, et al. (2013) Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med 10: e1001387. doi:10.1371/journal.pmed.1001387.
29. Ishiwa A, Komano T (2004) PilV Adhesins of Plasmid R64 Thin Pili Specifically Bind to the Lipopolysaccharides of Recipient Cells. J Mol Biol 343: 615–625. doi:10.1016/j.jmb.2004.08.059.
30. Dierikx C, van der Goot J, Fabri T, van Essen-Zandbergen A, Smith H, et al. (2013) Extended-spectrum-β-lactamase- and AmpC-β-lactamase-producing *Escherichia coli* in Dutch broilers and broiler farmers. J Antimicrob Chemother 68: 60–67. doi:10.1093/jac/dks349.
31. Dierikx C, van Essen-Zandbergen A, Veldman K, Smith H, Mevius D (2010) Increased detection of extended spectrum beta-lactamase producing *Salmonella enterica* and *Escherichia coli* isolates from poultry. Vet Microbiol 145: 273–278. doi:10.1016/j.vetmic.2010.03.019.
32. Börjesson S, Jernberg C, Brolund A, Edquist P, Finn M, et al. (2013) Characterization of plasmid-mediated AmpC-producing *E. coli* from Swedish broilers and association with human clinical isolates. Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis 19: E309–311. doi:10.1111/1469-0691.12192.
33. Baudry PJ, Mataseje L, Zhanel GG, Hoban DJ, Mulvey MR (2009) Characterization of plasmids encoding CMY-2 AmpC beta-lactamases from *Escherichia coli* in Canadian intensive care units. Diagn Microbiol Infect Dis 65: 379–383. doi:10.1016/j.diagmicrobio.2009.08.011.\
34. Villa L, García-Fernández A, Fortini D, Carattoli A (2010) Replicon sequence typing of IncF plasmids carrying virulence and resistance determinants. J Antimicrob Chemother 65: 2518–2529. 34.
35. Woodford N, Carattoli A, Karisik E, Underwood A, Ellington MJ, et al. (2009) Complete nucleotide sequences of plasmids pEK204, pEK499, and pEK516, encoding CTX-M enzymes in three major *Escherichia coli* lineages from the United Kingdom, all belonging to the international O25:H4-ST131 clone. Antimicrob Agents Chemother 53: 4472–4482. doi:10.1128/AAC.00688-09.
36. Kruse H, Sørum H (1994) Transfer of multiple drug resistance plasmids between bacteria of diverse origins in natural microenvironments. Appl Environ Microbiol 60: 4015–4021.
37. Warnes SL, Highmore CJ, Keevil CW (2012) Horizontal transfer of antibiotic resistance genes on abiotic touch surfaces: implications for public health. mBio 3. doi:10.1128/mBio.00489-12.
38. Cavaco LM, Abatih E, Aarestrup FM, Guardabassi L (2008) Selection and persistence of CTX-M-producing *Escherichia coli* in the intestinal flora of pigs treated with amoxicillin, ceftiofur, or cefquinome. Antimicrob Agents Chemother 52: 3612–3616. doi:10.1128/AAC.00354-08.
39. Goren MG, Carmeli Y, Schwaber MJ, Chmelnitsky I, Schechner V, et al. (2010) Transfer of carbapenem-resistant plasmid from *Klebsiella pneumoniae* ST258 to *Escherichia coli* in patient. Emerg Infect Dis 16: 1014–1017. doi:10.3201/eid1606.091671.
40. Wirth T, Falush D, Lan R, Colles F, Mensa P, et al. (2006) Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol 60: 1136–1151. doi:10.1111/j.1365-2958.2006.05172.x.
41. Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL, et al. (2005) Identification of plasmids by PCR-based replicon typing. J Microbiol Methods 63: 219–228. doi:10.1016/j.mimet.2005.03.018.
42. García-Fernández A, Chiaretto G, Bertini A, Villa L, Fortini D, et al. (2008) Multilocus sequence typing of IncI1 plasmids carrying extended-spectrum beta-lactamases in *Escherichia coli* and *Salmonella* of human and animal origin. J Antimicrob Chemother 61: 1229–1233. doi:10.1093/jac/dkn131.
43. Jolley KA, Maiden MCJ (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11: 595. doi:10.1186/1471-2105-11-595.
44. Cohen Stuart J, Dierikx C, Al Naiemi N, Karczmarek A, Van Hoek AHAM, et al. (2010) Rapid detection of TEM, SHV and CTX-M extended-spectrum beta-

lactamases in *Enterobacteriaceae* using ligation-mediated amplification with microarray analysis. J Antimicrob Chemother 65: 1377–1381. doi:10.1093/jac/dkq146.
45. Li Y, Hu Y, Bolund L, Wang J (2010) State of the art de novo assembly of human genomes from massively parallel sequencing data. Hum Genomics 4: 271–277.
46. Schatz MC, Phillippy AM, Sommer DD, Delcher AL, Puiu D, et al. (2013) Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies. Brief Bioinform 14: 213–224. doi:10.1093/bib/bbr074.
47. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, et al. (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9: 75. doi:10.1186/1471-2164-9-75.
48. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29: 22–28.
49. de Been M, van Schaik W, Cheng L, Corander J, Willems RJ (2013) Recent recombination events in the core genome are associated with adaptive evolution in *Enterococcus faecium*. Genome Biol Evol 5: 1524–1535. doi:10.1093/gbe/evt111.
50. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797. doi:10.1093/nar/gkh340.
51. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinforma Oxf Engl 25: 1972–1973. doi:10.1093/bioinformatics/btp348.
52. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinforma Oxf Engl 22: 2688–2690. doi:10.1093/bioinformatics/btl446.
53. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731–2739. doi:10.1093/molbev/msr121.
54. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359. doi:10.1038/nmeth.1923.
55. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, et al. (2012) Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol 50: 1355–1361. doi:10.1128/JCM.06094-11.
56. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biol 5: R12. doi:10.1186/gb-2004-5-2-r12.
57. Tekaia F, Yeramian E (2005) Genome trees from conservation profiles. PLoS Comput Biol 1: e75. doi:10.1371/journal.pcbi.0010075.
58. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinforma Oxf Engl 22: 1658–1659. doi:10.1093/bioinformatics/btl158.

Table S1. Assembly statistics of genome sequences of strains sequenced with Illumina technology.

Strain	# scaffolds	N50 (bp)	N90 (bp)	Longest scaffold (bp)	Average Nt coverage	GC%	Total assembled sequence (Mbp)	# genes
148	112	125332	30827	296899	120	50.57	4.91	4767
320	103	147419	37926	367231	117	50.69	5.11	4964
681	96	159817	42261	396973	117	50.73	5.08	4932
38.27	182	95667	18156	323445	112	50.29	5.20	5073
38.34	179	109891	19368	452240	110	50.30	5.33	5246
53A	89	197763	37922	833818	122	50.65	4.88	4805
85B	113	144724	37770	410503	116	50.34	5.11	5020
1240	147	194347	27095	631572	112	50.62	5.26	5024
1350	153	109054	28271	564784	113	50.59	5.21	5147
1365	112	135615	30837	475403	117	50.68	5.09	4947
38.16	106	176995	30927	331015	119	50.66	5.00	4830
897	130	217264	40739	601538	113	50.69	5.23	5083
1047	114	226645	32629	579834	115	50.73	5.14	4983
38.52	154	147220	19997	396795	113	50.57	5.20	5017
53C	138	158467	26710	464801	109	50.62	5.41	5276
435	184	160556	26937	485993	106	50.42	5.44	5230
328	126	136750	30766	456828	109	50.62	5.45	5335
597	62	318266	46813	536123	119	50.57	5.02	4924
668	111	198351	24899	511643	115	50.58	5.17	5012
606	151	119134	25096	391442	110	50.62	5.34	5187
1A	95	202867	31244	448803	115	50.54	5.11	4971
27A	132	175774	23021	411303	114	50.49	5.13	4983
9B	249	106534	14198	257148	107	50.61	5.41	5326
87A	242	99611	13081	209419	100	50.38	5.40	5041
FAH1	143	154416	30162	371825	114	50.67	5.13	5177
FAH2	98	155406	37441	318151	117	50.62	5.05	5029
FAP1	112	155398	33965	371715	116	50.46	5.12	5116
FAP2	106	155350	37417	371666	117	50.51	5.07	5067
FBH1	114	103452	23824	344819	118	50.73	5.00	4941
FBP1	139	100668	30761	286798	116	50.55	5.10	5152
FCH1	162	91851	27790	217493	116	50.66	5.06	5092
FCP1	98	118819	25417	297448	125	50.59	4.73	4641

All statistics shown are based on the scaffolded assemblies and scaffolds of size ≥ 500 bp.

Table S2. Assembled *bla*$_{TEM}$ genes and ambiguous nucleotide positions.

Strain	Typed *bla*$_{TEM}$ gene (screening for *bla*$_{TEM-1}$ was not performed)	*bla*$_{TEM}$ gene in WGS	Protein length	*bla*$_{TEM}$ average Nt coverage ± s.d.	*bla*$_{TEM}$ relative coverage *	Position of ambiguous Nt (& corresp. aa position)†	Coverage of ambiguous Nt	Most abundant Nts at ambiguous position (& resulting aa). Consensus (assembled) Nts are in bold.
320	TEM-52	TEM-20	280 aa	325.0 ± 79.4	2.8	310 (104)	337	175 × A (Lys), **162 × G (Glu)**
						402 (134)	437	**232 × G (Ala)**, 205 × T (Ala)
						545 (182)	347	**192 × C (Thr)**, 155 × T (Met)
						712 (238)	233	**120 × A (Ser)**, 113 × G (Gly)
681	TEM-52	TEM-20	280 aa	295.4 ± 70.8	2.5	310 (104)	295	151 × A (Lys), **144 × G (Glu)**
						402 (134)	403	**222 × G (Ala)**, 181 × T (Ala)
						545 (182)	323	**179 × C (Thr)**, 142 × T (Met)
						712 (238)	225	**119 × A (Ser)**, 106 × G (Gly)
38.34	TEM-52	TEM-1	280 aa	326.4 ± 73.9	3.0	310 (104)	311	**168 × G (Glu)**, 143 × A (Lys)
						402 (134)	399	**215 × T (Ala)**, 184 × G (Ala)
						545 (182)	426	**228 × T (Met)**, 198 × C (Thr)
						712 (238)	240	**120 × G (Gly)**, 116 × A (Ser)
85B	TEM-52	TEM-52	286 aa	181.2 ± 36.7	1.6	-	-	-
27A	TEM-52	TEM-52	286 aa	125.9 ± 23.8	1.1	-	-	-
148	-	TEM-1	286 aa	171.0 ± 24.0	1.4	-	-	-
1240	-	TEM-1	286 aa	413.0 ± 75.2	3.7	545 (182)	407	**221 × T (Met)**, 186 × C (Thr)
1350	-	TEM-1	286 aa	169.1 ± 26.5	1.5	-	-	-
1365	-	TEM-1	286 aa	170.6 ± 28.0	1.5	-	-	-
38.16	-	TEM-1	286 aa	315.7 ± 38.3	2.7	-	-	-
328	-	TEM-1	286 aa	174.1 ± 22.3	1.6	-	-	-
668	-	TEM-1	286 aa	106.6 ± 19.5	0.9	-	-	-
606	-	TEM-1	286 aa	120.4 ± 16.3	1.1	-	-	-
87A	-	TEM-1	286 aa	194.7 ± 37.6	1.9	-	-	-
FAH1	-	TEM-1	286 aa	142.6 ± 21.1	1.3	-	-	-
FAH2	-	TEM-1	286 aa	153.9 ± 22.0	1.3	-	-	-
FAP1	-	TEM-1	286 aa	155.0 ± 21.1	1.3	-	-	-
FAP2	-	TEM-1	286 aa	143.9 ± 26.3	1.2	-	-	-

* *bla*$_{TEM}$ relative coverage was calculated by dividing the average *bla*$_{TEM}$ coverage by the average genomic coverage (Table S1).

† Amino acid positions are in correspondence with the amino acid positions listed for TEM-1 by the Lahey clinic (www.lahey.org/Studies/temtable.asp).

Ambiguous positions were found as follows: first, the raw Illumina reads of each strain were mapped against its own assembly (allowing a maximum of one mismatch per seed). Second, positions were designated as ambiguous when at least 50 reads mapped to it and when the second most abundantly mapped residue had a share of ≥ 25% of all mapped residues at that position. Positions covered by indels were not taken into account. The four ambiguous positions found in the assembled *bla*$_{TEM}$ genes of strains 320, 681 and 38.34 exactly corresponded to the four SNPs that differentiate *bla*$_{TEM-1}$ and *bla*$_{TEM-52}$ in the same genetic region. This suggested that both *bla*$_{TEM-1}$ and *bla*$_{TEM-52}$ are present in these strains. In strains 320 and 681 this appears to have resulted in an erroneous *bla*$_{TEM-20}$ assembly (i.e. a hybrid between *bla*$_{TEM-1}$ and *bla*$_{TEM-52}$). The single ambiguous position found in the assembled *bla*$_{TEM}$ gene of strain 1240 may point to the presence of *bla*$_{TEM-1}$ and *bla*$_{TEM-135}$ in this strain.

Table S3 Core proteomes of IncI1 and IncK plasmids.

Inc group (nr of plasmids analysed)	Core protein locus tags of representative plasmids (R64 for IncI1, pCT for IncK)	Annotation	Protein name
IncI1 (50)	R64_p081	relaxase	NikB
	R64_p090	conjugal transfer protein	TraX
	R64_p123	type IV prepilin	PilS
	R64_p124	integral membrane protein	PilR
	R64_p125	type IV pilus ATPase	PilQ
	R64_p126	type IV pilus protein	PilP
	R64_p127	type IV pilus protein	PilO
	R64_p129	type IV pilus protein	PilM
IncK (17)	pCT_053	conjugal transfer protein	TrbC
	pCT_054	conjugal transfer protein	TrbB
	pCT_057	endonuclease	ParB
	pCT_064	conserved hypothetical plasmid protein	-
	pCT_068	conjugal transfer integral membrane protein	TraY
	pCT_069	conjugal transfer protein	TraX
	pCT_070	conjugal transfer protein	TraW
	pCT_071	conjugal transfer protein	TraV
	pCT_072	conjugal transfer nucleotide-binding protein	TraU
	pCT_073	conjugal transfer protein	TraT
	pCT_074	conjugal transfer protein	TraS
	pCT_075	conjugal transfer protein	TraR
	pCT_076	conjugal transfer protein	TraQ
	pCT_077	conjugal transfer protein	TraP
	pCT_078	conjugal transfer protein	TraO
	pCT_079	conjugal transfer protein	TraN
	pCT_080	conjugal transfer protein	TraM
	pCT_087	conjugal transfer protein	TraK
	pCT_088	conjugal transfer ATPase protein	TraJ
	pCT_089	conjugal transfer lipoprotein	TraI
	pCT_090	conjugal transfer protein	TraH
	pCT_092	conjugal transfer protein	TraE
	pCT_107	type IV pilus protein	PilP
	pCT_109	type IV pilus outer membrane protein	PilN
	pCT_110	type IV pilus protein	PilM
	pCT_111	type IV pilus lipoprotein	PilL
	pCT_113	type IV pilus protein	PilI

Table S4. Assembly statistics of genome sequences of strains 53C and FAP1 (sequenced with Pacific Biosciences long-read technology).

Strain	Contig	Circular?	Molecule type	Length (kbp)	Average Nt coverage	GC%	# genes
53C	1	No	Chromosome	2271.2	70	51.20	2266
	2	No	Chromosome	1367.7	58	50.38	1416
	3	No	Chromosome	880.1	71	50.78	890
	4	No	Chromosome	637.0	57	50.49	669
	5	Yes	IncF plasmid	134.8	88	49.53	143
	6	Yes	IncI1 plasmid	109.7	78	50.95	124
	7	Yes	IncK plasmid	86.0	105	52.67	106
	8	Yes	IncI2 plasmid	56.9	79	42.27	77
	9	No	Phage	18.7	20	55.19	40
	10	No	?	1.5	6.7	32.27	0
	11	No	?	0.9	2.7	31.77	1
FAP1	1	No	Chromosome	4874.7	140	50.71	4855
	2	Yes	IncF plasmid	141.8	185	50.49	155
	3	No	IncI1 plasmid	129.4	219	50.07	143
	4	Yes	IncI2 plasmid	62.4	112	42.32	81
	5	Yes	plasmid	46.2	90	44.24	62

Discusión

A lo largo de esta tesis se plantearon tres objetivos: (1) desarrollar un método para la reconstrucción de plásmidos a partir de experimentos de secuenciación de genomas completos (Illumina); (2) estudiar el plasmidoma de cepas UPEC *E. coli* O25b:H4-B2 ST131 y (3) estudiar los mecanismos de diseminación de plásmidos resistentes a antibióticos en una colección de cepas obtenidas de distintos nichos. Para ellos se han analizado 41 cepas de *E. coli* y 186 plásmidos, descritos en dos estudios independientes.

Escherichia coli plasmidomics

Aún tenemos mucha incertidumbre acerca de la biología de los plásmidos y su interacción con una especie como *E. coli*. Sabemos que son el principal método de transferencia horizontal (Halary et al. 2010), que están fuertemente asociados con la patogenicidad (Johnson and Nolan 2009) y con la resistencia a antibióticos (De la Cruz and Davies 2000; Ochman et al. 2000; Thomas 2002; Fondi and Fani 2010; Smillie et al. 2010). Sin embargo conocemos muy poco de como es la biología de los plásmidos en comunidades. Desconocemos como es la distribución de los plásmidos dentro de una especie tan bien estudiada como *E. coli*. En esta tesis hemos caracterizado 186 plásmidos frente a los 306 plásmidos secuenciados completamente y que están asociados a *Escherichia coli* (GenBank Diciembre 2014). Esto indica que la diversidad de los plásmidos está muy subestimada en las bases de datos y son necesarios muchos más estudios que den la importancia necesaria a los plásmidos para conocer cual es la población plasmídica real de *Escherichia coli*.

Teniendo en cuenta la variabilidad de los plásmidos en un grupo clonal como ST131, la transferencia de los plásmidos tiene que ser muy activa. Basándonos en los dendrogramas de los artículos (de Been et al. 2014; Lanza et al. 2014) vemos que existen plásmidos muy parecidos a los

encontrados en *Eschericha coli* y que por lo tanto su transferencia entre especies es casi segura. Lo que no sabemos es como son las distribuciones poblacionales de los plásmidos en nichos complejos. Sería interesante conocer si la variabilidad plasmídica es superior a la cromosómica en un determinado nicho.

Los resultados que hemos observado indican que la transferencia de plásmidos supera a la velocidad de evolución de los cromosomas. Las escalas evolutivas de los plásmidos y de los cromosomas son distintas. En un conjunto clonal como puede ser ST131 (Lanza et al. 2014) la variabilidad de los plásmidos indica que la evolución del plasmidoma es más rápida que los cromosomas. Sin embargo en el artículo (de Been et al. 2014) la conservación de los plásmidos es mayor que el de los cromosomas. Si ponemos ambos datos en conjunto lo que nos damos cuenta es de que la variación de los plásmidos, o el flujo de plásmidos, es muy superior a las tasas de mutación de los cromosomas, mientras que determinados plásmidos que son fuertemente seleccionados (por ejemplo con antibióticos) mantienen una estructura más estable de la esperada.

PLACNET

PLACNET ha demostrado en *E. coli* que es un método capaz de reconstruir plásmidos con buena eficacia y sensibilidad. En ambos estudios publicados en esta tesis doctoral, PLACNET fue validado frente a una secuenciación con PacBio (de Been et al. 2014), comparativa con datos moleculares de laboratorio (S1-PFGE) y frente a la simulación de una secuenciación por Illumina de distintas cepas de referencia (Lanza et al. 2014). El método ha demostrado una alta eficacia con una media de error de 3,7% en ambos conjuntos de datos. Además, la mayor parte de las secuencias desechadas durante el proceso son repeticiones como secuencias de inserción (IS) o transposasas, elementos repetidos en la mayoría de los cromosomas que no deberían de suponer una gran diferencia al analizar las

cepas. La asignación de los genes de resistencias a antibióticos funcionó correctamente incluso con la peculiaridad de que suelen estar en elementos móviles que dificultan siempre el funcionamiento de PLACNET. Hay un amplio margen de mejora para el método, por una parte si el *pruning*[9] de la red. La automatización de este paso, de vital importancia para la correcta definición de los plásmidos, permitiría implementar el método para estudios a gran escala (centenas de cepas). Esta mejora no es inviable pero es compleja. Por una parte se necesita una métrica que indique cuándo la red está definiendo mejor los plásmidos, para posteriormente identificar aquellos nodos que presentan una mayor conectividad (gran número de *scaffold links)* y son de pequeño tamaño; lo que generalmente son elementos repetidos. Una vez identificados, su duplicación y correcta asignación a todas las posibles conexiones, podría representar la solución óptima, acorde a la métrica antes mencionada. Evidentemente, encontrar dicha métrica es la clave. El otro punto es encontrar un *algoritmo de clustering*[10] que identifique cada plásmido de forma automática. Esta parte es más sencilla, existen múltiples algoritmos de *clustering*, como Markov Cluster Algorithm (MCL), Unweighted Pair Group Method with Arithmetic Mean (UPGMA) o *K-means*. Una vez que tengamos una red óptima, solo hay que tratar de encontrar el algoritmo que mejor se adecúe a nuestro tipo de datos.

En cuanto a las fortalezas y debilidades del método, podemos concluir que PLACNET es especialmente eficiente en la detección de plásmidos epidémicos o repetidos dentro de un conjunto de cepas. El principal inconveniente es la resolución de dos plásmidos de la misma familia que cohabitan la misma cepa. Existen genes muy conservados dentro de algunos elementos, como algunas proteínas del canal de conjugación, que a efectos prácticos funcionan como una repetición cuando hay dos plásmidos presentes de la misma familia, debido a su alta similitud (>95%). Estos

9 Prunning: Es el proceso durante el cual modificamos la red, duplicando y reasignando las repeticiones para que la red represente los elementos genéticos de forma optima.

10 *Algoritmos de clustering:* Son algoritmos que tratan de encontrar conjuntos de datos similares o más conectados entre si, dentro de un grafo o un conjunto de datos relacionados.

elementos que deberían de estar contenidos en dos *contigs* independientes, pero en la realidad suelen converger en un contig quimera/híbrido. Esto genera redes más densas, donde varios contigs comparten muchas referencias entre sí por ser de la misma familia, y que además se encuentran conectados mediante *scaffolds* por los elementos comunes muy conservados. Biológicamente hay que evaluar la posibilidad que tienen estos plásmidos a formar co-integrados. Por ejemplo en la cepa FAP1 del estudio (de Been et al. 2014) nos encontramos que la secuenciación con PacBio muestra que en realidad existe un co-integrado de dos plásmidos IncF. En este punto, deberíamos considerar si un plásmido co-integrado es un estado transitorio de dos plásmidos, así como considerar su distribución poblacional, es decir, cuántas bacterias presentan el co-integrado frente a las que presentan dos plásmidos independientes.

En ocasiones no hemos sido capaces de asignar algunos *contigs* a ningún elemento definido en PLACNET. Existen tres posibles explicaciones:

1. Elementos que nunca han sido observados en los plásmidos presentes en esa cepa. Esta explicación, que si bien es plausible, tiene el inconveniente que deberían existir *scaffolds* con el resto de los contigs. Si bien no existirían las conexiones con los plásmidos de referencia, si deberían de existir las conexiones con los propios contigs vecinos en el plásmido.
2. Contaminación. Podría deberse a una contaminación durante el procesamiento de las muestras para la secuenciación. Es una teoría difícil de demostrar pero que justificaría por qué estos elementos suelen tener menor profundidad de secuenciación.
3. Interferencias poblacionales. Podría suceder que no todas las bacterias secuenciadas contengan los genes que conforman estos *contigs* (población mixta). Esto explicaría una profundidad de secuenciación menor. Durante la asignación de los *scaffold,* PLACNET filtra los enlaces menos abundantes para evitar falsos positivos, lo que puede provocar la eliminación de estos posibles *scaffolds* por ser elementos menos

abundantes que el cromosoma (el cromosoma por ser más abundante condiciona el umbral de filtrado para los *scaffolds*).

También podemos considerar la opción de varios sucesos al mismo tiempo, por ejemplo, elementos poco comunes que además no se encuentran en toda la población. Debemos de tener en cuenta que la secuenciación de una bacteria no deja de ser un estado instantáneo de la población (que debería de ser clonal) y por lo tanto pueden existir ciertas desviaciones. El cultivo de las cepas sin medios selectivos, una dilatada manipulación en el laboratorio, o problemas en la extracción del ADN pueden ocasionar problemas posteriores en la secuenciación.

Ahora nos queda demostrar la utilidad de PLACNET en otras especies que no sean *Escherichia coli*, como pueden ser *Klebsiella, Pseudomonas,* o *Enterococcus*. Además se podría pensar en un PLACNET adaptado a metagenomas para intentar estudiar el moviloma de muestras complejas. Más factible aún es el estudio de plásmidos en muestras multi-especie, por ejemplo aplicar PLACNET a un conjunto de *proteobacterias* que procedan de las mismas muestras y usando la capacidad de PLACNET, identificar plásmidos diseminados y estudiar cuales son los hospedadores naturales de los plásmidos y qué bacterias son capaces de transmitir información genética entre sí.

Diseminación de plásmidos resistentes a antibióticos.

Los resultados obtenidos en el estudio (de Been et al. 2014) dejan claro la diseminación de dos plásmidos portadores de resistencias a antibióticos (IncI1 e IncK-B/O) en la cadena alimentaria (animales, comida, humanos). Esto no quiere decir que exista una transmisión directa, ya que existen alternativas que explicarían la presencia de dichos plásmidos en los distintos nichos sin necesidad de que exista un contacto directo (Singer 2015). Aunque las interpretaciones tradicionales pueden indicar la dispersión de un clon resistente, debido a que eran muestras con idénticos MLST, la secuenciación

completa reveló la existencia de un número elevado de mutaciones superiores a las esperadas dentro de un mismo clon. Estos resultados demuestran que la clasificación por MLST no es suficientemente resolutiva como para monitorizar una expansión clonal, recomendando la secuenciación masiva para este tipo de estudios (Köser et al. 2012; Roetzer et al. 2013; den Bakker et al. 2014). Sin embargo, en el caso de los plásmidos, el nivel de conservación superior al observado en el cromosoma, lo que puede explicarse por una diseminación plasmídica. La diseminación ha podido ocurrir por un contacto directo entre las cepas, a través de intermediarios que sean capaces de colonizar los distintos nichos: pollos, cerdos, comida y humanos, o bien por recursos comunes como puede ser el agua (Singer 2015).

Poco sabemos acerca del reemplazamiento de una cepa por otra (Singer 2015), o de un plásmido por otro. Para que una bacteria, por ejemplo una cepa específica de *E. coli*, colonice el intestino debería de tener una ventaja evolutiva sobre las *E. coli* que ya se encuentran en el intestino o bien que entren de forma masiva y desplacen a las demás. Se sabe que en la flora intestinal existe un conjunto de cepas (normalmente una cepa dominante) que son permanentes, mientras que otras son transitorias (Tenaillon et al. 2010). Los estudios existentes a este respecto datan entre los años 50 y 80 (Sears et al. 1950; Sears and Brownlee 1952; Sears et al. 1956; Bettelheim et al. 1972; Cooke et al. 1972; Caugant et al. 1981). Es necesaria una actualización de estos estudios aplicando los conocimientos y las tecnologías actuales (Singer 2015). Si atendemos a los datos que muestran que las cepas son distintas pero los plásmidos son los mismos, una de las hipótesis plausibles sería que los plásmidos entren con las poblaciones transitorias y por conjugación pasen a las cepas permanentes, estabilizándose en estas poblaciones.

Si aceptamos que es una diseminación plasmídica y no una expansión clonal, entonces ¿tiene algo de especial este plásmido para estar diseminado?. Como se describe en de Been et al. (2014) el plásmido no contiene ningún

gen que no este en plásmidos cercanos, por lo cual no existe ningún elemento diferenciador, a parte de una variedad de la *shufflon protein* del sistema de pilina tipo IV (Ishiwa and Komano 2004), de la cual se desconoce si puede ser determinante para una diseminación exitosa. A pesar de no tener un gen "especial" que pueda explicar el éxito en la diseminación, lo que si que se observa es una combinación única de factores. Puede suceder que, al igual que en la virulencia, no se pueda explicar su éxito en la diseminación por una característica específica sino por una combinación de varias de ellas. Tampoco hay que olvidar que estas cepas están seleccionadas específicamente por ser resistentes, por lo que no sabemos cuál es el ratio real de ocupación de este plásmido dentro de la población. Puede que simplemente este plásmido se encuentre en cantidades minoritarias frente a otros plásmidos, pero se encuentre sobrerrepresentado en la colección seleccionada para este estudio. Hay que resaltar también que el gen $bla_{CTX\text{-}M\text{-}1}$ solo fue encontrado en los plásmidos IncI1, lo cual podría indicar que la tasa de conjugación es mayor que los eventos de recombinación. Esto expliacría porque no encontramos esta resistencia insertada en otros plásmidos. Es interesante el caso de la cepa 435 en donde se atisba un proceso de la pérdida de $bla_{CTX\text{-}M\text{-}1}$. Esto suscita la pregunta de cómo de estables son estos plásmidos en ausencia de presión selectiva y cuáles pueden ser los reservorios. Como comentamos en la **Introducción,** los plásmidos tienen sus mecanismos de permanencia, pero pueden perder sus elementos móviles como los transposones. Esto nos lleva a plantearnos cuál es la frecuencia de pérdida de estos elementos en ausencia de presión. Esto significaría que las poblaciones resistentes tienen una frecuencia umbral definida por el ratio entre la presión y la frecuencia de pérdida de los elementos de resistencia a antibióticos.

Por otra parte, es interesante el hecho de que ambos plásmidos (IncI1 e IncK-B/O) se encuentren presentes en cepas de todos los filogrupos. Esto indica que estos plásmidos pueden, potencialmente, estar en cualquier *E. coli* patógena. No existen estudios que describan la frecuencia de los plásmidos

en distintos nichos. Es decir, ¿cuántas de las proteobacterias presentes en el intestino comparten el mismo plásmido?, ¿cuál es la permanencia media de un plásmido en una cepa?, ¿cómo es la dinámica real de los plásmidos en las poblaciones? (Sørensen et al. 2005). La respuesta a éstas y otras preguntas nos ayudarían a entender la dinámica de los plásmidos y así poder diseñar nuevas estrategias para prevenir su difusión.

E. coli O25b:H4-B2 ST131 y su plasmidoma.

Poco o nada conocemos de la relación entre colonización y el clon de *E. coli* ST131. Numerosos estudios muestran ST131 como la secuencia tipo más prevalente dentro del total de *E. coli* productora de BLEE, con frecuencias aproximadas de 17%-19% (Rogers et al. 2011). Estos datos varían sensiblemente en función del tipo de estudio: si las muestras son sobre cepas BLEES, cepas productoras de $bla_{\text{CTX-M-15}}$, resistentes a fluoroquinolonas o simplemente causantes de infección urinaria (UPEC). Sería interesante conocer si existen diferencias significativas entre la colonización por ST131 y su incidencia en infecciones. Es decir, parece que el 20% de las infecciones urinarias (resistentes) son por ST131, pero no sabemos si el 20% de la población está colonizada mayoritariamente por ST131. Los pocos estudios que hay al respecto muestran que en realidad ST131 está en menor frecuencia en voluntarios sanos de lo que se podría esperar atendiendo a los datos de infecciones (Leflon-Guibout et al. 2008). Los estudios son escasos, por lo que las conclusiones nos son consistentes. Son necesarios más estudios que esclarezcan si ST131 se caracteriza por ser un gran colonizador o un gran patógeno. La división entre cepas ExPEC (como ST131) y cepas comensales es difusa. ExPEC muestra muchas características de las cepas comensales fuertemente adaptadas al intestino (Leimbach et al. 2013; Singer 2015). Diversos estudios sostienen que ST131 presenta un alto contenido de factores de virulencia (Rogers et al. 2011; Petty et al. 2014), lo que haría pensar que es un patógeno especialmente virulento. Sin embargo los estudios realizados con modelos animales no muestran una virulencia especialmente alta (Banerjee and Johnson 2014; Mora et al. 2014). Dado que factores de

virulencia se encuentran frecuentemente asociados a mayor capacidad colonizadora, podríamos estar ante un colonizador más eficiente debido a sus factores de virulencia (Singer 2015).

Lo que parece claro, de acuerdo a los estudios disponibles, es la fuerte asociación entre infecciones resistentes a antibióticos y el clon ST131 (Rogers et al. 2011). En opinión de este autor a este respecto (y en ausencia de datos reales), existe una cierta sobrevaloración de la incidencia de este clon en infecciones urinarias. La gran mayoría de los estudios se plantean con muestras recogidas en hospitales. Sin embargo la gran mayoría de las infecciones urinarias se tratan en atención primaria en donde, a falta de complicaciones, no se deriva a los hospitales para realizar un urocultivo. Esto significa que en realidad la gran mayoría de las infecciones urinarias son sensibles a tratamientos antibióticos y solo aquellas en las que ha fallado el tratamiento llegan a los servicios de microbiología de los hospitales. Los estudios basados en pacientes hospitalizados también tienen este contratiempo ya que estamos hablando de pacientes que generalmente se encuentran bajo tratamiento antibiótico o lo han estado recientemente, en estos casos en normal que las cepas que persisten sean resistentes ya que ha existido una presión selectiva que les ha favorecido claramente. Por estas razones sería necesario realizar un estudio de infecciones urinarias en pacientes de atención primaria, previo a cualquier tratamiento con antibiótico y tratar de asociar las cepas infectivas con la flora intestinal para tratar de establecer la asociación que existe entre virulencia y colonización (Singer 2015). Solo si entendemos la naturaleza de ST131, patógeno y/o colonizador, podremos tomar acciones eficaces para atajar el problema.

Tanto el estudio de (Blanco et al. 2013), como los estudios de (Petty et al. 2014) y (Johnson et al. 2013) observaron un aumento significativo de las cepas portadoras del alelo *fimH30*. Estas cepas tienen varias características que podrían explicar su expansión clonal. La primera es que tienen una fuerte asociación con resistencias a fluoroquinolonas debido a mutaciones en los genes *gyrA* y *parC*. Además, la mayoría tienen una inserción en el gen *fimB*

que podría causar una mayor expresión de la adhesina y aumentar su virulencia y colonización (Totsika et al. 2011). No queda claro si el alelo *fimH30* es fenotípicamente distinto, así que desconocemos la causa real de esta expansión clonal, además de desconocer si es una expansión en colonización, en infección (sensibles y resistentes) o ambos (Johnson et al. 2013). También en este punto hay que recurrir al sesgo que se puede haber cometido en la selección de las muestras. Faltan también estudios de genómica comparativa que determinen otros factores que puedan explicar la expansión clonal.

Tanto los estudios con PFGE, como con genomas completos, muestran claramente cuatro grupos dentro de las ST131 (Blanco et al. 2013; Petty et al. 2014) que concuerdan con los virotipos descritos en el artículo que forma parte de esta tesis. Los factores de virulencia que determinan estos virotipos son *ibeA, sat, iroN* y *afa/dr*. En un principio estos factores de virulencia se encuentran integrados en el cromosoma de ST131, a excepción de *iroN* que está en plásmidos. Los factores *sat* y *afa/Dr* han sido también localizados en plásmidos (Guignot et al. 2007), por lo que esta clasificación podría variar en un futuro. Si bien es cierto que en estos momentos existe una correlación entre el virotipo y la filogenia, en un futuro esta asociación podría divergir si algunos de estos factores se integra en un plásmido en vez de en el cromosoma. Dada la heterogeneidad y el flujo de plásmidos en ST131, la correlación entre virotipos y filogenia puede deberse a una rápida expansión que exceda la capacidad de mutación en la cepa, restringiendo su diversidad y manteniendo la asociación.

Los plásmidos encontrados en las cepas analizadas de ST131 de esta tesis han resultado ser plásmidos muy parecidos a otros plásmidos antes descritos. Plásmidos que han sido identificados en otros filogrupos y que por lo tanto no parece que sean un rasgo característico de ST131. A diferencia de los plásmidos del estudio (de Been et al. 2014) en donde encontramos plásmidos epidémicos altamente conservados, en este caso encontramos una alta heterogeneidad, tan alta que incluso no somos capaces de identificar una

sola proteína que sea común a todos ellos. Esto se debe principalmente a que algunos de ellos han perdido la región de conjugación parcialmente, que suele ser el modulo más conservado en un plásmido (Fernández-López et al. 2006; Smillie et al. 2010). Es interesante observar que aunque todos las cepas contienen un plásmidos del tipo MOB_{F12}/IncF existe una cierta desviación en cuanto al tipo de plásmido (Fig 3 (Lanza et al. 2014)). Por un lado los virotipos A y C contienen plásmidos muy parecidos a los plásmidos referencia pEK499, pEC_L8 y pEC_L46. Estos tres plásmidos pertenecen a cepas ST131 (Woodford et al. 2009; Smet et al. 2010), desconocemos el virotipo de la cepa original. El éxito aparente de este plásmido puede deberse a un rápida expansión clonal de la cepa y por lo tanto a un caso de deriva génica, o que realmente es la combinación del cromosoma junto con el plásmido lo genera un fenotipo exitoso (entre los grupos A y C suman el 54% de las muestras de (Blanco et al. 2013)). Sin embargo, los plásmidos de los virotipos B y D parecen pertenecer a otra subfamilia dentro de los MOB_{F12}. Los datos no son concluyentes ya que la muestra no es lo suficientemente amplia como para poder realizar asociaciones significativas, además de solo contener cepas resistentes a antibióticos. Desconocemos qué tipo de plásmidos podemos hallar en cepas sensibles y si estos plásmidos son derivados de plásmidos de resistencia, pero carentes de este cargo, o si por el contrario pertenecen a familias distintas. Esto sería realmente interesante ya que podríamos distinguir si los plásmidos únicamente se encuentran por sus resistencias o si por el contrario su valor virulento es determinante para el fenotipo y la diseminación de la cepa.

Conclusiones

1. Esta tesis aporta el desarrollo y validación del primer método bioinformático, PLACNET, que permite la caracterización completa del plasmidoma en genomas bacterianos.

2. PLACNET ha permitido discriminación entre una diseminación plasmídica y una bacteriana.
3. PLACNET ha sido validado frente a métodos avanzados de secuenciación de tercera generación y simulaciones de secuenciación sobre cepas conocidas, con tasas de error cercanas al 3,7%.
4. PLACNET muestra una gran capacidad para determinar plásmidos epidémicos.
5. La variabilidad del plasmidoma del clon *E. coli* ST131 es superior a la estimada a partir de la distancia evolutiva de sus cromosomas.
6. Los grupos mayoritarios de plásmidos reconstruidos por PLACNET en cepas de *E. coli* ST131 son MOB_{F12}/IncF, MOB_{P12}/IncI y MOB_{P51}/ColE1-like. Plásmidos crípticos de pequeño tamaño (MOB_{Q4}, 4 kb; y no-MOB/Rep_HTH27/36, 1.5 kb), desestimados por métodos tradicionales, han sido determinados eficazmente siguiendo PLACNET.
7. Los plásmidos de *E. coli* ST131 son derivados muy cercanos a otros plásmidos de *E. coli* encontrados en cepas de los filogrupos A, B1, B2 y D..
8. La transmisión de los genes $bla_{CTX\text{-}M1}$ y $bla_{CMY\text{-}2}$, observados en una colección de 32 muestras de *E. coli*, se debe a una diseminación plasmídica y no a una expansión clonal, tal y como se había determinado anteriormente a partir de los métodos clásicos de tipado.
9. Se ha observado que existen eventos de transmisión directa de cepas de *E. coli* entre el ganado porcino y el personal encargado de su estabulación.

Bibliografía

Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12:402.

Altschul S. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Alvarado A, Garcillán-Barcia MP, de la Cruz F. 2012. A degenerate primer MOB typing (DPMT) method to classify gamma-proteobacterial plasmids in clinical and environmental settings. PLoS One 7:e40438.

Angiuoli S V., Salzberg SL. 2011. Mugsy: Fast multiple alignment of closely related whole genomes. Bioinformatics 27:334–342.

Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25:1968–1969.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, et al. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Den Bakker HC, Allard MW, Bopp D, Brown EW, Fontana J, Iqbal Z, Kinney A, Limberger R, Musser KA, Shudt M, et al. 2014. Rapid whole-genome sequencing for surveillance of salmonella enterica serovar Enteritidis. Emerg. Infect. Dis. 20:1306–1314.

Banerjee R, Johnson JR. 2014. A new clone sweeps clean: the enigmatic emergence of Escherichia coli sequence type 131. Antimicrob. Agents Chemother. 58:4997–5004.

Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19:455–477.

Barrick JE, Colburn G, Deatherage DE, Traverse CC, Strand MD, Borges JJ, Knoester DB, Reba A, Meyer AG. 2014. Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. BMC Genomics 15:1039.

De Been M, Lanza VF, de Toro M, Scharringa J, Dohmen W, Du Y, Hu J, Lei Y, Li N, Tooming-Klunderud A, et al. 2014. Dissemination of Cephalosporin Resistance Genes

between Escherichia coli Strains from Farm Animals and Humans by Specific Plasmid Lineages. PLoS Genet. 10:e1004776.

Besemer J, Lomsadze a, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 29:2607–2618.

Bettelheim KA, Faiers M, Shooter RA. 1972. Serotypes of Escherichia coli in normal stools. Lancet 2:1223–1224.

Blanco J, Mora A, Mamani R, López C, Blanco M, Dahbi G, Herrera A, Marzoa J, Fernández V, de la Cruz F, et al. 2013. Four Main Virotypes among Extended-Spectrum-β-Lactamase-Producing Isolates of Escherichia coli O25b:H4-B2-ST131: Bacterial, Epidemiological, and Clinical Characteristics. J. Clin. Microbiol. 51:3358–3367.

Bohlin J, Brynildsrud OB, Sekse C, Snipen L. 2014. An evolutionary analysis of genome expansion and pathogenicity in Escherichia coli. BMC Genomics 15:882.

Bouet J-Y, Nordström K, Lane D. 2007. Plasmid partition and incompatibility--the focus shifts. Mol. Microbiol. 65:1405–1414.

Brolund A, Franzén O, Melefors O, Tegmark-Wisell K, Sandegren L. 2013. Plasmidome-analysis of ESBL-producing escherichia coli using conventional typing and high-throughput sequencing. PLoS One 8:e65793.

Bruand C, Ehrlich SD. 2000. UvrD-dependent replication of rolling-circle plasmids in Escherichia coli. Mol. Microbiol. 35:204–210.

Buchholz U, Bernard H, Werber D, Böhmer MM, Remschmidt C, Wilking H, Deleré Y, an der Heiden M, Adlhoch C, Dreesman J, et al. 2011. German outbreak of Escherichia coli O104:H4 associated with sprouts. N. Engl. J. Med. 365:1763–1770.

Burian J, Guller L, Macor M, Kay WW. 1997. Small cryptic plasmids of multiplasmid, clinical Escherichia coli. Plasmid 37:2–14.

Burland V, Shao Y, Perna NT, Plunkett G, Sofia HJ, Blattner FR. 1998. The complete DNA sequence and analysis of the large virulence plasmid of Escherichia coli O157:H7. Nucleic Acids Res. 26:4196–4204.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. 2008. ALLPATHS: De novo assembly of whole-genome shotgun microreads. Genome … 18:810–820.

Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL, Threlfall EJ. 2005. Identification of plasmids by PCR-based replicon typing. J. Microbiol. Methods 63:219–228.

Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. 2014. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob. Agents Chemother. 58:3895–3903.

Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, Parkhill J. 2005. ACT: the Artemis Comparison Tool. Bioinformatics 21:3422–3423.

Cascales E, Buchanan SK, Duché D, Kleanthous C, Lloubès R, Postle K, Riley M, Slatin S, Cavard D. 2007. Colicin biology. Microbiol. Mol. Biol. Rev. 71:158–229.

Caugant DA, Levin BR, Selander RK. 1981. Genetic diversity and temporal variation in the E. coli population of a human host. Genetics 98:467–490.

Cervantes-Rivera R, Pedraza-López F, Pérez-Segura G, Cevallos M a. 2011. The replication origin of a repABC plasmid. BMC Microbiol. 11:158.

Chan JZ-M, Pallen MJ, Oppenheim B, Constantinidou C. 2012. Genome sequencing in clinical microbiology. Nat. Biotechnol. 30:1068–1071.

Chaudhuri RR, Henderson IR. 2012. The evolution of the Escherichia coli phylogeny. Infect. Genet. Evol. 12:214–226.

Che D, Hockenbury C, Marmelstein R, Rasheed K. 2010. Classification of genomic islands using decision trees and their ensemble algorithms. BMC Genomics 11 Suppl 2:S1.

Chevreux B, Wetter T, Suhai S. 1999. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Ger. Conf. Bioinforma.:45–56.

Clermont O, Bonacorsi S, Bingen E. 2000. Rapid and simple determination of the Escherichia coli phylogenetic group. Appl. Environ. Microbiol. 66:4555–4558.

Clermont O, Christenson JK, Denamur E, Gordon DM. 2013. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ. Microbiol. Rep. 5:58–65.

Cooke EM, Hettiaratchy IG, Buck AC. 1972. Fate of ingested Escherichia coli in normal persons. J. Med. Microbiol. 5:361–369.

Coque TM, Novais A, Carattoli A, Poirel L, Pitout J, Peixe L, Baquero F, Cantón R, Nordmann P. 2008. Dissemination of clonally related Escherichia coli strains expressing extended-spectrum beta-lactamase CTX-M-15. Emerg. Infect. Dis. 14:195–200.

Croxen M a., Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. 2013. Recent advances in understanding enteric pathogenic Escherichia coli. Clin. Microbiol. Rev. 26:822–880.

Croxen MA, Finlay BB. 2010. Molecular mechanisms of Escherichia coli pathogenicity. Nat. Rev. Microbiol. 8:26–38.

Dahbi G, Mora A, Mamani R, López C, Alonso MP, Marzoa J, Blanco M, Herrera A, Viso S, García-Garrote F, et al. 2014. Molecular epidemiology and virulence of Escherichia coli O16:H5-ST131: Comparison with H30 and H30-Rx subclones of O25b:H4-ST131. Int. J. Med. Microbiol. 304:1247–1257.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394–1403.

Deatherage DE, Barrick JE. 2014. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods Mol. Biol. 1151:165–188.

Deatherage DE, Traverse CC, Wolf LN, Barrick JE. 2014. Detecting rare structural variation in evolving microbial populations from new sequence junctions using breseq. Front. Genet. 5:468.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 27:4636–4641.

Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. 2012. Transforming clinical microbiology with bacterial genome sequencing. Nat. Rev. Genet. 13:601–612.

Didelot X, Méric G, Falush D, Darling AE. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. BMC Genomics 13:256.

Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. Trends Genet.:1–9.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2007. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res. 17:1697–1706.

Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo AC, Dong X, Lu P, Szafron D, Greiner R, Wishart DS. 2005. BASys: A web server for automated bacterial genome annotation. Nucleic Acids Res. 33.

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. 2005. Diversity of the human intestinal microbial flora. Science 308:1635–1638.

Edwards DJ, Holt KE. 2013. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microb. Inform. Exp. 3:2.

Escobar-Páramo P, Clermont O, Blanc-Potard a. B, Bui H, Le Bouguénec C, Denamur E. 2004. A specific genetic background is required for acquisition and expression of virulence factors in Escherichia coli. Mol. Biol. Evol. 21:1085–1094.

Evans DG, Silver RP, Evans DJ, Chase DG, Gorbach SL. 1975. Plasmid controlled colonization factor associated with virulence in Escherichia coli enterotoxigenic for humans. Infect. Immun. 12:656–667.

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

Fernández-López R, Garcillán-Barcia MP, Revilla C, Lázaro M, Vielva L, de la Cruz F. 2006. Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution. FEMS Microbiol. … 30:942–966.

Field CM, Summers DK. 2011. Multicopy plasmid stability: Revisiting the dimer catastrophe. J. Theor. Biol. 291:119–127.

Fischbach M a, Lin H, Liu DR, Walsh CT. 2006. How pathogenic bacteria evade mammalian sabotage in the battle for iron. Nat. Chem. Biol. 2:132–138.

Fondi M, Fani R. 2010. The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks. Environ. Microbiol. 12:3228–3242.

Fouts DE. 2006. Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res. 34:5839–5851.

Foxman B. 2003. Epidemiology of urinary tract infections: incidence, morbidity, and economic costs. Dis. Mon. 49:53–70.

Foxman B. 2010. The epidemiology of urinary tract infection. Nat. Rev. Urol. 7:653–660.

Francia MV, Varsaki A, Garcillán-Barcia MP, Latorre A, Drainas C, de la Cruz F. 2004. A classification scheme for mobilization regions of bacterial plasmids. FEMS Microbiol. Rev. 28:79–100.

Frank C, Werber D, Cramer JP, Askar M, Faber M, an der Heiden M, Bernard H, Fruth A, Prager R, Spode A, et al. 2011. Epidemic profile of Shiga-toxin-producing Escherichia coli O104:H4 outbreak in Germany. N. Engl. J. Med. 365:1771–1780.

Freitas AR, Novais C, Tedim AP, Francia MV, Baquero F, Peixe L, Coque TM. 2013. Microevolutionary events involving narrow host plasmids influences local fixation of vancomycin-resistance in Enterococcus populations. PLoS One 8:e60589.

Frost LS, Leplae R, Summers AO, Toussaint A. 2005. Mobile genetic elements: the agents of open source evolution. Nat. Rev. Microbiol. 3:722–732.

García-Fernández A, Fortini D, Veldman K, Mevius D, Carattoli A. 2009. Characterization of plasmids harbouring qnrS1, qnrB2 and qnrB19 genes in Salmonella. J. Antimicrob. Chemother. 63:274–281.

Garcillán-Barcia MP, Francia MV, de la Cruz F. 2009. The diversity of conjugative relaxases and its application in plasmid classification. FEMS Microbiol. Rev. 33:657–687.

Gerdes K, Christensen SK, Løbner-Olesen A. 2005. Prokaryotic toxin-antitoxin stress response loci. Nat. Rev. Microbiol. 3:371–382.

Gil R, Silva FJ, Peretó J, Moya A. 2004. Determination of the core of a minimal bacterial gene set. Microbiol. Mol. Biol. Rev. 68:518–537.

Grad YH, Godfrey P, Cerquiera GC, Mariani-Kurkdjian P, Gouali M, Bingen E, Shea TP, Haas BJ, Griggs A, Young S, et al. 2013. Comparative Genomics of Recent Shiga Toxin-Producing Escherichia coli O104:H4: Short-Term Evolution of an Emerging Pathogen. MBio 4:e00452–12.

Gradel KO, Schønheyder HC, Arpi M, Knudsen JD, Ostergaard C, Søgaard M. 2014. The Danish Collaborative Bacteraemia Network (DACOBAN) database. Clin. Epidemiol. 6:301–308.

Guglielmini J, de la Cruz F, Rocha EPC. 2013. Evolution of Conjugation and Type IV Secretion Systems. Mol. Biol. Evol. 30:315–331.

Guglielmini J, Quintais L, Garcillán-Barcia MP, de la Cruz F, Rocha EPC. 2011. The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. PLoS Genet. 7:e1002222.

Guignot J, Chaplais C, Coconnier-Polter M-H, Servin AL. 2007. The secreted autotransporter toxin, Sat, functions as a virulence factor in Afa/Dr diffusely adhering Escherichia coli by promoting lesions in tight junction of polarized epithelial cells. Cell. Microbiol. 9:204–221.

Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. 2010. Network analyses structure genetic diversity in independent genetic worlds. Proc. Natl. Acad. Sci. U. S. A. 107:127–132.

Hamrick TS, Harris SL, Spears P a, Havell E a, Horton JR, Russell PW, Orndorff PE. 2000. Genetic characterization of Escherichia coli type 1 pilus adhesin mutants and identification of a novel binding phenotype. J. Bacteriol. 182:4012–4021.

Hauser M, Mayer CE, Söding J. 2013. kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics 14:248.

Hayes F, Van Melderen L. 2011. Toxins-antitoxins: diversity, evolution and function. Crit. Rev. Biochem. Mol. Biol. 46:386–408.

Hayes F. 2003. Toxins-antitoxins: plasmid maintenance, programmed cell death, and cell cycle arrest. Science 301:1496–1499.

Hemmerich C, Buechlein A, Podicheti R, Revanna K V., Dong Q. 2010. An Ergatis-based prokaryotic genome annotation web server. Bioinformatics 26:1122–1124.

Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J. 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res. 18:802–809.

Hsiao W, Wan I, Jones SJ, Brinkman FSL. 2003. IslandPath: aiding detection of genomic islands in prokaryotes. Bioinformatics 19:418–420.

Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves T a, Alkan C, Dennis MY, et al. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. 24:688–696.

Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE. 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. 6:90.

Ishiwa A, Komano T. 2004. PilV adhesins of plasmid R64 thin pili specifically bind to the lipopolysaccharides of recipient cells. J. Mol. Biol. 343:615–625.

Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangl JL, Jones CD. 2007. Extending assembly of short DNA sequences to handle error. Bioinformatics 23:2942–2944.

Jensen LB, Garcia-Migura L, Valenzuela a JS, Løhr M, Hasman H, Aarestrup FM. 2010. A classification system for plasmids from enterococci and other Gram-positive bacteria. J. Microbiol. Methods 80:25–43.

Johnson JR, Johnston B, Clabots C, Kuskowski M a, Castanheira M. 2010. Escherichia coli sequence type ST131 as the major cause of serious multidrug-resistant E. coli infections in the United States. Clin. Infect. Dis. 51:286–294.

Johnson JR, Sannes MR, Croy C, Johnston B, Clabots C, Kuskowski MA, Bender J, Smith KE, Winokur PL, Belongia EA. 2007. Antimicrobial drug-resistant Escherichia coli from humans and poultry products, Minnesota and Wisconsin, 2002-2004. Emerg. Infect. Dis. 13:838–846.

Johnson JR, Tchesnokova V, Johnston B, Clabots C, Roberts PL, Billig M, Riddell K, Rogers P, Qin X, Butler-Wu S, et al. 2013. Abrupt emergence of a single dominant multidrug-resistant strain of Escherichia coli. J. Infect. Dis. 207:919–928.

Johnson TJ, DebRoy C, Belton S, Williams ML, Lawrence M, Nolan LK, Thorsness JL. 2010. Pyrosequencing of the Vir plasmid of necrotoxigenic Escherichia coli. Vet. Microbiol. 144:100–109.

Johnson TJ, Nolan LK. 2009. Pathogenomics of the virulence plasmids of Escherichia coli. Microbiol. Mol. Biol. Rev. 73:750–774.

Jolley KA, Maiden MCJ. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11:595.

Kaper JB, Nataro JP, Mobley HL. 2004. Pathogenic Escherichia coli. Nat. Rev. Microbiol. 2:123–140.

Kent WJ. 2002. BLAT - The BLAST-like alignment tool. Genome Res. 12:656–664.

Kisand V, Lettieri T. 2013. Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. BMC Genomics 14:211.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller C a, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22:568–576.

Koren S, Harhay GP, Smith TPL, Bono JL, Harhay DM, Mcvey SD, Radune D, Bergman NH, Phillippy AM. 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol. 14:R101.

Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR, et al. 2012. Rapid Whole-Genome Sequencing for Investigation of a Neonatal MRSA Outbreak. N. Engl. J. Med. 366:2267–2275.

Krekeler N, Marenda MS, Browning GF, Holden KM, Charles J a, Wright PJ. 2012. Uropathogenic virulence factor FimH facilitates binding of uteropathogenic Escherichia coli to canine endometrium. Comp. Immunol. Microbiol. Infect. Dis. 35:461–467.

Kunne C, Billion A, Mshana SE, Schmiedel J, Domann E, Hossain H, Hain T, Imirzalioglu C, Chakraborty T. 2012. Complete Sequences of Plasmids from the Hemolytic-Uremic Syndrome-Associated Escherichia coli Strain HUSEC41. J. Bacteriol. 194:532–533.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12.

De la Cruz F, Davies J. 2000. Horizontal gene transfer and the origin of species: Lessons from bacteria. Trends Microbiol. 8:128–133.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359.

Lanza VF, de Toro M, Garcillán-Barcia MP, Mora A, Blanco J, Coque TM, de la Cruz F. 2014. Plasmid Flux in Escherichia coli ST131 Sublineages, Analyzed by Plasmid Constellation

Network (PLACNET), a New Method for Plasmid Reconstruction from Whole Genome Sequences. PLoS Genet. 10:e1004766.

Lau SH, Reddy S, Cheesbrough J, Bolton FJ, Willshaw G, Cheasty T, Fox AJ, Upton M. 2008. Major uropathogenic Escherichia coli strain isolated in the northwest of England identified by multilocus sequence typing. J. Clin. Microbiol. 46:1076–1080.

Laupland KB, Church L. 2014. Population-Based Epidemiology and Microbiology of Community- Onset Bloodstream Infections. Clin. Microbiol. Rev. 27:647–664.

Lee D, Seo H, Park C, Park K. 2009. WeGAS: a web-based microbial genome annotation system. Biosci. Biotechnol. Biochem. 73:213–216.

Leflon-Guibout V, Blanco J, Amaqdouf K, Mora A, Guize L, Nicolas-Chanoine MH. 2008. Absence of CTX-M enzymes but high prevalence of clones, including clone ST131, among fecal Escherichia coli isolates from healthy subjects living in the area of Paris, France. J. Clin. Microbiol. 46:3900–3905.

Leimbach A, Hacker J, Dobrindt U. 2013. E. coli as an all-rounder: The thin line between commensalism and pathogenicity. Curr. Top. Microbiol. Immunol. 358:3–32.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. 2009. SNP detection for massively parallel whole-genome resequencing. Genome Res. 19:1124–1132.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009. SOAP2: An improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

Lloyd AL, Henderson TA, Vigil PD, Mobley HLT. 2009. Genomic islands of uropathogenic Escherichia coli contribute to virulence. J. Bacteriol. 191:3469–3481.

Lohman TM, Ferrari ME. 1994. Escherichia coli single-stranded DNA-binding protein: multiple DNA-binding modes and cooperativities. Annu. Rev. Biochem. 63:527–570.

Lozano C, García-Migura L, Aspiroz C, Zarazaga M, Torres C, Aarestrup FM. 2012. Expansion of a plasmid classification system for Gram-positive bacteria and determination of the diversity of plasmids in Staphylococcus aureus strains of human, animal, and food origins. Appl. Environ. Microbiol. 78:5948–5955.

Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. 2011. Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. Proc. Natl. Acad. Sci. U. S. A. 108:7200–7205.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu YY, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18.

MacLean D, Jones JDG, Studholme DJ. 2009. Application of "next-generation" sequencing technologies to microbial genetics. Nat. Rev. Microbiol. 7:287–296.

Mahillon J, Chandler M. 1998. Insertion sequences. Microbiol. Mol. Biol. Rev. 62:725–774.

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, et al. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. U. S. A. 95:3140–3145.

Makino K, Ishii K, Yasunaga T, Hattori M, Yokoyama K, Yutsudo CH, Kubota Y, Yamaichi Y, Iida T, Yamamoto K, et al. 1998. Complete nucleotide sequences of 93-kb and 3.3-kb plasmids of an enterohemorrhagic Escherichia coli O157:H7 derived from Sakai outbreak. DNA Res. 5:1–9.

Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J. 2012. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res. 40:e6.

Mau B, Glasner JD, Darling AE, Perna NT. 2006. Genome-wide detection and analysis of homologous recombination among sequenced strains of Escherichia coli. Genome Biol. 7:R44.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

McNally A, Cheng L, Harris SR, Corander J. 2013. The evolutionary path to extraintestinal pathogenic, drug-resistant Escherichia coli is marked by drastic reduction in detectable recombination within the core genome. Genome Biol. Evol. 5:699–710.

Meyer R. 2009. The R1162 mob proteins can promote conjugative transfer from cryptic origins in the bacterial chromosome. J. Bacteriol. 191:1574–1580.

Meyer RR, Laine PS. 1990. The single-stranded DNA-binding protein of Escherichia coli. Microbiol. Rev. 54:342–380.

Mikheyev AS, Tin MMY. 2014. A first look at the Oxford Nanopore MinION sequencer. Mol. Ecol. Resour. 14:n/a – n/a.

Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17:589–596.

Mora A, Dahbi G, López C, Mamani R, Marzoa J, Dion S, Picard B, Blanco M, Alonso MP, Denamur E, et al. 2014. Virulence Patterns in a Murine Sepsis Model of ST131 Escherichia coli Clinical Isolates Belonging to Serotypes O25b:H4 and O16:H5 Are Associated to Specific Virotypes.Rasko DA, editor. PLoS One 9:e87025.

Narzisi G, Mishra B. 2011. Comparing de novo genome assembly: the long and short of it. PLoS One 6:e19175.

Nataro JP, Kaper JB, Robins-Browne R, Prado V, Vial P, Levine MM. 1987. Patterns of adherence of diarrheagenic Escherichia coli to HEp-2 cells. Pediatr. Infect. Dis. J. 6:829–831.

Nicolas-Chanoine M-H, Blanco J, Leflon-Guibout V, Demarty R, Alonso MP, Caniça MM, Park Y-J, Lavigne J-P, Pitout J, Johnson JR. 2008. Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. J. Antimicrob. Chemother. 61:273–281.

Novais A, Pires J, Ferreira H, Costa L, Montenegro C, Vuotto C, Donelli G, Coque TM, Peixe L, Novais Â. 2012. Characterization of globally spread Escherichia coli ST131 isolates (1991 to 2010). Antimicrob. Agents Chemother. 56:3973–3976.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.

Olsen RJ, Long SW, Musser JM. 2012. Bacterial genomics in infectious disease and the clinical pathology laboratory. Arch. Pathol. Lab. Med. 136:1414–1422.

Otto TD, Dillon GP, Degrave WS, Berriman M. 2011. RATT: Rapid Annotation Transfer Tool. Nucleic Acids Res. 39:1–7.

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, et al. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 42:D206–D214.

Pallen MJ, Wren BW. 2007. Bacterial pathogenomics. Nature 449:835–842.

Pareja-Tobes P, Manrique M, Pareja-Tobes E, Pareja E, Tobes R. 2012. BG7: A New Approach for Bacterial Genome Annotation Designed for Next Generation Sequencing Data. PLoS One 7.

Parsot C. 2005. Shigella spp. and enteroinvasive Escherichia coli pathogenicity factors. FEMS Microbiol. Lett. 252:11–18.

Peigne C, Bidet P, Mahjoub-Messai F, Plainvert C, Barbe V, Médigue C, Frapy E, Nassif X, Denamur E, Bingen E, et al. 2009. The plasmid of Escherichia coli strain S88 (O45:K1:H7) that causes neonatal meningitis is closely related to avian pathogenic E. coli plasmids and is associated with high-level bacteremia in a neonatal rat meningitis model. Infect. Immun. 77:2272–2284.

Petersen J. 2011. Phylogeny and compatibility: plasmid classification in the genomics era. Arch. Microbiol. 193:313–321.

Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, Forde BM, Phan M-D, Gomes Moriel D, Peters KM, Davies M, et al. 2014. Global dissemination of a multidrug resistant Escherichia coli clone. Proc. Natl. Acad. Sci. U. S. A. 111:5694–5699.

Pósfai G, Plunkett G, Fehér T, Frisch D, Keil GM, Umenhoffer K, Kolisnychenko V, Stahl B, Sharma SS, de Arruda M, et al. 2006. Emergent properties of reduced-genome Escherichia coli. Science 312:1044–1046.

Rahman A, Pachter L. 2013. CGAL: computing genome assembly likelihoods. Genome Biol. 14:R8.

Rankin DJ, Turner L a., Heinemann J a., Brown SP. 2012. The coevolution of toxin and antitoxin genes drives the dynamics of bacterial addiction complexes and intragenomic conflict. Proc. R. Soc. B Biol. Sci. 279:3706–3715.

Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin C-S, Iliopoulos D, et al. 2011. Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. N. Engl. J. Med. 365:709–717.

Remm M, Storm CE, Sonnhammer EL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314:1041–1052.

Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, et al. 2013. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. PLoS Med. 10.

Rogers B a, Sidjabat HE, Paterson DL. 2011. Escherichia coli O25b-ST131: a pandemic, multiresistant, community-associated strain. J. Antimicrob. Chemother. 66:1–14.

Rosvoll TCS, Pedersen T, Sletvold H, Johnsen PJ, Sollid JE, Simonsen GS, Jensen LB, Nielsen KM, Sundsfjord A. 2010. PCR-based plasmid typing in Enterococcus faecium strains reveals widely distributed pRE25-, pRUM-, pIP501- and pHTbeta-related replicons associated with glycopeptide resistance and stabilizing toxin-antitoxin systems. FEMS Immunol. Med. Microbiol. 58:254–268.

Sadowy E, Luczkiewicz A. 2014. Drug-resistant and hospital-associated Enterococcus faecium from wastewater, riverine estuary and anthropogenically impacted marine catchment basin. BMC Microbiol. 14:66.

Sallet E, Gouzy J, Schiex T. 2014. EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes. Bioinformatics 30:2659–2661.

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, et al. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 22:557–567.

Schembri M a, Kjaergaard K, Sokurenko E V, Klemm P. 2001. Molecular characterization of the Escherichia coli FimH adhesin. J. Infect. Dis. 183 Suppl :S28–S31.

Schubert S, Darlu P, Clermont O, Wieser A, Magistro G, Hoffmann C, Weinert K, Tenaillon O, Matic I, Denamur E. 2009. Role of intraspecies recombination in the spread of pathogenicity islands within the Escherichia coli species. PLoS Pathog. 5.

Schwartz DJ, Kalas V, Pinkner JS, Chen SL, Spaulding CN, Dodson KW, Hultgren SJ. 2013. Positively selected FimH residues enhance virulence during urinary tract infection by altering FimH conformation. Proc. Natl. Acad. Sci. U. S. A. 110:15530–15537.

Schweiger MR, Kerick M, Timmermann B, Isau M. 2011. The power of NGS technologies to delineate the genome organization in cancer: From mutations to structural variations and epigenetic alterations. Cancer Metastasis Rev. 30:199–210.

Sears HJ, Brownlee I, Uchiyame JK. 1950. Persistence of individual strains of Escherichia coli in the intestinal tract of man. J. Bacteriol. 59:293–301.

Sears HJ, Brownlee I. 1952. Further observations on the persistence of individual strains of Escherichia coli in the intestinal tract of man. J. Bacteriol. 63:47–57.

Sears HJ, Janes H, Saloum R, Brownlee I, Lamoreaux LF. 1956. Persistence of individual strains of Escherichia coli in man and dog under varying conditions. J. Bacteriol. 71:370–372.

Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

Shrivastava S, Reddy CVSK, Mande SS. 2010. INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms. J. Biosci. 35:351–364.

Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34:D32–D36.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–1123.

Singer RS. 2015. Urinary tract infections attributed to diverse ExPEC strains in food animals: evidence and data gaps. Front. Microbiol. 6:1–9.

Sinha S, Redfield RJ. 2012. Natural DNA uptake by Escherichia coli. PLoS One 7:e35620.

Smet A, Van Nieuwerburgh F, Vandekerckhove TTM, Martel A, Deforce D, Butaye P, Haesebrouck F. 2010. Complete nucleotide sequence of CTX-M-15-plasmids from clinical Escherichia coli isolates: insertional events of transposons and insertion sequences. PLoS One 5:e11202.

Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. 2010. Mobility of plasmids. Microbiol. Mol. Biol. Rev. 74:434–452.

Snel B, Bork P, Huynen MA. 1999. Genome phylogeny based on gene content. Nat. Genet. 21:108–110.

Snipen L, Ussery DW. 2010. Standard operating procedure for computing pangenome trees. Stand. Genomic Sci. 2:135–141.

Sørensen SJ, Bailey M, Hansen LH, Kroer N, Wuertz S. 2005. Studying plasmid horizontal transfer in situ: a critical review. Nat. Rev. Microbiol. 3:700–710.

Sorsa LJ, Dufke S, Heesemann J, Schubert S. 2003. Characterization of an iroBCDEN gene cluster on a transmissible plasmid of uropathogenic Escherichia coli: evidence for horizontal transfer of a chromosomal virulence factor. Infect. Immun. 71:3285–3293.

Stewart AC, Osborne B, Read TD. 2009. DIYA: A bacterial annotation pipeline for any genomics lab. Bioinformatics 25:962–963.

Sullivan MJ, Petty NK, Beatson S a. 2011. Easyfig: a genome comparison visualizer. Bioinformatics 27:1009–1010.

Summers DK, Beton CW, Withers HL. 1993. Multicopy plasmid instability: the dimer catastrophe hypothesis. Mol. Microbiol. 8:1031–1038.

Tap J, Mondot S, Levenez F, Pelletier E, Caron C, Furet J-P, Ugarte E, Muñoz-Tamayo R, Paslier DLE, Nalin R, et al. 2009. Towards the human intestinal microbiota phylogenetic core. Environ. Microbiol. 11:2574–2584.

Tekaia F, Yeramian E. 2005. Genome trees from conservation profiles. PLoS Comput. Biol. 1:e75.

Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal Escherichia coli. Nat. Rev. Microbiol. 8:207–217.

Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat. Rev. Microbiol. 3:711–721.

Thomas CM. 2002. Paradigms of plasmid organization. Mol. Microbiol. 37:485–491.

Tobe T, Hayashi T, Han CG, Schoolnik GK, Ohtsubo E, Sasakawa C. 1999. Complete DNA sequence and structural analysis of the enteropathogenic Escherichia coli adherence factor plasmid. Infect. Immun. 67:5455–5462.

Totsika M, Beatson SA, Sarkar S, Phan M-D, Petty NK, Bachmann N, Szubert M, Sidjabat HE, Paterson DL, Upton M, et al. 2011. Insights into a multidrug resistant Escherichia coli pathogen of the globally disseminated ST131 lineage: genome analysis and virulence mechanisms. PLoS One 6:e26578.

Touchon M, Rocha EPC. 2007. Causes of insertion sequences abundance in prokaryotic genomes. Mol. Biol. Evol. 24:969–981.

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13:36–46.

Trevisi E, Zecconi A, Cogrossi S, Razzuoli E, Grossi P, Amadori M. 2014. Strategies for reduced antibiotic usage in dairy cattle farms. Res. Vet. Sci. 96:229–233.

Tu Q, Ding D. 2003. Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. FEMS Microbiol. Lett. 221:269–275.

Venturini C, Beatson S a, Djordjevic SP, Walker MJ. 2010. Multiple antibiotic resistance gene recruitment onto the enterohemorrhagic Escherichia coli virulence plasmid. FASEB J. 24:1160–1166.

Vernikos GS, Parkhill J. 2006. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics 22:2196–2203.

Vila J, Ruiz J, Marco F, Barcelo A, Goñi P, Giralt E, Jimenez de Anta T. 1994. Association between double mutation in gyrA gene of ciprofloxacin-resistant clinical isolates of Escherichia coli and MICs. Antimicrob. Agents Chemother. 38:2477–2479.

Villa L, García-Fernández A, Fortini D, Carattoli A. 2010. Replicon sequence typing of IncF plasmids carrying virulence and resistance determinants. J. Antimicrob. Chemother. 65:2518–2529.

Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 3:199–208.

Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, Surovcik K, Meinicke P, Merkl R. 2006. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics 7:142.

Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos G a, Tiedje JM, Whittam TS. 2009. Cryptic lineages of the genus Escherichia. Appl. Environ. Microbiol. 75:6534–6544.

Wardal E, Gawryszewska I, Hryniewicz W, Sadowy E. 2013. Abundance and diversity of plasmid-associated genes among clinical isolates of Enterococcus faecalis. Plasmid 70:329–342.

Wardal E, Markowska K. 2014. Molecular Analysis of VanA Outbreak of Enterococcus faecium in Two Warsaw Hospitals: The Importance of Mobile Genetic Elements. BioMed Res. ….

Warren RL, Sutton GG, Jones SJM, Holt RA. 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23:500–501.

Waters LS, Storz G. 2009. Regulatory RNAs in bacteria. Cell 136:615–628.

Weissman SJ, Beskhlebnaya V, Chesnokova V, Chattopadhyay S, Stamm WE, Hooton TM, Sokurenko E V. 2007. Differential stability and trade-off effects of pathoadaptive mutations in the Escherichia coli FimH adhesin. Infect. Immun. 75:3548–3555.

Wiles TJ, Kulesus RR, Mulvey M a. 2008. Origins and virulence mechanisms of uropathogenic Escherichia coli. Exp. Mol. Pathol. 85:11–19.

Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, et al. 2006. Sex and virulence in Escherichia coli: an evolutionary perspective. Mol. Microbiol. 60:1136–1151.

Woodford N, Carattoli A, Karisik E, Underwood A, Ellington MJ, Livermore DM. 2009. Complete nucleotide sequences of plasmids pEK204, pEK499, and pEK516, encoding CTX-M enzymes in three major Escherichia coli lineages from the United Kingdom, all

belonging to the international O25:H4-ST131 clone. Antimicrob. Agents Chemother. 53:4472–4482.

Wu J, Mao X, Cai T, Luo J, Wei L. 2006. KOBAS server: A web-based platform for automated annotation and pathway identification. Nucleic Acids Res. 34.

Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L. 2011. KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 39.

Yang X, Chockalingam SP, Aluru S. 2013. A survey of error-correction methods for next-generation sequencing. Brief. Bioinform. 14:56–66.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25:2865–2871.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.

Zhang LY, Chang SH, Wang J. 2010. How to make a minimal genome for synthetic minimal cell. Protein Cell 1:427–434.

Zhou Y, Call DR, Broschat SL. 2013. Using Protein Clusters from Whole Proteomes to Construct and Augment a Dendrogram. Adv. Bioinformatics 2013:1–8.

Zhou Y, Landweber LF. 2007. BLASTO: A tool for searching orthologous groups. Nucleic Acids Res. 35:678–682.

Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: A Fast Phage Search Tool. Nucleic Acids Res. 39:1–6.

Zong Z. 2013. Complete sequence of pJIE186-2, a plasmid carrying multiple virulence factors from a sequence type 131 Escherichia coli O25 strain. Antimicrob. Agents Chemother. 57:597–600.

Otras Publicaciones

PipX, the coactivator of NtcA, is a global regulator in cyanobacteria

Javier Espinosa[a], Francisco Rodríguez-Mateos[b], Paloma Salinas[a], Val F. Lanza[c], Ray Dixon[d], Fernando de la Cruz[c], and Asuncion Contreras[a,1]

[a]División de Genética and [b]Departamento de Matemática Aplicada, Universidad de Alicante, 03080 Alicante, Spain; [c]Departamento de Biología Molecular, Facultad de Medicina, Universidad de Cantabria and Instituto de Biomedicina y Biotecnología de Cantabria, 39011 Santander, Spain; and [d]Department of Molecular Microbiology, John Innes Centre, Norwich NR4 7UH, United Kingdom

Edited by Robert Haselkorn, University of Chicago, Chicago, IL, and approved May 6, 2014 (received for review March 4, 2014)

To modulate the expression of genes involved in nitrogen assimilation, the cyanobacterial P_{II}-interacting protein X (PipX) interacts with the global transcriptional regulator NtcA and the signal transduction protein P_{II}, a protein found in all three domains of life as an integrator of signals of the nitrogen and carbon balance. PipX can form alternate complexes with NtcA and P_{II}, and these interactions are stimulated and inhibited, respectively, by 2-oxoglutarate, providing a mechanistic link between P_{II} signaling and NtcA-regulated gene expression. Here, we demonstrate that PipX is involved in a much wider interaction network. The effect of *pipX* alleles on transcript levels was studied by RNA sequencing of *S. elongatus* strains grown in the presence of either nitrate or ammonium, followed by multivariate analyses of relevant mutant/control comparisons. As a result of this process, 222 genes were classified into six coherent groups of differentially regulated genes, two of which, containing either NtcA-activated or NtcA-repressed genes, provided further insights into the function of NtcA–PipX complexes. The remaining four groups suggest the involvement of PipX in at least three NtcA-independent regulatory pathways. Our results pave the way to uncover new regulatory interactions and mechanisms in the control of gene expression in cyanobacteria.

nitrogen regulation | transcription | translation | photosynthesis

Cyanobacteria are phototrophic organisms that perform oxygenic photosynthesis. Autotrophic growth requires the constant assimilation of ammonium via the glutamine synthetase–glutamate synthase cycle, resulting in consumption of the 2-oxoglutarate (2-OG) (1, 2) that accumulates during nitrogen starvation, making this metabolite an excellent indicator of the intracellular carbon-to-nitrogen balance (3, 4). The 2-OG, the signal of nitrogen deficiency, modulates the activity and/or binding properties of three key cyanobacterial nitrogen regulators: the signal transduction protein P_{II}; the transcriptional activator NtcA; and PipX, a regulatory factor that can interact with either NtcA or P_{II}.

The homotrimeric P_{II} protein, one of the most conserved and widespread signal transduction proteins in nature, plays key roles in nitrogen assimilatory processes (5). P_{II} contains three binding sites (one per subunit) for 2-OG and ATP (6, 7), and it regulates the activity of *N*-acetyl-glutamate-kinase (NAGK), a key enzyme for biosynthesis of arginine, by direct protein–protein interactions (3, 8, 9). The 2-OG stimulates binding of NtcA to target sites (10), transcription activation in vitro (11), and complex formation between NtcA and PipX (12). The interaction between PipX and NtcA is known to be relevant under nitrogen limitation for activation of NtcA-dependent genes in *Synechococcus elongatus* and *Anabaena* sp. PCC 7120 (hereafter, *Anabaena*) (12–14). The NtcA–PipX complex consists of one active (2-OG–bound) NtcA dimer and two PipX molecules. Each NtcA subunit binds one PipX molecule in such a way that it stabilizes the active NtcA conformation and probably helps recruit RNA polymerase without providing extra DNA contacts (15, 16). The tudor-like domain of PipX provides the contacts for both NtcA–PipX and P_{II}–PipX interactions. When nitrogen is abundant, intracellular levels of 2-OG are low and sequestration of PipX by P_{II} decreases NtcA–PipX complex formation. A summary of the interactions involving NtcA, PipX, or P_{II} is shown in Fig. 1.

The PipX partner-swapping model predicts that, at least under the physiological range of 2-OG levels, *pipX* mutations specifically impairing PipX–P_{II} complexes would favor formation of NtcA–PipX complexes. Crystal structures of PipX–P_{II} complexes, surface plasmon resonance, P_{II}-stimulated NAGK activity assays, and yeast two-hybrid analysis established the importance of PipX residues Y32 and E4 for interactions with P_{II} proteins and of Y32 for interactions with NtcA (15, 17). Reporter and transcript analyses indicated that both Y32A and E4A mutations had stimulatory effects on the NtcA-activated genes *glnB*, *glnN*, and *nblA* but did not address differences between the in vivo action of $PipX^{E4A}$ and $PipX^{Y32A}$, two proteins with different biochemical properties. Here, we show that the in vivo properties of $PipX^{E4A}$ and $PipX^{Y32A}$ are indeed very different and that these differences affect both NtcA-dependent and NtcA-independent genes.

The 2-OG–dependent partner swapping of PipX between P_{II} and NtcA provides a mechanistic link between P_{II} signaling and NtcA-regulated gene expression but does not exclude the possibility that PipX, either by itself or bound to P_{II}, could participate in additional protein–protein interactions influencing gene expression. To address the question of whether PipX affects *S. elongatus* gene expression in an NtcA-independent manner, we compared transcript profiles of *pipX* mutants in cultures grown with either ammonium or nitrate. In these conditions, and to a

Significance

P_{II}, a signal transduction protein involved in nitrogen control in bacteria and plants, and NtcA, the transcriptional nitrogen regulator of cyanobacteria, can form complexes with P_{II} interacting protein X (PipX). We demonstrate by a combination of genetic, transcriptomic, and multivariate analyses that PipX is involved in a much wider interaction network affecting nitrogen assimilation, translation, and photosynthesis. Two groups of genes differentially regulated by *pipX* provided further insights into the function of NtcA–PipX complexes and an improved definition of the consensus NtcA binding motif. The other four groups suggested the involvement of PipX in NtcA-independent regulatory pathways. Our results pave the way to uncover new regulatory interactions and mechanisms in the control of gene expression in cyanobacteria.

Author contributions: J.E., F.d.l.C., and A.C. designed research; J.E., P.S., F.d.l.C., and A.C. performed research; F.R.-M. contributed new reagents/analytic tools; J.E., F.R.-M., P.S., V.F.L., R.D., F.d.l.C., and A.C. analyzed data; and J.E., F.R.-M., R.D., and A.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: contrera@ua.es.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1404097111/-/DCSupplemental.

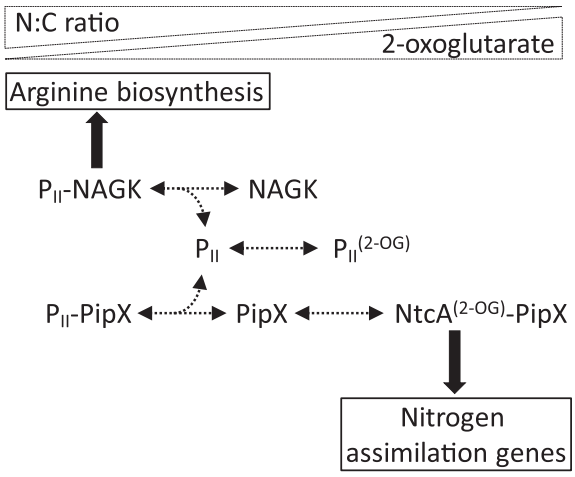


Fig. 1. PipX swapping model and the nitrogen interaction network. The functions and interactions mediated by PipX according to the 2-OG levels are schematized.

greater extent in ammonium, the levels of 2-OG are relatively low, NtcA is mainly inactive, and null *pipX* mutations are not expected to have a significant impact on the NtcA regulon (13). We reasoned that the $pipX^{E4A}$ and $pipX^{Y32A}$ mutations provide a means to bypass the need to use *ntcA* mutants and/or conditions of nitrogen deprivation to identify NtcA target genes. Most relevant to this work, the combined analyses of *pipX* mutant strains enabled us to identify additional regulatory targets of PipX in a context in which we could further characterize the NtcA regulon.

Results and Discussion

Global Phenotypic Impact of *pipX* Mutations Under Nitrogen-Rich Conditions. The involvement of PipX in gene expression and its potential roles in global regulation other than nitrogen control by NtcA were investigated in *S. elongatus* strains grown in either nitrate or ammonium and carrying either a null (Δ*pipX*) or point mutation ($pipX^{Y32A}$ or $pipX^{E4A}$) derivative. Under these nitrogen regimes, particularly in ammonium, the concentration of 2-OG is expected to be relatively low and transcriptional regulation by NtcA–PipX complexes would not have widespread importance in WT cells. Because the $pipX^{Y32A}$ or $pipX^{E4A}$ allele is associated with a resistance marker (C.S3 cassette) that decreases *pipX* gene expression (18), a WT derivative with the same insertion in the identical position (strain CS3X) was required to perform isogenic mutant/control comparisons. Therefore, two strains were used as WT controls: WT *S. elongatus* for mutant strain SA591, which carries the Δ*pipX* allele, and CS3X for strains $CS3X^{E4A}$ and $CS3X^{Y32A}$, which carry the $pipX^{Y32A}$ and $pipX^{E4A}$ alleles, respectively (*SI Appendix*, Table S1). A total of six global transcriptome comparisons were carried out, considering three mutant/control pairs for each of the two nitrogen regimes: ammonium and nitrate.

Scatter plots of $\log_2$ fold changes of *pipX* mutants vs. their respective controls are represented in Fig. 2. For each mutant/control comparison, subsets of genes that are up-regulated or down-regulated more than fourfold in the mutants only in nitrate, only in ammonium, or in both conditions are highlighted. Interestingly, many genes were differentially regulated in the Δ*pipX* mutant, and for each condition, there were more genes up-regulated than down-regulated in the absence of an active *pipX* allele, suggesting that PipX participates more frequently in negative regulation than in positive regulation under the experimental conditions tested (Fig. 2*A*). Whereas the global impact of $pipX^{Y32A}$ on gene expression seemed roughly similar to the effect of the Δ*pipX* allele, the effect of $pipX^{E4A}$ was restricted to a smaller number of genes and, importantly, most of those genes were up-regulated in both nitrate and ammonium (Fig. 2 *B* and *C*).

The results supported a role for PipX in both negative and positive regulation of multiple target genes in *S. elongatus* under conditions of nitrogen sufficiency and demonstrate that the mutations $pipX^{Y32A}$ and $pipX^{E4A}$ have very different effects on gene expression. To gain deeper insights into the functions of PipX in vivo while adding robustness to the analysis, we next analyzed combined information from all 10 transcript datasets.

Multivariate Analysis of *S. elongatus* Transcripts. To identify groups of genes with discrete expression patterns defined by the Δ*pipX*, $pipX^{E4A}$, or $pipX^{Y32A}$ alleles, we performed multivariate analysis with standardized residuals from linear regressions of data (log-transformed) from mutant vs. control strains (both $CS3X^{E4A}$ and $CS3X^{Y32A}$ vs. CS3X and SA591 vs. WT) cultured in the presence of either ammonium or nitrate. First, we demonstrated that only a small proportion of the transcriptome responded to the nitrogen source and/or to *pipX* alleles. A total of 1,663 genes with

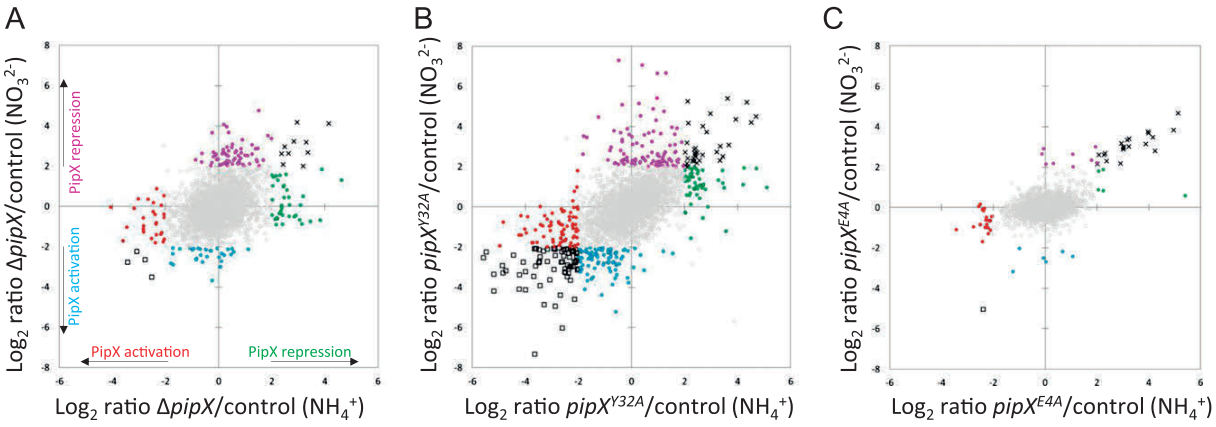


Fig. 2. Effect of *pipX* mutations on nitrate and ammonium transcriptomes. The mutant/control $\log_2$ ratios in nitrate vs. ammonium are represented as a scatter plot. (*A*) Δ*pipX*/WT with indication of the inferred PipX functions. (*B*) $CS3X^{Y32A}$/CS3X. (*C*) $CS3X^{E4A}$/CS3X. Genes shown in colors and in gray represent above and below, respectively, the cutoff for a $\log_2$ ratio >2 (absolute values). Positive values in nitrate (purple), ammonium (green), or both conditions (x) and negative values in nitrate (blue), ammonium (red), or both conditions (□) are shown.

residuals lower than 1.5 were considered as nonresponsive in any of the six mutant/control comparisons, and the distributions of their residuals were fitted well by truncated normal distributions (gray dots in *SI Appendix*, Fig. S1 and Table S3). For the ammonium and nitrate conditions, 1,958 and 2,052 genes, respectively, were nonresponsive in any of the three mutant/control comparisons. From the group of genes with residuals greater than 1.5, only those with residuals exceeding a threshold value of 2.5 for at least one of the six variables were selected as being differentially regulated. The resulting 282 genes, after excluding *pipX* itself, were individually analyzed to discard those with reads mapping mainly to the noncoding strand (Dataset S1), bringing the number down to 257 genes.

To explore the existence of different expression profiles within the 257 differentially regulated genes further, principal component analysis (PCA) of residuals was used to extract the first two principal components, accounting for about 70% of the total variance. The plot of the data for these two components (Fig. 3*A*) suggested the classification of the 257 genes into four main groups (classes 1, 2, 3, and 4), defined by using *k*-means cluster analysis. As shown in Fig. 3*A* (*Inset*), changes in expression for Δ*pipX* in both ammonium and nitrate and for $pipX^{Y32A}$ mainly in ammonium were associated with the first axis (PC1). In contrast, changes for $pipX^{E4A}$ in both ammonium and nitrate and, to a lesser extent, for $pipX^{Y32A}$ in nitrate, were associated with the second component (PC2).

The relatively large sizes of the groups and the wide distribution of the dataset, particularly in classes 2 and 3, prompted us to use additional classification criteria to obtain better-defined groups with smaller numbers of similarly regulated genes. This was done by defining an additional four groups according to independent clustering by hierarchical Ward's minimum variance and fuzzy c-means methods. Only those genes that were coherently grouped into the same class with the three clustering methods were retained. The cluster dendrogram from Ward's method was then cut to produce six coherent groups comprising a total of 222 genes.

As a result of this classification process, the four original groups became six groups. The 35 genes that fell outside these groups (Fig. 3*A*, gray dots, and Dataset S2) included the genes with the lowest expression levels (among the 257 differentially regulated genes), as well as genes with rather unique expression patterns. Only class 4 (43 genes) was left untouched, although class 1 lost some members (32 genes remained) and classes 2 and 3 (losing 10 and 17 genes, respectively) were each split into two new classes: 2.1 (37 genes), 2.2 (43 genes), 3.1 (18 genes), and 3.2 (49 genes) (Datasets S3–S8). Differences between the new classes originating from a common class, that is, 2.1 vs. 2.2 and 3.1 vs. 3.2, were largely related to the third principal component (Fig. 3*B*, PC3 axis), which accounted for an additional 13% of the total variance and was mainly correlated with differences in expression provided by Δ*pipX* and $pipX^{Y32A}$ alleles, with the ammonium conditions providing the greatest differences in these mutant control comparisons. Hereafter, each of the six classes was treated and analyzed as a distinct class.

NtcA-Activated Genes and NtcA Binding Motifs in *S. elongatus*. The distinctive feature of class 4 transcripts was their up-regulation in $CS3X^{E4A}$ in both nitrate and ammonium and in $CS3X^{Y32A}$ only in nitrate (Figs. 3*A* and 4*A*). Notably, a rather small, but still significant, down-regulation in the Δ*pipX* strain was observed in nitrate. Class 4 comprised most of the paradigmatic nitrogen assimilation genes or operons from *S. elongatus*, that is, gene targets known or predicted to be activated by NtcA–PipX complexes under conditions of nitrogen deficiency (*SI Appendix*, Table S2). They include *ntcA* itself, genes encoding the key glutamine synthetase enzyme *glnA*, ammonium transporters *amtB* and *amt1*, components of the nitrate transport and assimilation system *nirA*/*nrtABCD*/*narB*, and components of the cyanate transport and assimilation system *cynABD* and *cynS*. The transcriptional response in all five genetic backgrounds and two nitrogen regimes is illustrated in Fig. 4*B* for two paradigmatic NtcA target genes: *ntcA*, which is autogenously regulated, and *nirA*, which encodes nitrite reductase. Both genes were similarly affected by the mutant alleles Δ*pipX*, $pipX^{Y32A}$, and $pipX^{E4A}$ under each of the nitrogen regimes tested.

Class 4 expression patterns suggest that WT PipX can weakly activate some NtcA target genes in nitrate but not in ammonium-grown cells. In contrast, $PipX^{Y32A}$ and, to a greater extent, $PipX^{E4A}$ appear to be competent to interact with NtcA and coactivate target promoters at low concentrations of 2-OG. Our

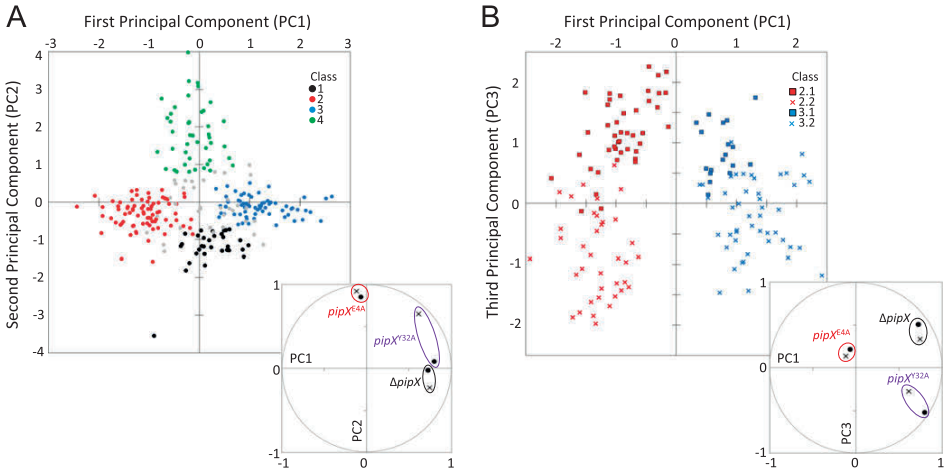


Fig. 3. Multivariate analysis and clustering of differentially expressed genes. (*A*) Scores for the two first principal components in the PCA of standardized residuals from mutant/control comparisons for the 257 genes with residuals larger than 2.5 in at least one comparison. Classified and nonclassified genes are colored and gray, respectively. (*Inset*) Scatter of mutant/control comparisons plotted as the correlation coefficients between them and the first two principal components in the unit circle. Nitrate (x) and ammonium (●) are shown. (*B*) Same as in *A*, but the first and third principal components in the PCA are represented only for genes classified originally in classes 2 and 3. Different symbols and colors identify genes from the final classes (2.1, 2.2, 3.1, and 3.2). (*Inset*) Same as in *A*, but PC2 is replaced by PC3.

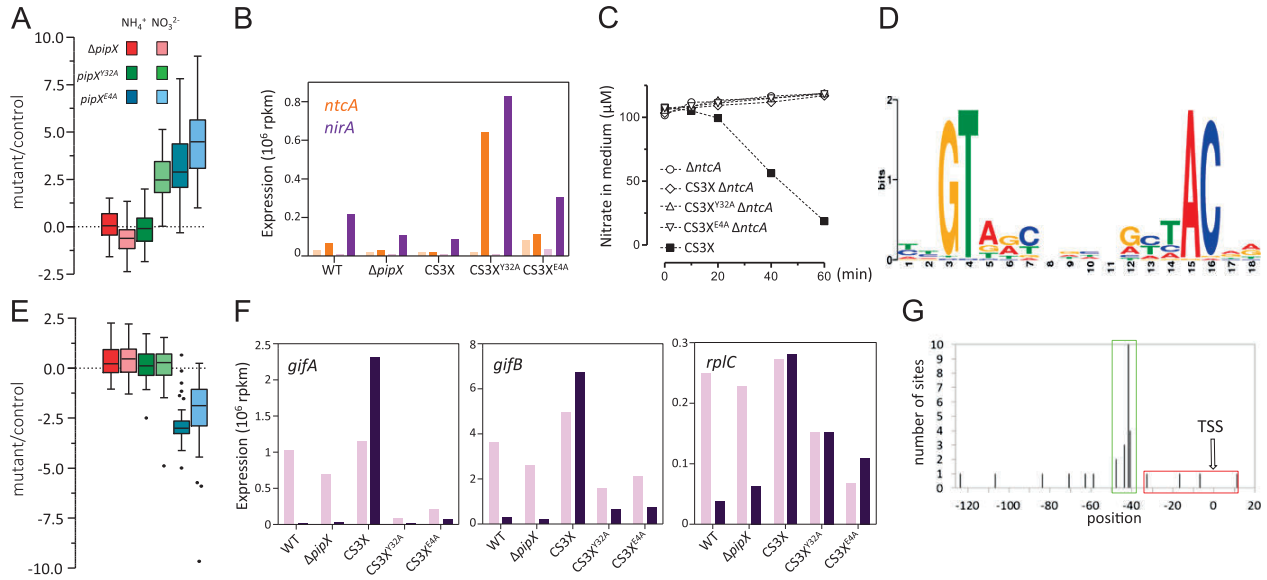


Fig. 4. Gain-of-function mutations $pipX^{E4A}$ and $pipX^{Y32A}$ target NtcA-regulated genes. (*A*) Box plot representation (median, box from first to third quartile and Tukey whiskers) of the within-group variance of mutant/control comparisons in class 4. (*B*) Expression pattern of the *ntcA* and *nirA* genes in ammonium (light-colored bars) and nitrate (dark-colored bars). (*C*) Effect of *ntcA* inactivation on nitrate uptake. Nitrate was added at time 0 to cell suspensions containing 20 μg of chlorophyll *a* per milliliter. A representative experiment is shown. (*D*) WebLogo based on 30 NtcA binding motifs identified with MEME upstream of sequences from class 4 genes. (*E*) Box plot representation of class 1 expression patterns. Outliers are represented by dots. (*F*) Expression patterns of class 1 genes *gifA*, *gifB*, and *rplC*. Light and dark bars are used for ammonium and nitrate, respectively. (*G*) Positioning of 29 NtcA boxes (*SI Appendix*, Table S2) relative to the TSS with activation (green) and repression (red) functions (details are provided in *SI Appendix*, Fig. S3 *A* and *B*).

interpretation of the effects of $pipX^{Y32A}$ and $pipX^{E4A}$ alleles on the expression of NtcA gene targets is that the lower affinity of $PipX^{E4A}$ or $PipX^{Y32A}$ (compared with PipX) for P_{II} would result in increased concentration of the mutant NtcA–PipX complexes, and consequent activation of NtcA targets in nitrate-grown cells. In addition, the lower affinity of $PipX^{Y32A}$ (compared with $PipX^{E4A}$) for NtcA would account for the nitrogen regulation observed in $CS3X^{Y32A}$ (i.e., the differences between nitrate and ammonium cultures). This result suggests that the mutant protein $PipX^{Y32A}$, despite its reduced affinity for P_{II}, is still engaged in partner swapping between NtcA and P_{II}. Strong support for the idea that the two mutant proteins still interact with P_{II} in vivo comes from the finding that the toxicity conferred by $pipX^{Y32A}$ and $pipX^{E4A}$ alleles is counteracted by P_{II} (17, 19).

To provide functional evidence that the mutant proteins $PipX^{Y32A}$ and $PipX^{E4A}$ exerted their effect on class 4 genes by interacting with NtcA rather than an NtcA-independent mechanism, we tested two functions that require NtcA in *S. elongatus*: growth on nitrate as a nitrogen source and nitrate transport. The prediction was that the *ntcA* null allele should be epistatic to the $pipX^{Y32A}$ and $pipX^{E4A}$ alleles.

When we attempted to inactivate *ntcA* by homologous recombination with the *ntcA::aphII* null allele, kanamycin-resistant clones carrying the inactive allele were obtained for CS3X, $CS3X^{E4A}$, or $CS3X^{Y32A}$ when transformants were selected on plates containing ammonium but not on nitrate. Furthermore, the ammonium-selected clones failed to grow on nitrate and to transport it (Fig. 4*C*). This therefore implies that the increased expression of the nitrate transport and assimilation systems in the $CS3X^{E4A}$ and $CS3X^{Y32A}$ strains requires NtcA. Thus, the results support the activation of class 4 by mutant NtcA–PipX complexes.

To provide additional evidence that the $PipX^{Y32A}$ and $PipX^{E4A}$ exerted their effect on class 4 genes by interacting with NtcA, and to gain further insights into the NtcA regulon in *S. elongatus*, we looked for NtcA binding motifs in promoter regions. The canonical NtcA-activated promoter is composed of an NtcA binding box, traditionally described with the consensus $GTAN_8TAC$, centered at ~41.5 nt upstream from the transcription start site (TSS), and separated 22–23 nt from a −10 box conforming to the consensus TAN_3T (20). In addition to this orthodox promoter structure, which matches that of the *Escherichia coli* class II Crp-dependent promoters, NtcA can activate from positions further upstream or from sequences not matching the reported consensus (21).

NtcA binding motifs were found in a few genes or operons experimentally characterized in *S. elongatus* (22–24) and in others predicted to be controlled by NtcA in the cyanobacterium *S. elongatus* PCC 6301 (25), which is almost identical to *S. elongatus*. Remarkably, NtcA binding motifs were found in association with 29 of the 30 transcription units in class 4 (Fig. 5 and *SI Appendix*, Table S2), a result indicating that class 4 contains almost exclusively operons directly activated by NtcA.

The extended NtcA binding consensus derived from the in silico analysis of class 4 genes (Fig. 4*D*) included the experimentally characterized binding site at the *glnN* promoter, previously referred to as atypical (23, 26). The high levels of transcripts found in strain $CS3X^{Y32A}$ in nitrate and in strain $CS3X^{E4A}$ in both ammonium and nitrate enabled us to map putative TSSs roughly from the RNA-sequencing read data (*SI Appendix*, Table S2 and Fig. S2). This, together with the predicted −10 elements, helped position the putative NtcA boxes. According to our analysis, most NtcA boxes appear to be centered at the canonical −41.5 position (Fig. 4*G* and *SI Appendix*, Fig. S3*A*).

Taking into account both the presence and positions of NtcA boxes and the expression patterns, only two class 4 genes were atypical: *Synpcc7942_1363* (the only gene in this class without recognizable NtcA boxes) and tRNA-Phe. Although direct regulation and binding in the absence of predictable NtcA binding sites has recently been suggested (27, 28), the rather attenuated NtcA-dependent pattern of *Synpcc7942_1363* could be explained

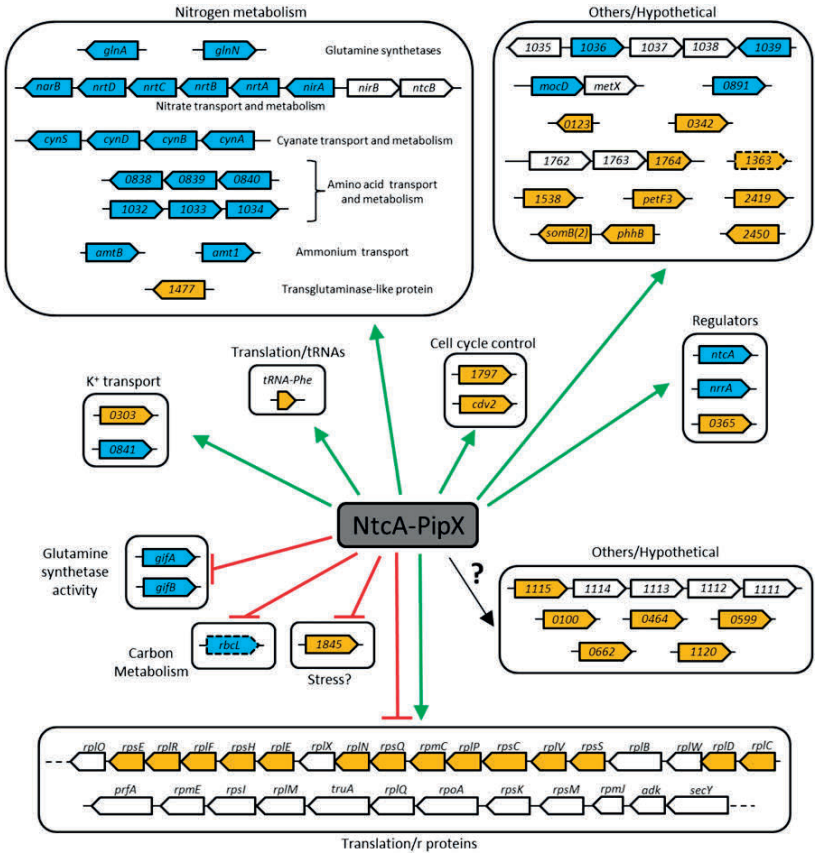


Fig. 5. Genomic organization of the *S. elongatus* NtcA–PipX regulon. Transcriptional units identified from the multivariate analysis as targets of NtcA–PipX complexes are shown in separate panels according to functional categories. ORFs and their orientation on the chromosome are shown as wide arrows in blue or orange, respectively, when there is or is not experimental or in silico evidence (22–25). Dashed arrows indicate no recognizable NtcA boxes. Green arrows and red blunt lines refer to activation and repression roles, respectively, inferred for NtcA–PipX. The question mark refers to genes with NtcA boxes of unknown function.

by indirect activation. On the other hand, tRNA-Phe was up-regulated in strains CS3X^{E4A} and CS3X^{Y32A} in nitrate and its promoter contains an NtcA binding site. However, in contrast to all other class 4 genes, which exhibit higher expression in nitrate, tRNA-Phe was induced in ammonium (Dataset S9). A precedent for a rather atypical regulation of a translation-related factor by NtcA is *gltX*, encoding the glutamyl-tRNA synthetase, which is expressed in both nitrate and ammonium in an NtcA-dependent manner (29).

In summary, transcriptomic analysis from a panel of five *S. elongatus* strains and two nitrogen conditions, followed by multivariate analyses, resulted in the identification of a coherent group of genes activated by the NtcA–PipX complex. These assignments, when considered alongside previously known NtcA target genes, identified new members of the NtcA regulon. In silico analysis allowed us to expand the consensus NtcA binding motif and to map NtcA boxes in promoter regions. The results provide further insights into the 2-OG–dependent partner swapping of PipX between NtcA and P_{II}, as well as the properties of the PipX^{Y32A} and PipX^{E4A} proteins.

NtcA-Repressed Genes. As happens with other members of the CRP family of transcriptional activators (30), NtcA can also mediate repression when its binding sites overlap the RNA polymerase binding region located between −40 and +20 nt relative to the TSS. Paradigms of this type of control are the *Synechocystis* sp. PCC 6803 genes *gifA* and *gifB*, encoding the glutamine synthetase-inactivating factors IF7 and IF17 (31).

If repression of target genes is also favored by mutations increasing NtcA–PipX complex formation, the prediction is that *pipX* mutations would have the opposite effect on NtcA-repressed and NtcA-activated transcripts. As shown in Fig. 3*A*, classes 1 and 4 occupy opposite positions along the PC2 axis, suggesting that NtcA may negatively control class 1 genes. Furthermore, the box plot of class 1 (Fig. 4*E*) can be regarded as a roughly inverted version of the class 4 box plot, with the main difference being exhibited by the *pipX*^{Y32A} allele.

The extended consensus from Fig. 4*D* was then used to search for NtcA boxes outside class 4, for which the complete set of 257 differentially regulated genes was used. Only 10 new hits (associated with nine genes) were found, four of which corresponded to three class 1 genes and included *gifA* and *gifB*, with NtcA boxes located at −32.5 and −16.5, respectively (*SI Appendix*, Fig. S3*B* and Table S2). Their expression pattern in WT *S. elongatus* (Fig. 4*F*) is consistent with the regulatory importance of IF7 and IF17 in the nitrogen metabolism of cyanobacteria (32, 33). Although PipX does not seem to be required for nitrate repression, the *pipX*^{E4A} and *pipX*^{Y32A} alleles confer repression in nitrate and, to a lesser extent, also in ammonium. Neither *gifA* or *gifB* showed ammonium induction in strain CS3X, a phenomenon that remains to be investigated.

Interestingly, 16 ribosomal genes were found in class 1. Two NtcA binding sites were found upstream of the coding region of *rplC*, the first gene of the main ribosome cluster. The site centered at −41.5, suggesting activation, is at odds with the inclusion of this gene in class 1. However, the second NtcA site at +10.5 is easier to reconcile with the negative effect of the gain-of-function alleles on transcript levels (Fig. 4 *E* and *F*). It is worth noting that the ribosomal genes that make a major contribution to the class 1 box plot were not down-regulated by CS3X^{Y32A} in nitrate. Taken together, these results suggest complex NtcA-dependent regulation at the ribosomal gene cluster and an intriguing connection between the nitrogen signaling system and the gene expression machinery.

No recognizable NtcA boxes were found at the remaining 16 genes found in class 1. These include *rbcL*, reported to be NtcA-repressed in *Anabaena* (34). Here, it is tempting to propose the involvement of the two orphan response regulators from class 4: NrrA, reported to be part of the NtcA regulatory cascade in *Anabaena* (35, 36), and the *Synpcc7942_0365* gene product, either or both of which may account for indirect NtcA-dependent repression of some class 1 genes.

A simple interpretation of the results presented so far on classes 4 and 1 (Fig. 4 and *SI Appendix*, Fig. S3 and Table S2) is that whereas positive regulation tends to be directly exerted by NtcA, negative regulation would tend to be exerted indirectly, via an NtcA-activated repressor.

NtcA Targets Outside Classes 1 and 4. The only gene outside class 1 matching functional and structural criteria for NtcA repression was *Synpcc7942_1845* from class 2.1 (*SI Appendix*, Fig. S4). Its expression differed significantly from the representative pattern of class 2.1 (Fig. 6). As expected for NtcA-repressed genes, it was down-regulated in CS3X^{E4A} and, to a greater extent, in CS3X^{Y32A}. It also contains an NtcA binding site centered at −2.5 (*SI Appendix*, Fig. S3*B*). However, the expression pattern of *Synpcc7942_1845* differed between the two control strains, being higher in the CS3X background, a result suggesting that additional regulatory mechanisms may prevent transcript accumulation in ammonium in the WT. Recent data indicate that *Synpcc7942_1845* acts as a general stress protein in *S. elongatus* (37), suggesting that it may be subjected to multiple regulatory controls.

Only six genes, belonging to class 2.2 (one gene), class 3.2 (three genes), and the group of nonclassified genes (two genes), had NtcA boxes for which we could not infer a particular function. Their atypical expression patterns may be due to the presence of cryptic NtcA sites or to the confluence of additional regulatory systems.

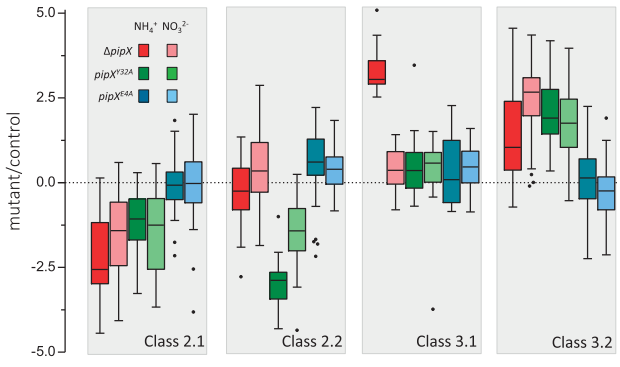


Fig. 6. Box plot representation of the within-group variance of the mutant/control comparisons in classes 2.1, 2.2, 3.1, and 3.2. Details are as described in Fig. 4*A*.

Recent transcriptomic and ChIP studies of the *Anabaena* NtcA regulon allowed the identification of large numbers of NtcA targets in this heterocyst-forming cyanobacterium (28, 38). In this context, it is worth noting that our multivariate analysis did not aim to provide a comprehensive study of NtcA targets but, instead, was designed to identify genes with paradigmatic and simple regulation by PipX, which may be representative of interactions with NtcA or with other transcriptional regulators. In this context, several NtcA target genes previously characterized in *S. elongatus* did not score above the cutoff levels. Notably, *glnB*, also expressed from a strong NtcA-independent promoter (39), *gltX*, subjected to both positive and negative regulation by NtcA (29) and *nblA*, controlled by several additional regulators (40–42), were not detected in our analysis.

NtcA-Independent Expression Patterns and Functions of PipX Regulons. The effects of mutations on the expression patterns of classes 2.1, 2.2, 3.1, and 3.2 are illustrated in Figs. 3*B* and 6. The distinctive feature of class 2.1 transcripts was their down-regulation by Δ*pipX* and *pipX*^{Y32A} alleles in both nitrate and ammonium. Exactly the opposite effect was found for class 3.2 genes: up-regulation by Δ*pipX* and *pipX*^{Y32A} alleles in both nitrate and ammonium. Class 2.2 was characterized by significant down-regulation by *pipX*^{Y32A}, especially in ammonium, and class 3.1 was characterized by up-regulation by Δ*pipX* in ammonium.

Most of the mutant/control changes detected in our analysis were similar in nitrate and ammonium cultures. The finding that the Δ*pipX* allele affected class 2.1 and class 3.2 genes similarly in nitrate and ammonium cultures suggests that PipX plays the same role in both nitrogen-rich conditions. On the other hand, only Δ*pipX* in class 3.1 and *pipX*^{Y32A} in class 4 affected gene expression specifically in one condition. Although the molecular basis of PipX regulation of class 3.1 genes remains elusive, the 2-OG–dependent partner swapping of PipX between NtcA and P_{II} (12) provides the background to interpret the class 4 expression pattern, further indicating that 2-OG was limiting for NtcA–PipX^{Y32A} complex formation but not for NtcA–PipX^{E4A} complex formation in our ammonium cultures.

The effect of Δ*pipX*, *pipX*^{E4A}, and *pipX*^{Y32A} alleles on the transcript levels of classes 2.1, 2.2, 3.1, and 3.2 cannot be reconciled with the involvement of NtcA–PipX complexes in the regulation of their corresponding genes. In particular, the drastic impact of *pipX* inactivation at classes 2.1, 3.1 (specifically in ammonium), and 3.2 and the lack of effect of the alleles *pipX*^{E4A} (in classes 2.2 and 3.1) and *pipX*^{Y32A} (in class 3.1) do not support the involvement of NtcA. Furthermore, with very few exceptions (in classes 2.2 and 3.2; *SI Appendix*, Table S2), NtcA boxes were absent from the genes or transcription units involved. Because both the response to the Δ*pipX*, *pipX*^{E4A}, and *pipX*^{Y32A} alleles and the promoter structure (discussed above) argued against the involvement of NtcA–PipX complexes in regulation of class 2.1, 2.2, 3.1, and 3.2 genes, our working hypothesis is that each of these four groups constitutes regulons influenced by PipX in an NtcA-independent manner.

To investigate further the internal coherence of the groups of genes obtained by multivariate analyses, which was based on the ability of *S. elongatus* gene transcripts to respond similar to *pipX* alleles, we followed the cluster of orthologous genes (COG) classification system to assign functions within groups.

The distribution of both COG categories and genes of unknown function differed greatly across the six groups (*SI Appendix*, Figs. S5 and S6). Genes of unknown function made similar contributions to the complete *S. elongatus* genome (*ca.* 37%) and to the 222 genes included in the six groups analyzed here (*ca.* 34%) but were especially abundant in class 2.1 (24 of 37 genes) and rare in class 1 (four of 32 genes). Translation was

overrepresented (14% vs. 7% in the complete *S. elongatus* genome), but ribosomal proteins were found exclusively in class 1 (18 of 32 genes) and tRNAs were found almost exclusively in class 3.2 (12 of 13 genes), thus revealing a strong connection between PipX and translation. The second most abundant category was energy production and conversion, which was particularly abundant in class 3.1 (9 of 18 genes were photosynthesis-related genes). Inorganic ion transport and metabolism and amino acid transport and metabolism were found almost exclusively in class 4 (21 of 43 genes in total), in close agreement with their role in nitrogen assimilation. Carbohydrate transport and metabolism was relatively well represented in class 2.2 (5 of 43 genes).

PipX Modulon, a Working Model. The results presented in this work provided important insights into the *S. elongatus* NtcA regulon while revealing unexpected functions of PipX (Fig. 7). The finding that the six groups of genes that emerged from the multivariate analysis showed very good internal coherence gave credit to the hypothesis that PipX is involved in processes other than coactivation/corepression of NtcA targets, which are mainly related to nitrogen assimilation (class 4 and some class 1 genes). In this context, the finding that the double-mutant *ntcA pipX* is less viable than the *ntcA* single mutant, as inferred by failure to segregate the Δ*pipX* allele in the *ntcA* null mutant (*SI Appendix*, Fig. S7), supports the involvement of PipX in NtcA-independent functions.

The finding that regulation by PipX could be observed in both or just one of the nitrogen conditions used (ammonium in class 3.1) supports the idea that signals other than 2-OG affect PipX interactions. The identification as members of the PipX modulon of highly expressed genes for ribosomal proteins (class 1), tRNAs (class 3.2), and photosynthesis (class 3.1) further suggests that NtcA-independent regulons participate in the adaptation of the cyanobacterial machineries for translation and photosynthesis to nutrient or other environmental changes. The emerging picture is that of PipX as a multifunctional protein involved in fine-tuning of different gene expression programs in response to different signals.

The effect of *pipX* inactivation on transcript levels indicated a positive role for PipX in class 2.1 and a negative role in classes 3.1 and 3.2. On the other hand, the impact of $pipX^{Y32A}$ on class 2.2 transcripts indicated that PipX has the potential to act as a negative regulator, probably under environmental conditions not tested in this work. How PipX exerts the different roles inferred here is still a matter of speculation. It may function by interacting with transcriptional, signal transduction, or even posttranscriptional regulators.

To account for the six groups of genes identified here with common expression patterns, we propose the involvement of PipX in a minimum of four types of regulatory complexes, of which only one (NtcA–PipX complex) is presently characterized.

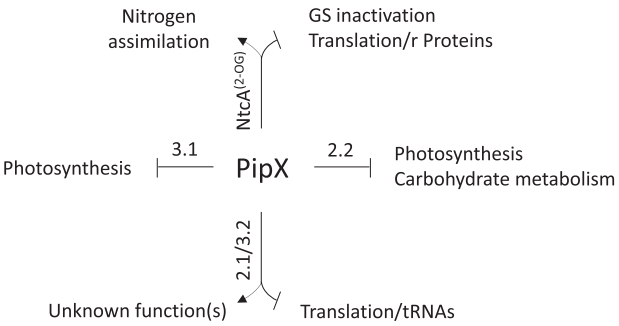


Fig. 7. PipX regulatory functions inferred from this work. Positive regulation and negative regulation are depicted by arrowheads and blunt lines, respectively.

The rather symmetrical effects of *pipX* mutations in the expression patterns of classes 1 and 4, reflecting the involvement of NtcA–PipX complexes in both groups, have a parallel in classes 2.1 and 3.2, which gave remarkably symmetrical box plots; thus, it is tempting to speculate that the same regulatory complex could be involved in activating and repressing these genes. Different complexes would be involved in repression of class 2.2 and 3.1 genes.

The intracellular availability of PipX and its specific interactions in particular environmental conditions or genetic backgrounds are probably determined by a network of interactions involving NtcA, P_{II}, their binding proteins and effectors (Fig. 1), and perhaps P_{II}-modifying enzymes (43, 44). This complexity makes it difficult to infer a potential role for P_{II} in the proposed PipX complexes. Notably, P_{II} proteins acting in complexes with DNA-binding transcriptional regulators have been reported in other systems (45–47).

The crucial impairment of binding to P_{II} caused by the PipX substitution Y32A argues against the involvement of P_{II} in the regulation of class 2.2 genes, where the $pipX^{Y32A}$ allele causes gain of function (as a repressor). Direct involvement of the tudor-like domain, and therefore of P_{II}, also appears unlikely in class 3.1 regulatory complexes, because $pipX^{Y32A}$ and $pipX^{E4A}$ alleles had no effect on the repressive function of PipX; here, it is tempting to propose a role for the C-terminal α-helices of PipX in repression. Finally, although the (complete) loss of function conferred by $pipX^{Y32A}$ would support the involvement of the tudor-like domain in regulation of class 2.1 and 3.2 genes, the silent effect of the $pipX^{E4A}$ allele argues against direct involvement of P_{II}.

Whatever mechanisms are involved, our results suggest that PipX participates in at least three novel regulatory scenarios, different from the known NtcA–PipX complexes, to modulate gene expression.

Materials and Methods

Strains, Growth Conditions, and Nitrate Transport Assays. *S. elongatus* strains and plasmids are listed in *SI Appendix*, Table S1. Growth, genetic manipulations and nitrate uptake assays were as described (13, 17).

Preparation and RNA Analysis. For RNA preparations, 100-mL cultures of each strain were grown in BG11 or $BG11^{A}$ until a OD_{750} nm of ~0.9 was attained. RNA was purified with a Qiagen RNeasy Protect Bacteria Mini Kit and on-column RNase-free DNase I digestion. Samples were assayed for RNA integrity using an Agilent 2100 Bioanalyzer and quantified with a Qubit fluorometer (Life Technologies). Removal of 16S and 23S rRNA from total RNA was performed using a MicrobExpress Bacterial mRNA Purification Kit (Ambion) or treatment with a Ribo-Zero Magnetic Kit (Epicenter). RNA samples were divided into multiple aliquots of ≤5 μg of RNA, and separately enriched mRNA samples were pooled, run on the 2100 Bioanalzyer to confirm reduction of 16S and 23S rRNA before preparation of cDNA fragment libraries with a ScriptSeq v2 RNA-Seq Library Preparation Kit, and sequenced (Illumina HiSeq2000; Macrogene).

Computational Methods. Ten datasets of unique mappable reads covered each nucleotide strand specifically for an average of ~120 to 500 times for the chromosome and 1.5 to 190 times for the plasmid. Read alignments were performed using Bowtie2 (48) against an *S. elongatus* reference chromosome and endogenous plasmid (GenBank accession nos. CP000100 and CP000101, respectively). Gene expression, represented as reads per kilobase, was determined by Samtools (49), the Artemis Genome Browser (Wellcome Trust Sanger Institute) (50), and homemade Perl scripts. The data were normalized by quantiles (51). Statistical analysis was performed using the DESeq package (52) and R software (www.r-project.org/). Normalized reads per kilobase per million data are provided in Dataset S9.

NtcA motifs were first identified with MEME (53) (150 nt upstream of the TSSs or, when unpredicted, of initiation codons and a background consisting of a fourth-order Markov model of the entire genome) and were used to search for palindromic motifs between 6 and 20 bp. FIMO was used (54) to identify NtcA boxes outside class 4. The position-specific probability matrix for the motif was derived from the 30 matches provided by MEME. Searches were performed in sequences comprising 250 nt upstream and 50 nt

downstream of the TSSs or 250 nt upstream of the initiation codon. The hits were filtered using two criteria: *P* value <0.0001 and q-value <0.7.

A COG list was downloaded from Cyanobase (http://genome.microbedb.jp/cyanobase/SYNPCC7942), with manual assignation where relevant.

Multivariate analyses were carried out with SPSS (IBM) and R. PCA was applied on the correlation matrix, with varimax rotation of the first two principal components. Package cluster (55) was used for fuzzy c-means clustering (56).

ACKNOWLEDGMENTS. We thank J. I. Labella for carrying out MEME and FIMO analysis and Circos Perl programming, A. Obrebska and S. Burillo for generation of mutants, K. Forchhammer for plasmid pNTCA-KAN, and V. Rubio for critically reading the manuscript. This work was supported by the Spanish Ministry of Economy and Competitivity (Grants BFU2009-07371, BFU2012-33364, and BFU2011-26608) and the European Seventh Framework Program (Grants 289326/FP7-KBBE-2011-5 and 282004/FP7-HEALTH-2011-2.3.1-2).

1. Muro-Pastor MI, Reyes JC, Florencio FJ (2001) Cyanobacteria perceive nitrogen status by sensing intracellular 2-oxoglutarate levels. *J Biol Chem* 276(41):38320–38328.
2. Muro-Pastor MI, Reyes JC, Florencio FJ (2005) Ammonium assimilation in cyanobacteria. *Photosynth Res* 83(2):135–150.
3. Forchhammer K (2004) Global carbon/nitrogen control by PII signal transduction in cyanobacteria: From signals to targets. *FEMS Microbiol Rev* 28(3):319–333.
4. Laurent S, et al. (2005) Nonmetabolizable analogue of 2-oxoglutarate elicits heterocyst differentiation under repressive conditions in *Anabaena* sp. PCC 7120. *Proc Natl Acad Sci USA* 102(28):9907–9912.
5. Leigh JA, Dodsworth JA (2007) Nitrogen regulation in bacteria and archaea. *Annu Rev Microbiol* 61:349–377.
6. Fokina O, Chellamuthu VR, Forchhammer K, Zeth K (2010) Mechanism of 2-oxoglutarate signaling by the *Synechococcus elongatus* PII signal transduction protein. *Proc Natl Acad Sci USA* 107(46):19760–19765.
7. Truan D, et al. (2010) A new P(II) protein structure identifies the 2-oxoglutarate binding site. *J Mol Biol* 400(3):531–539.
8. Burillo S, Luque I, Fuentes I, Contreras A (2004) Interactions between the nitrogen signal transduction protein PII and N-acetyl glutamate kinase in organisms that perform oxygenic photosynthesis. *J Bacteriol* 186(11):3346–3354.
9. Llácer JL, et al. (2007) The crystal structure of the complex of PII and acetylglutamate kinase reveals how PII controls the storage of nitrogen as arginine. *Proc Natl Acad Sci USA* 104(45):17644–17649.
10. Vázquez-Bermúdez MF, Herrero A, Flores E (2002) 2-Oxoglutarate increases the binding affinity of the NtcA (nitrogen control) transcription factor for the *Synechococcus glnA* promoter. *FEBS Lett* 512(1-3):71–74.
11. Tanigawa R, et al. (2002) Transcriptional activation of NtcA-dependent promoters of *Synechococcus* sp. PCC 7942 by 2-oxoglutarate *in vitro*. *Proc Natl Acad Sci USA* 99(7): 4251–4255.
12. Espinosa J, Forchhammer K, Burillo S, Contreras A (2006) Interaction network in cyanobacterial nitrogen regulation: PipX, a protein that interacts in a 2-oxoglutarate dependent manner with PII and NtcA. *Mol Microbiol* 61(2):457–469.
13. Espinosa J, Forchhammer K, Contreras A (2007) Role of the *Synechococcus* PCC 7942 nitrogen regulator protein PipX in NtcA-controlled processes. *Microbiology* 153(Pt 3): 711–718.
14. Valladares A, et al. (2011) Specific role of the cyanobacterial PipX factor in the heterocysts of *Anabaena* sp. strain PCC 7120. *J Bacteriol* 193(5):1172–1182.
15. Llácer JL, et al. (2010) Structural basis for the regulation of NtcA-dependent transcription by proteins PipX and PII. *Proc Natl Acad Sci USA* 107(35):15397–15402.
16. Zhao MX, et al. (2010) Structural basis for the allosteric control of the global transcription factor NtcA by the nitrogen starvation signal 2-oxoglutarate. *Proc Natl Acad Sci USA* 107(28):12487–12492.
17. Laichoubi KB, Espinosa J, Castells MA, Contreras A (2012) Mutational analysis of the cyanobacterial nitrogen regulator PipX. *PLoS ONE* 7(4):e35845.
18. Espinosa J, Castells MA, Laichoubi KB, Forchhammer K, Contreras A (2010) Effects of spontaneous mutations in PipX functions and regulatory complexes on the cyanobacterium *Synechococcus elongatus* strain PCC 7942. *Microbiology* 156(Pt 5):1517–1526.
19. Espinosa J, Castells MA, Laichoubi KB, Contreras A (2009) Mutations at *pipX* suppress lethality of PII-deficient mutants of *Synechococcus elongatus* PCC 7942. *J Bacteriol* 191(15):4863–4869.
20. Herrero A, Muro-Pastor AM, Flores E (2001) Nitrogen control in cyanobacteria. *J Bacteriol* 183(2):411–425.
21. Luque I, Forchhammer K (2008) Nitrogen assimilation and C/N balance sensing. *The Cyanobacteria: Molecular Biology, Genetics and Evolution*, ed Herrero EFA (Caister Academic, Norwich, UK), pp 335–382.
22. Luque I, Flores E, Herrero A (1994) Molecular mechanism for the operation of nitrogen control in cyanobacteria. *EMBO J* 13(23):5794.
23. Sauer J, Dirmeier U, Forchhammer K (2000) The *Synechococcus* strain PCC 7942 *glnN* product (glutamine synthetase III) helps recovery from prolonged nitrogen chlorosis. *J Bacteriol* 182(19):5615–5619.
24. Vázquez-Bermúdez MF, Paz-Yepes J, Herrero A, Flores E (2002) The NtcA-activated *amt1* gene encodes a permease required for uptake of low concentrations of ammonium in the cyanobacterium *Synechococcus* sp. PCC 7942. *Microbiology* 148(Pt 3):861–869.
25. Su Z, Olman V, Mao F, Xu Y (2005) Comparative genomics analysis of NtcA regulons in cyanobacteria: Regulation of nitrogen assimilation and its coupling to photosynthesis. *Nucleic Acids Res* 33(16):5156–5171.
26. Aldehni MF, Forchhammer K (2006) Analysis of a non-canonical NtcA-dependent promoter in *Synechococcus elongatus* and its regulation by NtcA and PII. *Arch Microbiol* 184(6):378–386.
27. Camargo S, Valladares A, Flores E, Herrero A (2012) Transcription activation by NtcA in the absence of consensus NtcA-binding sites in an *anabaena* heterocyst differentiation gene promoter. *J Bacteriol* 194(11):2939–2948.
28. Picossi S, Flores E, Herrero A (2014) ChIP analysis unravels an exceptionally wide distribution of DNA binding sites for the NtcA transcription factor in a heterocyst-forming cyanobacterium. *BMC Genomics* 15(1):22.
29. Luque I, Contreras A, Zabulon G, Herrero A, Houmard J (2002) Expression of the glutamyl-tRNA synthetase gene from the cyanobacterium *Synechococcus* sp PCC 7942 depends on nitrogen availability and the global regulator NtcA. *Mol Microbiol* 46(4):1157–1167.
30. Kolb A, Busby S, Buc H, Garges S, Adhya S (1993) Transcriptional regulation by cAMP and its receptor protein. *Annu Rev Biochem* 62:749–795.
31. García-Domínguez M, Reyes JC, Florencio FJ (2000) NtcA represses transcription of *gifA* and *gifB*, genes that encode inhibitors of glutamine synthetase type I from *Synechocystis* sp. PCC 6803. *Mol Microbiol* 35(5):1192–1201.
32. Galmozzi CV, Fernández-Avila MJ, Reyes JC, Florencio FJ, Muro-Pastor MI (2007) The ammonium-inactivated cyanobacterial glutamine synthetase I is reactivated *in vivo* by a mechanism involving proteolytic removal of its inactivating factors. *Mol Microbiol* 65(1):166–179.
33. García-Domínguez M, Reyes JC, Florencio FJ (1999) Glutamine synthetase inactivation by protein-protein interaction. *Proc Natl Acad Sci USA* 96(13):7161–7166.
34. Ramasubramanian TS, Wei TF, Golden JW (1994) Two *Anabaena* sp. strain PCC 7120 DNA-binding factors interact with vegetative cell- and heterocyst-specific genes. *J Bacteriol* 176(5):1214–1223.
35. Ehira S, Ohmori M (2006) NrrA directly regulates expression of *hetR* during heterocyst differentiation in the cyanobacterium *Anabaena* sp. strain PCC 7120. *J Bacteriol* 188(24):8520–8525.
36. Ehira S, Ohmori M (2006) NrrA, a nitrogen-responsive response regulator facilitates heterocyst development in the cyanobacterium *Anabaena* sp. strain PCC 7120. *Mol Microbiol* 59(6):1692–1703.
37. Ruffing AM (2013) RNA-Seq analysis and targeted mutagenesis for improved free fatty acid production in an engineered cyanobacterium. *Biotechnol Biofuels* 6(1):113.
38. Mitschke J, Vioque A, Haas F, Hess WR, Muro-Pastor AM (2011) Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in *Anabaena* sp. PCC7120. *Proc Natl Acad Sci USA* 108(50):20130–20135.
39. Lee HM, Vázquez-Bermúdez MF, de Marsac NT (1999) The global nitrogen regulator NtcA regulates transcription of the signal transducer PII (GlnB) and influences its phosphorylation level in response to nitrogen and carbon supplies in the Cyanobacterium *synechococcus* sp. strain PCC 7942. *J Bacteriol* 181(9):2697–2702.
40. López-Redondo ML, et al. (2010) Environmental control of phosphorylation pathways in a branched two-component system. *Mol Microbiol* 78(2):475–489.
41. Luque I, Zabulon G, Contreras A, Houmard J (2001) Convergence of two global transcriptional regulators on nitrogen induction of the stress-acclimation gene *nblA* in the cyanobacterium *Synechococcus* sp. PCC 7942. *Mol Microbiol* 41(4):937–947.
42. Salinas P, et al. (2007) The regulatory factor SipA provides a link between NblS and NblR signal transduction pathways in the cyanobacterium *Synechococcus* sp. PCC 7942. *Mol Microbiol* 66(6):1607–1619.
43. Chellamuthu VR, Alva V, Forchhammer K (2013) From cyanobacteria to plants: Conservation of PII functions during plastid evolution. *Planta* 237(2):451–462.
44. Forchhammer K (2010) The network of P(II) signalling protein interactions in unicellular cyanobacteria. *Adv Exp Med Biol* 675:71–90.
45. Castellen P, Rego FG, Portugal ME, Benelli EM (2011) The *Streptococcus* mutans GlnR protein exhibits an increased affinity for the *glnRA* operon promoter when bound to GlnK. *Braz J Med Biol Res* 44(12):1202–1208.
46. Kayumov A, Heinrich A, Fedorova K, Ilinskaya O, Forchhammer K (2011) Interaction of the general transcription factor TnrA with the PII-like protein GlnK and glutamine synthetase in *Bacillus subtilis*. *FEBS J* 278(10):1779–1789.
47. Oliveira MA, et al. (2012) Interaction of GlnK with the GAF domain of Herbaspirillum seropedicae NifA mediates NH_4^+- regulation. *Biochimie* 94(4):1041–1047.
48. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.
49. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
50. Rutherford K, et al. (2000) Artemis:Sequence visualization and annotation. *Bioinformatics* 16(10):944–945.
51. Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94.
52. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.
53. Bailey TL, et al. (2009) MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res* 37(Web Server issue):W202–W208.
54. Grant CE, Bailey TL, Noble WS (2011) FIMO: Scanning for occurrences of a given motif. *Bioinformatics* 27(7):1017–1018.
55. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2014) Cluster Analysis Basics and Extensions. R package version 1.15.2. Available at http://cran.stat.ucla.edu/web/packages/cluster/index.html. Accessed April 20, 2014.
56. Kaufman L, Rousseeuw PJ (1990) *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley, New York).

PipX, the coactivator of NtcA, is a global regulator in cyanobacteria

Espinosa J*, Rodríguez-Mateos F†, Salinas P*, Lanza VF‡, Dixon R§, de la Cruz F‡ and Contreras A*.

* División de Genética, Universidad de Alicante, Apartado de Correos 99, 03080, Alicante, Spain; † Departamento de Matemática Aplicada, Universidad de Alicante, Apartado de Correos 99, 03080, Alicante, Spain; ‡ Departamento de Biología Molecular. Facultad de Medicina. Universidad de Cantabria & Instituto de Biomedicina y Biotecnología de Cantabria. Avenida Cardenal Herrera Oria, s/n, 39011, Santander, Spain; § Department of Molecular Microbiology, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK.

Corresponding Author: Contreras A, División de Genética, Universidad de Alicante, Apartado de Correos 99, 03080, Alicante, Spain.

Supplementary information

1. Methods

2. Supplementary Figures

3. Supplementary Tables

4. Supplementary Material References

1. Methods

Computational methods. To carry out multivariate analysis of all three mutant vs control comparisons (6 variables = 3 mutants/control x 2 nitrogen culture conditions) the standardized residuals from linear regressions of (log-transformed) data for mutant vs control strains were calculated. Genes with residuals lower than 1.5 for all comparisons were considered as non-responding genes, (Fig. S1, gray dots) and truncated normal distributions were fitted to the distributions of their residuals (Table S3). A total of 282 genes, with residuals exceeding a threshold value of 2.5 for at least one variable, were considered as responding genes and selected for further analysis (the probability to be in this group would be lower than 0.005, with less than 13 genes expected by chance). Artemis Genome Browser was used to exclude genes with reads mapping mainly to the non-coding strand. The resulting 257 genes were subjected to different statistical analysis carried out with SPSS and R software's. First, the residuals of the 257 genes for the 6 comparisons were subjected to principal component analysis (PCA) and Varimax rotation was used to enhance the interpretability of the first two principal components. Classification into four main groups was obtained using k-means cluster analysis (1). Additional independent classifications into four groups were obtained by hierarchical Ward's minimum variance clustering and fuzzy c-means clustering, carried out using function fanny in package cluster (2). Genes coherently grouped with the three clustering methods (222 genes) were selected, and 6 groups were obtained by cutting the cluster dendrogram from Ward's method (classes 1, 2.1, 2.2, 3.1, 3.2 and 4).

Determination and positioning of NtcA binding sites. To determine the Transcription start site (TSS), reads were pooled and the position where the greatest number of reads began was considered the most likely TSS (Fig. S2). The resulting TSS differ to those determined by primer extension on 1 nt (for *ntcA*, *amt1* and *nirA*) 6 nt (for *glnA*) and 8 nt (for *glnN*). NtcA motifs present upstream Class 4 genes were identified with MEME (3). 150 nt upstream of the TSSs or, when unpredicted, of initiation codons for each gene and a background consisting of a fourth-order Markov model of the entire genome was used to search for palindromic motifs between 6 bp and 20 bp. The first statistically significant motif found among members of group 4 (*E*-value $1.3e^{-009}$) corresponded to a previously reported NtcA binding site. 30 out of 31 sequences presented one NtcA site according to MEME. To identify NtcA sites outside of class 4, FIMO was used (4). The position-specific probability matrix (PSPM) for the motif was derived from the 30 matches provided by MEME:

Position	A	C	G	T
1	0.166667	0.366667	0.083333	0.383333
2	0.166667	0.333333	0.133333	0.366667
3	0.000000	0.016667	0.966667	0.016667
4	0.000000	0.016667	0.000000	0.983333
5	0.616667	0.050000	0.183333	0.150000
6	0.383333	0.050000	0.416667	0.150000
7	0.083333	0.583333	0.100000	0.233333
8	0.333333	0.250000	0.183333	0.233333
9	0.283332	0.099999	0.349999	0.266665
10	0.266666	0.350000	0.100000	0.283333
11	0.233332	0.183332	0.249999	0.333332
12	0.233332	0.099999	0.583332	0.083332
13	0.149999	0.416666	0.049999	0.383332
14	0.149999	0.183332	0.049999	0.616666
15	0.983332	0.000000	0.016666	0.000000
16	0.016666	0.966666	0.016666	0.000000
17	0.366666	0.133332	0.333332	0.166666
18	0.383332	0.083332	0.366666	0.166666

Possible -10 elements were searched at positions 6-8 nucleotides upstream of determined TSS. Positioning of putative NtcA binding sites and -10 elements is shown in Figs. S3A-B.

Functional classification of genes within the PipX modulon. The COG list downloaded from Cyanobase on 30th (October of 2013) was manually re-annotated when relevant (see Dataset S1, Tables S1-8). Heat maps of functional groups amongst the different groups of genes were generated with the gplots library of R (Fig. S5). Correlations between list of genes and COGs were represented using linkage maps plotted with Circos (5) (Fig. S6).

Inactivation of *pipX* in an *ntcA* null background. *S. elongatus* WT and the *ntcA* null strain CS37 (*ntcA::Cm*^{R}, (6)) were transformed in parallel with plasmid pUAGC59.1 (7). Transformant clones selected on BG11^{A} plates containing kanamycin were subjected to PCR analysis to verify segregation of *pipX* alleles (Fig. S7).

2. Supplementary Figures

Figure S1. Analysis of regulated and non-regulated genes obtained after mutant/control comparisons. The value of the residuals (ZRE) with highest dispersion for the mutant/control comparisons, grown in nitrate (*x* axis) or ammonia (*y*-axis), was calculated for each gene and represented as a scatterplot. In gray are depicted non-responsive genes with a ZRE<1.5 for all 6 comparisons (Δ*pipX*/WT, CS3X^{Y32A}/CS3X and CS3X^{E4A}/CS3X in nitrate or ammonia). In purple and red are, respectively, the genes considered as non-responsive in ammonia and nitrate (ZRE<1.5 in either nitrate or ammonia for all 3 mutant/control comparisons). In green are depicted genes with ZRE>1.5. Note that this group contain the 282 genes considered as being differentially regulated (ZRE>2.5).

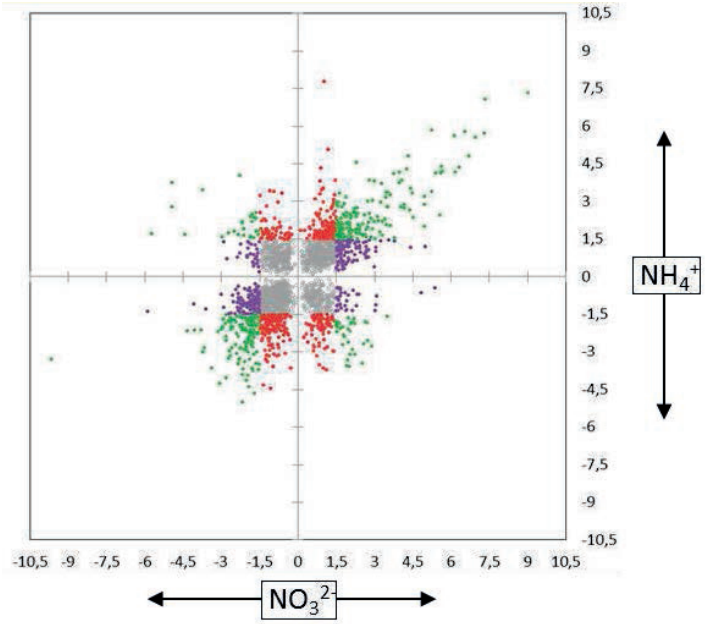


Figure S2. Determination of the possible transcription start site (TSS) of *ntcA* (*Synpcc7942_0127*). The image corresponds to Artemis software loaded with the *S. elongatus* chromosome (GenBank reference CP000100) and the BAM files of the CS3X^{E4A} and CS3X^{Y32A} strains corresponding to the nitrogen conditions indicated. The graphical view corresponds to coverage by strand of the entire *ntcA* genomic region. The nucleotide sequence of the *ntcA* promoter is shown below. The TSS, in this example -108 nt, was determined according to the position where reads began relative to the ATG initiation codon (+1).

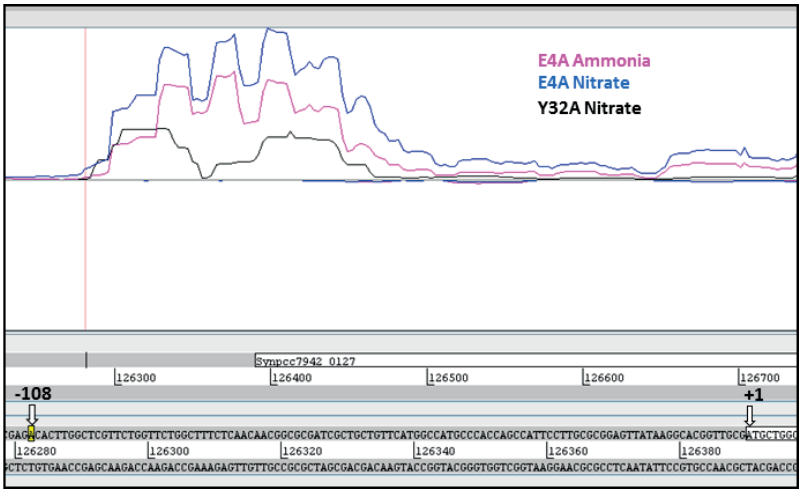


Figure S3. Promoter sequences of genes with NtcA sites at activation (green boxed) or repression (red boxed) compatible positions. A) Promoters of genes found in class 4 with NtcA boxes at canonical positions and -10 elements highlighted in green and blue, respectively. Boxed or underlined nucleotides refer, respectively, to TSS reported by primer extension (8-10) or inferred from this work. Numbers refer to *Synpcc7942* locus tags. B) Promoter of three class 1 (*gifA*, *gifB* and *rplC*) and one class 2.1 (locus tag *1845*) genes with their NtcA repressor sites and -10 elements (when recognizable) highlighted in red and blue, respectively. A putative NtcA box at activation position in *rplC* is highlighted in green. Initiation codons are in bold. Note that the NtcA site in *gifA* overlaps the -10 element.

A

nirA	GTT	GTAGTTTCTGTTAC	CAATTGCGAATCGAGAACTGCC	TAATCT	GCCGAGTATGCAAGCTGCTT
amt1	ACT	GTTACATCGATTAC	AAAACAACCTTGAGTCTCGCTG	AATGCT	TACAGAGATCTCACAAGGAT
ntcA	AAA	GTAGCAGTTGCTAC	AAGCAGCAGCTAGGCTAGGCCG	TACGGT	AACGAGACACTTGGCTCGTT
glnA	TAT	GTATCAGCTGTTAC	AAAAGTGCCGTTTCGGGCTACC	TAGGAT	GAAAGCGGTCAGCAATGCTT
glnN	TCT	GTATCTTTTCTAGC	GATCGAGCTGGTCACCATTGAG	TACGAT	CAATTGACTAGCTTTTTTGG
0840	TTA	GTAGTCACCGTTAC	AGTCATTCCTTAGAACTTGTTT	TAGTTA	ATTGGTAGCTGTTTACACCA
1036	ACA	GTAGCTACAGCTAC	GAATTGAAAGCTGGTGCCAGCC	TAGCTT	GGATGGTTAAGCACTTTCAG
1538	CTC	GTATCAAGGGTAAC	GGTTCTTATTGGTTGACTTTAA	TACTTT	AAAATTCTGACTCACCTCTC
cynA	GTT	GTAACGACGGCTAC	ATTTTGACCCTGGGGTTACTAC	TACCAT	TCGCCCTTAACTGAGGAAAG
amtB	AAA	GTAGCAAAAGTTAC	GTATATCACCAGTCTGCCTAGCC	AGAGTT	GTGAGATCTCCGAACCTTC
1797	AGT	GTGGTGGCGGTAAC	AGTTCTGAGCTAGACAGGGCCGT	TAAAGT	AGCGCCAACTTGTGATGGC
mocD	CAG	GTAGCGATCGCTAC	AGCAGCAACAAAGATTGACACGA	TTGGTT	AGAAAGAGACCTTGGGCGA
0342	AAA	GTGGTGTATGCGAC	AAATGAAGTGCTGCATCTCTTGC	TAAGAC	TGTGGCAAAAGCCTTCGAA
1032	GCC	GTATCCCGAACTAC	AGAAGTGGACTCTGAGCGATTTC	TATAGT	CCTTTCATGACATTTATGG
nrrA	CTC	GTAAAGGCGAATAC	AGAAGCCACAATGGACAGCTTGC	TAGGTT	AAAGTCACAACTCCCAAGA
phhB	TCC	GTAGCAAAAAGCAC	GAGATTACTCGTCTCAAGTCGC	TACTTT	TAAATGCACCTCGTGTTGAC
1039	TTC	GTAGCCTTGCTTAC	ACTTTGCCGATCTGGCCTCTTC	TAGGAT	GCCAAAGCTGTGGCTTAGCC

B

gifA		TCC	GTAGCATTTGCTAC	AGTTTAAGGGCGAGGGTAATCGTTGGCCCAGTTTATCTGGG..N_{70}..**ATG**
gifB		ACG	GTAGCAATTTATAC	AGAAAAATATTGCTAGATTTAAATCAAGCCTAGTTTTGCTT..N_{16}..**ATG**
1845				TTT GTATCCGTTGCTAC CAAAGGTGGCCC **ATG**
rplC	ACC GTATCGGTAGACAC	GATTCAGCAAAATGGTGTATTGT	TTATAT	TTGTCGCTCT GTTGCGGTAAGCAC GAC..N_{270}..**GTG**

Figure S4. Expression patterns of Class 2.1 locus tag *1845*. Light and dark bars are used for ammonium and nitrate.

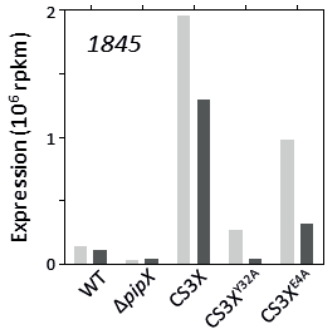


Figure S5. COG functions within the PipX modulon. A heatmap plot of gene distribution according to COG categories in the genome and the 6 classes defined here. The heat scale is the frequency of ORFs assigned to each individual COG category listed to the right.

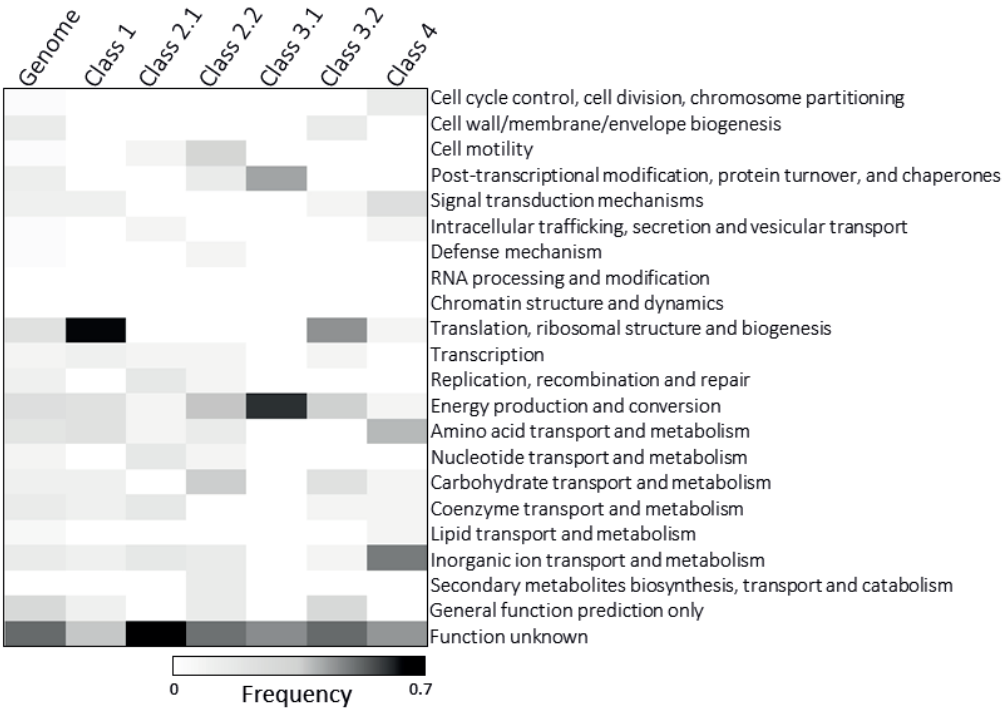


Figure S6. Functional categories in the 6 groups defined by multivariate analysis. Multiple maps connecting genes in the six groups with COGs categories are shown. Listed genes are named according to Synpcc7942 locus tag ID. COGs are depicted in different colors. C, Energy production and conversion; D, Cell cycle control, cell division and chromosome partitioning; E, Amino acid transport and metabolism; F, Nucleotide transport and metabolism; G, Carbohydrate transport and metabolism; H, Coenzyme transport and metabolism; I, Lipid transport and metabolism; J, Translation, ribosomal structure and biogenesis; K, Transcription; L, Replication, recombination and repair; M, Cell wall/membrane/envelope biogenesis; N, Cell motility; O, Post-translational modification, protein turnover, chaperones; P, Inorganic ion transport and metabolism; Q, Secondary metabolites biosynthesis, transport and catabolism; R, General function; T, Signal transduction mechanisms; U, Intracellular trafficking, secretion and vesicular transport; V, Defense mechanism; S, Function unknown.

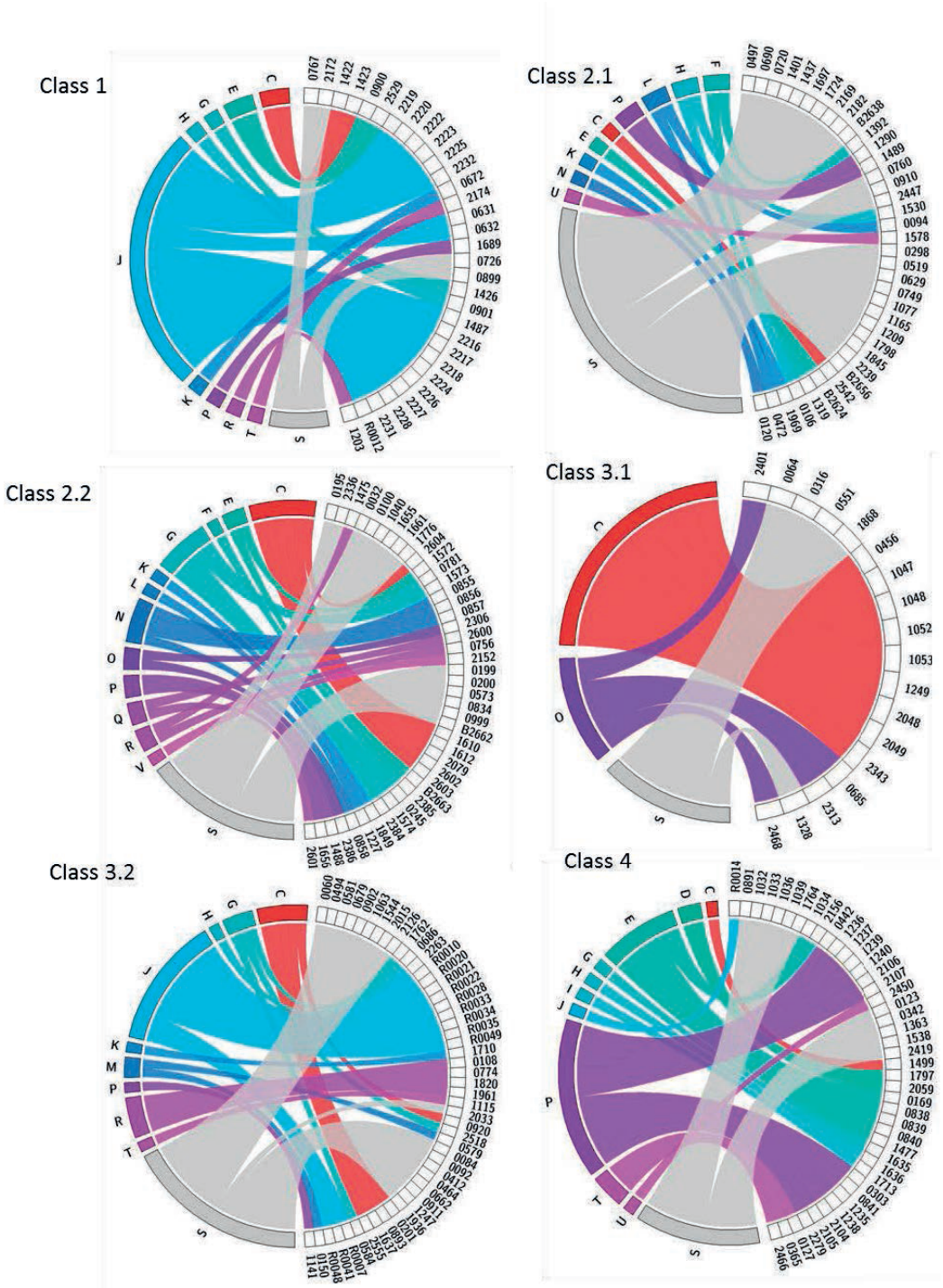
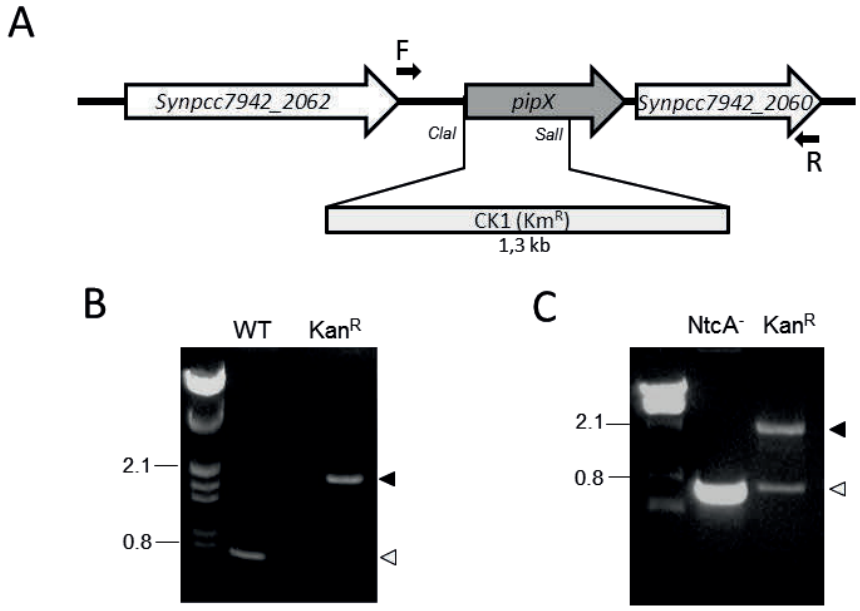


Figure S7. Segregation of *pipX* inactivation allele in the *ntcA* null strain. (A) Schematic representation of the *pipX::CK1* allele generated by insertion of the kanamycin-resistance cassette C.K1 with indication of relevant restriction sites and positions of primers (black arrows) used to verify allele replacement. (B) PCR verification of *pipX* alleles with primers PipX-126F (F) and PipX-5R (R) in *S. elongatus* (WT) and a representative kanamycin-resistant clone transformed with plasmid pUAG59.1 (Kan^R). (C). Idem in CS37 ($NtcA^-$). PCR products corresponding to *pipX* wild-type and *pipX::CK1* alleles are indicated as white or black arrowheads, respectively. Size markers (λ *Hind*III *Eco*RI) with indication of relevant band sizes (in Kilobases) are shown at the left.



3. Supplementary Tables

Table S1. Strains and Plasmids used in this work.

Strain/Plasmid	Genotype or relevant characteristics	Reference
S. elongatus	Wild-type *Synechococcus elongatus* PCC 7942	Pasteur Culture Collection
SA591	(Δ*pipX*) PipX$^-$, (*pipX::C.K1*), Km^r	(7)
CS3X	PipX, ϕ(*C.S3-pipX*), Sm^r	(11)
CS3X^{E4A}	PipX^{E4A}, ϕ(*C.S3-pipX^{E4A}*), Sm^r	(12)
CS3X^{Y32A}	PipX^{Y32A}, ϕ(*C.S3-pipX^{Y32A}*), Sm^r	(12)
MNtcA	NtcA$^-$, (*ntcA::aphII*), Km^r	(13)
CS3X-MNtcA	PipX NtcA$^-$, ϕ(*C.S3-pipX*), (*ntcA::aphII*), Sm^r Km^r	This work
CS3X^{E4A}-MNtcA	PipX^{E4A} NtcA$^-$, ϕ(*C.S3-pipX^{E4A}*), (*ntcA::aphII*), Sm^r Km^r	This work
CS3X^{Y32A}-MNtcA	PipX^{Y32A} NtcA$^-$, ϕ(*C.S3-pipX^{Y32A}*), (*ntcA::aphII*), Sm^r Km^r	This work
pUAGC59.1	(*pipX::C.K1*), Ap^r Km^r	(7)
pUAGC393	ϕ(*C.S3-pipX*), Ap^r Sm^r	(11)
pUAGC375	ϕ(*C.S3-pipX^{E4A}*), Ap^r Sm^r	(12)
pUAGC380	ϕ(*C.S3-pipX^{Y32A}*), Ap^r Sm^r	(12)
pNTCA-KAN	(*ntcA::aphII*), Km^r	(13)

Table S2. NtcA target genes in *S. elongatus*

Transcription unit	Gene name	Published TSS	Predicted TSS (this work)	NtcA site	NtcA site position
Synpcc7942_0123		n.d	n.d.	GTGGCGAAGAGTAC	n.d.
Synpcc7942_0127	*ntcA*	-108 (*-109*)	-108	GTAGCAGTTGCTAC	(*-40.5*)
Synpcc7942_0169	*glnN*	-30 (*-22*)	-30	GTATCTTTTCTAGC	(*-41.5*)
Synpcc7942_0303		n.d	+51	GTGATGAATGGCAC	-106.5
Synpcc7942_0342		-82	-45[(a)]	GTGGTGTATGCGAC	-41.5
Synpcc7942_0365		-51	n.d.	GTAATCTTTGTTAG	n.d.
Synpcc7942_0442	*amt1*	-102 (*-103*)	-102	GTTACATCGATTAC	(*-40.5*)
Synpcc7942_0840-0838		n.d.	-44	GTAGTCACCGTTAC	-41.5
Synpcc7942_0841		n.d.	+26[(a)]	GTCGCGATTGATAC	-70.5
Synpcc7942_0891		-34	-16[(a)]	GTAACTGGATACAC	-58.5
Synpcc7942_1032-1034		-25	-26	GTATCCCGAACTAC	-41.5
Synpcc7942_1036		-23	-24	GTAGCTACAGCTAC	-41.5
Synpcc7942_1039		-11	-51[(a)]	GTAGCCTTGCTTAC	-47.5
Synpcc7942_1240-1235	*nirA nrtA nrtB nrtC nrtD narB*	-56 (*-31*)	-30	GTAGTTTCTGTTAC	(*-41.5*)
Synpcc7942_1363		-105	n.d.	*Not found*	n.d.
Synpcc7942_1477		0	n.d.	GTAACATCTGACAC	n.d.
Synpcc7942_1499	*petF3*	-61	-64	GTGATTACCCCTAC	-123.5
Synpcc7942_1538		n.d.	-20	GTATCAAGGGTAAC	-41.5
Synpcc7942_1635	*somB(2)*	-63	n.d	GTTGACTAGGCGAC	n.d.
Synpcc7942_1636	*phhB*	-53	-25[(a)]	GTAGCAAAAAGCAC	-41.5
Synpcc7942_1713	*mocD*	-39	-40	GTAGCGATCGCTAC	-41.5
Synpcc7942_1764		0	n.d.	GTAACAGAGACAAC	n.d.
Synpcc7942_1797		-31	-31	GTGGTGGCGGTAAC	-47.5
Synpcc7942_2059	*cdv2* (*sepF*)	-84	n.d.	GTTTGTGCTATTAC	n.d.
Synpcc7942_2107-2104	*cynA cynB cynD cynS*	-16	-16	GTAACGACGGCTAC	-43.5
Synpcc7942_2156	*glnA*	0* (*-147*)	-141	GTATCAGCTGTTAC	(*-40.5*)
Synpcc7942_2279	*amtB*	-72	-36	GTAGCAAAAGTTAC	-43.5
Synpcc7942_2419		n.d.	n.d.	GTTGCCCCTTCTAA	n.d.
Synpcc7942_2450	*gspD*	-70	n.d.	GTGAGTGAATTTAC	n.d.
Synpcc7942_2466	*nrrA*	-23	-23	GTAAAGGCGAATAC	-43.5
Synpcc7942_R0014	tRNA-Phe	n.d.	n.d.	GTTGCTCCTCTCAC	n.d.
Synpcc7942_2529	*gifA*	-102	-102	GTAGCATTTGCTAC	-16.5
Synpcc7942_0900	*gifB*	-32	-32	GTAGCAATTTATAC	-32.5
Synpcc7942_2232-2203	*rplC…prfA*	-67*	-292	GTATCGGTAGACAC	-41.5
				GTTGCGGTAAGCAC	+11.5
Synpcc7942_1845		-24	-13[(a)]	GTATCCGTTGCTAC	-6.5
Synpcc7942_0100		-63	-25[(a)]	GTAGCCTAAGTCAC	-40.5
Synpcc7942_1115-1111		-174	n.d.	GTAATCATTATTAC	n.d.
Synpcc7942_0662		-59	-59	GTGGAAACCGTTAC	-62.5
Synpcc7942_0464		0	n.d.	GTAGCCACAGTCAC	n.d.
Synpcc7942_0599		-29	-29	GTTGCAGTGGCAAC	-83.5
Synpcc7942_1120		-19	n.d.	GTGGATGTCGTTAC	n.d.

NtcA boxes in differentially regulated genes. The majority of the genes listed belong to Class 4, except for: *gifA*, *gifB* and the operon *Synpcc7942_2232-2203* (belonging to Class 1), *Synpcc7942_1845* (Class 2.1), *Synpcc7942_0100* (Class 2.2), *Synpcc7942_0662*, *_0464* and the operon *_1115-1111* (Class 3.2) and *Synpcc7942_0599* and *_1120* (original Class 3). Sequence of NtcA sites, identified using either MEME or FIMO (underlined genes), correspond to strand +. Transcription start sites (TSS) are relative to the initiation codon. When available, TSS obtained previously by RNAseq (14) or primer extension (in brackets and italics) is indicated in the column labelled as published TSS. The predicted TSS obtained in this work (see details in methods and supplementary Fig. 1) are also provided (n.d. means not determined and [(a)] the provided TSS could not be determined in our wild-type datasets). The position of the NtcA sites is relative to the TSS either determined by primer extension or obtained in this work. The asterisk (*) represents discrepancies with the first gene of the transcription unit reported by (14) (the TSS is relative to the start codon of ORFs *Synpcc7942_2157* and *_2233*).

Table S3: Fitting of truncated normal distributions to the residuals of mutant/control comparisons for core genes.

	Δ*pipX*		*pipX*^{Y32A}		*pipX*^{E4A}	
	ammonium	nitrate	ammonium	nitrate	ammonium	nitrate
$\hat{\mu}$	-0.0686	-0.1099	0.0565	-0.0453	-0.0657	-0.0703
$\hat{\sigma}$	0.7956	0.7771	0.6749	0.7325	0.7142	0.5919
D	0.0133	0.0099	0.0114	0.0192	0.0187	0.0204
p	0.7456	0.9970	0.9826	0.5725	0.6043	0.4912

Estimated values of the parameters mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) by maximum likelihood (15)) and Kolmogorov-Smirnov goodness-of-fit statistics (*D*) and corresponding *p*-values. Parameters were estimated from the data, therefore, *p*-values are approximate.

4. Supplementary Material References

1. Kaufman L & Rousseeuw PJ (1990) *Finding Groups in Data: An Introduction to Cluster Analysis* (John Wiley & Sons, Inc., New York).
2. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2014). Cluster Analysis Basics and Extensions. R package version 1.15.2.
3. Bailey TL*, et al.* (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37(Web Server issue):W202-208.
4. Grant CE, Bailey TL, & Noble WS (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27(7):1017-1018.
5. Krzywinski M*, et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19(9):1639-1645.
6. Vega-Palas MA, Flores E, & Herrero A (1992) NtcA, a global nitrogen regulator from the cyanobacterium *Synechococcus* that belongs to the Crp family of bacterial regulators. *Mol Microbiol* 6(13):1853-1859.
7. Espinosa J, Forchhammer K, Burillo S, & Contreras A (2006) Interaction network in cyanobacterial nitrogen regulation: PipX, a protein that interacts in a 2-oxoglutarate dependent manner with PII and NtcA. *Mol Microbiol* 61(2):457-469.
8. Luque I, Flores E, & Herrero A (1994) Molecular mechanism for the operation of nitrogen control in cyanobacteria. *Embo J* 13(23):5794.
9. Sauer J, Dirmeier U, & Forchhammer K (2000) The *Synechococcus* strain PCC 7942 *glnN* product (glutamine synthetase III) helps recovery from prolonged nitrogen chlorosis. *J Bacteriol* 182(19):5615-5619.
10. Vazquez-Bermudez MF, Paz-Yepes J, Herrero A, & Flores E (2002) The NtcA-activated *amt1* gene encodes a permease required for uptake of low concentrations of ammonium in the cyanobacterium *Synechococcus* sp. PCC 7942. *Microbiology* 148(Pt 3):861-869.
11. Espinosa J, Castells MA, Laichoubi KB, Forchhammer K, & Contreras A (2010) Effects of spontaneous mutations in PipX functions and regulatory complexes on the cyanobacterium *Synechococcus elongatus* strain PCC 7942. *Microbiology* 156(Pt 5):1517-1526.
12. Laichoubi KB, Espinosa J, Castells MA, & Contreras A (2012) Mutational Analysis of the Cyanobacterial Nitrogen Regulator PipX. *PLoS One* 7(4):e35845.
13. Sauer J, Gorl M, & Forchhammer K (1999) Nitrogen starvation in *Synechococcus* PCC 7942: involvement of glutamine synthetase and NtcA in phycobiliprotein degradation and survival. *Arch Microbiol* 172(4):247-255.
14. Vijayan V, Jain IH, & O'Shea EK (2011) A high resolution map of a cyanobacterial transcriptome. *Genome Biol* 12(5):R47.
15. Johnson N, Kotz S, & Balakrishnan N (1994) *Continuous Multivariate Distributions* (John Wiley & Sons, New York) 2nd Ed.

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Biotechnology

Membrane-associated nanomotors for macromolecular transport

Elena Cabezon, Val F Lanza and Ignacio Arechaga

Nature has endowed cells with powerful nanomotors to accomplish intricate mechanical tasks, such as the macromolecular transport across membranes occurring in cell division, bacterial conjugation, and in a wide variety of secretion systems. These biological motors couple the chemical energy provided by ATP hydrolysis to the mechanical work needed to transport DNA and/or protein effectors. Here, we review what is known about the molecular mechanisms of these membrane-associated machines. Sequence and structural comparison between these ATPases reveal that they share a similar motor domain, suggesting a common evolutionary ancestor. Learning how these machines operate will lead the design of nanotechnology devices with unique applications in medicine and engineering.

Address
Departamento de Biología Molecular, Universidad de Cantabria, and Instituto de Biomedicina y Biotecnología de Cantabria (IBBTEC), UC-SODERCAN-CSIC, C. Herrera Oria s/n, 39011 Santander, Spain

Corresponding authors: Cabezon, Elena (cabezone@unican.es), Arechaga, Ignacio (arechagai@unican.es)

Current Opinion in Biotechnology 2012, **23**:537–544

This review comes from a themed issue on **Nanobiotechnology**

Edited by **Fernando de la Cruz** and **Geoffrey M Gadd**

For a complete overview see the Issue and the Editorial

Available online 19th December 2011

0958-1669/$ – see front matter,

http://dx.doi.org/10.1016/j.copbio.2011.11.031

Introduction

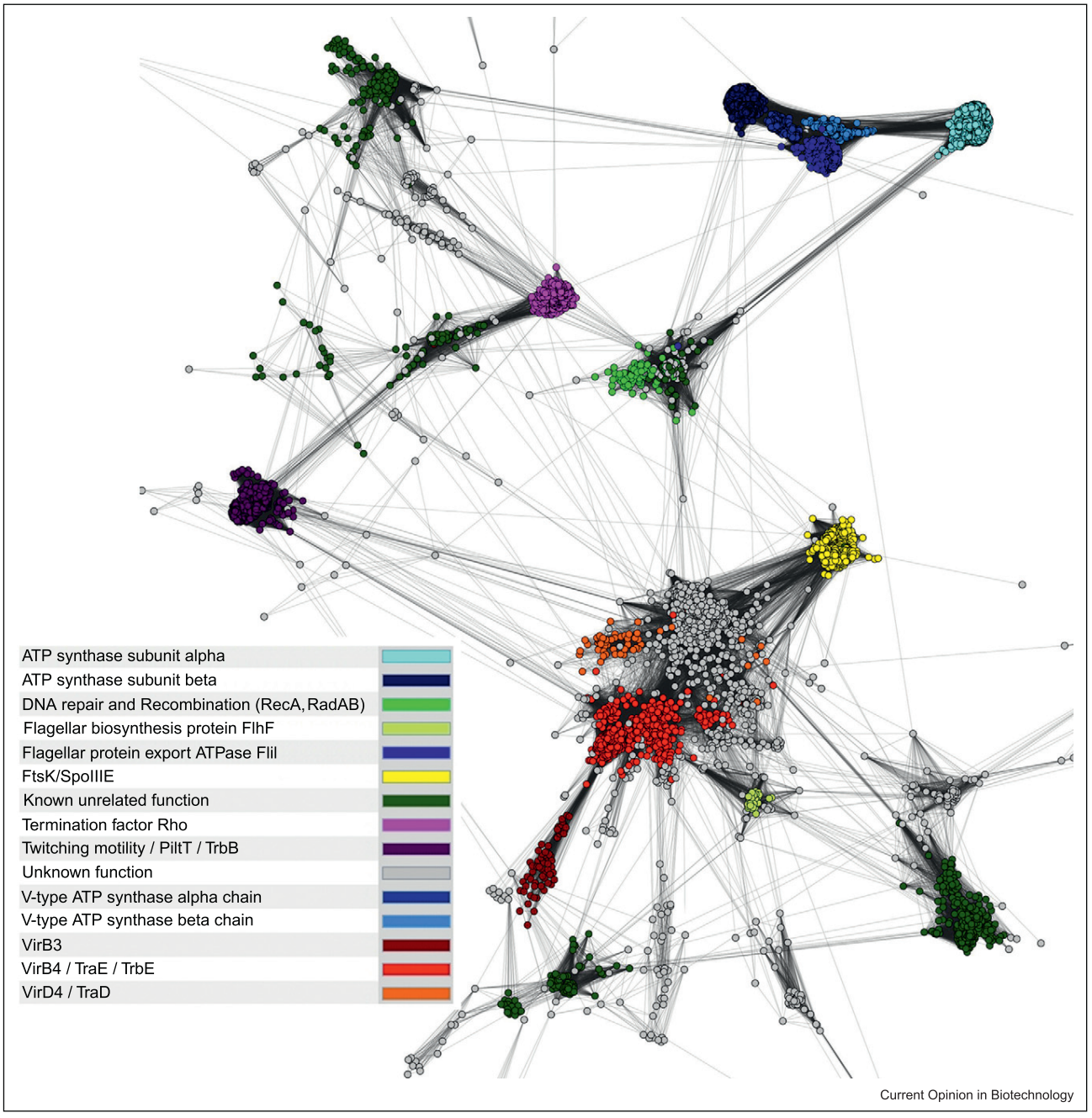
Bacterial conjugation and chromosome segregation during cell division are biological processes that need the participation of specific machineries that use ATP to pump DNA and protein effectors across membranes [1,2•]. Under rapid growth conditions, proper chromosome segregation is accomplished by FtsK/SpoIIIE proteins, which use the energy released from ATP hydrolysis to move the remainder of the bacterial chromosome across the constricting septal membranes [3,4]. The exchange of genetic material in bacterial conjugation and the delivery of oncogenic DNA and virulence effectors into eukaryotic cells are mediated by type IV secretion systems (T4SS) [5]. These systems consist of sophisticated machineries [6] involved in ssDNA and protein transport among cells; conducted by TrwB-like or VirB4-like ATP-driven motors, respectively [5]. In this article, we will focus on the structural similarities and evolution of these DNA and protein transport systems and other well known biological motors like the F-type ATPase. Understanding how these machineries perform their tasks in living cells is a key step to develop hybrid nanodevices with extraordinary applications.

Evolution of F/V-type ATPases and membrane DNA/protein transporters

Membrane-associated machines involved in macromolecular transport across cell membranes are members of the RecA-like family of hexameric P-loop ATPases [7]. In order to gather further understanding of the relationships between these motors, we performed a novel computational sequence analysis that resulted in the Protein Homology Network (PHN) shown in Figure 1. To this purpose, we generated a comprehensive library of related P-loop ATPases sequences (see Figure 1 for further details). Sequences within this library were compared using a BLAST All-vs-All analysis, clustered according to the sequence identity, and represented graphically as a network in which distances between different clusters are proportional to the e-values of the BLAST analysis; thus, the length of the edges is related to the phylogenetic distance among the connected proteins. Interestingly, clusters consisting of sequences corresponding to FtsK/SpoIIIE motors appear together with VirD4-like and VirB4-like proteins, which is in accordance with previous comparative genomic studies [7]. These results indicate that membrane-associated DNA and protein transporters share similar sequence motifs, suggesting a common evolutionary ancestor.

The evolution of membrane-associated DNA and protein transporters is intimately related to the evolution of F-type and V-type ATPases. It has been suggested that these ATPases emerged as a combination of a RNA/DNA helicase and a proton channel [8]. Following this idea, Mulkidjanian *et al.* [9••] proposed an interesting evolutionary scenario, where ancestors of ion-translocating ATPases, such as F-type ATPases, would have been membrane translocases that coupled ATP hydrolysis to RNA/DNA translocation across the membrane and, subsequently, to protein translocation (Figure 2). In both cases, the translocated polymer would occupy the place of the central stalk. The evolution of F/V-type ATPases from this hypothetical protein translocase ancestor could have resulted from mutations in the proteolipid membrane channel that would have impeded protein translocation [9]. According to this hypothesis, it is enticing to speculate that F/V–type ATPases could have evolved from a VirB3/VirB4-like ancestor.

Figure 1

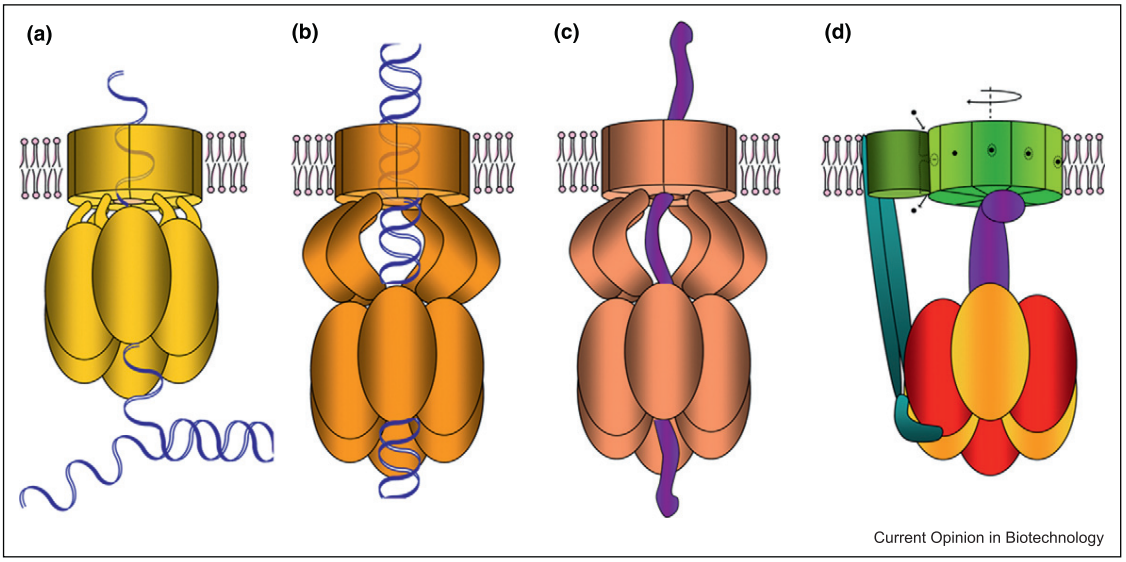


Protein Homology Network of membrane-associated DNA/protein translocases and related P-loop ATPases. Protein sequences were retrieved from Uniprot (Bacteria and Archaea) by psi-blast, clustered with CD-HIT [57], and aligned with MUSCLE [58]. A HMM profile was built for each cluster and new sequences were retrieved to generate a library of 30,038 P-loop ATPase sequences. Redundant sequences were clustered at 95% identity, resulting in a final library of 6902 clustered sequences. A BLAST All-vs_All analysis of sequences within this library resulted in a protein homology edge-weighted network, in which distances between nodes are proportional to the BLAST e-value.

Structural comparison of macromolecular transporters

FtsK/SpoIIIE DNA transporters are anchored to the membrane by several amino-terminal transmembrane segments [10] that appear to be essential for cell division [11], as they mediate interaction with other division machinery proteins [12]. Likewise, conjugative VirD4 DNA coupling proteins localize in the membrane and interact with other components of the T4SS core complex through its transmembrane domain [13]. Following the

Figure 2



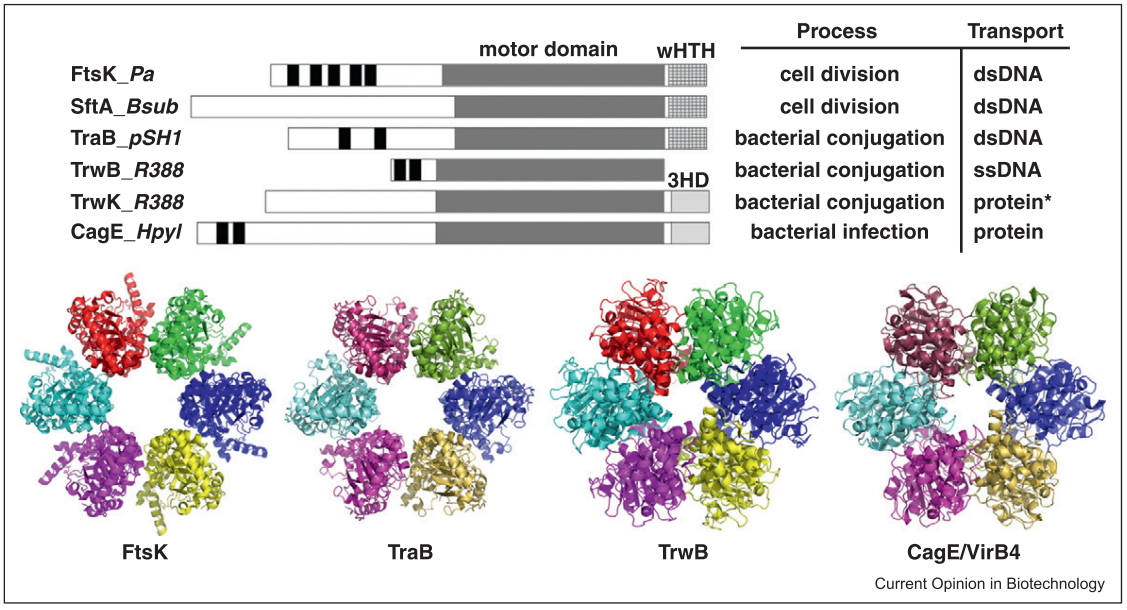
Evolution of membrane-anchored DNA and protein translocases (adapted from [9]). Comparison of single-stranded DNA pumps, such as TrwB-like coupling proteins **(a)**, double stranded DNA transporters of the FtsK/SpoIIIE protein family **(b)** and protein/effector (purple) translocators like VirB4-like proteins **(c).** F-type ATPases **(d)** have been proposed to have evolved from a protein translocator ancestor that in turn evolved from a primordial RNA/ssDNA translocase [9]. Black spots in **(d)** represent H^+ or Na^+ ions passing through the interface of subunit a and the c-ring, which is coupled to the rotation of the central γ subunit (purple) to drive the synthesis of ATP from ADP and Pi.

transmembrane segments, FtsK transporters present a linker domain of variable length, whose precise function is still unknown. This linker domain is shorter in SpoIIIE and it is absent in TrwB-like conjugative motors (Figure 3). The carboxy-terminal regions of these transporters are, however, very well conserved [14]. This region includes the ATPase motifs and it contains the DNA translocation domain. Crystallographic structures of the C-terminal domain of FtsK [15] and TrwB [16] show ring hexamers with a RecA-like fold that differ in the size of the central pore, which is 30 Å in diameter in FtsK and 20 Å in TrwB to accommodate dsDNA and ssDNA, respectively (Figure 3). Based on the structural analysis of FtsK, a rotary inchworm mechanism has been proposed for dsDNA translocation [15]. The structure suggests a sequential model of ATPase activity around the ring; a model originally proposed for F_1-ATPase [17] and later adapted for six catalytic subunits in several DNA translocating motors, such as the T7 gp4 helicase [18] or the φ29 DNA packaging motor [19].

FtsK/SpoIIIE proteins present an extra γ domain with a winged-helix fold [20•] that recognizes a specific 8 bp sequence in the chromosomal DNA (KOPS and SRS sequences for FtsK and SpoIIIE, respectively) [21,22]. Binding to this sequence guides the motor in the appropriate direction. Such a domain is not present in TrwB-like conjugative transporters, which are supposed to translocate ssDNA without specific sequence recognition. Instead, TrwB is a structure-specific DNA binding protein that recognizes with very high affinity a G-quadruplex DNA structure [23•], which could act as a loading site for the motor. Interestingly, the conjugative DNA transporter TraB of *Streptomyces*, a gram positive bacteria, resembles septal DNA translocators of the FtsK/SpoIIIE family in sequence and domain organization. This protein specifically binds to repeated 8 bp motifs on the conjugative plasmid, a binding mode reminiscent of the FtsK interaction with KOPS [24••]. Homology modeling of the TraB DNA translocase domain based on the FtsK crystallographic structure (2iut.pdb) (here rebuilt and shown in Figure 3) suggests that this protein has evolved from an FtsK-like ancestor protein and it is now adapted to translocate a circular dsDNA molecule by conjugation [24••].

Until recently, it was widely accepted that all chromosomal transporters in the FtsK/SpoIIIE family comprise transmembrane segments at their amino-terminal domain. However, a soluble SpoIIIE paralog in *B. subtilis* has been found to act at the early states of cell division [25•]. This motor, named SftA, aids in moving DNA away from the closing septum, whereas FtsK/SpoIIIE transporters would be involved in DNA translocation only when septum closure precedes chromosome segregation. Interestingly, an intriguing recent report showing that *E. coli* FtsK transmembrane domain is not required for efficient chromosome dimer resolution suggests that a membrane pore is not needed for DNA transport [26•], which is in contradiction with the pre-established idea

Figure 3



Domain structure of membrane-associated DNA and protein translocators. *Upper panel*, schematic comparison between *Pseudomonas aeruginosa* FstK, *Bacillus subtilis* SftA (a soluble form of SpoE III), *Streptomyces venezolae* TraB of plasmid pSVH1, the coupling protein TrwB of plasmid R388, and TrwK and CagE, the VirB4 homologs in plasmid R388 and *Helicobacter pylori*, respectively. The biological functions and the substrates translocated by these machines are indicated. *Black bars*, N-terminal transmembrane helices as predicted with TMHMM (www.cbs.dtu.dk/services/TMHMM-2.0). The motor domain common to all the proteins is depicted in gray. DsDNA translocators contain a C-terminal wHTH domain, which is absent in TrwB-like proteins. VirB4 proteins, instead, present a different three helical domain (3HD). *Bottom panel*, the structures of Ftsk_Pa (2iuu.pdb) and TrwB_R388 (1e9r.pdb) revealed that the inner diameter of both proteins (30 Å and 20 Å, respectively) is different to accommodate either dsDNA or ssDNA. Threading modeling of the motor domains of TraB-pSH1 using FtsK_Pa as template, and *H. pylori* CagE, using TrwB as template, indicate a high degree of structure conservation, which suggests a common evolutionary ancestor.

that the transmembrane domains form a pore in the membrane across which DNA is translocated [27].

Contrary to the structural information available on DNA transporters, little is known about the atomic structure of VirB4 proteins. This is the largest and most conserved family of proteins in Type IV secretion systems [28], playing an essential role in pilus biogenesis [29] and protein transport. Homology modeling of the C-terminus of *Agrobacterium tumefaciens* VirB4 [30], TrwK from the R388 plasmid [31], and *Helicobacter pylori* CagE (this work, see Figure 3) using the atomic coordinates of the coupling protein TrwB as a template, suggests that VirB4 subunits assemble as homohexamers and work as docking sites for substrate transport. Topology prediction for most of the VirB4 proteins indicates that they are soluble proteins, with the exception of VirB4 homologs belonging to the IncX branch and TrbE of RP4 plasmid [32]. Interestingly, a sequence analysis of proteins of this clade has identified members of the VirB4 family codifying a unique polypeptide composed of VirB3 and VirB4 proteins fused together [33]. VirB3 is an integral membrane protein located at the inner bacterial membrane [34], probably involved in anchoring VirB4 to the membrane. Interestingly, VirB3 proteins appear very close to the VirB4 cluster in the sequence analysis shown in Figure 1, which reinforces the idea of a co-evolution between VirB3 and VirB4 proteins. It is tempting to speculate that VirB3 forms a membrane channel coupled to VirB4 for protein transport.

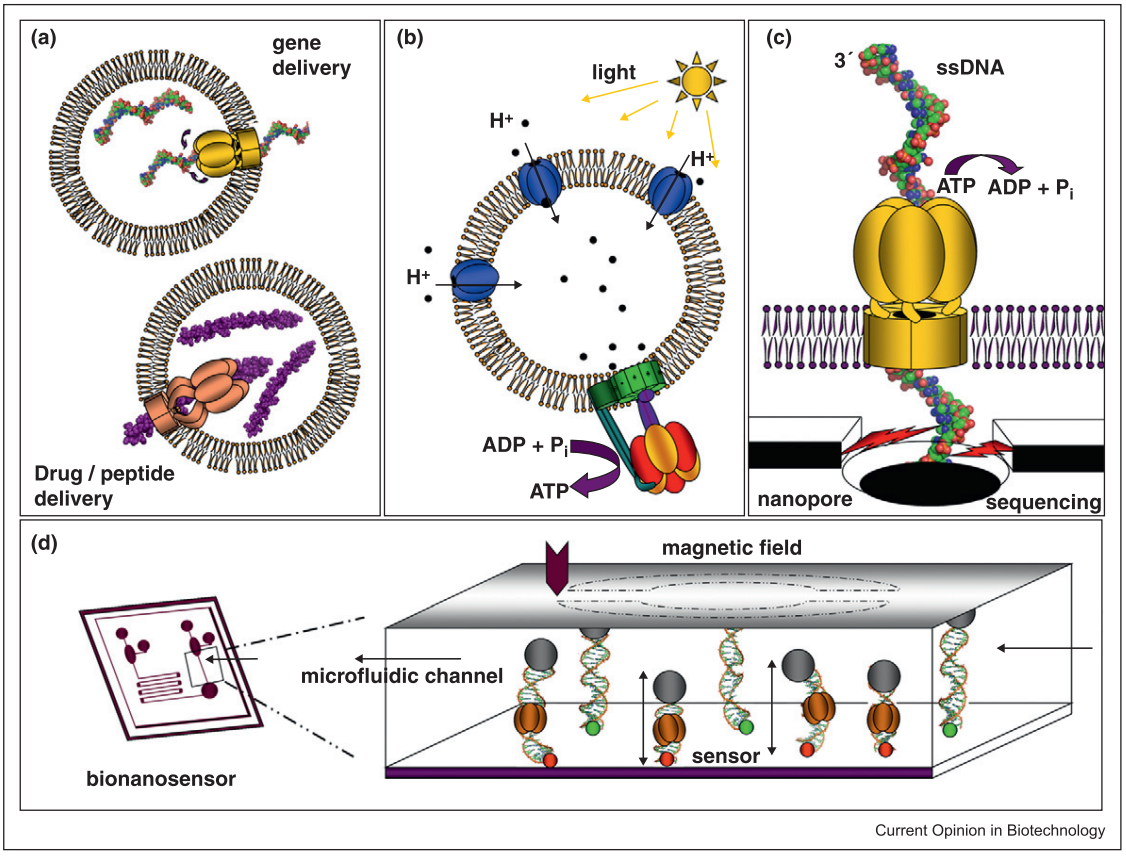
Applications in nanotechnology and nanomedicine

In recent years, there have been important advances in the understanding of how motor proteins work, unveiling remarkable properties. These molecular machines can couple chemical energy provided by ATP hydrolysis into mechanical work at nearly 100% efficiency, being able to move objects thousands of times their weight at high speed. The biophysical properties of these motors can be analyzed by single-molecule techniques, such as optical and magnetic tweezers or atomic force microscopy [35,36]. These techniques have provided information, for instance, on the translocation rate of FtsK on dsDNA [37–39] or in the study of the phi29 portal motor, which can package the whole DNA viral genome into the viral procapsid in ~5.5 min and against loads of up to 57 pN, making it one of the strongest nanomotors known to date [40].

On these bases, one of the aims on the emerging bionanotechnology field is to incorporate these molecular motors into useful devices [41,42]. One of the main problems in the construction of these nanodevices is to insert them into an artificial environment while retaining their function, being their long-term stability a limitation for their use in nanotechnology. For instance, the phi29 connector, redesigned to include regions of hydrophobicity, can act as a conductive channel to allow the translocation of double stranded DNA when reconstituted into liposomes [43•]. In that sense, the membrane-associated motors described here have the potential advantage of presenting transmembrane segments to be anchored to the membrane and, therefore, they do not require engineering. Following this idea, TrwB has been successfully reconstituted into liposomes [44]. Reconstitution of functional DNA/peptide transporters in lipid vesicles might have a tremendous potential for biomedical applications, such as gene or drug delivery (Figure 4a).The use of bacterial conjugation as a potential tool for genomic engineering to deliver DNA into virtually any type of cell has been envisaged [45] and T4SS-mediated heterologous DNA transfer into human cells has been reported [46,47]. Furthermore, T-DNA transfer from *Agrobacterium tumefaciens* into plant cells has been widely used to genetically modify crops [48]. However, a simplified delivery system including only the components needed for gene transformation would have the advantage of eliminating those elements that introduce uncontrolled variability into the system.

A limiting step in the development of these motor devices is that they use ATP as chemical fuel, which means that the device needs to have the ATP supply continuously refreshed. To this end, a natural ATP-regeneration system, consisting of bacteriorhodopsin plus ATP synthase, has been successfully reconstructed *in vitro* by inserting both complexes in lipid vesicles [49]. By using a light

Figure 4



Applications of membrane-associated motors in nanotechnology. **(a)** Reconstitution of a functional DNA or protein transporter into liposomes to perform DNA or peptide transfer assays in biomedical applications, such as gene or drug delivery; **(b)** artificial vesicles containing F_0F_1-ATP synthase and bacteriorhodopsin can be used to produce chemical ATP fuel through a light-driven proton gradient [49]; **(c)** application of a membrane-associated ssDNA translocating motor in third generation DNA sequencing technology. The motor is embedded in a lipid bilayer, driving the release of the ssDNA molecule at a constant ATP-dependent measurable velocity across a nanopore; **(d)** example of a Lab-on-a-Chip_device with a single-molecule signaling system (adapted from www.bionano-switch.info). A dsDNA molecule is attached to the chip and connected to a sensor that is able to detect changes on the distance of the DNA molecule to the chip, produced by the action of the FtsK DNA translocating motor.

source, bacteriorhodopsin creates a proton gradient across the artificial membrane that enables ATP synthase to produce ATP. The incorporation of a modular fuel supply component into the liposomes will result into self-contained, fuel-independent devices (Figure 4b); a step forward towards the creation of artificial organelles with wide nanotechnological applications [50].

Nanodevices have also potential applications in third generation DNA sequencing technologies. A new approach to read single DNA molecules in real time consists of reading the sequence as a DNA strand transits through a solid-state or a biological nanopore [51,52]. Protein nanopores, such as the heptameric α-hemolysin, have been genetically engineered to distinguish individual nucleotides within a single-stranded DNA sequence [53•] and, hybrid pores, formed by insertion of α-hemolysin into solid-state nanopores, have been designed [54]. However, one of the challenges of these technologies is temporal resolution, as DNA moves too rapidly through the nanopore. A simple setup device could be envisaged where a ssDNA translocating motor, such as TrwB, embedded in a lipid bilayer coupled to a solid-state nanopore, would drive the release of the single stranded DNA molecule at a constant ATP-dependent measurable velocity across the pore (Figure 4c). Future research in this field, as well as economical aspects, will determine the success of these technologies.

In addition to DNA sequencing, biological protein pores could have innumerable applications in nanomedicine, nanoelectronics and biosensing (see [55•] for a comprehensive review). For instance, biological molecular motors could also be used as bio-nanoactuators, working as switching devices. With this idea, a device consisting in a molecular motor, such as FtsK, acting on a DNA molecule attached to a microfluidic sensor has been designed (www.bionano-switch.info). The DNA molecule is attached to the sensor by one end and to a magnetic bead in the other and stretched by an external magnetic field. Each time FtsK pulls the magnetic bead towards the sensor, an electronic output is generated (Figure 4d). It is likely that the integration of these molecular switches with suitable electronic sensors will provide a major step forward in developing Lab-on-Chip technologies.

Outlook and perspectives

The potential use in nanotechnology of rotary motors like the F-type ATPases or the flagellum motor, as well as linear-motion motors, such as kinesin or myosin, has been extensively described elsewhere [41,42,56]. Here, we have focused on potential applications of membrane-associated DNA and protein translocases. Many of the applications of molecular motors in nanotechnology are mainly proof-of-concept examples that are far from being applied to ready-to-use devices. Progress in the commercial use of these machines will probably come from the integration of the biological understanding of the mode of action of these motors with the engineering development of functional devices.

Acknowledgement

This work was supported by Spanish MCINN Grant BFU2008-00806.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Chen I, Christie PJ, Dubnau D: **The ins and outs of DNA transfer in bacteria.** *Science* 2005, **310**:1456-1460.

2. • Burton B, Dubnau D: **Membrane-associated DNA transport machines.** *Cold Spring Harb Perspect Biol* 2010, **2**:a000406.
A comprehensive review of molecular motors involved in DNA transport across membranes in bacterial conjugation, transformation and cell division.

3. Bigot S, Sivanathan V, Possoz C, Barre FX, Cornet S F: **FtsK, a literate chromosome segregation machine.** *Mol Microbiol* 2007, **64**:1434-1441.

4. Sherratt DJ, Arciszewska LK, Crozat E, Graham JE, Grainge I: **The *Escherichia coli* DNA translocase FtsK.** *Biochem Soc Trans* 2010, **38**:395-398.

5. Alvarez-Martinez CE, Christie PJ: **Biological diversity of prokaryotic type IV secretion systems.** *Microbiol Mol Biol Rev* 2009, **73**:775-808.

6. Fronzes R, Christie PJ, Waksman G: **The structural biology of type IV secretion systems.** *Nat Rev Microbiol* 2009, **7**:703-714.

7. Iyer LM, Makarova KS, Koonin EV, Aravind L: **Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging.** *Nucleic Acids Res* 2004, **32**:5260-5279.

8. Walker J: **ATP synthesis by rotary catalysis (Nobel Lecture).** *Angew Chem Int Ed Engl* 1998, **37**:2309-2319.

9. •• Mulkidjanian AY, Makarova KS, Galperin MY, Koonin EV: **Inventing the dynamo machine: the evolution of the F-type and V-type ATPases.** *Nat Rev Microbiol* 2007, **5**:892-899.
The authors propose an evolutionary scenario where F/V-type ATPases originated from membrane protein translocases, which, in turn, evolved from ancestral RNA translocases.

10. Dorazi R, Dewar SJ: **Membrane topology of the N-terminus of the *Escherichia coli* FtsK division protein.** *FEBS Lett* 2000, **478**:13-18.

11. Draper GC, McLennan N, Begg K, Masters M, Donachie WD: **Only the N-terminal domain of FtsK functions in cell division.** *J Bacteriol* 1998, **180**:4621-4627.

12. Di Lallo G, Fagioli M, Barionovi D, Ghelardini P, Paolozzi L: **Use of a two-hybrid assay to study the assembly of a complex multicomponent protein machinery: bacterial septosome differentiation.** *Microbiology* 2003, **149**:3353-3359.

13. Llosa M, Zunzunegui S, de la Cruz F: **Conjugative coupling proteins interact with cognate and heterologous VirB10-like proteins while exhibiting specificity for cognate relaxosomes.** *Proc Natl Acad Sci USA* 2003, **100**:10465-10470.

14. Barre FX: **FtsK and SpoIIIE: the tale of the conserved tails.** *Mol Microbiol* 2007, **66**:1051-1055.

15. Massey TH, Mercogliano CP, Yates J, Sherratt DJ, Lowe J: **Double-stranded DNA translocation: structure and mechanism of hexameric FtsK.** *Mol Cell* 2006, **23**:457-469.

16. Gomis-Ruth FX, Moncalian G, Perez-Luque R, Gonzalez A, Cabezon E, de la Cruz F, Coll M: **The bacterial conjugation**

protein TrwB resembles ring helicases and F1-ATPase. *Nature* 2001, **409**:637-641.

17. Boyer PD: **The ATP synthase – a splendid molecular machine.** *Annu Rev Biochem* 1997, **66**:717-749.

18. Singleton MR, Sawaya MR, Ellenberger T, Wigley DB: **Crystal structure of T7 gene 4 ring helicase indicates a mechanism for sequential hydrolysis of nucleotides.** *Cell* 2000, **101**:589-600.

19. Chemla YR, Aathavan K, Michaelis J, Grimes S, Jardine PJ, Anderson DL, Bustamante C: **Mechanism of force generation of a viral DNA packaging motor.** *Cell* 2005, **122**:683-692.

20. • Lowe J, Ellonen A, Allen MD, Atkinson C, Sherratt DJ, Grainge I: **Molecular mechanism of sequence-directed DNA loading and translocation by FtsK.** *Mol Cell* 2008, **31**:498-509.
The crystallographic structure of FtsK γ-domain bound to KOPS, a specific 8 bp DNA signature, shows how three γ-domains recognize the sequence. The structure suggests that only 3 domains per hexamer are needed for KOPS recognition and *dif* recombination.

21. Bigot S, Saleh OA, Lesterlin C, Pages C, El Karoui M, Dennis C, Grigoriev M, Allemand JF, Barre FX, Cornet F: **KOPS: DNA motifs that control *E. coli* chromosome segregation by orienting the FtsK translocase.** *EMBO J* 2005, **24**:3770-3780.

22. Sivanathan V, Allen MD, de Bekker C, Baker R, Arciszewska LK, Freund SM, Bycroft M, Lowe J, Sherratt DJ: **The FtsK gamma domain directs oriented DNA translocation by interacting with KOPS.** *Nat Struct Mol Biol* 2006, **13**:965-972.

23. • Matilla I, Alfonso C, Rivas G, Bolt EL, de la Cruz F, Cabezon E: **The conjugative DNA translocase TrwB is a structure-specific DNA-binding protein.** *J Biol Chem* 2010, **285**:17537-17544.
Identification of G4-quadruplex DNA structures as the preferred substrate of the coupling protein TrwB. This structured DNA could work as a loading site for DNA pumping mediated by TrwB.

24. •• Vogelmann J, Ammelburg M, Finger C, Guezguez J, Linke D, Flotenmeyer M, Stierhof YD, Wohlleben W, Muth G: **Conjugal plasmid transfer in Streptomyces resembles bacterial chromosome segregation by FtsK/SpoIIIE.** *EMBO J* 2011, **30**:2246-2254.
The authors demonstrate that TraB, the coupling protein of a conjugative plasmid in Streptomyces, resembles septal DNA translocators of the FtsK/SpoIIIE family. The protein binds to repeated 8 bp motifs on the plasmid, a binding mode reminiscent of the FtsK interaction with 8 bp KOPS. Data suggest that the protein has evolved from an FtsK-like ancestor protein and it is now adapted to translocate a circular dsDNA molecule by conjugation.

25. • Kaimer C, Gonzalez-Pastor JE, Graumann PL: **SpoIIIE and a novel type of DNA translocase, SftA, couple chromosome segregation with cell division in Bacillus subtilis.** *Mol Microbiol* 2009, **74**:810-825.
Identification of a soluble paralog within the Ftsk/SpoIIIE family, probably involved in early stages of chromosome segregation, prior septum formation.

26. • Dubarry N, Barre FX: **Fully efficient chromosome dimer resolution in *Escherichia coli* cells lacking the integral membrane domain of FtsK.** *EMBO J* 2010, **29**:597-605.
A soluble truncated variant of FtsK, lacking the transmembrane segment, is sufficient for efficient chromosome dimer resolution. Based on these results, the authors claim that FtsK does not need to transport DNA through a pore formed by its integral membrane domain.

27. Burton BM, Marquis KA, Sullivan NL, Rapoport TA, Rudner DZ: **The ATPase SpoIIIE transports DNA across fused septal membranes during sporulation in Bacillus subtilis.** *Cell* 2007, **131**:1301-1312.

28. Fernandez-Lopez R, Garcillan-Barcia MP, Revilla C, Lazaro M, Vielva L, de la Cruz F: **Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution.** *FEMS Microbiol Rev* 2006, **30**:942-966.

29. Kerr JE, Christie PJ: **Evidence for VirB4-mediated dislocation of membrane-integrated VirB2 pilin during biogenesis of the Agrobacterium VirB/VirD4 type IV secretion system.** *J Bacteriol* 2010, **192**:4923-4934.

30. Middleton R, Sjolander K, Krishnamurthy N, Foley J, Zambryski P: **Predicted hexameric structure of the Agrobacterium VirB4 C terminus suggests VirB4 acts as a docking site during type IV secretion.** *Proc Natl Acad Sci USA* 2005, **102**:1685-1690.

31. Pena A, Ripoll-Rozada J, Zunzunegui S, Cabezon E, de la Cruz F, Arechaga I: **Autoinhibitory regulation of TrwK, an essential VirB4 ATPase in type IV secretion systems.** *J Biol Chem* 2011, **286**:17376-17382.

32. Arechaga I, Pena A, Zunzunegui S, del Carmen Fernandez-Alonso M, Rivas G, de la Cruz F: **ATPase activity and oligomeric state of TrwK, the VirB4 homologue of the plasmid R388 type IV secretion system.** *J Bacteriol* 2008, **190**:5472-5479.

33. Batchelor RA, Pearson BM, Friis LM, Guerry P, Wells JM: **Nucleotide sequences and comparison of two large conjugative plasmids from different Campylobacter species.** *Microbiology* 2004, **150**:3507-3517.

34. Mossey P, Hudacek A, Das A: ***Agrobacterium tumefaciens* type IV secretion protein VirB3 is an inner membrane protein and requires VirB4, VirB7, and VirB8 for stabilization.** *J Bacteriol* 2010, **192**:2830-2838.

35. Seidel R, Dekker C: **Single-molecule studies of nucleic acid motors.** *Curr Opin Struct Biol* 2007, **17**:80-86.

36. Allemand JF, Maier B: **Bacterial translocation motors investigated by single molecule techniques.** *FEMS Microbiol Rev* 2009, **33**:593-610.

37. Saleh OA, Perals C, Barre FX, Allemand JF: **Fast, DNA-sequence independent translocation by FtsK in a single-molecule experiment.** *EMBO J* 2004, **23**:2430-2439.

38. Pease PJ, Levy O, Cost GJ, Gore J, Ptacin JL, Sherratt D, Bustamante C, Cozzarelli NR: **Sequence-directed DNA translocation by purified FtsK.** *Science* 2005, **307**:586-590.

39. Saleh OA, Bigot S, Barre FX, Allemand JF: **Analysis of DNA supercoil induction by FtsK indicates translocation without groove-tracking.** *Nat Struct Mol Biol* 2005, **12**:436-440.

40. Smith DE, Tans SJ, Smith SB, Grimes S, Anderson DL, Bustamante C: **The bacteriophage straight phi29 portal motor can package DNA against a large internal force.** *Nature* 2001, **413**:748-752.

41. Hess H: **Engineering applications of biomolecular motors.** *Annu Rev Biomed Eng* 2011, **13**:429-450.

42. van den Heuvel MG, Dekker C: **Motor proteins at work for nanotechnology.** *Science* 2007, **317**:333-336.

43. • Wendell D, Jing P, Geng J, Subramaniam V, Lee TJ, Montemagno C, Guo P: **Translocation of double-stranded DNA through membrane-adapted phi29 motor protein nanopores.** *Nat Nanotechnol* 2009, **4**:765-772.
In this work, the phi-29 connector protein is redesigned to include regions of hydrophobicity and inserted it in planar lipid bilayers while retaining its function. The engineered protein can act as a conductive channel to translocate double stranded DNA.

44. Vecino AJ, Segura RL, Ugarte-Uribe B, Aguila S, Hormaeche I, de la Cruz F, Goni FM, Alkorta I: **Reconstitution in liposome bilayers enhances nucleotide binding affinity and ATP-specificity of TrwB conjugative coupling protein.** *Biochim Biophys Acta* 2010, **1798**:2160-2169.

45. Llosa M, de la Cruz F: **Bacterial conjugation: a potential tool for genomic engineering.** *Res Microbiol* 2005, **156**:1-6.

46. Fernandez-Gonzalez E, de Paz HD, Alperi A, Agundez L, Faustmann M, Sangari FJ, Dehio C, Llosa M: **Transfer of R388 derivatives by a pathogenesis-associated type IV secretion system into both bacteria and human cells.** *J Bacteriol* 2011, **193**:6257-6265.

47. Schroder G, Schuelein R, Quebatte M, Dehio C: **Conjugative DNA transfer into human cells by the VirB/VirD4 type IV secretion system of the bacterial pathogen Bartonella henselae.** *Proc Natl Acad Sci USA* 2011, **108**:14643-14648.

48. Tzfira T, Citovsky V: **Agrobacterium-mediated genetic transformation of plants: biology and biotechnology.** *Curr Opin Biotechnol* 2006, **17**:147-154.

49. Richard P, Pitard B, Rigaud JL: **ATP synthesis by the F0F1-ATPase from the thermophilic Bacillus PS3 co-reconstituted with bacteriorhodopsin into liposomes. Evidence for stimulation of ATP synthesis by ATP bound to a noncatalytic binding site.** *J Biol Chem* 1995, **270**:21571-21578.

50. Choi HJ, Montemagno CD: **Artificial organelle: ATP synthesis from cellular mimetic polymersomes.** *Nano Lett* 2005, **5**:2538-2542.

51. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X *et al.*: **The potential and challenges of nanopore sequencing.** *Nat Biotechnol* 2008, **26**:1146-1153.

52. Kowalczyk SW, Blosser TR, Dekker C: **Biomimetic nanopores: learning from and about nature.** *Trends Biotechnol* 2011, **29**:607-614.

53. • Stoddart D, Heron AJ, Klingelhoefer J, Mikhailova E, Maglia G, Bayley H: **Nucleobase recognition in ssDNA at the central constriction of the alpha-hemolysin pore.** *Nano Lett* 2010, **10**:3633-3637.

Modification with streptavidin of two specific regions within the lumen of α-hemolysin results in the immobilization of biotinylated ssDNA, which can confer specificity for different nucleotides in protein-nanopore DNA sequencing.

54. Hall AR, Scott A, Rotem D, Mehta KK, Bayley H, Dekker C: **Hybrid pore formation by directed insertion of alpha-haemolysin into solid-state nanopores.** *Nat Nanotechnol* 2010, **5**:874-877.

55. • Majd S, Yusko EC, Billeh YN, Macrae MX, Yang J, Mayer M: **Applications of biological pores in nanomedicine, sensing, and nanoelectronics.** *Curr Opin Biotechnol* 2010, **21**:439-476.

An extensive and complete description of the potential applications of protein pores in medicine and engineering.

56. Korten T, Mansson A, Diez S: **Towards the application of cytoskeletal motor proteins in molecular detection and diagnostic devices.** *Curr Opin Biotechnol* 2010, **21**:477-488.

57. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22**:1658-1659.

58. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792-1797.

 DOI 10.1002/bies.200900164

Numbers on the edges: A simplified and scalable method for quantifying the Gene Regulation Function

Raul Fernandez-Lopez,[1,2] *Irene del Campo,*[1] *Raúl Ruiz,*[1] *Val Lanza,*[1] *Luis Vielva,*[3] *and Fernando de la Cruz*[1]*

[1]Instituto de Biomedicina y Biotecnología de Cantabria (IBBTEC), Universidad de Cantabria-CSIC-IDICAN, Cardenal Herrera Oria s/n, 39011 Santander, Spain
[2]Department of Systems Biology, Harvard Medical School, 200 Longwood Ave, Boston, MA, USA
[3]Departamento de Ingeniería de Comunicaciones, Universidad de Cantabria, Avda. de Los Castros, 39005 Santander, Spain

The gene regulation function (GRF) provides an operational description of a promoter behavior as a function of the concentration of one of its transcriptional regulators. Behind this apparently trivial definition lies a central concept in biological control: the GRF provides the input/output relationship of each edge in a transcriptional network, independently from the molecular interactions involved. Here we discuss how existing methods allow direct measurement of the GRF, and how several trade-offs between scalability and accuracy have hindered its application to relatively large networks. We discuss the theoretical and technical requirements for obtaining the GRF. Based on these requirements, we introduce a simplified and easily scalable method that is able to capture the significant parameters of the GRF. The GRF is able to predict the behavior of a simple genetic circuit, illustrating how addressing the quantitative nature of gene regulation substantially increases our comprehension on the mechanisms of gene control.

Keywords: gene regulation function; network parameterization; plasmid regulatory network; transcriptional regulator

Abbreviations: α, growth rate; β, production rate for repressor-free promoter; β_0, production rate for repressor-bound promoter; 3′-UTR, 3′ untranslated region; AFU, arbitrary fluorescent units; ara, arabinose; aTc, anhydrotetracycline; CFP, cyan fluorescent protein; GFP, green fluorescent protein; GRF, gene regulation function; K, dissociation constant; OD, optical density; RBS, ribosome binding site; X, concentration of transcriptional repressor; X_{ss}, concentration of transcriptional repressor at the steady state; Y, product of the target promoter; YFP, yellow fluorescent protein.

Introduction

In their seminal work on adaptation, Jacob and Monod proposed that there are two kinds of genes.(1) The first kind encodes proteins involved in catalytic or scaffolding activities; these were called structural genes. Genes of the second kind are involved in controlling in what amounts and under what circumstances structural proteins are produced; they were therefore called regulatory genes. Transcriptional regulation was recognized at that time as a core process of cell physiology,(2) and has been a central topic of biological research ever since. Our perspective on how transcriptional activity is regulated has changed substantially over the past 60 years.(3) Today there are two, apparently paradoxical, research trends on transcriptional control. On the one hand, analysis of transcriptional regulation has become ''global''. Originally, transcriptional regulators were thought to control the production of a certain set of proteins in response to a specific stimulus.(4) Now we know that regulators are interconnected in intricate transcriptional networks,(5–9) and that the outcomes of these networks to changes in the environment are often complicated to assess. On the other hand, transcriptional control has become ''individual.'' The original description of the *lac*$^{+/-}$ phenotype consisted of a graph showing how a bacterial population smoothly switched its *lac* phenotype in response to a diauxic shift.(10) This behavior is still referred in textbooks as the canonical example of deterministic control: a culture of *E. coli* grown with glucose and lactose as carbon sources will always exhibit diauxie. However, at the single cell level, a culture of genetically identical cells grown in the same culture medium often exhibits a mixture of *lac*$^{+}$ and *lac*$^{-}$ phenotypes, and the transition between them is purely stochastic.(11) Transcriptional regulation is therefore a probabilistic process, and this fact has many implications for cell adaptation and differentiation.(12–14)

Although this situation might seem contradictory, actually it is not: regulatory networks are holistic, but cells within a population show strong individualism, based on the probabilistic nature of control processes. There is a common dimension in both aspects of transcriptional regulation. The final outcome of a regulatory network or stochastic switch within the cell depends not only on which factors are produced

***Correspondence to:** Dr. F. de la Cruz, C. Herrera Oria s/n, Santander, Cantabria, 39011 Spain
E-mail: delacruz@unican.es

 BioEssays 32:346–355,

but also on the amounts of that factor. To understand how cells self-regulate, we need to describe quantitatively the nature of cell responses.[15] While it remains technically challenging to assess how many mRNAs and proteins a cell produces as a response to a certain input,[16,17] it is feasible to determine how much a promoter activity changes as a consequence of a change in the concentration of its regulators.

The gene regulation function (GRF) indicates the relationship between the intracellular concentration of a given regulator and the transcriptional activity of its target promoter.[18] Ideally, it is a continuous bijective function that renders a certain promoter activity for any given regulator concentration. This function is characteristic for a regulator/promoter pair and each promoter has as many GRFs as the number of transcriptional regulators acting on it. The GRF is an input/output relationship that depends on a number of intermediate biochemical events (regulator oligomerization, binding site searching kinetics, DNA binding, interactions with RNApol, *etc*. . .). Therefore, it can be considered a metafunction: a mathematical abstraction that has no real process counterpart, but represents the convolution of many. This is in fact its main value, since it provides a complete description of the regulator/promoter response space regardless of the underlying mechanism. Disentangling how each intervening molecular process contributes to the final observable GRF is a quintessential problem of transcriptional regulation,[19] and has attracted extensive research efforts. A detailed analysis of the molecular mechanisms of transcriptional regulation is beyond the scope of the present work. We do not discuss how to predict the GRF from our knowledge of each individual molecular step, but instead on how to measure it experimentally. Specifically, we describe how existing methods for accurate expression profiling and controlled protein expression allow a direct inference of the GRF.

The GRF is a powerful tool for systems and synthetic biology. The behavior of a given network depends not only on its topology but also on its parameterization. For relatively large networks it is usually complicated to measure every relevant biochemical parameter, like dissociation constants or cooperativity indexes. Moreover, it is even more complicated to establish how parameters measured *in vitro* reflect the *in vivo* behavior.[20] A systematic method for measuring the GRF simplifies the task of assessing the relevant parameters of a given network, and also provides an *in vivo* framework to interpret *in vitro* results. One of the ultimate goals of synthetic biology is to generate a repertoire of standard biological parts that can be used in the design of higher order devices.[21] One important set of such parts should be a series of well-characterized transcriptionally regulated promoters.[22] Without a precise description of the response dynamics of each component, the integration of complex devices becomes a trial-and-error instead of a design-based process.

Precise measurement of the GRF: The λ-cascade method

The gold standard for measuring the GRF is the method developed by Rosenfeld *et al.*[18] In this work the authors measured the GRF for the P_R promoter of bacteriophage lambda in response to variations in the concentration of protein cI, its transcriptional repressor. Regulator concentration (input) was monitored by translational fusion of gene *cI* to the YFP (yellow fluorescent protein) gene. The transcriptional activity of P_R (output) was determined using a transcriptional fusion with the CFP (cyan fluorescent protein) gene (Fig. 1). The regulator concentration was modulated by making its expression dependent upon an upstream Tet promoter in a $tetR^+$ background strain. Cells were grown in the presence of anhydrotetracycline to achieve full induction of repressor cI. Then the inducer was washed away, so that Tet promoter activity decreased over time to a final off state. Consequently, the levels of cI-YFP were shown to decrease analogically due to protein degradation and dilution in growing cells. The

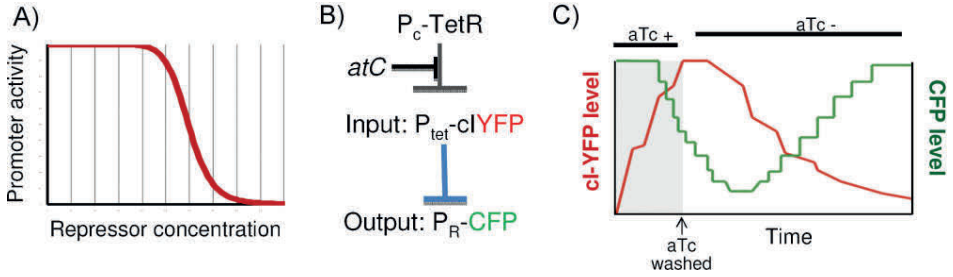


Figure 1. The λ-cascade method developed by Rosenfeld and Alon. **A:** Ideal representation of the GRF. Repressor concentration is decreased analogically (*x*-axis), and the GFP production rate (*y*-axis) is measured as an indicator of target promoter transcriptional activity. **B:** Scheme of the λ-cascade regulatory circuit. The transcriptional repressor cI is tagged by translational fusion with YFP and expressed from the Ptet promoter. TetR (constitutively produced) blocks Ptet expression. Repression can be lifted by the action of aTc. The target promoter (PR) transcriptional activity is tracked by following the CFP fluorescence. **C:** Plot showing cI-YFP (red line, regulator) and CFP (green line, target promoter) dynamics in a single cell. The inducer aTc is kept at high concentration for 2 h (gray area) and then washed away (white area). Concentration of cI starts to decrease as judged by the decrease in YFP fluorescence. Simultaneously, CFP fluorescence increases. Figure modified from Rosenfeld *et al.* [18]

decrease in cI-YFP levels and the subsequent variation of P_R-driven CFP expression were tracked over time for each individual cell by real-time fluorescence microscopy, thus obtaining the GRF for the P_R/cI promoter/regulator pair. The biological meaning of the GRF and its predictive power were demonstrated later,[23] when the behavior of a simple circuit was predicted upon the measurement of the component's GRF. The main advantage of the λ-cascade method was its ability to make single-cell measurements in real time. This allowed an accurate evaluation of the contribution of stochasticity in the dynamics of the system and prevented underestimation of the cooperativity index, a problem that bulk population measurements might cause.[18] As pointed out by the authors, the only caveat was that its applicability to other regulatory systems might be complicated by the refinement of the required measurements.[23] Nevertheless, it served as a powerful proof of concept that it is possible to obtain quantitative models of biological systems with predictive power.

A scalable method for obtaining the GRF: The two-plasmid method

The main obstacle for the applicability of the λ-cascade method is that it requires both extensive genetic manipulation and technically demanding measurements. Regulators need to be translationally fused with fluorescent reporters, which often results in non functional proteins. Quantifying each interaction, even in small networks with a discrete number of arrows, would be challenging. Nevertheless, the λ-cascade method demonstrated that the fundamental bases for resolving the GRF are apparently simple, *i.e.*, the ability to precisely measure promoter activities over time and to control the induction levels of a given regulator. In this work we illustrate how existing methods can lead to a simplified procedure to measure the GRF. We will show that important trade-offs exist between the exactness of the measurement and the scalability of the procedure. In this respect, the method that we propose here must be considered optimized for scalability, as the λ-cascade is optimized for accuracy. In terms of scalability, the λ-cascade paradigm was complemented by efforts to determine quantitative kinetic parameters using bulk culture measurements in model regulatory networks.[24]

We propose a variation of the Rosenfeld method, based on a two-plasmid system (one for expressing the regulator and one for measuring the target promoter response) (Fig. 2). The system consists of an arabinose inducible expression vector (the input plasmid) and a GFP reporter plasmid to precisely determine the response in the target promoter (the output plasmid). This method presents two main differences with respect to the original parameterization of the GRF in the λ-cascade system:

(i) Single-cell measurements are replaced by bulk population averages: This allows the use of simple equipment such as fluorescence plate readers and flow cytometers as measuring instruments, instead of time-lapse microscopy facilities, which are of limited availability. On the other hand, it decreases the precision of the measurement, since the population GRF might show differences to single-cell GRF due to underestimation of cooperativity indexes (see *Trade-offs and limitations in experimental measurements of the GRF* below).

(ii) Regulator concentrations are estimated, not directly measured: the major handicap for extending the λ-cascade

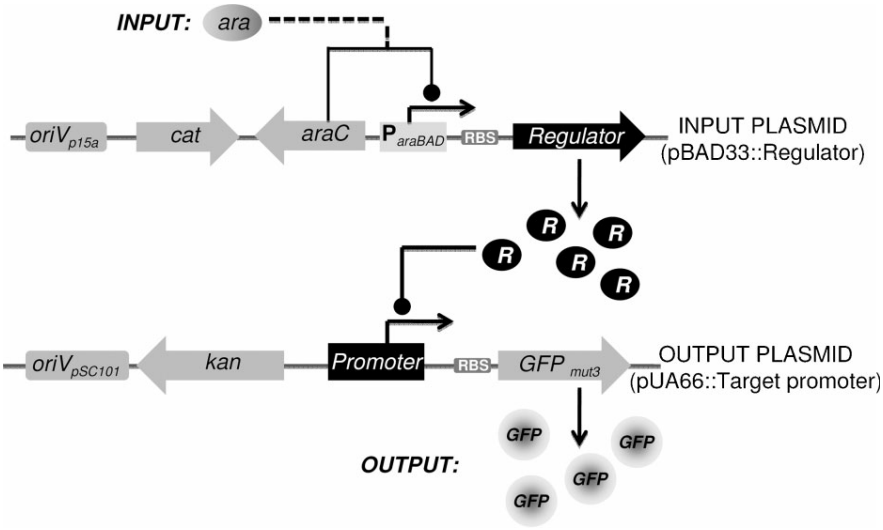


Figure 2. Scheme of the two-plasmid measuring system. The input plasmid consists of an expression vector (pBAD33) in which the desired regulator is cloned under the control of P_{ara}BAD promoter (and thus is arabinose inducible). The output plasmid is a pSC101 derivate (pUA66) engineered for analyzing the activity of a given promoter. The target promoter is cloned so that it drives expression of the GFPmut2 gene.

method to the analysis of other regulatory systems is the need for a fluorescent protein/transcriptional regulator translational fusion. Translational fusions often result in non-functional proteins. To overcome this limitation we avoided the direct measurement of the regulator concentration. For this purpose, the regulator gene was cloned under the inducible promoter of the arabinose operon P_{ARA}BAD. Expression from this promoter is inducible, so increasing concentrations of arabinose in the culture medium produce increasing levels of P_{ARA}BAD expression. Regulators were cloned so they always shared a strong consensus ribosomal binding site and a common 3′-UTR. This was intended to minimize differential translational efficiencies when different regulators were cloned. Since it is still possible that post-translational modifications and protein degradation rates have an important effect on regulator concentrations, measurements were made during steady state. Under steady-state conditions an invariant average concentration of regulator can be assumed. Steady-state levels depend on production and degradation rates. These rates are assumed to be independent from each other, at least under a range of non-extreme conditions (*e.g.*, extremely high protein production could saturate the protein degradation machinery). Although degradation rates are unknown, they can be considered constant for a given regulator on a given genetic background. Variation in the steady-state concentration for two different *ara* induction levels therefore depend only on differences in production rates. Since it is possible to measure the response of P_{ARA}BAD to increasing *ara* concentrations, changes in the regulator concentration can be inferred. Thus, the two-plasmid method measures relative changes: the regulator concentration is changed in a known proportion and the relative response of the target promoter is then measured.

Measuring the GRF I: Determining promoter activities

The first technical requisite for determining the GRF is to precisely measure promoter activities. An additional requirement for this simplified method is that these measurements must be done during steady state. Fluorescent expression profiling allows tracking promoter activity along time with error levels below 20%, accuracy unmatched by any other transcription estimation technique.[25–27] Moreover, it measures promoter activities, not RNA concentrations. This makes the technique insensitive to differences in mRNA processing or degradation of the natural transcript, focusing on the rate that is effectively modulated by repressors or activators. Fluorescent expression profiling for bulk *E. coli* cultures follows the general protocol described in Refs.[24–26] Promoters are cloned in the pUA66 reporter vector, which contains the GFP-mut2 gene downstream of a strong ribosomal binding site. Fluorescence production and cell growth (OD_{600}) are simultaneously tracked so it is possible to determine the GFP/OD change rate. Since GFP-mut2 is a stable protein, the death rate can be assumed to be caused exclusively by dilution in the growing cells. Under these conditions, promoter activity becomes equivalent to the maximum of the time derivative of GFP levels per OD unit [(dGFP/d*t*)/OD].[24,25,27] To ensure that promoter activities were measured during steady-state conditions, several modifications were applied to the original procedure (Box 1). Cells were grown for longer times, starting from low OD levels ($\ll$0.01). This ensures that the culture undergoes a number of replication events that dilute the GFP levels originating from previous growth. Under our conditions, GFP/OD levels reach a plateau in which d(GFP/OD)/d$t = 0$, indicative of the steady-state regime (Fig. 3A). GFP levels at that point equal, by definition, the production rate divided by

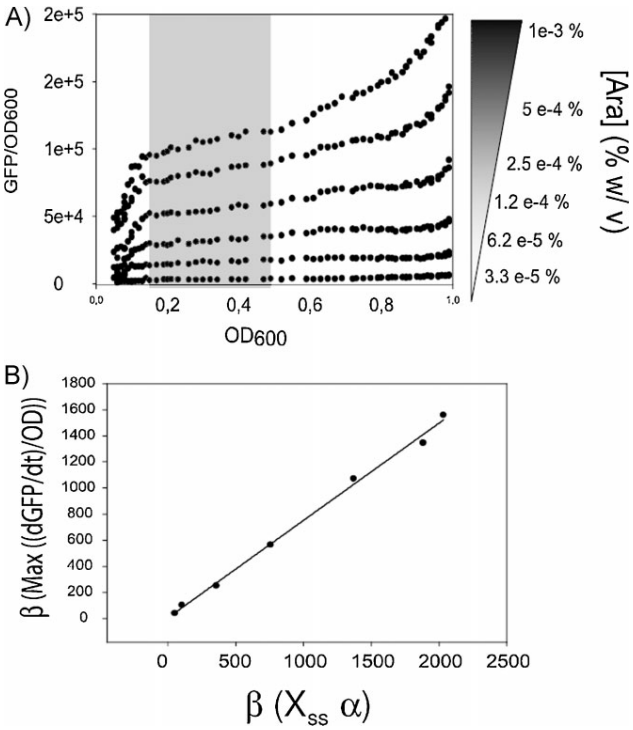


Figure 3. Expression profiling. **A:** Expression profiles of cultures containing pBAD33::GFPmut2 exposed to different arabinose concentrations. Fluorescence levels per OD unit (GFP/OD_{600}) are plotted against cell growth (OD_{600}). Each dot line represents one arabinose concentration, indicated by the scale at the right of the figure. After several rounds of replication, cells reach steady state (gray box). **B:** Comparison between promoter activities (β) estimated by taking the time derivative of the GFP signal per OD unit [*y*-axis, $\beta = \text{Max}$ (dGFP/d*t*/OD)] plotted against the promoter activities calculated as the product of growth rate by steady-state concentration (*x*-axis, $\beta = Xss^{*}\alpha$). Linear regression yielded an *r* square value of 0.99 and a slope of 0.79.

the degradation/dilution rate. Since the protein is stable and the dilution rate can be obtained from the OD_{600} growth curves, promoter activities can thus be directly estimated:

$$\frac{d\left(\frac{GFP}{OD}\right)}{dt} = \beta - \alpha\left(\frac{GFP}{OD}\right) \quad (1.1)$$

$$\left\{\frac{d\left(\frac{GFP}{OD}\right)}{dt}\right\}_{SS} = \beta - \alpha\left(\frac{GFP}{OD}\right)_{SS} = 0 \quad (1.2)$$

$$\beta = \alpha\left(\frac{GFP}{OD}\right)_{SS}$$

where GFP are the fluorescence levels, β the production rate per OD unit, and α the growth rate. This method yields promoter activities equivalent to those described in Refs.[24,25,27](Fig. 3B). The advantage is that, since the steady-state phase can be monitored precisely, many data points can be averaged to estimate steady-state levels more precisely. The reproducibility of these measurements in different experiments was better than SD $<$ 15%.

Box 1:
Experimental protocol for expression profiling

Step 1: Cell growth

Reporter strains were inoculated in M9 Medium plus kanamycin (25 mg/mL), casaminoacids (0.2%) and glycerol (0.5%). Cells were grown for 16 h, at 37°C with vigorous aeration.

Step 2: Measurement

Cultures were diluted 1:10 000 in the same medium in 96-well plates and incubated in a Victor3 fluorimeter at 37°C with orbital shaking for about 6 h. Fluorescence and absorbance were determined for each well every 5 min. To counterbalance evaporation, water was injected after each three steps of fluorescence and absorbance measurement. Water evaporated in the Victor3 multiplate reader at 37°C was 0.28 μL/min/well.

Step 3: Data processing

Absorbance data points had background (data obtained from culture medium with no cells) subtracted. Absorbance values were transformed into OD_{600} equivalents using a calibrated curve between Victor 3 readouts and a regular spectrophotometer (width 1 cm). Fluorescence background was obtained from cells containing the promoter-less plasmid pUA66. Fluorescence background was subtracted from the values obtained for the reporter strains at the same OD_{600} (not necessarily at the same time points). Growth rate (a) was calculated from the OD_{600} data. Fluorescence/OD_{600} was plotted against OD_{600} and the steady-state level was obtained by averaging values during the steady-state.

Measuring the GRF II: Inducing the regulator

The second technical challenge was to induce expression of the transcriptional regulator in a controlled and measurable way. This was achieved by cloning the regulator under the control of an inducible promoter. However, not every inducible promoter is suitable for this purpose. Population measurements of dose/response curves for inducible promoters yield an apparently continuous curve, indicative of an analogical increase in the output as the inducer concentration increases. However, in many cases these curves do not reflect the single-cell dynamics, and the apparent analogical behavior is the product of the averaging of different proportions of cells being either in on or in off state.[28] This is a relevant issue for promoters that respond to a certain chemical modulator (an inducer or a corepressor). On many occasions the intracellular concentration of the modulator is dependent on active transport and the transporter is often activated by the transported metabolite. Transporter activation creates a positive feedback loop that drives system bi-stability.[29] To overcome this issue, we took advantage of the engineered P_{ARA}BAD system developed by Keasling and coworkers, which uses an *E. coli* strain (BW27783) that constitutively expresses *araE*, the arabinose transporter.[30] The P_{ARA}BAD promoter can be induced by adding different concentrations of arabinose to the growing medium. Since the transporter is no longer under the control of the transported metabolite, the positive feedback is broken. The quasi-analogical dynamics of this system was confirmed by measuring the expression levels of *E. coli* BW27783 bearing a P_{ARA}BAD::GFP transcriptional fusion by flow cytometry (Fig. 4A). As shown in Fig. 4, the whole population displaces along the *X*-axis as the *ara* concentration is increased. The population response to increasing *ara* concentrations was measured by expression profiling as indicated in the previous section. Results are shown in Fig. 3B. P_{ARA}BAD::GFP fluorescence values were plotted against the corresponding *ara* concentrations (Fig. 4B). Data were fitted to a Michaelis–Menten function ($R^2 = 0.99$), thus obtaining an expression that infers the P_{ARA}BAD expression level as a function of the arabinose concentration. When a transcriptional regulator is cloned in the pBAD33 vector under expression of P_{ARA}BAD, relative changes in steady-state concentrations can be inferred according to the standard curve obtained for the GFP.

Mathematical framework for determining the GRF

Since the GRF is a metafunction that does not represent a single process, different levels of abstraction can be applied to

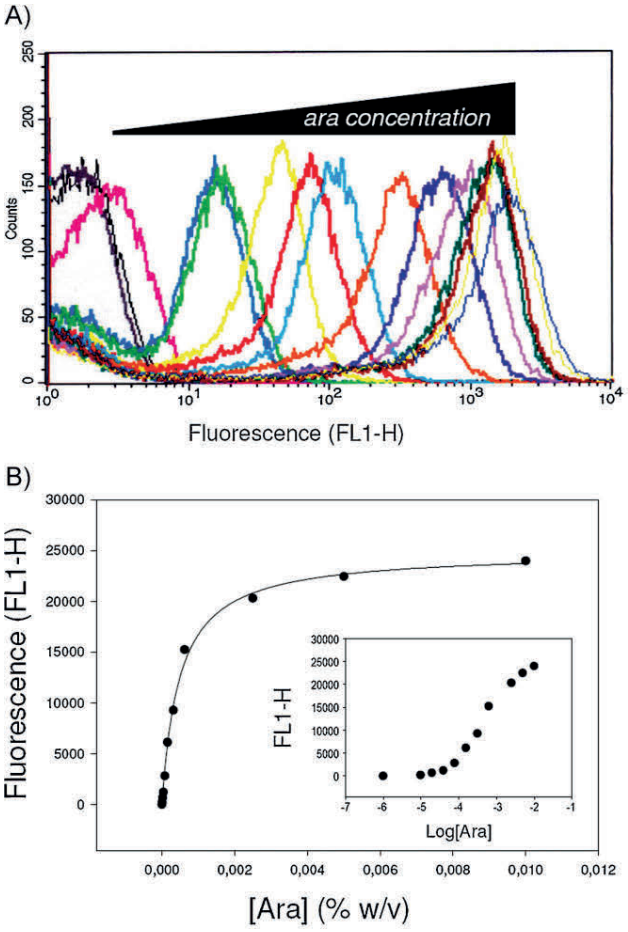


Figure 4. Response of the input plasmid to arabinose induction. **A:** Fluorescence distributions (*x*-axis indicates fluorescence, *y*-axis indicates cell number) obtained by flow cytometry analysis of a culture of *E. coli* BW27783 containing pBAD33::GFPmut2. Each color indicates a different arabinose concentration. **B:** Median fluorescence values obtained from the distributions shown above plotted against the Ara concentration used (in % w/v). Inner chart: arabinose concentrations indicated in log scale (% w/v).

obtain the significant parameters that conform it. The simplest model that, in our hands, successfully captured the GRF dynamics is illustrated in Fig. 5. Three assumptions were made to implement the model: constant growth rate (cells are in steady-state exponential growth), separation of timescales (protein binding to operators and promoters is much faster than transcription), and an invariant relationship between mRNA and protein production (since proteins are translated from the same RBS and 3′-UTR).

Under these assumptions, the dynamics of the product of the target promoter (Y) response to the concentration of

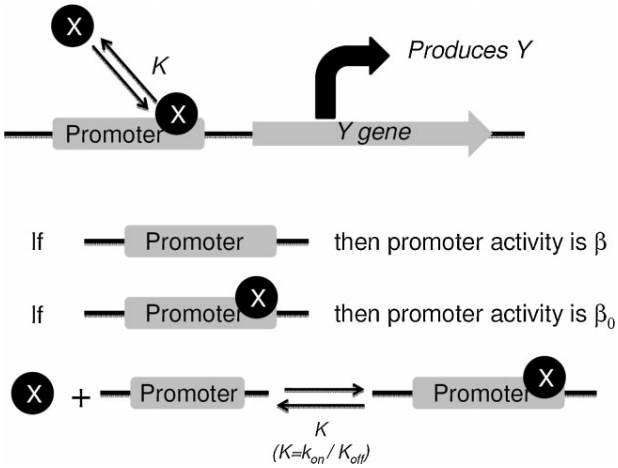


Figure 5. Model for repressor-promoter binding. Reporter gene Y production is assumed to be a one-step process with birth rate β or β_0 depending on the occupancy of the promoter. Binding of repressor X to its target promoter is considered to be in equilibrium, with dissociation constant *K*.

transcriptional repressor (X) follows:

$$\frac{dY}{dt} = \beta \underbrace{\frac{K}{K+X}}_{\substack{X \text{ free} \\ \text{promoter} \\ \text{fraction}}} + \beta_0 \underbrace{\frac{X}{K+X}}_{\substack{X \text{ bound} \\ \text{promoter} \\ \text{fraction}}} - \alpha Y \tag{2.1}$$

where β and β_0 correspond respectively to the activity rates for repressor-free and repressor-bound promoters, K stands for the dissociation constant and α indicates the growth rate of the culture. In steady state ($dY/dt = 0$):

$$Y_{SS} = \frac{\beta K + \beta_0 X_{SS}}{\alpha(K + X_{SS})} \tag{2.2}$$

If the concentration goes to zero, Y reaches a maximum

$$Y_{max}(x) = Y_{(x \to 0)} = \frac{\beta K}{\alpha K} = \frac{\beta}{\alpha} \tag{2.3}$$

If the concentration of repressor is sufficiently high, Y reaches a minimum

$$Y\min(x) = Y_{(x \to \infty)} = \frac{\beta_0 X_{SS}}{\alpha X_{SS}} = \frac{\beta_0}{\alpha}$$
$$\begin{cases} \beta K + \beta_0 X_{SS} \approx \beta_0 X_{SS}; & \beta_0 X_{SS} \gg \beta K \\ \alpha(K + X_{SS}) \approx \alpha X_{SS}; & X_{SS} \gg K \end{cases} \tag{2.4}$$

We define R as the regulatory rank of a given promoter Y for a given repressor X:

$$R_Y^X = \frac{Y_{\max}(x)}{Y_{\min}(x)} \tag{2.5}$$

This parameter is promoter/repressor dependent, and indicates the expression space that a certain repressor can produce when acting upon a target promoter. Therefore, a given promoter with two repressors will show different R values depending on the ability of each repressor to interfere with transcription initiation.

Introducing $Y_{\max} = \beta/\alpha$ and $Y_{\min} = \beta_0/\alpha$ into Eq. (2.2) we obtain

$$Y_{SS} = \frac{Y_{\max} K + Y_{\min} X_{SS}}{K + X_{SS}} \tag{2.6}$$

This expression indicates how the target promoter Y responds to changes in the repressor X, and therefore can be considered directly proportional to the GRF. However, it is probably easier to interpret the relative change of Y from its $Y_{\max}$ levels for each X concentration; this expression also renders a dimensionless value. The transformation follows:

$$\frac{Y_{SS}}{Y_{\max}} = \frac{K + {Y_{\min}}/{Y_{\max}} X_{SS}}{K + X_{SS}} \tag{2.7}$$

$$\frac{Y_{\max}}{Y_{SS}} = \frac{K + X_{SS}}{K + \left({X_{SS}}/{R_Y^X}\right)} \tag{2.8}$$

This expression is fairly intuitive: given an X concentration of repressor, the target promoter is $Y_{\max}/Y$ times repressed. Parameters $Y_{\max}$ and $Y_{\min}$ (and therefore R) can be experimentally measured. For measuring $Y_{\max}$ the promoter activity is determined in a regulator-defective genetic background. $Y_{\min}$ can in turn be measured by inducing the regulator expression to maximal levels [so $X/(X+K) \to 1$]. For this model to work, $Y_{\min}$ must be attained within the experimental rank of concentrations of X. Although the former expression was derived for the case of a transcriptional repressor, the derivation for a transcriptional activator is trivial taking into consideration that:

$$Y_{\min}(x) = Y_{(x=0)} = \frac{\beta K}{\alpha K} = {\beta}/{\alpha}$$
$$Y_{\max}(x) = Y_{(x=\infty)} = {\beta_0}/{\alpha}$$

The former expressions assume a unique binding event of the regulator to the promoter. For the more general case in which n regulator molecules bind the promoter, we can use the Hill approximation to say that the product of n different dissociation constants ($K_1 K_2 K_3 \ldots K_n$) can be made equal to a general constant K to the power of n (K^n). Therefore, the generalized expression follows:

$$\frac{Y_{\max}}{Y_{SS}} = \frac{K^n + X_{SS}^n}{K^n + \left(X_{SS}^n/R_Y^X\right)} \tag{2.9}$$

It should be noticed that a value of $n > 1$ in the GRF does not imply that the number of binding events is precisely n. It only indicates, as in the Hill function, the apparent cooperativity grade in the final observable dynamics.

A case study: Determining the GRF for the transcriptional repressor KorA

To demonstrate the validity of the two-plasmid method, we obtained the GRF for KorA, a transcriptional repressor from the broad host range plasmid R388.[31] KorA is the major transcriptional modulator for the conjugative mating system, a multiprotein secretion system that mediates DNA transfer during plasmid conjugation.[32] KorA represses four promoters in the plasmid, binding to a conserved sequence called the KorA box. KorA controls its own expression, therefore establishing a negative feedback loop.[32] We analyzed the GRF of KorA on its own promoter ($\mathbf{P}_{korA}$) using the procedures described above. Briefly, *korA* gene was PCR-amplified and cloned in expression vector pBAD33, under the control of P_{ARA}BAD promoter. To monitor $\mathbf{P}_{korA}$ expression levels, the promoter was PCR-amplified and cloned in the reporter vector pUA66. Maximal $\mathbf{P}_{korA}$ promoter activity ($Y_{\max}$) was measured by expression profiling of *E. coli* BW27783 containing pUA66::$\mathbf{P}_{korA}$, thus measuring the promoter activity in absence of its cognate regulator. Minimal $\mathbf{P}_{korA}$ promoter activity ($Y_{\min}$) was measured by introducing pBAD33::KorA and inducing KorA expression at saturating concentrations of arabinose. Saturating concentrations were attained at 3.5E-5% w/v. Higher *ara* concentrations impaired cell growth (data not shown). $\mathbf{P}_{korA}$ expression levels (Y) were measured as indicated in Box 1 under different concentrations of arabinose. KorA levels were estimated by running in parallel a set of *E. coli* BW27783 cultures containing pBAD33::GFP-*mut2* subjected to the same *ara* concentrations. $Y_{\max}/Y$ levels were plotted against estimated concentrations of KorA (X) expressed in GFP fluorescence equivalents (Fig. 6A), and data were adjusted to Eq. (2.9) using the Levenberg–Marquardt algorithm for non-linear regression. Regression yielded $K = 22 \pm 10$ arbitrary fluorescent units (AFU). The regulatory rank was $R = 33$, indicating that KorA was able to repress $\mathbf{P}_{korA}$ 33-fold from its native promoter activity. This value was in accordance to the experimental determination of R ($Y_{\max}/Y_{\min}$). The cooperativity index was $n = 1.3 \pm 0.4$, indicating a slight cooperative effect of KorA in $\mathbf{P}_{korA}$

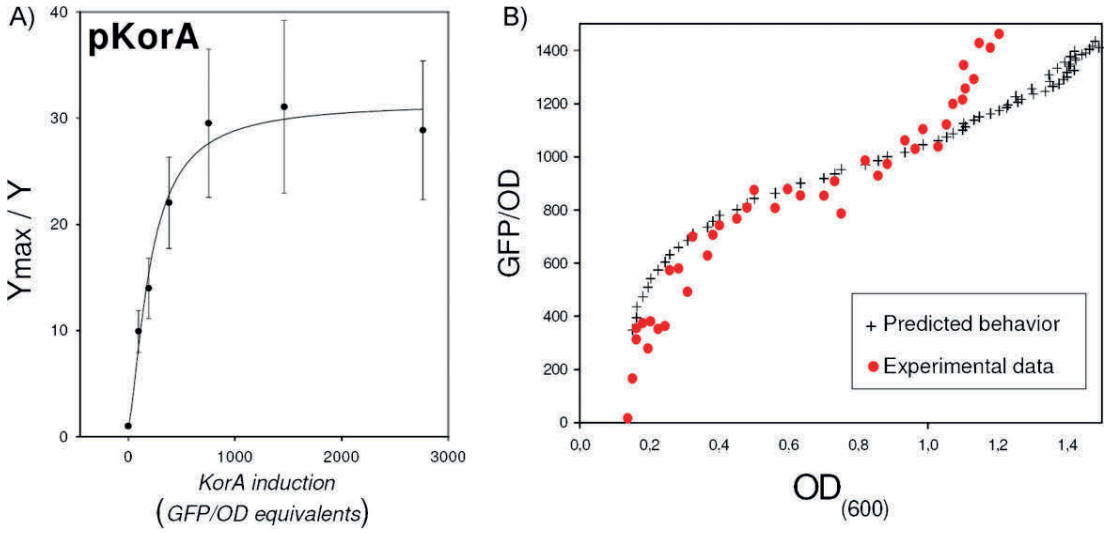


Figure 6. Measurement and performance of the GRF for KorA regulator acting on $P_{kor}A$ promoter. **A:** Repression ratio (Y_{max}/Y) of $P_{kor}A$ promoter (*y*-axis) as a function of apparent KorA repressor concentration (*x*-axis). KorA concentrations is modified by inducing pBAD33::korA with different levels of arabinose, and apparent KorA concentrations are calculated from calibrating curves (pBAD33::GFP) obtained in parallel. Maximal steady-state fluorescence level achieved by $P_{kor}A$ promoter (Y_{max}, [KorA] = 0) is divided by steady-state fluorescence levels achieved by $P_{kor}A$ (Y), and plotted in ordinates. Error bars indicate the standard deviation of eight independent experiments. Data were fitted to Eq. (2.9) with $R^2 = 0.98$. **B:** A comparison between the predicted behavior of $P_{kor}A$-KorA feedback loop (black crosses) with an experimental measurement of $P_{kor}A$-KorA dynamics (red dots). Experimental measurements were done in *E. coli* BW27783 cultures containing plasmid R388 (that contributes the $P_{kor}A$-KorA feedback loop) and pUA66: $P_{kor}A$-GFP (reporter plasmid that infers $P_{kor}A$ expression levels). Predicted dynamics were obtained from Eq. (2.1) using GFP/OD measured at $OD_{600} = 0.2$ as initial X value. All parameters in the simulation were determined experimentally [K and n obtained from data from (A), growth rate (α) obtained from growth curve and production rates β and β_0 obtained from expression profiles without KorA and at maximal KorA concentration, respectively].

transcriptional activity. The R^2 value of the adjusted curve was $R^2 = 0.980$.

To check the predictive power of the KorA/$\mathbf{P}_{korA}$ GRF we simulated the time behavior of the negative feedback loop. We measured the behavior of this feedback loop experimentally by introducing plasmid R388 (that carries KorA under $\mathbf{P}_{korA}$ expression) into an *E. coli* BW27783 strain containing also pUA66::$\mathbf{P}_{korA}$ and profiling the expression levels along time. We fed the computational simulation with the initial fluorescence value experimentally obtained at $OD_{600} = 0.2$. We then compared how the theoretical prediction and the experimental measurement behave as the culture grew (Fig. 6B). The OD_{600} start point was chosen because fluorescence values for $OD_{600} < 0.2$ were found to be extremely noisy. As shown in the figure, the computational prediction based on the measured GRF follows the dynamics of the experimental data until the cell culture enters stationary phase ($OD_{600} = 1$). At this point both curves separate sharply, indicating that the GRF performs well as a dynamical predictor for cells in constant growth but not in stationary cultures. This was to be expected since one of the assumptions of our model was that cells were in exponential growth.

Trade-offs and limitations in experimental measurements of the GRF

As we have shown, the two-plasmid method is able to extract sufficient information to determine the GRF using simple measurement techniques and a basic mathematical treatment. It is based on two previously reported experimental procedures[25,30] and might be useful for systematic parameterization of relatively large networks. However, as stated above, several caveats should be taken into consideration when applying this method. The fundamental one refers to bulk averaging in contrast with single-cell measurements. Cell heterogeneity is a possible source of conflict: population measurements tend to underestimate cooperativity indexes due to convolution of the regulator concentration distribution and the promoter response.[33] The two-plasmid method alleviates this problem using an induction system that exhibits a unimodal distribution (Fig. 4A). However, since both regulator concentration and promoter activity distributions are averaged, the method will still underestimate the steepness of the curve. To overcome this problem, a simultaneous measurement at the single-cell level of both regulator concentration and promoter activity must be carried

out. Flow cytometry can determine both distributions if the transcriptional regulator and target promoter are tagged with fluorescent proteins with enough spectral separation and similar maturation timescales. However, due to the small size and bacillary shape of *E. coli*, this technique is unsuitable for assessing bacterial size. As a consequence flow cytometry retrieves a convoluted distribution of ill-defined cell sizes and fluorescence values. If a higher level of accuracy is desired, measurements can be made by fluorescence microscopy, which would render an equivalent procedure to the λ-cascade method. Scalability and precision in the measurement are therefore inversely related. According to the size of the network and the level of refinement needed, the researcher will have to decide which method to use. In that respect, further research is needed to assess the comparative performances of both methods.

Common to any GRF determination method is the need for a regulator-defective genetic background. In the λ-cascade, the authors used a phage protein so the host bacteria were naturally cI^{-}.[18] Similarly, we have used a plasmid-borne regulator, so the host strain is naturally KorA defective. If one wants to determine the GRF for a chromosomally encoded regulator, a knockout background is always needed, since the full possible range of regulator concentrations must be sampled. If the two-plasmid method described here were to be used, this mutation must be introduced into the *E. coli* BW27783 strain, since this genetic background is needed for appropriate induction of the P_{ara}BAD promoter.[30] For that purpose, extensive knockout libraries that allow P1 transduction, like the Keio collection,[34] can be used.

Conclusions and prospects

Despite its abstract nature, the GRF has proven predictive power.[23] Although we have focused on methods used for estimating the GRF in prokaryotes, similar approaches have been successfully employed in eukaryotes, where chromatin remodeling was found to be an essential player determining the GRF.[35] These results illustrate the possibility of calculating the output of a certain regulatory network if we know the transfer functions of its individual components.

Besides being a powerful tool for network analysis, the GRF opens the possibility to analyze a burning question in synthetic biology: How do promoters integrate the action of two or more regulators? It might happen that the signals deployed from two different transcription factors to the same target promoter act independently from each other, so the final function can be calculated as the sum of the individual contributions. This is usually referred to as orthogonality between two inputs. It is also possible that both regulators display mutual interactions, so the function of regulators A and B is not equal to the sum of their independent GRFs. Since the GRF is a function, and function orthogonality is strictly defined, it can be used to test for this principle. It has been demonstrated that the multiple input functions for sugar metabolism in *E. coli* can sometimes be separated,[36] so it seems that, at least in some cases, biological control systems are orthogonal. The circumstances and the kinds of promoter architectures with which orthogonality in the GRF is allowed still need to be elucidated.

Parameterizing global regulatory networks is a formidable task, and massive parallel efforts are needed. In that sense, the development of a standard unit system for promoter activities is urgently needed so that results obtained by different methods and in different laboratories can be compared. Efforts are being made to obtain absolute scales based on the effective counting of mRNAs and proteins,[17] and to establish a set of reference promoters.[37] Developing standard scales would be useful for Systems Biology and is probably essential for the development of standardized parts in synthetic biology. It is difficult to envisage how a new engineering discipline can be founded without an operational metric system.

Another crucial question is to determine what forces shape the GRF. Cost-benefit theory for the GRF identifies three distinguishable levels of selection: efficiency, economy, and noise adaptation.[38] Efficiency indicates the ability of the GRF to sense and respond in an appropriate timescale to changes in the input. Economy refers to the ability to produce responses only when the benefits for the bacterial fitness exceed the cost of producing the response. Noise adaptation indicates the capacity of the GRF to suppress or enhance cell-to-cell variation to achieve efficient and economic responses. Interestingly, there are severe trade-offs between these three levels.[38] As an example, a strict GRF with a high cooperativity index would be efficient (the cell would respond according to a certain threshold) and economic (no suboptimal responses under the threshold), but it would be extremely vulnerable to input noise, since noise transmission increases linearly with cooperativity indexes.[38] Network topology comes to play at this point, since network motifs such as negative feedback and feed-forward loops can decrease noise.[38–40] In other cases, in which input fluctuations are fast or undetectable (so the cell has to rely on "blind" decisions), noise in gene expression can not be deleterious but beneficial.[41] Selective landscapes will be the ultimate force shaping the GRF,[38] a function in which topological properties, noise in gene expression, and response time requirements concur. Therefore, the GRF exceeds any utilitarian view for design purposes: it is the keystone of gene regulation. It establishes how inputs and outputs are computed by the cell, and how these responses are modified and evolve by environmental constraints.

BioEssays 32:346–355, © 2010 Wiley Periodicals, Inc.

Acknowledgments: Work in the FdlC laboratory was supported by grants BFU2008-00995/BMC (Spanish Ministry of Education), RD06/0008/1012 (RETICS research network, Instituto de Salud Carlos III, Spanish Ministry of Health) and LSHM-CT-2005_019023 (European VI Framework Program).

References

1. **Jacob F, Monod J.** 1959. Genes of structure and genes of regulation in the biosynthesis of proteins. *C R Hebd Seances Acad Sci* **249**: 1282–4.
2. **Umbarger HE.** 1964. Intracellular regulatory mechanisms: regulation in multicellular forms may be an elaboration upon the pattern evolved in microorganisms. *Science* **145**: 674–9.
3. **Balleza E, Lopez-Bojorquez LN, Martinez-Antonio A,** *et al.* 2009. Regulation by transcription factors in bacteria: beyond description. *FEMS Microbiol Rev* **33**: 133–51.
4. **Stent GS.** 1964. The Operon: on its third anniversary. Modulation of transfer RNA species can provide a workable model of an operator-less operon. *Science* **144**: 816–20.
5. **Shen-Orr SS, Milo R, Mangan S,** *et al.* 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* **31**: 64–8.
6. **Ihmels J, Friedlander G, Bergmann S,** *et al.* 2002. Revealing modular organization in the yeast transcriptional network. *Nat Genet* **31**: 370–7.
7. **Kashtan N, Itzkovitz S, Milo R,** *et al.* 2004. Topological generalizations of network motifs. *Phys Rev E Stat Nonlin Soft Matter Phys* **70**: 031909.
8. **Madan Babu M, Teichmann SA.** 2003. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res* **31**: 1234–44.
9. **Gutierrez-Rios RM, Rosenblueth DA, Loza JA,** *et al.* 2003. Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res* **13**: 2435–43.
10. **Jacob F, Monod J.** 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**: 318–56.
11. **Choi PJ, Cai L, Frieda K,** *et al.* 2008. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science* **322**: 442–6.
12. **Paulsson J.** 2004. Summing up the noise in gene networks. *Nature* **427**: 415–8.
13. **Elowitz MB, Levine AJ, Siggia ED,** *et al.* 2002. Stochastic gene expression in a single cell. *Science* **297**: 1183–6.
14. **Suel GM, Garcia-Ojalvo J, Liberman LM,** *et al.* 2006. An excitable gene regulatory circuit induces transient cellular differentiation. *Nature* **440**: 545–50.
15. **Kim HD, Shay T, O'Shea EK,** *et al.* 2009. Transcriptional regulatory circuits: predicting numbers from alphabets. *Science* **325**: 429–32.
16. **Raj A, van Oudenaarden A.** 2009. Single-molecule approaches to stochastic gene expression. *Annu Rev Biophys* **38**: 255–70.
17. **Golding I, Paulsson J, Zawilski SM,** *et al.* 2005. Real-time kinetics of gene activity in individual bacteria. *Cell* **123**: 1025–36.
18. **Rosenfeld N, Young JW, Alon U,** *et al.* 2005. Gene regulation at the single-cell level. *Science* **307**: 1962–5.
19. **Segal E, Widom J.** 2009. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet* **10**: 443–56.
20. **Cases I, de Lorenzo V.** 2005. Promoters in the environment: transcriptional regulation in its natural context. *Nat Rev Microbiol* **3**: 105–18.
21. **Endy D.** 2005. Foundations for engineering biology. *Nature* **438**: 449–53.
22. **Hasty J, McMillen D, Collins JJ.** 2002. Engineered gene circuits. *Nature* **420**: 224–30.
23. **Rosenfeld N, Young JW, Alon U,** *et al.* 2007. Accurate prediction of gene feedback circuit behavior from component properties. *Mol Syst Biol* **3**: 143.
24. **Ronen M, Rosenberg R, Shraiman BI,** *et al.* 2002. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci U S A* **99**: 10555–60.
25. **Zaslaver A, Bren A, Ronen M,** *et al.* 2006. A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nat Methods* **3**: 623–8.
26. **Kalir S, McClure J, Pabbaraju K,** *et al.* 2001. Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science* **292**: 2080–3.
27. **Setty Y, Mayo AE, Surette MG,** *et al.* 2003. Detailed map of a *cis*-regulatory input function. *Proc Natl Acad Sci U S A* **100**: 7702–7.
28. **Ozbudak EM, Thattai M, Lim HN,** *et al.* 2004. Multistability in the lactose utilization network of *Escherichia coli*. *Nature* **427**: 737–40.
29. **Santillan M, Mackey MC, Zeron ES.** 2007. Origin of bistability in the lac Operon. *Biophys J* **92**: 3830–42.
30. **Khlebnikov A, Datsenko KA, Skaug T,** *et al.* 2001. Homogeneous expression of the P(BAD) promoter in *Escherichia coli* by constitutive expression of the low-affinity high-capacity AraE transporter. *Microbiology* **147**: 3241–7.
31. **Fernandez-Lopez R, Garcillan-Barcia MP, Revilla C,** *et al.* 2006. Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution. *FEMS Microbiol Rev* **30**: 942–66.
32. **Seubert A, Hiestand R, de la Cruz F,** *et al.* 2003. A bacterial conjugation machinery recruited for pathogenesis. *Mol Microbiol* **49**: 1253–66.
33. **Cluzel P, Surette M, Leibler S.** 2000. An ultrasensitive bacterial motor revealed by monitoring signaling proteins in single cells. *Science* **287**: 1652–5.
34. **Baba T, Ara T, Hasegawa M,** *et al.* 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**: 2006 0008.
35. **Kim HD, O'Shea EK.** 2008. A quantitative model of transcription factor-activated gene expression. *Nat Struct Mol Biol* **15**: 1192–8.
36. **Kaplan S, Bren A, Zaslaver A,** *et al.* 2008. Diverse two-dimensional input functions control bacterial sugar genes. *Mol Cell* **29**: 786–92.
37. **Kelly JR, Rubin AJ, Davis JH,** *et al.* 2009. Measuring the activity of BioBrick promoters using an *in vivo* reference standard. *J Biol Eng* **3**: 4.
38. **Kalisky T, Dekel E, Alon U.** 2007. Cost-benefit theory and optimal design of gene regulation functions. *Phys Biol* **4**: 229–45.
39. **Becskei A, Serrano L.** 2000. Engineering stability in gene networks by autoregulation. *Nature* **405**: 590–3.
40. **Paulsson J, Ehrenberg M.** 2001. Noise in a minimal regulatory network: plasmid copy number control. *Q Rev Biophys* **34**: 1–59.
41. **Wolf DM, Vazirani VV, Arkin AP.** 2005. Diversity in times of adversity: probabilistic strategies in microbial survival games. *J Theor Biol* **234**: 227–53.