Elucidating the Molecular Factors Implicated in the Persistence and Evolution of Transferable Antibiotic Resistance

Porse, Andreas

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Elucidating the Molecular Factors Implicated in the **Persistence and Evolution of Transferable Antibiotic Resistance**

Andreas Porse

PhD Thesis
November 2017
The Novo Nordisk Foundation Center
for Biosustainability, DTU

Preface

This PhD was conducted from 1th December 2014 to 30th November 2017. The work was carried out at the Technical University of Denmark (DTU); first at the institute of Systems Microbiology and later at the Novo Nordisk Foundation Center for Biosustainability (CFB). Throughout the work, I was supervised by professor Morten Otto Alexander Sommer and co-supervised by professor Lars Jelsbak and PhD Christian Munck.

The PhD was funded by the Lundbeck Foundation, the European Union and the Danish Council for Independent Research.

Andreas Porse, November 2017.

Novo Nordisk Foundation Center for Biosustainability,

DTU

The role of the infinitely small in nature is infinitely great.
- Louis Pasteur

Abstract

Being the most diverse and abundant domain of life, bacteria exemplify the remarkable ability of evolution to fit organisms into almost any imaginable niche on the planet. Although the capacity of bacteria to diversify and adapt is fundamental to natural ecosystems and modern biotechnology, the same adaptive mechanisms constantly threaten human health. Less than a century ago, infectious disease was among the most common causes of mortality, but luckily this situation was drastically improved with the introduction of vaccination and effective antimicrobial drugs. Unfortunately, this situation is changing with the rapid emergence of multidrug resistant bacteria that do not respond to our current treatments. This process is to a large extent driven by gene exchange that allows bacteria to rapidly acquire ready-made adaptive features. The aim of this thesis has been to understand the adaptive mechanisms governing the dynamics of bacterial gene-sharing. Specifically, the focus has been on antibiotic resistance genes and their genetic vectors due to the profound implications of these genetic elements in human health.

To observe the extend and impact of gene transfer events in a highly relevant natural environment, we looked into the genomes of *Escherichia coli* longitudinally sampled from the infant gut over the first year of life. Sequence analysis revealed a high degree of genomic plasticity, with frequent gene acquisition and loss events. While the acquisition of new genetic material is often deleterious, we show that plasmids encoding resistance and virulence factors may indeed be stably maintained in the gut despite imposing a measurable fitness cost to their bacterial hosts *in vitro*.

In two studies investigating the stability of genetic elements, we zoom in on the molecular mechanisms enabling conflict resolution between incoming genetic elements and naïve recipient genomes. In both studies, the burden of initially costly genetic elements is ameliorated via adaptive evolution over time. In the case of a large multi-drug resistance plasmid, adaptation happens through IS26 mediated deletions of costly genes that (collaterally) sacrifice the transfer proficiency of the plasmid. For the industrially relevant mevalonate production pathway, we observe similar population-level loss dynamics. Using ultra-deep sequencing we show that the cost-attenuated pathway variants are interrupted by different IS-element insertions that enrich over time due to the fitness benefit of production loss. For both studies, the compensatory activity depends on the host background, and we suggest measures that can harness evolution to increase genetic stability of the costly production pathway.

The final study of this thesis investigates the phenotypic effects of expressing 200 antibiotic resistance

genes in *E. coli*. As the currency of evolution, genes are subject to selection at different levels that may promote or limit their success when transferred to a new host. Through sequence analysis and experimental interrogations, we suggest that functional constrains, rather than sequence composition, is the main challenge that acquired genes encounter when transferred across phylogeny.

The work conducted in this thesis provides novel insight into the persistence and evolution of highly relevant genetic elements *in vitro*, *In vivo* and *in situ*. The conclusions shed light on fundamental evolutionary questions of genome dynamics and bacterial adaptation, which may ultimately improve our ability to predict and prevent the spread of antibiotic resistance and guide the engineering of robust biological systems.

Dansk sammenfatning

Trods deres beskedne størrelse repræsenterer bakterier den største mangfoldighed af liv på jordkloden. Bakteriernes enorme diversitet skyldes deres evne til hurtigt at tilpasse sig nye omstændigheder. Denne evolutionære kapacitet er vigtig for jordens økosystemer og udnyttes til produktion af billige, og mere miljørigtige, kemikalier af den bioteknologiske of farmaceutiske industri. Ligesom tilpasning til andre miljøer, har nogle bakterier udviklet evnen til at forårsage infektioner hos mennesker, og før opfindelsen af antibiotika var infektionssygdomme blandt den hyppigste dødsårsag. Desværre oplever vi en udvikling der igen øger truslen fra bakterielle infektioner, fordi mange bakterier er blevet resistente overfor de antibiotika som før blev brugt til at bekæmpe dem. Dette skyldes tilpasningen af bakterier til vores forbrug af antibiotika, og en stor del af svaret på deres hurtige tilpasning skal findes i deres evne til at dele mobile DNA-elementer med hinanden.

I denne afhandling har jeg undersøgt de molekylære faktorer, som er med til at afgøre spredningen og tilpasningen af mobile DNA-elementer involveret i antibiotikaresistens hos sygdomsfremkaldende bakterier.

I de første to studier har vi undersøgt omfanget og indflydelsen af mobile genetiske elementer i tilpasningen af *Escherichia coli* bakterier isoleret fra spædbørns tarmflora i løbet af det første leveår. Via genomiske DNA-analyser detekterede vi hyppig overførsel, samt tab, af plasmider og bakteriofag-elementer. Dette var tilfældet ligegyldigt om barnet modtog antibiotika eller ej. Hvor nogle plasmider, som bidrager til virulens eller antibiotikaresistens kan overleve over længere tidsperioder i et tarmmiljø

uden direkte nytteværdi for værtsbakterien, observerede vi at et plasmid involveret i antibiotikaresistens tilmed kunne øge værtsbakteriens overlevelse i tarmen selv uden tilstedeværelsen af antibiotika.

I de næste to studier zoomer vi ind på plasmider som evolutionære enheder og belyser nogle af de mekanismer, som afgør deres tilpasning til nye værtsbakterier. I det første studie har vi undersøgt, hvordan et multiresistens plasmid tilpasser sig til kliniske *Klebsiella pneumoniae* og *Escherichia coli* isolater. Her ser vi at plasmidet kan opnå højere stabilitet ved, via rekombination af IS-elementer, at fjerne plasmidregioner involveret i dets horisontale overførsel. I det andet studie viser vi at IS-medierede adaptive mekanismer også har betydning for et syntetisk plasmid involveret i bioteknologisk produktion af mevalonat. Her omgås cellernes dannelse af toksiske produkter ved at indsætte IS-elementer fra kromosomet i produktionsgenerne. I begge studier, observerede vi en væsentlig sammenhæng mellem stammebaggrund og omfanget af kompenserende begivenheder.

Formålet med det sidste studie var at belyse de faktorer, som gør at nogle gener har sværere ved at sprede sig end andre. Til dette formål benyttede vi 200 af de mest hyppige antibiotikaresistensgener fundet i offentlige resistensgen-databaser. Vi viser eksperimentelt at sekvenssammensætningen har lille betydning, men at funktion af det protein som genet koder for, spiller en større rolle for resistensfunktion of biologisk fitness. Vi ser at specielt proteiner, som er afhængige af cellulære komponenter, kan være begrænsede når de overføres imellem fjernt beslægtede bakterier.

Arbejdet i denne afhandling er primært udført på bakterier og gener med relevans for antibiotikaresistens, men de samme basale mekanismer bidrager også generelt til livets udvikling på det molekylære plan. Det er min forhåbning at resultaterne vil bidrage til en dybere forståelse af bakteries tilpasning, som på sigt kan give os et forspring i kampen mod antibiotikaresistente bakterier, samt bidrage til en mere rationel konstruktion af produktionsorganismer. Bidrag til disse områder er hårdt tiltrængte, hvis vi ønsker at skabe et samfund som er både medicinsk og miljømæssigt bæredygtigt.

List of publications

Overview of (published and unpublished) scientific articles resulting from the work of this thesis.

* Denotes equal contribution.

I. Andreas Porse*, Heidi Gumpert*, Jessica Z. Kubicek-Sutherland, Nahid Karami, Ingegerd Adlerberth, Agnes E. Wold, Dan I. Andersson, Morten O.A. Sommer (2017).
Genome Dynamics of Escherichia coli during Antibiotic Treatment: Transfer, Loss, and Persistence of Genetic Elements In situ of the Infant Gut.
Frontiers in Cellular and Infection Microbiology, Vol 7.

II. Heidi Gumpert*, Jessica Z. Kubicek-Sutherland*, Andreas Porse*, Nahid Karami*, Christian Munck, Marius Linkevicius, Ingegerd Adlerberth, Agnes E. Wold, Dan I. Andersson, Morten O.A. Sommer (2017). *Transfer and persistence of a multi-drug resistance plasmid in situ of the infant gut microbiota in the absence of antibiotic treatment.*
Frontiers in Microbiology, Vol 8.

III. Andreas Porse, Kristian Schønning, Christian Munck and Morten O. A. Sommer (2016).
Survival and evolution of a large multidrug resistance plasmid in new clinical bacterial hosts.
Molecular Biology and Evolution. Vol: 33, issue: 11, pages: 2860-2873.

IV. Peter Rugbjerg, Nils Myling-Petersen, Andreas Porse, Kira Sarup-Lytzen, Morten O.A. Sommer (2017).
Diverse genetic error modes constrain large-scale bio-based production.
Accepted for publication in Nature Communications.

V. Andreas Porse, Thea S. Schou, Christian Munck, Mostafa M. H. Ellabaan, Morten O.A. Sommer (2017).
Biochemical mechanisms determine the functional compatibility of heterologous genes.
Accepted for publication in Nature Communications.

Published work that is not included in this thesis:

Ida Lauritsen*, Andreas Porse*, Morten O.A. Sommer, Morten Nørholm (2017).
A versatile one-step CRISPR-Cas9 based approach to plasmid-curing.
Microbial Cell Factories. Vol 16.

Acknowledgements

Conducting a PhD is indeed an exciting adventure of scientific and personal development. Expanding the edge of human knowledge obviously entails lots of hard work and stressful times, but the rewards are unique and manifold. One of the greatest joys of science is sharing knowledge, collaborating or giving and receiving help from wonderful colleagues.

Throughout my PhD I have had the pleasure of interacting with many inspiring scientists, without whom the work of this thesis would not be possible.

First, I would like to dedicate a huge "thank you" to **Christian** who introduced me to the fascinating field of plasmids and antibiotic resistance (which primarily took place over a beer in the shady bars of Copenhagen…). You have been, and still are, a profound mentor and a friend that I truly appreciate.

Morten, you opened my eyes to the challenge of antibiotic resistance and I am immensely grateful for getting the opportunity to write this thesis in your group. I appreciate the ambitious and optimistic atmosphere you create, and I thrive with the freedom and trust you provide.

I am really grateful to all of the members of the Sommerlab for being good colleagues and friends! Of the current members, I would like to dedicate a special thanks to **Peter**, for being an uncompromised scientist - you have a rare eye for the details and a curios mind that has made all our discussions and collaborative efforts truly enjoyable! **Gitte**, you deserve a huge thank you for fixing all the administrative stuff that can drain the energy of even the most dedicated scientists. I would also like to dedicate special thanks to the antibiotic minority of the lab: **Mari**, **Mostafa**, **Lejla, Maria-Anna, Scott** and **Leonie. Mari**, thank you for taking care of everything and keeping the order of the lab – and for our interesting food discussions! **Mostafa**, you big-hearted computer wiz! Thank you for always carrying a smile and a helping hand. **Lejla**, your efforts always surprise me and you constantly push my definition of "high-throughput"! **Maria-Anna** thanks for answering my inferior Linux questions – what's actually going on with that blood you and **Eugene** took from me? **Scoot**, my favorite twitter guy! I am sure you will do a great job on your PhD. **Leonie**, you are among the most curious and good-hearted scientists that I know. You have always shown interest in my projects, asking good questions, and you deserve special thanks for several occasions of critical proof reading; including that of this thesis!

Among the rest of the people in the lab, my fellow PhD-office mates deserve special mentioning: **Carina, Ruben, Kira, Michael** and **Gonzalo**. **Carina**, I am glad you joined our office and are taking good care of us

with Portuguese candy! **Ruben** – my favorite Mexican guy! Thank you for all our interesting discussions on synthetic biology and CRISPR, for being so curious and socially including, and for teaching me about Mexico and Valentina! **Kira**, thank you for suddenly appearing (feet or eyes) above the table and sparking interesting (sometimes far-out) scientific discussions and other office fun! **Michael** – you program your slideshows (everything?) and you do it well. You are my go-to guy for bioinformatic discussions and graphical suggestions and I hope you're still up for a beer after starting your own lab at MIT one day! **Gonzalo** – galán! Thank you for watering my chili-plant while I am writing this and for being who you are…

I have enjoyed numerous discussions with the metabolic engineers of the lab. **Mareike**, thank you for interesting discussions on the engineering of natural plasmids for stability and for proving pFREE's ability to cure endogenous replicons. **Felipe**, my midnight-cloning lab-mate! I appreciate the smell (beer) of the "complex medium" that you are brewing. **Sang-Woo**, I hope your chilies are growing and your phages behaving! **Eric** and **Jakob**, I appreciate your interest in my F-plasmids and "super colonizers" for your secret experiments!

Among past members of the lab, I really appreciate the fun times with **Rachel**, the great personality and Chinese lessons from **Lumeng** as well as **Hans** for his great enthusiasm and scientific creativity! I have also greatly valued the company of **Nils** on many occasions; in and outside the lab. I hope we will have many more good fishing trips and lengthy discussions over a beer or a few cups of Winters Glow!

During my time in the Sommerlab I have been lucky to work with and supervise several aspiring scientists as part of their studies, including: **Minna**, **Carola**, **Camilla**, **Sara, Christoffer** and **Thea**. A particular thank you goes to **Thea** for your massive cloning efforts and uplifting social nature that I, and everyone at CFB, will surely remember you for! **Carola**, you are still here and I really appreciate your dry sense of humor and the occasional use of your bench for my over-flow plates.

Apart from people associated with the Sommerlab and Hvidovre Hospital (**Heidi** and **Kristian**), I am grateful to have met many inspiring people at DTU; including those of CFB and 301. From 301 I am especially thankful to my co-supervisor **Lars** and his group members and from CFB I devote a special thanks **Ida** for our joint plasmid-curing endeavors, for being much smarter than she realizes, and for so much more.

I am also very grateful to **Dan Andersson** for our collaborations and for having me in his group during my external stay in Uppsala. I would like to thank the people of the D7:3 corridor for being so welcoming

and for making me feel at home in your very inspiring research environment. Here, I am especially thankful to **Marius** for being extremely helpful and getting me in place.

Apart from my colleagues I am lucky to have a long list of friends that have contributed to a fulfilling life outside the lab. **Dennis**, I am thankful for our countless hours in the gym, discussing everything from plasmids to switched cap converters and meta-PhD related matters. **Jakob**, I truly appreciate our long music-jam-art-mathematical modelling sessions and discussions ranging from Dire Straits to plasmid-loss dynamics. **Lauge**, I always enjoy the time we spend together, whether it is going "hunting" with our cameras, cooking exotic dishes or getting lost in the Swedish mountains! **Filip**, I think we discovered our interest for science simultaneously forever ago! I would also like to thank all the amazing people of KK23: **Joachim, Ida, Kristian, Two-beers, Vino, Skat, Laura, Helene, Victor, Jesper, Farhad, Andrea, Jing…** for endless times of fun and hygge as well as your authentic interest and support in all aspects of my life!

Finally I would like to dedicate the biggest gratitude towards my family. Especially my mother **Birgit**, my father **Gert**, and my brother **Simon** for their all-time presence and support, and also my uncle **Jens** for showing interest in my work.

Table of contents

Abbreviations

CAI	Codon Adaptation Index
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DNA	Deoxyribonucleic acid
ESBL	Extended Spectrum Beta Lactamase
GC	Guanine-Cytosine
GFP	Green Fluorescent Protein
HGT	Horizontal Gene Transfer
H-NS	Histone-like Nucleoid Structuring protein
ICE	Integrative and Conjugative Element
IS	Insertion Sequence
Kb	Kilo base pair
mRNA	messenger RiboNucleic Acid
ORF	Open Reading Frame
ROS	Reactive Oxygen Species
R-plasmid	(Antibiotic) Resistance Plasmid
Spp.	Species (plural)
T4SS	Type-4 Secreation System
TA	Toxin-Antitoxin

Introduction

> "*For the first half of geological time our ancestors were bacteria.*
> *Most creatures still are bacteria..."* – Richard Dawkins (1996)

Evolution is the process through which present-day organisms arose and are constantly changing. Since the first self-replication entities arose on our planet, a myriad of life forms have emerged. While human existence represents only a minute fraction in the history of life, the majority of life forms are single-celled organisms that have inhabited the earth for billions of years.

Being the most diverse and abundant domain of life, bacteria exemplify the remarkable ability of evolution to fit organisms into almost any imaginable ecological niche. Although the capacity of bacteria to undertake a wide range of metabolic functions is fundamental to natural ecosystems modern biotechnology, the same adaptive mechanisms strongly influence human health. While many bacteria support a healthy life, the pathogenic minority that cause disease has received most of the attention in modern medicine[1]. Less than a century ago, infectious disease was the most common cause of mortality, but luckily this situation was drastically improved with the introduction of vaccination and effective antimicrobial drugs. Unfortunately, this was just another bump on the evolutionary road of our microbial enemies, and we are currently witnessing the full capability of bacterial evolution to overcome our most critical medical inventions[2].

Interbacterial transfer of antibiotic resistance genes is among the main events driving the emergence of multidrug resistant pathogens[3]. Therefore, in order to predict, and prevent, the further emergence of so-called "superbugs", we need understand the molecular forces driving the dissemination of mobile genetic elements and their genetic content.

The aim of this thesis work has been to obtain knowledge on the adaptive mechanisms and evolutionary dynamics of bacteria at the molecular level. Specifically, the focus has been on plasmids implicated in the dissemination of antibiotic resistance and virulence within Enterobacteria. I have used high-throughput sequencing and diverse laboratory techniques to follow, reproduce, and measure the effect of adaptive trajectories followed by bacteria *in situ* and *in vitro*. In addition, I employed gene synthesis technology to better understand the selective constrains experienced by transferable antibiotic resistance genes upon entering a naïve host. While most results have been obtained at different genetic levels in the laboratory, I have strived to couple *in vitro* insights to *in vivo* observations for a broader picture of the complex reality experienced by bacterial strains and plasmids residing in the human gut.

Chapter 1

"*Never underestimate an adversary that has a three-point-five-billion-year head start.*"
- Abigail Salyers (2002)

1 Antibiotic resistance

The discovery and development of effective antibacterial compounds is one of our greatest medical innovations. By curing and preventing bacterial infections, that were previously untreatable, antibiotics have saved countless lives and allowed for invasive surgical procedures[4]. Unfortunately, we are now witnessing the power of microbial evolution, and HGT in particular, as a response to our extensive antibiotic use[5].

The first widely used antibiotic was the synthetic arsenic based compound Salvarsan developed to treat syphilis (*Treponema pallidium)* in 1909[6]. The same systematic screening approach led to the discovery of the sulphonamide family of antibiotics in 1935, which is still in use today[7]. Nevertheless, the introduction of penicillin in 1940'es marked the introduction of β-lactam antibiotics, which quickly became the most widely used class of antibiotics due to their high potency and low toxicity[8]. As a natural product, the discovery of penicillin inspired efforts into screening bacteria and fungi for their ability to produce antibiotic compounds. Indeed, important drugs such as streptomycin, tetracycline and chloramphenicol were discovered within a decade after penicillin by systematic screening and isolation approaches[9]. While a few classes of synthetic antibiotics were developed, most of the compounds in use today are (derivatives of) natural compounds[9].

The production of antibiotic drugs by natural organisms has led to numerous speculations on the functions of these compounds in nature[10]. It is tempting to believe that antibiotics have always been agents of chemical warfare produced by organisms to avoid or prey upon bacteria. However, many natural environments do not reach the high concentrations of antibiotics required for significant growth inhibition of competitors, and some studies suggest a more complex regulatory role of antibiotics as signalling molecules not directly involved in the killing of competitors[10].

For many years, antibiotics were considered wonder drugs that would free us from the burden of bacterial infections. Nevertheless, soon after the introduction of the first antibiotics, it was realized that bacteria would suddenly become resistant to previously effective treatments. Today, important pathogens such as *Staphylococcus aureus*, *Klebsiella pneumoniae* and *Escherichia coli* are highly resistant to the first generation of penicillin and many of its derivatives (**Figure 1**). Although resistant pathogens emerged soon after the introduction of penicillin, the first β-lactamase enzyme, able to

hydrolyse penicillin, was discovered even earlier; prior to penicillin's clinical use[11]. Similarly, more recent phylogenetic analysis, and investigations of metagenomic DNA, has revealed that most antibiotic resistance genes are ancient and ubiquitously present in natural environments; including those of no previous human contact[12–18]. It has been hypothesised that these resistance genes mainly evolved in, or as a response to, antibiotic producing organism and subsequently mobilised and spread to human pathogens facilitated by anthropogenic selection[4,19]. While the enrichment of antibiotic resistance genes by humans is clear, the connection from the clinic to environmental reservoirs is still not fully understood[12–17].

β-lactamases - an evolving threat to our most important antibiotics

Historically, the β-lactams is the most widely used antibiotic drug class[20,21]. Unfortunately, β-lactamase enzymes able to hydrolyse all know β-lactam drugs are now widely disseminated and pose a growing challenge in clinical antibiotic resistance. While β-lactamases have mobilised and spread extensively throughout the 19th century, they are also widely present in environmental isolates and have likely been around for millions of years[16,22]. The first β-lactamases to emerge among clinical isolates belonged to the TEM and SHV families that would confer resistance to the first generations of penicillins. By the 1970s these enzymes would be highly disseminated among especially *E. coli* and *K. pneumoniae* isolates that were still sensitive to the cephalosporin class of β-lactams introduced in the 1980s[20]. However, already in 1983 plasmid-borne mutants of the original SHV and TEM enzymes, able to hydrolyse cephalosporins, arose[20]. These improved β-lactamases were termed "Extended Spectrum Beta Lactamases" (ESBL) and the SHV and TEM variants were the dominant ESBLs in enterobacteria up until the 1990s. However, by the end of the 1990s, other β-lactamases had emerged, including the novel CTX-M class that were not directly related to the TEM and SHV β-lactamases. This CTX-M class quickly arose to become the most prevalent ESBL class carried by *E. coli* and *K. pneumoniae* clinical isolates, and no particular reason for their success has been proposed[23].

Because the ESBL prevalence has reached critical levels in most countries (**Figure 1**) the carbapenems is a critical class of β-lactam drugs used as last resort therapy against ESBL positive infections[24]. However, the selection imposed by carbapenems have encouraged the rapid emergence and spread of carbapenemases, with some European countries experiencing resistance rates of > 50% among invasive Gram-negative isolates[25]. In these isolates, carbapenemases are most often encoded from plasmids by enzymes such as the Klebsiella Pneumonia Carbapenemase (KPC) and New Delhi Metallo-β-lactamase-1 (NDM-1) that, along with VIM and OXA types, demonstrate how the emergence and spread of certain resistance genes, as a response to increased drug usage, can rapidly compromise medical treatments[26].

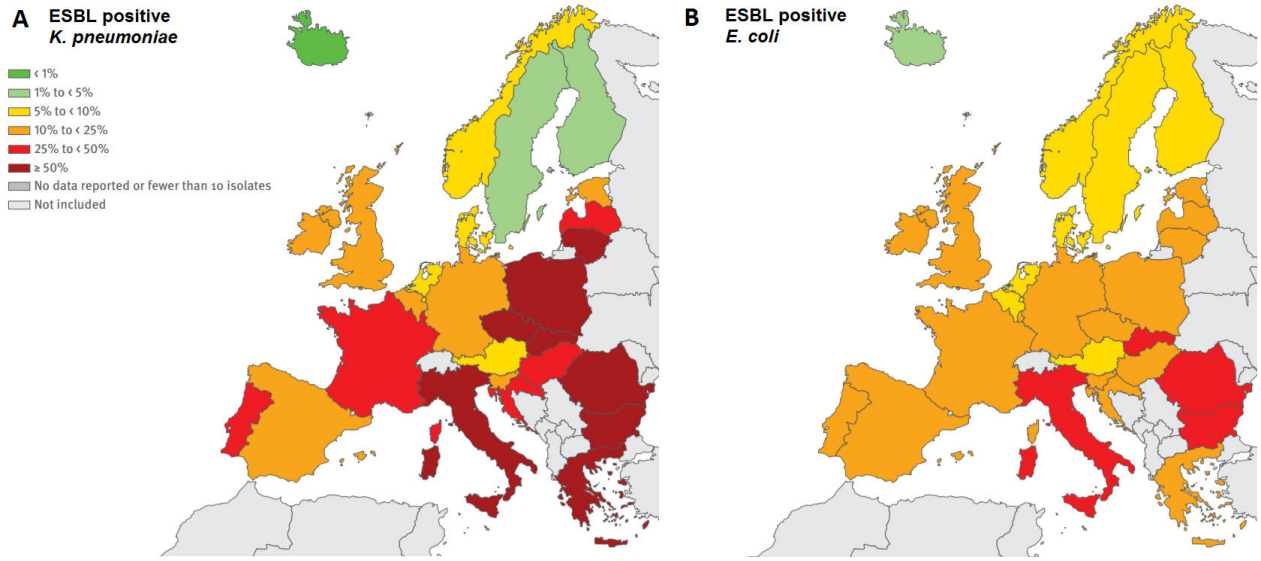


Figure 1. The prevalence of invasive (blood and cerebrospinal fluid) *K. pneumoniae* (**A**) and *E. coli* (**B**) isolates resistant to third-generation cephalosporins in Europe (2015). Figures adapted from the ECDC surveillance report 2015[25].

Mechanisms of antibiotic resistance

Genetic innovations leading to antibiotic resistance is caused by either modification (e.g. mutation or amplification) of endogenous genes or, more commonly, the acquisition of pre-evolved resistance genes. Mutational resistance is most often caused by alterations of conserved antibiotic targets, up-regulation of chromosomally encoded efflux pumps or down-regulation of uptake mechanisms, and will not be detailed further due to the focus of this thesis on transferable resistance[27].

Biochemical research into the mechanistic basis of transferable (primarily plasmid-borne) resistance has revealed numerous mechanisms of resistance[4]. The mechanisms by which resistance genes confer resistance are diverse and can be classified into five major categories (**Figure 2**) detailed in the following.

Drug modification

An important class of diverse resistance mediators are those interacting directly with the drug to chemically inactivate it by cleavage or modification. Drug hydrolysis is the most common cause of β-lactam resistance and is mediated by numerous β-lactamase variants of different spectrum[28]. Drug modification is another widespread strategy resistance employed by especially aminoglycoside and chloramphenicol resistance enzymes (acetyl-, phosphor-, and adenyltransferases); but are also common for macrolides (e.g. *ereA*) and have, more recently, emerged for tetracycline (*tetX*)[29].

Efflux

Because most drugs work inside the cell, preventing the drug from reaching its target by decreased permeability or increased efflux, serves as a widespread resistance strategy[27,30]. Whereas decreased permeability is intrinsic to some bacteria, and can be achieved via mutational down-regulation of membrane porins, active efflux of antibiotic compounds is also a common mechanism for transferred resistance genes[29,31]. Efflux pumps are membrane spanning structures that vary in their complexity and substrate specificity[30]. While some efflux pumps e.g. the *tetA* tetracycline/proton antiporter are fairly specific, others e.g. the chromosomal AcrAB-TolC and MexAB-OprM are more promiscuous efflux systems intrinsic to *E. coli* and *P. aeruginosa* conferring low-level resistance to multiple drug-classes when up-regulated. Efflux-mediated resistance has been reported for most drug classes, but are less frequently involved in resistance towards drugs acting on cell envelope structures (outside the cell) compared to those acting inside the cell e.g. ribosome targeting antibiotics[30].

Target modification

As for mutational resistance, drug-targets can be modified to avoid binding of the drug. There are several ways by which enzymatic modifications or physical protection can, similar to amino acid changes, prevent the drug from binding to its target site[27]. Ribosome modifying enzymes are often found on plasmids in both Gram-positive and Gram-negative bacteria and a prevalent class is encoded by the *erm* (erythromycin ribosomal methylation) genes[29]. These enzymes confer macrolide resistance via methylation of the adenine 2058 residue of the 23S rRNA in the 50S ribosomal subunit, which also affect the binding of lincosamides and streptogramin B[29]. The vancomycin resistance (*van*) gene-clusters encode cell wall restructuring enzymes that modify the cell wall of Gram-positives to allow peptidoglycan cross-linking despite the presence of vancomycin[32]. Other proteins do not chemically modify the drug target, but may displace (e.g. TetO and TetM) or stabilize the drug-bound structure (gyrase binding Qnr peptides) to avoid lethal effects via close interaction with the target[29].

Resistant target (target replacement)

Another way to circumvent (recessive) lethal effects of drug inhibition, is to supply a resistant target replacement that retains its native functionality[29]. An important example is the *mecA*-encoded penicillin binding proteins with lower affinity for β-lactams widely present in *S. aureus* (MRSA)[33]. Trimethoprim and sulfamethoxazole drugs target two essential enzymes in the folate synthesis pathway: the dihydrofolate reductase and dihydropteroic acid synthase respectively[29,34]. Resistance

to these drugs can be accomplished by acquisition of resistant variants, or mutations in, these genes, but might also be achieved via overexpression of the native enzymes to titrate the drugs[29].

Regulatory mechanism

Bacteria have evolved systems, which in reaction to external stimuli elicits a response, to environmental stimuli e.g. xenogenic compounds, high temperature or nutrient starvation. These phenotypic adaptations happen through rewiring of cellular networks that increases resistance via existing features of the genome[29]. For example, the CpxA sensor kinase of *E. coli* is activated upon envelope stress and activates a kinase cascade to up-regulate efflux pumps that excrete toxic chemicals[35]. Although plasmid-encoded global regulators e.g. H-NS homologs may influence endogenous or acquired resistance mechanisms[36], regulatory changes leading to resistance are mostly achieved via mutations e.g. of efflux repressors or porin expression[27]. The lack of trans-acting regulators on plasmids is likely a result of the complex contexts in which they have to function, as well as the generally low level of resistance conferred by these resistance mechanisms in most species (**manuscript V**)[27,35].

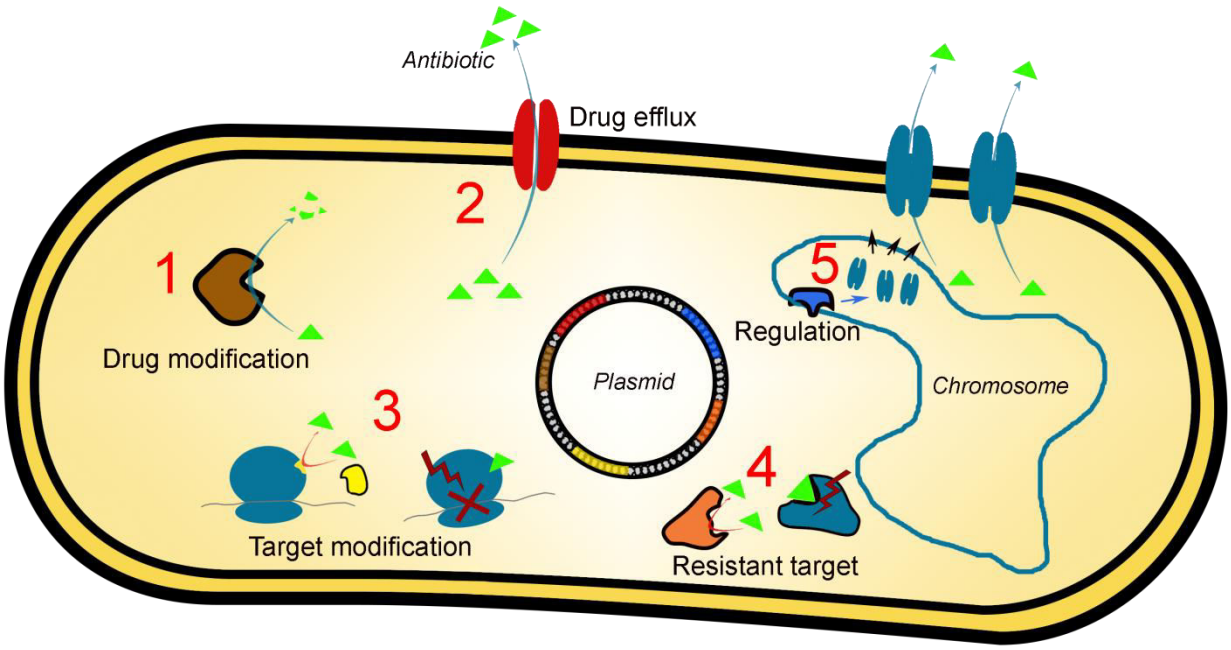


Figure 2. Molecular mechanisms of transferable antibiotic resistance. Genetic vectors may carry genes that confer resistance through five major categories: **1**) Modification or degradation of the drug. **2**) Elimination of the drug through active efflux. **3**) Modifications of cellular structures to avoid inhibition or lethal effects of antibiotic binding. **4**) Replacement of the drug target by a resistant variant. **5**) Regulatory interference to activate endogenous resistance mechanisms.

Chapter 2

2 Horizontal gene transfer and bacterial evolution

The beauty of evolution is that simple principles, acting on genetic entities, allow us to understand complex biological phenomena such as disease dynamics and antibiotic resistance. Neo-Darwinistic theory rests on the notion that new traits arise by gradual change and selection imposed by the environment[37]. The molecular basis for this theory is the generation of *de novo* mutations in existing genetic material caused by factors such as uncorrected DNA-replication errors, intra-genomic recombination or exposure to mutagens. Such mutations can either be deleterious, neutral or beneficial to the reproductive success, or "fitness", of the organism in a given environment. This paradigm is limited by the assumption that two diverging phylogenetic groups always share (variations of) the genetic material found in their common ancestor. However, these principles of "gradual change" do not explain the large innovative "jumps" sometimes observed in nature[38].

in the late 1940s Tatum and Lederberg challenged the assumption of strictly vertical evolution in bacteria by demonstrating the "sexual" recombination of genetic traits, later identified as plasmid transfer, between two variants of *E. coli*[39]. However, the extent to which genetic transfer played a role in evolution was not known until several decades later[40].

From initial sequencing efforts applied in molecular phylogeny, it became clear that different genetic markers would yield conflicting phylogenetic relationships between organisms[41]. This realization, along with the numerous observations of multiple antibiotic resistance plasmid-transfers, lead to the increasing recognition of HGT as an important driver of evolution[40,42,43]. When bacterial whole genome sequences became available around the turn of the millennium, the horizontal aspect of evolution was further supported, and a surprising variation in genomic content within species was revealed[38,44–47]. One of the first comparative genomics papers compared three *E. coli* genomes available at that time (2002) and suggested that these three strains shared less than 40% of their genomic content, despite their common species affiliation[48]. Transferred genes are typically identified from their divergent sequence composition, from homology networks or incongruent phylogenetic gene-signals compared to established, e.g. ribosomal, phylogenetic markers[49–51]. Although such techniques have their limitations, and only successful transfer event can be detected, it has been inferred that at least 81 ± 15% of all bacterial gene families have undergone horizontal transfer at some point in their evolutionary history[52].

Of these transfer events, several examples of cross-domain gene transfers exist e.g. between archaea and bacteria as well as between eukaryotes and prokaryotes and vice versa[53–55]. Discovering such cross-clade transfer events challenged the idea of a unidirectional tree of life and offered the "network of life" as a more meaningful metaphor describing the pervasive gene sharing across phylogeny[56].

Bacteria often reside in communities where HGT can readily take place, and many genes are associated more with a particular environment than a specific host[57,58]. The massive genetic flux experienced by prokaryotic organisms has blurred the species definition, and even closely related species can vary substantially in their genomic content[59]. This realisation has encouraged the use of terms such as "accessory genes" and the "pan-genome". Whereas the "core-genome" refers to the genes shared between all members of a species, the pan-genome refers to all genes within a species; including those harboured by only a subset of strains. For example, while the core-genome of sequenced *E. coli* isolates is made up from around 3188 genes, the current pan-genome of this species exceeds 60.000 gene families[60]. These "extra" accessory genes represent a pool of transferable genetic material, which bacteria can potentially draw for their adaptation[61].

Gene acquisition and genomic rearrangement/deletion events are thought to occur more frequently, and have a higher selective impact, compared to single nucleotide substitutions in most bacterial species, and have likely played a major role in evolution (**Figure 3**)[62]. However, it has been challenging to understand how prokaryotic genomes can be so highly energetically optimized, and functionally dense, while still tolerating the constant flux and potential disruptive effects of horizontally acquired genes[62,63]. While isolated populations of e.g. intracellular pathogens, usually have small "closed" pan-genomes, species that experience fluctuating environmental conditions tend to have large "open" pan-genomes, and this might reflect the benefit of flexible adaptation that HGT provides[64,65]. While stress, e.g. induced by antibiotic exposure, may modulate recombination and HGT frequencies, to what extent the rates of HGT is subject to selection is not clear[66,67]. Although high rates of gene acquisition increases the chance of acquiring rare beneficial genes, it is often deleterious to receive novel genetic material, and the pervasive rates of HGT in nature could also indicate that (selfish) mobile genetic elements are simply overwhelming their unwilling hosts[62]**.**

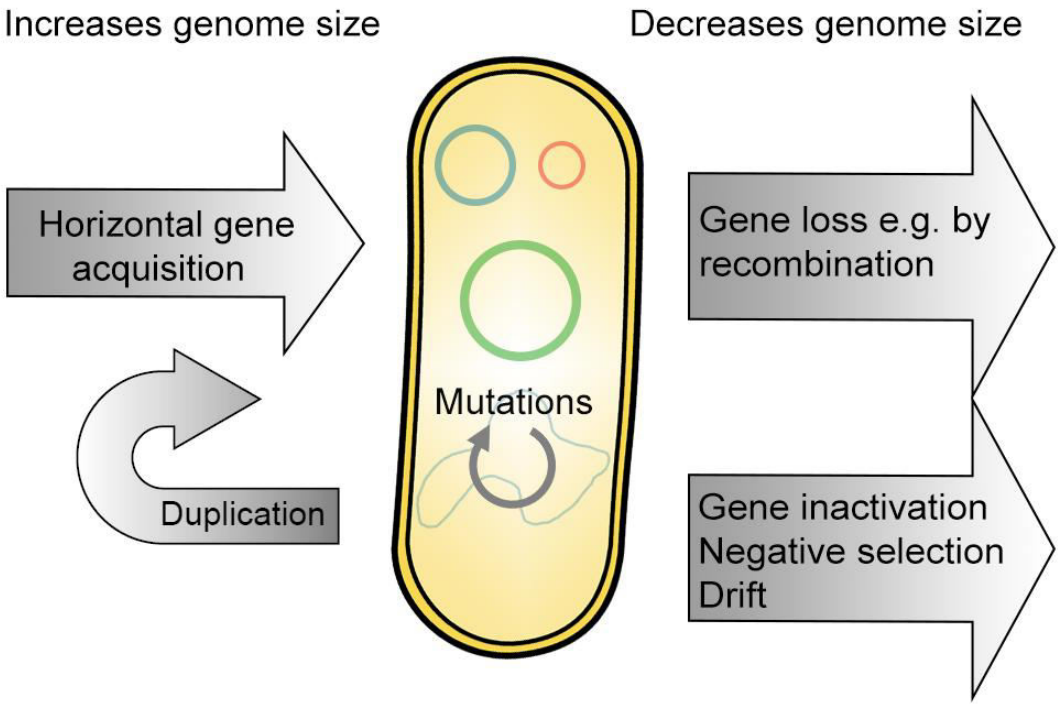


Figure 3. Microbial genomes are subject to constant genetic flux facilitated by multiple genomic events. Deletions, amplifications, single nucleotide mutations (SNPs), recombination and the acquisition of numerous mobile entities through HGT allow for adaptive innovations. Adapted from Mira et al. 2001[68].

2.1. Modes and elements of gene transfer

A first crucial step in the transfer of genetic material is the physical delocalisation of DNA from one cell to another. For this purpose, the most well studied mechanisms are: Transformation, phage transduction and conjugation[69].

Transformation

Transformation is the direct uptake of DNA from the surroundings[70]. The term was originally coined by the British microbiologist Frederick Griffith following his famous experiment with *Streptococcus pneumoniae* in 1928. He infected mice with a live, but avirulent, strain along with a heat inactivated virulent strain of *S. pneumoniae*. While he did not expect the establishment of an infection, a morphological "transformation" of the avirulent strain into the virulent strain was observed[71]. He attributed this phenomenon to a "transformation factor", which was later recognized as DNA[72]. Transformation requires several competence factors to be expressed by the host, and the efficiency of DNA uptake relies exclusively on the recipient bacterium[73]. The competence machinery facilitate the uptake, stabilization and processing of the exogenous DNA[70]. Although their expression conditions vary substantially, competence factors have been identified in more than 82 bacterial species[73]. For these species, few of which are pathogenic, natural transformation can be the major

source of HGT and genome evolution, however, our knowledge on the contribution of natural transformation to antibiotic resistance dissemination is still limited[74].

Phage transduction

Transduction is bacteriophages-mediated transfer of DNA[69]. Phages are believed to be the most abundant biological entities on the planet; estimated to outnumber their bacterial prey by 10:1 [75]. The successful infection of a temperate phage leads to either cell lysis, resulting in host killing and immediate propagation of the phage, or chromosomal integration (lysogeny). The latter manifests as so called "prophages", which are highly abundant in bacterial genomes[75]. While the natural life cycle of phages does not normally involve the packaging and transfer of bacterial DNA, accidental uptake of host DNA by phages can happen by either generalized or specialised transduction[5]. Specialised transduction involves the inclusion of flanking chromosomal DNA in the excised prophage; restricting transfer to genes in close proximity of the integrated phage. Generalized transduction is the accidental packaging of random host DNA during the lytic cycle, which potentially allows for all genes in the genome to be transferred[76].

While transducing phages generally also encode protein components of the phage particle itself, certain prophage-like genomic regions almost exclusively encapsulate random genomic DNA to form transducing particles termed "Gene Transfer Agents"[77]. Phage mediated HGT is widely documented and important examples of genes disseminated through phages include the photosynthesis genes in cyanobacteria as well as virulence factors of *E. coli*, *Vibrio cholerae* and *Staphylococcus aureus*[78–81]. Being encapsulated, phage-mediated transfer is believed to have a long reach in terms of time and distance compared to other transfer mechanisms, however, phages are often very host specific and tend to mediate transfer within a narrow host-range[82]. Although broad host-range phages do exist, recent findings suggest that transduction may also take place without strict host (replication) compatibility[83,84].

Conjugation

In contrast to transformation and transduction, the process of conjugation requires cell-to-cell contact, because a junction, through which DNA is transferred into recipient cells, needs to be formed[69]. In Gram-negative bacteria, this happens via a single stranded DNA intermediate that is channelled through a type-IV secretion system (T4SS). By virtue of its dedicated role in DNA transfer, conjugation is perhaps the most sophisticated and relevant mechanism mediating the transfer of adaptive factors, such as antibiotic resistance, virulence and metabolic genes, in nature[85–87]. Many large plasmids encode full conjugational transfer systems mediating their own transfer, but these might also aid the transfer of, often smaller, mobilizable plasmids in trans[88]. Conjugation is widely

involved in the transfer of genetic elements ranging from small mobilizable plasmids to large chromosomally located integrative conjugative elements (ICEs) as well as entire bacterial chromosomes[89]. In contrast to the other mechanisms of transfer, some conjugation systems are extremely broad in their transfer range and allow for cross kingdom transfer of plasmid DNA e.g. from bacteria to plants and other eukaryotes[89,90].

Due to the biological cost of expressing and integrating the large structures involved in conjugation, they are often subjected to tight regulation that allow expression only temporarily such as immediately following arrival in the recipient cell or in certain growth stages in response to optimal cell densities[91]. These (transient) costs can be afforded by a mobile element if the horizontal propagation outweighs the negative effects on vertical inheritance[92]. However, the structural components of conjugation systems may also provide a direct selective benefit for the host cell due to its role in adhesion and biofilm formation[93].

2.2. Plasmids and their genetic content

The work of this thesis has focused largely on the transfer and evolution of plasmids, and these important genetic elements will be described in detail below.

Some years after their discovery the term "plasmid" was defined by Joshua Lederberg in 1952 as "*a generic term for any extrachromosomal heredity determinant*"[94]. This term is still valid today where thousands of plasmid sequences have been deposited in public databases. Apart from being abundant natural vectors of gene transfer, plasmids have also played a detrimental role in the development of recombinant DNA technology and their usefulness for heterologous gene expression makes them fundamental to modern biotechnology[43]. Because of their highly mosaic structure and self-sustaining genetic features, plasmids are perhaps the most important vehicles of HGT and are highly implicated in the dissemination of traits such as virulence or antibiotic resistance[85,95,96]. Plasmids are often modular with conserved regions involved in stability and replication, and other, more variable, "adaptive" regions are susceptible to rearrangements or insertions of mobile genetic elements carrying accessory features (**Figure 4**). Many plasmids harbour transposable elements, some of which have carried genes or integron structures, able to capture novel genetic entities, into the plasmid backbone[61].

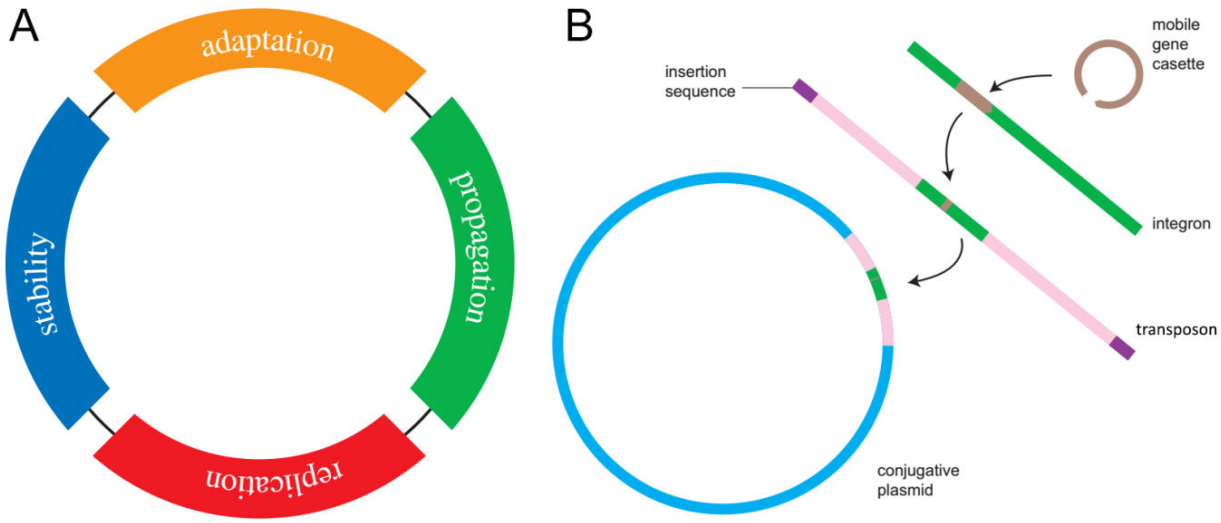


Figure 4. A) Plasmid gene-functions can be categorized into four groups contributing to different aspects of plasmid survival; each of which are described in the main text. **B)** Plasmids are themselves host replicons for lesser mobile genetic elements that are acquired in a hierarchical fashion. The structure of plasmids often shows a pattern of sequentially acquired mobile genetic elements; sometimes collected in "hot-spot" regions with a high genetic turnover. Figures from Norman et al. 2009[61].

Transposable elements

The simplest mobile genetic elements are insertion sequences (IS)[97]. They employ transposase enzymes to copy (replicative transposition) or cut (conservative transposition) themselves from one place in the genome to another[98]. For some transposase families, the recombination mechanism recognises specific target site motifs, whereas others seem to jump more or less randomly[98]. Even though ISs do not confer any direct adaptive benefits, transposase genes are the most ubiquitous and abundant genetic elements in nature, where they play an important role in genome reorganisation[99]. IS elements provide genome plasticity by catalysing insertions, deletions and rearrangements at much higher frequencies than the genetic innovations achievable by DNA replication errors alone (**manuscript IV**)[100,101]. Their importance in HGT rely on the ability of proximal IS elements to transfer genetic cargo as transposons, and the wide presence of transposons on plasmids reveals their essential role in plasmid genesis and (re-)organization[61]. Although the adaptive content of transposons has supported their wide dissemination, they can also aid in adaptation via disruptive insertions or recombination events to delete costly regions as is exemplified by in the work of this thesis (**manuscript III** and **IV**)[101].

Integrons

Another important class of genetic structures promoting gene mobility and expression are integrons. Although integrons are often found in transposons, the controlled integration and expression of gene cassettes by integrons has many benefits compared to the potentially disruptive nature of transposon insertions[102]. Genetic material is recognised by the presence of an *attC*-site and inserted into the integron structure via site-specific recombination with the integron *attI* site; a reaction that is catalysed by the integron encoded Int1 integrase of the tyrosine integrase family[102]. Apart from acquiring genes, a key feature of integrons lie in their ability to express the acquired genetic material from a strong integron-associated promoter located upstream of the *attI* recombination site. While integron structures found in environmental isolates are highly diverse, the integrons of pathogenic bacteria typically belong to the class 1 type embedded in IS-flanked genetic contexts on plasmids[102,103]. Class 1 integrons are believed to originate from environmental variants that were transferred to, and highly enriched in, pathogens upon resistance gene acquisition as a response to increasing antibiotic exposure[103].

Replication, copy-number and host-range

Plasmids need to carefully control their replication in order to maintain sufficient copy numbers for stable inheritance, but at the same time avoiding negative selection from excessive burdening of the host[104]. Therefore, replication is often coordinated with the replication of the host chromosome and takes place by the strand displacement, rolling circle or theta mechanism. All replicons examined in this thesis replicate by the theta replication mechanism, which progress in a similar manner to chromosomal replication. Theta replicating plasmids may depend entirely on host initiation proteins (e.g. ColE1) but more often carry their own replication "Rep" initiator protein that binds to directly repeated "iteron" recognition sites. The binding of the rep protein to these sites serves to melt the DNA duplex and recruit host replication factors, but protein-protein interactions between Rep proteins, on different plasmids within the cell, may inhibit this initiation if the copy number (and thus amount of Rep protein) is already high enough[105]. The overlap of dosage-dependent regulation is also the reason for incompatibility between similar replicons; a feature widely used for plasmid classification. The host-range of plasmids is determined primarily by the compatibility of host-factors with the plasmid replicon[106]. Therefore, some plasmids encode more replicons or additional replication factors to broaden their host compatibility. However, because replication does not equal stable inheritance, plasmids have evolved distinct ways to avoid loss upon cell division.

Plasmid stability systems

There seems to be a trade-off between plasmid copy number and size for natural plasmids, and large plasmids are often maintained in very few copies per cell. Low copy numbers, however, comes with a higher risk of plasmid segregational loss because the chance of random dispersal of plasmids to both daughter cells upon cell-division decreases with decreasing plasmid copies. Therefore, partitioning systems are of vital importance to ensure stable inheritance of low-copy plasmids. These systems employ two trans-acting proteins and a (centromere-like) binding site to actively bring plasmids into each daughter cell upon cell-division. An example of such a system consist of the "ParA" filament protein that connects to the plasmids "*parS"* site via the "ParB" anchoring protein[107]. If these first-line segregation systems fail, many plasmids employ second-line rescue systems that limit the competitive effects of plasmid-free cells on the plasmid-bearing population. These "addiction systems" were discovered due to their effect on plasmid stability and are widespread on mobile elements as well as chromosomes. They encode a stable toxin along with a less stable anti-toxin that, under normal conditions, neutralizes the effect of the toxin. If a plasmid encoded toxin-antitoxin (TA) system is lost, the relatively stable toxin will inhibit the growth of the plasmid-free cell due to the quicker turnover of the antitoxin. TA systems are diverse and work through many different mechanisms. While the toxic component is generally a protein that targets essential processes of the cell, the anti-toxin can also be inhibitory RNAs. For example, type I and III TA systems encode protein toxins that are inhibited by RNA anti-toxins binding directly to the toxin protein or its encoding mRNA. Interestingly, plasmid-borne restriction/modification systems may also act as TA systems with the added benefit of restricting incoming plasmid competitors[108].

The continuum of plasmid-host compatibility

When is a plasmid compatible? Rather than being a deterministic quality, the host compatibility of plasmids is determined more fluently by transfer, replication and stability mechanisms, which may be somewhat dynamic features[109,110,106]. Whereas some plasmids are only associated with a very limited set of hosts, others can seemingly propagate in a broad range of species[109,111]. It is evident that even closely related plasmids can co-exist to some degree, despite overlap in replication features that traditionally categorized them as "incompatible"[106]. Importantly, many plasmids can replicate in a number of hosts, but the degree to which they are stably maintained over time can vary substantially[101,110], and instability is a plastic feature, which may be modulated through adaptive evolution to expand or shift effective host compatibility[101,112]. The notion of host-range is complicated further because many plasmids can transfer much beyond the hosts that allows for their replication[89,90,101,110,113]. The implications of such transfer events are not well understood, but delivery of bacterial DNA to eukaryotes does play a role in bacterial virulence[113].

This plasticity calls for an updated definition of host-range to consider not only replication-proficient hosts, but to also include the hosts in which a plasmid is (and has the evolutionary potential to be) stably maintained (**Figure 5**). For example, initially very unstable plasmid-host pairs may be maintained under sufficient positive selection for plasmid-borne traits to encourage the evolution of a more stable association in the absence of selection[101,114].

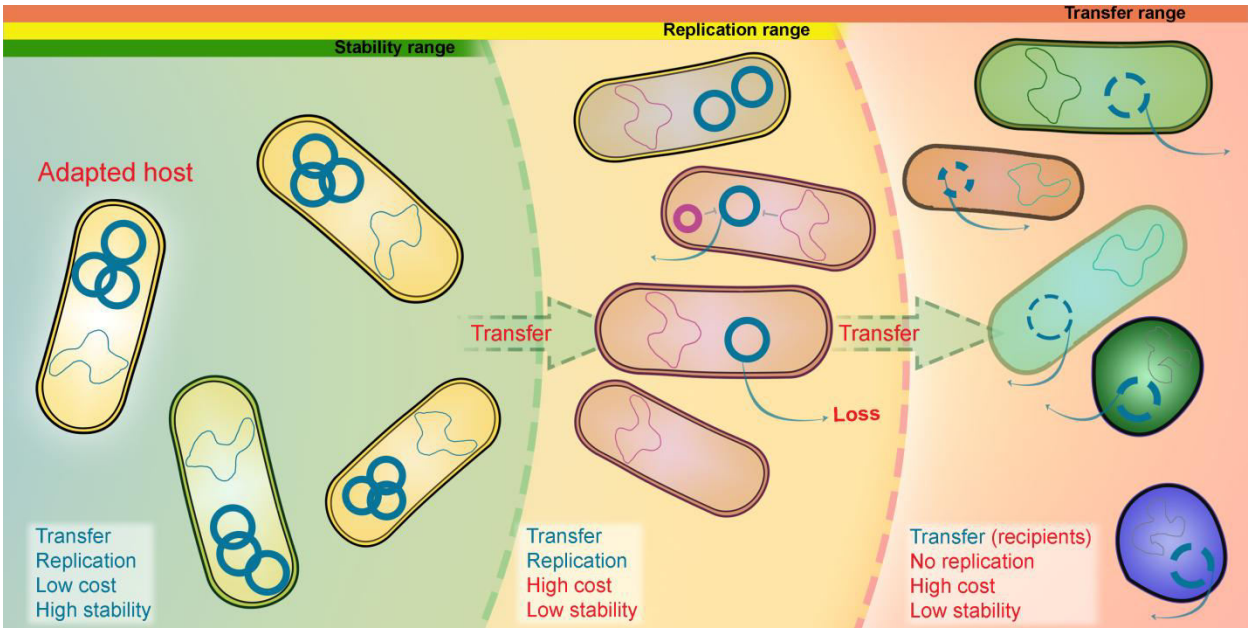


Figure 5. Plasmid-host compatibility can be nuanced across replication, stability and transfer proficiency. Rather than being equally "compatible" with all potential plasmid recipients, a plasmid may experience varying degrees of stability due to differences in replication proficiency, costs or segregational efficiency. Plasmids typically resides within a domestic group of hosts, representing the most stable plasmid-host combinations (green area), but may occasionally transfer to, and reside in, less optimal hosts (yellow and red areas). While transfer proficiency can be high, naïve plasmid-host combinations are often unstable due to suboptimal, or costly, interactions with new hosts. Given enough time, i.e. via positive selection of plasmid-borne traits, a plasmid may evolve to shift its stable host range. The transfer range of conjugation systems is often broader than the replication range of the plasmids encoding them; allowing for plasmid delivery to hosts where immediate loss is likely unless recombination with endogenous replicons or rapid adaptation takes place.

2.3. The role of plasmids in host adaptation

Despite being unstable, a plasmid can be maintained in a population if it carries beneficial traits, and the most variable and interesting feature of plasmids is perhaps the genetic cargo that they carry[61,115,116]. These adaptive genes are not directly involved in plasmids maintenance, but provide an indirect benefit to the plasmid by improving the survival of the host. Accessory elements are often found in defined regions of the plasmid backbone that are separated from those involved in stability, replication and transfer[61,115]. Depending on the selective benefit of inserted traits in a given environment, variants of the plasmid backbone created by continuous recombination, insertion, deletion and mutational events will continuously be selected (**manuscript III** and **IV**)[101,115]. Although the role of many plasmid-carried genes has yet to be elucidated, some adaptive traits can lead to massive innovative jumps exemplified by plasmid-borne metabolic pathways or multidrug resistance[101,117]. There are important examples of adaptive plasmid traits that influence the ecology of their hosts to an extent where the presence of the plasmid largely defines the species. Such plasmid-driven instant speciation events have played a tremendous role in the evolution of important pathogens including *Yersenia pestis*, *Bacillus antrhacis* and *Shigella* spp.[118–121]. These species all encode their virulence/colonization factors from large stable plasmids and there is evidence of sequential acquisition of adaptive virulence genes e.g. by the pMT1 plasmid of *Y. pestis*, that support sudden ecological jumps[120]. In theory, most genes can be carried on plasmids and quasi-essential plasmids carrying ribosomal operons exist in certain species[122]. While most genes have probably been associated with plasmids throughout history, some genes are more likely to be selected in a plasmid context[123]. For example, genes involved in local adaptation and social traits are believed to benefit from the mobility, and perhaps the copy-number plasticity, of plasmids that might help explain the somewhat puzzling existence of non-transmissible plasmids[124].

Plasmids and antibiotic resistance

Historically, plasmids probably owe their fame to their involvement in antibiotic resistance and most of our serious resistance problems can be attributed to plasmids[125–127]. Whereas the treatment of persistent infections, such as tuberculosis or *Pseudomonas aeruginosa* implicated in cystic fibrosis, is mainly compromised by in-host evolution of mutational resistance, HGT of plasmid carried antibiotic resistance genes is largely responsible for the current pandemic of multidrug resistant pathogens[126]. When plasmid-borne resistance was first reported some 60 years ago, transferable multidrug resistance was not anticipated by the research community, and it came as a big surprise that pre-made resistance packages could be transferred between pathogens[127]. The first observations of transferable resistance were made by Japanese researches in the mid-1950s following an outbreak of *Shigella dysenteria* resisting treatment of up to four antibiotics[127]. The demonstration that these

so-called “R-factors” were transferrable to other Enterobacteriaceae, and curable using acridine orange, made the connection to concurrent research on the F-plasmid that defined R-factors as plasmids[128]. These studies were published in Japanese journals and received little attention in the west. Eventually, the Japanese work on R-plasmids was published in western journals but were met with scepticism[127]. It was not until the early 1960s that the British geneticist *Naomi Datta* discovered similar phenomena of transferable multi-drug resistance in *Salmonella typhimurium* during an outbreak in London[129]. Subsequent studies showed that plasmids were common across bacterial species, and their presence and resistance profiles would often reflect antibiotic usage rather than specific host strains[127]. Such observations led to the worrisome realisation that antibiotic resistance was not easily confined, and that the problem was likely amplified by the massive use of antibiotics outside the clinic. Today, plasmid-borne multidrug resistance is widespread and highly problematic for clinical outcomes[96,130]. Important Gram-negative pathogens approaching a pan-resistant level due to plasmid acquisition are: *Acinetobacter baumanii*, *P. aeruginosa*, *K. pneumonia* and *E. coli*. Especially nosocomial *E. coli* and *K. pneumoniae* infections are prevalent and are becoming increasingly resistant due to plasmid carriage, which places them among the most urgent bacterial threats in our health care system today[2,3].

Important bacterial clones implicated in plasmid-borne antibiotic resistance

While any pathogenic strain can in principle acquire a multidrug resistance plasmid and become a clinical problem, there seems to be a bias towards specific clone-types in the dissemination of multidrug resistance plasmids[3]. Two important high-risk clones that seem to be especially good at disseminating and acquiring resistance plasmids are *E. coli* ST131 and *K. pneumonia* ST258[131,132]. Within the last 20 years, *E. coli* ST131 has become the dominant extraintestinal pathogenic *E. coli* sequence type in the world[133]. Its evolutionary history suggest a role of early mutational fluoroquinolone resistance and CTX-M-15 β-lactamase encoding plasmid acquisition as important adaptive steps, however, high virulence and increased host-to-host transmission are also hallmarks of the ST131 clone[3]. *K. pneumonia* ST258 is often implicated in nosocomial bloodstream and lung infections, and its tendency to carry plasmid encoded carbapenemases renders this clone-type especially problematic. The ST258 clone is believed to be a hybrid of *K. pneumonia* ST11 and ST442 with an added genomic island (ICEp258.2) that encode type IV pili and restriction systems[3]. Compared to the ST11 ancestor, lacking this genomic island, ICEp258.2 seems to facilitate the carriage of carbapenemase encoding IncF plasmids to a higher degree[132].

Although the reason for the success of these clone-types is still not fully understood, stable plasmid carriage seems to play an important role in their adaptation. Interestingly, recent comparative genomics analysis of ST131 genomes suggest that adaptation of regulatory regions in the core

genome, as a response to plasmid acquisition, have increased the plasmid hosting qualities of this clone[134]. However, we still need to figure out if other factors, e.g. high virulence, colonization or transmissibility, have associated these clones more frequently with clinical environments, and therefore increased plasmid exposure and subsequent adaptation, or whether it is the other way around (improved plasmid carriage has led to clinical success).

2.4. Why do plasmids exist?

While there can be obvious benefits for pathogenic hosts in carrying plasmids encoding virulence or antibiotic resistance genes, we know fairly little about how selfish genetic elements persist when they do not confer an advantage to their hosts. This long standing question in plasmid biology, more recently referred to as the "plasmid paradox", has been subject to much attention in recent years, but was originally taken on by researches in the 1970s[135]. Plasmids encode inherently selfish features (conjugation, partitioning and TA-systems etc.) that promote plasmid survival at the expense of host fitness[136]. Therefore it was proposed that, because plasmid-loss will always occur at some frequency, long-term plasmid survival requires either positive selection for plasmid-borne traits or high rates of conjugational transfer[116].

Initial efforts employed mathematical modelling of plasmid invasion experimental setups to explore the population dynamics and existence conditions of bacterial plasmids[135,137]. These pioneering studies were carried out on constitutively conjugating F-plasmids in the absence of selection and predicted a wide range of conditions under which these plasmids could be maintained as infectious parasites through conjugative transfer alone[135,137]. However, most initial studies were carried out on very few plasmids, which had been maintained in the laboratory for years, and these studies did not explain the existence of non-conjugative plasmids in nature[138]. As more plasmids, including "freshly isolated" natural plasmids were subjected to similar investigations in the following years, it became clear that most natural plasmids encode strictly controlled transfer machinery and were not always capable of parasitic maintenance; even under optimal conditions[138].

Because only some, very infectious, plasmids could survive as purely parasitic entities, and only under the favourable conditions of e.g. high population densities used in laboratory experiments, the view on plasmids as purely selfish entities was relaxed[139]. Due to the high cost of de-repressed conjugative plasmids compared to regulated ones, it seemed that the excessive cost of constitutive transfer would seldom pay off in nature[138–141]. Instead, it was suggested that plasmids must have evolved towards mutualism with their hosts to lessen their burden at expense of their horizontal propagation[139]. The idea that some degree of mutualism was required for plasmid survival could also explain why costly or non-conjugative plasmids would often encode beneficial features such as antibiotic resistance[139,142]. While most conjugative plasmids are probably maintained by a

combination of conjugative transfer and positive selection for plasmid-borne traits, the role of infectious transfer, including the influence of host background and external environments, in parasitic plasmid maintenance is still debated[116,143–146].

Due to the tendency of bacterial genomes to expel unnecessary genetic elements over time[63], it has been puzzling why highly selected e.g. antibiotic resistance genes do not jump from the plasmid into the chromosome more often to get rid of the (costly) plasmid[116]. Plasmid specific qualities have been proposed to argue for the selection of plasmid located over chromosomally located genes, and underline the importance of plasmid transfer in answering this dilemma (**Figure 6**)[116,147–149]. One idea is that plasmid-borne traits are often only locally beneficial, i.e. in a local niche or in a limited time-span, and would be able to invade faster than chromosomal genes during sporadic selection[149]. Such reasoning can also help explain why universally beneficial (i.e. essential) genes are almost exclusively encoded from the chromosome. A similar hypothesis was put forward by *Bergstrom et al.* suggesting that the long-term survival of plasmids is facilitated by their ability to "sample" potentially superior hosts immigrating into a local plasmid bearing population; a scenario where stationary chromosomal genes would be outcompeted. This reasoning is in line with the conventional argument for sexual reproduction, stating that the higher diversity created by genetic recombination is beneficial because it increases the chance of creating novel superior gene combinations[150].

Another appealing aspect of transferable plasmids is their ability to promote social traits[116,148]. It has been argued that genes, such as those involved in iron acquisition, virulence or bacteriocin production, are carried on plasmids because it allows for re-infection of cheaters to strengthen cooperation within bacterial populations[147–149,151]. Another important cooperative characteristic that plasmids are known to promote is biofilm formation; a life style that may also allow for higher rates of conjugation[93,152,153]. While the selective advantages of biofilms are manifold, they have also been suggested to promote plasmid-survival due to their spatiotemporal properties[145]. The idea is that plasmids are retained in the lower, nutrient depleted, levels of a biofilm structure where cells are not (or very slowly) dividing, and that this "reservoir" may rescue the population under sporadic plasmid selective conditions (**Figure 6**)[145].

While the biofilm-reasoning would also support the persistence of non-mobile plasmids, the remaining models for plasmid survival rely on the ability of plasmids to spread by conjugation. Therefore, it has been challenging to explain the high number of plasmids without known transfer systems found in natural bacterial isolates[88]. These plasmids are often small (< 10kb) and replicate by RNA-controlled mechanisms, e.g. colE1-type replicons, that are kept in relatively high copy numbers[105]. *San Millan et al.* showed that genes can benefit from a location on multi-copy plasmids, due to the increased dosage and evolvability experienced by plasmid-borne genes compared to

single-copy chromosomal genes[124]. The same study also demonstrate an important trade-off between copy number and fitness cost of plasmids, which highlights the necessity of positive selection in the maintenance of these plasmids[124]. In another study, *San Millan et al.* demonstrate that the fitness landscapes experienced by plasmids can be more complex than previously thought, with some plasmids modulating the costs of co-residing plasmids to increase their combined fitness[154].

Plasmids tend to be costly upon arrival in a new host and with no transfer mechanisms to compensate their cost, non-transferable plasmids require positive selection for their long-term maintenance[114,124]. However, a pivotal study by *Bouma and Lenski* published in 1988 showed that initially costly plasmids can co-evolve with their host to become cost-free, and even beneficial, in the absence of selection[155]. These observations have since been repeated for many different plasmids[101,114,156–160] and imply that plasmids can in principle persist with minimal transfer or selection (**Figure 6**)[161]. It is not known to what extend these cost ameliorations happen in nature, and whether they entail tradeoffs in other (natural) environments that we are not aware of. Therefore, much work is still needed if we want to understand the (long-term) role of these mechanisms on the survival of natural plasmids in their native habitats.

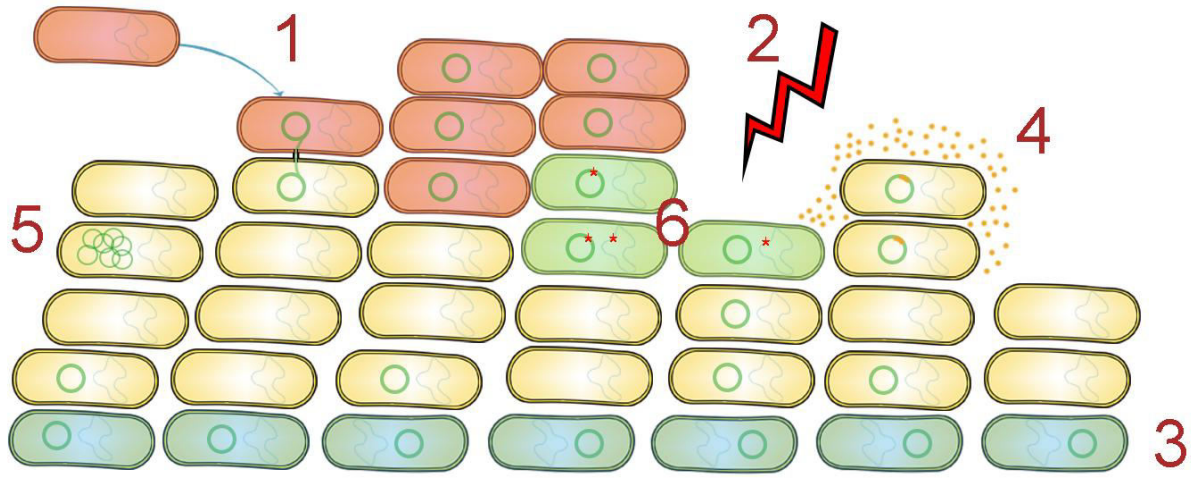


Figure 6. Plasmid-located genes may be selected over chromosomal genes in spite of the (initial) fitness burden imposed by plasmid-carriage. Plasmid-survival has been attributed to several qualities of plasmids and natural phenomena that increase their benefit compared to chromosomal genes. **1**) Mobile plasmids may transfer to incoming superior genetic backgrounds to improve their survival[116]. **2**) Sporadic selection for plasmid-borne traits allows plasmid-enrichment e.g. from stable reservoirs such as metabolically inactive layers of a biofilm (**3**)[145,149]. **4**) Social traits are more easily maintained on mobile plasmids due to their re-infective properties[147]. **5**) Multi-copy plasmids allow for higher gene-dosage and evolvability, which may benefit certain genotypes[124]. **6**) Plasmid-host adaptations can occur (even for very costly plasmids if they are temporarily kept by selection) to increase plasmid persistence[101,114,155,162].

Chapter 3

3 Barriers to horizontal gene transfer

A fundamental aim of evolutionary biology has been to understand why some genes are more widely disseminated compared to others. Because extensive genetic transfer has shaped many modern life forms and occurs readily across the tree of life, by transfer of small genetic segments to whole chromosomes, it has been proposed that no insurmountable barriers to HGT exist[163]. However, from laboratory experiments we know that some genes can be fully resistant to transfer between certain hosts due to their toxic effects[164]. Other studies, on environmental antibiotic resistance genes, show that the total functional gene pool of resistance genes far exceeds those found in human pathogens; even though functional-selection studies show that these genes can confer resistance in a relevant species[14,165–167]. While these studies are often experimentally biased in terms of unnaturally high expression levels, they indicate that lateral gene flow is restricted by more than absolute functionality. In the following, I will discuss what is currently believed to be the most important factors implicated in the successful transfer and persistence of genetic elements.

Mobilisation

If a gene is never mobilised from its original host genome, it is unlikely to experience transfer. As summarized earlier in this chapter, there are several modes and elements of transfer that can carry a gene from one cell to another (**Figure 4B**)[69]. Although some of the mechanisms (transformation and generalized transduction) do not depend on a mobile genetic context, the remaining transfer mechanisms are highly dependent on the association with elements that are able to mobilise a gene from its stationary position on the host chromosome. Intracellular mobilisation can happen via transposable elements or integrons that may move genes between the chromosome and e.g. a plasmid vector able to mediate intercellular transfer between genomes. Studies comparing genes of the environmental resistome to that of human pathogens, suggest the main difference between these two groups to be the increased frequency of mobility elements near resistance genes in human pathogens[14,168,169]. Looking into the genomic context of the genes used in **manuscript V** we classified every gene as either mobile or non-mobile depending on the presence or absence of flanking genes associated with mobility e.g. transposase, integron, conjugation or phage related genes. This classification was a strong predictor of dissemination (Mann-Whitney U-test, $P < 0.001$); supporting the idea that initial mobilisation is a crucial step in gene dissemination (**Figure 7**).

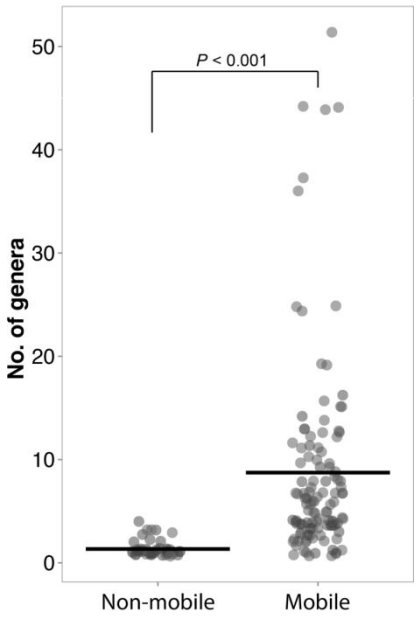


Figure 7. The association of antibiotic resistance genes with mobile elements is a strong predictor of dissemination across bacterial genera (Mann-Whitney U-test, P < 0.001). Resistance genes from **manuscript V**.

DNA entry and genomic defence mechanisms

The next step towards successful gene dissemination is the association of a gene with a transfer mechanism able to carry the gene between cells. Due to the trade-offs between evolvability and genetic integrity, most prokaryotes have evolved strategies to prevent excessive genetic flux from polluting their genomes[69]. As only a minority of species are able to take up naked DNA from the environment, successful gene transfer relies largely on recipient compatibility with the vectors implicated in HGT[69]. The likelihood of transfer by these vectors depend on e.g. the presence of phage receptors, replication (stability) and conjugation proficiency of a plasmid or DNA homology supporting recombination if the vector is not autonomously replicating[69,170].

Enzymatic restriction is a classical example of a more active defence mechanism present in many bacterial species that functions to protect the recipient genome against incoming DNA[108]. These systems are encoded on chromosomes as well as on plasmids, and come in pairs of restriction and modification enzymes that ensure, via specific DNA methylation patterns, that only foreign DNA is recognized and cut by the restriction nuclease[108]. Consequently, although restriction enzymes do not impose an absolute barrier[171], some phage and plasmid genomes encode anti-restriction proteins that mimic certain DNA-motifs to decoy restriction enzymes and improve infection rates[172]. Many plasmids also encode restriction-antirestriction systems themselves, presumably to fight off competing mobile elements and increase their stability[173].

Another way that plasmids can avoid competition from similar plasmids in the same cell, is via surface exclusion mechanisms[174]. These mechanisms are encoded within the conjugation operon and act on the cell envelope to reduce pilus binding and DNA entry[69].

More recently, the sequence specific and highly adaptable properties of CRISPR-Cas systems have been recognized for their usefulness as a tool in genetic engineering[175]. The CRISPR-Cas system functions as an adaptive immune system that protects bacterial and archaeal hosts against phages and plasmid intruders[176]. While the CRISPR-Cas family is diverse, all members rely on the same principles of incorporating DNA "spacers" into chromosomal CRISPR-arrays that are subsequently transcribed to precisely target foreign DNA[176]. Tight CRISPR-based immunity can be highly advantageous in the presence of lytic phages, but might also hinder the entry of beneficial genetic elements[177]. However, this trade-off seems balanced via the frequent escapers emerging from the recombination prone repetitive nature of CRISPR arrays; creating the flexibility in population immunity that allows the occasional sampling of beneficial genetic elements[177]. An interesting example of this plasticity of CRISPR-Cas defence was demonstrated by sequencing of CRISPR loci in enterococci isolated before and after the common use of antibiotics[178]. In this study, *Palmer* and *Gilmore* find a strong inverse correlation between the presence of CRISPR-Cas loci and acquired antibiotic resistance genes; suggesting that the advantage of resistance outweighed the benefit of CRIPSR-Cas defence in this case[178].

3.1. The biological cost of gene acquisition

When a gene has successfully entered a new host, it will either be retained or lost through purifying selection or genetic drift. The probability of a genetic element to persist in a population is partly determined by its contribution to the fitness of the cell, and the potential cost of resistance genes and their plasmids is believed to play a key role in the persistence of antibiotic resistance[179,180].

The biological fitness of an organism is defined as its reproductive success in a certain environment[180]. Strictly, fitness is defined by the number of extra individuals per generation: $|\omega| = N_{aftet} / N_{before}$, where N is the number of individuals (carrying a certain trait) before and after one generation[181]. Fitness can be manifested through many different phenotypes and includes the ability to reproduce (grow), survive and transmit. From a gene-centric point of view, fitness is determined by the success of the replicative unit, e.g. genome, plasmid or transposon, in which the gene is located[149].

A gene can contribute to its own reproductive success by encoding a beneficial phenotype that increases survival of its reproductive locus. A metabolic gene that allows more efficient utilization of resources or an antibiotic resistance gene mediating detoxification will enable faster growth of their

host when nutrients are limited or xenogenic compounds are present. Some genes may increase overall survival while not directly affecting growth e.g. by allowing efficient adhesion in a dynamic bladder or gut environment where rapid elimination from a preferred niche would otherwise reduce the population size[182,183]. In contrast, horizontal transmission of genetic elements e.g. plasmids or phages often comes at the expense of reduced vertical transmission, but might still pay off because overall vector propagation can still be supported through decupling from vertical inheritance[92,138].

It has long been known that the introduction of foreign genetic elements, such as cloning vectors, inflict changes to host homeostasis that affects measures of fitness[184–186]. Such effects have been suggested to stem from the "metabolic burden" imposed by plasmid replication and gene expression that draws on cellular resources, but might also stem from competition with, or disruption of, essential processes in a broader sense[187–190].

DNA maintenance is cheap but gene expression is costly

While there is some correlation between the number of resistance genes on plasmids and their cost, there is no strong correlation between host fitness and plasmid size[191]. It is also intriguing that a whole (1.66 Mb) bacterial chromosome can be transplanted into yeast cells without noticeable growth defects[192]. Although DNA replication does require cellular energy, and drains on carbon metabolism, the cost of DNA replication itself is most often negligible compared to downstream processes; especially in nutrient rich environments[193–195].

Because the informational content in DNA is expressed via an amplification cascade of transcription and translation, the cost of these processes is inevitably higher than the cost of replication. Furthermore, their individual contribution to the total cost of protein expression depends on the nutrient availability; indicating that limited metabolic resources play a role in the cost of these processes[196]. While mRNA synthesis does spend cellular resources, a more pronounced cost of transcribing incoming gens lies in the sequestering of RNA polymerases that might be limiting for the transcription of vital processes[197]. This was supported by *Lamberte et al.* showing that the introduction of AT-rich DNA is especially costly due to the higher probability of intragenic promoter motifs, and by *Yona et al*. demonstrating that the *E. coli* genome has evolved to avoid (expression from) intragenic promoters[197,198].

Similarly, the cost of translation has been attributed to titration of ribosomes and tRNAs form more vital cellular processes[195]. Theoretical work suggests that the material cost of amino acids can be subject to selection, and that inefficient translation, caused by suboptimal codon usage, can decrease fitness[199,200]. Suboptimal codon usage may also cause translational problems due to, for

example, tRNA depletion and ribosomal stalling. From the idea that codons employing the most abundant tRNAs allow for a higher expression, the term “codon adaptation index” (CAI) has been established to reflect the codon preferences of highly expressed gens within a genome[201]. However, while selection for optimal codon usage in highly expressed genes of fast growing bacteria is well established, it has been challenging to identify a universal selective force acting on codon bias across organisms and life styles[202–204]. This lack of a clear selective signal is likely due to selection for other qualities such as optimal protein folding, reduced mRNA interactions or improved stability (**Figure 8**)[199,205].

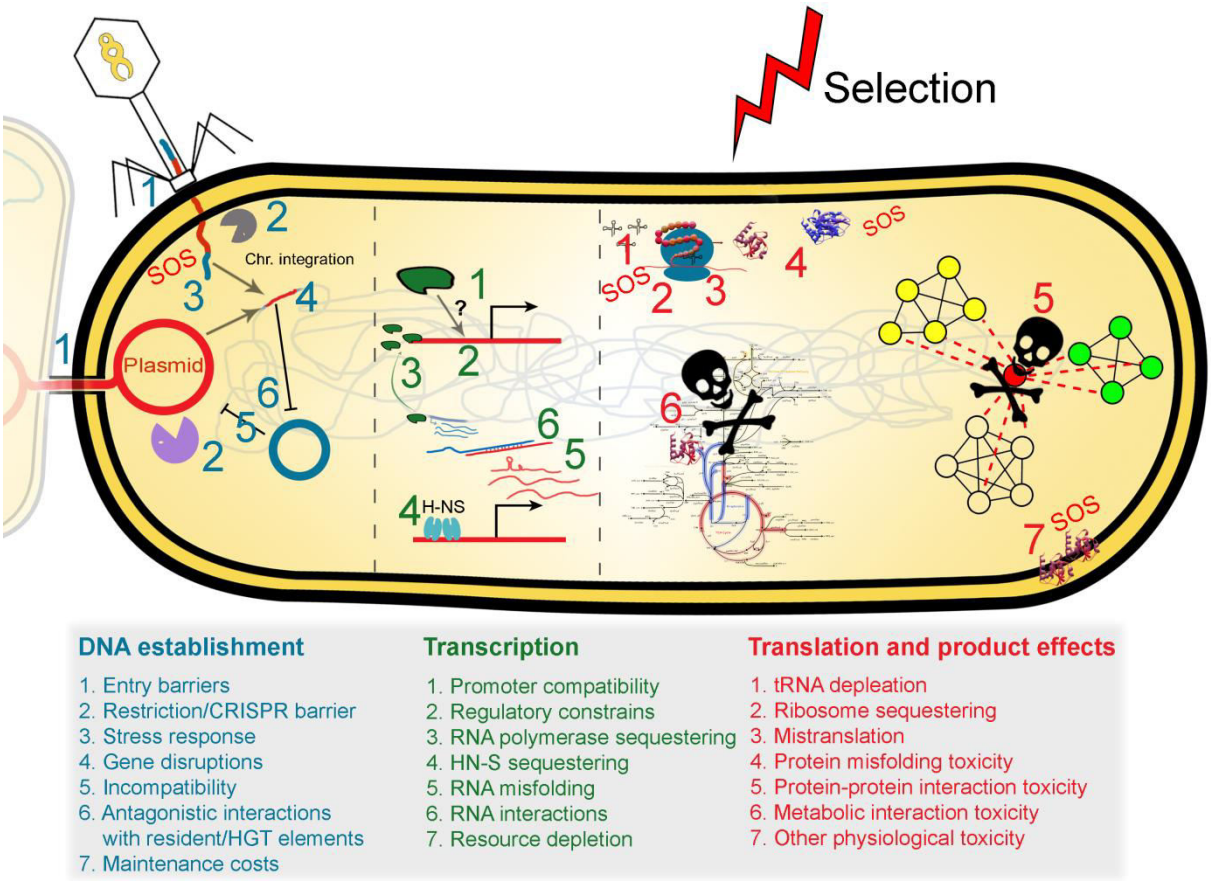


Figure 8. Molecular factors affecting the establishment and persistence of horizontally transferred genes. Apart from initial mobilisation and vector compatibility, the selection of transferred genes is determined at multiple functional levels by the effects on host fitness and external (selective) forces.

Costs originating from disruptive protein behaviour

Although the costs associated with replication, transcription or translation are of evolutionary importance (e.g. as evident from codon selection[203]), these costs are often low compared to the potentially deleterious effects inflicted by the encoded proteins of incoming DNA (**Figure 8**)[190,206–208]. From protein production efforts, it is well known that too highly expressed proteins tend to misfold and might lead to the formation of insoluble "inclusion bodies" that can be toxic to the cell[206,209]. These effects seem to impose evolutionary constrains on highly expressed proteins to optimize their folding properties[209]. This is in line with the observations of *Tomala and Korona* who used an expression library of *Saccharomyces cerevisiae* native proteins to show that the cost of proteins did not rely on translational features such as mRNA-folding or CAI but rather on structural features of the expressed proteins[207]. The authors show that especially transmembrane and unstructured protein regions, together with protein length, were the main predictors of fitness cost related to native protein expression in *S. cerevisiae* [207].

Numerous proteins have been shown to incur a cost when heterologously expressed for production purposes, due to the direct or indirect (e.g. product) toxicity of their function[190,210–213]. The costs of disruptive protein functions are also evident from mobile elements[189,214–217]. For example, plasmids costs can arise from the interaction between chromosomal and plasmid proteins. For example, the TrfA1 plasmid replication initiation protein reduces fitness of the *Shewanella oneidensis* host by antagonistic interactions with the DNA helicase DnaB[189], and similar costly interactions between chromosomal and plasmid encoded proteins have been observed for other host-plasmid combinations[92,101,214,218].

The cost of transferable resistance genes

It is commonly assumed that acquiring (mutational or horizontal) antibiotic resistance entails a burden on the host[179,191,219]. However, in contrast to mutational resistance, little effort has been made to quantify and study the fitness effects of transferable resistance genes (**Table 1 and manuscript V**)[179,180,219]. Although the costs of resistance are offset by the strong selection of antibiotics and occasional tight regulation, they likely play a key role in the long term persistence and dissemination of resistance[166,220,221]. While few reports assessing the cost of transferable resistance exist, these reveal a span of fitness effects where some genes are associated with high costs (> 20% fitness reduction) and others have no measureable cost, and may even be beneficial, in the absence of antibiotics (**Table 1 and manuscript V**). To compensate for a potentially high cost of resistance, the expression of some resistance genes is regulated to ensure a low cost of carriage when antibiotics are absent[222].

While there are examples of regulated drug-modifying genes e.g. chromosomal *ampC* β-lactamase regulated by the ampR repressor, tight regulation is more often seen for genes within the efflux or target-modifying mechanistic classes[222]. Important examples are the vancomycin resistance conferring Van-clusters, the tetracycline efflux encoding *tetA* gene and many enzymes involved in ribosomal methylation[29,222,223].

Table 1. Fitness effects of heterologously expressed resistance genes reported in the literature. All fitness cost estimates were acquired by *in vitro* measurements (growth rate or competition assays).

Antibiotic class	Gene	Fitness cost	Expression level	Implicated mechanism	Host bacterium	References
Tetracycline	*tetA*	Medium	High (unregulated)	Potassium metabolism	*E. coli*	[215,224]
Vancomycin	*VanA and VanB clusters*	High	High (unregulated)	Cell wall synthesis	*S. aureus, E. faecalis, E. faecium*	[217,223]
Linezolid, Macrolides	*cfr + ermB*	High	Medium	Ribosomal methylation	*S. aureus*	[225]
Linezolid	*cfr*	Low	Medium	Ribosomal methylation	*S. aureus*	[225]
β-lactams	*SME-1*	High	Very high (pUC)	Signal peptide	*E. coli*	[226]
β-lactams	*OXA10, OXA24*	High	Very high (pUC)	Cell wall	*E. coli*	[227]
β-lactams	*SFO1*	High	Very high (pUC)	Cell wall	*E. coli*	[227]
β-lactams	*AmpC*	Medium	High (unregulated)	Unknown	*S. enterica*	[228]
β-lactams	*AmpC*	Low/none	High (unregulated)	Unknown	*E. coli, E. cloacae, P. aeruginosa*	[228,229]
Aminoglycosides	*rmtC*	Low/none	Low	Ribosomal methylation	*E. coli*	[230]
Quinolones	*qnr* genes	Low/beneficial	High (pBR322)	Gyrase interactions	*E. coli*	[224,231]

3.2. The context matters

"*Everything is everywhere, but the environment selects*"

- Lourens Gerhard Marinus Baas Becking (1934)[232]

It was recently suggested that universally fit resistance-conferring mutants exist[233], however, such notions are likely an artefact of unexplored fitness trade-offs in known or unknown environmental conditions[234]. Disseminated antibiotic resistance genes are often found in variety of different genetic, genomic and environmental contexts; and have likely experienced many more on the way to their present location. Therefore, the cost of a gene can only be viewed in relation to its context, and there are numerous examples of variability in fitness effects of mutations and gene acquisition between hosts and growth media (**Table 1**)[101,126,191,228,235,126]. Apart from the modulation of fitness effects, the genetic and extracellular environmental context may also affect evolutionary trajectories that compensate the costs of resistance[101,126,236].

The external environment

Antibiotic resistance genes are most beneficial in the presence of antibiotics, but their fitness effects in the absence of antibiotics may depend on the external as well as the internal (host) environment (**Table 1**). This is especially important when considering co-selection by adjacent genes on multi-gene units such as plasmids, were the combined selective effects are hard to predict[126]. For example, many pathogen-carried antibiotic resistance plasmids often contain genes annotated to have metabolic or virulence functions (**manuscript I**)[115]. Such features may be beneficial during infection or when certain nutrients (e.g. iron) are limited, but may also provide broader benefits of e.g. adhesion and biofilm formation, which does not depend on antibiotic selection[93].

Fitness is often measured as growth rates or head-to-head competitions in the laboratory[191,237]. While these measures may overlap well with more complex *in vivo* environments[115,191], as the reader will see in **manuscript II** of this thesis, there are exceptions. Ideally, fitness should be measured in exactly the situation we wish to know more about such as the human body for pathogenic bacteria. By continuous sampling from natural environments, we obtain a direct measure of fitness through e.g. clone abundance, but isolating the factors responsible for selection is complicated and require simplified experimental setups[115,238]. However, there are obvious ethical complications if we want to carry out controlled experiments in humans, and it is still challenging to account for differences between individuals, e.g. in their microbiota or immune factors[239]. In addition, most natural bacterial isolates studied in the context of antibiotic resistance are likely to experience environmental fluctuations, e.g. those of a human digestive tract or immune response, that makes it hard to obtain one "true measure" of fitness in simplified experimental setups.

The human gut environment

"*It has been shown that (gene) transfer in the intestinal tracts of animals and humans occurs ad libitum; it's a bordello down there!*" – Davies and Davies 2010[4]

The human gut exemplifies an especially important microbial environment that facilitates HGT, and it has been suggested as a significant reservoir through which pathogens acquire resistance and virulence traits[58,115,169,238]. Being in close contact with our body, pathogens may cause infection in the gut or be transmitted from the gut to infection-prone areas such surgery wounds or the urinary tract [115,240,241].

The colonization of the gut by bacteria takes place within a few days after birth, where facultative anaerobic bacteria, mainly *E. coli* and enterococci, are the first to colonize followed by anaerobic genera of *Bifidobacterium*, *Clostridium* and *Bacteroides* in response to lower oxygen levels and the nutrients from breast milk[242]. The speed and diversity with which the gut is colonized depends on the route of delivery and the general hygiene of the environment, and 45% of Swedish infants are colonized by a mean of 2.1 distinct *E. coli* lineages three days after birth[243]. As a result of lower hygiene, these numbers are much higher in developing countries, where a similar group of infants carried 8.5 *E. coli* lineages on average[244]. *E. coli* is highly adapted to the gut environment and readily reaches numbers of $>10^{10}$ CFU/g faeces in the infant gut after which it declines to 10^{6}-10^{8} CFU/g in most adult carriers[242]. Other members of the *Enterobacteriaceae* family e.g. *Klebsiella* and *Enterobacter* species are common in the infant gut, but lack the competitive fitness to persist in high numbers into adulthood[242,245]. As solid foods are introduced, the infant microbiota shifts towards the adult composition with increased abundance of *Bacteroides*, *Ruminococcus, Lachnospiraceae, Prevotellaceae* and *Clostridium*, accompanied with a decrease in *Enterobacteriaceae*, *Bifidobacterium* and *Lactobacillales*[245].

The human gut is a complex environment, far from that of a Petri-dish, and its microbial inhabitants are subjected to fluctuations in oxygen levels, pH, nutrients, physical forces, suppression from the host immune system and competition from other microbial inhabitants[242,246]. In response to these harsh conditions, gut inhabiting bacteria have evolved numerous strategies to optimize their colonization[246]. Such strategies may provide unambiguous benefits (e.g. gut-specific nutrient utilisation) that would not be detected in most laboratory setups.

The physical turnover of intestinal content promoted by peristaltic movements expels bacteria from the intestines unless they are able to attach physically to the intestinal surface. Therefore, outer membrane adhesion structures such as pili or lectins that attach to epithelial cells or mucus layers are important colonization factors[247]. Due to their accessibility and role in pathogenesis, these

surface structures are also preferred recognition sites and antibody targets for the immune system. Although similar mechanisms are widely used by harmless commensals, many colonization factors may therefore also be categorised as "virulence" factors or *vice versa*[246]. Due to the co-localisation of colonization factors and antibiotic resistance genes, on the same mobile elements within the gut microbiota, co-selection may play a role in their persistence[248].

A general strategy employed by epithelial immune cells to regulate the microbiota is the secretion of antimicrobial peptides (AMPs)[249]. While these peptides targets broadly by interacting with conserved cell envelope structures, some strains have developed increased tolerance to AMPs e.g. by reducing negatively charged residues of the outer membrane[246]. In addition to modulation by the immune system, direct competition also takes place between gut inhabiting strains. Many bacteria produce their own antimicrobial peptides (bacteriocins) to directly inhibit competitors and these are often encoded on plasmids; likely due to their social nature and role in plasmid stability[250].

Apart from physical niche competition and immune evasion, a major competitive factor of gut microbes is the ability to utilize limited nutrients. For example, some successful species of the infant gut are able to utilize oligosaccharides from breast milk or digest intestinal mucus layers, and such differential feeding strategies are important host strategies in the establishment of a healthy microbiota[245]. Such strategies are especially important in response to infection, where the immune systems can actively remove nutrients such as iron, zink, and manganese by secretion of sequestering proteins to limit the growth of pathogens[246]. Increased secretion of reactive oxygen species (ROS) by the infected host tissue also serves to limit pathogen colonization and many superior colonisers express enzymes (e.g. catalase) to reduce oxidative damage[246]. Interestingly, some enteropathogens take advantage of this mechanism to outcompete the normal gut microbiota by anaerobic respiration of products formed from ROS oxidised enterocyte metabolites upon inflammation[251].

The environment as a barrier to HGT

Even if a gene is successfully mobilised and is able to contribute to the shared gene pool of an isolated community, physical distance between donor and recipient is an obvious barrier to HGT[252]. Whereas conjugation requires direct cell-to-cell contact, the distance reached by transformation and transduction depends on the stability of the free DNA or phage particles[253]. Nonetheless, with the broad transfer ability of some HGT mechanisms it is not unlikely that a chain of transfer events may connect ecologically different gene pools to overcome geographical distances[166]. However, if organisms occupy a narrow environmental niches, they are less likely to encounter potential donor organisms, which is evident from the reduced and uniform genomes of most endosymbionts[65].

Gene exchange network analysis have shown ecological preferences for most transfer events

happening within the human microbiome, soil habitats and other environments (**Figure 9**)[13,254,255]; however genes conferring broadly beneficial traits such as antibiotic resistance are disseminated widely across ecological niches. These observations suggest that the barrier posed by ecological differences may largely be a question of functional displacement, rather than ineffective transfer, where genes conferring niche specific or universal benefits are less likely to be lost in a given environment[58,254]. As such, the external environment becomes critical to the propagation of transferable genes, but the selective effects, and thus overall stability of transferred elements, can also be highly dependent on the internal, genomic environment, in which they arrive[101,110].

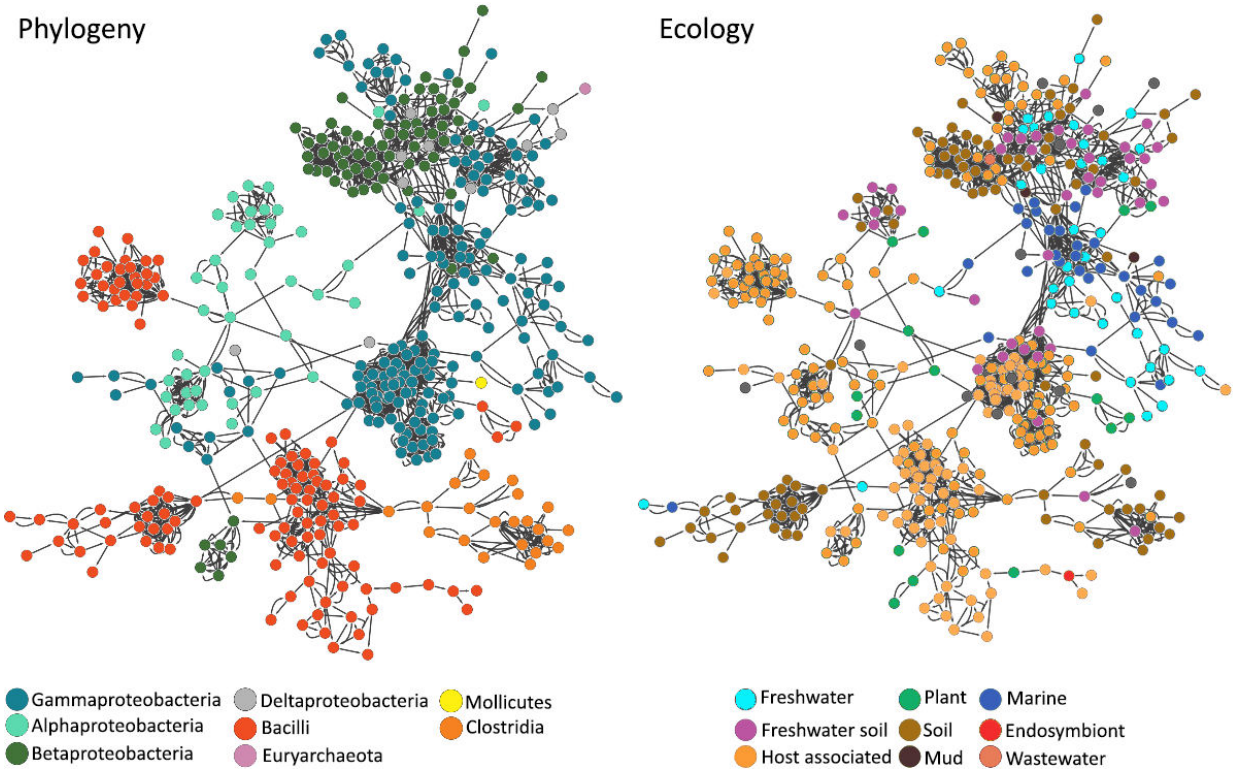


Figure 9. Gene exchange networks colored by phylogenetic and ecological associations. Networks adapted from Popa and Dagan 2011[252].

3.3. Phylogenetic factors affecting HGT

Although antibiotic resistance genes are present in many environments, the do not move unhindered across ecological niches[13], and more recent studies suggest phylogeny as a stronger barrier for genes of the human, animal and soil resistomes[12,168]. In addition, gene exchange network analysis across gene categories suggests that HGT primarily happens between closely related genomes (**Figure 9**)[82,168,252,255,256]. Reasons for this phylogenetic confinement have been discussed at different levels: from vector entry to compatibility of expression and propagation machinery, to selection at different levels[252].

All transfer mechanisms show some restrictions in their host-range due to, for example, receptor availability or recombination proficiency[69]. While narrow host-range plasmids can only replicate in a few species with similar hosts replication proteins, many broad host-range e.g. IncP plasmid vectors are able to transfer much beyond the observed gene-exchange networks of most transferred genes[110,111,255,257]. DNA similarity is important for transformation events, where homologous recombination is needed for stable genomic integration[51,255]. Whereas other ways to integrate DNA e.g. transposons, integrons, site-specific or illegitimate recombination exist, plasmid vectors and certain phages are not dependent on stable integration events for their propagation[255,258].

Sequence level features as barriers to HGT

The sequence similarity of donor and recipient genomes is a strong predictor of transfer frequency, suggesting that fundamental genetic barriers exist[82,255,259–261]. Apart from the effects of sequence homology on recombination efficiency, more general features such as the GC-content and codon usage have been suggested to influence compatibility of transferred genes with the recipient genome[255,259,260,262].

GC-content

An interesting observation is that the majority (86%) of predicted recent transfers have taken place between genomes differing less than 5% in their GC-content, which suggests a role of sequence composition in the establishment of foreign genetic elements[255]. One relatively well studied mechanism used by proteobacteria to control the expression of incoming genes, is the binding of foreign DNA by nucleoid structuring proteins; including H-NS and its homologs[263]. These proteins recognize sequences of lower GC-content, compared to the genome average to bind and silence potentially costly heterologous genes[264]. This cost can be partly attributed to the intra- or extragenic sequestering of RNA polymerase by AT-rich DNA in the acquired DNA[197]. Apart from its physical binding of transferred DNA, the role of H-NS in HGT is further supported by the existence of plasmid encoded homologs and the, seemingly co-evolved, role of H-NS in the regulation of conjugative transfer genes[265,266]. Furthermore, transferred genes tend to have slightly lower GC-content compared to their host genomes; suggesting a selection for some degree of H-NS silencing[267].

A study by *Raghavan et al.* suggests a direct effect of the GC-content on the fitness of *E. coli*[268]. In this study, the growth rate of highly expressed synonymous *gfp* constructs, spanning a GC-range from 40.4–53.7%, was measured to be significantly increased for the high-GC variants[268]. While GC-dependent fitness effects are interesting, they have only been demonstrated for (unnaturally) high expression setups in *E. coli* and a mechanistic basis is still lacking[269].

Codon usage

There seems to be a connection between codon usage and the transferability of genes within microbial communities[259,260,270]. Computational studies suggest that high codon similarity between donor and recipient genomes is crucial for HGT[260,270]. However, other studies propose that this is only true to a certain point, because excessive expression may be deleterious (in agreement with the H-NS model) to gene retention[259].

Functional studies performed on synonymous libraries of highly expressed *gfp* variants have revealed that an elevated CAI of the entire ORF was beneficial for growth, but the CAI did not predict expression very well[200]. In contrast, the codons determining mRNA-folding energy of the N-terminal (–4 to +37 positions) strongly predicted the expression of codon shuffled *gfp*[200,271]. A strong folding i.e. secondary structure of the mRNA close to the translation initiation site is thought to slow ribosome progression with a much more profound effect than that of e.g. suboptimal tRNA accessibility experienced by low CAI genes[272]. Therefore the influence of mRNA-folding could also be a valuable predictor to include when investigating barriers of HGT (**manuscript V**).

A general problem arising when correlating the influence of sequence composition to transferability predictions is that causality is hard to establish. Because the sequence composition is similar within closely related organisms, there could be several other explanations for why sequence similarity between donor and recipient genomes is frequent. A common evolutionary and/or ecological history might entail overlapping physiology, niche occupancy, transfer vector compatibility and selective conditions, which likely have a much more profound influence on HGT than that of the sequence composition itself (**manuscript V**)[252]. Another explanation could be that organisms more engaged in HGT with each other, may have co-evolved their translational machinery to support a common gene-pool of e.g. certain functional qualities[262,270].

Functional barriers to HGT

The integration of genetic innovations that have (co-)evolved with distantly related genomes can pose a functional challenge[252]. Phylogenetic barriers might be formed by incompatibility with functional elements, other than the open reading frame itself, which may hinder proper functional integration of transferred genetic elements. For example, the expression of a newly acquired gene might be compromised if the native expression platform is incompatible with recipient machinery[252]. However, recent findings show that mutations readily occur to achieve *de novo* promoter function or to compensate the influence of suboptimal sequence composition on expression levels[198,273]. Such observations along with the broad host-range of certain plasmids, imply that the compatibility of transcriptional machinery might not be a major barrier to HGT[257].

A repeated observation in HGT literature is a strong bias in the biological functions of transferred genes[274]. This bias was proposed already in the early genome sequencing days, following the observation that so called "informational" genes were underrepresented among recently transferred genes compared to "operational" genes[275]. Whereas the first of these gene categories covers the genomic core genes involved in information processing tasks such as translation, transcription and replication, the latter "operational" class include more peripheral biosynthetic, metabolic and regulatory functions[275]. There are, however, important exceptions of e.g. polymerases and DNA repair genes widely present on plasmids, and the original hypothesis was later refined by *Jain et al.* in 1999[276–279]. In this paper, they detail their original hypothesis by arguing that the distinction between informational and operational genes could be attributed to differences in the complexity of the contexts in which they operate[277]. This "complexity hypothesis" states that gene products highly dependent on (co-evolved) interaction partners were less likely to be transferred alone, because their functional integration in a foreign genomic context would require compatibility with all interaction partners in the recipient genome to be successful[280]. That is, the more interactions a protein require for correct functionality, the more interactions can go wrong to damage existing, potentially vital, interactions resulting in lower host fitness[281]. These constraints are also evident from the limited adaptive freedom of highly connected genes[282]. Several years later, the notion was further supported by the availability of protein-protein interaction data that showed a strong signal for connectivity and the propensity of a gene to undergo HGT[274,280].

Whereas the complexity hypothesis has received substantial support *in silico*[255,276,277,280,282–284], few studies have subjected its claims to experimental testing and have yielded somewhat conflicting results[164,281,285–287]. These studies investigate the functional compatibility and fitness effects of supplying, or replacing, informational genes with distant homologs. *Sorek et al.* find that ribosomal genes were significantly underrepresented in the cloneable fraction of whole genome sequencing libraries from multiple Sanger-sequenced organisms when expressed in *E. coli*; suggesting that these genes are toxic, at least when supplied in a multi-copy vector[164]. A recent study by *Kacar and Garmendia et al.* shows an increased fitness cost of replacing elongation factor Tu with its distant homologs in *E. coli,* and similar results have been obtained by *Lind et al.* when replacing ribosomal subunits in *Salmonella typhimurium*[281,287]. In contrast, the supplementation of distant homologs of acetyl-CoA carboxylase or RNA polymerase b subunits did not impose significant costs in *E. coli*[285,286]. Furthermore, both *Lind et al.* and *Kacar and Garmendia et al.* demonstrate that increased gene dosage can ameliorate the initial costs of foreign gene replacement[281,288]. Hence, it seems likely that

highly connected genes are generally, but not always, costly and that this barrier may be more flexible than previously thought.

Given the preceding discussions, it is evident that gene dissemination is restricted in numerous ways. However, considering the time of prokaryotic existence and the speed with which antibiotic resistance genes have spread in response to antibiotic pressure[289], it seems likely that most genes can be mobilized and disseminate to some extent. Even if these genes, or their mobile elements, are subject to strong selective or functional barriers, compensatory adaptive strategies may enhance fitness to encourage proper integration into new genomic contexts[136,162,198,223,273,290]. Finally, second-order selection and sheer chance (founder effects) may also play a substantial role in the transfer patterns observed. For example, some genes might be disseminated due to co-selection by other beneficial or addictive (TA-systems) genes and some may, by chance, be frequent in an environment just because they arrived first[166].

Hitherto, most studies have been largely observational and leave room for a more holistic understanding of the factors governing successful HGT. To encourage this progress, the work conducted in this PhD thesis sought to map HGT *in situ* and experimentally address the barriers proposed by *in silico* observations using a large-scale and diverse synthetic library of naturally occurring genes.

Concluding remarks and future directions

Life evolves from a set of common principles dictated by molecular mechanisms, environmental selection and a fierce competition for survival. Research into these principles has allowed us to predict, and even control, the evolutionary trajectories of life forms and that knowledge has advanced our existence in many areas including agricultural breeding, modern biotechnology and disease control[291–293]. However, for the fast growing microbes, evolution takes place at a rate that sometimes leaves us as vacant spectators to the fascinatingly threatening plasticity of these ever-adapting life forms.

Imposing a strong selection pressure on bacteria will eventually lead to their adaptation. While controlled adaptive evolution experiments are exploited in the engineering of strains for more sustainable bio-based production of chemicals, unharnessed evolution can also be detrimental to large-scale production scenarios (**manuscript IV**). In medicine, unintended evolution is exemplified by the important lesson of antibiotic resistance that is increasingly compromising the treatment of critical pathogens[2]. This medical crisis is largely caused by the overuse of antibiotics, and therefore (global) regulatory policies, encouraging more rational use of antibiotics along with accelerated drug development, are the main efforts necessary to halt resistance development[21].

At the genetic level, evolutionary principles are shaping the way bacteria become resistant, and understanding the limitations of evolutionary processes involved in mutational and acquired antibiotic resistance is critical if we want to predict and prevent future resistance spread and *de novo* evolution of pathogenic traits. For instance, the assessment of the propensity of relevant genes, and their mobile elements, to engage successfully in HGT should be an integral part of developing new antibiotics[294].

Being widely implicated in HGT, the role of plasmids in our current pandemic of antibiotic resistant pathogens should not be underestimated. I invite the reader to consider a broader view of plasmids as plastic entities navigating dynamic fitness landscapes that continuously shape, and still are being shaped by, different genomic and environmental contexts. Therefore, steps towards elucidating the evolution and interaction of endemic clones and plasmids, in their natural habitats, is needed to fully comprehend the relevance of different environmental reservoirs and selective factors in the dissemination of antibiotic resistance among critical bacterial clones. Such insight will not only improve our understanding of antibiotic resistance, and bacterial evolution in general, but the harnessing of unwanted evolution will also aid in our ability to engineer more robust biological systems for industrial and medical applications (**manuscript IV**)[292].

References

1. Microbiology by numbers. *Nat. Rev. Microbiol.* **9,** 628–628 (2011).

2. WHO. *Antimicrobial resistance. Bulletin of the World Health Organization* **61,** (2014).

3. Andersson, D. I. & Hughes, D. Selection and Transmission of Antibiotic-Resistant Bacteria. *Microbiol. Spectr.* **5,** 1–17 (2017).

4. Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* **74,** 417–433 (2010).

5. Von Wintersdorff, C. J. H. *et al.* Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* **7,** 1–10 (2016).

6. Ehrlich P., H. S. *Die Experimentelle Chemotherapie der Spirilosen. Berlin: Julius Springer* (Springer, Julius, 1910).

7. Domagk, G. Ein Beitrag zur Chemotherapie der bakteriellen Infekionen. *Dtsch. Medizinische Wochenschrift* **61,** 250–253 (1935).

8. Fleming, A. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B.influenzae. *Br. J. Exp. Pathol.* **10,** 226–236 (1929).

9. Aminov, R. I. A brief history of the antibiotic era: Lessons learned and challenges for the future. *Front. Microbiol.* **1,** 1–7 (2010).

10. Sengupta, S., Chattopadhyay, M. K. & Grossart, H. P. The multifaceted roles of antibiotics and antibiotic resistance in nature. *Front. Microbiol.* **4,** 1–13 (2013).

11. ABRAHAM, E. P. & CHAIN, E. An Enzyme from Bacteria able to Destroy Penicillin. *Nature* **146,** 837–837 (1940).

12. Hu, Y. *et al.* The Bacterial Mobile Resistome Transfer Network Connecting the Animal and Human Microbiomes. **82,** 6672–6681 (2016).

13. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9,** 1–10 (2014).

14. Munck, C. *et al.* Limited dissemination of the wastewater treatment plant core resistome. *Nat. Commun.* **6,** 8452 (2015).

15. Pehrsson, E. C. *et al.* Interconnected microbiomes and resistomes in low-income human habitats. *Nature* **533,** 212–216 (2016).

16. Brandt, C. *et al.* In silico serine β-lactamases analysis reveals a huge potential resistome in environmental and pathogenic species. *Sci. Rep.* **7,** 43232 (2017).

17. Crofts, T. S., Gasparrini, A. J. & Dantas, G. Next-generation approaches to understand and combat the antibiotic resistome. *Nat. Rev. Microbiol.* (2017). doi:10.1038/nrmicro.2017.28

18. D'Costa, V. M. *et al.* Antibiotic resistance is ancient. *Nature* **477,** 457–61 (2011).

19. Jiang, X. *et al.* Dissemination of antibiotic resistance genes from antibiotic producers to pathogens. *Nat. Commun.* **8,** 15784 (2017).

20. Demain, A. L. & Elander, R. P. The β-lactam antibiotics: past, present, and future. *Antonie Van Leeuwenhoek* **75,** 5–19 (1999).

21. World Health Organization. The evolving threat of antimicrobial resistance: Options for action. *WHO Publ.* 1–119 (2014).

22. Barlow, M. & Hall, B. G. Phylogenetic analysis shows that the OXA beta-lactamase genes have been on plasmids for millions of years. *J. Mol. Evol.* **55,** 314–21 (2002).

23. Livermore, D. M. *et al.* CTX-M: changing the face of ESBLs in Europe. *J. Antimicrob. Chemother.* **59,** 165–74 (2007).

24. Breilh, D., Texier-Maugein, J., Allaouchiche, B., Saux, M.-C. & Boselli, E. Carbapenems. *J. Chemother.* **25,** 1–17 (2013).

25. European Centre for Disease Prevention and Control. *Antimicrobial resistance surveillance in Europe 2015. Annual Report of the European Antimicrobial Resistance Surveillance Network (EARS-Net). Www.Ecdc.Europa.Eu* (2015). doi:10.2900/39777

26. Livermore, D. M. Has the era of untreatable infections arrived? *J. Antimicrob. Chemother.* **64 Suppl 1,** i29-36 (2009).

27. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O. & Piddock, L. J. V. Molecular mechanisms of antibiotic resistance. *Nat. Rev. Microbiol.* **13,** 42–51 (2015).

28. Bush, K. & Fisher, J. F. Epidemiological expansion, structural studies, and clinical challenges of new β-lactamases from gram-negative bacteria. *Annu. Rev. Microbiol.* **65,** 455–78 (2011).

29. Munita, J. M., Arias, C. A., Unit, A. R. & Santiago, A. De. Mechanisms of Antibiotic Resistance. *Microbiol Spectr* **4,** 1–37 (2016).

30. Poole, K. *Efflux-mediated antimicrobial resistance*. *J Antimicrob Chemother* **56,** (2005).

31. Munck, C., Ellabaan, M., Klausen, M. S. & Sommer, M. O. A. The Resistome Of Important Human Pathogens. *bioRxiv* (2017). doi:http://dx.doi.org/10.1101/140194

32. Courvalin, P. Vancomycin Resistance in Gram-Positive Cocci. *Clin. Infect. Dis.* **42,** S25–S34 (2006).

33. Hiramatsu, K. *et al.* Genomic Basis for Methicillin Resistance in Staphylococcus aureus. *Infect. Chemother.* **45,** 117–36 (2013).

34. Sköld, O. Sulfonamide resistance: mechanisms and trends. *Drug Resist. Updat.* **3,** 155–160 (2000).

35. Weatherspoon-Griffin, N., Yang, D., Kong, W., Hua, Z. & Shi, Y. The CpxR/CpxA Two-component regulatory system up-regulates the multidrug resistance cascade to facilitate Escherichia coli resistance to a model antimicrobial peptide. *J. Biol. Chem.* **289,** 32571–32582 (2014).

36. Nishino, K. & Yamaguchi, A. Role of Histone-Like Protein H-NS in Multidrug Resistance of Escherichia coli Role of Histone-Like Protein H-NS in Multidrug Resistance of Escherichia coli. *Society* **186,** 1423–1429 (2004).

37. Darwin, C. The Origin of Species by means of Natural Selection. *Murray, London* **6,** 504 (1968).

38. Lawrence, J. G. Gene Transfer in Bacteria: Speciation without Species? *Theor. Popul. Biol.* **61,** 449–460 (2002).

39. Tatum, E. L. & Lederberg, J. Gene Recombination in the Bacterium Escherichia coli. *J. Bacteriol.* **53,** 673–684 (1947).

40. Syvanen, M. Cross-species gene transfer; implications for a new theory of evolution. *J. Theor. Biol.* **112,** 333–43 (1985).

41. Boto, L. Horizontal gene transfer in evolution: facts and challenges. *Proc. R. Soc. B Biol. Sci.* **277,** 819–827 (2010).

42. Hilario, E. & Gogarten, J. P. Horizontal transfer of ATPase genes--the tree of life becomes a net of life. *Biosystems.* **31,** 111–9 (1993).

43. Cohen, S. N. DNA cloning: a personal view after 40 years. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 15521–9 (2013).

44. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19,** 2226–38 (2002).

45. Ochman, H., Lawrence, J. G. & Groisman, E. a. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405,** 299–304 (2000).

46. Martin, W. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays* **21,** 99–104 (1999).

47. Lawrence, J. G. & Ochman, H. Molecular archaeology of the Escherichia coli genome. *Proc. Natl. Acad. Sci. U. S. A.*

95, 9413–7 (1998).

48. Welch, R. A. *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 17020–4 (2002).

49. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23,** 254–267 (2006).

50. Dufraigne, C. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res.* **33,** e6–e6 (2005).

51. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* **16,** 472–482 (2015).

52. Dagan, T., Artzy-Randrup, Y. & Martin, W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci.* **105,** 10039–10044 (2008).

53. Gophna, U., Charlebois, R. L. & Doolittle, W. F. Have archaeal genes contributed to bacterial virulence? *Trends Microbiol.* **12,** 213–219 (2004).

54. Watkins, R. F. & Gray, M. W. The frequency of eubacterium-to-eukaryote lateral gene transfers shows significant cross-taxa variation within amoebozoa. *J. Mol. Evol.* **63,** 801–814 (2006).

55. Guljamow, A. *et al.* Horizontal gene transfer of two cytoskeletal elements from a eukaryote to a cyanobacterium. *Curr. Biol.* **17,** 757–759 (2007).

56. Ragan, M. a, McInerney, J. O. & Lake, J. a. The network of life: genome beginnings and evolution. Introduction. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **364,** 2169–75 (2009).

57. Frigaard, N.-U., Martinez, A., Mincer, T. J. & DeLong, E. F. Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439,** 847–850 (2006).

58. Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **544,** 124–124 (2017).

59. Polz, M. F., Alm, E. J. & Hanage, W. P. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.* **29,** 170–175 (2013).

60. Binnewies, T. T. *et al.* Ten years of bacterial genome sequencing: Comparative-genomics-based discoveries. *Funct. Integr. Genomics* **6,** 165–185 (2006).

61. Norman, A., Hansen, L. H. & Sørensen, S. J. Conjugative plasmids: vessels of the communal gene pool. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **364,** 2275–89 (2009).

62. Vos, M., Hesselman, M. C., te Beek, T. A., van Passel, M. W. J. & Eyre-Walker, A. Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends Microbiol.* **23,** 598–605 (2015).

63. Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17,** 589–596 (2001).

64. Brzuszkiewicz, E., Gottschalk, G., Ron, E., Hacker, J. & Dobrindt, U. Adaptation of Pathogenic *E . coli* to Various Niches : Genome Flexibility is the Key. *Microb. Pathog.* **6,** 110–125 (2009).

65. McCutcheon, J. P. & Moran, N. a. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10,** 13–26 (2011).

66. Jutkina, J., Rutgersson, C., Flach, C. F. & Joakim Larsson, D. G. An assay for determining minimal concentrations of antibiotics that drive horizontal transfer of resistance. *Sci. Total Environ.* **548–549,** 131–138 (2016).

67. Gillings, M. R. & Stokes, H. W. Are humans increasing bacterial evolvability? *Trends Ecol. Evol.* **27,** 346–352 (2012).

68. Patel, S. Drivers of bacterial genomes plasticity and roles they play in pathogen virulence, persistence and drug resistance. *Infection, Genetics and Evolution* **45,** 151–164 (2016).

69. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat.*

Rev. Microbiol. **3,** 711–21 (2005).

70. Chen, I. & Dubnau, D. DNA uptake during bacterial transformation. *Nat. Rev. Microbiol.* **2,** 241–9 (2004).

71. Griffith, F. The significance of pneumococcal types. *J. Hyg. (Lond).* **XXVII,** (1928).

72. McCarty, M. & Avery, O. INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES II. EFFECT OF DESOXYRIBONUCLEASE ON THE BIOLOGICAL ACTIVITY OF THE TRANSFORMING. *J. Exp. Med.* 89–96 (1946).

73. Johnston, C., Martin, B., Fichant, G., Polard, P. & Claverys, J.-P. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat. Rev. Microbiol.* **12,** 181–96 (2014).

74. Domingues, S., Nielsen, K. M. & da Silva, G. J. Various pathways leading to the acquisition of antibiotic resistance by natural transformation. *Mob. Genet. Elements* **2,** 257–260 (2012).

75. Fortier, L.-C. & Sekulovic, O. Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence* **4,** 354–65 (2013).

76. Jiang, S. & Paul, J. Gene transfer by transduction in the marine environment. *Appl. Environ. Microbiol.* **64,** (1998).

77. Lang, A. S., Zhaxybayeva, O. & Beatty, J. T. Gene transfer agents: phage-like elements of genetic exchange. *Nat. Rev. Microbiol.* **10,** 472–482 (2012).

78. Fogg, P. C. M., Saunders, J. R., McCarthy, A. J. & Allison, H. E. Cumulative effect of prophage burden on Shiga toxin production in Escherichia coli. *Microbiology* **158,** 488–97 (2012).

79. Novick, R. P., Christie, G. E. & Penadés, J. R. The phage-related chromosomal islands of Gram-positive bacteria. *Nat. Rev. Microbiol.* **8,** 541–551 (2010).

80. Waldor, M. & Mekalanos, J. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science (80-.).* **272,** 1910–1914 (1996).

81. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438,** 86–89 (2005).

82. Popa, O., Landan, G. & Dagan, T. Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. *ISME J.* 1–12 (2016). doi:10.1038/ismej.2016.116

83. Ross, A., Ward, S. & Hyman, P. More is better: Selecting for broad host range bacteriophages. *Front. Microbiol.* **7,** 1–6 (2016).

84. Haaber, J. *et al.* Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells. *Nat. Commun.* **7,** 13333 (2016).

85. Halary, S., Leigh, J. W., Cheaib, B., Lopez, P. & Bapteste, E. Network analyses structure genetic diversity in independent genetic worlds. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 127–32 (2010).

86. van der Meer, J. R. & Sentchilo, V. Genomic islands and the evolution of catabolic pathways in bacteria. *Curr. Opin. Biotechnol.* **14,** 248–254 (2003).

87. de la Cruz, F. & Davies, J. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* **8,** 128–33 (2000).

88. Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & de la Cruz, F. Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* **74,** 434–52 (2010).

89. Amabile-Cuevas, C. F. & Chicurel, M. E. Bacterial plasmids and gene flux. *Cell* **70,** 189–199 (1992).

90. Bundock, P., den Dulk-Ras, a, Beijersbergen, a & Hooykaas, P. J. Trans-kingdom T-DNA transfer from Agrobacterium tumefaciens to Saccharomyces cerevisiae. *EMBO J.* **14,** 3206–14 (1995).

91. Frost, L. S. & Koraimann, G. Regulation of bacterial conjugation: balancing opportunity with adversity. *Future Microbiol.* **5,** 1057–71 (2010).

92. Turner, P., Cooper, V. & Lenski, R. Tradeoff between horizontal and vertical modes of transmission in bacterial

plasmids. *Evolution (N. Y).* **52,** 315–329 (1998).

93. Ghigo, J. M. Natural conjugative plasmids induce bacterial biofilm development. *Nature* **412,** 442–5 (2001).

94. Joshua Lederberg. Cell Genetics and Hereditary symbiosis. *Physiol Rev* **32(4),** 403–30 (1952).

95. Johnson, T. J. & Nolan, L. K. Pathogenomics of the virulence plasmids of Escherichia coli. *Microbiol. Mol. Biol. Rev.* **73,** 750–74 (2009).

96. Carattoli, A. Resistance plasmid families in Enterobacteriaceae. *Antimicrob. Agents Chemother.* **53,** 2227–38 (2009).

97. Siguier, P., Filée, J. & Chandler, M. Insertion sequences in prokaryotic genomes. *Curr. Opin. Microbiol.* **9,** 526–531 (2006).

98. Mahillon, J. & Chandler, M. Insertion sequences. *Microbioliology Mol. Biol. Rev.* **62,** 725–74. (1998).

99. Aziz, R. K., Breitbart, M. & Edwards, R. A. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* **38,** 4207–4217 (2010).

100. Durfee, T. *et al.* The complete genome sequence of Escherichia coli DH10B: Insights into the biology of a laboratory workhorse. *J. Bacteriol.* **190,** 2597–2606 (2008).

101. Porse, A., Schønning, K., Munck, C. & Sommer, M. O. A. Survival and Evolution of a Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts. *Mol. Biol. Evol.* **33,** 2860–2873 (2016).

102. Gillings, M. R. Integrons: past, present, and future. *Microbiol. Mol. Biol. Rev.* **78,** 257–77 (2014).

103. Gillings, M. *et al.* The Evolution of Class 1 Integrons and the Rise of Antibiotic Resistance. *J. Bacteriol.* **190,** 5095–5100 (2008).

104. Paulsson, J. Multileveled selection on plasmid replication. *Genetics* **161,** 1373–84 (2002).

105. del Solar, G., Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M. & Díaz-Orejas, R. Replication and control of circular bacterial plasmids. *Microbiol. Mol. Biol. Rev.* **62,** 434–64 (1998).

106. Velappan, N., Sblattero, D., Chasteen, L., Pavlik, P. & Bradbury, A. R. M. Plasmid incompatibility: More compatible than previously thought? *Protein Eng. Des. Sel.* **20,** 309–313 (2007).

107. Ebersbach, G. & Gerdes, K. Plasmid segregation mechanisms. *Annu. Rev. Genet.* **39,** 453–79 (2005).

108. Kobayashi, I. Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic acids research* **29,** 3742–56 (2001).

109. Suzuki, H., Yano, H., Brown, C. J. & Top, E. M. Predicting plasmid promiscuity based on genomic signature. *J. Bacteriol.* **192,** 6045–55 (2010).

110. De Gelder, L., Ponciano, J. M., Joyce, P. & Top, E. M. Stability of a promiscuous plasmid in different hosts: no guarantee for a long-term relationship. *Microbiology* **153,** 452–63 (2007).

111. Jain, A. & Srivastava, P. Broad host range plasmids. *FEMS Microbiol. Lett.* **348,** 87–96 (2013).

112. Stalder, T. *et al.* Emerging patterns of plasmid-host coevolution that stabilize antibiotic resistance. *Sci. Rep.* **7,** 4853 (2017).

113. Lacroix, B. & Citovsky, V. Transfer of DNA from bacteria to eukaryotes. *MBio* **7,** 1–9 (2016).

114. San Millan, A. *et al.* Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. *Nat. Commun.* **5,** 5208 (2014).

115. Porse, A. *et al.* Genome Dynamics of Escherichia coli during Antibiotic Treatment: Transfer, Loss, and Persistence of Genetic Elements In situ of the Infant Gut. *Front. Cell. Infect. Microbiol.* **7126,** 1–12 (2017).

116. Bergstrom, C. T., Lipsitch, M. & Levin, B. R. Natural selection, infectious transfer and the existence conditions for bacterial plasmids. *Genetics* **155,** 1505–1519 (2000).

117. Dennis, J. J. The evolution of IncP catabolic plasmids. *Curr. Opin. Biotechnol.* **16,** 291–298 (2005).

118. Keim, P. *et al.* NIH Public Access. **30,** 397–405 (2011).

119. Lindler, L. E., Plano, G. V, Burland, V., Mayhew, G. F. & Blattner, F. R. Complete DNA sequence and detailed analysis of the Yersinia pestis KIM5 plasmid encoding murine toxin and capsular antigen. *Infect. Immun.* **66,** 5731–42 (1998).

120. Rasmussen, S. *et al.* Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago. *Cell* **163,** 571–582 (2015).

121. Venkatesan, M. M. *et al.* Complete DNA Sequence and Analysis of the Large Virulence Plasmid of Shigella flexneri. *Society* **69,** 3271–3285 (2001).

122. Anda, M. *et al.* Bacterial clade with the ribosomal RNA operon on a small plasmid rather than the chromosome. *Proc. Natl. Acad. Sci.* **112,** 14343–14347 (2015).

123. Hall, J. P. J., Brockhurst, M. A. & Harrison, E. Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. *Philos. Trans. R. Soc. London B Biol. Sci.* **372,** (2017).

124. San Millan, A., Escudero, J. A., Gifford, D. R., Mazel, D. & MacLean, R. C. Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat. Ecol. Evol.* **1,** 10 (2016).

125. Svara, F. & Rankin, D. J. The evolution of plasmid-carried antibiotic resistance. *BMC Evol. Biol.* **11,** 130 (2011).

126. Hughes, D. & Andersson, D. I. Environmental and genetic modulation of the phenotypic expression of antibiotic resistance. *FEMS Microbiol. Rev.* 1–18 (2017). doi:10.1093/femsre/fux004

127. Davies, J. Vicious circles: looking back on resistance plasmids. *Genetics* **139,** 1465–1468 (1995).

128. Campbell, A. M. episomes. 101–145 doi:https://doi.org/10.1016/S0065-2660(08)60286-2

129. Datta, N. Transmissible drug resistance in an epidemic strain of Salmonella typhimurium. *J. Hyg. (Lond).* **60,** 301–10 (1962).

130. Navon-Venezia, S., Kondratyeva, K. & Carattoli, A. Klebsiella pneumoniae: A major worldwide source and shuttle for antibiotic resistance. *FEMS Microbiol. Rev.* **41,** 252–275 (2017).

131. Johnson, J. R., Johnston, B., Clabots, C., Kuskowski, M. a & Castanheira, M. Escherichia coli sequence type ST131 as the major cause of serious multidrug-resistant E. coli infections in the United States. *Clin. Infect. Dis.* **51,** 286–94 (2010).

132. Chen, L. *et al.* Carbapenemase-producing Klebsiella pneumoniae: Molecular and genetic decoding. *Trends Microbiol.* **22,** 686–696 (2014).

133. Stoesser, N. *et al.* Evolutionary History of the Global Emergence of the *Escherichia coli* Epidemic Clone ST131. *MBio* **7,** e02162-15 (2016).

134. McNally, A. *et al.* Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. *PLoS Genet.* **12,** e1006280 (2016).

135. Stewart, F. M. & Levin, B. R. The Population Biology of Bacterial Plasmids: A PRIORI Conditions for the Existence of Conjugationally Transmitted Factors. *Genetics* **87,** 209–28 (1977).

136. San Millan, A. & MacLean, R. C. Fitness Costs of Plasmids: a Limit to Plasmid Transmission. *Microbiol. Spectr.* **5,** 1–12 (2017).

137. Levin, B. R. & Stewart, F. M. Probability of establishing chimeric plasmids in natural populations of bacteria. *Science* **196,** 218–20 (1977).

138. Lundquist, P. D. & Levin, B. R. Transitory derepression and the maintenance of conjugative plasmids. *Genetics* **113,** 483–497 (1986).

139. Levin, B. R. & Lenski, R. E. *Coevolution in bacteria and their viruses and plasmids*. *Coevolution* (1983).

140. Simonsen, L. The existence conditions for bacterial plasmids: Theory and reality. *Microb. Ecol.* **22,** 187–205 (1991).

141. Freter, R., Freter, R. R. & Brickner, H. Experimental and mathematical models of Escherichia coli plasmid transfer in vitro and in vivo. *Infect. Immun.* **39,** 60–84 (1983).

142. Zünd, P. & Lebek, G. Generation time-prolonging R plasmids: correlation between increases in the generation time of Escherichia coli caused by R plasmids and their molecular size. *Plasmid* **3,** 65–9 (1980).

143. Lili, L. N., Britton, N. F. & Feil, E. J. The persistence of parasitic plasmids. *Genetics* **177,** 399–405 (2007).

144. Imran, M., Jones, D. & Smith, H. Biofilms and the plasmid maintenance question. *Math. Biosci.* **193,** 183–204 (2005).

145. Madsen, J. S., Burmølle, M. & Sørensen, S. J. A spatiotemporal view of plasmid loss in biofilms and planktonic cultures. *Biotechnol. Bioeng.* **110,** 3071–3074 (2013).

146. Bahl, M. I., Hansen, L. H., Licht, T. R. & Sørensen, S. J. Conjugative transfer facilitates stable maintenance of IncP-1 plasmid pKJK5 in Escherichia coli cells colonizing the gastrointestinal tract of the germfree rat. *Appl. Environ. Microbiol.* **73,** 341–343 (2007).

147. Nogueira, T. *et al.* Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Curr. Biol.* **19,** 1683–91 (2009).

148. Rankin, D. J., Rocha, E. P. C. & Brown, S. P. What traits are carried on mobile genetic elements, and why? *Heredity (Edinb).* **106,** 1–10 (2011).

149. Eberhard, W. G. Evolution in bacterial plasmids and levels of selection. *Q. Rev. Biol.* **65,** 3–22 (1990).

150. Hadany, L. & Comeron, J. M. Why are sex and recombination so common? *Ann. N. Y. Acad. Sci.* **1133,** 26–43 (2008).

151. Mc Ginty, S. E., Rankin, D. J. & Brown, S. P. Horizontal gene transfer and the evolution of bacterial cooperation. *Evolution* **65,** 21–32 (2011).

152. Madsen, J. S., Burmølle, M., Hansen, L. H. & Sørensen, S. J. The interconnection between biofilm formation and horizontal gene transfer. *FEMS Immunol. Med. Microbiol.* **65,** 183–195 (2012).

153. Molin, S. & Tolker-Nielsen, T. Gene transfer occurs with enhanced efficiency in biofilms and induces enhanced stabilisation of the biofilm structure. *Curr. Opin. Biotechnol.* **14,** 255–261 (2003).

154. San Millan, A., Heilbron, K. & Maclean, R. C. Positive epistasis between co-infecting plasmids promotes plasmid survival in bacterial populations. *ISME J.* 1–12 (2013). doi:10.1038/ismej.2013.182

155. Bouma, J. & Lenski, R. Evolution of a bacteria/plasmid association. *Nature* **335,** (1988).

156. Modi, R. & Adams, J. Coevolution in bacterial-plasmid populations. *Evolution (N. Y).* **45,** 656–667 (1991).

157. Dionisio, F., Conceição, I. C., Marques, a C. R., Fernandes, L. & Gordo, I. The evolution of a conjugative plasmid and its ability to increase bacterial fitness. *Biol. Lett.* **1,** 250–2 (2005).

158. Dahlberg, C. & Chao, L. Amelioration of the cost of conjugative plasmid carriage in Eschericha coli K12. *Genetics* **165,** 1641–9 (2003).

159. Harrison, E., Guymer, D., Spiers, A. J., Paterson, S. & Brockhurst, M. A. Parallel Compensatory Evolution Stabilizes Plasmids across the Parasitism-Mutualism Continuum. *Curr. Biol.* **25,** 2034–2039 (2015).

160. Sota, M. *et al.* Shifts in the host range of a promiscuous plasmid through parallel evolution of its replication initiation protein. *ISME J.* **4,** 1568–80 (2010).

161. Peña-Miller, R., Rodríguez-González, R., MacLean, R. C. & San Millan, A. Evaluating the effect of horizontal transmission on the stability of plasmids under different selection regimes. *Mob. Genet. Elements* 00–00 (2015). doi:10.1080/2159256X.2015.1045115

162. Harrison, E. & Brockhurst, M. a. Plasmid-mediated horizontal gene transfer is a coevolutionary process. *Trends Microbiol.* **20,** 262–7 (2012).

163. Hotopp, J. C. D. *et al.* Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes.

Science (80-.). **317,** 1753–1756 (2007).

164. Sorek, R. *et al.* Genome-Wide Experimental Determination of Barriers to Horizontal Gene Transfer. *Science (80-.).* **318,** 1449–1452 (2007).

165. Martínez, J. L., Coque, T. M., Lanza, V. F., de la Cruz, F. & Baquero, F. Genomic and metagenomic technologies to explore the antibiotic resistance mobilome. *Ann. N. Y. Acad. Sci.* **1388,** 26–41 (2017).

166. Martínez, J. L. Bottlenecks in the transferability of antibiotic resistance from natural ecosystems to human bacterial pathogens. *Front. Microbiol.* **2,** 1–6 (2012).

167. Forsberg, K. J. *et al.* The Shared Antibiotic Resistome of Soil Bacteria and Human Pathogens. *Science (80-.).* **337,** 1107–1111 (2012).

168. Forsberg, K. J. *et al.* Bacterial phylogeny structures soil resistomes across habitats. *Nature* **509,** 612–6 (2014).

169. Sommer, M. O. a, Dantas, G. & Church, G. M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* **325,** 1128–1131 (2009).

170. Fox, R. E., Zhong, X., Krone, S. M. & Top, E. M. Spatial structure and nutrients promote invasion of IncP-1 plasmids in bacterial populations. *ISME J.* **2,** 1024–39 (2008).

171. Roer, L., Aarestrup, F. M. & Hasman, H. The EcoKI type I restriction-modification system in Escherichia coli affects but is not an absolute barrier for conjugation. *J. Bacteriol.* **197,** 337–342 (2015).

172. McMahon, S. a *et al.* Extensive DNA mimicry by the ArdA anti-restriction protein and its role in the spread of antibiotic resistance. *Nucleic Acids Res.* **37,** 4887–97 (2009).

173. Mruk, I. & Kobayashi, I. To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. *Nucleic Acids Res.* **42,** 70–86 (2014).

174. Achtman, M., Kennedy, N. & Skurray, R. Cell--cell interactions in conjugating Escherichia coli: role of traT protein in surface exclusion. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5104–5108 (1977).

175. Singh, V., Braddick, D. & Dhar, P. K. Exploring the potential of genome editing CRISPR-Cas9 technology. *Gene* **599,** 1–18 (2017).

176. Houte, S. van, Buckling, A. & Westra, E. R. Evolutionary Ecology of Prokaryotic Immune Mechanisms. *Microbiol. Mol. Biol. Rev.* **80,** 745–763 (2016).

177. Jiang, W. *et al.* Dealing with the Evolutionary Downside of CRISPR Immunity: Bacteria and Beneficial Plasmids. *PLoS Genet.* **9,** e1003844 (2013).

178. Palmer, K. L. & Gilmore, M. S. Multidrug-Resistant Enterococci Lack CRISPR- cas. *MBio* **1,** 1–10 (2010).

179. Andersson, D. I. & Levin, B. R. The biological cost of antibiotic resistance. *Curr. Opin. Microbiol.* **2,** 489–93 (1999).

180. Hughes, D. & Andersson, D. I. Evolutionary consequences of drug resistance: shared principles across diverse targets and organisms. *Nat. Rev. Genet.* (2015). doi:10.1038/nrg3922

181. Orr, H. A. Fitness and its role in evolutionary genetics. *Nat. Rev. Genet.* **10,** 531–539 (2009).

182. Adlerberth, I. *et al.* P fimbriae and other adhesins enhance intestinal persistence of Escherichia coli in early infancy. *Epidemiol Infect* **121,** 599–608 (1998).

183. Flores-Mireles, A. L., Walker, J. N., Caparon, M. & Hultgren, S. J. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nat. Rev. Microbiol.* **13,** 269–284 (2015).

184. Diaz Ricci, J. C. & Hernández, M. E. Plasmid Effects on *Escherichia coli* Metabolism. *Crit. Rev. Biotechnol.* **20,** 79–108 (2000).

185. Gonçalves, G. A. L., Bower, D. M., Prazeres, D. M. F., Monteiro, G. A. & Prather, K. L. J. Rational engineering of Escherichia coli strains for plasmid biopharmaceutical manufacturing. *Biotechnol. J.* **7,** 251–261 (2012).

186. Birnbaum, S. & Bailey, J. E. Plasmid presence changes the relative levels of many host cell proteins and ribosome

components in recombinant Escherichia coli. *Biotechnol. Bioeng.* **37,** 736–745 (1991).

187. Wu, G. *et al.* Metabolic Burden: Cornerstones in Synthetic Biology and Metabolic Engineering Applications. *Trends Biotechnol.* **34,** 652–664 (2016).

188. Diaz Ricci, J. C. & Hernández, M. E. Plasmid effects on Escherichia coli metabolism. *Crit. Rev. Biotechnol.* **20,** 79–108 (2000).

189. Yano, H. *et al.* Evolved plasmid-host interactions reduce plasmid interference cost. *Mol. Microbiol.* **101,** 743–756 (2016).

190. Glick, B. R. Metabolic load and heterologous gene expression. *Biotechnol. Adv.* **13,** 247–261 (1995).

191. Vogwill, T. & MacLean, R. C. The genetic basis of the fitness costs of antimicrobial resistance: a meta-analysis approach. *Evol. Appl.* **8,** 284–295 (2015).

192. Tagwerker, C. *et al.* Sequence analysis of a complete 1.66 Mb Prochlorococcus marinus MED4 genome cloned in yeast. *Nucleic Acids Res.* **40,** 10375–10383 (2012).

193. Stoebel, D. M., Dean, A. M. & Dykhuizen, D. E. The cost of expression of Escherichia coli lac operon proteins is in the process, not in the products. *Genetics* **178,** 1653–60 (2008).

194. Bragg, J. G. & Wagner, A. Protein material costs: single atoms can make an evolutionary difference. *Trends Genet.* **25,** 5–8 (2009).

195. Shachrai, I., Zaslaver, A., Alon, U. & Dekel, E. Cost of Unneeded Proteins in E. coli Is Reduced after Several Generations in Exponential Growth. *Mol. Cell* **38,** 758–767 (2010).

196. Kafri, M., Metzl-Raz, E., Jona, G. & Barkai, N. The Cost of Protein Production. *Cell Rep.* **14,** 22–31 (2016).

197. Lamberte, L. E. *et al.* Horizontally acquired AT-rich genes in Escherichia coli cause toxicity by sequestering RNA polymerase. *Nat. Microbiol.* **2,** 16249 (2017).

198. Yona, A. H., Alm, E. J. & Gore, J. Random Sequences Rapidly Evolve into De Novo Promoters. *bioRxiv* (2017). doi:https://doi.org/10.1101/111880

199. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. TL - 12. *Nat. Rev. Genet.* **12 VN-r,** 32–42 (2011).

200. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science (80-.).* **324,** 255–258 (2009).

201. Sharpl, P. M. & Li, W. The codon Adaptation Index - A measure of directional synonymous codon usage bias, and its possible applications. *Nucleic Acids Res.* **15,** 1281–1295 (1987).

202. Hershberg, R. & Petrov, D. A. Selection on Codon Bias. *Annu. Rev. Genet.* **42,** 287–299 (2008).

203. Sharp, P. M., Emery, L. R. & Zeng, K. Forces that influence the evolution of codon bias. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **365,** 1203–12 (2010).

204. Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F. & Sockett, R. E. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* **33,** 1141–1153 (2005).

205. Lao, P. J. & Forsdyke, D. R. Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome Res.* **10,** 228–236 (2000).

206. Geiler-samerotte, K. A., Dion, M. F., Budnik, B. A., Wang, S. M. & Hartl, D. L. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc Natl Acad Sci U S A* **108,** 680–85 (2010).

207. Tomala, K. & Korona, R. Evaluating the fitness cost of protein expression in saccharomyces cerevisiae. *Genome Biol. Evol.* **5,** 2051–2060 (2013).

208. Singh, G. P. & Dash, D. Electrostatic Mis-Interactions Cause Overexpression Toxicity of Proteins in E. coli. *PLoS One* **8,** 1–6 (2013).

209. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A* **102,** 14338–14343 (2005).

210. Dunlop, M. J. Engineering microbes for tolerance to next-generation biofuels. *Biotechnol. Biofuels* **4,** 32 (2011).

211. Hansen, E. H. *et al.* De novo biosynthesis of Vanillin in fission yeast (schizosaccharomyces pombe) and baker's yeast (saccharomyces cerevisiae). *Appl. Environ. Microbiol.* **75,** 2765–2774 (2009).

212. Pitera, D. J., Paddon, C. J., Newman, J. D. & Keasling, J. D. Balancing a heterologous mevalonate pathway for improved isoprenoid production in Escherichia coli. *Metab. Eng.* **9,** 193–207 (2007).

213. Rozkov, A. *et al.* Characterization of the metabolic burden on Escherichia coli DH1 cells imposed by the presence of a plasmid containing a gene therapy sequence. *Biotechnol. Bioeng.* **88,** 909–915 (2004).

214. San Millan, A., Toll-Riera, M., Qi, Q. & MacLean, R. C. Interactions between horizontally acquired genes create a fitness cost in Pseudomonas aeruginosa. *Nat. Commun.* **6,** 6845 (2015).

215. Hellweger, F. L. Escherichia coli adapts to tetracycline resistance plasmid (pBR322) by mutating endogenous potassium transport: in silico hypothesis testing. *FEMS Microbiol. Ecol.* **83,** 622–31 (2013).

216. Starikova, I. *et al.* Fitness costs of various mobile genetic elements in Enterococcus faecium and Enterococcus faecalis. *J. Antimicrob. Chemother.* **68,** 2755–65 (2013).

217. Foucault, M. L., Courvalin, P. & Grillot-Courvalin, C. Fitness cost of VanA-type vancomycin resistance in methicillin-resistant Staphylococcus aureus. *Antimicrob. Agents Chemother.* **53,** 2354–2359 (2009).

218. Ingmer, H., Miller, C. & Cohen, S. N. The RepA protein of plasmid pSC101 controls Escherichia coli cell division through the SOS response. *Mol. Microbiol.* **42,** 519–526 (2001).

219. Beceiro, A., Tomás, M. & Bou, G. Antimicrobial resistance and virulence: A successful or deleterious association in the bacterial world? *Clin. Microbiol. Rev.* **26,** 185–230 (2013).

220. Martínez, J. L., Coque, T. M. & Baquero, F. What is a resistance gene? Ranking risk in resistomes. *Nat. Publ. Gr.* **13,** 116–123 (2014).

221. Levin, B. R. *et al.* The population genetics of antibiotic resistance. *Clin. Infect. Dis.* **24 Suppl 1,** S9-16 (1997).

222. Depardieu, F., Podglajen, I., Leclercq, R., Collatz, E. & Courvalin, P. Modes and modulations of antibiotic resistance gene expression. *Clin. Microbiol. Rev.* **20,** 79–114 (2007).

223. Foucault, M.-L., Depardieu, F., Courvalin, P. & Grillot-Courvalin, C. Inducible expression eliminates the fitness cost of vancomycin resistance in enterococci. *Proc. Natl. Acad. Sci.* **107,** 16964–16969 (2010).

224. Michon, A. *et al.* Plasmidic qnrA3 enhances Escherichia coli fitness in absence of antibiotic exposure. *PLoS One* **6,** (2011).

225. LaMarre, J. M., Locke, J. B., Shaw, K. J. & Mankin, A. S. Low fitness cost of the multidrug resistance gene cfr. *Antimicrob. Agents Chemother.* **55,** 3714–3719 (2011).

226. Marciano, D. C., Karkouti, O. Y. & Palzkill, T. A fitness cost associated with the antibiotic resistance enzyme SME-1 β-lactamase. *Genetics* **176,** 2381–2392 (2007).

227. Fernández, A. *et al.* Expression of OXA-type and SFO-1 β-lactamases induces changes in peptidoglycan composition and affects bacterial fitness. *Antimicrob. Agents Chemother.* **56,** 1877–84 (2012).

228. Morosini, M. I., Ayala, J. A., Baquero, F., Martinez, J. L. & Blazquez, J. Biological cost of AmpC production for Salmonella enterica serotype typhimurium. *Antimicrob. Agents Chemother.* **44,** 3137–3143 (2000).

229. Moya, B., Juan, C., Albertí, S., Pérez, J. L. & Oliver, A. Benefit of having multiple ampD genes for acquiring β-lactam resistance without losing fitness and virulence in Pseudomonas aeruginosa. *Antimicrob. Agents Chemother.* **52,** 3694–3700 (2008).

230. Gutierrez, B. *et al.* Fitness cost and interference of Arm/Rmt aminoglycoside resistance with the RsmF housekeeping methyltransferases. *Antimicrob. Agents Chemother.* **56,** 2335–2341 (2012).

231. Machuca, J. *et al.* Interplay between plasmid-mediated and chromosomal-mediated fluoroquinolone resistance and bacterial fitness in Escherichia coli. *J. Antimicrob. Chemother.* **69,** 3203–3215 (2014).

232. De Wit, R. & Bouvier, T. 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environ. Microbiol.* **8,** 755–758 (2006).

233. Reding-Roman, C. *et al.* The unconstrained evolution of fast and efficient antibiotic-resistant bacterial genomes. *Nat. Ecol. Evol.* **1,** 50 (2017).

234. Reznick, D. & King, K. Antibiotic resistance: Evolution without trade-offs. *Nat. Ecol. Evol.* **1,** 66 (2017).

235. Paulander, W., Maisnier-Patin, S. & Andersson, D. I. The fitness cost of streptomycin resistance depends on rpsL mutation, carbon source and RpoS (sigmaS). *Genetics* **183,** 539–46, 1SI–2SI (2009).

236. Björkman, J. Effects of Environment on Compensatory Mutations to Ameliorate Costs of Antibiotic Resistance. *Science (80-.).* **287,** 1479–1482 (2000).

237. Concepción-Acevedo, J., Weiss, H. N., Chaudhry, W. N. & Levin, B. R. Malthusian Parameters as Estimators of the Fitness of Microbes: A Cautionary Tale about the Low Side of High Throughput. *PLoS One* **10,** e0126915 (2015).

238. Gumpert, H. *et al.* Transfer and Persistence of a Multi-Drug Resistance Plasmid in situ of the Infant Gut Microbiota in the Absence of Antibiotic Treatment. *Front. Microbiol.* **8,** (2017).

239. Debelius, J. *et al.* Tiny microbes, enormous impacts: what matters in gut microbiome studies? *Genome Biol.* **17,** 217 (2016).

240. Chen, S. L. *et al.* Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 5977–5982 (2006).

241. C. Sasakawa. *Molecular mechanisms of bacterial infection via the gut. Current topics in microbiology and immunology* **337,** (2009).

242. Adlerberth, I. & Wold, A. E. Establishment of the gut microbiota in Western infants. *Acta Paediatr. Int. J. Paediatr.* **98,** 229–238 (2009).

243. Nowrouzian, F. *et al.* Escherichia coli in infants' intestinal microflora: Colonization rate, strain turnover, and virulence gene carriage. *Pediatr. Res.* **54,** 8–14 (2003).

244. Adlerberth, I. *et al.* High turnover rate of Escherichia coli strains in the intestinal flora of infants in Pakistan. *Epidemiol. Infect.* **121,** 587–598 (1998).

245. Arrieta, M.-C., Stiemsma, L. T., Amenyogbe, N., Brown, E. M. & Finlay, B. The Intestinal Microbiome in Early Life: Health and Disease. *Front. Immunol.* **5,** 1–18 (2014).

246. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14,** 20–32 (2015).

247. Spaulding, C. N. *et al.* Selective depletion of uropathogenic E. coli from the gut by a FimH antagonist. *Nature* **546,** 528–532 (2017).

248. Karami, N., Nowrouzian, F., Adlerberth, I. & Wold, A. E. Tetracycline Resistance in Escherichia coli and Persistence in the Infantile Colonic Microbiota Tetracycline Resistance in Escherichia coli and Persistence in the Infantile Colonic Microbiota. *Society* **50,** 156–161 (2006).

249. Ostaff, M. J., Stange, E. F. & Wehkamp, J. Antimicrobial peptides and gut microbiota in homeostasis and pathology. *EMBO Mol. Med.* **5,** 1465–1483 (2013).

250. Inglis, R. F., Bayramoglu, B., Gillor, O. & Ackermann, M. The role of bacteriocins as selfish genetic elements. *Biol. Lett.* **9,** 20121173–20121173 (2013).

251. McNally, A., Thomson, N. R., Reuter, S. & Wren, B. W. 'Add, stir and reduce': Yersinia spp. as model bacteria for pathogen evolution. *Nat. Rev. Microbiol.* **14,** 177–190 (2016).

252. Popa, O. & Dagan, T. Trends and barriers to lateral gene transfer in prokaryotes. *Curr. Opin. Microbiol.* **199,** 615–623 (2011).

253. Jacek Majewski. Sexual Isolation in Bacteria. *FEMS Microbiol Lett* **6,** 161–169 (2001).

254. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480,** 241–244 (2011).

255. Popa, O., Hazkani-Covo, E., Landan, G., Martin, W. & Dagan, T. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res.* **21,** 599–609 (2011).

256. Andam, C. P. & Gogarten, J. P. Biased gene transfer in microbial evolution. *Nat. Rev. Microbiol.* **9,** 543–555 (2011).

257. Klümper, U. *et al.* Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *ISME J.* **9,** 934–945 (2014).

258. Ikeda, H. & Tomizawa, J. -i. Prophage P1, an extrachromosomal replication unit. *Cold Spring Harb. Symp. Quant. Biol.* **33,** 791–798 (1968).

259. Park, C. & Zhang, J. High expression hampers horizontal gene transfer. *Genome Biol. Evol.* **4,** 523–532 (2012).

260. Medrano-Soto, A., Moreno-Hagelsieb, G., Vinuesa, P., Christen, J. A. & Collado-Vides, J. Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. *Mol. Biol. Evol.* **21,** 1884–1894 (2004).

261. Singh, K., Milstein, J. N. & Navarre, W. W. Xenogeneic Silencing and Its Impact on Bacterial Genomes. *Annu. Rev. Microbiol.* **70,** 199–213 (2016).

262. Tuller, T. Codon bias, tRNA pools, and horizontal gene transfer. *Mob. Genet. Elements* **1,** 75–77 (2011).

263. Dorman, C. J. H-NS-like nucleoid-associated proteins, mobile genetic elements and horizontal gene transfer in bacteria. *Plasmid* **75,** 1–11 (2014).

264. Dorman, C. J. H-NS: a universal regulator for a dynamic genome. *Nat. Rev. Microbiol.* **2,** 391–400 (2004).

265. Doyle, M. *et al.* An H-NS-like stealth protein aids horizontal DNA transmission in bacteria. *Science* **315,** 251–2 (2007).

266. Will, W. & Frost, L. Characterization of the opposing roles of H-NS and TraJ in transcriptional regulation of the F-plasmid tra operon. *J. Bacteriol.* **188,** (2006).

267. Daubin, V. *et al.* The source of laterally transferred genes in bacterial genomes. *Genome Biol.* **4,** R57 (2003).

268. Raghavan, R., Kelkar, Y. D. & Ochman, H. A selective force favoring increased G+C content in bacterial genes. *Proc. Natl. Acad. Sci.* **109,** 14504–14507 (2012).

269. Kelkar, Y. D., Phillips, D. S. & Ochman, H. Effects of Genic Base Composition on Growth Rate in G+C-rich Genomes. *G3 GENES, GENOMES, Genet.* **5,** 1247–1252 (2015).

270. Tuller, T. *et al.* Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids Res.* **39,** 4743–4755 (2011).

271. Goodman, D. B., Church, G. M. & Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science (80-.).* **342,** 475–479 (2013).

272. Boël, G. *et al.* Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature* **529,** 358–363 (2016).

273. Amorós-Moya, D., Bedhomme, S., Hermann, M. & Bravo, I. G. Evolution in regulatory regions rapidly compensates the cost of nonoptimal codon usage. *Mol. Biol. Evol.* **27,** 2141–2151 (2010).

274. Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet.* **36,** 760–766 (2004).

275. Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 6239–44 (1998).

276. Abby, S. S., Tannier, E., Gouy, M. & Daubin, V. Lateral gene transfer as a support for the tree of life. *Proc. Natl.*

Acad. Sci. **109,** 4962–4967 (2012).

277. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 3801–3806 (1999).

278. Maor-Shoshani, a, Reuven, N. B., Tomer, G. & Livneh, Z. Highly mutagenic replication by DNA polymerase V (UmuC) provides a mechanistic basis for SOS untargeted mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 565–570 (2000).

279. Runyen-Janecky, L. J., Hong, M. & Payne, S. M. The virulence plasmid-encoded impCAB operon enhances survival and induced mutagenesis in Shigella flexneri after exposure to UV radiation. *Infect. Immun.* **67,** 1415–1423 (1999).

280. Cohen, O., Gophna, U. & Pupko, T. The complexity hypothesis revisited: Connectivity Rather Than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* **28,** 1481–1489 (2011).

281. Kacar, B., Garmendia, E., Tuncbag, N., Andersson, D. I. & Hughes, D. Functional constraints on replacing an essential gene with its ancient and modern homologs. *MBio* **8,** 1–30 (2017).

282. Aris-Brosou, S. Determinants of adaptive evolution at the molecular level: The extended complexity hypothesis. *Mol. Biol. Evol.* **22,** 200–209 (2005).

283. Leigh, J. W., Schliep, K., Lopez, P. & Bapteste, E. Let them fall where they may: Congruence analysis in massive phylogenetically messy data sets. *Mol. Biol. Evol.* **28,** 2773–2785 (2011).

284. Lercher, M. J. & Pál, C. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol. Biol. Evol.* **25,** 559–567 (2008).

285. Omer, S., Kovacs, A., Mazor, Y. & Gophna, U. Integration of a foreign gene into a native complex does not impair fitness in an experimental model of lateral gene transfer. *Mol. Biol. Evol.* **27,** 2441–2445 (2010).

286. Wellner, A. & Gophna, U. Neutrality of foreign complex subunits in an experimental model of lateral gene transfer. *Mol. Biol. Evol.* **25,** 1835–1840 (2008).

287. Lind, P. A., Tobin, C., Berg, O. G., Kurland, C. G. & Andersson, D. I. Compensatory gene amplification restores fitness after inter-species gene replacements. *Mol. Microbiol.* **75,** 1078–1089 (2010).

288. Knöppel, A., Lind, P. a, Lustig, U., Näsvall, J. & Andersson, D. I. Minor fitness costs in an experimental model of horizontal gene transfer in bacteria. *Mol. Biol. Evol.* **31,** 1220–7 (2014).

289. Hughes, V. M. & Datta, N. Conjugative plasmids in bacteria of the 'pre-antibiotic' era. *Nature* **302,** 725–726 (1983).

290. Levin, B. R. The accessory genetic elements of bacteria: existence conditions and (co)evolution. *Curr. Opin. Genet. Dev.* **3,** 849–54 (1993).

291. Omenn, G. S. Evolution and public health. *Proc. Natl. Acad. Sci.* **107,** 1702–1709 (2010).

292. Bull, J. J. & Barrick, J. E. Arresting Evolution. *Trends Genet.* xx**,** 1–11 (2017).

293. Dragosits, M. & Mattanovich, D. Adaptive laboratory evolution – principles and applications for biotechnology. *Microb. Cell Fact.* **12,** 64 (2013).

294. Sommer, M. O. A., Munck, C., Toft-Kehler, R. V. & Andersson, D. I. Prediction of antibiotic resistance: time for a new preclinical paradigm? *Nat. Rev. Microbiol.* **15,** 689–696 (2017).

Acknowledgements:
I thank Leonie Jahn and Christian Munck for proof reading of the preceding sections.

Present investigations

The research conducted for this thesis couples molecular insight to the behaviour and evolution of genetic elements, at different environmental and genetic levels, and is presented in three themes:

Evolution of *E. coli* in the infant gut

Manuscript I and **II** investigate the evolution of co-existing *E. coli* lineages residing in the infant gut. The gut microbiota has received much attention for its role in health and disease, and the included studies underline its role as a hot-spot for HGT. While the gut is a complex selective environment, we sought to explain several genomic events by controlled *in vitro* and *in vivo* interrogations. Although some genomic events could be explained by beneficial effects on fitness, others did not show a selective benefit in our model-setups. These studies exemplify the high genetic turnover of gut-residing *E. coli* isolates and underline the ability of virulence and antibiotic resistance genes to persist for extended periods of time in spite of no direct selection.

Host-plasmid adaptation in the clinic and industry

In the two studies of **manuscript III** and **IV**, focusing on the adaptive events leading to increased stability of initially costly plasmids, we demonstrate how transposable elements constitute an efficient adaptive resource able to target expensive plasmid regions. In a large multidrug resistance plasmid, originating from a clinical *Klebsiella pneumoniae* strain, this cost amelioration happened through intramolecular recombination events that compromise the ability of plasmid transfer. Similarly, we observed that transposon mediated loss of function dynamics severely compromises production capacity in the industrially relevant case of mevalonate biosynthesis. For both studies, we emphasize the importance of the host background and show that these events, although universally beneficial, occur at much lower frequencies in certain hosts.

Molecular factors governing the success of transferable resistance genes

In **manuscript V** we explore the functional integration and biological costs of HGT. To do this, we employed a synthetic library of 200 antibiotic resistance genes representatively sampled from public databases. Though experimental interrogations, we find that the sequence composition of a resistance gene is of minor importance compared to the mechanistic constrains imposed by the encoded protein product. We demonstrate that drug-modifying proteins are less restricted by previous host affiliations in their functionality and costs. In contrast, proteins employing mechanisms with a high degree of metabolic or physiological interactions were less likely to work, and imposed a cost that scaled with the phylogenetic distance of their customary hosts when tested in *E. coli*.

Manuscript I [redacted]

frontiers in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 12 April 2017
doi: 10.3389/fcimb.2017.00126

Genome Dynamics of *Escherichia coli* during Antibiotic Treatment: Transfer, Loss, and Persistence of Genetic Elements *In situ* of the Infant Gut

Andreas Porse[1†], Heidi Gumpert[2†], Jessica Z. Kubicek-Sutherland[3], Nahid Karami[4], Ingegerd Adlerberth[4], Agnes E. Wold[4], Dan I. Andersson[3] and Morten O. A. Sommer[1]*

[1] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark, [2] Department of Clinical Microbiology, Hvidovre University Hospital, Hvidovre, Denmark, [3] Department of Medical Biochemistry and Microbiology, Uppsala University Biomedical Centre, Uppsala, Sweden, [4] Department of infectious Diseases, University of Gothenburg, Sahlgrenska Academy, Gothenburg, Sweden

OPEN ACCESS

Edited by:
Alfredo G. Torres, University of Texas Medical Branch, USA

Reviewed by:
Swaine Chen, Genome Institute of Singapore, Singapore
Timothy James Wells, The University of Queensland, Australia

***Correspondence:**
Morten O. A. Sommer
msom@bio.dtu.dk

[†] These authors have contributed equally to this work.

Received: 14 February 2017
Accepted: 28 March 2017
Published: 12 April 2017

Citation:
Porse A, Gumpert H, Kubicek-Sutherland JZ, Karami N, Adlerberth I, Wold AE, Andersson DI and Sommer MOA (2017) Genome Dynamics of Escherichia coli during Antibiotic Treatment: Transfer, Loss, and Persistence of Genetic Elements In situ of the Infant Gut. Front. Cell. Infect. Microbiol. 7:126. doi: 10.3389/fcimb.2017.00126

Elucidating the adaptive strategies and plasticity of bacterial genomes *in situ* is crucial for understanding the epidemiology and evolution of pathogens threatening human health. While much is known about the evolution of *Escherichia coli* in controlled laboratory environments, less effort has been made to elucidate the genome dynamics of *E. coli* in its native settings. Here, we follow the genome dynamics of co-existing *E. coli* lineages *in situ* of the infant gut during the first year of life. One *E. coli* lineage causes a urinary tract infection (UTI) and experiences several alterations of its genomic content during subsequent antibiotic treatment. Interestingly, all isolates of this uropathogenic *E. coli* strain carried a highly stable plasmid implicated in virulence of diverse pathogenic strains from all over the world. While virulence elements are certainly beneficial during infection scenarios, their role in gut colonization and pathogen persistence is poorly understood. We performed *in vivo* competitive fitness experiments to assess the role of this highly disseminated virulence plasmid in gut colonization, but found no evidence for a direct benefit of plasmid carriage. Through plasmid stability assays, we demonstrate that this plasmid is maintained in a parasitic manner, by strong first-line inheritance mechanisms, acting on the single-cell level, rather than providing a direct survival advantage in the gut. Investigating the ecology of endemic accessory genetic elements, in their pathogenic hosts and native environment, is of vital importance if we want to understand the evolution and persistence of highly virulent and drug resistant bacterial isolates.

Keywords: *Escherichia coli*, genome evolution, virulence plasmid dynamics, plasmid persistence, horizontal gene transfer, antibiotic treatment, urinary tract infections, infant gut

INTRODUCTION

The human gut is home to a dense microbial ecosystem, the human gut microbiota, playing an important role in human health and physiology (Marchesi et al., 2015). As a commensal constituent of the gut microbiota in warm-blooded animals, *Escherichia coli* is highly adapted to the gut and colonizes the gastrointestinal tract within the first hours of life (Drasar and Hill, 1974). However, some environmental and commensal *E. coli* isolates have acquired genetic factors that allow them to cause disease within the digestive tract or when transferred to other body sites such as the blood, brain, and urinary tract (Smith et al., 2007). While diarrheagenic *E. coli* are a common cause of gastro intestinal infections in third world countries and travelers, extraintestinal pathogenic *E. coli* (ExPEC) are facultative pathogens that reside in the human gut microbiota but occasionally establish in extra-intestinal body sites (Köhler and Dobrindt, 2011). Here, urinary tract infections casued by ExPEC are among the most common bacterial infections in developed countries, were patients are often infected via transmission of strains from their own intestinal flora to their urinary tract (Foxman, 2010).

The broad adaptation of *E. coli* to the gut environment and extraintestinal body sites is reflected in the remarkable genetic diversity within the species. This genetic flexibility is largely facilitated by horizontal gene transfer (HGT) of accessory genetic elements including plasmids and phages (Brzuszkiewicz et al., 2009). These elements are widely present within the gut microbiota and can provide their bacterial hosts with antibiotic resistance or virulence factors (Salyers et al., 2004; Sommer et al., 2010). Acquiring virulence genes might not only influence the risk and severity of infections caused by the pathogen, but has also been suggested to assist in the general persistence of commensal bacterial strains of the gut (Diard et al., 2010; Chen et al., 2013). Indeed, virulence determinants such as those involved in adhesion, biofilm formation and iron acquisition correlate with prolonged colonization in the digestive tract (Adlerberth et al., 1998; Nowrouzian et al., 2003).

Recent studies into the dynamics of clinical bacterial genomes at genomic resolution have been carried out with time-series sampling and underlines the high plasticity of plasmids and their host associations *in situ* (Conlan et al., 2014, 2016). Conjugative plasmids are of particular interest, as they are the main vehicles of HGT in *E. coli,* playing an essential role in the adaptation toward antibiotics or specific host niches (Johnson and Nolan, 2009; Norman et al., 2009).

Whereas, much effort has been devoted to study the survival conditions of plasmids *in vitro* (Slater et al., 2008) our knowledge on the behavior of plasmids *in situ* of their native hosts and natural environment is limited (Karami et al., 2007; Conlan et al., 2014, 2016). In order for a plasmid to persist in the long term, it needs to either be stably segregated upon cell division, confer a fitness advantage to its host, or transfer at high enough rates to compensate the lack of the latter two (Simonsen, 1991; Slater et al., 2008). As most plasmids do not exhibit sufficient rates of transfer to survive without selection, stable inheritance, and adaptive traits are key to their long term survival (Simonsen, 1991).

To elucidate the genome dynamics of *E. coli* in its native environment of the gut, we genome sequenced individual *E. coli* isolates over the first year of an infant's life. We conduct *in vitro* and *in vivo* competition assays to elucidate the selective drivers of the observed dynamics, and gain a deeper understanding of the endemic mobile elements contributing to the dissemination of virulence and antibiotic resistance factors.

MATERIALS AND METHODS

Genome Sequencing of *E. coli* Lineages

The strains were isolated and typed as part of a previous study by Karami et al. (2007). These were cultured in LB broth and genomic DNA was isolated using an UltraClean© Microbial DNA Isolation Kit (MoBio Laboratories, Inc., California). Sequencing libraries were prepared using the TruSeq and Nextera XT (Illumina, California) protocols. Illumina HiSeq sequencing was performed by Partners HealthCare Center for Personalized Genetic Medicine (Cambridge, Massachusetts).

Sequence Analysis

Genomes for each sequenced isolate were assembled using Velvet (v1.2.10; Zerbino and Birney, 2008) and annotated via RAST (Aziz et al., 2008). Reads from the isolates were mapped onto the reference, e.g., earliest isolated, genome via Bowtie2 (2.1.0; Langmead et al., 2009), and single nucleotide polymorphisms (SNPs) were enumerated via SAMTools (0.1.19; Li et al., 2009). The SNP threshold was set to include SNPs with a phred score of above 30 and at least 90% of the high-quality reads at the site as the variant. Additionally, to ensure that all isolates within a lineage consisted of the same genomic content as the representative isolates, genomic areas lacking mapped read coverage were identified using BEDTools (2.18.2; Quinlan and Hall, 2010).

The pNK29 plasmid was assembled into a circular plasmid with aid from plasmid alignments produced using MUMer (Kurtz et al., 2004). Contigs belonging to the pNK29 antibiotic resistance plasmid were first identified in lineage B as the new genetic material of the isolate at 32 days, and then used to identify the corresponding contigs in the lineage A genome. The RAST annotations for this plasmid were refined based on homologous genes in pOLA52 (NC_010378.1) that were either missing or incorrect in pNK29.

Plasmid Identification and Comparison

Other plasmids were identified by first separating contigs based coverage to infer copy-number relative to genomic contigs and then by grouping contigs together with similar abundances. The average coverage of each contig was determined using BEDTools (Quinlan and Hall, 2010). Plasmid incompatibility grouping was done using the PlasmidFinder tool (Carattoli et al., 2014). Homologous previously sequenced plasmids were identified using BLAST and the NCBI nt database (Altschul et al., 1990). Circular plasmid diagrams were created using the BLAST ring image generator (BRIG; Alikhan et al., 2011).

For pNK29-2, blastn searches of the plasmid contigs revealed 14 plasmids with very high identity (99%) and hits with a pNK29-2 coverage of >97% where selected (**Figure 3** and **Table S4**). As an exception, pECO-bc6 was also included despite its lower coverage (88%) to illustrate deletion of plasmid accessory genes flanked by inverted repeats. The EasyFigure software was used for linear comparison of plasmid sequences displayed in **Figure 4** (Sullivan et al., 2011).

The core genome of the *E. coli* hosts listed in **Figure 3** was estimated using ROARY via annotations from PROKKA (Seemann, 2014; Page et al., 2015). The aligned, ungapped core genome was used to construct a maximum likelihood phylogenetic tree using the RAxML software (Stamatakis, 2014). MLST types were assigned using MLSTfinder (Larsen et al., 2012), and *fimH* types were assigned using the sequences referred to by Dias et al. (2010).

Strain Tagging and pNK29-2 Plasmid Curing

Lineage A and B strains isolated at the first time point were tagged with antibiotic resistance markers to allow quantification during competitive fitness experiments, plasmid loss experiments, assessment of conjugation ability, and plasmid curing. Resistance cassettes conferring resistance to Chloramphenicol and Kanamycin respectively were amplified from cloning vectors of the pZ system (Lutz and Bujard, 1997) and inserted into the chromosomal *araB* gene of the Lineage A and B strains using the Lambda Red recombineering system of pTKRED (Kuhlman and Cox, 2010). The following regions of homology were used for insertions into *araB*: 5′-GTAGCGAGGTTAAGATCGGTAATC ACCCCTTTCAGGCGTTGGTTAGCGTT-3′ and 5′-GCCTAA CGCACTGGTAAAAGTTATCGGTACTTCCACCTGCGACA TTCTGA-3′.

The pNK29-2 plasmid was tagged with the Sh *ble* Zeocin resistance gene in a transposase gene located at 61 kb using the following homology ends: 5′-CTTCGGGAACGCTGTAACGAT TACCACCAACCTCGATATAGCTGTCCCGG-3′ and 5′-TAA CAACGGGAAAGTCGTGTTCAACTCCGGATTCCTGTTGC TGGCCGACC-3′.

To cure pNK29-2 we disrupted the *stbA* gene of the *stbAB* stability operon with a Kanamycin resistance marker using the following homology ends: 5′-CATAAATGTGATGTGT GAAGTATGATGATATTTTGACACGGTAACCTGAGTA GGGATAACAGGGTAAT-3′ and 5′-TTCATTTTAAGAC GCACATCATTCATTGCCTCCTGCACCGAATCAGTAG CTAGGGATAACAGGGTAAT-3′. These contain the 18 bp I-SceI endonuclease site enabling *in situ* double digestion of the recombinant plasmid to enhance plasmid loss via induction of the I-SceI endonuclease from pTKRED. Following recombineering, recombinants were selected on Kanamycin containing plates incubated at 30°C to retain the pTKRED plasmid. These were verified with PCR to confirm insertion into the *stbA* gene. Verified colonies were inoculated directly into LB medium containing 0.5% w/v L-arabinose and grown at 30°C for 2 h to induce expression of the pTKRED encoded I-SceI endonuclease and subsequently switched to 42°C for 3 h to allow curing of the temperature sensitive pTKRED vector. The culture was diluted and plated on LB to obtain single colonies that were confirmed for plasmid curing by the absence of growth on Kanamycin (pNK29-2) and Spectinomycin (pTKRED). The curing of pNK29-2 was validated by PCR using primers targeting two distinct loci of the pNK29-2 backbone: Upstream *stb*: 5′-CTCAACAAGGGTTATTGC-3′, downstream *stb*: 5′-GAATGGCAAATGAAACG-3′ and Upstream 61 kb transposase: 5′-GAATGGCAAATGAAACG-3′, downstream 61 kb transposase: 5′-AGAAGGCTGCGGTGCTGAAG-3′. The plasmid-cured variant of the lineage A strain was re-transformed with the Zeocin tagged pNK29-2 plasmid to control for potential effects of the curing process.

Conjugative Transfer Assessment

In order to test the ability of pNK29-2 to conjugate, outgrown over-night (O/N) cultures as well as exponentially growing cultures of the lineage A strain and *E. coli* MG1655::tetA was mixed equally and incubated O/N. Incubations were done at 37°C and 30°C on a solid agar surface as well as in liquid cultures without shaking. Additional tagging at the 32 and 26 kb positions were carried out to ensure that the initial insertion at the 61 kb position of pNK29-2 was not the cause of the dysfunctional conjugation ability.

In vitro Competition, Growth Rate, and Plasmid Loss Assays

Two O/N cultures were diluted to the same OD and mixed 1:1 in LB medium. The competition was carried out in 1.5 ml cultures and a volume of 1.5 μl was transferred to a fresh well every 20 h. From OD measurements the number of generations was estimated to ∼10 generations/transfer. The ratio of the competitors was determined as the fraction of colonies on Chloramphenicol agar plates compared to plates containing Kanamycin. Plasmid loss was assessed by comparing plate-counts of at least 100 colonies on LB and Zeocin agar plates. In addition, colonies from LB plates were streaked on Zeocin containing plates and PCR (using the primers listed in the plasmid curing section) was performed on a subset of colonies to verify plasmid presence. Competitions in iron-limited medium was carried out as for the LB competitions described above, except that M9 medium [M9 minimal medium (standard), 2010] with 5 μM $FeCl_3$ was used instead or LB. OD measurements of growth rates in M9 5 μM $FeCl_3$ were conducted in 96-well plates containing 150 μl medium/well using a *ELx808* plate reader (BioTek, USA). Breathe-Easy (Sigma-Aldrich) film was applied to minimize evaporation during measurements. OD at 600 nms was measured with 5-min intervals for 24 h and incubated with shaking at 37°C between measurements.

In vivo Competition Experiments

Female BALB/c mice (5–6 weeks old) were used in all *in vivo* studies (Charles River Laboratories, distributed by Scanbur). All mice were pre-treated orally with streptomycin as described previously (Lasaro et al., 2014). Briefly, streptomycin sulfate salt (Sigma-Aldrich) was added to the drinking water at 5 g/L, along with 5 g/L of glucose, to enhance taste, for 72 h followed by 24

h of fresh water (no drug or glucose) to allow the streptomycin to clear the animal's system prior to infection. A single colony of each *E. coli* strain was grown in LB shaking overnight at 37°C. Cells were pelleted, washed once in PBS and re-suspended in PBS (13 mM phosphate with 137 mM NaCl at pH 7.4), and then mixed in a 1:1 ratio of *E. coli* lineage A isolate with pNK29-2 to the cured variant. Fifteen mice that were pre-treated with streptomycin followed by 24 h without drug were administered 100 μL containing 2×10^8 CFU of this 1:1 *E. coli* mixture by oral gavage. Feces was collected at days 2, 4, 7, and 12 post-infection. Additionally, on day 12 following termination of the experiment, a segment of the small intestine was removed. The feces and small intestine segment were homogenized in PBS, serially diluted, and equal amounts were plated on LA-Cam (25 μg/ml chloramphenicol, selecting for the pNK29-2 containing strain) and LA-Kan (50 μg/ml kanamycin, selecting for the cured strain). Following overnight incubation at 37°C, CFUs were enumerated and subsequently replica plated from both LA-Cam and LA-Kan to LA-Zeo (40 ug/ml zeocin) to screen for the presence and transfer of the pNK29-2 plasmid containing a Zeocin-resistance marker. CFU values were normalized per gram of tissue (CFU/g). The competitive index was calculated by dividing the OUTPUT on days 2, 4, 7, and 12 (CamR CFU/g divided by KanR CFU/g) by the INPUT on day 0 (CamR CFU/g divided by KanR CFU/g). The input value was 1.04 indicating a 1:1.04 initial ratio of *E. coli* lineage A isolate with pNK29-2 relative to the cured *E. coli* lineage A isolate. The non-parametric Mann-Whitney *U*-Test was used to compare the sample populations. The *P*-values indicate the probability of falsely rejecting the null-hypothesis of equal population means.

Ethics Statement

Animal experiments were performed in accordance with national (regulation SJVFS 2012:26) and institutional guidelines. The Uppsala Animal Experiments Ethics Review Board in Uppsala, Sweden approved all mouse protocols undertaken in this study under reference no. 154/14. Animal experiments were performed at the Swedish National Veterinary Institute (SVA) in Uppsala, Sweden.

RESULTS

The current study material was obtained from an infant enrolled in the ALLERGYFLORA study with the original purpose of correlating the composition of the gut microbiota to the development of allergies later in life (Adlerberth et al., 2006).

This infant was subjected to long-term antibiotic treatment as a consequence of a urinary tract infection, and was selected for this study due to an observed change in the resistance profile of *E. coli* strains isolated from this infant (Karami et al., 2007).

Fecal samples were obtained from the infant at 2, 9, 16, and 32 days, and 2, 6, and 12 months after birth (**Figure 1**). *E. coli* lineages were identified by morphological and biochemical characteristics as well as subsequent confirmation by PFGE and random amplified polymorphic DNA (RAPD) typing. The two main lineages were designated "A" and "B." *E. coli* lineage A was recovered at all the sampling time points and lineage B was only transiently present in the samples collected from day 9 to 32 days of age (Karami et al., 2007; **Figure 1**). At 11 days of age, a UTI infection was diagnosed and trimethoprim administered intravenously (i.v.) for 5 days. However, due to the subsequent presence of enterococci in addition to *E. coli*, the antibiotic treatment was changed to i.v. ampicillin for 5 days followed by oral amoxicillin treatment for an additional 8 days. Lastly, the infant was administered trimethoprim prophylactically for the following 7 months (**Figure 1**).

Transfer of an Antibiotic Resistance Plasmid between Two Distinct *E. coli* Lineages Co-Colonizing the Infant Gut

We sequenced the genomes of the *E. coli* isolates, obtained from the fecal and urine samples of the infant, which confirmed the lineages previously identified via RAPD and PFGE typing. Comparing the genomic similarity of the two lineages revealed that the lineages did indeed originate from two different strains; with the initial isolate of lineage A (4.91 Mb) sharing only 77% of lineage B's (5.45 Mb) genomic content.

To assess the genomic divergence of the individual lineage A isolates, during the sampling period, a SNP-based phylogenetic tree was constructed (**Figure 2**, **Table S1**). Only one SNP was found when comparing the genomes of lineage A isolates collected at 2, 9, and 16 days to the UTI isolate (**Figure 2**). Given the very high sequence similarities between these isolates, lineage A colonizing the gut microbiota was assumed to be the cause of the UTI.

From the annotated genomes of lineage A and lineage B isolates, we identified several factors that could contribute to the pathogenicity of these strains. The genome of the uropathogenic lineage A encoded the type 1 fimbriae FimH among other adhesion factors (*AidA-I* and *yqi* encoded adhesions), siderophore (enterobactin and yersiniabactin) transporters and hemin receptors (TonB-system) as well as enterotoxins (*senB* and *vat*) and Hemolysin E. Although the lineage B isolates did not cause infection, its genomic content reveals similar virulence factors such as the type 1 fimbriae (*fimH*), serum survival factors (*iss*) and iron acquisition (aerobactin synthesis and transport), but no enterotoxins.

While no SNPs were detected in the three isolates from lineage B, we report the sequence of pNK29, a 42.2 kb TEM-1b encoding conjugative IncX plasmid, that transferred from lineage A to lineage B *in situ* of the gut (Karami et al., 2007; **Figure S1**, **Table S2**). This plasmid was first detected in lineage B at 32 days, and coincided with high resistance toward ampicillin compared to earlier isolates (**Figure 1**). Similar conjugative plasmids of the IncX family are prevalent in pathogenic *E. coli,* as well as other Enterobacteriaceae isolated from humans and animals, playing an important role in the dissemination of antibiotic resistance genes (Norman et al., 2008; Toro et al., 2014).

Karami et al. reported an increase in lineage A counts from $10^{6.4}$ CFU/g fecal matter to a density of 10^{11} CFU/g as the infant was switched from trimethoprim to ampicillin and amoxicillin treatment during the UTI infection from day 16 to 32 (Karami et al., 2007). Such events can increase population size, and thus the probability of plasmid transfer and enrichment of pNK29

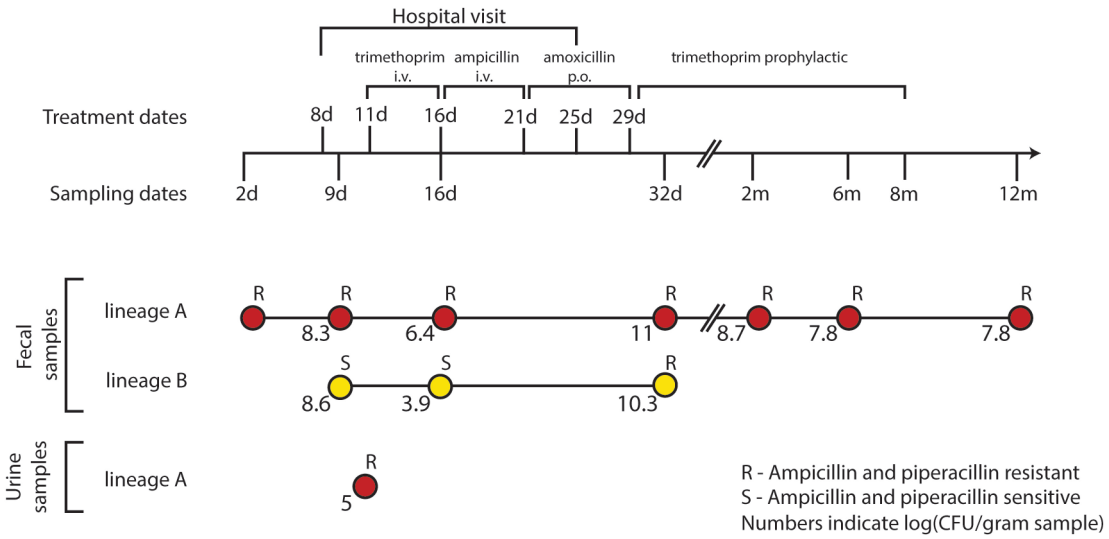


FIGURE 1 | Resistance profiles of co-existing *E. coli* lineages during the course of a urinary tract infection and antibiotic treatment. *E. coli* isolates were obtained during the first 12 months of an infant's life. At 8 days of age, the infant was admitted to the hospital due to a urinary tract infection from which the lineage A clone was isolated at day 11. After 5 days of i.v. trimethoprim the treatment was switched to 5 days of i.v. ampicillin followed by an additional 8 days of oral (p.o.) amoxicillin. Trimethoprim was administered to prevent reoccurring infections for the following 7 months. Along the course of treatment, lineage B acquired a TEM-1b encoding IncX plasmid from lineage A; rendering both lineages resistant to the β-lactam treatment at day 32.

bearing cells. A similar increase in population counts from $10^{3.9}$ to $10^{10.3}$ CFU/g was observed for lineage B as a result of pNK29 acquisition and antibiotic selection (**Figure 1**).

Interestingly, pNK29-bearing lineage B isolates were no longer detected in the subsequent samples collected after cessation of amoxicillin treatment (**Figure 1**). Plasmids often impose a fitness cost upon first encounter with new host backgrounds which could render lineage B less fit in the absence of selection (Porse et al., 2016). Measuring the *in vitro* competitive fitness of the initial plasmid-carrying lineage B isolate revealed a burden of carriage (-4.9%, $sd \pm 4.1\%$, $P = 0.046$); indicating that a counterselection of linage B, due to plasmid invasion, might have taken place after discontinuation of amoxicillin treatment.

Major Genomic Events of Lineage a during Gut Colonization

Apart from the pNK29 plasmid, all isolates of lineage A carried a large virulence plasmid, designated pNK29-2, which was detected throughout the year of sampling. In addition to these two large plasmids, a novel plasmid-element was detected in the 6 m_2 and 12 m isolates (**Figure 2**). This small (2,545 bp) cryptic plasmid was termed pNK29-3 and had a low GC content of 33.4%. Two open reading frames were identified on the plasmid, which encode putative mobilization and replication proteins (**Figure S2**). By comparing the coverage depth of the plasmid to the average coverage depth of the genome, we estimate the copy number to be around nine plasmids per cell. BLAST-analysis revealed a high resemblance to the pIGMS31 previously isolated from *Klebsiella pneumoniae* as well as pEA1 (DQ659147.1) isolated from a Brazilian *Pantoea agglomerans* strain (**Figure S2**; Smorawinska et al., 2012; Carattoli et al., 2014). It was shown that while pIGMS31 can be mobilized to Alpha- and Gamma-proteobacteria, it replicates via a rolling circle mechanism that functions in Gammaproteobacteria only (Smorawinska et al., 2012). Therefore, pNK29-3 likely originates from other Gammaproteobacteria constituents of the gut flora.

Coinciding with the acquisition of pNK29-3 by lineage A, the 6 m_2 and 12 m isolates of lineage A were also missing ∼54 kb of their chromosome compared to the previous isolates. Annotations and flanking attR and attL sites of this region suggested that it was an integrated phage. Using the PHAST server and BLAST searches against NCBI GenBank, we detected a high similarity to the Siphoviridae prophage of *E. coli* strains FH199 and UMNK88 (Zhou et al., 2011). Losing this prophage did not alter the resistance profile of the strain, but might have been a result of negative selection imposed by increased excision activity as a consequence of the cellular stress imposed by antibiotic exposure (Beaber et al., 2004).

Lineage a Harbored a Highly Disseminated and Stable Virulence Plasmid

The pNK29-2 plasmids of lineage A displayed very high sequence identity to several widely disseminated plasmids deposited in NCBI's Genbank (**Figure 3**). Interestingly, 12 highly similar plasmids were previously isolated from various pathogenic *E. coli* strains and one originated from *Klebsiella pneumoniae*. In particular, the endemic pUTI89 plasmid was found to have only 7 SNP differences to the pNK29-2 plasmid, and aligning the contigs from this plasmid showed that there were no additional insertions (Chen et al., 2006). Only two of the SNPs lead to non-synonymous changes. These are located in the *rsvB* gene, a resolvase, and in the *traE* gene, a conjugal transfer protein for F

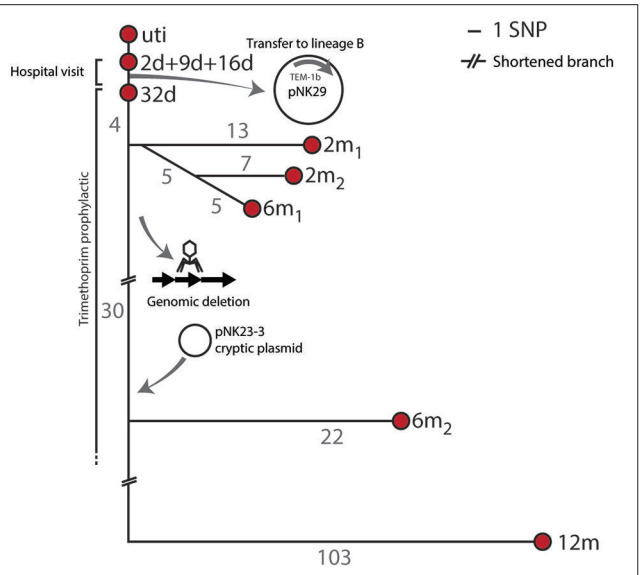


FIGURE 2 | SNP tree and genomic events of lineage A. Branch-numbers show the amount of SNPs separating the isolates obtained from 2 days to 12 months after birth. The majority of samples were isolated during hospitalization and prophylactic treatment. No SNPs were detected between the isolates collected at days 2, 9, and 16, and only one SNP (in the blaTEM1b promoter) was detected between these isolates and the isolate collected at 32 days. Two isolates were included at the 2 months sampling point (2 m1 and 2 m2) and these shared 5 SNPs in common compared to the 32 days isolate. Two 6 month isolates were also included. The infant was subjected to antibiotic treatment from day 11 to 8 months after birth. Apart from SNP differences one sub-lineage had undergone major genomic changes by means of chromosomal deletions and acquisition of new plasmid DNA.

pilus assembly (**Table S3**). We tested the ability of pNK29-2 to conjugate to *E. coli* MG1655, in liquid and on solid media, and we did not detect any transconjugants; implying that the mutated *traE* transfer gene is dysfunctional in pNK29-2.

pNK29-2 Resembles Virulence Plasmids Found in a Diverse Set of Pathogenic *E. coli*

The *E. coli* strain UTI89, harboring the pUTI89 plasmid, is an archetypical uropathogenic *E. coli* (UPEC) strain isolated from a patient with an acute bladder infection (Mulvey et al., 2001). The pUTI89 plasmid belongs to the IncFIB/IIA incompatibility group and shares several characteristics with the F-plasmid including a full *tra* operon for conjugative transfer (Chen et al., 2006). Additionally, its core backbone includes stability mediating genes such as the *ccdA-ccdB* toxin-antitoxin system and the *stbAB* operon ensuring stable inheritance upon cell division (Cusumano et al., 2010). Several of the lesser conserved plasmid regions can be related to virulence and overall adaptation to the human host. These encode the enterotoxicity (*senB*), copper tolerance (*scsC*/*scsD*) and iron acquisition factors. The *cjrABC* operon encodes proteins involved in iron transport that also cause sensitivity to colicin and has shown involvement in UTI virulence (Smajs and Weinstock, 2001; Cusumano et al., 2010).

While a majority of the bacterial hosts carrying these virulence plasmids is associated with UTIs, extremely similar plasmids have been isolated across *E. coli* patho- and sequencetypes from all over the world (**Figure 3**). For example, the prototype neonatal meningitis *E. coli* strain RS218, isolated in 1974, carries a virulence plasmid virtually identical to pUTI89 and pNK29-2, which has been shown to play an important role in its pathogenicity (Wijetunge et al., 2014).

Likewise, the genomic backgrounds hosting these plasmids are diverse (**Figure 3**) and include strains of the dominating extraintestinal pathogenic *E. coli* ST131 of major clinical importance (Stoesser et al., 2016). The geographical locations where these strains have been isolated varies, with highly similar plasmids isolated from different strains in the US, Japan, Canada, and Sweden, suggesting that these pUTI89-like plasmids are globally disseminated in a non-clonal fashion (**Figure 3**).

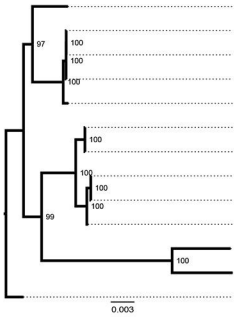
Plasmids Similar to pNK29-2 Carry Antibiotic Resistance Genes Obtained from Independent Insertion Events

Although these plasmids share substantial homology, some show variation within defined but variable "genetic load" regions of their plasmid backbone (**Figure 4**).

Antibiotic use is a strong selection force and some of the plasmids similar to pNK29-2 encode one or more, antibiotic resistance genes organized among mobile element associated cassettes. For instance, the p1ESCUM and pKPN-7c3 plasmids have acquired the *tetA* gene, conferring tetracycline resistance, at different positions within their genetic cargo region, but associated with the same mobile elements (**Figure 4**; Johnson et al., 2016). pMVAST0167_1 and pECO-fce encode multiple antibiotic resistance genes from their genetic load region and share a TEM-1b gene in the same location and accommodate a similar integron with genes conferring resistance toward aminoglycosides (*aadA5*), sulphonamides (*sul1*) and trimethoprim (*dfrA17*). Compared to pMVAST0167_1, pECO-fce encodes several additional resistance genes (*sul2*, *tetA*, *strA*, *strB*, and *tmrB*) from a cassette inserted between the TEM-1b gene and the *int1* integron (**Figure 4**).

pNK29-2 Confers a Fitness Cost to its Host *In vitro*

While the pNK29 plasmid, encoding the TEM-1b β-lactamase, was strongly selected for during the β-lactam treatment administered here (Karami et al., 2007), it is not obvious how large virulence plasmids persist in the gut. The lineage A strain persistently colonized the gut for the entire duration of the study, from 2 days to 1 year after birth, suggesting that it is well adapted to the human host. Such long term survival could be supported by general persistence factors of UPEC strains, suggesting an overlap between UTI virulence and gut persistence factors (Nowrouzian et al., 2005; Chen et al., 2013). Only minor selection is required for a plasmid to survive if it does not impose a significant fitness cost on its host. We used antibiotic resistance markers to tag the plasmid and the host chromosome of lineage A in order to separate the plasmid-bearing variant from the cured variant and



Host strain	ST	Plasmid	Size (kb)	Cov. %	ID %	SNPs	Disease	Location	Year
E. coli Ec29	420	pNK29-2	114	100	100	0	UTI	Sweden	1999
E. coli UTI89	95	pUTI89	114	100	99	7	UTI	US	2001
E. coli RS218	95	RS218	114	100	99	13	NMEC	US	1974
E. coli UM146	643	pUM146	114	100	99	216	Crohn's disease	Canada	2010
E. coli SF-166	95	pSF-166-1	114	100	99	15	Blood infection	US	2008
E. coli SaT040	131 H22	pSaT040	114	100	99	16	ExPEC	US	2007
E. coli ZH063	131 H35	pZH063-1	114	100	99	11	ExPEC	Canada	2002
E. coli MVAST0167	131 H41	pMVAST0167	128	97	99	34	ExPEC	US	2010
E. coli Eco889	131 H41	pECO-fce	212	97	99	283	Unknown	US	2014
E. coli SE15	131 H41	pECSF1	122	99	99	15	ExPEC	Japan	2010
E. coli EcC14	NA	pEC14	114	100	99	50	UTI	US	2009
E. coli UMN026	643	p1ESCUM	122	97	99	4	UTI	Sweden	2010
E. coli ECONIH2	127	pECO-bc6	101	88	99	29	ExPEC	US	2014
K. pneumoniae Kpn555	NA	pkPN-7c3	142	97	99	29	Unknown	US	2014

FIGURE 3 | Overview of plasmids resembling pNK29-2 and their host phylogeny. Thirteen plasmids with high similarity to pNK29-2 were downloaded from Genbank (see **Table S4** for accession numbers and references). Info on host strain, its disease associations and geographical origin was obtained from the literature. The "Year" column denotes the first mentioning of the host strain in the literature unless the isolation year was clearly stated. *In silico* MLST typing was performed and FimH types were added to the ST131 clade highlighting their internal diversity. A core-genome based maximum-likelihood tree was constructed to illustrate the diversity of the *E. coli* plasmid hosts. Node numbers are bootstrap confidence values.

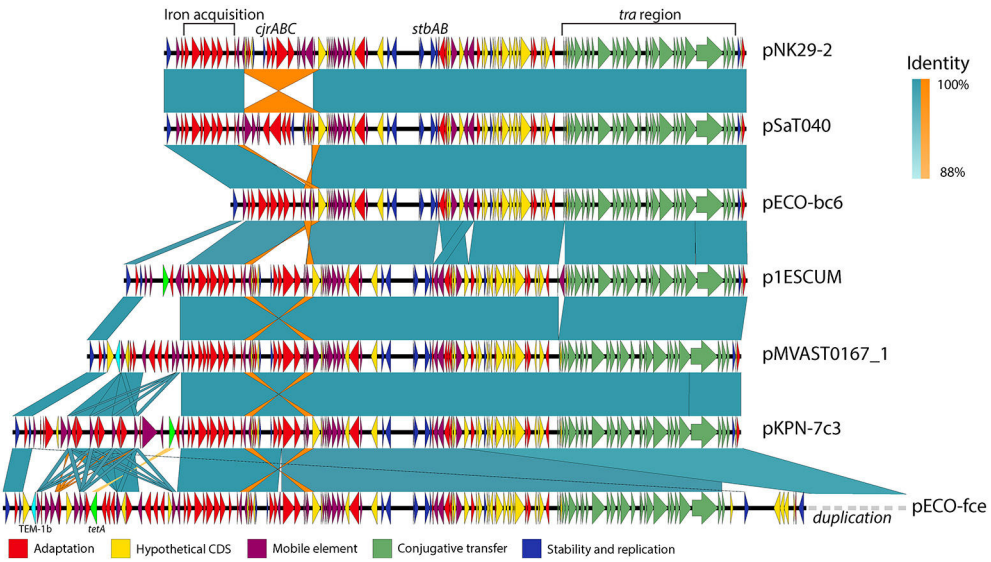


FIGURE 4 | Genetic variability and conservation within similar virulence plasmid backbones. six out of the Thirteen virulence plasmids similar to pNK29-2 showed signs of major restructuring events and are illustrated here. While these plasmids show some degree of variation, they also share a conserved core of transfer (green), stability (blue), and virulence genes (red). Mobile elements constituting inverted repeats (highlighted in orange) allow for instability of the *cjrABC*-containing region exemplified in pSaT040 and pECO-bc6. Additional genes involved in antibiotic resistance have been inserted upstream the fully conserved iron acquisition cluster of the genetic load region in p1ESCUM, pMVAST0167_1, pKPN7c3, and pECO-fce. The TEM-1b and *tetA* genes are highlighted in cyan and light green respectively. pECO-fce is significantly larger than the remaining plasmids due to a duplication of its transfer region which has been condensed for simplicity. Colored shades illustrate BLAST identity of pairwise comparisons and orange shades highlight inverted regions.

detect plasmid loss as well as transfer events. We used markers previously used for tagging in similar fitness experiments, where they did not impose a measurable cost (Chen et al., 2013). Similarly, we also confirmed that the Sh *ble* Zeocin resistance gene, used for plasmid tagging, did not impose a significant cost during 5 days of competitive growth against the non-tagged variant (two sample *t*-test, $P = 0.11$).

The cured and plasmid bearing strains were mixed in equal proportions and propagated in LB medium for 8 days without selection. From CFU quantifications on selective agar plates, the average *in vitro* fitness cost of the plasmid-carrying strain was measured to be 0.92 sd $\pm$ 0.25% per generation. Although this cost of plasmid carriage is low, one would expect a steady decline of the plasmid bearing cells under these growth conditions. Due to the presence of putative iron acquisition systems on the pNK29-2 plasmid, we also tested the competitive fitness as well as absolute growth rates in iron-limited M9 minimal medium containing 5 μM $FeCl_3$ mimicking the lower range of

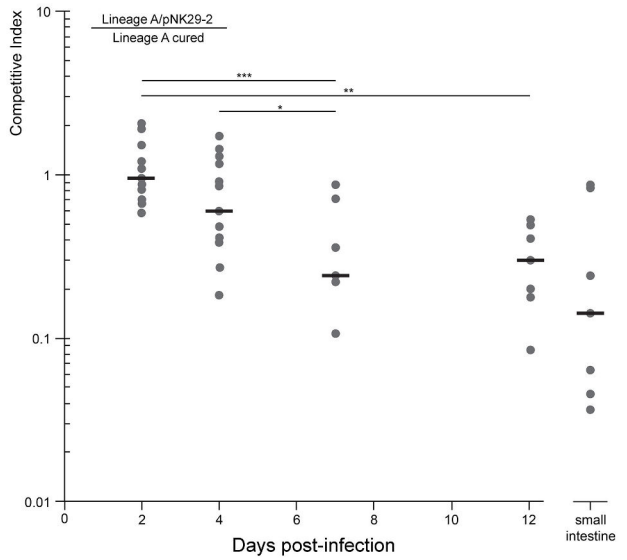


FIGURE 5 | Competitive growth of lineage a strain against a pNK29-2 cured variant within streptomycin treated mice. Mice where inoculated by oral gavage with equal amounts of plasmid carrying and plasmid-free cells. The competitive index (top left equation) is the ratio of CFUs obtained from selective plating. This ratio was also quantified for the small intestine at the endpoint (day 12) of the experiment to assess the potential role of adhering cells. Replicates were excluded when the strains were no longer detected. Significance indicators illustrate pairwise comparisons of means (Mann-Whitney *U*-Test: $^{*}P = 0.014$, $^{**}P = 0.0015$, $^{***}P = 0.001$)

physiological concentrations (Wang, 1996). We were not able to detect a significant advantage in terms of growth rate (Two sample *t*-test, $P = 0.43$—**Figure S3**) or competitive fitness of the plasmid carrying strain when mixed 1:1 in the same medium ($1.07 \pm 2\%$, one sample *t*-test, $P = 0.45$).

The Plasmid Carrying Strain is Outcompeted *In vivo* of the Mouse Gut

Because the plasmid encodes a diverse set of factors thought to be involved in iron acquisition, toxin production as well as several hypothetical proteins, that might be selected for in more complex *in vivo* settings, we set out to test the competitive fitness in a gut environment of streptomycin treated mice. The plasmid-cured strain was competed against its plasmid-carrying ancestor in the mouse gut and we measured the proportion of plasmid-carrying to plasmid-free cells in the feces over the course of 12 days (**Figure 5**). Here, we observed a steady decline in plasmid-carrying cells compared to plasmid-free cells with the plasmid-free cells dominating the average population after 7 days of direct competition. Interestingly, the plasmid-carrying populations avoid complete out-competition by plasmid-free cells in this time-span (**Figure 5**).

To investigate whether the measurements obtained from feces were representative of the intestinal contents, including cells adhering to the intestinal wall, we measured the proportion of plasmid-bearing cells sampled directly from the small intestine of the mice at the end of the 12 days experiment. There was no significant difference between the direct sampling compared to fecal counts (Mann-Whitney *U*-Test, $P = 0.27$) and >10% of the cells contained the plasmid in the average replicate population at this stage (**Figure 5**).

From the competition experiment we calculated the average fitness cost per generation of the plasmid in the mouse gut to be 0.83 *sd* $\pm$ 0.2%, assuming that the competing *E. coli* underwent 10 generations per day (Rang et al., 1999; Lee et al., 2010; Myhrvold et al., 2015). This cost is slightly lower, but not significantly different, from the plasmid cost measured *in vitro* (Two sample *t*-test: $P = 0.61$).

pNK29-2 is Stably Inherited Despite its Cost

Although the plasmid-bearing strain is less fit than the plasmid-cured variant *in vivo*, it seems possible that a minor subpopulation of plasmid-bearing cells can persist for extended time periods; especially if competition from plasmid-free daughter cells is postponed via stable inheritance mechanisms. As for all the virulence plasmids analyzed here (**Figure 3**), pNK29-2 carries the highly conserved *stbAB* stability operon (**Figure 4**) that encodes active segregation machinery; ensuring that the plasmid is stably segregated to both daughter cells during cell division (Guynet et al., 2011). To test the segregational stability of pNK29-2 in its native lineage A host (day 0 isolate), we conducted 14 days of serial passaging in LB medium, corresponding to ~140 generations of growth. In such a setup, plasmid-free segregants would eventually take over the population due to the cost of plasmid carriage. However, we did not observe any plasmid-free cells within this time span with a detection limit of plasmid-free cells of ~1%. Similarly, no plasmid-free cells were detected in the *in vivo* competition assays, during the 12 days of gut colonization.

Taken together, these results imply that virulence plasmids, such as pNK29-2, have no direct advantage in gut colonization but are able to persist in spite of a small, but significant, fitness cost due to efficient plasmid inheritance mechanisms.

DISCUSSION

Culture independent methods based on metagenomic sequencing have been used to investigate the abundance profiles of strains colonizing the gut over time (Morowitz et al., 2011; Brown et al., 2013; Sharon et al., 2013). However, these methods are limited in their ability to observe genomic events at high resolution such as horizontal transfer and single nucleotide variations.

Due to our longitudinal sampling and the high resolution of single isolate genome sequencing, we were able to observe a glimpse of the complex genome dynamics of *E. coli* in its native settings. We confirm a gut-inhabiting strain as the origin of a bladder infection; supporting the general belief that UTIs are caused by gut inhabiting *E. coli* strains that eventually enters the urethra (Chen et al., 2013). Furthermore, we report the sequence of pNK29, a novel 42.2 kb IncX plasmid carrying a TEM-1b β-lactamase, which was transferred between the two co-existing *E. coli* lineages of the gut. Few phenotypic reports exist documenting plasmid mediated HGT of antibiotic resistance genes between bacteria in the human gut and our data supports

that transfer of resistance genes take place in the gut, and may be enhanced by antibiotic treatment (Bidet et al., 2005; Karami et al., 2007; Trobos et al., 2009; Goren et al., 2010). The recipient lineage was only sampled at one time-point after the termination of β-lactam antibiotic administration (day 32) and declined to undetectable levels thereafter. This could indicate a negative selection of pNK29-carrying lineage B isolates in the absence of antibiotic selection; however, confirming the role of pNK29 in the counterselection of lineage B in the gut will require further *in vivo* competition experiments. Lineage B showed a large drop in population counts when subjected to the first round of trimethoprim treatment upon hospitalization (**Figure 1**). The prophylactic administration of trimethoprim coincided with the disappearance of lineage B, and could be another likely explanation for its absence in the consecutive time points.

While we did not observe any genomic alterations of lineage B apart from the acquisition of pNK29, lineage A experienced chromosomal deletions and acquired a cryptic plasmid as well as a high number of SNPs during the sampling period (**Figure 2**).

Genome plasticity is believed to play a crucial role in the adaptation of pathogens to the selective forces imposed by the immune system or the remaining microbiota within a human host (Brzuszkiewicz et al., 2009). The mutation rate observed for lineage A was high and resemble that of mutator phenotypes that are often enriched among UPEC isolates (Labat et al., 2005). Such increased rates of mutation and recombination events might also be the result of antibiotic treatment of the infant; leading to induction of the bacterial SOS response, which has been shown to increase mutation rates (Beaber et al., 2004; Michel, 2005).

Non-synonymous mutations were indeed detected in genes related to antibiotic tolerance, e.g., those involved in folate metabolism (*folA*—targeted by trimethoprim), fusaric acid resistance (*fusB*), ABC-transport and membrane permeability (porins; **Table S1**). Equivalently, the genomic deletion of the 54 kb region might have resulted from antibiotic mediated stress known to induce prophage excision and increase horizontal gene transfer in general (Beaber et al., 2004; Nanda et al., 2015).

Apart from small cryptic plasmids providing no obvious selective advantage to their bacterial host, gut-inhabiting *E. coli* isolates often carry plasmids that allow adaptation toward the human host by contributing virulence or antibiotic resistance factors (Johnson and Russo, 2005). In addition to the β-lactamase carrying pNK29 IncX plasmid, we identified a 114 kb plasmid (pNK29-2) in lineage A that was strikingly similar to other previously sequenced virulence plasmids from a diverse set of pathogenic *E. coli* strains (**Figure 3**). These plasmids have been shown to play a role in the initial stages of UTI infection in a mouse model in a different genetic host background, and could have provided lineage A with the necessary virulence factors leading to the successful UTI infection observed in the studied infant. Although it is generally believed that specific pathotypes of *E. coli* carry different virulence plasmids, plasmid backbones virtually identical to pNK29-2 have recently been found in *K. pneumoniae* as well as several divergent *E. coli* strains; with the earliest isolate dating back to 1974 (**Figure 3**). The high conservation of these plasmids suggests that they provide a universal adaptive benefit to their ExPEC hosts regardless of infection site (Johnson and Nolan, 2009; Cusumano et al., 2010; Wijetunge et al., 2014).

When comparing the genetic composition of the currently sequenced virulence plasmids with high similarity to pNK29-2, it is clear that certain regions tend to preserve genetic features across host, geography and time (**Figures 3**, **4**). These include mediators of iron acquisition, toxin production and putative copper resistance mediators (*scsC* and *scsD*; DebRoy et al., 2010). Carrying genes implicated in virulence, these plasmids could confer a survival advantage to their bacterial host during infection. Prior studies have examined the role of pUTI89 and pRS218 in urinary tract infections and neonatal meningitis, respectively (Cusumano et al., 2010; Wijetunge et al., 2014). These studies did not observe any phenotypic differences *in vitro* between the plasmid bearing and cured host in terms of growth rate, type 1 pilus expression or biofilm formation. However, they did observe a significant difference in infectivity using rodent infection models (Cusumano et al., 2010; Wijetunge et al., 2014).

A vital defense mechanism of the human body is to restrict iron from pathogens, thus acquisition and transport of iron is an important survival mechanism for ExPEC strains *in vivo* (Andrews et al., 2003). Therefore, iron acquisition could be beneficial for survival in many niches of the human body, including the densely populated gut microbiota, where access to iron is limited (Andrews et al., 2003; Nowrouzian et al., 2003). While Cusumano et al. speculated that the *cjr* operon of pUTI89 was beneficial in a UTI infection scenario due to its putative involvement in iron acquisition, we could not detect an advantage in neither absolute growth rate nor competitive fitness of lineage A carrying the pNK29-2 plasmid when grown in iron-limited conditions (Cusumano et al., 2010). Although the effect was small or non-existent in our experimental setups, pNK29-2 might provide an advantage by other means. For example, the pNK29-2 genes encoding siderophore receptors or transporters might provide an advantage, only if the available iron is on a certain form e.g., bound by its respective siderophore. As other iron acquisition systems are located on the chromosome of lineage A, the pNK29-2 encoded systems could be redundant in this strain, but might be selected in other hosts to encourage plasmid maintenance in a communal context.

Previous studies have shown that even minor differences in host genomes can be highly influential in determining plasmid establishment as well as subsequent adaptation and long term persistence (Humphrey et al., 2012; Porse et al., 2016). Thus, it is intriguing that virulence plasmids imposing a minor, but significant, fitness cost without providing any strongly selected phenotypes, can persist in a competitive environment such as the human gut. From our *in vivo* competition experiment, it is clear that the pNK29-2 carrying strain was less fit in the murine gut but does reach stable counts from day 7 to 12 (**Figure 5**). This stagnation in competition could encourage extended plasmid persistence and might be explained by changes in the growth rate of gut inhabiting strains over time; indicating a non-constant selection pattern

(Rang et al., 1999). Such selection patterns are known to occur in mixed bacterial populations encompassing social iron acquisition phenotypes and similar dynamics could take place in the competition between pNK29-2 carrying and plasmid-free strains in the gut (Stojiljkovic et al., 1993; Ross-gillespie et al., 2007). However, this does not seem to be the case from our growth rate measurements in iron-limiting conditions, were the plasmid carrying strain did not have an advantage on its own (**Figure S3**).

The measured fitness cost of the pNK29-2 in the lineage A strain was surprisingly similar between *in vitro* and *in vivo* setups, and is low compared to previous observations of plasmid costs (Vogwill and MacLean, 2015). While plasmids can in theory compensate their loss and fitness cost by re-infection of plasmid-free hosts (Slater et al., 2008), this is unlikely to be the case for pNK29-2 as we did not observe any transconjugants in our *in vitro* conjugation assays nor during the *in vivo* competition experiment. The inability of plasmids to conjugate might be explained by the fitness constraints a functional conjugation machinery can impose on certain hosts; supported by the existence of plasmid variants, such as pCE10A from Lu et al., lacking the conjugative transfer operon, that are otherwise identical to pNK29-2 (Lu et al., 2011; Porse et al., 2016).

Loss of pNK29-2 was not observed, neither *in vivo* nor *in vitro*, suggesting that primary stability mechanisms such as active segregation and toxin-antitoxin systems are the most important persistence parameters for these plasmids. This is consistent with the high degree of conservation of stability systems among all 14 plasmids examined here and further supported by the fact that we, as well as previous studies, have experienced considerable trouble curing strains from these plasmids; unless the *stbAB* operon is disrupted (Cusumano et al., 2010; Wijetunge et al., 2014).

By characterizing the genomes of persistent lineages of *E. coli* colonizing the gut of an infant, we observed substantial dynamics, highlighting that strains colonizing the human gut undergo continuous change. While genomic plasticity can lead to improved persistence, some elements are surprisingly stable. Our cost and stability characterizations suggest that a low cost and a high segregational stability, combined with plasmid-encoded universal virulence factors, presumed to increase fitness in a broad range of infection scenarios (Cusumano et al., 2010; Wijetunge et al., 2014), are likely the main parameters governing the success of endemic virulence plasmids. Further understanding of the factors contributing to genomic variation of gut colonizing pathogens will aid in rational interventions against the virulence and antibiotic resistance determinants widely disseminated among these isolates.

DATA AVAILABILITY

All sequenced genomes can be accessed via the Bioproject PRJNA352659.

AUTHOR CONTRIBUTIONS

NK, IA, and AW provided the *E. coli* isolates. AP did the *in vitro* work and strain tagging. JK performed the *in vivo* competition experiment. HG processed the sequencing data. AP, HG, and MS analyzed the sequencing data results and AP and HG wrote the manuscript with input from MS, JK, DA, NK, and IA.

FUNDING

This research was funded by the EU H2020 ERC-20104-STG LimitMDR (638902) and the Danish Council for Independent Research Sapere Aude programme DFF -4004-00213, the Medical Faculty of the University of Göteborg (ALFGBG138401) and the Swedish Medical Research Council.

ACKNOWLEDGMENTS

We thank Mari Cristina Rodriguez de Evgrafo and Marius Faza for aid in sequence library preparation. Christian Munck and Lejla Imamovic are thanked for helpful discussions. MS acknowledges support from the Novo Nordisk Foundation, and the Lundbeck Foundation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fcimb.2017.00126/full#supplementary-material

Figure S1 | Plasmid map and BLAST comparison of pNK29. pNK29 compared to IncX1 plasmids pOLA52 (outer ring, green) and pRPEC180_47(middle ring, blue). Open reading frames are drawn directionally (inner ring, black). Selected annotations are labeled outside the ring (see **Table S2** for full annotation list).

Figure S2 | Plasmid pNK29-3 compared to most similar plasmids in GeneBank. The two ORFs encode putative replication and mobilization proteins and open reading frames are drawn directionally (inner ring, black).

Figure S3 | Growth rate of lineage A with and without pNK29-2 under iron-limited conditions.

Table S1 | SNPs in different isolates of lineage A. Table containing the SNPs from lineage A, including the annotation and whether the animo acid change was synonymous or non-synonymous. The isolates from days 2, 9, and 16 are left out as they did not have any SNP differences from the representative genome for the lineage (isolate taken at 2d).

Table S2 | Annotations for pNK29. Names in square brackets indicate homologs in pOLA52.

Table S3 | SNPs in the pNK29-2 plasmid of lineage A compared to pUTI89. SNPs were identified by comparing the initial linage A isolate (day 2) with pUTI89 (CP000244.1).

Table S4 | Plasmids similar to pNK29-2 and information on their origin. The table presented in **Figure 3**, but more detailed information on size, location, and references. The GenBank ID refers to the plasmid sequence entry in GenBank and the Pubmed ID (PMID) to the study where the plasmid was initially described.

REFERENCES

Adlerberth, I., Svanborg, C., Carlsson, B., Mellander, L., Hanson, L. A., Jalil, F., et al. (1998). P fimbriae and other adhesins enhance intestinal persistence of *Escherichia coli* in early infancy. *Epidemiol. Infect.* 121, 599–608. doi: 10.1017/S0950268898001137

Adlerberth, I., Lindberg, E., Aberg, N., Hesselmar, B., Saalman, R., Strannegård, I. L., et al. (2006). Reduced enterobacterial and increased staphylococcal colonization of the infantile bowel: an effect of hygienic lifestyle? *Pediatr. Res.* 59, 96–101. doi: 10.1203/01.pdr.0000191137.12774.b2

Alikhan, N.F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. doi: 10.1186/1471-2164-12-402

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Andrews, S. C., Robinson, A. K., and Rodríguez-Quiñones, F. (2003). Bacterial iron homeostasis. *FEMS Microbiol. Rev.* 27, 215–237. doi: 10.1016/S0168-6445(03)00055-X

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Beaber, J. W., Hochhut, B., and Waldor, M. K. (2004). SOS response promotes horizontal dissemination of antibiotic resistance genes. *Nature* 427, 72–74. doi: 10.1038/nature02241

Bidet, P., Burghoffer, B., Gautier, V., Brahimi, N., Mariani-Kurkdjian, P., El-Ghoneimi, A., et al. (2005). *In vivo* transfer of plasmid-encoded ACC-1 AmpC from *Klebsiella pneumoniae* to *Escherichia coli* in an infant and selection of impermeability to imipenem in *K.* pneumoniae. *Antimicrob. Agents Chemother.* 49, 3562–3565. doi: 10.1128/AAC.49.8.3562-3565.2005

Brown, C. T., Sharon, I., Thomas, B. C., Castelle, C. J., Morowitz, M. J., and Banfield, J. F. (2013). Genome resolved analysis of a premature infant gut microbial community reveals a *Varibaculum cambriense* genome and a shift towards fermentation-based metabolism during the third week of life. *Microbiome* 1:30. doi: 10.1186/2049-2618-1-30

Brzuszkiewicz, E., Gottschalk, G., Ron, E., Hacker, J., and Dobrindt, U. (2009). Adaptation of pathogenic *E. coli* to various niches : genome flexibility is the key. *Microb. Pathog.* 6, 110–125. doi: 10.1159/000235766

Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., et al. (2014). *In silico* detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* 58, 3895–3903. doi: 10.1128/AAC.02412-14

Chen, S. L., Hung, C. S., Xu, J., Reigstad, C. S., Magrini, V., Sabo, A., et al. (2006). Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5977–5982. doi: 10.1073/pnas.0600938103

Chen, S. L., Wu, M., Henderson, J. P., Hooton, T. M., Hibbing, M. E., Hultgren, S. J., et al. (2013). Genomic diversity and fitness of *E. coli* strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection. *Sci. Transl. Med.* 5, 184ra60. doi: 10.1126/scitranslmed.3005497

Conlan, S., Thomas, P. J., Deming, C., Park, M., Lau, A. F., Dekker, J. P., et al. (2014). Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Sci. Transl. Med.* 6, 254ra126. doi: 10.1126/scitranslmed.3009845

Conlan, S., Park, M., Deming, C., Thomas, P. J., Young, A. C., Coleman, H., et al. (2016). Plasmid dynamics in KPC-positive klebsiella pneumoniae during long-term patient colonization. *MBio* 7, e00742-16. doi: 10.1128/mBio.00742-16

Cusumano, C. K., Hung, C. S., Chen, S. L., and Hultgren, S. J. (2010). Virulence plasmid harbored by uropathogenic *Escherichia coli* functions in acute stages of pathogenesis. *Infect. Immun.* 78, 1457–1467. doi: 10.1128/IAI.01260-09

DebRoy, C., Sidhu, M. S., Sarker, U., Jayarao, B. M., Stell, A. L., Bell, N. P., 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. doi: 10.1016/j.plasmid.2009.10.003

Diard, M., Garry, L., Selva, M., Mosser, T., Denamur, E., and Matic, I. (2010). Pathogenicity-associated islands in extraintestinal pathogenic *Escherichia coli* are fitness elements involved in intestinal colonization. *J. Bacteriol.* 192, 4885–4893. doi: 10.1128/JB.00804-10

Dias, R. C., Moreira, B. M., and Riley, L. W. (2010). Use of fimH single-nucleotide polymorphisms for strain typing of clinical isolates of *Escherichia coli* for epidemiologic investigation. *J. Clin. Microbiol.* 48, 483–488. doi: 10.1128/JCM.01858-09

Drasar, B. S., and Hill M. J. (1974). *Human Intestinal Flora*. London: Academic Press Inc.

Foxman, B. (2010). The epidemiology of urinary tract infection. *Nat. Rev. Urol.* 7, 653–660. doi: 10.1038/nrurol.2010.190

Goren, M. G., Carmeli, Y., Schwaber, M. J., Chmelnitsky, I., Schechner, V., and Navon-Venezia, S. (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

Guynet, C., Cuevas, A., Moncalián, G., and de la Cruz, F. (2011). The stb operon balances the requirements for vegetative stability and conjugative transfer of plasmid R388. *PLoS Genet.* 7:e1002073. doi: 10.1371/journal.pgen.1002073

Humphrey, B., Thomson, N. R., Thomas, C. M., Brooks, K., Sanders, M., Delsol, A. A., et al. (2012). Fitness of *Escherichia coli* strains carrying expressed and partially silent IncN and IncP1 plasmids. *BMC Microbiol.* 12:53. doi: 10.1186/1471-2180-12-53

Johnson, T. J., and Nolan, L. K. (2009). Pathogenomics of the virulence plasmids of *Escherichia coli*. *Microbiol. Mol. Biol. Rev.* 73, 750–774. doi: 10.1128/MMBR.00015-09

Johnson, J. R., and Russo, T. A. (2005). Molecular epidemiology of extraintestinal pathogenic (uropathogenic) *Escherichia coli*. *Int. J. Med. Microbiol.* 295, 383–404. doi: 10.1016/j.ijmm.2005.07.005

Johnson, T. J., Danzeisen, J. L., Youmans, B., Case, K., Llop, K., Munoz-Aguayo, J., et al. (2016). Separate F-Type plasmids have shaped the evolution of the H 30 subclone of *Escherichia coli* sequence type 131. *mSphere* 1, e00121–e00116. doi: 10.1128/mSphere.00121-16

Karami, N., Martner, A., Enne, V. I., Swerkersson, S., Adlerberth, I., and Wold, A. E. (2007). Transfer of an ampicillin resistance gene between two *Escherichia coli* strains in the bowel microbiota of an infant treated with antibiotics. *J. Antimicrob. Chemother.* 60, 1142–1145. doi: 10.1093/jac/dkm327

Köhler, C. D., and Dobrindt, U. (2011). What defines extraintestinal pathogenic *Escherichia coli*? *Int. J. Med. Microbiol.* 301, 642–647. doi: 10.1016/j.ijmm.2011.09.006

Kuhlman, T. E., and Cox, E. C. (2010). Site-specific chromosomal integration of large synthetic constructs. *Nucleic Acids Res.* 38:e92. doi: 10.1093/nar/gkp1193

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12. doi: 10.1186/gb-2004-5-2-r12

Labat, F., Pradillon, O., Garry, L., Peuchmaur, M., Fantin, B., and Denamur, E. (2005). Mutator phenotype confers advantage in *Escherichia coli* chronic urinary tract infection pathogenesis. *FEMS Immunol. Med. Microbiol.* 44, 317–321. doi: 10.1016/j.femsim.2005.01.003

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25. doi: 10.1186/gb-2009-10-3-r25

Larsen, M. V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R. L., et al. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. *J. Clin. Microbiol.* 50, 1355–1361. doi: 10.1128/JCM.06094-11

Lasaro, M., Liu, Z., Bishar, R., Kelly, K., Chattopadhyay, S., Paul, S., et al. (2014). *Escherichia coli* isolate for studying colonization of the mouse intestine and its application to two-component signaling knockouts. *J. Bacteriol.* 196, 1723–1732. doi: 10.1128/JB.01296-13

Lee, S. M., Wyse, A., Lesher, A., Everett, M. L., Lou, L., Holzknecht, Z. E., et al. (2010). Adaptation in a mouse colony monoassociated with *Escherichia coli* K-12 for more than 1,000 Days. *Appl. Environ. Microbiol.* 76, 4655–4663. doi: 10.1128/AEM.00358-10

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lu, S., Zhang, X., Zhu, Y., Kim, K. S., Yang, J., and Jin, Q. (2011). Complete genome sequence of the neonatal-meningitis-associated *Escherichia coli* strain CE10. *J. Bacteriol.* 193, 7005. doi: 10.1128/JB.06284-11

Lutz, R., and Bujard, H. (1997). Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res.* 25, 1203–1210.

M9 minimal medium (standard) (2010). M9 minimal medium (standard). *Cold Spring Harb. Protoc.* 2010:pdb.rec12295. doi: 10.1101/pdb.rec12295

Marchesi, J. R., Adams, D. H., Fava, F., Hermes, G. D., Hirschfield, G. M., Hold, G., et al. (2015). The gut microbiota and host health: a new clinical frontier. *Gut* 65, 330–339. doi: 10.1136/gutjnl-2015-309990

Michel, B. (2005). After 30 years of study, the bacterial SOS response still surprises us. *PLoS Biol.* 3, 1174–1176. doi: 10.1371/journal.pbio.0030255

Morowitz, M. J., Denef, V. J., Costello, E. K., Thomas, B. C., Poroyko, V., Relman, D. A., et al. (2011). Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1128–1133. doi: 10.1073/pnas.1010992108

Mulvey, M. A., Schilling, J. D., and Hultgren, S. J. (2001). Establishment of a persistent *Escherichia coli* reservoir during the acute phase of a bladder infection. *Infect. Immun.* 69, 4572–4579. doi: 10.1128/IAI.69.7.4572-4579.2001

Myhrvold, C., Kotula, J. W., Hicks, W. M., Conway, N. J., and Silver, P. A. (2015). A distributed cell division counter reveals growth dynamics in the gut microbiota. *Nat. Commun.* 6:10039. doi: 10.1038/ncomms10039

Nanda, A. M., Thormann, K., and Frunzke, J. (2015). Impact of spontaneous prophage induction on the fitness of bacterial populations and host-microbe interactions. *J. Bacteriol.* 197, 410–419. doi: 10.1128/JB.02230-14

Norman, A., Hansen, L. H., She, Q., and Sørensen, S. J. (2008). Nucleotide sequence of pOLA52: a conjugative IncX1 plasmid from *Escherichia coli* which enables biofilm formation and multidrug efflux. *Plasmid* 60, 59–74. doi: 10.1016/j.plasmid.2008.03.003

Norman, A., Hansen, L. H., and Sørensen, S. J. (2009). Conjugative plasmids: vessels of the communal gene pool. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 2275–2289. doi: 10.1098/rstb.2009.0037

Nowrouzian, F., Hesselmar, B., Saalman, R., Strannegård, I. L., Aberg, N., Wold, A. E., et al. (2003). *Escherichia coli* in infants' intestinal microflora: colonization rate, strain turnover, and virulence gene carriage. *Pediatr. Res.* 54, 8–14. doi: 10.1203/01.PDR.0000069843.20655.EE

Nowrouzian, F. L., Wold, A. E., and Adlerberth, I. (2005). *Escherichia coli* strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora of infants. *J. Infect. Dis.* 191, 1078–1083. doi: 10.1086/427996

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

Porse, A., Schoønning, K., Munck, C., and Sommer, M. O. (2016). Survival and evolution of a large multidrug resistance plasmid in new clinical bacterial hosts. *Mol. Biol Evol.* 33, 2860–2873. doi: 10.1093/molbev/msw163

Quinlan, A. R., and Hall, I. M. (2010). Bed tools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rang, C. U., Licht, T. R., Midtvedt, T., Conway, P. L., Chao, L., Krogfelt, K. A., et al. (1999). Estimation of growth rates of *Escherichia coli* BJ4 in streptomycin-treated and previously germfree mice by in situ rRNA hybridization. *Clin. Diagn. Lab. Immunol.* 6, 434–436. doi: 10.1128/JB.01581-07

Ross-gillespie, A., Gardner, A., West, S. A., and Griffin, A. S. (2007). Frequency dependence and cooperation: theory and a test with bacteria. *Am. Nat.* 170, 331–342. doi: 10.1086/519860

Salyers, A. A., Gupta, A., and Wang, Y. (2004). Human intestinal bacteria as reservoirs for antibiotic resistance genes. *Trends Microbiol.* 12, 412–416. doi: 10.1016/j.tim.2004.07.004

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., and Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* 23, 111–120. doi: 10.1101/gr.142315.112

Simonsen, L. (1991). The existence conditions for bacterial plasmids: theory and reality. *Microb. Ecol.* 22, 187–205.

Slater, F. R., Bailey, M. J., Tett, A. J., and Turner, S. L. (2008). Progress towards understanding the fate of plasmids in bacterial communities. *FEMS Microbiol. Ecol.* 66, 3–13. doi: 10.1111/j.1574-6941.2008.00505.x

Smajs, D., and Weinstock, G. M. (2001). The iron- and temperature-regulated cjrBC genes of Shigella and enteroinvasive *Escherichia coli* strains code for colicin Js uptake. *J. Bacteriol.* 183, 3958–3966. doi: 10.1128/JB.183.13.3958-3966.2001

Smith, J. L., Fratamico, P. M., and Gunther, N. W. (2007). Extraintestinal pathogenic *Escherichia coli*. *Foodborne Pathog. Dis.* 4, 134–163. doi: 10.1089/fpd.2007.0087

Smorawinska, M., Szuplewska, M., Zaleski, P., Wawrzyniak, P., Maj, A., Plucienniczak, A., et al. (2012). Mobilizable narrow host range plasmids as natural suicide vectors enabling horizontal gene transfer among distantly related bacterial species. *FEMS Microbiol. Lett.* 326, 76–82. doi: 10.1111/j.1574-6968.2011.02432.x

Sommer, M. O., Church, G. M., and Dantas, G. (2010). The human microbiome harbors a diverse reservoir of antibiotic resistance genes. *Virulence* 1, 299–303. doi: 10.4161/viru.1.4.12010

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Stoesser, N., Sheppard, A. E., Pankhurst, L., De Maio, N., Moore, C. E., Sebra, R., et al. (2016). Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. *MBio* 7, e02162–e02115. doi: 10.1128/mBio.02162-15

Stojiljkovic, I., Cobeljic, M., and Hantke, K. (1993). *Escherichia coli* K-12 ferrous iron–uptake mutants are impaired in their ability to colonize the mouse intestineNo Title. *FEMS Microbiol. Lett.* 108, 111–115.

Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. - PubMed - NCBI. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039

de Toro, M., Garcillán-barcia, M. P., and De La Cruz, F. (2014). Plasmid diversity and adaptation analyzed by massive sequencing of *Escherichia coli* plasmids. *Microbiol. Spectr.* 2, 1–16. doi: 10.1128/microbiolspec.PLAS-0031-2014

Trobos, M., Lester, C. H., Olsen, J. E., Frimodt-Møller, N., and Hammerum, A. M. (2009). Natural transfer of sulphonamide and ampicillin resistance between *Escherichia coli* residing in the human intestine. *J. Antimicrob. Chemother.* 63, 80–86. doi: 10.1093/jac/dkn437

Vogwill, T., and MacLean, R. C. (2015). The genetic basis of the fitness costs of antimicrobial resistance: a meta-analysis approach. *Evol. Appl.* 8, 284–295. doi: 10.1111/eva.12202

Wang, C.T. (1996). Concentration of arsenic, selenium, zinc, iron and copper in the urine of blackfoot disease patients at different clinical stages. *Eur. J. Clin. Chem. Clin. Biochem.* 34, 493–497.

Wijetunge, D. S., Karunathilake, K. H., Chaudhari, A., Katani, R., Dudley, E. G., Kapur, V., et al. (2014). Complete nucleotide sequence of pRS218, a large virulence plasmid, that augments pathogenic potential of meningitis-associated *Escherichia coli* strain RS218. *BMC Microbiol.* 14:203. doi: 10.1186/s12866-014-0203-9

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Res.* 39, 347–352. doi: 10.1093/nar/gkr485

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary material

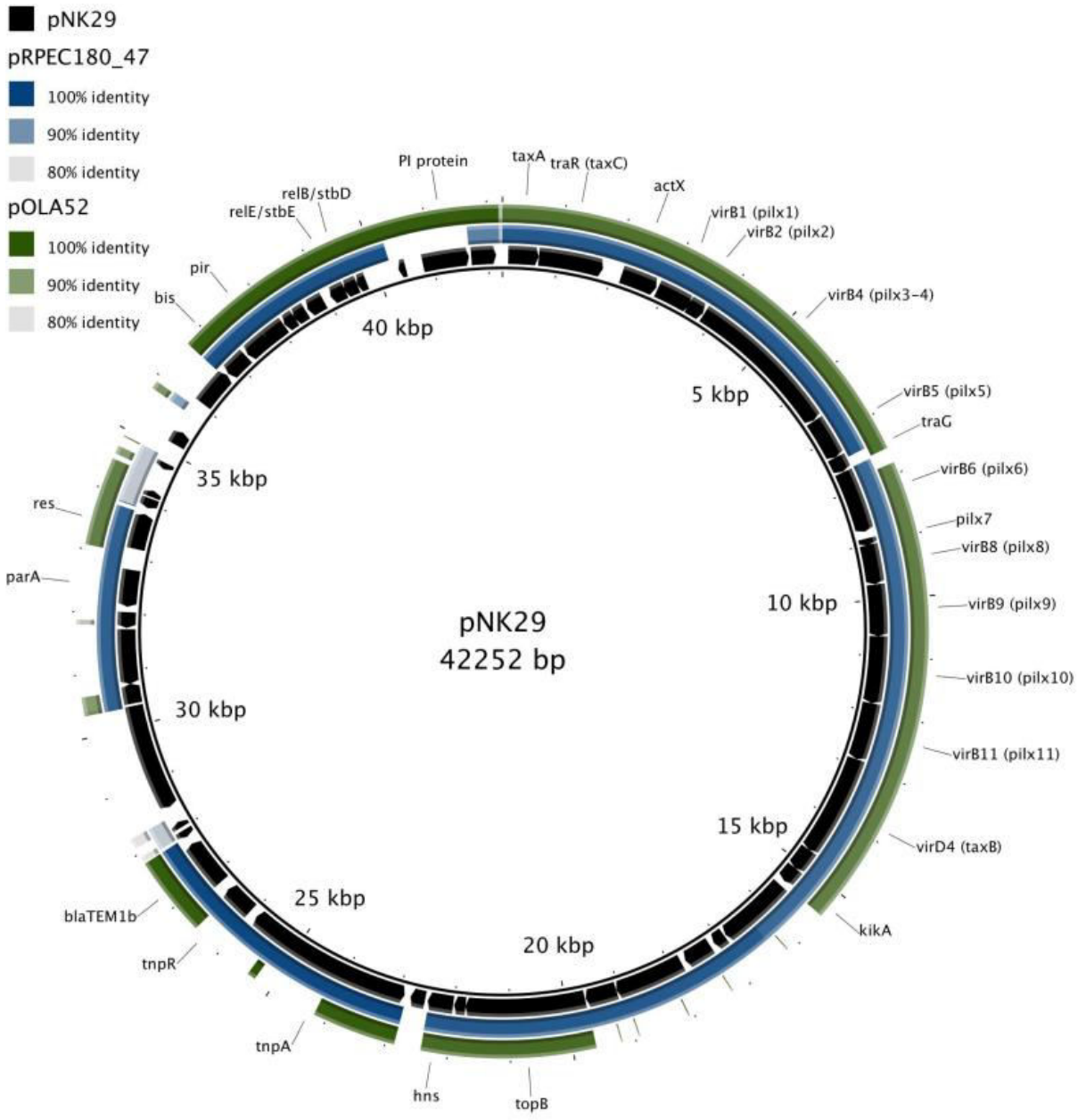


Figure S1. Plasmid map and BLAST comparison of pNK29. pNK29 compared to IncX1 plasmids pOLA52 (outer ring, green) and pRPEC180_47(middle ring, blue). Open reading frames are drawn directionally (inner ring, black). Selected annotations are labeled outside the ring (see Table S2 for full annotation list).

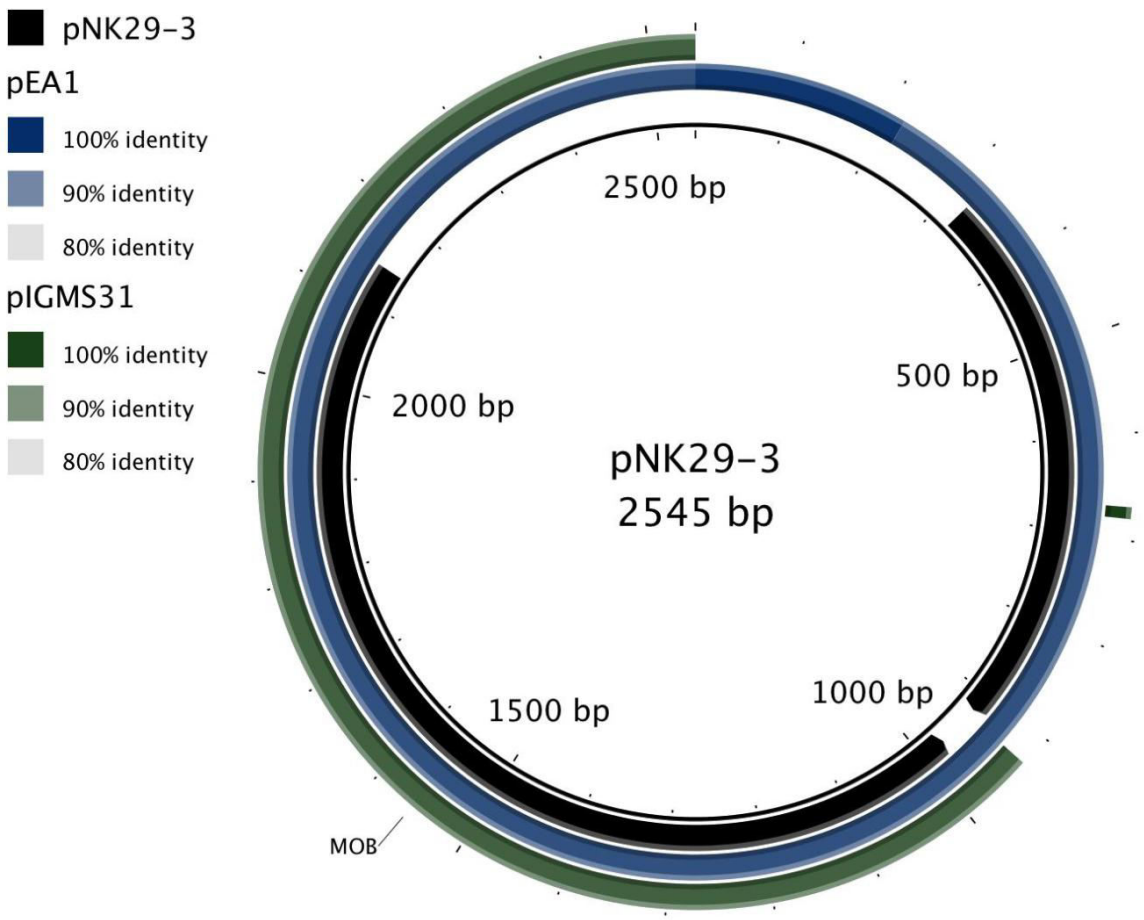


Figure S2. Plasmid pNK29-3 compared to most similar plasmids in GeneBank. The two ORFs encode putative replication and mobilization proteins and open reading frames are drawn directionally (inner ring, black).

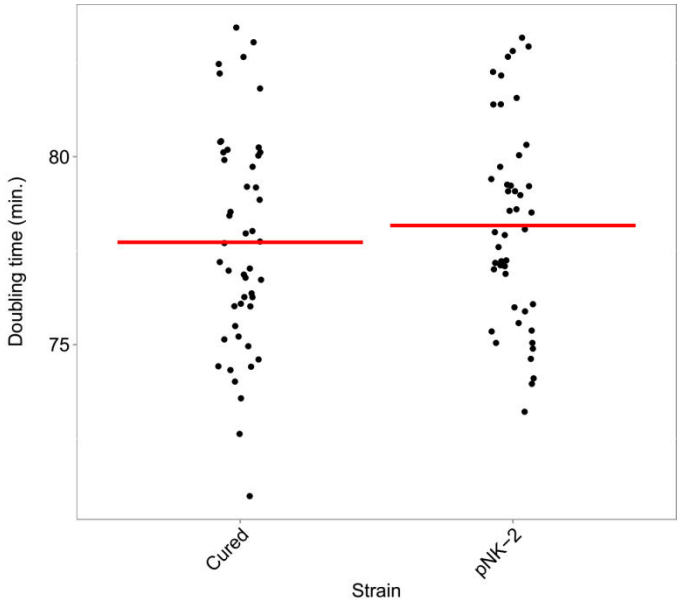


Figure S3. Growth rate of lineage A with and without pNK29-2 under iron-limited conditions.

Table S1. SNPs in different isolates of lineage A. Table containing the SNPs from lineage A, including the annotation and whether the animo acid change was synonymous or non-synonymous. The isolates from days 2, 9, and 16 are left out as they did not have any SNP differences from the representative genome of the lineage (isolated at 2d)**. Table continuous on next page!**

Contig no.	Position	Day 2	UTI	Day 32	2 m1	2 m2	6 m1	6 m2	12 m	Annotation	mutation
61	61701	A	G							6-phosphogluconate dehydrogenase, decarboxylating (EC 1.1.1.44)	nonsynonymous
56	6755	G		A	A	A	A	A	A	No coding sequence identified	
5	176310	G			A	A	A	A	A	Dihydrofolate reductase (EC 1.5.1.3)	synonymous
93	189	G			A	A	A			No coding sequence identified	
42	18614	G			A		A			Primosomal protein I	nonsynonymous
93	95	G			A		A			No coding sequence identified	
93	104	G			A		A			No coding sequence identified	
93	105	G			A		A			No coding sequence identified	
93	171	G			A		A			No coding sequence identified	
15	137066	G			A					Radical SAM family enzyme, similar to coproporphyrinogen III oxidase	synonymous
24	77511	G			A					Transcriptional activator of maltose regulon, MalT	nonsynonymous
49	14209	G			A					hypothetical protein	synonymous
4	85958	G			A					3-oxoacyl-acyl-carrier protein reductase (EC 1.1.1.100)	synonymous
62	46358	G			A					No coding sequence identified	
5	175792	T			C	C	C	C	C	Dihydrofolate reductase (EC 1.5.1.3)	nonsynonymous
20	275987	T			C		C			Glutamate synthase	nonsynonymous
20	297334	T			C					FUSARIC ACID RESISTANCE PROTEIN FUSB	nonsynonymous
21	19073	T			C					IS	nonsynonymous
24	25690	T			C					Nitrite reductase NAD(P)H small subunit (EC 1.7.1.4)	nonsynonymous
34	44455	A			G					LSU ribosomal protein L19p	synonymous
70	8108	A			G					Putative Ton-B dependent hemine receptor	nonsynonymous
62	60490	C			T	T	T	T	T	Alpha-2-macroglobulin	synonymous
93	150	C			T		T			No coding sequence identified	
10	126386	C			T					Ferredoxin-type protein NapG (periplasmic nitrate reductase)	nonsynonymous
14	3707	C			T					Putative GTP-binding protein	nonsynonymous
20	214976	C			T					LppC putative lipoprotein	nonsynonymous
23	43422	C			T					Putative phosphatase YieH	nonsynonymous
16	43695	G				A				Inner membrane protein YiaA	nonsynonymous
20	74596	G				A				Hexuronate utilization operon transcriptional repressor ExuR	nonsynonymous
42	9944	G				A				putative tail protein	nonsynonymous
46	5067	G				A				Putative stability/partitioning protein encoded within prophage CP-933T	synonymous
5	123037	G				A				Cell division protein FtsQ	nonsynonymous
17	98463	T				C				Holliday junction DNA helicase RuvB	synonymous
24	60742	T				C				Transcription elongation factor GreB	nonsynonymous
30	120639	G				C				expressed protein	nonsynonymous
10	92885	A				G				Putative membrane protein	nonsynonymous
26	193630	A				G				Putative inner membrane protein	synonymous
26	218585	A				G				No coding sequence identified	
30	97163	A				G				Glutamate Aspartate transport ATP-binding protein GltL (TC 3.A.1.3.4)	nonsynonymous
83	6188	A				G				FIG01045219: hypothetical protein	synonymous
26	171451	G					A			Inner membrane ABC transporter permease protein YcjP	stop
3	137385	T					C			FIG138056: a glutathione-dependent thiol reductase	synonymous
27	15149	C					T			IncF plasmid conjugative transfer pilus assembly protein TraU	nonsynonymous
51	40378	C					T			Chitinase (EC 3.2.1.14)	synonymous
10	196403	C						A	A	DedA family inner membrane protein YohD	nonsynonymous
15	10933	G						A	A	Protease III precursor (EC 3.4.24.55)	synonymous
16	34911	G						A	A	Outer membrane protein A precursor	nonsynonymous
20	182830	G						A	A	Inner membrane protein	synonymous
23	37891	G						A	A	6-phospho-beta-glucosidase (EC 3.2.1.86)	nonsynonymous
25	57550	G						A	A	ATP-dependent hsl protease ATP-binding subunit HslU	nonsynonymous
35	4250	G						A	A	O-antigen ligase	synonymous
39	14389	G						A	A	Ubiquinone biosynthesis monooxygenase UbiB	nonsynonymous
3	26626	G						A	A	Enoyl-CoA hydratase (EC 4.2.1.17) / 3-hydroxyacyl-CoA dehydrogenase (EC 1.	nonsynonymous
4	39270	G						A	A	Arginine/ornithine antiporter ArcD	nonsynonymous
4	66817	G						A	A	Uncharacterized protein YtfN	stop
5	5863	G						A	A	Prolyl-tRNA synthetase (EC 6.1.1.15), bacterial type	nonsynonymous
10	145962	G						A		Putative ABC transporter ATP-binding protein	nonsynonymous
15	147808	G						A		Ornithine decarboxylase (EC 4.1.1.17)	nonsynonymous
17	309028	T						A		FIG00638170: hypothetical protein	nonsynonymous
18	136370	G						A		Hydroxylamine reductase (EC 1.7.-.-)	synonymous
19	96733	G						A		Per-activated serine protease autotransporter enterotoxin EspC	nonsynonymous
1	53133	G						A		PTS system, sorbose-specific IIC component (EC 2.7.1.69)	nonsynonymous
4	95471	G						A		No coding sequence identified	
53	22400	G						A		Single-stranded DNA-binding protein	synonymous
5	191026	G						A		Carnitine operon protein CaiE	nonsynonymous
147	554	T						C	C	hypothetical protein	nonsynonymous
20	183063	T						C	C	Inner membrane protein	nonsynonymous
3	70340	T						C	C	No coding sequence identified	
3	122654	T						C	C	NADP-dependent malic enzyme (EC 1.1.1.40)	synonymous
43	100362	T						C	C	Flagellar hook-associated protein FlgK	nonsynonymous
49	16016	T						C	C	LysR family transcriptional regulator YafC	nonsynonymous
5	120353	T						C	C	Cell division protein FtsZ (EC 3.4.24.-)	nonsynonymous
5	170421	T						C	C	Outer membrane protein Imp, required for envelope biogenesis / Organic sc	nonsynonymous
62	31047	T						C	C	Uncharacterized protein YphG, TPR-domain containing	nonsynonymous
94	267	T						C	C	No coding sequence identified	
13	17707	T						C		Phage head-to-tail joining protein	nonsynonymous
25	1495	T						C		Outer membrane vitamin B12 receptor BtuB	synonymous
53	13735	T						C		Holin-like protein CidA	synonymous
10	178934	A						G	G	GTP cyclohydrolase I (EC 3.5.4.16) type 1	nonsynonymous
21	56232	A						G	G	adherence and invasion outermembrane protein (Inv,enhances Peyer's patc	nonsynonymous
83	198	A						G	G	No coding sequence identified	
35	2808	A						G		No coding sequence identified	
43	106731	A						G		Ribosomal large subunit pseudouridine synthase C (EC 4.2.1.70)	nonsynonymous
5	62832	A						G		glutamyl-Q-tRNA synthetase	synonymous
77	4289	A						G		hypothetical protein	nonsynonymous
147	436	C						T	T	No coding sequence identified	
17	79481	C						T	T	Flagellar biosynthesis protein FlhA	synonymous
17	216927	C						T	T	PTS system, chitobiose-specific IIC component (EC 2.7.1.69)	nonsynonymous
17	299242	C						T	T	No coding sequence identified	
30	65147	C						T	T	Ribonuclease I precursor (EC 3.1.27.6)	nonsynonymous
43	39245	C						T	T	No coding sequence identified	
17	232929	C						T		Threonyl-tRNA synthetase (EC 6.1.1.3)	synonymous
26	200370	C						T		No coding sequence identified	
3	94889	C						T		Sulfate and thiosulfate import ATP-binding protein CysA (EC 3.6.3.25)	synonymous
43	29182	C						T		No coding sequence identified	
5	225896	C						T		Aspartokinase (EC 2.7.2.4) / Homoserine dehydrogenase (EC 1.1.1.3)	synonymous
7	65266	C						T		ATP-dependent Clp protease proteolytic subunit (EC 3.4.21.92)	synonymous
10	14792	G							A	Phosphate acetyltransferase (EC 2.3.1.8)	synonymous
122	2192	G							A	Pyruvate formate-lyase (EC 2.3.1.54)	nonsynonymous
15	104208	G							A	Protein involved in stability of MscS mechanosensitive channel	nonsynonymous

15	115558	G							A	PTS system, mannitol-specific IIB component (EC 2.7.1.69) / PTS system, man	nonsynonymous
17	296387	G							A	Purine nucleotide synthesis repressor	synonymous
18	176760	G							A	Dipeptide transport system permease protein DppC (TC 3.A.1.5.2)	nonsynonymous
18	200735	T							A	Glutamate transport membrane-spanning protein	nonsynonymous
19	113564	G							A	Flagellar biosynthesis protein FlhA	synonymous
20	29027	G							A	Malate synthase G (EC 2.3.3.9)	synonymous
21	38804	G							A	iron aquisition yersiniabactin synthesis enzyme (Irp2)	nonsynonymous
23	7535	G							A	Ribose ABC transport system, ATP-binding protein RbsA (TC 3.A.1.2.1)	synonymous
23	14472	G							A	No coding sequence identified	
23	135025	G							A	3-deoxy-D-manno-octulosonic-acid transferase (EC 2.-.-.-)	nonsynonymous
24	31001	G							A	FIG006427: Putative transport system permease protein	synonymous
25	1435	G							A	Outer membrane vitamin B12 receptor BtuB	synonymous
26	254547	T							A	hypothetical protein	nonsynonymous
26	254550	T							A	hypothetical protein	nonsynonymous
30	39378	G							A	Enterobactin synthetase component F, serine activating enzyme (EC 2.7.7.-)	nonsynonymous
32	23323	G							A	No coding sequence identified	
32	70369	G							A	PhnJ protein	synonymous
42	3849	G							A	No coding sequence identified	nonsynonymous
51	23078	G							A	LSU ribosomal protein L4p (L1e)	synonymous
53	1415	G							A	Glutamate-aspartate carrier protein	synonymous
53	6083	G							A	NrfC protein	synonymous
5	25753	G							A	Intramembrane protease RasP/YluC, implicated in cell division based on FtsL	nonsynonymous
5	103289	G							A	VgrG protein	nonsynonymous
5	296526	G							A	Putative inner membrane protein	synonymous
7	15002	G							A	Glucokinase, ROK family (EC 2.7.1.2)	synonymous
7	35837	G							A	Nucleoside-specific channel-forming protein Tsx precursor	synonymous
10	41534	T							C	FIG00638014: hypothetical protein	synonymous
10	155429	T							C	Putative metal chaperone, involved in Zn homeostasis, GTPase of COG0523 f	nonsynonymous
15	25680	T							C	Lysophospholipid transporter LplT	nonsynonymous
15	59616	T							C	Predicted oxidoreductase, Fe-S subunit	nonsynonymous
15	91830	T							C	2-octaprenyl-3-methyl-6-methoxy-1,4-benzoquinol hydroxylase (EC 1.14.13.	synonymous
15	102370	T							C	No coding sequence identified	
17	149347	T							C	FIG00639292: hypothetical protein	nonsynonymous
17	257077	T							C	Acyl-CoA dehydrogenase (EC 1.3.99.3)	nonsynonymous
19	125833	T							C	FIG00641604: hypothetical protein	synonymous
20	29055	T							C	Malate synthase G (EC 2.3.3.9)	nonsynonymous
20	166907	T							C	Integral membrane protein TerC	synonymous
20	260128	T							C	Uncharacterized ABC transporter, ATP-binding protein YrbF	nonsynonymous
21	13870	T							C	hypothetical protein	synonymous
24	56129	T							C	FIG00638818: hypothetical protein	nonsynonymous
24	106762	T							C	Aspartate-semialdehyde dehydrogenase (EC 1.2.1.11)	nonsynonymous
26	131738	T							C	DNA topoisomerase I (EC 5.99.1.2)	nonsynonymous
27	43829	T							C	No coding sequence identified	
30	47089	T							C	Isochorismate synthase (EC 5.4.4.2) of siderophore biosynthesis	nonsynonymous
31	39231	T							C	Branched-chain amino acid ABC transporter, amino acid-binding protein (TC	nonsynonymous
32	37751	T							C	Lysyl-tRNA synthetase (class II) (EC 6.1.1.6)	synonymous
39	73336	T							C	Arylsulfatase (EC 3.1.6.1)	nonsynonymous
39	109139	T							C	Protein yifE	synonymous
4	23661	T							C	FIG000988: Predicted permease	nonsynonymous
56	21039	T							C	ATPase provides energy for both assembly of type IV secretion complex and	nonsynonymous
5	61314	T							C	Sugar/maltose fermentation stimulation protein homolog	synonymous
5	85250	T							C	Blue copper oxidase CueO precursor	synonymous
5	127065	T							C	Cell division protein FtsW	nonsynonymous
5	278788	T							C	TRAP dicarboxylate transporter, DctM subunit, unknown substrate 8	nonsynonymous
5	324641	T							C	Phosphoenolpyruvate-dihydroxyacetone phosphotransferase (EC 2.7.1.121),	nonsynonymous
61	31544	T							C	Propanediol dehydratase reactivation factor large subunit	nonsynonymous
61	51403	T							C	No coding sequence identified	
7	53127	T							C	FIG01057005: hypothetical protein	nonsynonymous
10	106855	A							G	Sensory histidine kinase AtoS	nonsynonymous
11	4616	A							G	Single-stranded DNA-binding protein	nonsynonymous
14	17663	A							G	GTP-binding protein TypA/BipA	nonsynonymous
15	15860	A							G	FIG004819: Prepilin peptidase dependent protein B precursor	nonsynonymous
15	34683	A							G	No coding sequence identified	
16	65330	A							G	L-xylulose/3-keto-L-gulonate kinase (EC 2.7.1.-)	nonsynonymous
17	102945	A							G	Cell wall endopeptidase, family M23/M37	nonsynonymous
17	117869	A							G	Putative amidohydrolase	synonymous
17	143930	A							G	Rtn protein	synonymous
17	164143	A							G	Putative uncharacterized protein YeaK	nonsynonymous
17	240500	A							G	Vitamin B12 ABC transporter, permease component BtuC	synonymous
18	227304	A							G	Putative membrane protein	synonymous
19	74778	A							G	oxidoreductase, aldo/keto reductase family	nonsynonymous
20	82805	A							G	Putative cell division protein precursor	synonymous
20	161873	A							G	Putative cytoplasmic protein	synonymous
24	48644	A							G	Multimodular transpeptidase-transglycosylase (EC 2.4.1.129) (EC 3.4.-.-)	synonymous
26	254549	A							G	hypothetical protein	synonymous
28	32170	A							G	BarA sensory histidine kinase (VarS GacS)	nonsynonymous
28	53920	A							G	Putative electron transfer flavoprotein subunit YgcQ	synonymous
28	87481	A							G	Formate hydrogenlyase transcriptional activator	nonsynonymous
30	59369	A							G	LysR-family transcriptional regulator YbeF	nonsynonymous
30	134664	A							G	SeqA protein, negative modulator of initiation of replication	nonsynonymous
31	21929	A							G	Nickel ABC transporter, periplasmic nickel-binding protein NikA (TC 3.A.1.5.3	nonsynonymous
32	75262	A							G	FIG00638130: hypothetical protein	nonsynonymous
37	18162	A							G	Protein yhjK	nonsynonymous
39	100861	A							G	Threonine dehydratase biosynthetic (EC 4.3.1.19)	nonsynonymous
3	158670	A							G	Exopolyphosphatase (EC 3.6.1.11)	nonsynonymous
43	66383	A							G	Major curlin subunit precursor CsgA	synonymous
4	86155	A							G	3-oxoacyl-acyl-carrier protein reductase (EC 1.1.1.100)	nonsynonymous
55	5078	A							G	Respiratory nitrate reductase beta chain (EC 1.7.99.4)	nonsynonymous
5	230361	A							G	Inner membrane protein CreD	nonsynonymous
62	46866	A							G	Iron-sulfur cluster regulator IscR	synonymous
62	59308	A							G	Alpha-2-macroglobulin	synonymous
7	2996	A							G	putative lipoprotein	nonsynonymous
7	55881	A							G	Cytochrome O ubiquinol oxidase subunit IV (EC 1.10.3.-)	nonsynonymous
9	29805	A							G	Protein ydgH precursor	synonymous
9	31986	A							G	NAD(P) transhydrogenase alpha subunit (EC 1.6.1.2)	synonymous
9	37980	C							G	Permeases of the major facilitator superfamily	nonsynonymous
10	192408	C							T	tRNA-dihydrouridine synthase C	nonsynonymous
10	228596	C							T	Scaffold protein for 4Fe-4S cluster assembly ApbC, MRP-like	synonymous
10	232712	C							T	Fimbriae usher protein StcC	nonsynonymous
10	300700	C							T	Mannose-1-phosphate guanylyltransferase (GDP) (EC 2.7.7.22) / Mannose-6-	nonsynonymous

Table S2. Annotations for pNK29. Names in square brackets denote homologs in pOLA52.

ORF No.	Start	Stop	Strand	Annotation
1	105	650	+	DNA distortion protein 1 [taxA]
2	653	1819	+	IncQ plasmid conjugative transfer DNA nicking endonuclease TraR (pTi VirD2 homolog) [taxC]
3	2172	2690	+	actX, homologous to actX on pOLA52
4	2620	2832	+	hypothetical protein
5	2865	3512	+	Peptidoglycan hydrolase VirB1, involved in T-DNA transfer [pilx1]
6	3490	3786	+	Major pilus subunit of type IV secretion complex, VirB2 [pilx2]
7	3808	6561	+	ATPase provides energy for both assembly of type IV secretion complex and secretion of T-DNA complex (VirB4) [pilx3-4]
8	6572	7330	+	Minor pilin of type IV secretion complex (VirB5) [pilx5]
9	7331	7618	+	IncQ plasmid conjugative transfer protein TraG
10	7627	8757	+	Integral inner membrane protein of type IV secretion complex (VirB6) [pilx6]
11	8885	9013	+	pilx7, homologous to pilx7 on pOLA52
12	9003	9716	+	Inner membrane protein forms channel for type IV secretion of T-DNA complex, VirB8 [pilx8]
13	9721	10650	+	Forms the bulk of type IV secretion complex that spans outer membrane and periplasm (VirB9) [pilx9]
14	10647	11852	+	Inner membrane protein of type IV secretion of T-DNA complex, TonB-like, VirB10 [pilx10]
15	11854	12885	+	ATPase provides energy for both assembly of type IV secretion complex and secretion of T-DNA complex (VirB11) [pilx11]
16	12888	14723	+	Type IV secretion system protein VirD4 [taxB]
17	14720	15133	+	hypothetical lipoprotein, homologous to pOLA52
18	15130	15432	+	IncN plasmid KikA protein
19	15533	16858	+	Phage tail fiber protein
20	16887	17114	+	hypothetical protein
21	17194	17712	+	putative nuclease
22	17794	19056	+	Cell division protein FtsH (EC 3.4.24.-)
23	19060	19626	+	FIG01045518: hypothetical protein
24	19640	21793	+	DNA topoisomerase III (EC 5.99.1.2)
25	21790	21996	+	Haemolysin expression modulating protein
26	22012	22473	+	DNA-binding protein H-NS [hns]
27	22524	22787	+	hypothetical protein
28	22898	25699	-	Mobile element protein, homologous to tnpA of Tn3 pRPEC180_47
29	25670	25921	+	hypothetical protein
30	26067	26624	+	Mobile element protein, homologous to tnpR of pRPEC180_47 (transposon Tn3 resolvase)
31	26807	27667	+	Beta-lactamase (EC 3.5.2.6)
32	27814	27948	-	hypothetical protein
33	28012	28128	+	hypothetical protein
34	28432	30354	-	L. lactis predicted coding region ORF00041
35	30398	30754	+	DNA distortion protein 3
36	30751	31728	-	hypothetical protein
37	31753	32034	-	hypothetical protein
38	32132	32794	-	Chromosome partitioning protein ParA [par]
39	33175	33819	+	Resolvase [res]
40	33935	34114	-	hypothetical protein
41	34145	34264	+	hypothetical protein
42	34687	34803	-	hypothetical protein
43	35175	35429	+	hypothetical protein
44	36056	36724	+	L. lactis predicted coding region ORF00041
45	36758	37204	-	bis, homologous to pOLA52 (FIG01048616: hypothetical protein)
46	37244	38080	-	pir, homologous to pOLA52
47	37673	38002	-	YagA protein
48	38095	38310	-	FIG01048886: hypothetical protein
49	38300	38545	-	FIG01047979: hypothetical protein
50	38590	38913	-	FIG01047054: hypothetical protein
51	39059	39340	-	RelE/StbE replicon stabilization toxin
52	39330	39581	-	RelB/StbD replicon stabilization protein (antitoxin to RelE/StbE)
53	39578	39751	-	hypothetical protein
54	40368	40499	-	hypothetical protein
55	40816	41652	+	PI protein
56	41692	42138	+	FIG01047678: hypothetical protein

Table S3. SNPs in the pNK29-2 plasmid of lineage A compared to pUTI89. SNPs were identified by comparing the initial linage A isolate (day 2) with pUTI89 (CP000244.1).

Position	pUTI89	pNK29-2	Original amino acid	Amino acid change	annotation
17995	A	G	G		hypothetical protein
51718	G	T	D	Y	rsvB
53060	C	T			
68939	G	A	Q		hypothetical protein
69562	C	T	Q		hypothetical protein
80619	G	A	G	D	traE
91269	T	G	L		trbC

Table S4. Plasmids similar to pNK29-2 and information on their origin. The table presented in Figure 3, but more detailed information on size, location, and references. The GenBank ID refers to the plasmid sequence entry in GenBank and the Pubmed ID (PMID) to the study where the plasmid was initially described.

Name	Size (bp)	GenBank ID	Host	ST	Study (PMID)	Coverage	ID %	SNPs	Disease	Location	Year
pNK29-2	114,230		E. coli Ec29	ST420	*This study*	100	100	0	UTI	Sweden	1999
pUTI89	114,230	*CP000244.1*	E. coli UTI89	ST95	*16585510*	100	99	7	UTI	US	2001
RS218	114,231	*CP007150.1*	E. coli RS218	ST95	*25164788*	100	99	13	neonatal meningitis	US	1974
pEC14	114,222	*GQ398086.1*	E. coli EcC14	*NA*	*19887083*	100	99	50	UTI	US (Illinois)	2009
pUM146	114,550	*CP002168.1*	E. coli UM146	ST643	*21075930*	100	99	216	ileal Crohn's disease	Canada	2010
pZH063-1	114,223	*CP014523.1*	E. coli strain ZH063	ST131	*27390780*	100	99	11	ExPEC	Canada (Winnipeg)	2002
pSF-166-1	114,221	*CP012634.1*	E. coli SF-166	ST95	*26543109*	100	99	15	Bloodstream Infection	US (San Francisco)	2008
pSaT040	114,223	*CP014496.1*	E. coli SaT040	ST131	*27390780*	100	99	16	ExPEC	US (Burlington)	2007
pECO-bc6	101,201	*CP014668.1*	E. coli strain ECONIH2	ST127	*27353756*	88	99	29	ExPEC	US	2014
pMVAST0167	128,305	*CP014493.1*	E. coli SaT0167	ST131	*27390780*	97	99	34	ExPEC	US (Minneapolis)	2010
pECO-fce	212,180	*CP015160.1*	E. coli Eco889	ST313	*Unpublished*	97	99	283	Unknown	US	2014
pECSF1	122,345	*AP009379.1*	E. coli SE15	ST131	*20008064*	99	99	15	ExPEC	Japan	2010
pkPN-7c3	142,858	*CP015131.1*	K. pneumoniae Kpn555	*NA*	*Unpublished*	97	99	29	Unknown	US	2014
p1ESCUM	122,301	*CU928148.1*	E. coli UMN026	ST643	*23785449*	97	99	4	UTI	Sweden	2010

Manuscript II [redacted]

frontiers in Microbiology

ORIGINAL RESEARCH
published: 26 September 2017
doi: 10.3389/fmicb.2017.01852

Transfer and Persistence of a Multi-Drug Resistance Plasmid *in situ* of the Infant Gut Microbiota in the Absence of Antibiotic Treatment

Heidi Gumpert[1, 2†], Jessica Z. Kubicek-Sutherland[3†], Andreas Porse[4†], Nahid Karami[5†], Christian Munck[4], Marius Linkevicius[3], Ingegerd Adlerberth[5], Agnes E. Wold[5], Dan I. Andersson[3] and Morten O. A. Sommer[4]*

[1] Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark, [2] Department of Clinical Microbiology, Hvidovre Hospital, University of Copenhagen, Hvidovre, Denmark, [3] Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden, [4] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark, [5] Department of Infectious Diseases, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

OPEN ACCESS

Edited by:
Feng Gao,
Tianjin University, China

Reviewed by:
Swaine Chen,
Genome Institute of Singapore, Singapore
Christopher Morton Thomas,
University of Birmingham, United Kingdom

***Correspondence:**
Morten O. A. Sommer
msom@bio.dtu.dk

†These authors have contributed equally to this work.

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *01 August 2017*
Accepted: *11 September 2017*
Published: *26 September 2017*

Citation:
Gumpert H, Kubicek-Sutherland JZ, Porse A, Karami N, Munck C, Linkevicius M, Adlerberth I, Wold AE, Andersson DI and Sommer MOA (2017) Transfer and Persistence of a Multi-Drug Resistance Plasmid in situ of the Infant Gut Microbiota in the Absence of Antibiotic Treatment. Front. Microbiol. 8:1852. doi: 10.3389/fmicb.2017.01852

The microbial ecosystem residing in the human gut is believed to play an important role in horizontal exchange of virulence and antibiotic resistance genes that threatens human health. While the diversity of gut-microorganisms and their genetic content has been studied extensively, high-resolution insight into the plasticity, and selective forces shaping individual genomes is scarce. In a longitudinal study, we followed the dynamics of co-existing *Escherichia coli* lineages in an infant not receiving antibiotics. Using whole genome sequencing, we observed large genomic deletions, bacteriophage infections, as well as the loss and acquisition of plasmids in these lineages during their colonization of the human gut. In particular, we captured the exchange of multidrug resistance genes, and identified a clinically relevant conjugative plasmid mediating the transfer. This resistant transconjugant lineage was maintained for months, demonstrating that antibiotic resistance genes can disseminate and persist in the gut microbiome; even in absence of antibiotic selection. Furthermore, through *in vivo* competition assays, we suggest that the resistant transconjugant can persist through a fitness advantage in the mouse gut in spite of a fitness cost *in vitro*. Our findings highlight the dynamic nature of the human gut microbiota and provide the first genomic description of antibiotic resistance gene transfer between bacteria in the unperturbed human gut. These results exemplify that conjugative plasmids, harboring resistance determinants, can transfer and persists in the gut in the absence of antibiotic treatment.

Keywords: *Escherichia coli*, horizontal gene transfer, infant gut, genome dynamics, plasmid transfer, *in vivo* fitness, mouse models, antibiotic resistance

INTRODUCTION

The evolution of multidrug resistant bacteria through horizontal gene transfer (HGT) is resulting in human pathogens that are no longer amenable to antibiotic therapy (Davies and Davies, 2010). It is believed that antibiotic resistance genes are frequently exchanged between bacteria within the human microbiome, where the intestinal bacterial community in particular is considered a hub

for HGT (Liu et al., 2012). Transfer of antibiotic resistance genes within the gut microbiota is believed to happen primarily via conjugative plasmids and has been demonstrated to occur in both animals (McConnell et al., 1991; Schjørring et al., 2008), and humans (Lester et al., 2006; Trobos et al., 2009). Due to the low transfer frequency, and initial instability of plasmids in the absence of selection, previous studies have utilized experimental set-ups where the host was inoculated with a high number of bacteria, with subsequent monitoring to detect if the antibiotic resistance genes had been transferred from the donor strain (McConnell et al., 1991; Lester et al., 2006; Schjørring et al., 2008; Trobos et al., 2009).

We and others have documented the transfer of antibiotic resistance genes amongst naturally occurring bacteria in the human gut microbiota, and these reports describe changes in the antibiotic resistance profiles of strains collected from patients undergoing antibiotic treatment (Bidet et al., 2005; Karami et al., 2007; Conlan et al., 2014, 2016; Porse et al., 2017). Additionally, a retrospective study examining *Bacteroides* isolates, collected over a period of 40 years, demonstrated that extensive resistance gene exchange occurred between species of *Bacteroides* and other genera in the human colon (Shoemaker et al., 2001). Yet, while the unperturbed gut microbiome has been the subject of numerous metagenomic studies (Balzola et al., 2010; Huttenhower et al., 2012; Forslund et al., 2013), including the construction of complete genomes of various species and strains from metagenomic data (Sharon et al., 2013), the use of metagenomics is not well-suited to detect HGT events due to difficulties in associating mobile genetic elements with individual genomes.

To investigate the dynamics of horizontal gene exchange between *Escherichia coli* of the unperturbed gut microbiota, we use whole genome sequencing to characterize co-existing *E. coli* lineages isolated over the first year of an infant's life. Observing the transfer and enrichment of a conjugative antibiotic resistance plasmid, along with subsequent genomic events, in the absence of antibiotic treatment, we performed *in vivo* fitness assays indicating that this resistance plasmid is maintained in a gut environment despite being costly *in vitro*.

MATERIALS AND METHODS

Strain Isolation and Population Counts

Fecal samples were obtained from an infant enrolled in the ALLERGYFLORA study (Nowrouzian et al., 2003). A sample of the rectal flora was obtained using a cotton-tipped swab at 3 days after birth. The infant's parents collected fecal samples at 1, 2, and 4 weeks, and 2, 6, and 12 months of age. Samples were plated on Drigalski agar plates for the isolation of *Enterobacteriaceae* with a detection limit of $10^{2.5}$ CFU/g fecal matter. Each morphotype was enumerated separately, and strain identities of the enumerated morphotypes were confirmed using random amplified polymorphic DNA (RAPD) typing (Nowrouzian et al., 2003). Initial confirmation of the RAPD-typing was confirmed by pulsed-field gel electrophoresis (PFGE). Isolated strains were subjected to complete serotyping (O:K:H) (Statens Serum Institute, Copenhagen, Denmark).

From the 5 sampling times positive for *E. coli*, a total of 13 isolates were selected and stored for further analysis.

Antibiotic Susceptibility and Minimum Inhibitory Concentration (MIC) Determination

All isolates were tested for their susceptibility to the following antibiotics using the disc diffusion method (Oxoid, Sweden): ampicillin, amoxicillin/clavulanic acid, piperacillin, mecillinam, cefadroxil, ceftazidime, cefuroxime, cefoxitin, chloramphenicol, gentamicin, tobramycin, streptomycin, nitrofurantoin, nalidixic acid, tetracycline, trimethoprim, and sulphonamide. From the saved isolates, the exact MICs of one isolate per lineage per sampling point were determined using the broth dilution method (**Table S1**; Wiegand et al., 2008).

Genome Sequencing

Genomic DNA from each of the 13 isolates was obtained using the UltraClean® Microbial DNA Isolation Kit (Mobio Laboratories, Inc.). Sequencing was performed by Partners HealthCare Center for Personalized Genetic Medicine (Massachusetts, USA) or at the Novo Nordisk Foundation Centre for Biosustainability (Lyngby, Denmark).

Sequence Analysis

Reads from each isolate were assembled using Velvet (Zerbino and Birney, 2008). Contigs with <500 bp were filtered and corrected by aligning reads using Bowtie2 (version 2.1.0) (Langmead et al., 2009). Single-nucleotide polymorphisms (SNPs) were called using SAMtools (version 0.1.19) (Li et al., 2009), and edited using custom biopython scripts (Cock et al., 2009). Contigs were annotated using the RAST server (Aziz et al., 2008). SAMtools were also used to determine the number of SNPs between isolates, where identified variants had a phred quality score of at least 50 and >90% of the high-quality reads as the variant. The assemblies from the following isolates where used as references for SNP-calling: lineage A $2w_2$, lineage B $2m_2$ and lineage C $12m_2$. SNPs occurring in short homologous regions after genomic deletions or acquisitions were also filtered. BEDtools (Quinlan and Hall, 2010) was used to calculate read coverage across genomes and thus identify acquired or deleted genomic information. MUMmer was used to align sequences (Khan et al., 2009). Multi-locus sequence type (MLST) groups were determined using the database hosted at http://mlst.warwick.ac.uk/mlst/dbs/Ecoli (Wirth et al., 2006).

Phage Identification

The PHAST phage search tool server (Zhou et al., 2011) was used to identify possible intact phages in the contigs. In addition, BLAST was used to identify similar previously described phages. Phage integration sites were determined by aligning contigs containing the flanking regions of the phage to an earlier isolate not containing the prophage.

Plasmid Analysis

The PlasmidFinder web-service (http://cge.cbs.dtu.dk/services/PlasmidFinder) was used to identify replicons in the assembled

contigs and classify plasmids into incompatibility groups (Carattoli et al., 2014). Plasmid diagrams depicting read coverage were drawn in R via custom scripts, and plasmid ring diagrams were drawn using BLAST Ring Image Generator (BRIG; Alikhan et al., 2011) with the "-task megablast" option to BLAST. Additionally, contigs belonging to plasmids (that had a copy number greater than one) were identified based on their relative abundance to the genome via BEDTools (Quinlan and Hall, 2010).

Genomic Deletion Verification by PCR

Based on the alignment of contigs to the genome of CFT 073 (NC_004431), flanking primers were designed to show that the deletion in lineage A was a chromosomal excision. In addition to show contiguity prior to the deletion, controls were included to show the occurrence of the deletion only in the lineage A isolate sampled at 6 months.

In Vitro Conjugation Assay

To test the ability of Lineage B to transfer the pHUSEC41-1-like plasmid to the plasmid free Lineage A, outgrown overnight cultures of two lineages were mixed equally and incubated for 12 h. Incubations were done at 37°C on a solid agar surface as well as in liquid cultures without shaking. Mating cultures were plated on LB containing chloramphenicol and ampicillin to select for transconjugants.

In Vitro Competition Experiments to Assess the Fitness Costs of the pHUSEC41-1-Resembling Plasmid

To assess fitness cost, pairwise growth competition experiments in Davis minimal medium with 25 mg/mL glucose (DM25) were performed using isolates of lineage A sampled at 2 weeks and 2 months, respectively, the latter which had acquired the plasmid closely resembling pHUSEC41-1 (Künne et al., 2012). The experiment was performed as previously described (Enne et al., 2005), but in brief, the two isolates were grown overnight in nutrient broth, and then inoculated into DM25 at a dilution of $1{:}10^4$ and grown for 24 h. The cultures were then mixed together in a ratio of 1:1, and then diluted 1:100 into fresh DM25. The serial passage step was continued for 6 days, corresponding to ∼60 generations of competition. After initially mixing the two cultures together, and after each 24 h period, the cultures were diluted appropriately and 100 μL were added to Iso-Sensitest plates (Oxoid, Sweden) in triplicate, with and without 50 mg/L of ampicillin. Colonies were counted after over-night incubation at 37°C, where the mean number of colonies on ampicillin plates was subtracted from the plates without ampicillin to determine the mean number of colonies lacking the pHUSEC41-1-like plasmids. Six replicates of the fitness experiment were conducted.

In Vivo Competitive Fitness Assays

Isolates used in the competitive fitness studies were tagged with chloramphenicol (Cam^R) and kanamycin (Kan^R) resistance markers, *cat* and *aph(3′)-II* genes, respectively, amplified from cloning vectors of the pZ vector system (Lutz and Bujard, 1997): lineage A $2w_1$—Cam^R, 2m—Kan^R, $6m_1$—Cam^R, and lineage C at $12m_1$—Kan^R. The markers were inserted into the chromosomal *araB* gene of the lineage A and B strains using the Lambda Red recombineering system of pTKRED (Kuhlman and Cox, 2010). The following regions of homology were used for insertions into *araB*: 5′- GTAGCGAGGTTAAGATCGGTAATCACCCCTTTC AGGCGTTGGTTAGCGTT-3′ and 5′-GCCTAACGCACTGGT AAAAGTTATCGGTACTTCCACCTGCGACATTCTGA-3′.

Previous studies have shown that the inactivation of *araB* is fitness neutral in a murine model and that the Cam^R and Kan^R markers do not significantly affect the growth of *E. coli* (Chen et al., 2013; Linkevicius et al., 2016).

Female BALB/c mice (5–6 weeks old) were used in all *in vivo* experiments (Charles River Laboratories, distributed by Scanbur). All mice were pre-treated orally with streptomycin as described previously (Lasaro et al., 2014). Briefly streptomycin sulfate salt (Sigma-Aldrich) was added to the drinking water at 5 g/L, along with 5 g/L of glucose to enhance taste, for 72 h followed by 24 h of fresh water (no drug or glucose) to allow the streptomycin to be cleared from the animal's system prior to inoculation. No streptomycin was administered during the course of infection. Ten mice were administered 100 μL containing a 1:1 *E. coli* mixture by oral gavage of the examined strains. Feces were homogenized in PBS, serially diluted, and equal amounts were plated on LA-Cam (25 μg/ml chloramphenicol, selecting for the chromosomal marker), LA-Kan (50 μg/ml kanamycin, selecting for the chromosomal marker) and either LA-Kan, Amp or LA-Cam, Amp (50 μg/ml kanamycin or 25 μg/ml chloramphenicol, and 100 μg/ml ampicillin selecting for the pHUSEC41-1 plasmid) to determine the number of viable bacterial cells as well as the fraction containing the pHUSEC41-1 plasmid. CFU values were normalized per gram of tissue (CFU/g). The plasmid was maintained stably in all competitions and conjugational transfer between competing strains was assessed through replica-plating. The competitive index was calculated by dividing the output on days 2, 4, and 7 by the input on day 0.

Ethics Statement

Animal experiments were performed in accordance with national (regulation SJVFS 2012:26) and institutional guidelines. The Uppsala Animal Experiments Ethics Review Board in Uppsala, Sweden approved all mouse protocols undertaken in this study under reference no. 154/14. Animal experiments were performed at the Swedish National Veterinary Institute (SVA) in Uppsala, Sweden.

RESULTS AND DISCUSSION

Study Material

Our study material was selected from an infant enrolled in the ALLERGYFLORA study, which was designed to examine the link between the infant gut microbial colonization pattern, over the first year of life, and the development of allergies (Nowrouzian et al., 2003). Fecal samples were cultured for *E. coli* and various colony types were assigned to specific lineages via random amplified polymorphic DNA and enumerated separately (RAPD; **Figure 1**). Sampling at 3 days and 1 week

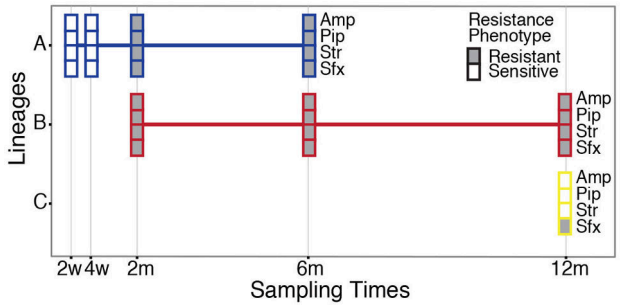


FIGURE 1 | Sampling and antibiotic resistance profile of the *E. coli* isolates. A total of three *E. coli* lineages **(A–C)** were sampled from the infant's intestinal microbiota over the first year of life. Boxes indicate both the presence of the lineages and their antibiotic resistance profile to ampicillin (Amp), piperacillin (Pip), streptomycin (Str), and sulfamethoxazole (Sfx) at the sampling points. Filled, boxes indicate resistant isolates, and empty boxes indicate sensitive isolates (see **Table S1** for MIC values).

yielded no *E. coli* isolates. Between 2 weeks and 12 months a total of three distinct lineages were identified: A, B and C (**Figure 1**). The sampling at 2 and 4 weeks after birth yielded only colonies belonging to lineage A, which were sensitive to all antibiotics tested (**Table S1**). At the 2 month sampling time, lineage B appeared and resistance to the antibiotics: ampicillin, piperacillin, streptomycin, and sulfamethoxazole was measured. At this sampling time, the antibiotic resistance profile of lineage A changed, and subsequent isolates were now resistant to ampicillin, piperacillin, streptomycin, and sulfamethoxazole, matching the resistance profile of lineage B. lineages A and B were both present at the 6 months sampling time with no changes in the antibiotic resistance profile. At the 12 month sampling time, only lineage B remained, with the addition of lineage C, which was resistant to sulfamethoxazole. From plate-count estimations, we observed a consistent decrease in population numbers of *E. coli* in the gut of the infant over the first year of life (**Figure 2A**). This is in line with the other infants enrolled in the ALLERGYFLORA study and in parallel with the establishment of a microbiota dominated by anaerobic bacteria (Nowrouzian et al., 2003; Palmer et al., 2007).

Genomic Relationship of the Lineages Isolated from the Gut

A total of 13 isolates from lineages A, B, and C were genome sequenced with at least one isolate sequenced per lineage per sampling point. Lineage A included two isolates from the 2 week sampling time ($2w_1$ and $2w_2$), one from 4 weeks (4w) and 2 months (2m) and two from 6 months ($6m_1$ and $6m_2$) lineage B included two isolates form 2 months ($2m_1$ and $2m_2$) one from 6 months (6m) and two from 12 months ($12m_1$ and $12m_2$) and lineage C isolates included two from 12 months ($12m_1$ and $12m_2$). To confirm lineage identities of the isolates, we assessed both the number of SNPs and the amount of total genomic content shared between lineages by comparing to the first isolate sampled from each lineage.

Both lineages A and C had ∼90,000 single nucleotide differences when compared to lineage B (**Table S2**). Interestingly, lineages A and C were less different with an order of magnitude fewer SNP when compared to each other; having ∼7,000 SNPs. Similarly, when comparing the percentage of the genomic content shared between the lineages, lineages A and C shared between 79.7 and 82.3% in common with lineage B, whereas lineage A and C shared at least 93.6% of the genomic content (**Table S3**). While these results indicate that lineages A and C are more closely related to each other than to lineage B, the number of SNPs and the differences in genomic content reveal that they are different lineages. While RAPD-typing of the isolates was sensitive enough to successfully classify the isolates into the three distinct lineages, MLST typing assigned both lineage A and C isolates to ST12, whereas the lineage B isolates belonged to ST782.

Evolutionary relationships amongst the isolates within each lineage were established based on the SNPs identified by aligning reads to an isolate from the first time point the lineage was sampled (**Table S4**). SNPs identified in isolates from lineages A, B, and C produced consistent phylogenetic relations that show a progression in the acquisition of SNPs; indicating that the samples were representative clones of the lineages (**Figure 2B**).

Multiple Antibiotic Resistance Plasmid Transfer *in situ* of the Gut in the Absence of Antibiotic Pressure

To identify the genomic changes underlying the acquisition of antibiotic resistance in lineage A, sequence data collected from the sensitive isolates ($2w_1$, $2w_2$, and 4w) were compared to sequence data from the resistant isolates (2m, $6m_1$, and $6m_2$). Two non-conservative genomic mutations in the betaine aldehyde dehydrogenase (*betB*) and phosphoenolpyruvate carboxylase (*pckA*) genes were identified; however, these mutations would not be expected to contribute to antibiotic resistance. Instead, additional genetic information, totaling 90 kb, was found in the resistant lineage A isolates compared to sensitive lineage A isolates (**Figure 1**). The newly acquired genetic information had a read coverage two times greater than the chromosome, and included conjugative transfer genes; suggesting a newly acquired plasmid with ∼2 copies per chromosome. Additionally, the following resistance genes were identified: the β-lactamase $bla_{\mathrm{TEM-1c}}$, an aminoglycoside 3′-phosphotransferase (*strA*), and streptomycin phosphotransferase (*strB*), as well as the dihydropteroate synthase gene (*sul2*), conferring resistance to sulfonamides.

The phenotypic resistance patterns (**Figure 1**) suggested that the horizontally acquired resistance was transferred from lineage B to lineage A. Aligning reads from lineage B to the newly acquired plasmid in lineage A resulted in 100% identity with only one identified SNP variant. Although we cannot out rule that the plasmid was already present in lineage A, or transferred from other constituents of the microbiota, the high degree of identity between the plasmids, the co-appearance of lineage B and a matching resistance profile is consistent with lineage B transferring its antibiotic resistance plasmid to lineage A.

Querying sequence databases yielded the clinically important conjugative, IncI1-type pHUSEC41-1 plasmid of 91,942 bp (Grad et al., 2013). Contigs from the isolates in this study aligned

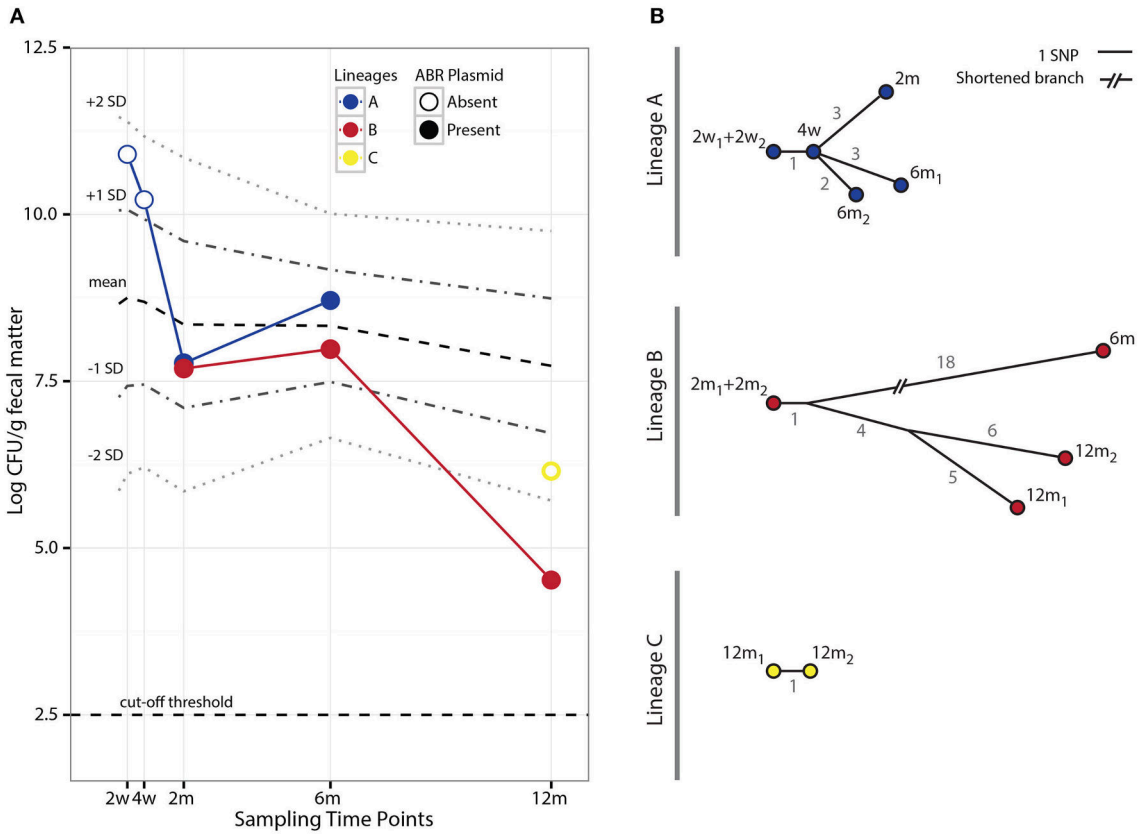


FIGURE 2 | Population counts and SNP evolution of co-existing *E. coli* lineages. **(A)** Fecal population counts of *E. coli* lineages A, B, and C at different sampling points during the first year of life of the infant studied. Filled circles indicate the presence of the pHUSEC41-1-like antibiotic resistance plasmid. For comparison, the mean population levels and ±1 and 2 standard deviations (*SD*) at the same sampling points for 272 *E. coli* strains isolated from 128 infants in the ALLERGYFLORA cohort are indicated in the figure. **(B)** Phylogenetic trees based on the number of SNPs found in each of the isolates of lineages A and B. The gray values next to each branch indicates the number of SNPs between isolates.

to pHUSEC41-1 resulted in 99.3% coverage of the plasmid with an average of 99.0% identity (**Figure 3**). The alignment also showed that there were no insertions in the transferred plasmid compared to pHUSEC41-1. The pHUSEC41-1 plasmid was initially identified in the *E. coli* serotype O104:H4 strain HUSEC41 isolated from a child in Germany with hemolytic-uremic syndrome (HUS; Künne et al., 2012). This plasmid has additionally been found in other sequenced *E. coli* isolates of different serotypes isolated from patients in France and Finland (Grad et al., 2013); highlighting the wide dissemination of this multiple antibiotic resistance plasmid amongst geographically dispersed *E. coli* strains.

The pHUSEC41-1-Like Resistance Plasmid Is Costly *in Vitro* but Beneficial *in Vivo* of the Mouse Gut

Interestingly, the acquisition of the pHUSEC41-1-like resistance plasmid by lineage A was associated with an initial steep drop in population counts, from $10^{10.2}$ CFU/g of fecal matter in the 4 week sample to $10^{7.8}$ CFU/g in the 2 month sample (**Figure 2A**). To determine whether this decrease related to a fitness cost imposed on lineage A from carrying the resistance plasmid, we conducted pair-wise *in vitro* competition experiments comparing the growth of a lineage A before and after the acquisition of the plasmid; namely lineage A 4w and 2m isolates. In these experiments, carriage of the plasmid incurred a cost of 6.3% (±1.9%) per generation on lineage A. However, despite the *in vitro* fitness cost of plasmid carriage, and the lack of obvious selection, the lineage persisted in the gut for at least another 4 months; showing an increase in cell counts during this time (**Figure 2A**).

We speculated that while the pHUSEC41-1-like plasmid slowed the growth of lineage A host *in vitro,* these conditions do not reflect the natural habitat of the strains and important environmental factors might contribute to fitness advantage of plasmid-carried genes *in vivo*. Therefore, to assess whether the plasmid provided a fitness advantage in a model gut environment, we tested the fitness of the plasmid bearing strain in the mouse gut before and after acquisition of the plasmid. Here we observed that the plasmid-carrying isolate out-competed the plasmid free isolate, and that the plasmid conferred a fitness advantage to lineage A in the mouse gut ($P > 0.01$; **Figure 4A**).

The fact that the lineage A transconjugant survived, increased its population counts, and exhibited a fitness advantage *in vivo*,

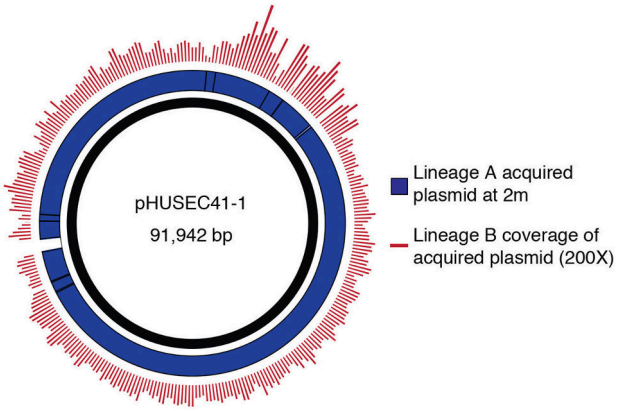


FIGURE 3 | Transfer of a plasmid mediating antibiotic resistance. Contigs corresponding to the newly acquired plasmid were identified by analyzing differences in the read alignment coverage before and after the change in the resistance profile. Reads from lineage B are mapped to the acquired plasmid contigs of A, displaying coverage depth. High coverage and identity between the strains was observed. The acquired plasmid contigs of A were aligned to the sequence of pHUSEC41-1.

highlights that resistance genes may more readily disseminate, and persist in healthy individuals never treated with antibiotics, than previously believed. However, studies examining the cost of plasmid carriage are often performed *in vitro*, and disagreement between *in vitro* and *in vivo* fitness measurements observed here, emphasizes the importance of investigating the persistence of antibiotic resistance in more natural settings.

While efforts have been devoted to studying the persistence of multidrug resistance plasmids in clinical *E. coli* isolates (Porse et al., 2016), our knowledge on the behavior of natural plasmids *in situ* of their native environment is limited (Conlan et al., 2014, 2016; Porse et al., 2017). While some studies show that stable inheritance and adaptive traits are crucial for long term plasmid survival (Simonsen, 2010), others suggest that certain conjugative plasmids can maintain themselves if present in their natural habitat of structured biofilms (Fox et al., 2008; Madsen et al., 2013). A substantial portion of pHUSEC41-1 encodes the *tra* genes involved in conjugative transfer. In addition to the effect of horizontal dissemination on plasmid persistence, conjugative transfer systems of plasmids have previously been shown to enhance adhesion and biofilm formation; features that may provide a survival advantage in the densely populated and structured environment of the gut (Ghigo, 2001; Fox et al., 2008; Madsen et al., 2013).

We tested the conjugative ability of the plasmid *in vitro* as well as *in vivo* of the mouse gut and found that the pHUSEC41-1-like plasmid conjugates at frequencies above 10^{-6} transconjugants per donor in all the tested conditions. In the mouse gut, an average of 10.8% of lineage A population had received the pHUSEC41-1-like plasmid from the lineage B strain at the final day 7 time point, indicating that the plasmid is actively conjugating in this environment.

In addition, pHUSEC41-1 encodes numerous proteins of unknown function that could potentially benefit its host *in vivo*, but further molecular analysis would be required to elucidate their role in plasmid persistence. However, candidate genes mediating the *in vivo* selection of pHUSEC41-1 could be factors involved in cobalamin biosynthesis (*cbiX*), DNA repair (*impCAB*; (Runyen-Janecky et al., 1999; Bali et al., 2014)), and conjugational transfer (*tra*). The CbiX protein can function as the terminal enzyme in siroheme biosynthesis in *E. coli*, which is known to aid iron utilization by its host (Bali et al., 2014). Iron is often restricted in the human body, and the ability to exploit these limited iron resources has been linked to increased persistence of *E. coli in vivo* (Andrews et al., 2003). pHUSEC-41-1 also harbors the *imp* operon, encoding an error-prone DNA repair system, that has been linked to increased survival following mutagenesis in a *Shigella* host and could similarly enhance the survival of *E. coli* hosts exposed to stressful conditions of the gut (Runyen-Janecky et al., 1999).

A Large Deletion Observed in Lineage a Was Associated with an Increase in Population Counts *in situ*

After the acquisition of the pHUSEC41-1-like plasmid, a large deletion was detected in lineage A isolates at the 6-month sampling point (**Figure 5**). The deletion totaled 100.4 kb, aligned to a contiguous region in *E. coli* strain CFT 073 (NC_004431) and PCR assays confirmed the deletion (**Figure S1**). Annotated genes located in the region included iron scavenging genes, such as the *iroA* gene cluster and the hemolysin activator protein, peptide antibiotic genes microcin H47 and colicin-E1, which target *E. coli,* and antigen 43, which may have a role in adhesion (Cascales et al., 2007; Selkrig et al., 2012). Lastly, genes involved in fatty-acid synthesis, carbohydrate, and amino acid metabolism were also lost as a result of the deletion (See **Table S5** for a complete list). A smaller chromosomal deletion was also identified in the lineage B $12m_1$ isolate (**Figure 5**). The deleted region totaled 26 kb and included genes characteristic for horizontally acquired DNA; including P fimbriae encoded by the *pap* genes as well as mobile element genes (See **Table S6** for complete list).

At the 2 month sampling time, when lineage B was first sampled, lineages A and B had roughly the same population counts, at $10^{7.8}$ and $10^{7.7}$ CFU/g, respectively (**Figure 2A**). However, in contrast to lineage B, the population counts of lineage A increased by an order of magnitude at 6m. Upon receiving a foreign plasmid, antagonistic interactions between horizontally acquired chromosomal and plasmid factors might lower the fitness of the host e.g., due to overlapping gene or regulatory functions and these may be compensated through deletions (San Millan et al., 2015; Porse et al., 2016). To assess whether the large deletion, that occurred in lineage A between the 2 and 6 months sampling point, served as an adaptive response for lineage A, we performed an *in vivo* fitness assessment in the mouse gut between lineage A 2m and 6m, representing isolates before and after the large deletion. We did not find a statistically significant difference in fitness of the lineage A isolates with and without the large deletion (**Figure 4B**), suggesting that the deletion did not drive the increased population counts more than

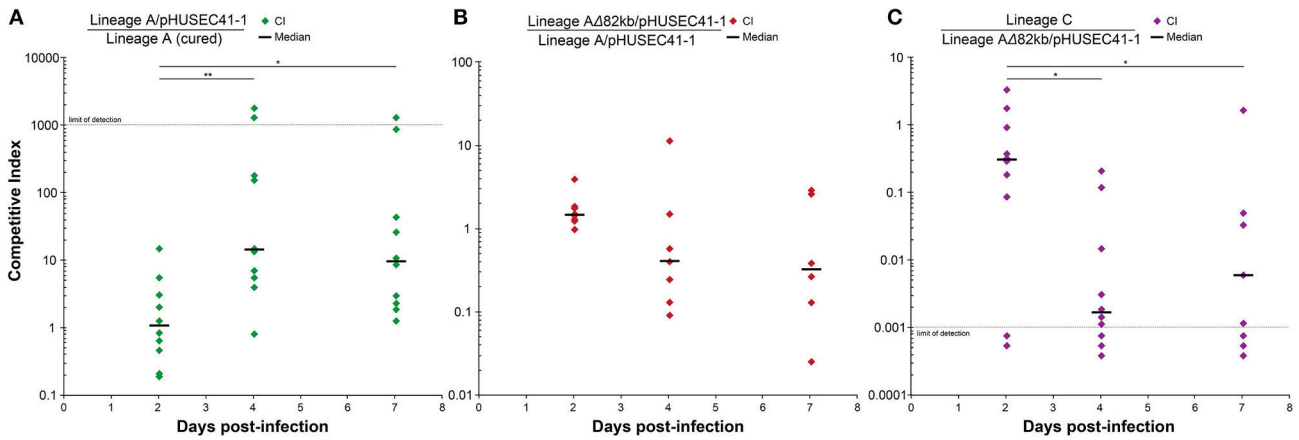


FIGURE 4 | *In vivo* competition experiments. For each experiment, 10 mice were inoculated orally with equal amounts of two strains to quantify their relative fitness in a mouse gut environment. **(A)** The two lineage A isolates, with and without the pHUSEC41-1-like plasmid, were competed to assess the fitness effect of plasmid carriage *in vivo*. **(B)** The fitness effect of the 82kb deletion in lineage A, occurring between the 2 and 6 month time points, was hypothesized to be advantageous, but no significant fitness increase was measured for the deletion-isolate. **(C)** Lineage C was competed against the 6 months lineage A isolate to assess the potential role of lineage C in the disappearance of lineage A. Competitive indexes were analyzed relative to day 2 using the non-parametric Mann–Whitney *U*-test with a $P < 0.05$ considered significant and the degrees of statistical significance presented as $^{**}P < 0.01$ or $^{*}P < 0.05$.

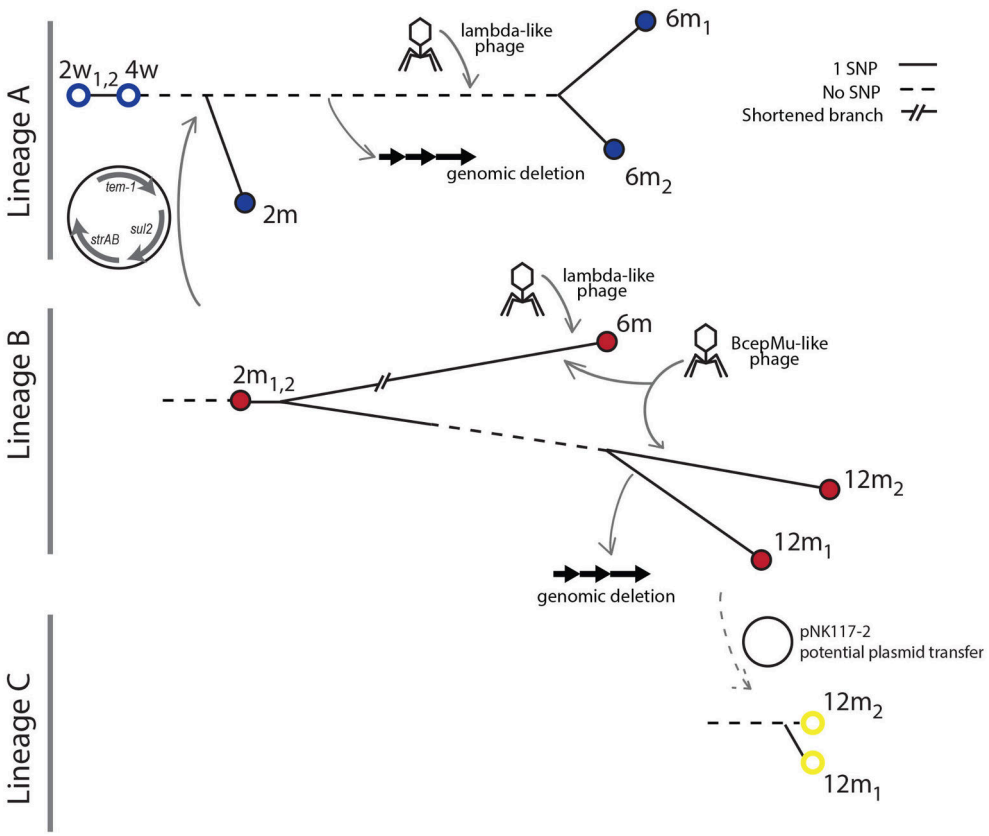


FIGURE 5 | Overview of lineage genome dynamics. The transfer of a multidrug resistance plasmid from lineage B to lineage A occurred prior to the 2 month sampling time. The transfer occurred before diversification of the A lineage. At the 6 month sampling point, a Bcep-mu like phage infecting the B lineage was detected. In addition, both the A and B lineages were infected by lambda-like phages at this time point. A large genomic deletion occurred in the A lineage after the 2 month but before the 6 month sampling point. No isolates of lineage A were obtained at the final sampling time at 12 months, but a new isolate from lineage C is sampled along with lineage B.

plasmid carriage in itself. As samples were not obtained between the 2m and 6m time points, and a Lambda-like phage infection also occurred within this time-span, a potential beneficial effect of the deletion could be outweighed by the subsequent phage acquisition. Just as acquisition of plasmid DNA can alter cell homeostasis, features of the acquired plasmid such as conjugation are known to induce the SOS response, which can increase the rate of genome rearrangements (Baharoglu and Mazel, 2014).

Incoming Lineage C Shares an IncX Plasmid with Lineage B and Establishes in the Gut despite Inferior Fitness in *in Vivo* Experiments

Lineage C was sampled for the first time at the final 12m sampling time point. As the related lineage A was not sampled at this time and the counts of lineage B were the lowest sampled, we hypothesized that lineage C could be superior in terms of its ability to survive and compete in the gut. Therefore, we performed an *in vivo* fitness assessment in the mouse gut between the previously most abundant lineage A 6m and lineage C 12m strains. We found, in contrast to our hypothesis, that the lineage A 6m isolate out-competed the lineage C isolate in the mouse gut ($P < 0.05$; **Figure 4C**). Although the fitness of *E. coli* lineages is likely to vary between the human and mouse intestine, this result indicates that other factors of the complex gut environment not related to the appearance of lineage C, such as interactions with the remaining constituents of the microbiota or phage predators, may have played a more prominent role in the disappearance of lineage A. In addition, lineage C harbored a plasmid of 35.8 kbp, termed pNK117-2, which contained the *pilx* conjugation system similar to that of pOLA52 (Norman et al., 2008; **Figure S2**). Interestingly, pNK117-2 had 100% similarity to a plasmid from lineage B and might have been transferred from lineage B to lineage C. While we cannot demonstrate a second *in situ* transfer event, as we did not sample lineage C previously without pNK117-2, the presence of pNK117-2 in both lineage B and C with 100% sequence identity further exemplifies how plasmids can experience rapid dissemination in the absence of obvious selection.

CONCLUSIONS

This work highlights the advantages of studying the longitudinal dynamics of co-existing bacterial lineages in the gut microbiota as a complement to metagenomic sequencing efforts. The power of this approach is expected to increase as cultivation methods for representative sampling of the gut microbiota improves further, and we anticipate that studies augmenting metagenomic sequencing with genomic sequencing and *in vivo* fitness models will provide a richer and more detailed view of the highly dynamic nature of individual genomes and HGT in the human gut microbiota. The substantial genome plasticity captured in this study highlights the dynamic nature of individual genomes of the gut microbiota. Of particular interest, we identify the transfer of a multi-drug resistance plasmid at the genomic level between co-existing bacterial lineages in the unperturbed human gut. Our findings suggest that, even though antibiotic resistance genes are not considered beneficial in the absence of antibiotic selection, they may hitchhike along with other selected traits. Further studies investigating the molecular mechanisms responsible for host compatibility and persistence of endemic antibiotic resistance plasmids *in situ* will refine our knowledge on the existence conditions of mobile elements, which will allow a better understanding of their role in the epidemiology and evolution of pathogenic bacteria.

AUTHOR CONTRIBUTIONS

MS, HG, AP, DA, and JK conceived and designed the study. IA and AW designed the ALLERGYFLORA study and isolated the strains used in the present study. HG conducted the genomic analysis and strain phenotyping. NK performed the initial typing of the *E. coli* lineages, the phenotypic resistance testing, and the *in vitro* fitness cost assays. AP aided in strain sequencing, did *in vitro* conjugation assays, finalized the manuscript and tagged the isolates with resistance markers that JK and ML used to perform *in vivo* fitness experiments. HG, AP, and CM wrote the manuscript with input from JK, MS, ML, NK, IA, AW, and DA.

FUNDING

This work was supported by the Danish Free Research Councils for Health and Disease, the European Union FP7-HEALTH-2011-single-stage grant agreement 282004, EvoTAR (MS and DA), the Medical Faculty of the University of Göteborg (ALFGBG138401) and the Swedish Medical Research Council (DA). MS further acknowledges support from the Novo Nordisk Foundation and the Lundbeck Foundation.

DATA AVAILABILITY

All sequenced genomes can be accessed via the Bioproject PRJNA396689.

ACKNOWLEDGMENTS

We thank Mari Rodriguez de Evgrafov for preparing single-end sequencing libraries and Lejla Imamovic for advice regarding phages.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2017.01852/full#supplementary-material

Figure S1 | Large deletion in the genome of lineage A. Contigs from strain A were aligned to reference genome CFT073. Dark gray colored contigs represent regions flanking the excision. Pale colored contigs represent the region lost due to the deletion. Arrows indicate the position of the primers designed based on the CFT073 genome used to confirm the genomic excision.

Figure S2 | Plasmid map of pHK117-2**.** Plasmid NK117-2 identified in both lineage B and C compared to IncX1 plasmids pRPEC180_47 (middle ring, blue) and pOLA52 (outer ring, green). Open reading frames (ORFs) identified on pNK117-2 are drawn in the inner most ring in black, with arrows indicating the reading direction. Annotations for selected ORFs are labeled outside of the rings.

Table S1 | Antibiotic susceptibility tests of selected isolates. MIC values of lineages A, B, and C for sampling time points. One isolate per lineage per sampling time was selected for MIC testing.

Table S2 | SNP comparison lineage A and B. Number of SNPs between the lineages using selected isolates. The rows of the table indicate the reads that were used aligned to contigs of the isolate as indicated in the column.

Table S3 | SNP comparison lineage A and C. Coverage of the total genomic content between the lineages using selected isolates. The rows of the table indicate the reads that were used aligned to contigs of the isolate as indicated in the column.

Table S4 | Within lineage SNPs. Table containing the SNPs from lineage A, B, and C, respectively, including the annotation and whether the amino acid change was synonymous or non-synonymous.

Table S5 | Deleted genes from Linage A. List of annotated genes identified in the deleted chromosomal region of lineage A.

Table S6 | Deleted genes from Linage B. List of annotated genes identified in the deleted chromosomal region of lineage B.

REFERENCES

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. doi: 10.1186/1471-2164-12-402

Andrews, S. C., Robinson, A. K., and Rodríguez-Quiñones, F. (2003). Bacterial iron homeostasis. *FEMS Microbiol. Rev.* 27, 215–237. doi: 10.1016/S0168-6445(03)00055-X

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Baharoglu, Z., and Mazel, D. (2014). SOS, the formidable strategy of bacteria against aggressions. *FEMS Microbiol. Rev.* 38, 1126–1145. doi: 10.1111/1574-6976.12077

Bali, S., Rollauer, S., Roversi, P., Raux-Deery, E., Lea, S. M., Warren, M. J., et al. (2014). Identification and characterization of the "missing" terminal enzyme for siroheme biosynthesis in α-proteobacteria. *Mol. Microbiol.* 92, 153–163. doi: 10.1111/mmi.12542

Balzola, F., Bernstein, C., Ho, G. T., and Lees, C. (2010). A human gut microbial gene catalogue established by metagenomic sequencing: Commentary. *Inflamm. Bowel Dis. Monit.* 11:28. doi: 10.1038/nature08821

Bidet, P., Burghoffer, B., Gautier, V., Brahimi, N., Mariani-Kurkdjian, P., El-Ghoneimi, A., et al. (2005). *In vivo* transfer of plasmid-encoded ACC-1 AmpC from *Klebsiella pneumoniae* to *Escherichia coli* in an infant and selection of impermeability to imipenem in *K. pneumoniae*. *Antimicrob. Agents Chemother.* 49, 3562–3565. doi: 10.1128/AAC.49.8.3562-3565.2005

Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., et al. (2014). *In silico* detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* 58, 3895–3903. doi: 10.1128/AAC.02412-14

Cascales, E., Buchanan, S. K., Duche, D., Kleanthous, C., Lloubes, R., Postle, K., et al. (2007). Colicin biology. *Microbiol. Mol. Biol. Rev.* 71, 158–229. doi: 10.1128/MMBR.00036-06

Chen, S. L., Wu, M., Henderson, J. P., Hooton, T. M., Hibbing, M. E., Hultgren, S. J., et al. (2013). Genomic diversity and fitness of *E. coli* strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection. *Sci. Transl. Med.* 5, 184ra60. doi: 10.1126/scitranslmed.3005497

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Conlan, S., Park, M., Deming, C., Thomas, P. J., Young, A. C., Coleman, H., et al. (2016). Plasmid dynamics in KPC-positive Klebsiella pneumoniae during long-term patient colonization. *MBio* 7, e00742–e00716. doi: 10.1128/mBio.00742-16

Conlan, S., Thomas, P. J., Deming, C., Park, M., Lau, A. F., Dekker, J. P., et al. (2014). Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Sci. Transl. Med.* 6:254ra126. doi: 10.1126/scitranslmed.3009845

Davies, J., and Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* 74, 417–433. doi: 10.1128/MMBR.00016-10

Enne, V. I., Delsol, A. A., Davis, G. R., Hayward, S. L., Roe, J. M., and Bennett, P. M. (2005). Assessment of the fitness impacts on *Escherichia coli* of acquisition of antibiotic resistance genes encoded by different types of genetic element. *J. Antimicrob. Chemother.* 56, 544–551. doi: 10.1093/jac/dki255

Forslund, K., Sunagawa, S., Kultima, J. R., Mende, D. R., Arumugam, M., Typas, A., et al. (2013). Country-specific antibiotic use practices impact the human gut resistome. *Genome* 23, 1163–1169. doi: 10.1101/gr.155465.113

Fox, R. E., Zhong, X., Krone, S. M., and Top, E. M. (2008). Spatial structure and nutrients promote invasion of IncP-1 plasmids in bacterial populations. *ISME J.* 2, 1024–1039. doi: 10.1038/ismej.2008.53

Ghigo, J. M. (2001). Natural conjugative plasmids induce bacterial biofilm development. *Nature* 412, 442–445. doi: 10.1038/35086581

Grad, Y. H., Godfrey, P., Cerquiera, G. C., Mariani-Kurkdjian, P., Gouali, M., Bingen, E., et al. (2013). Comparative genomics of recent Shiga toxin-producing *Escherichia coli* O104:H4: short-term evolution of an emerging pathogen. *MBio* 4, 1–10. doi: 10.1128/mBio.00452-12

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Karami, N., Martner, A., Enne, V. I., Swerkersson, S., Adlerberth, I., and Wold, A. E. (2007). Transfer of an ampicillin resistance gene between two *Escherichia coli* strains in the bowel microbiota of an infant treated with antibiotics. *J. Antimicrob. Chemother.* 60, 1142–1145. doi: 10.1093/jac/dkm327

Khan, Z., Bloom, J. S., Kruglyak, L., and Singh, M. (2009). A practical algorithm for finding maximal exact matches in large sequence datasets using sparse suffix arrays. *Bioinformatics* 25, 1609–1616. doi: 10.1093/bioinformatics/btp275

Kuhlman, T. E., and Cox, E. C. (2010). Site-specific chromosomal integration of large synthetic constructs. *Nucleic Acids Res.* 38:e92. doi: 10.1093/nar/gkp1193

Künne, C., Billion, A., Mshana, S. E., Schmiedel, J., Domann, E., Hossain, H., et al. (2012). Complete sequences of plasmids from the hemolytic-uremic syndrome-associated *Escherichia coli* strain HUSEC41. *J. Bacteriol.* 194, 532–533. doi: 10.1128/JB.06368-11

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25. doi: 10.1186/gb-2009-10-3-r25

Lasaro, M., Liu, Z., Bishar, R., Kelly, K., Chattopadhyay, S., Paul, S., et al. (2014). *Escherichia coli* isolate for studying colonization of the mouse intestine and its application to two-component signaling knockouts. *J. Bacteriol.* 196, 1723–1732. doi: 10.1128/JB.01296-13

Lester, C. H., Frimodt-Møller, N., Sørensen, T. L., Monnet, D. L., and Hammerum, A. M. (2006). *In vivo* transfer of the vanA resistance gene from an Enterococcus faecium isolate of animal origin to an *E. faecium* isolate of human origin in the intestines of human volunteers. *Antimicrob. Agents Chemother.* 50, 596–599. doi: 10.1128/AAC.50.2.596-599.2006

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Linkevicius, M., Anderssen, J. M., Sandegren, L., and Andersson, D. I. (2016). Fitness of *Escherichia coli* mutants with reduced susceptibility to tigecycline. *J. Antimicrob. Chemother.* 71, 1307–1313. doi: 10.1093/jac/dkv486

Liu, L., Chen, X., Skogerbø, G., Zhang, P., Chen, R., He, S., et al. (2012). The human microbiome: a hot spot of microbial horizontal gene transfer. *Genomics* 100, 265–270. doi: 10.1016/j.ygeno.2012.07.012

Lutz, R., and Bujard, H. (1997). Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res.* 25, 1203–1210. doi: 10.1093/nar/25.6.1203

Madsen, J. S., Burmølle, M., and Sørensen, S. J. (2013). A spatiotemporal view of plasmid loss in biofilms and planktonic cultures. *Biotechnol. Bioeng.* 110, 3071–3074. doi: 10.1002/bit.25109

McConnell, M. A., Mercer, A. A., and Tannock, G. W. (1991). Transfer of plasmid pAMβl between members of the normal microflora inhabiting the murine digestive tract and modification of the plasmid in a *Lactobacillus reuteri* host. *Microb. Ecol. Health Dis.* 4, 343–355. doi: 10.3109/08910609109140149

Norman, A., Hansen, L. H., She, Q., and Sørensen, S. J. (2008). Nucleotide sequence of pOLA52: a conjugative IncX1 plasmid from *Escherichia coli* which enables biofilm formation and multidrug efflux. *Plasmid* 60, 59–74. doi: 10.1016/j.plasmid.2008.03.003

Nowrouzian, F., Hesselmar, B., Saalman, R., Strannegård, I. L., Åberg, N., Wold, A. E., et al. (2003). *Escherichia coli* in infants' intestinal microflora: colonization rate, strain turnover, and virulence gene carriage. *Pediatr. Res.* 54, 8–14. doi: 10.1203/01.PDR.0000069843.20655.EE

Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., and Brown, P. O. (2007). Development of the human infant intestinal microbiota. *PLoS Biol.* 5, 1556–1573. doi: 10.1371/journal.pbio.0050177

Porse, A., Gumpert, H., Kubicek-Sutherland, J. Z., Karami, N., Adlerberth, I., Wold, A. E., et al. (2017). Genome dynamics of *Escherichia coli* during antibiotic treatment: transfer, loss, and persistence of genetic elements *in situ* of the infant gut. *Front. Cell. Infect. Microbiol.* 7:126. doi: 10.3389/fcimb.2017.00126

Porse, A., Schønning, K., Munck, C., and Sommer, M. O. A. (2016). Survival and evolution of a large multidrug resistance plasmid in new clinical bacterial hosts. *Mol. Biol. Evol.* 33, 2860–2873. doi: 10.1093/molbev/msw163

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Runyen-Janecky, L. J., Hong, M., and Payne, S. M. (1999). The virulence plasmid-encoded impCAB operon enhances survival and induced mutagenesis in *Shigella flexneri* after exposure to UV radiation. *Infect. Immun.* 67, 1415–1423.

San Millan, A., Toll-Riera, M., Qi, Q., and MacLean, R. C. (2015). Interactions between horizontally acquired genes create a fitness cost in *Pseudomonas aeruginosa*. *Nat. Commun.* 6:6845. doi: 10.1038/ncomms7845

Schjørring, S., Struve, C., and Krogfelt, K. A. (2008). Transfer of antimicrobial resistance plasmids from *Klebsiella pneumoniae* to *Escherichia coli* in the mouse intestine. *J. Antimicrob. Chemother.* 62, 1086–1093. doi: 10.1093/jac/dkn323

Selkrig, J., Mosbahi, K., Webb, C. T., Belousoff, M. J., Perry, A. J., Wells, T. J., et al. (2012). Discovery of an archetypal protein transport system in bacterial outer membranes. *Nat. Struct. Mol. Biol.* 19, 506–510. doi: 10.1038/nsmb.2261

Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., and Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* 23, 111–120. doi: 10.1101/gr.142315.112

Shoemaker, N. B., Vlamakis, H., Hayes, K., and Salyers, A. A. (2001). Evidence for extensive resistance gene transfer among *Bacteroides* spp. and among bacteroides and other genera in the human colon evidence for extensive resistance gene transfer among *Bacteroides* spp. and among bacteroides and other genera in the human C. *Appl. Env. Microbiol.* 67, 561–568. doi: 10.1128/AEM.67.2.561-568.2001

Simonsen, L. (2010). The Existence conditions for bacterial plasmids: theory and reality. *Microb. Ecol.* 22, 187–205. doi: 10.1007/BF02540223

Trobos, M., Lester, C. H., Olsen, J. E., Frimodt-Møller, N., and Hammerum, A. M. (2009). Natural transfer of sulphonamide and ampicillin resistance between *Escherichia coli* residing in the human intestine. *J. Antimicrob. Chemother.* 63, 80–86. doi: 10.1093/jac/dkn437

Wiegand, I., Hilpert, K., and Hancock, R. E. W. (2008). Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat. Protoc.* 3, 163–175. doi: 10.1038/nprot.2007.521

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L. H., 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

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Res.* 39, 347–352. doi: 10.1093/nar/gkr485

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary material

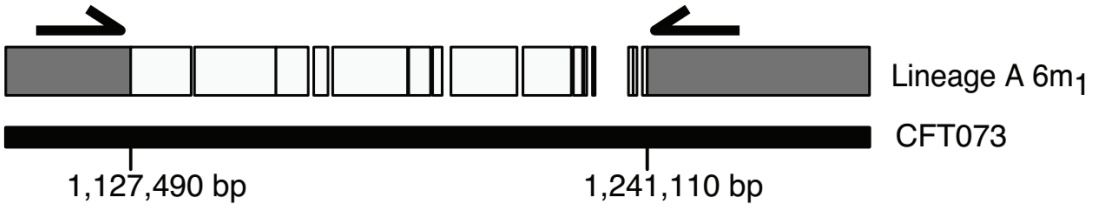


Figure S1. Large deletion in the genome of lineage A. Contigs from strain A were aligned to reference genome CFT073. Dark gray colored contigs represent regions flanking the excision. Pale colored contigs represent the region lost due to the deletion. Arrows indicate the position of the primers designed based on the CFT073 genome used to confirm the genomic excision.

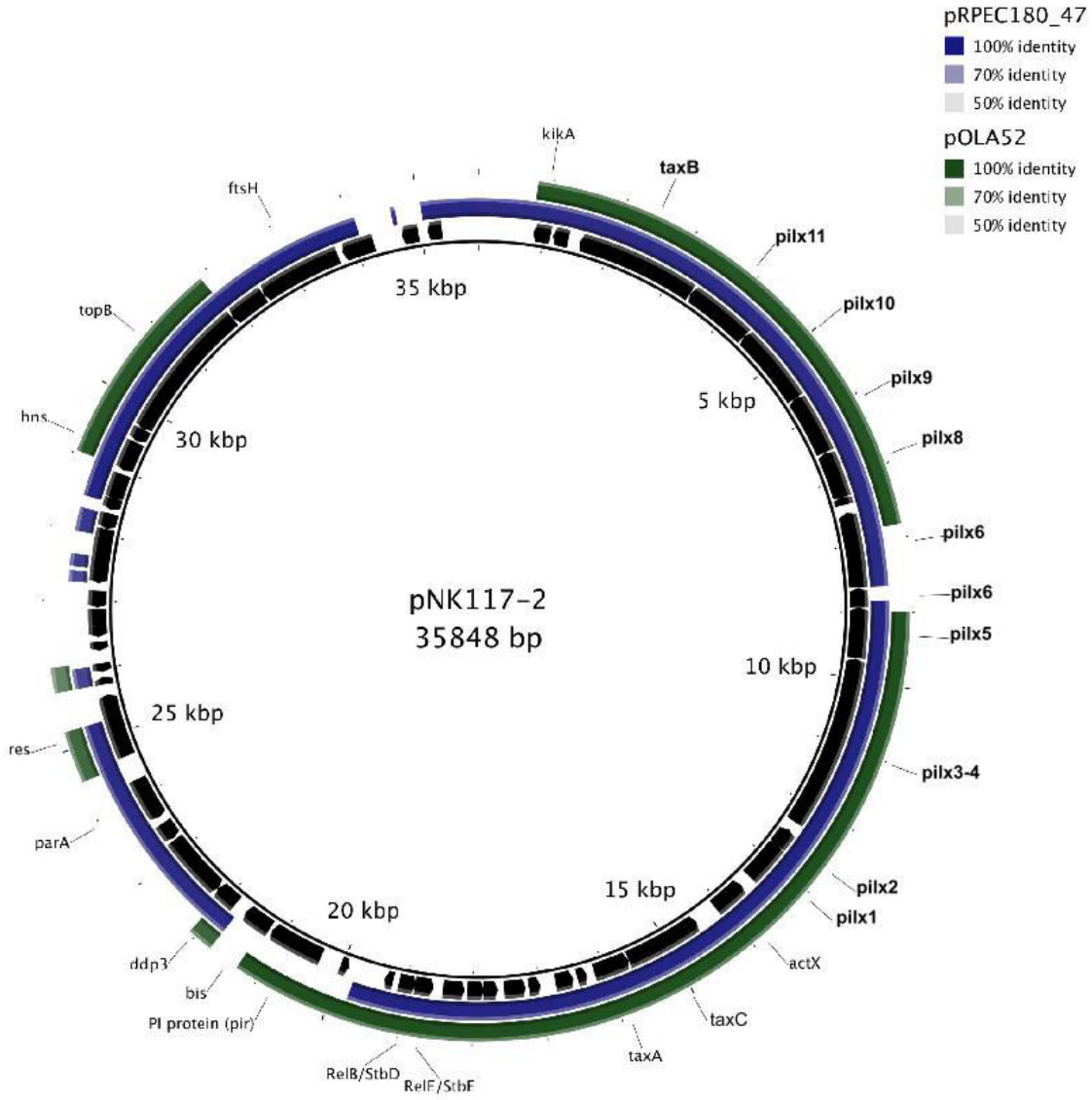


Figure S2. Plasmid map of pHK117-2. Plasmid NK117-2 identified in both lineage B and C compared to IncX1 plasmids pRPEC180_47 (middle ring, blue) and pOLA52 (outer ring, green). Open reading frames (ORFs) identified on pNK117-2 are drawn in the inner most ring in black, with arrows indicating the reading direction. Annotations for selected ORFs are labeled outside of the rings.

Table S1. Antibiotic susceptibility tests of selected isolates. MIC values of lineages A, B, and C for sampling time points. One isolate per lineage per sampling time was selected for MIC testing.

Lineage	**Lineage A**				**Lineage B**			**Lineage C**
Sampling Time	2w1	4w	2m	6m1	2m1	6m	12m1	12m1
Antibiotic (ug/mL)								
Ampicillin	2	2	512	512	512	512	512	2
Piperacillin	1	1	64	64	32	32	32	0.5
Mecillinam	0.125	0.125	1	1	1	1	0.5	0.125
Ceftazidime	0.25	0.25	0.25	0.25	0.0625	0.125	0.125	0.125
Cefuroxime	8	8	8	8	2	4	2	2
Cefoxitin	2	4	4	4	8	4	8	2
Chloramphenicol	1	1	1	1	1	1	1	1
Gentamicin	0.5	0.5	0.5	0.5	1	1	1	1
Tobramycin	0.25	0.25	0.5	0.125	0.5	0.25	0.5	0.5
Streptomycin	2	2	64	128	128	128	128	4
Nalidixic Acid	1	1	1	1	1	0.5	1	0.5
Tetracycline	0.5	0.5	0.5	0.5	1	1	0.5	1
Trimethoprim	4	4	2	4	0.125	0	0.125	0.125
Sulfamethoxazole	8	16	1024	1024	512	1024	1024	512

Table S2. SNP comparison lineage A and B. Number of SNPs between the lineages using selected isolates. The rows of the table indicate the reads that were used aligned to contigs of the isolate as indicated in the column.

		Aligned to contigs from		
		Lineage A 2w2	Lineage B 2m2	Lineage C 12m2
Reads from	Lineage A 2w2		94346	7324
	Lineage B 2m2	95215		94797
	Lineage C 12m2	7297	92100	

Table S3. SNP comparison lineage A and C. Coverage of the total genomic content between the lineages using selected isolates. The rows of the table indicate the reads that were used aligned to contigs of the isolate as indicated in the column.

		Aligned to contigs from		
		Lineage A 2w2	Lineage B 2m2	Lineage C 12m2
Reads from	A 2w2		80.0%	93.6%
	B 2m2	82.3%		82.0%
	C 12m2	95.5%	79.7%	

Table S4. Within lineage SNPs. Table containing the SNPs from lineage A, B, and C, respectively, including the annotation and whether the amino acid change was synonymous or non-synonymous.

Lineage A

Contig number	Position	2w2	2w1	4w	2m	6m1	6m2	Annotation	AA Change
5	27035	A		G	G	G	G		
7	282598	G			A			Formate hydrogenlyase subunit 2	No
32	38337	A			C			Betaine aldehyde dehydrogenase	Yes
13	108313	C			T			Phosphoenolpyruvate carboxylase	Yes
12	157912	G				A		Putative inner membrane protein	No
52	28325	C				A		Protein fdrA	Yes
239	107	T				C			
93	2581	T					C		
8	11719	G					T	type 1 fimbriae (FimD)	Yes

Lineage B

Contig number	Position	2m2	2m1	6m	12m1	12m2	Annotation	AA Change
12	273551	T		A			Putative amidohydrolase	Yes
20	13798	T		A			FIG00639301: hypothetical protein	Yes
590	436	G		A			Phage major tail tube protein	No
91	20935	G		A			Alcohol dehydrogenase	Yes
709	463	G		A				Yes
141	121	T		C				Yes
64	121	T		C				Yes
237	2376	C		G			IncF plasmid conjugative transfer pilin TraA	Yes
37	40257	A		G			LysR family transcriptional regulator lrhA	No
389	490	T		G			Phage capsid scaffolding protein	Yes
56	41072	T		G			Dipeptide transport system permease DppB	Yes
58	4945	A		G				Yes
24	63963	A		T	T	T	Oligopeptide transport system permease OppB	Yes
28	44283	C		T			Fe2+-dicitrate sensor, membrane component	Yes
35	41468	A		T			C-terminal domain of CinA type S	Yes
511	121	C		T			Adenine-specific methyltransferase	Yes
63	19848	A		T			Type III secretion inner membrane protein	Yes
593	514	C		T				Yes
778	418	G		T				Yes
36	2740	T			A	A	Chaperone protein DnaK	Yes
37	78302	T			A	A	Putative transporting ATPase	Yes
50	33170	C			A		Putative vimentin	Yes
80	12348	G			A		D-Galactonate repressor DgoR	Yes
12	327228	G			A			Yes
42	91228	G			C	C	FUSARIC ACID RESISTANCE PROTEIN FUSB/FUSC	Yes
18	98805	A			T	T	N-succinyl-L,L-diaminopimelate desuccinylase	Yes

Lineage C

Contig number	Position	12m1	12m2	Annotation	AA Change
35	132	T	C		

Table S5. Deleted genes from Linage A. List of annotated genes identified in the deleted chromosomal region of lineage A.

Category	Annotated genes
Fatty Acid Synthesis	(3R)-hydroxymyristoyl-[ACP] dehydratase (EC 4.2.1.-)
	3-hydroxydecanoyl-[ACP] dehydratase (EC 4.2.1.60)
	3-oxoacyl-[ACP] reductase (EC 1.1.1.100)
	3-oxoacyl-[ACP] synthase
	3-oxoacyl-[ACP] synthase (EC 2.3.1.41) FabV like
	Acyl carrier protein
	Acyl carrier protein (ACP1)
	FIG002571: 4-hydroxybenzoyl-CoA thioesterase domain protein
	FIG018329: 1-acyl-sn-glycerol-3-phosphate acyltransferase
	FIG143263: Glycosyl transferase / Lysophospholipid acyltransferase
	FIGfam138462: Acyl-CoA synthetase, AMP-(fatty) acid ligase
Carbohydrate and Amino Acid Metabolism	D-galactarate dehydratase (EC 4.2.1.42)
	D-galactarate dehydratase (EC 4.2.1.42)
	Tagatose-6-phosphate kinase AgaZ (EC 2.7.1.144)
	Putative O-methyltransferase
	Aspartate aminotransferase (EC 2.6.1.1)
	2-keto-3-deoxy-D-arabino-heptulosonate-7-phosphate synthase I alpha
Iron Scavenging	Glycosyltransferase IroB
	ABC transporter protein IroC
	Trilactone hydrolase IroD
	Periplasmic esterase IroE
	Outer Membrane Siderophore Receptor IroN
	Hemolysin activator protein precursor
	Heme utilization or adhesion of ShlA/HecA/FhaA family
Bacteriocins and Virulence	antigen 43 precursor
	Colicin-E1
	mannose-specific adhesin FimH
	microcin H47 secretion/processing ATP-binding mchF
	MchC protein Putative F1C and S fimbrial switch Regulatory protein
	type 1 fimbrae adaptor subunit FimF
	type 1 fimbriae anchoring protein FimD
	type 1 fimbriae major subunit FimA
	type 1 fimbriae protein FimI2C
	YeeV toxin protein
Other	Transposase (X4)
	Putative Transposase
	Mobile element protein
	Integrase IS, phage, Tn: Transposon-related functions
	Putative metal chaperone, involved in Zn homeostasis

	entry exclusion protein 2
	membrane: Transport of small molecules (Cations)
	FIG021862: membrane protein, exporter
	FIG027190: Putative transmembrane protein
	putative membrane protein (X2)
	putative secretion permease
	putative regulatory protein
	NgrB
	RNA-Arg-TCT

Table S6. Deleted genes from Linage B. List of annotated genes identified in the deleted chromosomal region of lineage B.

Category	Annotated genes
Virulence	putative regulator PapX protein
	PapG protein
	PapF protein
	PapE protein
	Fimbrial adapter papK precursor
	Periplasmic fimbrial chaperone
	Fimbriae usher protein StfC
	minor pilin subunit PapH
	major pilin subunit PapA
	PapI protein
Iron Scavenging	TonB-dependent receptor
Other	Mobile element protein (X9)
	conserved hypothetical protein; putative exported protein
	hypothetical protein (X4)
	FIG00638040: hypothetical protein
	FIG00638658: hypothetical protein
	FIG00640659: hypothetical protein
	FIG01070050: hypothetical protein
	FIG00641173: hypothetical protein

Manuscript III [redacted]

Survival and Evolution of a Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts

Andreas Porse,[1] Kristian Schønning,[2] Christian Munck,*[,1] and Morten O.A. Sommer*[,1]

[1]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark

[2]Department of Clinical Microbiology, Hvidovre University Hospital, Hvidovre, Denmark and Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

*Corresponding authors: E-mails: chrmu@bio.dtu.dk; msom@bio.dtu.dk.

Associate Editor: Miriam Barlow

Abstract

Large conjugative plasmids are important drivers of bacterial evolution and contribute significantly to the dissemination of antibiotic resistance. Although plasmid borne multidrug resistance is recognized as one of the main challenges in modern medicine, the adaptive forces shaping the evolution of these plasmids within pathogenic hosts are poorly understood. Here we study plasmid–host adaptations following transfer of a 73 kb conjugative multidrug resistance plasmid to naïve clinical isolates of *Klebsiella pneumoniae* and *Escherichia coli*. We use experimental evolution, mathematical modelling and population sequencing to show that the long-term persistence and molecular integrity of the plasmid is highly influenced by multiple factors within a 25 kb plasmid region constituting a host-dependent burden. In the *E. coli* hosts investigated here, improved plasmid stability readily evolves via IS26 mediated deletions of costly regions from the plasmid backbone, effectively expanding the host-range of the plasmid. Although these adaptations were also beneficial to plasmid persistence in a naïve *K. pneumoniae* host, they were never observed in this species, indicating that differential evolvability can limit opportunities of plasmid adaptation. While insertion sequences are well known to supply plasmids with adaptive traits, our findings suggest that they also play an important role in plasmid evolution by maintaining the plasticity necessary to alleviate plasmid–host constrains. Further, the observed evolutionary strategy consistently followed by all evolved *E. coli* lineages exposes a trade-off between horizontal and vertical transmission that may ultimately limit the dissemination potential of clinical multidrug resistance plasmids in these hosts.

Key words: **clinical isolates, antibiotic resistance, horizontal gene transfer, ESBL plasmid evolution, IS26 restructuring, experimental evolution.**

Article

Introduction

Conjugative plasmids are key contributors to horizontal gene transfer and carry a wide variety of accessory genetic elements important for the ecology and adaptation of bacterial species (Frost et al. 2005; Norman et al. 2009; Soucy et al. 2015). The role of plasmids in the dissemination of antibiotic resistance is increasingly worrisome for human health; allowing pathogenic bacteria to obtain multiple resistance genes in a single transfer event (Carattoli 2013). Indeed, strains of *Klebsiella pneumoniae* and *Escherichia coli* carrying multidrug resistance plasmids are currently recognized as one of the most urgent antibiotic resistance problems (WHO 2014). In these strains, plasmids encoding carbapenemases and extended spectrum β-lactamases (ESBLs) are of particular concern because they greatly limit effective treatment options (Davies and Davies 2010; Dhillon and Clark 2012).

Currently, the abundance and persistence of large plasmids in competitive environments with little or no selection pressure remains an evolutionary puzzle (Simonsen 1991; Bergstrom et al. 2000). While the range of hosts in which plasmids can successfully replicate is fairly well understood (Mazodier and Davies 1991; Carattoli 2009; Jain and Srivastava 2013), our knowledge of the influence of different host backgrounds on the long-term stability and evolution of natural plasmids remains limited. A number of studies have demonstrated that plasmids confer a cost upon entering a naïve host and that this can be compensated through adaptive evolution (Bouma and Lenski 1988; Dahlberg and Chao 2003; Dionisio et al. 2005; Harrison and Brockhurst 2012). Furthermore, recent studies have used next-generation sequencing to investigate the genetic basis underlying such plasmid–host adaptations (Sota et al. 2010; Harrison and Brockhurst 2012; San Millan et al. 2014; Harrison et al. 2015; Loftie-Eaton et al. 2015; San Millan et al. 2015). In the case of small nonconjugative plasmids, plasmid–host adaptations have been shown to occur via mutations in the plasmid replication machinery (Sota et al. 2010), in chromosomal genes interacting with replication proteins (San Millan et al. 2015) or by acquisition of stabilizing traits via interplasmid transposition (Loftie-Eaton et al. 2015). For plasmid–host evolution of a conjugative mercury-resistance plasmid to *Pseudomonas fluorescens*, adaptation occurred via translational down regulation caused by inactivation of a host-encoded two-component system (Harrison et al. 2015). However, less effort has been made to study the dynamics

and adaptation of large multidrug resistance plasmids in the bacterial hosts they are likely to encounter in a clinical environment.

Here, we investigate the molecular basis for adaptations of a large plasmid, isolated from a clinical *K. pneumoniae* strain, to naïve clinical isolates of *E. coli* and *K. pneumoniae*; two species highly implicated in plasmid mediated dissemination of multidrug resistance. To determine the factors implicated in the long-term plasmid survival, we characterize the influence of the plasmid on maximum growth rate, overall stability, and evolutionary potential as well as genetic adaptations in three novel plasmid–host combinations (fig. 1). We observe a consistent and rapid adaptation pattern in *E. coli* that is dependent on plasmid-borne insertion sequences (ISs) and driven by the cost a plasmid region encompassing the main conjugational machinery. We are the first to describe the role of IS mediated intramolecular restructuring in plasmid host-expansion and how these reactions, although broadly beneficial, were only observed in certain host backgrounds.

Results

The pKP33 plasmid (fig. 2) was obtained from a *K. pneumoniae* clinical host (Kp33) and sequenced using *Pacific Biosciences RS II* single molecule real time sequencing as well as *Illumina MiSeq* technology. While the short reads (~150 bp) obtained from the paired-end *Illumina MiSeq* run were not sufficient to capture the complex repetitive nature of the pKP33 plasmid, such as the localization and orientation of identical mobile genetic elements, the longer reads offered by single molecule real time sequencing enabled complete assembly of the plasmid.

The sequence of pKP33 revealed that it belongs to the IncN incompatibility group. Although IncN plasmids are able to replicate in a variety of enterobacterial pathogens, they are most frequently observed in *E. coli* and *K. pneumoniae* isolates where backbones similar to pKP33 contribute to the global dissemination of cephalosporin and carbapenem resistance (Miriagou et al. 2010; Eikmeyer et al. 2012; Carattoli 2013; Conlan et al. 2014). pKP33 carries eleven antibiotic resistance genes including the endemic CTX-M-15 ESBL variant (Bush and Fisher 2011). These genes confer resistance to multiple major drug classes including β-lactams (CTX-M-15, TEM-1, OXA-1), aminoglycosides (*aac-(6′)-III*, *aac(6′)-Ib and strA*, *strB*), quinolones (*qnrB*), sulfonamides (*sul2*) and dihydrofolate reductase inhibitors (*folA*). These were functionally confirmed by antibiotic susceptibility testing (supplementary table S1, Supplementary Material online). To support stable maintenance, pKP33 contains the *stb* operon known to encode factors involved in active segregation and regulation of conjugative transfer (Guynet et al. 2011). In addition, an *ecoRII-dcm* restriction–antirestriction pair may function as a toxin–antitoxin stability system by inhibiting growth of plasmid-free segregants (Mruk and Kobayashi 2014). The plasmid backbone shows a typical mosaic structure where functionally similar gene clusters are flanked by mobile genetic elements. In pKP33, the IS*6* and Tn*3* family of ISs were highly abundant; exemplified by several occurrences of identical IS*26* copies throughout the backbone.

Initial stability of pKP33 in the original *K. pneumoniae* host (Kp33) was investigated by serial passaging in the laboratory for 35 days in the absence of antibiotic selection (supplementary fig. S1, Supplementary Material online). We estimate, based on the transfer volumes and maximum culture density that this period corresponds to approximately 280 bacterial generations. Remarkably, plasmid loss was never detected for pKP33 in its native host, emphasizing that large plasmids carrying multiple antibiotic resistance genes can persist for extended time periods without selection given the right host environment.

Three clinical strains that did not carry multidrug resistance plasmids (naïve) were obtained to investigate their response to acquisition of pKP33. The *E. coli* isolates Ec37 (ST127) and Ec38 (ST1170) originate from clinical UTIs, belong to the B2 *E. coli* phylogroup, and share 80% of their protein families. The naïve Kp08 (ST36) *K. pneumonia* strain shares 54% of its protein families with both *E. coli* strains and 83% with the original pKP33 plasmid ancestor Kp33 (ST301).

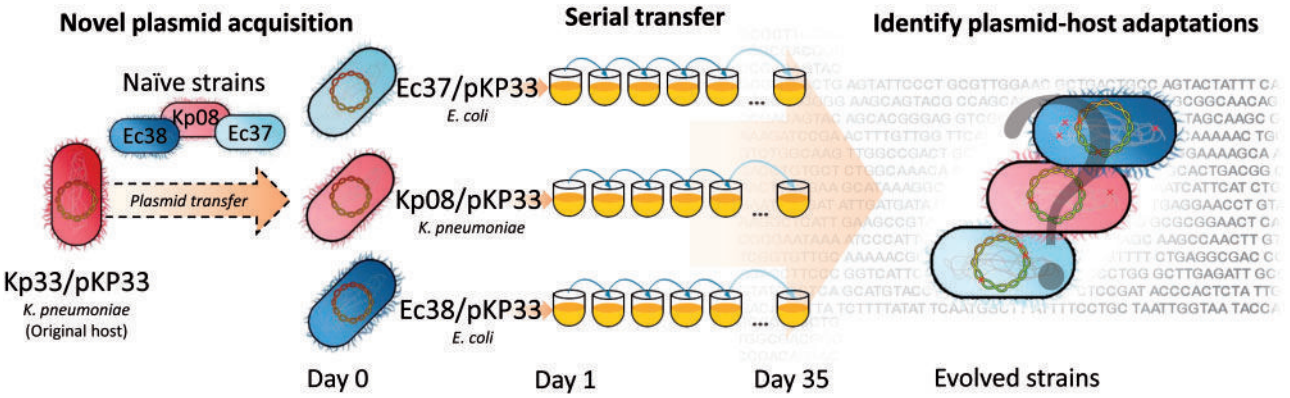


Fig. 1. Experimental overview. A large conjugative plasmid (pKP33) originating from a clinical *Klebsiella pneumoniae* isolate (Kp33) was transferred by conjugation into two different *Escherichia coli* (Ec37 and Ec38) strains and one *K. pneumoniae* (Kp08) strain isolated from urinary tract infections (UTIs) and blood infections, respectively. Quantification of the growth rate and plasmid stability was done immediately after plasmid acquisition (Day 0) as well as following 35 days of serial transfer in plasmid-selective conditions (Day 35—evolved strains). The genome sequences of the evolved strains were analyzed to determine the genetic adaptations involved in plasmid–host adaptations.

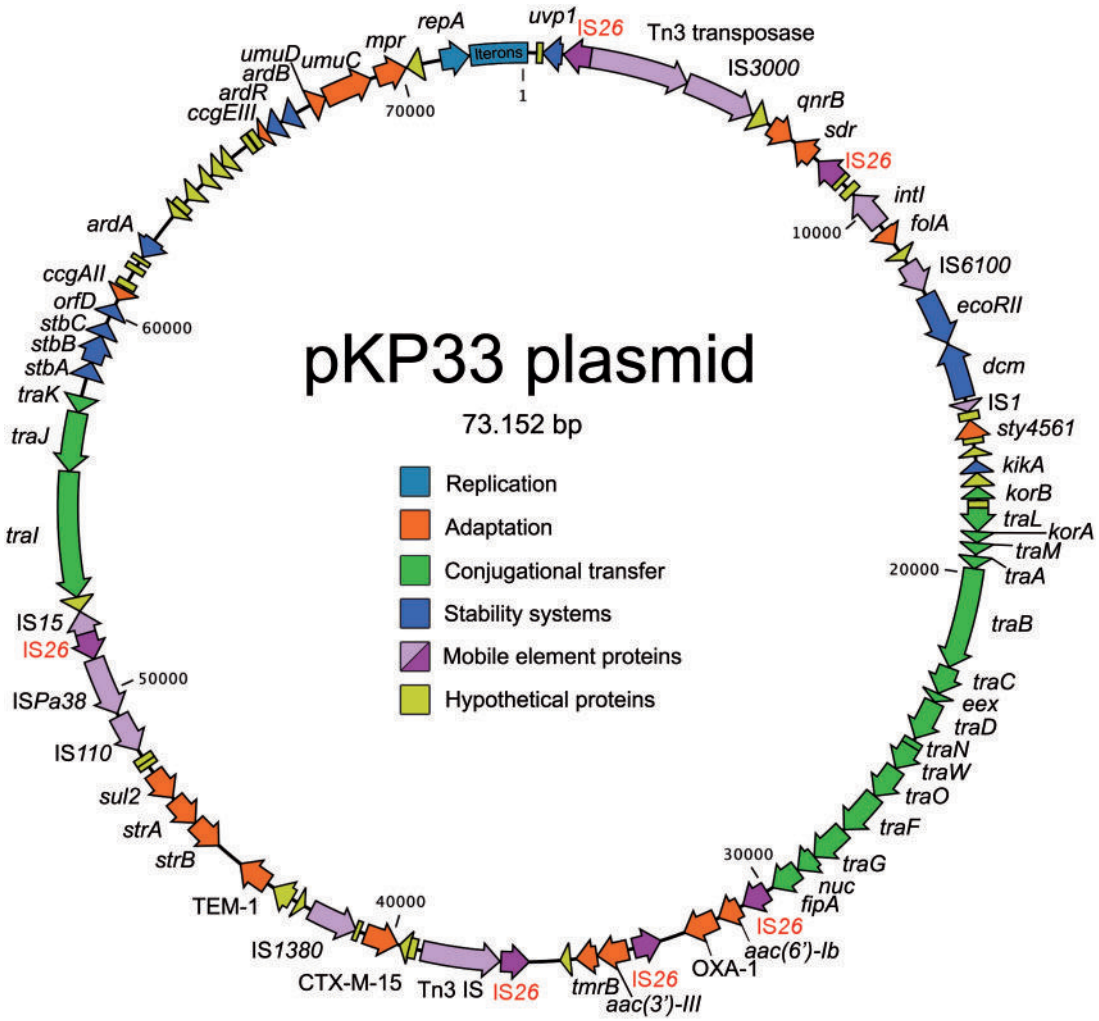


Fig. 2. Genetic map of the pKP33 ESBL plasmid originating from the clinical Kp33 strain. The plasmid belongs to the IncN incompatibility group and contains accessory genes (orange) involved in metabolism and antibiotic resistance along with several stability mechanisms (blue) and conjugational transfer machinery (green). Three β-lactamases, including the endemic CTX-M-15 ESBL, are encoded by the plasmid along with several other antibiotic resistance genes. Furthermore, the plasmid is sectioned by a number of mobile genetic elements (purple) dominated by the IS*26* insertion sequence (dark purple) as well as predicted genes that encode proteins of unknown function (yellow).

Additional information on these strains can be found in table 1 and the full proteome comparison is summarized in supplementary table S2, Supplementary Material online.

Plasmid Cost and Stability in pKP33-Naïve Clinical Isolates

While pKP33 was stably maintained in its native host strain in the absence of selection, this is likely a result of long-term co-evolution. To study the processes that contribute to plasmid success in new hosts, we set out to investigate the stability of pKP33 in naïve clinical *E. coli* and *K. pneumoniae* strains. We transferred pKP33 by conjugation to one *K. pneumoniae* (Kp08) and two *E. coli* (Ec37 and Ec38) clinical isolates (see table 1). To characterize the influence of the host strain on the cost of pKP33 carriage we measured the maximum growth rate as a fitness proxy for UTI pathogens (Gordon and Riley 1992; Nilsson et al. 2003). The plasmid imposed a significant burden on all three strains quantified as a reduction in growth rate of 6%, 11%, and 12.5% for Kp08, Ec37, and Ec38, respectively, compared to the plasmid-free ancestor (one-sample *t*-test. Ec38: $P < 0.001$; Ec37: $P < 0.001$; Kp08: $P < 0.001$) (fig. 3). While there was no significant difference in relative growth rate between Ec37/pKP33 and Ec38/pKP33 (two-sample *t*-test $P = 0.3$), the burden imposed by pKP33 on the Kp08 *K. pneumoniae* strain was significantly lower than for the *E. coli* isolates (two-sample *t*-test Ec38: $P = 0.0034$; Ec37: $P = 0.01$), implying that host specific factors influence the fitness cost of the newly acquired plasmid.

Due to the presence of various plasmid borne stability systems, pKP33 might persist in spite of the burden imposed on naïve hosts. To investigate the stability of pKP33 in the naïve strains, serial passaging of five lineages of each strain in nonselective medium was carried out for 35 days, corresponding to 280 generations of growth (fig. 4A, triangles). The stability of pKP33 differed between the isolates with the naïve Kp08 *K. pneumoniae* strain being the most stable (50% plasmid-free cells after 12.5 days) and *E. coli* the least stable (50% plasmid-free cells after 2.5 and 6 days for Ec37 and Ec38, respectively) (fig. 4B). Notably, while still detectable in all Ec37 and Kp08 populations, pKP33 was present in less than

Table 1. Strains and Plasmids. Clinical isolates were obtained from the Department of Clinical Microbiology at Hvidovre Hospital (Hvidovre, Denmark).

Strain or Plasmid	Designation	Characteristics and Comments	Reference or Source
Escherichiacoli TOP10	TOP10	Conjugation donor similar to DH10B (leucine auxotroph)	Invitrogen
Escherichia coli DY329	DY329	Used for recombineering. W3110 derivative with the genotype: Δ*lacU169 nadA*:*Tn10 gal490* λ*cI857* Δ(*cro-bioA*)	Yu et al. (2000)
Klebsiella pneumoniae CI ESBL 33	Kp33	ST*: *301*; Chromosomal bla: SHV27; Plasmids: pKP33	Clinical isolate (urine)
ESBL plasmid from Kp33	pKP33	Size: 67kb; IncN; Resistance genes: TEM-1, OXA-1, CTX-M 15, *strA*,*strB*, *aac(6′)Ib-cr*, *aac(3)-Iia*, *qnrB*, *sul2*, *folA*	Clinical plasmid
Escherichia coli CI 37	Ec37	ST*: *127*, Phylogroup: *B2**; *Resistance profile*: *None*; Plasmids replicons*: *FII* (61 kb)	Clinical isolate (urine)
Escherichia coli CI 38	Ec38	ST*: *1170*, Phylogroup: *B2**; Resistance profile: *chloramphenicol*; Plasmids replicons*: *FIB* (9.5 kb) & *FIC* (136 kb)	Clinical isolate (urine)
Klebsiella pneumoniae CI 08	Kp08	ST*: *36*; Resistance genes: *fosA*, SHV11, *oqxB*, *oqxA*;Plasmids replicons*: *FIB(K)* (22kb) & *IncR* (90 Kb)	Clinical isolate (blood)

NOTE.—Asterisks indicate computationally derived information obtained through the CGE Servers (MLST, PlasmidFinder and ResFinder available at: https://cge.cbs.dtu.dk/services, last accessed August 8, 2016).

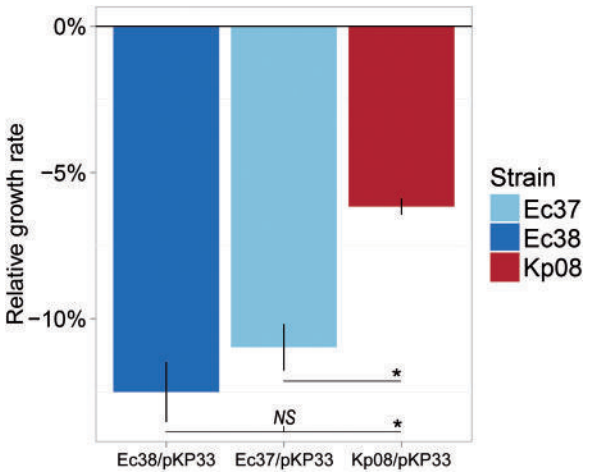


FIG. 3. Growth rate effects of novel plasmid acquisition. Plasmid acquisition leads to a reduction in maximum growth rate relative to the plasmid-free ancestor. Error bars show 95% confidence intervals of 16 pKP33 carrying biological replicates error propagated with the variation of 16 biological replicates of the plasmid-free ancestor. Asterisks indicate significant differences ($P < 0.05$) and *NS* a nondetectable difference ($P > 0.05$).

1% of the population for all three hosts after 35 days of serial passaging, demonstrating that these new plasmid–host combinations are markedly less stable compared to the original pKP33 host (supplementary fig. S1, Supplementary Material online).

In order to gain insight into the origin of the observed instability, we fitted an established mathematical model of plasmid stability to the measured plasmid stability data (Proctor 1994). This model describes the ratio of plasmid bearing to plasmid-free cells as a function of time and assumes the rate of plasmid loss upon cell division and the difference in growth rate between the plasmid bearing and plasmid-free cells as the main drivers of plasmid dynamics (equation 1—Materials and Methods). We did not include a parameter for conjugative transfer as pKP33 displayed negligible rates of conjugation in our liquid setup (see supplementary table S3, Supplementary Material online), which is in agreement with previous studies of IncN plasmid transfer systems (Bradley et al. 1980). The parameter estimates (see supplementary table S4, Supplementary Material online) for segregational loss ranged from 0.0008 to 0.0147 and due to uncertainty of the estimates, these were not significantly different from 0 for Ec37 and Ec38 (one-sample *t*-test, $P = 0.13$ and $P = 0.45$ for Ec37 and Ec38, respectively) but only for Kp08 (one-sample *t*-test, $P = 0.0013$). This indicates that the influence of the segregational loss rate on overall plasmid stability is minor compared to the effect of growth competition from plasmid-free segregants. Consistent with the growth rate reduction imposed by the plasmid (fig. 3), the fitness cost of pKP33 predicted by the model was statistically significant for all three strains (one-sample *t*-test; $P = 0.0054$, $P = 0.00019$, and $P = 0.012$ for Ec37, Ec38, and Kp08, respectively) and determined to be 8.3% (±4.14%), 14% (±2.85%), and 4.6% (±2.94%) for Ec37, Ec38, and Kp08, respectively. These estimates of plasmid fitness costs correspond well with the relative growth rates measured and confirm that the main driver of plasmid loss was competition between plasmid-bearing and plasmid-free segregants.

Adaptive Evolution during Antibiotic Selection Compensates Plasmid Cost in *E. coli* but not *K. pneumoniae*

Based on the plasmid stability experiments without selection, it seems unlikely that pKP33 can be stably maintained even in strains closely related to the original host (fig. 4 and supplementary table S2, Supplementary Material online). However, in clinical settings, plasmid transfer and subsequent host adaptation is likely promoted by selection for plasmid-encoded functions such as antibiotic resistance. In the presence of antibiotic selection pKP33 would confer a strong fitness advantage to an antibiotic sensitive host cell (supplementary table S1, Supplementary Material online). Such periods of selection can provide sufficient time for plasmid–host adaptations to occur, improving plasmid persistence once selection is removed (San Millan et al. 2014).

To investigate plasmid–host adaptation during positive selection, we repeated the serial transfer experiment in liquid medium containing cefuroxime (16 μg/ml) selecting for pKP33 carriage. Five lineages of each host carrying the

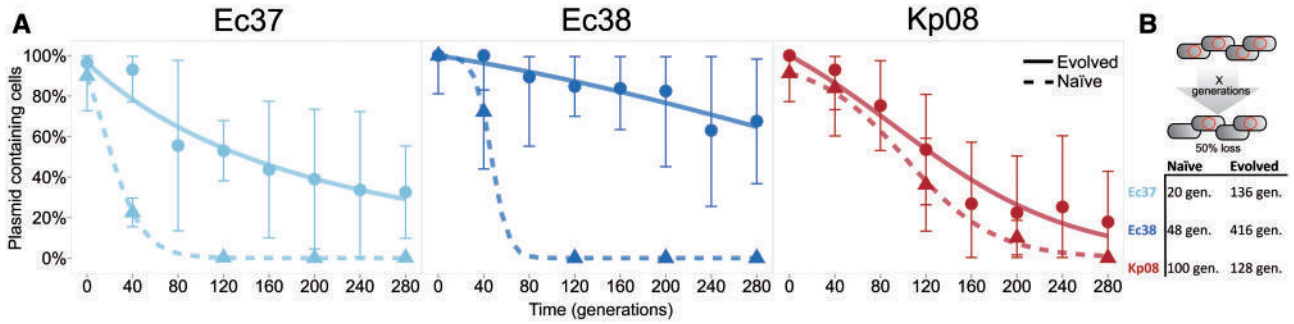


Fig. 4. Adaptive evolution increases stability of pKP33 in *Escherichia coli* hosts. (*A*) The proportion of plasmid-containing cells in initially plasmid bearing populations grown without plasmid selection as a function of time. Immediately after receiving pKP33, the naïve Ec37, Ec38, and Kp08 were propagated in nonselective medium for 35 days to quantify the stability of the newly received plasmid before any adaptations had taken place (triangles). To measure the stabilizing effect of plasmid–host adaptations, the serial passaging procedure was applied to the same plasmid–host combinations evolved for 35 days under plasmid selective conditions (circles). A mathematical model was fitted (lines) to the data points. Error bars illustrate the standard deviation of five independent lineages. (*B*) Plasmid half-life comparison. The average half-life of pKP33 in naïve and evolved lineages was determined from the fitted model and summarized as the number of generations needed for half of the population to be plasmid-free.

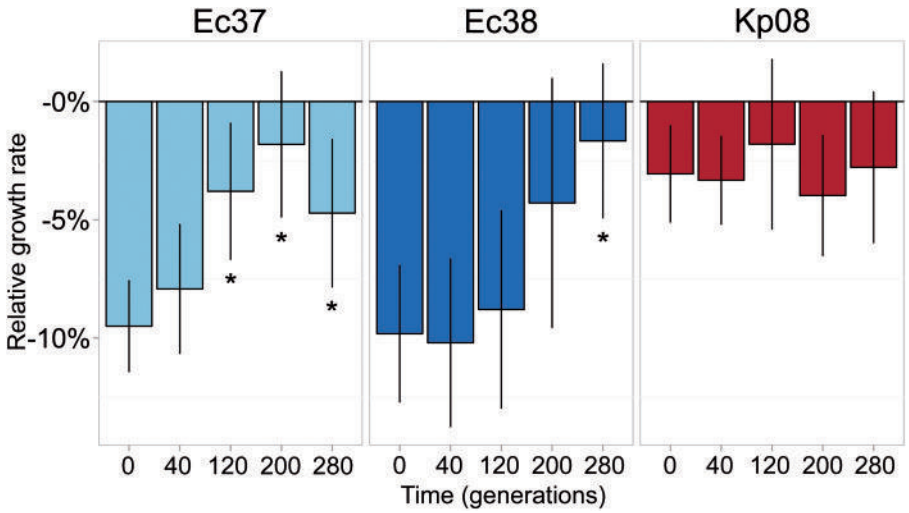


Fig. 5. Adaptive evolution reduces plasmid cost in *Escherichia coli*. The growth rate of each plasmid–host combination was measured throughout the evolution experiment and is shown relative to the plasmid-free host that was evolved in parallel. Means represent 40 randomly picked clones (8 from each evolving lineage) and error bars show the standard deviation of lineage means. Asterisks indicate a significant improvement from Day 0 (*$P < 0.05$, $n = 5$, Dunnett's test).

pKP33 plasmid were evolved in parallel and the growth rates of individual clones from the evolving populations were measured throughout the ~280 generations of the experiment and growth rates were normalized to the plasmid-free host grown in parallel to account for the effect of general medium adaptations (fig. 5. See supplementary fig. S2, Supplementary Material online for absolute values).

A significant increase in relative growth rate was observed for both *E. coli* carrying pKP33 when comparing the starting point (Day 0) to the evolved endpoint (Day 35) (two-sample *t*-test: Ec38: $P = 0.00215$, $n = 10$; Ec37: $P = 0.0154$, $n = 10$). At the end of the adaptation experiment the average evolved Ec38/pKP33 lineage, now referred to as Ec38/pKP33 D. 35, was indistinguishable from the plasmid-free host (one-sample *t*-test: $P = 0.186$, $n = 5$) . In contrast, the *K. pneumoniae* strain did not experience a significant improvement during the experiment (two-sample *t*-test: $P = 0.380$, $n = 10$).

Our growth rate measurements and stability assays (figs. 4 and 5) revealed that a less costly plasmid–host combination is more stable on the population level. To test whether the stability of the evolved plasmid–host combinations were higher than for the nonevolved ancestors, plasmid stability was measured again by serial passaging in nonselective medium for 35 days or roughly 280 bacterial generations (fig. 4A, circles). A pronounced improvement in the stability of the evolved *E. coli* lineages was evident from the much slower decline in the proportion of plasmid-bearing cells over time compared to the stability of the nonevolved lineages (fig. 4B). The strains evolved under antibiotic selection were able to maintain the plasmid for much longer, with a significant proportion of the *E. coli* populations still carrying

pKP33 after 35 days of serial passaging without antibiotic selection (one-sample *t*-test: Ec37: $P = 0.032$, $n = 5$; Ec38: $P = 0.0075$, $n = 5$). The two *E. coli* strains shows similar improvements in stability with a 6.8- and 8.6-fold increases in pKP33 half-life, respectively compared to the naïve hosts (fig. 4B). In contrast, we could not detect a significant improvement in stability of Kp08/pKP33 D.35 compared to the ancestral Kp08/pKP33 lineage. Comparing the values for ancestral and evolved Kp08/pKP33 lineages at different time points reveal that the plasmid containing fraction of the evolved lineages were not significantly higher than the ancestral plasmid–host combination (two-sample *t*-test: Day 15: $P = 0.31$, $n = 5$; Day 35: $P = 0.18$, $n = 5$) which is in accordance with the overlap of parameter estimates from the mathematical model fitted to data points of the ancestral and evolved Kp08/pKP33 lineages (supplementary table S4, Supplementary Material online).

Population Sequencing Reveals Large Deletions in *E. coli* Evolved Plasmids

To further examine the underlying factors responsible for the dramatic increase in plasmid stability we sequenced all evolved lineages at the population level. Sequencing reads from the evolved strains was mapped to the genome of the nonevolved ancestor representing the starting point before plasmid acquisition and a control population of each strain, evolved without the plasmid, enabling omission of the most frequent general medium adaptations (see supplementary table S5, Supplementary Material online). However, the majority of the chromosomal mutations occurred in virulence genes and none of them could be directly associated with plasmid stability. See supplementary material S1, Supplementary Material online for a discussion of genomic variants.

While no single nucleotide polymorphisms (SNPs) were identified in the *E. coli* evolved plasmids, a consistent deletion of approximately 25 kb (position 8–33 kb, fig. 6A) in pKP33 was observed for all *E. coli* lineages. Similar deletions were never observed in any of the evolved *K. pneumoniae* lineages (fig. 6B). The consistently deleted region encodes three resistance determinants, a restriction-antirestriction system and the main conjugation machinery. See supplementary fig. S3, Supplementary Material online for a detailed view of the region. Interestingly, deletions were always flanked by IS*26* ISs, suggesting that intramolecular transposition or recombination between these elements allowed for the observed loop-out dynamics. IS*26* is over-represented on bacterial plasmids, often associated with antibiotic resistance genes, and believed to catalyse plasmid reorganization through intramolecular replicative transposition events (Cullik et al. 2010; Curiao et al. 2011; Partridge et al. 2011; He et al. 2015).

Deletion Dynamics during Adaptation

Due to the positioning of IS*26* sequences in the pKP33 backbone, partial deletion of the 25 kb region was never observed. Therefore the deletion dynamics of the entire region were assumed to correlate with the presence of individual genes within the region. The trimethoprim (TMP) resistance gene *folA* located inside the region (highlighted in fig. 6A) allowed us to investigate the dynamics of the major deletion by selective plating of the antibiotic evolved populations on TMP containing agar plates (fig. 6B). Here, trimethoprim (TMP) resistance was never lost in neither the original Kp33 *K. pneumoniae* host nor the evolving Kp08/pKP33 lineages, consistent with a high intramolecular stability of pKP33 in these *K. pneumoniae* strains. Conversely, the *E. coli* populations showed a rapid decline in TMP resistance with TMP sensitive mutants dominating the population within 60 generations of evolution.

Plasmid Deletions Are Responsible for the Observed Growth Rate Improvements

The rapid fixation of the 25 kb deletion in all evolved *E. coli* lineages indicated that the main genetic structure involved in adaptation was the plasmid backbone. To verify that the observed improvements were due to the plasmid deletions and not chromosomal mutations, a representative evolved pKP33 (now referred to as pKP33evo) was purified from lineage four of the antibiotic evolved Ec38/pKP33 D.35 strain and transferred back into the ancestral strains. Growth rate measurements showed that the burden of plasmid carriage was indistinguishable between the ancestral Ec37 and Ec38 strains harboring pKP33evo and the evolved plasmid–host combinations (fig. 7). This confirms that the plasmid was the main locus of adaptation and no chromosomal mutations were needed for the observed improvements.

We transferred pKP33evo to the naïve Kp08 to investigate whether the major deletion would confer an advantage beyond the host in which it evolved (fig. 8).

Interestingly, Kp08 cells harboring the *E. coli* evolved plasmid displayed a marginal advantage in terms of a higher growth rate, which was borderline significant ($P = 0.063$, $n = 48$). To gain resolution and test the impact of the presumed fitness benefit of pKP33evo in Kp08, we conducted a 1:1 pairwise competition experiment by mixing Kp08/pKP33 and Kp08/pKP33 and subjecting these cocultures to 110 generations of serial transfer. Here, Kp08 carrying the ancestral pKP33 plasmid was consistently outcompeted in four parallel competitions against the *E. coli* evolved pKP33evo (fig. 8B). These results indicate that the deletion event rather than the adaptive advantage was the bottleneck preventing plasmid adaptation in *K. pneumoniae*.

No Single Locus Is Responsible for the Adaptive Benefit of the Major Plasmid Deletion

The majority of the deleted region encodes the type 4 secretion system (T4SS) comprising the conjugational transfer machinery that allows for autonomous horizontal transfer of the plasmid.

It is generally believed that horizontal transfer comes at the expense of reduced vertical transfer (Turner et al. 1998) and that the conjugational transfer machinery imposes a burden on the bacterial host cell (Zahrl et al. 2006; Fernandez-Lopez et al. 2014). We hypothesized that the T4SS was the main component responsible for the fitness cost of the deleted plasmid region and that the remaining part of the 25 kb

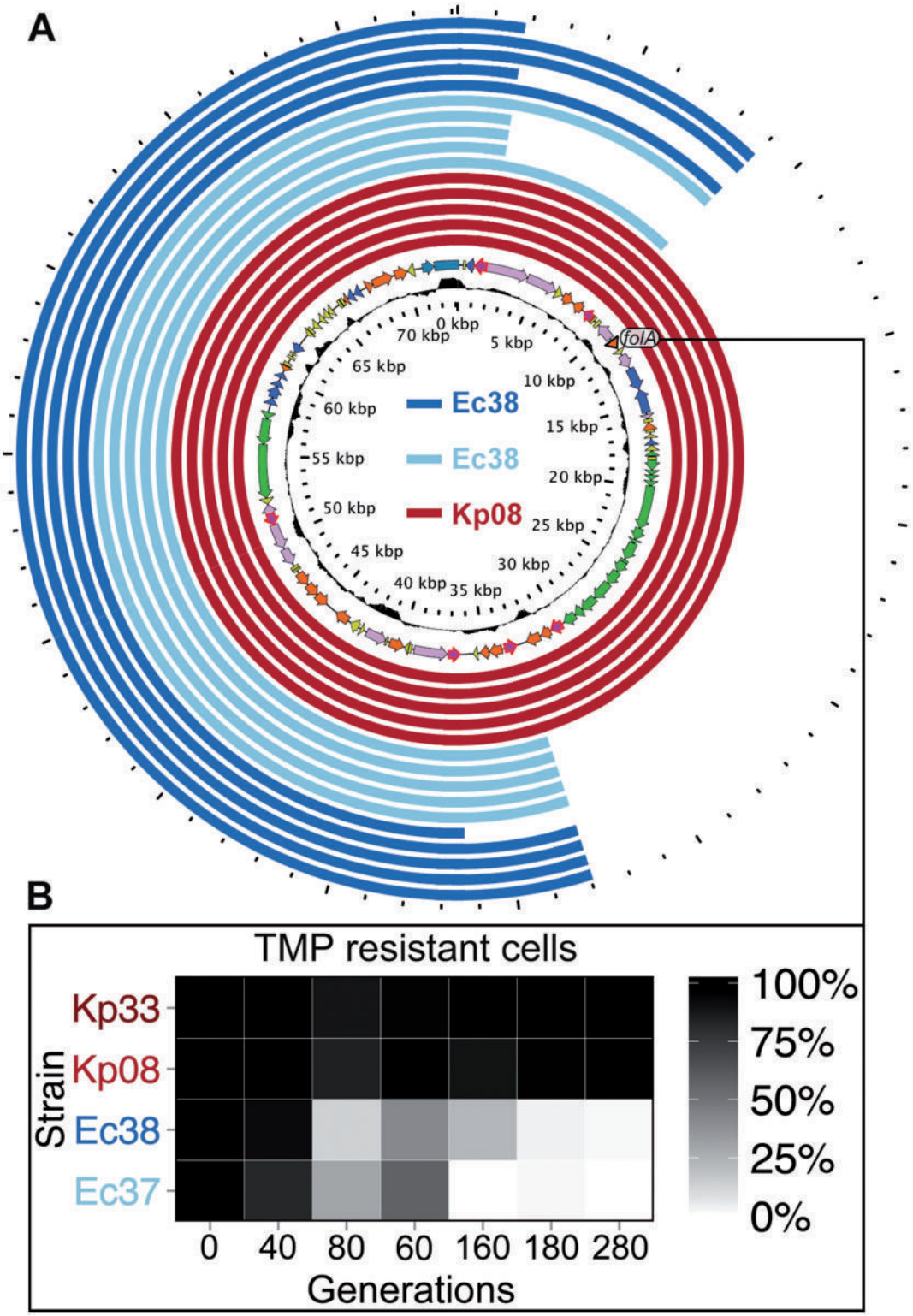


Fig. 6. Large deletions in the pKP33 plasmid backbone occurred in *Escherichia coli* but not *Klebsiella pneumoniae*. (*A*) The ancestral pKP33 plasmid compared to the evolved plasmid genomes displayed as a BLAST atlas. All evolved populations of each strain were sequenced at the 280-generation end point. IS*26* elements are indicated in purple with a red outline and the position of the trimethoprim (TMP) resistance gene *folA* is highlighted. Lineages are depicted in ascending order from the center. The inner-most circle shows the relative GC content. (*B*) Heat-map showing the abundance of TMP resistant cells in the populations grown with plasmid selection as a measure of the major deletion event. Each tile represents the average of five evolving lineages. None of the pKP33 free strains where resistant to TMP.

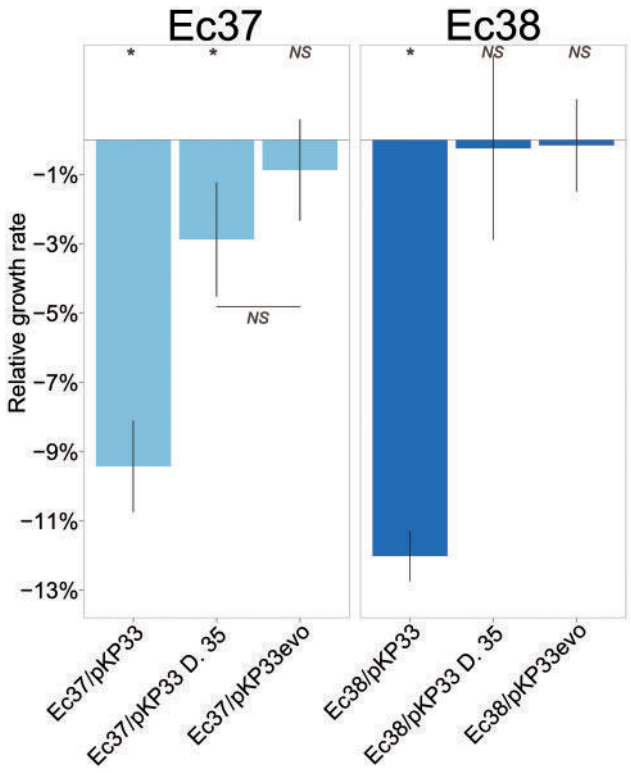


FIG. 7. Plasmid adaptations are responsible for the increased growth rate in the evolved *Escherichia coli* hosts. Growth rates of plasmid–host combinations before evolution (Day 0), after evolution (Day 35), and the evolved plasmid (pKP33evo) transferred back into the ancestral hosts, shown relative to the growth rate of the ancestral plasmid-free ancestors. A significant difference determined by two-sample *t*-test statistics is indicated with asterisks ($P < 0.05$) or NS (nonsignificant, $P > 0.05$). Top indicators are obtained through comparison to the plasmid-free ancestor (no difference). All measurements were done for 24 individually picked colonies and error bars display 95% confidence intervals of the mean.

region was codeleted merely as a result of constraints in IS*26* localization. To investigate this hypothesis, two pKP33 knockout mutants were constructed. In one pKP33 mutant (pKP33ΔT4SS) we deleted the T4SS only; while in the other mutant (pKP33Δ9-17 kb) the majority of the remaining 25 kb region containing the k*ikA*, the *ecoRII* endonucleases as well as the *folA* gene was deleted, leaving the T4SS intact. We tested the ability of the two mutants to transfer autonomously and while the pKP33ΔT4SS did not yield any transconjugants, the pKP33Δ9-17 kb mutant retained the ability to transfer by conjugation on par with the intact pKP33 plasmid (see supplementary table S3, Supplementary Material online). In both mutants, the intermediate part of the 25 kb region, containing the *kor* regulatory genes, were preserved to avoid a growth bias from excessive expression of the *kik*/*kil* or T4SS genes due to the absence of negative regulation by the KorA and KorB regulators (Moré et al. 1996).

The influence of the targeted deletions on the growth rate of the *E. coli* strains was measured along with the ancestral pKP33 and a 9–30 kb deletion mutant encompassing both partial knockouts (fig. 9).

There was no significant difference in terms of growth rate between carrying pKP33 with a targeted knockout of the 9–30 kb region when compared to the evolved plasmid (two-sample *t*-test: Ec38: $P = 0.374$, $n = 32$; Ec37: $P = 0.158$, $n = 32$) emphasizing that this deleted region was responsible for the observed plasmidcosts.

Interestingly, the effect of partial knockouts within the major deleted region differed between the two hosts. Knockout of the transfer machinery alone (pKP33ΔT4SS) resulted in significantly higher growth rates in both strains compared to the ancestral pKP33 plasmid (two-sample *t*-test: Ec37: $P < 0.001$, $n = 32$; Ec38: $P < 0.001$, $n = 32$) that were not distinguishable from the 9–30 kb knockout mutants (two-sample *t*-test: Ec37: $P = 0.115$, $n = 32$; Ec38: $P = 0.20$, $n = 32$). This indicates that the T4SS did indeed impose a cost on the naïve *E. coli* strains (fig. 9). Surprisingly, targeted knockout of the remaining region (pKP33Δ9-17 kb), leaving the T4SS region intact, also increased the growth rate of both *E. coli* strains and completely ameliorated the cost in Ec37 (one-sample *t*-test: $P = 0.388$, $n = 16$) but not in Ec38 (one-sample *t*-test: $P = < 0.001$, $n = 24$) where the T4SS knockout was more beneficial (two-sample *t*-test: $P = 0.0087$, $n = 40$) compared to the 9–17 kb deletion in this strain. See supplementary table S5, Supplementary Material online for an overview of the statistical comparisons made in this section.

Taken together, this implies that the reduction in growth rate imposed by the wild type pKP33 plasmid cannot be attributed to a single factor of the T4SS nor the remaining part of the 25 kb region alone, and that the exact nature of the cost varies between closely related strains.

Discussion

Conjugative plasmids are important mediators of horizontal gene transfer, enabling the direct exchange of multiple genes to allow rapid adaptation of bacteria to dynamic environments (Norman et al. 2009; Tamminen et al. 2012; Soucy et al. 2015). While a plasmid might replicate in, and transfer between, a diverse set of hosts, little is known about the factors determining the long-term persistence of endemic antibiotic resistance plasmids within their compatible range of host species (De Gelder et al. 2007). Here we simulate the scenario of novel host invasion of three pathogenic *E. coli* and *K. pneumoniae* isolates by a large clinically relevant multidrug resistance plasmid. Such transfer events are likely to take place in dense multispecies communities such as the human intestinal microbiota, supplying potential pathogens with a wide array of antibiotic resistance factors (Shoemaker et al. 2001; Sommer et al. 2009; Sommer and Dantas 2011).

We demonstrate that the host genetic background substantially influences initial plasmid cost and stability. The pKP33 plasmid was highly stable in its original Kp33 *K. pneumoniae* host, where the plasmid was maintained throughout 280 generations of culturing in the absence of selection (see supplementary fig. S1, Supplementary Material online). In contrast, pKP33 displayed a high degree of instability in the

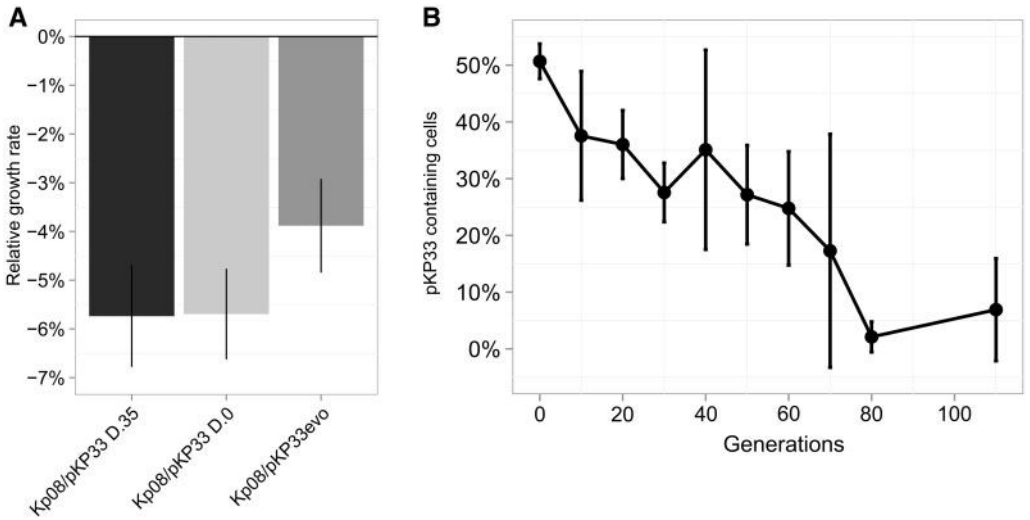


Fig. 8. The ancestral pKP33 as well as pKP33 evolved in *Escherichia coli* (pKP33evo) were transferred to the naïve Kp08 *Klebsiella pneumoniae* strain. (*A*) The exponential growth rate was measured for 24 individual colonies of Kp08 carrying the ancestral plasmid before (Day 0) and after (Day 35) 35 days of serial passaging in plasmid selective medium. These were compared to Kp08 carrying pKP33evo, transferred from the evolved *E. coli*. The growth rate improvement of Kp08/pKP33evo compared to Kp08/pKP33 Day 0 was borderline significant ($P = 0.063$). (*B*) Direct competition of Kp08 carrying the wild-type pKP33 against the Kp08 carrying the *Escherichia coli* evolved pKP33evo. Kp08/pKP33evo and Kp08/pKP33 Day 0 were equally mixed and serially transferred under antibiotic selection to select for both plasmids. The ratio of pKP33 to pKP33evo was quantified throughout the experiment by spotting on TMP and cefotaxime containing agar plates. Error bars depict the standard deviation of four replicate experiments.

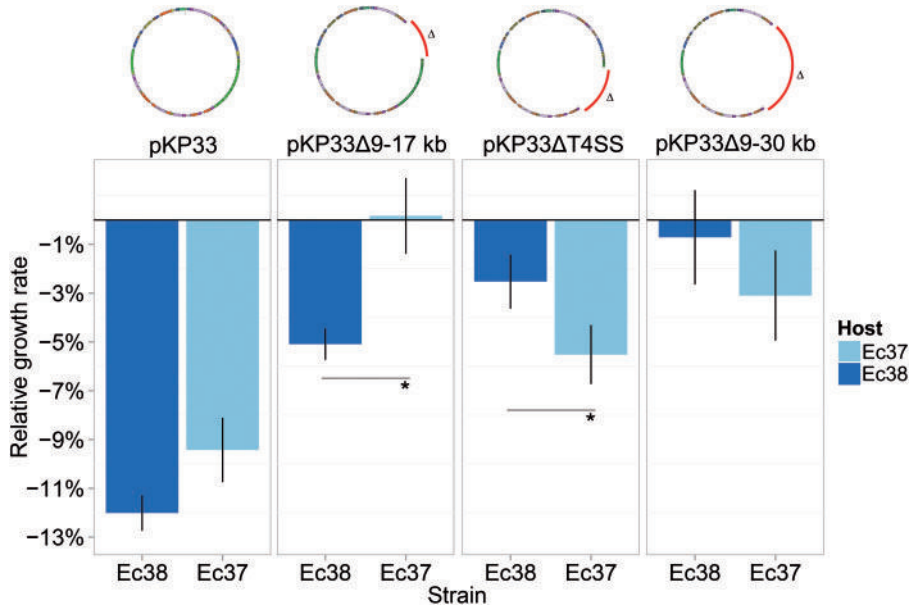


Fig. 9. Growth rate effects of targeted pKP33 knockouts in ancestral hosts. Engineered deletion mutants of pKP33 were generated in the ancestral hosts and the growth rate of the resulting variants was measured and depicted relative to the host without pKP33. The 19–30 kb part of pKP33 was deleted (pKP33ΔT4SS) comprising the conjugational transfer region and compared to deletion of the adjacent 9–17 kb (pKP33Δ9-17 kb) as well as a knockout of both regions (pKP33Δ9-30 kb). The top panel illustrates the position of the deleted parts (red) in pKP33. Error bars depict the 95% confidence interval of the mean measurements performed on at least 16 individual clones. A significant difference is indicated by asterisks (two-sample *t*-test, $P < 0.05$).

novel hosts, where it was lost in 50% of the population after 20, 48, and 100 generations of growth for Ec37, Ec38, and Kp08, respectively (fig. 4B).

Confirming previous observations of plasmid evolution (Heuer et al. 2007; De Gelder et al. 2008; San Millan et al. 2014; Harrison et al. 2015; Loftie-Eaton et al. 2015) we show that adaptive evolution in conditions selecting for plasmid carriage can significantly increase plasmid stability by reducing the fitness cost of an initially costly plasmid–host association. In addition, we show that nature of these

costs and the ability to adapt depends on the host background.

Sequencing of the evolved *E. coli* populations showed that the reduced fitness cost of pKP33evo was achieved by deleting a 25 kb plasmid region containing the T4SS involved in conjugational transfer; an event that was clearly dictated by the positioning of IS*26* elements in the plasmid backbone. Conjugation is considered costly in terms of vertical transfer (Turner et al. 1998), and targeted knockout of the main transfer machinery was indeed beneficial in terms of growth rate for both *E. coli* strains. Although liquid setups are commonly used in experimental evolution to rapidly assess a high number of cell divisions in a controlled environment, we note that such a setup might overemphasize the trade-offs observed here, because the conjugational transfer machinery of IncN plasmids works poorly in liquid environments (see supplementary table S3, Supplementary Material online). By measuring the conjugational transfer rates of pKP33 for all strains in a solid-surface setup we were able assess the potential of conjugative transfer to compensate the observed plasmid loss in a structured environment compared to our liquid setup and evaluate these values against the theoretical minimum requirement for parasitic maintenance of pKP33 in each strain (see supplementary table S3, Supplementary Material online). While pKP33 does show notably higher conjugation rates when surface associated, these rates are still orders of magnitude lower than the theoretical minimum rates required for parasitic plasmid maintenance in all naïve isolates (see supplementary table S3, Supplementary Material online), rendering the conjugational machinery disadvantages in these strains, even in conditions allowing high transfer rates.

Although it is unlikely that conjugation alone will maintain pKP33 even in dense, rapidly growing, homogenous populations similar to the ones investigated here, we note that there might be other selective advantages of carrying a conjugational transfer system in natural environments, even if transfer rates are low. We demonstrate that the original Kp33 host can indeed maintain a plasmid for much longer than even closely related naïve hosts. As suggested by Bergstrom et al., the ability to "sample" different host backgrounds for compatibility in diverse bacterial populations increases the opportunity for a stable plasmid–host relation, which along with other benefits of the T4SS such as increased biofilm formation, might improve long-term plasmid survival (Bergstrom et al. 2000; Ghigo 2001).

Surprisingly, the pKP33 knockout mutants lacking the 9–17 kb part whilst retaining the T4SS showed an equal or higher benefit in terms of growth rate compared to the pKP33ΔT4SS mutant, indicating that the cost is not exclusive to the T4SS but stems from multiple loci. While these costs displayed an additive nature, the relative contribution of each plasmid region varied between the two closely related *E. coli* UTI isolates (fig. 9, table 1 and supplementary table S2, Supplementary Material online). Whereas the 25 kb region of pKP33 had a half-life of roughly 60 generations in both *E. coli* strains evolved under plasmid selection, Kp08/pKP33 as well as the original Kp33/pKP33 strain retained the intact plasmid backbone throughout the evolution experiment. These observations emphasize the role of host genetic background in plasmid–host adaptation and highlight the potential role of stable plasmid donors as reservoirs in long-term plasmid survival. This is an important realization because it underlines the irregular fitness landscape navigated by transmissible plasmids in heterogeneous populations; a landscape that is eventually evened out during periods of selection for plasmid-borne traits, thus increasing plasmid frequency and opportunity for dissemination and adaptation towards new hosts. Interestingly, the *E. coli* evolved plasmid was also beneficial in Kp08 and although selection for compensatory mutations is expected to be weaker in Kp08 than for *E. coli* due to the lower cost of plasmid carriage (Qi et al. 2016), the evolved plasmid dominates the population within 50 generations of direct competition against Kp08 carrying the ancestral pKP33 plasmid (fig. 8B). Despite this potential competitive benefit of the large plasmid deletion in Kp08, restructuring was never observed in Kp08 nor in the original Kp33 host (fig. 6B), suggesting that the *E. coli* host were better catalysts of these adaptations than *K. pneumoniae*.

While it is reasonable to assume that the majority of intramolecular restructuring events lead to deleterious outcomes, we show that these events can indeed provide important plasticity for natural selection to work on. There are indications that the frequency at which recombination events occur can vary substantially within bacterial species. For example, Rodríguez-Beltrán et al. show that there is a tendency towards higher frequencies of homologous recombination in uropathogenic *E. coli* strains compared to nonpathogenic *E. coli* strains (Rodríguez-Beltrán and Tourret 2015). This has been suggested to increase adaptability in their native environment and such differences in mutational events, and in particular homologous recombination, could explain the differential evolvability of the *K. pneumoniae* strain compared to *E. coli* UTI isolates. These frequencies of recombination are suggested to be much higher ($>10^{-4}$) than point mutations ($\sim10^{-8}$), which might explain why we did not observe any SNPs able to compensate the plasmid imposed costs. Although we can only speculate on the mechanistic basis for these differences, they seem to correlate with the number of virulence genes and could also be influenced by host global regulators known to target horizontally acquired DNA and influence transposition frequency (Shiga et al. 2001; Doyle et al. 2007; Rodríguez-Beltrán and Tourret 2015). However, the cause is likely multifactorial and further investigations are needed to elucidate the exact nature of these differences.

As demonstrated here, the advancement of long read sequencing technologies aids greatly in the elucidation of the complex repetitive plasmid patterns attributed to abundant IS elements and it has become clear that IS elements, IS*26* in particular, play a major role in the organization of naturally occurring plasmids implicated in health-care associated outbreaks (Conlan et al. 2014; Harmer et al. 2014; He et al. 2015). Whereas it was recently demonstrated that transposition of beneficial genes into the plasmid backbone can improve stability (Loftie-Eaton et al. 2015), we are the first to describe IS mediated deletions as a driving force of plasmid

persistence in novel hosts. Whereas sacrificing important traits such as the conjugational transfer machinery can seem like an evolutionary dead-end strategy, such rapid recombination mediated restructuring might confer an advantage in certain environments (Labat et al. 2005). For a uropathogenic strain residing in the liquid environment of the urine bladder, getting rid of costly transfer machinery to increase growth rate can be a beneficial survival strategy (Bradley et al. 1980; Gordon and Riley 1992; Nilsson et al. 2003).

While domestication of costly plasmid traits seems like a common theme in plasmid–host evolution, the means of cost amelioration are diverse (Modi et al. 1991; Dahlberg and Chao 2003; Doyle et al. 2007; De Gelder et al. 2008; San Millan et al. 2014; Harrison et al. 2015). Recent work shows that the cost associated with carriage of a large conjugative plasmid in *Pseudomonas fluorescens* can be compensated by translational down-regulation attained via chromosomal mutations alone (Harrison et al. 2015). In contrast, we show that IS*26* mediated deletion of costly plasmid genes can provide similar benefits.

It is clear that strain specific factors can be highly influential, and extrapolating observations from single-host experiments to general plasmid behavior in clinical isolates is not a trivial task (De Gelder et al. 2007). The differences in stability observed for pKP33 in the *K. pneumoniae* and *E. coli* strains investigated here, suggests that transfer well within the expected natural host range of a plasmid might be more difficult than generally assumed. Although the initial tolerance, in terms of cost and stability, does seem to correlate with the relatedness of the strains to the original plasmid host (see supplementary table S2, Supplementary Material online), predicting the potential for stability improvements and long-term survival is not straight forward.

While species dependent constrains have been located to replication proteins in a nonconjugative broad host-range plasmid (Sota et al. 2010), our results suggest that dispensable factors of the plasmid backbone can impose similar limitations to the long-term host range of conjugative plasmids. Although chromosomal mutations as well as the presence of native plasmids in the naïve strains might have influenced the cost of pKP33 carriage, we expect these effects to be minimal compared to the major plasmid deletion event. Chromosomal mutations occurred almost exclusively in genes associated with virulence. Such genes are likely acquired by horizontal gene transfer and might be involved in the host specific cost patterns observed; possibly through regulatory interference or antagonistic interactions between for example membrane associated components of the deleted plasmid region (Doyle et al. 2007; San Millan et al. 2015). Although we did not observe any correlation between pKP33 carriage and retention of native plasmids, and we did not find any significant functional overlap in plasmid carried genes when compared to pKP33, we cannot out rule that the native plasmids of the strains have influenced plasmid cost and evolutionary outcomes (see supplementary material S1, Supplementary Material online). We acknowledge that such minor effects on fitness are hard to detect given the relatively low resolution of the growth measurements used here, but might be detectable by more sensitive fitness assays such as direct competition.

Although IS mediated cost amelioration effectively improves plasmid persistence, intramolecular restructuring events might come at the expense of conditionally useful components such as genes involved in antibiotic resistance or horizontal transfer, that might ultimately limit plasmid dissemination. While the evolved plasmid kept most of its resistance determinants, the trimethoprim resistance gene *folA* was consistently deleted along with the T4SS machinery involved in conjugation. These deletion patterns suggest that nonconjugative plasmids can evolve as a result of a strain dependent selection process dictated by transfer associated costs or collateral deletions to improve short term fitness at the expense of future niche expansion potential.

Our data suggests that while the selective forces against plasmid carriage can vary between even closely related isolates, the evolutionary solution might be more general and even broadly beneficial in strains to which genetic resolution is not immediately accessible.

Although the radical restructuring observed here ultimately narrows the plasmid host range by constraining its dissemination, such ongoing dynamics might help explain the dominance of nonconjugative plasmids in sequence databases (Smillie et al. 2010; Shintani et al. 2015). Further investigations of the mechanisms underlying plasmid persistence and the role of transposable elements herein are important to understand and prevent our current epidemic of multidrug resistance. Such studies shall preferentially be carried out on relevant specimen of clinical importance to directly improve our understanding of the barriers and opportunities implicated in successful plasmid dissemination that will provide us with the knowledge necessary to control and predict the rapid emergence of multidrug resistant pathogens.

Materials and Methods

Bacterial Strains, Plasmids, and Culture Conditions

The *E. coli* and *K. pneumoniae* strains used in this study were isolated from urinary tract and blood infections by the Department of Clinical Microbiology at Hvidovre Hospital, Denmark. The antibiotic resistance profile of the strains was evaluated by broth dilution antibiotic susceptibility testing (Wiegand et al. 2008) and summarized along with genotypic information in table 1 and supplementary table S1, Supplementary Material online. Culturing was generally done with shaking in lysogeny broth (LB) medium at 37°C. M9 minimal medium lacking leucine was used for counterselection of donors in the conjugation experiments. Cefotaxime (2 μg/ml) was added to ensure stable plasmid maintenance during culturing of strains harboring ESBL plasmids and TMP (16 μg/ml) was added when appropriate to prevent deletions of the 25 kb plasmid region containing the *folA* gene.

Plasmid Transfer and Conjugation Assays

Large plasmids were transferred to clinical isolates via either conjugation or direct electroporation when possible. For transfer by conjugation, the *TOP10* strain (Life Technologies, USA) was used as donor strain due to its high competence and leucine auxotrophy. Following incubation with the donor strain, cells were scraped of the agar surface, washed in isotonic salt water and plated on M9 minimal medium agar plates lacking leucine but containing cefotaxime for plasmid selection. Transconjugants were verified as β-galactosidase positive (blue colonies) on *X-gal* containing LB plates. For all conjugation experiments, the donor and recipient strains were grown with appropriate antibiotics to midexponential phase and washed in isotonic salt water before mixing 1:1. Conjugation was performed in either liquid LB medium or on a solid LB agar surface incubated without shaking at 37°C. For transfer rate quantification, conjugation was carried out for 2 h and the cells were plated on selective plates to quantify donors, recipients, and transconjugants.

Plasmid Evolution and Stability Experiments

Five parallel lineages of each plasmid-carrying strain were passaged once daily, in selective medium to ensure plasmid inheritance during the evolution experiment, and subsequently without selection to assess plasmid stability. Serial passaging was performed in 96-well plates containing 150 µl LB medium. Each day of the experiment, 1 µl culture was transferred to the corresponding well of a new plate by pin replication. Plates were incubated at 37°C and subjected to medium shaking at 400 rpm. The plasmid selection (cefuroxime, 16 µg/ml) used for the evolution experiment did not reduce the growth rate of plasmid bearing strains compared to growth in LB without antibiotics.

Growth Rate Measurements

Growth rate was measured as the maximum increase in optical density (OD) over time during exponential growth. OD measurements were conducted in 96-well plates containing 150 µl LB/well by the *ELx808* plate reader (BioTek, USA). Breathe-Easy (Sigma-Aldrich) film was applied to minimize evaporation during measurements. OD at 600 nms was measured with 5 min intervals for maximum 16 h and incubated with medium shaking at 37°C between measurements.

Colonies were picked into a pre-inoculation plate and were grown for 2–3 h with shaking at 37°C before inoculation of the final measurement plate.

Pairwise Competition

Two O/N cultures were diluted to the same OD and mixing equally in LB medium. Cefotaxime (1 µg/ml) was added to prevent plasmid loss. The competition was carried out in 1.5 ml cultures and a volume of 1.5 µl was transferred to a fresh well every 24 h. From OD measurements the number of generations was estimated to ~10 generations/day. The ratio of the competitors was determined as the fraction of colonies on TMP agar plates compared to plates containing cefotaxime (performed as in "Plasmid deletion and loss quantification").

Plasmid Deletion and Loss Quantification

Plasmid loss was measured as sensitivity to cefotaxime; assuming loss of the CTX-M-15 gene. While quantification was done at the population level, periodic verification of plasmid-free colonies was carried out by testing for the presence of the remaining plasmid carried resistance determinants. Deletion of the 9–30 kb plasmid region was measured as sensitivity to TMP, indicating the absence of the *folA* TMP resistance gene. Frozen culturing plates were thawed completely and 10-fold dilutions (10^{-1}–10^{-8}) were made in isotonic salt water. After thorough mixing of each dilution, 5 µl spots were placed on selective and nonselective agar plates (2 µg/ml cefotaxime or 16 µg/ml TMP). After absorption of the liquid, plates were incubated O/N at 30°C to ensure countable colony sizes. CFUs of the lowest countable dilution (<70 CFU) were quantified and the ratio of resistant to total cells was calculated.

Mathematical Modelling of Plasmid Stability

A mathematical model derived in Proctor (1994) was fitted to plasmid loss data using the *nls2* package in *R* (version 3.0.1). The model assumes that the population dynamics of a plasmid bearing population can be described by differential equations modelling the dynamics of plasmid containing $dPc/dt = (\gamma_{Pc} \cdot N_{Pc}) - (\mu \cdot N_{Pc})$, and plasmid-free cells $dPf/dt = (\gamma_{Pf} \cdot N_{Pf}) + (\mu \cdot N_{Pc})$ respectively.

Where *Pc* denotes plasmid containing cells and *Pf* denotes plasmid-free cells. Here, γ is the growth rate and μ is the segregational plasmid-loss rate. An equation describing the plasmid-free fraction as a function of time can be derived:

$$F_{pc} = \frac{(\mu+\rho)\cdot F_{pc,t0}}{(\mu+\rho\cdot(1-F_{pc,t0}))\cdot e^{(\mu+\rho)\cdot t} + \rho\cdot F_{pc,t0}} \quad (1)$$

Here, the starting number of plasmid containing cells is $F_{pc,t0}$ and the differential growth rate $(\gamma_{Pf} - \gamma_{Pc})$ is summarized in the parameter ρ.

Whole Genome and Plasmid Sequencing

Genomes and whole population DNA samples were extracted using the *DNeasy* Blood & Tissue Kit from *QIAGEN* (QIAGEN, Netherlands). Sample libraries for Illumina sequencing were prepared using the *TrueSeq Nano* and *Nextera XT* kits (Illumina, USA). All reference genomes and plasmids were prepared using the *TrueSeq Nano* kit and mechanical shearing. The remaining sequencing libraries were prepared using the *Nextera XT* employing enzymatic fragmentation. Sample concentrations were measured on a *Qubit* fluorometer (Life Technologies, USA) and analysed with the *Agilent 2100 Bioanalyzer* (Agilent, USA) for fragment length distributions. Libraries were sequenced paired-end on a *MiSeq* sequencer (Illumina, USA) with sample amounts ensuring coverage of >30X. In addition, pKP33 was sequences using the *Pacific Biosciences RS II* single molecule real time sequencing technology. PacBio library preparation and sequencing was performed by The Norwegian High-Throughput Sequencing Centre (NSC)

(Oslo, Norway). Sequenced genomes and short read data from the pKP33 evolved populations are deposited within the BioProject: PRJNA325878.

Sequence Analysis

All sequencing data was processed and analysed using the *CLC Genomics Workbench* software from *CLC Bio* (QIAGEN, Netherlands). Further analysis and preparation of final graphics was conducted in *R* (version 3.0.1). Annotation was done using the *RAST* server (Aziz et al. 2008). All annotations were manually inspected and updated via BLAST searches if deemed necessary. Reads from sequenced evolved genomes were mapped to reference genomes via the "Map reads to reference" feature in *CLC*. SNP's and small INDELS were detected automatically using the "Quality based variant detection" tool. Read mappings were manually inspected for each contig to identify larger INDELS. Variants resulting from ambiguous read-mappings were excluded from the variant analysis.

BLAST Analysis and Sequence Comparison Using *GView*

Consensus sequences of the evolved plasmids were extracted from read mappings to pKP33 and the cut-off for inclusion was set to 5× coverage, corresponding to approximately 10% of the average read coverage. These were blasted against the pKP33 plasmid reference using the *GView* web-server available at: https://server.gview.ca/, last accessed August 8, 2016. The blastn algorithm was used to generate a circular BLAST atlas using the following settings: *e* value cut-off was set to 10^{-10}, alignment length cut-off to 100 bp, percent identity cut-off to 85% and the genetic code to "Bacterial and Plant Plastid". The layout of the output was edited in the *GView* Java stand-alone application accessible from the results page.

Generating Plasmid Knock Out Mutants

Recombineering in the *E. coli DY329* strain was performed as previously described (Yu et al. 2000). Targeted deletions were made by introducing a chloramphenicol resistance cassette from the *pKD3* vector (Datsenko and Wanner 2000) into the pKP33 plasmid backbone. The T4SS region was targeted using primers with the following flanking regions: 5′-TTAAA TCTGCAATCAACAGAAGATAGTGAGTAAGGAGAAAGT ATGACCAC-3′ and 5′-AGAAATATAGCCTGCGTCAATCG TTTCTGCCGTGAGGGTACCGCTTTCCC -3′; deleting the approximately 10 kb part of pKP33 containing the main conjugational transfer region. A plasmid mutant with the 9–17.6 kb (pKP33Δ9-17 kb) region deleted was created using the following homologous regions: 5′-CTTCCATTCCGCC CATTTTTAGAAAATTTCGTGTCCATGCGATCAGGTTA-3′ and 5′-GATTTACGTGCATAGCCGATTTTCATTCTTTTCT CGCTAATTAGTTATGG-3′. A combination of both deletions from 9 to 30 kb was made using: 5′-GATTTACGTGCATA GCCGATTTTCATTCTTTTCTCGCTAATTAGTTATGG-3′ and 5′-AGAAATATAGCCTGCGTCAATCGTTTCTGCCGTG AGGGTACCGCTTTCCC-3′ as homology regions. All deletions were confirmed by PCR and subsequent *Sanger* sequencing (Macrogen, Korea).

Supplementary Material

Supplementary material S1, tables S1–S5, and figures S1–S3, are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Lundbeck Foundation and the Novo Nordisk Foundation. We thank Anna Koza for sequencing assistance and Bruce Levin for helpful comments on the manuscript.

References

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

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

Bouma J, Lenski R. 1988. Evolution of a bacteria/plasmid association. *Nature* 335:

Bradley DE, Taylor DE, Cohen DR. 1980. Specification of surface mating systems among conjugative drug resistance plasmids in *Escherichia coli* K-12. *J Bacteriol*. 143:1466–1470.

Bush K, Fisher JF. 2011. Epidemiological expansion, structural studies, and clinical challenges of new β-lactamases from Gram-negative bacteria. *Annu Rev Microbiol*. 65:455–478.

Carattoli A. 2009. Resistance plasmid families in Enterobacteriaceae. *Antimicrob Agents Chemother*. 53:2227–2238.

Carattoli A. 2013. Plasmids and the spread of resistance. *Int J Med Microbiol*. 303:298–304.

Conlan S, Thomas PJ, Deming C, Park M, Lau AF, Dekker JP, Snitkin ES, Clark TA, Luong K, Song Y, et al. 2014. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Sci Transl Med*. 6:254ra126.

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.

Curiao T, Canton R, Garcillan-Barcia MP, De La Cruz F, Baquero F, Coque TM. 2011. Association of composite IS26-sul3 elements with highly transmissible IncI1 plasmids in extended-spectrum-β-lactamase-producing *Escherichia coli* clones from humans. *Antimicrob Agents Chemother*. 55:2451–2457.

Dahlberg C, Chao L. 2003. Amelioration of the cost of conjugative plasmid carriage in *Escherichia coli* K12. *Genetics* 165:1641–1649.

Datsenko K, Wanner BL. 2000. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci U S A*. 97:6640–6645.

Davies J, Davies D. 2010. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev*. 74:417–433.

De Gelder L, Ponciano JM, Joyce P, Top EM. 2007. Stability of a promiscuous plasmid in different hosts: no guarantee for a long-term relationship. *Microbiology* 153:452–463.

De Gelder L, Williams JJ, Ponciano JM, Sota M, Top EM. 2008. Adaptive plasmid evolution results in host-range expansion of a broad-host-range plasmid. *Genetics* 178:2179–2190.

Dhillon RH-P, Clark J. 2012. ESBLs: a clear and present danger? *Crit Care Res Pract*. 2012:625170.

Dionisio F, Conceição IC, Marques CR, Fernandes L, Gordo I. 2005. The evolution of a conjugative plasmid and its ability to increase bacterial fitness. *Biol Lett*. 1:250–252.

Doyle M, Fookes M, Ivens A, Mangan MW, Wain J, Dorman CJ. 2007. An H-NS-like stealth protein aids horizontal DNA transmission in bacteria. *Science* 315:251–252.

Eikmeyer F, Hadiati A, Szczepanowski R, Wibberg D, Schneiker-Bekel S, Rogers LM, Brown CJ, Top EM, Pühler A, Schlüter A. 2012.

The complete genome sequences of four new IncN plasmids from wastewater treatment plant effluent provide new insights into IncN plasmid diversity and evolution. *Plasmid* 68:13–24.
Fernandez-Lopez R, del Campo I, Revilla C, Cuevas A, de la Cruz F. 2014. Negative feedback and transcriptional overshooting in a regulatory network for horizontal gene transfer. *PLoS Genet*. 10.
Frost LS, Leplae R, Summers AO, Toussaint A. 2005. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol*. 3:722–732.
Ghigo JM. 2001. Natural conjugative plasmids induce bacterial biofilm development. *Nature* 412:442–445.
Gordon DM, Riley M. 1992. A theoretical and experimental analysis of bacterial growth in the bladder. *Mol Microbiol*. 6:555–562.
Guynet C, Cuevas A, Moncalián G, de la Cruz F. 2011. The stb operon balances the requirements for vegetative stability and conjugative transfer of plasmid R388. *PLoS Genet*. 7:e1002073.
Harmer CJ, Moran R, Hall RM, Harmer CJ, Moran R, Hall RM. 2014. Movement of IS 26 -associated antibiotic resistance genes occurs via a translocatable unit that includes a single IS 26 and preferentially inserts adjacent to another IS 26. 5:1–9.
Harrison E, Brockhurst M. 2012. Plasmid-mediated horizontal gene transfer is a coevolutionary process. *Trends Microbiol*. 20:262–267.
Harrison E, Guymer D, Spiers AJ, Paterson S, Brockhurst MA. 2015. Parallel compensatory evolution stabilizes plasmids across the parasitism-mutualism continuum. *Curr Biol*. 25:2034–2039.
He S, Hickman B, Varani AM, Siguier P, Chandler M, Dekker JP. 2015. Insertion sequence IS 26 reorganizes plasmids in clinically isolated multidrug-resistant bacteria by replicative transposition. *MBio* 6:1–14.
Heuer H, Fox RE, Top EM. 2007. Frequent conjugative transfer accelerates adaptation of a broad-host-range plasmid to an unfavorable *Pseudomonas putida* host. *FEMS Microbiol Ecol*. 59:738–748.
Jain A, Srivastava P. 2013. Broad host range plasmids. *FEMS Microbiol Lett*. 348:87–96.
Labat F, Pradillon O, Garry L, Peuchmaur M, Fantin B, Denamur E. 2005. Mutator phenotype confers advantage in *Escherichia coli* chronic urinary tract infection pathogenesis. *FEMS Immunol Med Microbiol*. 44:317–321.
Loftie-Eaton W, Yano H, Burleigh S, Simmons RS, Hughes JM, Rogers LM, Hunter SS, Settles ML, Forney LJ, Ponciano JM, et al. 2015. Evolutionary paths that expand plasmid host-range: implications for spread of antibiotic resistance. *Mol Biol Evol*. 33:885–897.
Mazodier P, Davies J. 1991. Gene transfer between distantly related bacteria. *Annu Rev Genet*. 25:147–171.
Miriagou V, Papagiannitsis CC, Kotsakis SD, Loli a, Tzelepi E, Legakis NJ, Tzouvelekis LS. 2010. Sequence of pNL194, a 79.3-kilobase IncN plasmid carrying the blaVIM-1 metallo-beta-lactamase gene in *Klebsiella pneumoniae*. *Antimicrob Agents Chemother*. 54:4497–4502.
Modi RI, Wilke CM, Rosenzweig RF, Adams J. 1991. Plasmid macroevolution: selection of deletions during adaptation in a nutrient-limited environment. *Genetica* 84:195–202.
Moré MI, Pohlman RF, Winans SC. 1996. Genes encoding the pKM101 conjugal mating pore are negatively regulated by the plasmid-encoded KorA and KorB proteins. *J Bacteriol*. 178:4392–4399.
Mruk I, Kobayashi I. 2014. To be or not to be: regulation of restriction–modification systems and other toxin–antitoxin systems. *Nucleic Acids Res*. 42:70–86.
Nilsson AI, Berg OG, Aspevall O, Andersson DI, Kahlmeter G. 2003. Biological costs and mechanisms of fosfomycin resistance in *Escherichia coli*. *Society* 47:2850–2858.
Norman A, Hansen LH, Sørensen SJ. 2009. Conjugative plasmids: vessels of the communal gene pool. *Philos Trans R Soc Lond B Biol Sci*. 364:2275–2289.
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.
Proctor GN. 1994. Mathematics of microbial plasmid instability and subsequent differential growth of plasmid-free and plasmid-containing cells, relevant to the analysis of experimental colony number data. *Plasmid* 32:101–130.
Qi Q, Toll-Riera M, Heilbron K, Preston GM, MacLean RC. 2016. The genomic basis of adaptation to the fitness cost of rifampicin resistance in *Pseudomonas aeruginosa*. *Proc R Soc B Biol Sci*. 283:20152452.
Rodríguez-Beltrán J, Tourret J. 2015. High recombinant frequency in extraintestinal pathogenic *Escherichia coli* strains. *Mol Biol Evol*. 32:1708–1716.
San Millan A, Peña-Miller R, Toll-Riera M, Halbert ZV, McLean R, Cooper BS, MacLean RC. 2014. Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. *Nat Commun*. 5:5208.
San Millan A, Toll-Riera M, Qi Q, MacLean RC. 2015. Interactions between horizontally acquired genes create a fitness cost in *Pseudomonas aeruginosa*. *Nat Commun*. 6:6845.
Shiga Y, Sekine Y, Kano Y, Ohtsubo E. 2001. Involvement of H-NS in transpositional recombination mediated by IS1. *J Bacteriol*. 183:2476–2484.
Shintani M, Sanchez ZK, Kimbara K. 2015. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Front Microbiol*. 6:1–16.
Shoemaker NB, Vlamakis H, Hayes K, Salyers AA. 2001. Evidence for extensive resistance gene transfer among Bacteroides spp. and among Bacteroides and other genera in the human colon. *Appl Environ Microbiol*. 67:561–568.
Simonsen L. 1991. The existence conditions for bacterial plasmids: theory and reality. *Microb Ecol*. 187–205.
Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. 2010. Mobility of plasmids. *Microbiol Mol Biol Rev*. 74:434–452.
Sommer MO, Dantas G. 2011. Antibiotics and the resistant microbiome. *Curr Opin Microbiol*. 14:556–563.
Sommer MO, Dantas G, Church GM. 2009. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* 325:1128–1131.
Sota M, Yano H, Hughes JM, Daughdrill GW, Abdo Z, Forney LJ, Top EM. 2010. Shifts in the host range of a promiscuous plasmid through parallel evolution of its replication initiation protein. *ISME J*. 4:1568–1580.
Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nat Rev Genet*. 16:472–482.
Tamminen M, Virta M, Fani R, Fondi M. 2012. Large-scale analysis of plasmid relationships through gene-sharing networks. *Mol Biol Evol*. 29:1225–1240.
Turner P, Cooper V, Lenski R. 1998. Tradeoff between horizontal and vertical modes of transmission in bacterial plasmids. *Evolution (N. Y.)* 52:315–329.
World Health Organization. 2014. Antimicrobial resistance: global report on surveillance 2014. Geneva: WHO Press. Available from: http://apps.who.int/iris/handle/10665/112642.
Wiegand I, Hilpert K, Hancock REW. 2008. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat Protoc*. 3:163–175.
Yu D, Ellis HM, Lee EC, Jenkins N, Copeland NG, Court DL. 2000. An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc Natl Acad Sci U S A*. 97:5978–5983.
Zahrl D, Wagner M, Bischof K, Koraimann G. 2006. Expression and assembly of a functional type IV secretion system elicit extracytoplasmic and cytoplasmic stress responses in *Escherichia coli*. *J Bacteriol*. 188:6611–6621.

Supplementary material

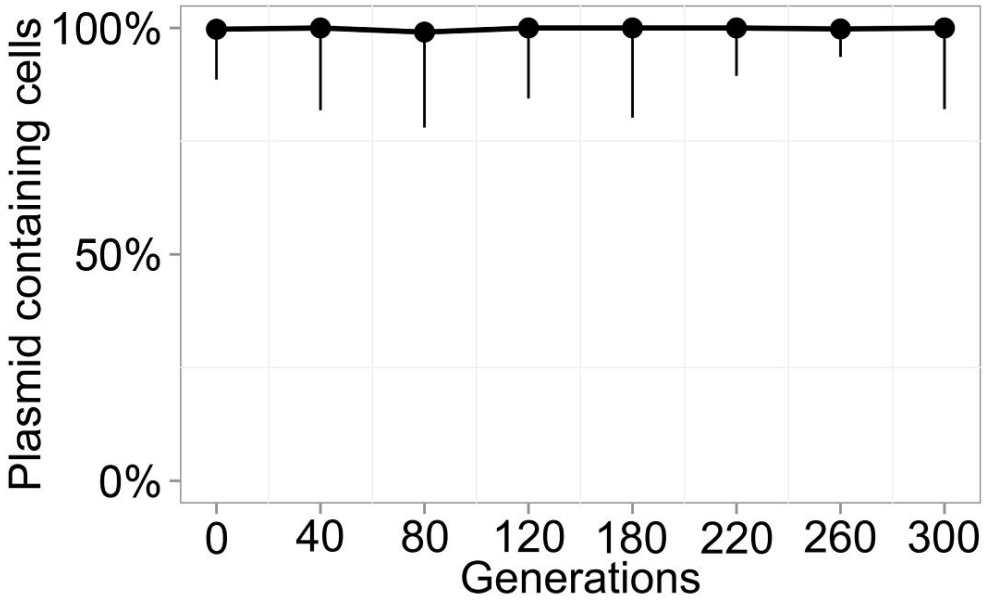
Supplementary table S1. Minimal inhibition concentrations selected antibiotics. The resistance phenotype of pKP33 was tested in a non-resistant strain background. Red numbers indicate increased tolerance compared to the plasmid-free host. In addition, the MIC values for the naïve clinical isolates carrying pKP33 or pKP33evo was tested for cefutaxime, trimethoprim and streptomycin.

These values were: >1024 µg/ml for cefotaxime for the evolved as well as the ancestral plasmid in all strains. Trimethoprim was >16 µg/ml for the ancestral plasmid in all strains and 2 µg/ml (same as the plasmid free strains) for the evolved plasmid in all isolates carrying the evolved plasmids. The MICs for streptomycin were 256 µg/ml for both the ancestral and evolved plasmid in Ec37 and Kp08, whereas a value of 512 µg/ml was measured for both plasmid variants in Ec38.

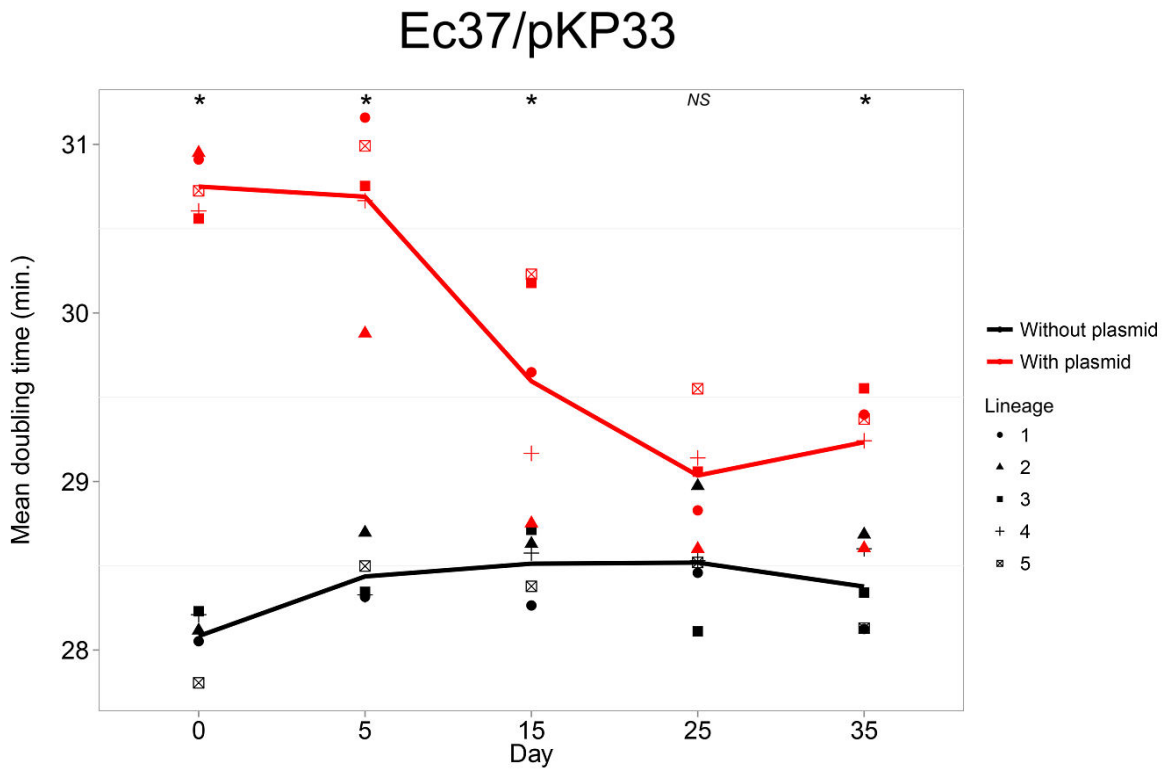
Drug	MIC value (µg/ml)
Cefotaxime	>16
Piperacillin	>128
Piperacillin+Tazobactam	128
Gentamicin	>32
Ciprofloxacin	>1
Trimethoprim	>16
Kanamycin	>50
Amikacin	>32
Streptomycin	>32
Tetracycline	1
Chloramphenicol	2
Meropenem	0.125

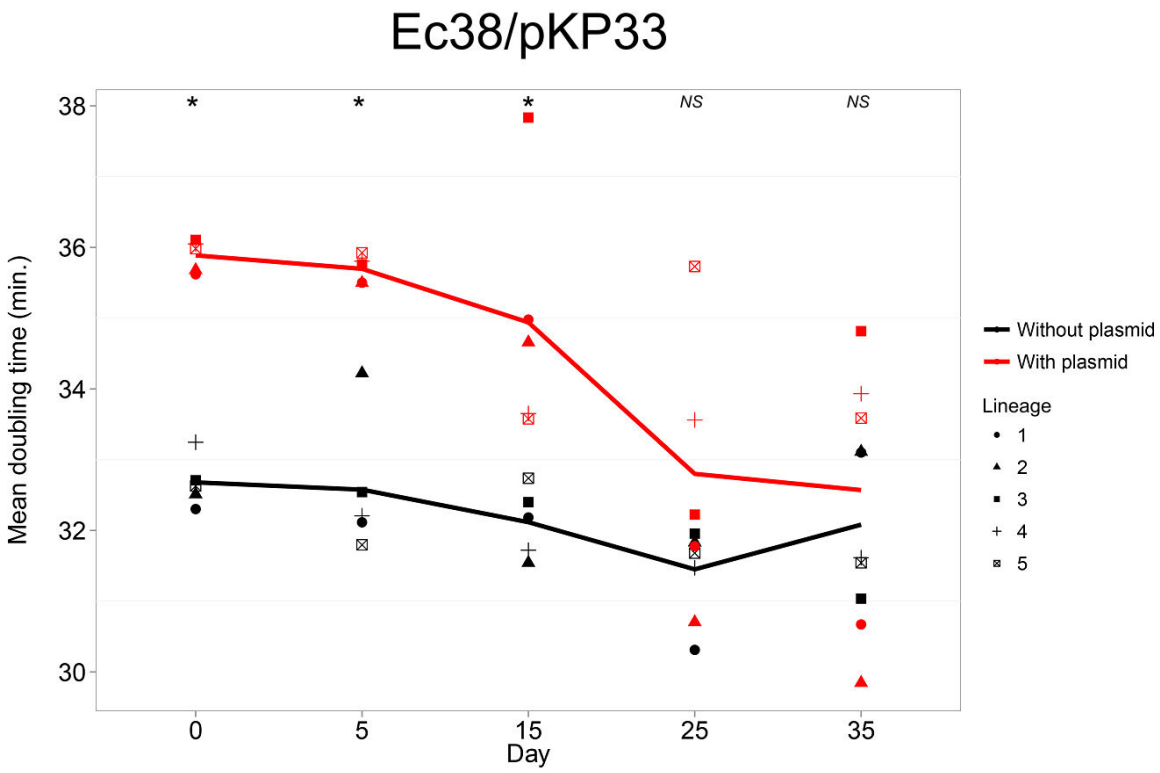
Supplementary fig. S1. Stability of pKP33 in the native KP33 host.

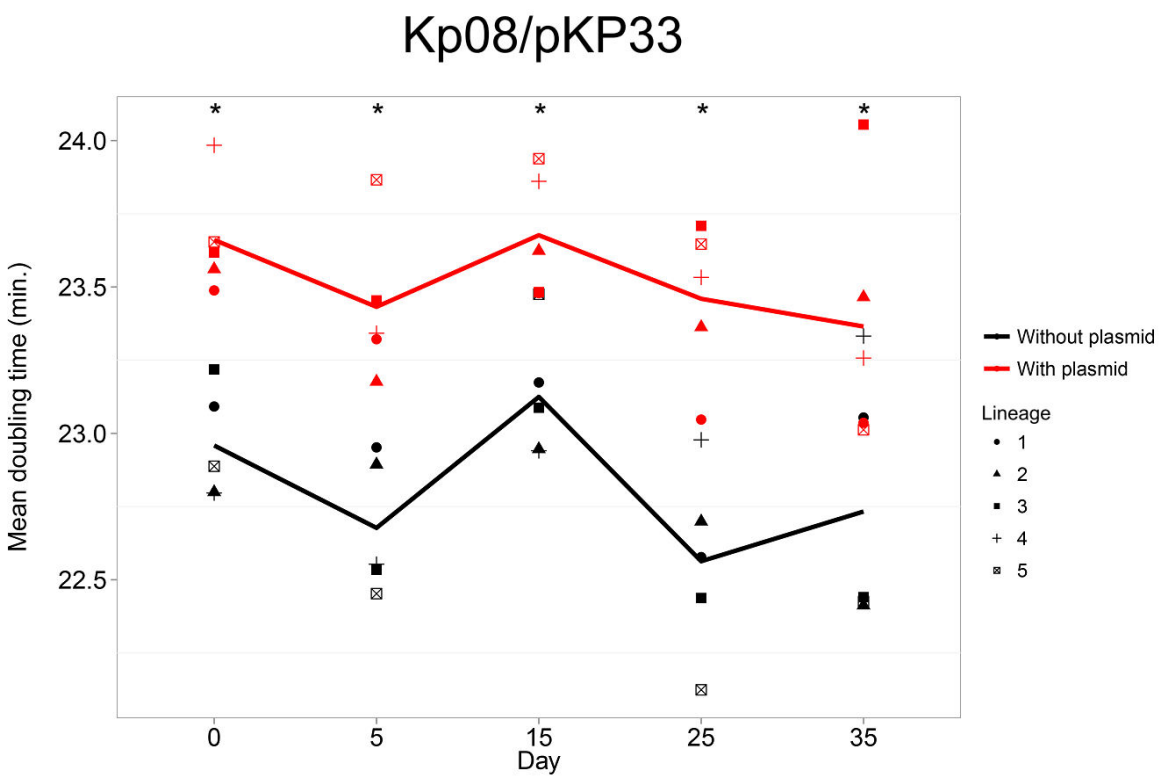
Stability of the pKP33 plasmid in the *K. pneumoniae* host from which it was initially isolated. Plasmid stability was quantified via spot assaying on plasmid selective and non-selective agar plates. The end-point was validated by streaking 100 colonies on selective medium (cefotaxime and Trimethoprim) to confirm 100% plasmid presence. The results are shown as the proportion of resistant to total cells during 35 days (280 generations) of serial transfer in non-selective medium. Error bars show standard deviation of five replicates and depict the variation involved in the measurements.



Supplementary fig. S2. Absolute growth rate measurements during adaptive evolution. Doubling times was calculated from measurements obtained after day 0, 5, 15, 25, 35 of the evolution experiment; corresponding to app. 0, 40, 120, 200, 280 bacterial generations. Each point represents the mean measurement of eight randomly picked clones from each lineage during evolution. Asterisks indicate a significant difference (*P < 0.05, n = 5, two-sample t-test) from the evolving plasmid-free lineages.




Ec38/pKP33
Mean doubling time (min.)
38
36
34
32
30
*
*
*
NS
NS
0
5
15
25
35
Day
Without plasmid
With plasmid
Lineage
1
2
3
4
5


Kp08/pKP33
Mean doubling time (min.)
24.0
23.5
23.0
22.5
*
*
*
*
*
0
5
15
25
35
Day
Without plasmid
With plasmid
Lineage
1
2
3
4
5

Supplementary table S2. Proteome comparison.

An *in silico* proteome comparison was performed using the CMG-Biotools software package (Vesth et al. 2013) to assess the number of shared protein families between the strains studied. Here a shared family is considered if the alignment is 50% identical and the length of the alignment is at least 50% of the longest protein. The table shows the number of shared protein families between each pair of strains (lower left corner) as well as the fraction calculated in per cent of the largest proteome in the comparison (top right corner).

	Ec37	**Ec38**	**Kp08**	**Kp33**	**Proteome size**
Ec37		80%	54%	56%	5001
Ec38	4004		54%	55%	4718
Kp08	2833	2812		83%	5253
Kp33	2814	2792	4346		5032

Supplementary table S3. Conjugation frequencies.

Conjugation experiments were performed in liquid LB (without shaking) as well as on a solid LB agar surface to estimate transfer rates (γ) using the methodology by Simonsen et al. 1990. According to Stewart and Levin's criterion (Levin and Stewart 1977):

$\gamma \widehat{N} > \alpha\omega + \tau$

the minimum requirement in terms of conjugative transfer rate (γ_{min}) for a plasmid to be parasitic depends on the population density ($\widehat{N}$), the fitness cost (α) of plasmid carriage, the growth rate (ω) and the rate of loss by segregation (τ). As we have estimated these parameters for all the naïve clinical isolates, we were able to approximate the minimal rate of conjugation needed to compensate the fitness cost and segregational loss rates observed for pKP33 in these strains (assuming that increased conjugation does not increase the fitness cost) and compare them to the actual transfer rate of pKP33 in each strain.

Strain	**Liquid transfer rate (γ_l)**	**Solid transfer rate (γ_s)**	**Mininum transfer rate γ_{min}**
Ec37/pKP33	5.03E-14	4.81E-13	1.29E-09
Ec38/pKP33	4.21E-13	1.35E-12	1.83E-09
Kp08/pKP33	7.17E-14	1.23E-11	1.01E-09
Kp33/pKP33	9.57E-12	2.19E-11	NA
Ec37/pKP33Δ9-17kb	2.23E-14	3.79E-12	1.40E-10
Ec38/pKP33Δ9-17kb	7.80E-14	6.23E-12	6.58E-10

Supplementary table S4. Best fit parameter values for a plasmid stability model of ancestral (day 0) and evolved (day 35) plasmid-host combinations.

A mathematic model assuming segregational loss (μ) and differential growth rate (ρ) as the main drivers of plasmid instability was fitted to plasmid loss rate obtained from either ancestral or evolved plasmid-host combinations (fig. 4*A*). Summery statistics of models fitted to plasmid loss-data from each lineage is shown. We were unable to fit the model to data from two lineages (ancestral Ec38/33 lineage 3 and evolved Ec37/33 lineage 4) due to high variation in the measurements or too few non-zero data points. The estimates for μ are relatively uncertain owing to the few data points obtained in the beginning of each datasets and negative estimates were set to 0. Here the estimate of μ is biased by the rapid outgrowth of plasmid bearing cells, making it hard for the model to distinguish between the two parameters. *P*-values were determined by a two-sample *t*-test comparing parameter values from day 0 and day 35 to assess the effect of evolution.

Strain	**Day**	**Segregation rate (μ)**	**SD (μ)**	***P*-value μ**	**Plasmid cost (ρ)**	**SD ρ**	***P*-value ρ**
Ec37/p33	0	0.0147	±0.026		8.3 %	±4.14 %	
Ec37/p33	35	0.0079	±0.0033	0.57	-0.3 %	±1.23 %	0.001
Ec38/p33	0	0.0008	±0.016		14.0 %	±2.85 %	
Ec38/p33	35	0.0008	±0.0061	0.99	0.4 %	±3.37 %	<0.0001
Kp08/p33	0	0.0060	±0.002		4.6 %	±2.94 %	
Kp08/p33	35	0.0067	±0.0033	0.68	1.6 %	±3.35 %	0.17

Supplementary S1

Discussion of genomic variants

Although plasmid adaptations had the dominant role in the improved persistence of pKP33 in *E. coli*, a minor or transient role of chromosomal mutations cannot be out ruled.

Most non-synonymous mutations detected in the remaining genome of the strains were found in genes annotated as virulence factors such as adhesions, fimbriae or iron scavenging proteins and could not be directly associated with plasmid stability. No mutations in non-coding regions of the genomes were consistently associated with all lineages or with any known regulatory elements e.g. promoter regions. Attenuation of virulence is often observed in asymptomatic urinary tract *E. coli* isolates where deletions or point mutations are often present in adhesion factors (Klemm et al. 2006; Roos et al. 2006). As virulence factors are redundant in our setup, and inactivation hereof might lead to growth rate improvements or perhaps a stronger association with the liquid phase due to decreased adhesion, attenuation is likely to impose a selective advantage (Roos and Ulett 2006). Furthermore, plasmid cost has been associated with excessive gene expression and interaction with horizontally acquired genes (Doyle et al. 2007; Harrison et al. 2015; San Millan et al. 2015). This, along with the presence of pKP33 encoded UmuC mutagenic DNA polymerase, might explain the increased tendency to mutate in virulence and phage-derived proteins in plasmid bearing lineages (Maor-Shoshani et al. 2000). Another interesting observation is that many chromosomal variants occur in genes encoding periplasmic or surface associated proteins (e.g. FtsK, PapC, KpsD and fimbrial proteins) which could be involved in antagonistic interactions with the membrane protruding T4SS or the putitative membrane protein STY461 of the deleted 25kb region in pKP33. If such interactions took place, the effect of chromosomal mutations must have been significantly lower than improvements attained from the deletion of plasmid genes, as the evolved plasmid alone is able to compensate the growth rate reduction to the same degree as the evolved plasmid-host combination (fig. 7). One consistent mutation, resulting in amino acid change from tyrosine to histidine, in the *ftsK* gene occurred in the chromosome of three out of five evolved Kp08 lineages. FtsK is a membrane located multi-domain protein involved in coordination of chromosomal segregation (Crozat et al. 2015). Considering the position of this variant in a variable non-essential linker domain and lack of correlation with stability or measureable growth rate improvements we do not assume an important role of this mutation in plasmid adaptation.

The indigenous plasmids of the naive isolates ranged from 7kb to 138kb in size and the IncFII of Ec37 as well as the IncFIC of Ec38 contained genes involved in conjugative transfer. The large (IncR of Kp08, IncFII of Ec37 and IncFIC of Ec38) plasmids contained genes associated with virulence such as toxin, siderophore and adhesions/fimbriae encoding genes. An all against all blast comparison at the protein level revealed that the two Kp08 plasmids (IncR and IncFIB(K)) were the only plasmids containing genes with similar function to those found in pKP33. Both plasmids contained the error-

prone polymerases UmuC and UmuD, which might lead to elevated mutation rates (Maor-Shoshani et al. 2000). Apart from these genes, no other e.g. similar toxin-antitoxin systems, IS*26* elements, or genes involved in gene regulation, that could have implications for the stability and evolution of pKP33, were found in the native plasmids of the naïve strains. While all Ec37 lineages, including the lineage evolved without pKP33, retained their native IncFII plasmid throughout the evolution experiment, the endogenous incF plasmids (IncFIC and IncFIB) of Ec38 were lost in all but one pKP33 evolved lineage, which retained the IncFIC plasmid. However, the loss of endogenous plasmids was not exclusively associated with the presence of pKP33, as the Ec38 lineage evolved without pKP33 had also lost both plasmids by the end of the experiment. None of the two native (IncFIB(K) and IncR) plasmids of Kp08 where lost in neither the pKP33 carrying lineages nor the pKP33 free control. For Ec38, we note that the native plasmids did not have a substantial effect on the growth rate upon pKP33 acquisition as the evolved plasmid-host combination (which had lost all its native plasmids) was indistinguishable from the ancestral strain (containing the native plasmids) transformed with the evolved pKP33 plasmid.

As for the remaining genomic content of these strains, we cannot rule out the possibility that factors such as titration of global regulators or antagonistic interference between virulence mediators could have influenced the stability and evolutionary outcome observed here.

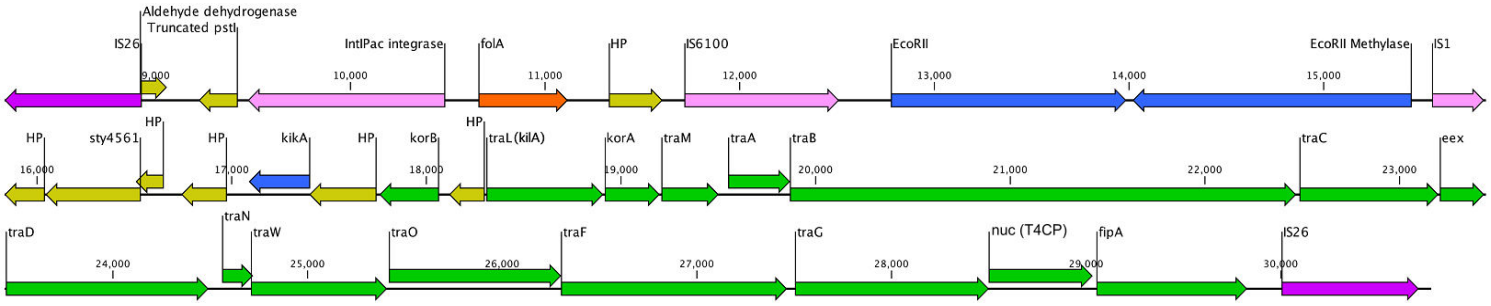
Supplementary table S5. Non-synonymous genetic variants.

Non-synonymous SNP and small in-del variants found by population sequencing and comparison to the ancestral reference genome of each strain. The position and amino acid change, insertion (Ins) or deletion (Del) is indicated relative to the reference genome sequenced before adaptive evolution. "Total variants" is the total number of variants across all lineages. Annotations were derived from the Rapid Annotations using Subsystems Technology (RAST) server (Aziz et al. 2008) and manual BLAST analysis via NCBI.

Strain	Position	Nature of mutation	Gene product	Function	Total variants	No. of mutated lineages
Kp08	Contig14, Pos: 185456 & 185528	Tyr440His & Tyr464His	Cell division protein FtsK	Coordinates cell division	4	3
Kp08	Contig13, Pos: 121476	Ala653Thr	RnfC	Involved in electron transport	3	3
Kp08	Contig12, Pos: 200895	Ala280Thr	FKBP-type peptidyl-prolyl cis-trans isomerase FkpA precursor	Chaperone; protein folding	2	2
Kp08	Contig27, Pos: 20285	Ala183Val	Alanine transaminase (EC 2.6.1.2)	Amino acid metabolism	1	1
Kp08	Contig14, Pos: 195113 & 195116	Met11Val & Val12Leu	Anaerobic dimethyl sulfoxide reductase chain C	Anaerobic respiration	2	1
Kp08	Contig45, Pos: 9844	Frameshift	Putative cell division protein precursor		1	1
Kp08	Contig12, Pos: 65936	Frameshift	Transcriptional activator NhaR	Osmotic stress	1	1
Ec37	Contig34, 5616	Gln106Arg	Periplasmic fimbrial chaperone	Involved in fimbriae synthessis	4	4
Ec37	Contig4, Pos: 4920, 5885,5890 & 6269	Ins, Del, Tyr158Asn & Pro284Leu	KpsD	Exopolysaccharide synthesis	4	3
Ec37	Contig11, Pos: 98610	Val12Ala	Type 1 fimbriae regulatory protein FimE	fimbriae, pathogenicity	1	1
Ec38	Contig25, Pos: 375	Val710Ala	Outer membrane usher protein PapC	P-fimbriae, pathogenicity	2	2
Ec38	Contig13, Pos: 104962	Leu99Phe	Prophage tail fiber protein	Phage component	2	2
Ec38	Contig1, Pos: 372253	Leu74Phe	Leucine-responsive regulatory protein Lrp	Global regulator	1	1
Ec38	Contig1, Pos: 111174	Pro46Leu	Lipoate synthase LipA	Lipoate biosynthesis	1	1
Ec38	Contig5, Pos: 85177	Cys130Tyr	Redox-sensitive transcriptional activator SoxR	Oxidative stress response	1	1

Supplementary fig. S3. Linear representation of the 8-33kb part of pKP33

A 25kb part of the plasmid flanked by IS*26* insertion sequences was deleted in all evolved *E. coli* lineages. Annotated genes are classified as accessory (orange), stability (blue), conjugational transfer (green) or mobile genetic elements (pink/purple).



Supplementary table S5. Overview of statistical comparisons.

A summary of the statistical comparisons made to assess the effect of plasmid knockout mutants on the growth rate of the naïve *E. coli* isolates. Two-sample *t*-test comparisons were done unless the comparison was made to the plasmid-free ancestor to which the growth rate was normalized and the variance propagated. Here a one-sample *t*-test (to test a difference from 0) was made.

Strain 1		**Strain 2**	***n***	***P*-value**
Kp08/pKP33	x	Kp08/pKP33evo	48	*0.063*
Ec37/pKP33Δ9-30kb	x	Ec37/pKP33evo	32	*0.158*
Ec38/pKP33Δ9-30kb	x	Ec38/pKP33evo	32	*0.374*
Ec37/pKP33	x	Ec37/pKP33ΔT4SS	32	*< 0.001*
Ec38/pKP33	x	Ec38/pKP33ΔT4SS	32	*< 0.001*
Ec37/pKP33ΔT4SS	x	Ec37/pKP33Δ9-30kb	32	*0.115*
Ec38/pKP33ΔT4SS	x	Ec38/pKP33Δ9-30kb	32	*0.2*
Ec37	x	Ec37/pKP33Δ9-17 kb	16	*0.388*
Ec38	x	Ec38/pKP33Δ9-17 kb	24	*< 0.001*
Ec38/pKP33ΔT4SS	x	Ec38/pKP33Δ9-17 kb	40	*0.0087*
Ec37/pKP33Δ9-17 kb	x	Ec38/pKP33Δ9-17 kb	40	*9.61E-05*
Ec37/pKP33ΔT4SS	x	Ec38/pKP33ΔT4SS	48	*0.016*
Ec37/pKP33 D. 35	x	Ec37/pKP33evo	48	*0.172*
Ec38/pKP33 D. 35	x	Ec38/pKP33evo	48	*0.396*
Ec37/pKP33ΔT4SS	x	Ec37/pKP33evo	48	*0.00169*
Ec38/pKP33ΔT4SS	x	Ec38/pKP33evo	48	*0.0608*
Ec37/pKP33Δ9-17 kb	x	Ec37/pKP33evo	48	*0.31*
Ec38/pKP33Δ9-17 kb	x	Ec38/pKP33evo	48	*< 0.001*

Supplement references:

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

Crozat E, Rousseau P, Fournes F, Cornet F. 2015. The FtsK Family of DNA Translocases Finds the Ends of Circles. J. Mol. Microbiol. Biotechnol. 24:396–408.

Doyle M, Fookes M, Ivens A, Mangan MW, Wain J, Dorman CJ. 2007. An H-NS-like stealth protein aids horizontal DNA transmission in bacteria. Science 315:251–252.

Harrison E, Guymer D, Spiers AJ, Paterson S, Brockhurst MA. 2015. Parallel Compensatory Evolution Stabilizes Plasmids across the Parasitism-Mutualism Continuum. Curr. Biol. 25:2034–2039.

Klemm P, Roos V, Ulett GC, Schembri M a, Svanborg C. 2006. Molecular Characterization of the Escherichia coli Asymptomatic Bacteriuria Strain 83972 : the Taming of a Pathogen Molecular Characterization of the Escherichia coli Asymptomatic Bacteriuria Strain 83972 : the Taming of a Pathogen. 74:781–785.

Maor-Shoshani a, Reuven NB, Tomer G, Livneh Z. 2000. Highly mutagenic replication by DNA polymerase V (UmuC) provides a mechanistic basis for SOS untargeted mutagenesis. Proc. Natl. Acad. Sci. U. S. A. 97:565–570.

Roos V, Schembri M a., Ulett GC, Klemm P. 2006. Asymptomatic bacteriuria Escherichia coli strain 83972 carries mutations in the foc locus and is unable to express F1C fimbriae. Microbiology 152:1799–1806.

Roos V, Ulett G. 2006. The asymptomatic bacteriuria Escherichia coli strain 83972 outcompetes uropathogenic E. coli strains in human urine. Infect. Immun. 74:615–624.

San Millan A, Toll-Riera M, Qi Q, MacLean RC. 2015. Interactions between horizontally acquired genes create a fitness cost in Pseudomonas aeruginosa. Nat. Commun. 6:6845.

Simonsen L, Gordon DM, Stewart FM, Levin BR. 1990. Estimating the rate of plasmid transfer: an end-point method. J. Gen. Microbiol. 136:2319–2325.

Vesth T, Lagesen K, Acar Ö, Ussery D. 2013. CMG-biotools, a free workbench for basic comparative microbial genomics. PLoS One 8:e60120.

Manuscript IV [redacted]

Diverse genetic error modes constrain large-scale bio-based production

Authors:

Peter Rugbjerg[a] – petru@biosustain.dtu.dk

Nils Myling-Petersen[a] – nimyp@biosustain.dtu.dk

Andreas Porse[a] – anpor@biosustain.dtu.dk

Kira Sarup-Lytzen[a] – kisa@biosustain.dtu.dk

Morten O. A. Sommer[a *] - msom@bio.dtu.dk

a) The Novo Nordisk Foundation Center for Biosustainability,
Technical University of Denmark, Building 220,
DK-2800 Kongens Lyngby, Denmark

* Corresponding author and lead contact

A few confidential passages have been modified/removed in the current manuscript version.

Abstract

A transition towards sustainable bio-based chemical production is important for green growth. However, productivity and yield frequently decrease as large-scale microbial fermentation progresses, commonly ascribed to phenotypic variation. Yet, given the high metabolic burden and toxicities, evolutionary processes might also constrain bio-based production. We experimentally simulated large-scale fermentation with mevalonic acid-producing *Escherichia coli*. By tracking growth rate and production, we uncovered how populations fully sacrifice production to gain fitness within 70 generations. Using ultra-deep (>1000x) time-lapse sequencing of the pathway populations, we identified multiple recurring intra-pathway genetic error modes. This genetic heterogeneity is only detected using deep sequencing and new population-level bioinformatics, suggesting that the problem is underestimated. A quantitative model explains the population dynamics based on enrichment of spontaneous mutant cells. We validate our model by tuning production load and escape rate of the production host and apply multiple orthogonal strategies for postponing genetically-driven production declines.

Keywords: Genetic heterogeneity, scale-up, evolutionary robustness, metabolic burden, pathway stability, population half-life, insertion sequence, polymerase slippage

Introduction

Bio-based production of chemicals and fuels is important to develop a more sustainable society. However, it remains difficult to scale up many processes that rely on engineered organisms to produce industrially relevant quantities of bio-compounds, which frequently require 100 m^3 fermentation volumes. Indeed, a lack of robustness of synthetic production strains is considered a main challenge for implementing large-scale bioprocesses [1,2]. Furthermore, despite advantages such as higher volumetric productivity, the industrial implementation of continuous fermentation is often limited by appearance of non-producer cells [3–5]. Indeed, declining productivity constrains the economic feasibility of most fermentation reactions to shorter fed-batch operations [6], ultimately limiting our societal transition towards bio-based chemical and fuel production.

Poor performance of bio-based processes is speculated to arise from phenotypic cell-to-cell variation rather than single nucleotide polymorphisms (SNPs) [7,8]. Suboptimal physical reactor conditions such as limited aeration and stochastic gene expression are thought to underlie population heterogeneities [9–11]. As such, sub-populations have been observed to temporarily cease production, then resume production at an unpredictable time [12]. In addition, the high-level cellular biosynthetic activity required for economically viable bioprocesses might reduce the fitness of producer cells enough to select for non-producing mutant cells during industrially relevant time-scales. Such genetic heterogeneity would be more detrimental than temporal phenotypic variations, as genetic heterogeneity results in the irreversible loss of production from a subpopulation in the fermentation tank.

The fitness cost of biosynthesis is pathway-specific and arises from metabolic loads such as enzyme synthesis, DNA synthesis, protein misfolding and drains on endogenous metabolites (required for glycolysis and redox power), but it can also result from the accumulation of toxic intermediates and by-products [13–18]. We employ the term “production load” to the sum of these effects, which present a selective disadvantage for productive cells in direct competition with non-productive cells. The fitness of a production organism can be improved in a variety of ways, including rational engineering [19], adaptive laboratory evolution [20,21], functional metagenomics [22], and fermentation optimization [23,24]. Despite recent progress, production organisms still retain a fitness cost that cannot be eliminated that is directly linked to the burden of non-natural biosynthetic productivity. Accordingly, production cells may be selected against in competition with more fit non-producing cells. However, the extent to which such evolutionary processes limit fermentation output remains unclear and

depends on eventual population size, production load, and the number of cell divisions required to reach industrial fermentation scales.

Generating the fermentation population inside an industrially sized 200 m^3 fed-batch bioreactor involves a gradual scale-up from a master cell bank aliquot and requires approximately 60-80 cell generations to reach population sizes of approximately 10^{20} cells. Such time-scales and population sizes could allow for both the generation and selection of non-producing organisms and might allow these organisms to reach substantial densities in the final fermentation population.

One mechanism that led to non-producing cells in early engineered bioprocesses is the loss of plasmids that encode components of the biosynthetic pathway. Strategies have been developed to limit the loss of plasmid-borne pathway cassettes, including punishing mis-segregation using plasmid-encoded selection genes, toxin-antitoxin systems and chromosomal integration of the pathway genes [25–27]. However, maintenance of the biosynthetic pathway cassette does not preclude the accumulation of genetic errors targeting pathway genes or central metabolic host genes *in trans,* which leads to a loss of biosynthetic activity and potentially improved fitness. Indeed, limiting the mutation rate in *Escherichia coli* by deleting error-prone DNA polymerases and chromosomal insertion sequences (ISs) has led to higher end-point L-threonine productivity and overexpressed recombinant protein titer [28,29]. Such reports suggest that genetic heterogeneity resulting from processes other than gene loss might play a key role in limiting fermentation productivity. However, the actual mechanism and population-level dynamics of such genetically-driven production disruption remains poorly understood, preventing the establishment of a framework for explaining and addressing such production failure modes.

In this study, we investigate the phenotypic and genotypic dynamics of *E. coli* strains engineered to produce mevalonic acid over time-scales relevant to industrial-scale fermentations. Mevalonic acid is a precursor to the important secondary metabolite class of isoprenoids, acting as a chemical building block for colorants, medicines, flavors, fuels and fragrances [30]. Using ultra-deep, time-lapse sequencing of the fermentation populations, we resolve diverse, previously difficult-to-decipher and non-canonical IS transposition events that limit production.

Results

Stability of the mevalonic acid-producing phenotype

We wanted to study the phenotypic dynamics of mevalonic acid-producing *E. coli* over industrially relevant time-scales. Inoculation of large fermenters typically involves gradual scale-up from an

aliquot of a master cell bank by serial growth in vessels of increasing volume [4]. During these cultivations, the original clone, giving rise to the master cell bank aliquots, proliferates through > 60 cell generations (Supplementary Table 1). To experimentally simulate this growth process, we serially transferred production strain lineages every eight hours for a total of nine times, corresponding to approximately 80 cell divisions (generations) (Fig. 1A). Specifically, we cultured five parallel lineages of an *E. coli* TOP10 clone harboring an induced mevalonic acid pathway plasmid (pMevT) maintained under constant antibiotic selection to prevent plasmid loss (Online Methods). To analyze phenotypic and genetic population dynamics, we sampled and freeze-stocked the growing populations every eight hours.

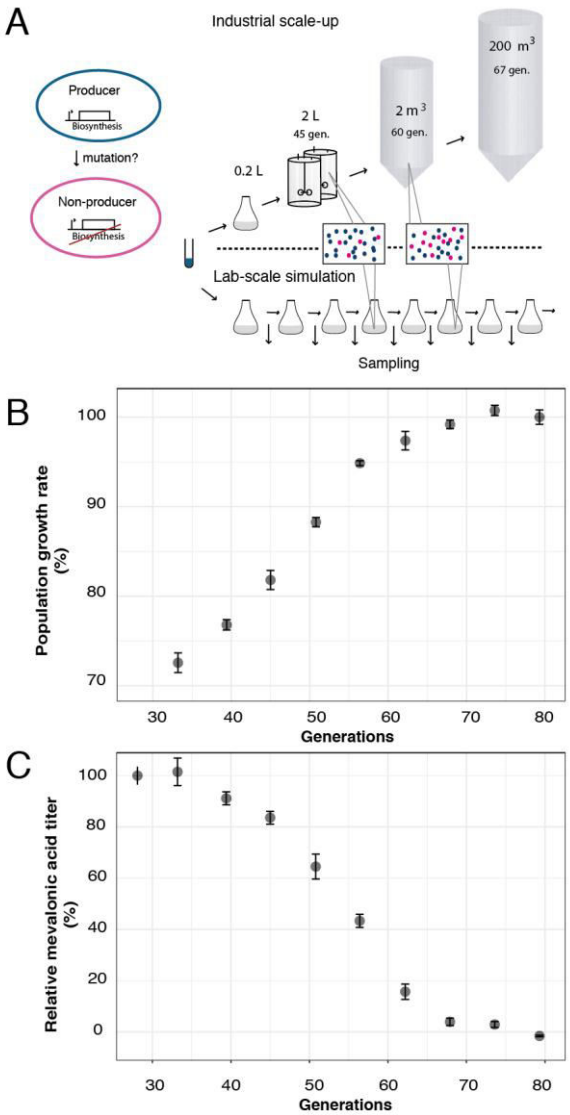


Figure 1. Stability of the mevalonic acid-producing phenotype. A) Large-scale industrial production of mevalonic acid was simulated through serial transfer of five parallel mevalonic acid-producing populations. The length of the fermentation simulation was chosen to mimic the generation number of a fermentation

population in a 200 m^3 fermentation tank. Production populations were sampled every eight hours for subsequent phenotypic and genotypic testing (Supplementary Table 2). B) Population-level average local growth rates were determined for parallel populations over the course of the experiment (Online Methods). The means are shown relative to the last time point (absolute value 0.84 hr^{-1}). A transition of the mean population growth rate is observed after 35 generations, in which the population growth rate increases to a stable phenotypic state after 70 generations, alleviating the measured production load (Supplementary Fig. 1). C) Mevalonic acid titers during the simulated fermentations. The means are shown relative to the earliest time point (Supplementary Table 3) and were calculated from five parallel lineages of the *E. coli* TOP10 mevalonic acid-producing clone. Error bars denote the standard error of the means.

Engineered biosynthesis of mevalonic acid proceeds from acetyl-CoA units through condensation to acetoacetyl-CoA (by AtoB) and HMG-CoA (by ERG13) prior to reduction with NADPH to form mevalonic acid (by tHMGR) [30–32]. Growth impairment from mevalonic acid production in *E. coli* arises largely from the 3-hydroxy-3-methyl-glutaryl-CoA (HMG-CoA) intermediate, which interferes with central fatty acid metabolism and the cell membrane [19]. As a result of high-level mevalonic acid production, our production strain had a 30 % production load, measured relative to a non-producing control strain harboring a pathway-excised plasmid (Supplementary Fig. 1).

To assess the dynamics of the population fitness during the experiment, we evaluated population growth rates (Online Methods). The average population growth rate gradually increased as a function of generation number, following a sigmoidal pattern that stabilized at a new level after 60-75 generations (Fig. 1B). The population growth rate at the beginning of the experiment was 28 % below the final population growth rate, highlighting a considerable change in fitness of the simulated fermentation populations (Fig. 1B). This factor combines all fitness changes over the simulated fermentations, e.g. also possible remaining loads of the IS-disrupted pathway. Notably still, the difference was similar to the measured production load (30 %).

Next, we determined the mevalonic acid titer of each sampled population throughout the simulated fermentation (Fig. 1C). Starting from generation 34, product titers began a decline by several percent per generation before leveling off at undetectable concentrations around generation 70. The onset of the decline in mevalonic acid production coincided with the increase in population growth rate and followed an inversely proportional pattern to the increased growth rate. Population-level growth rates were negatively highly correlated with production titers; following an exponential decline ($R^2 = 0.99$) (Supplementary Fig. 2). This correlation demonstrates how mevalonic acid production was reduced as more fit non-producers took over the fermentation population.

Modeling and measuring production decline during large-scale fermentations

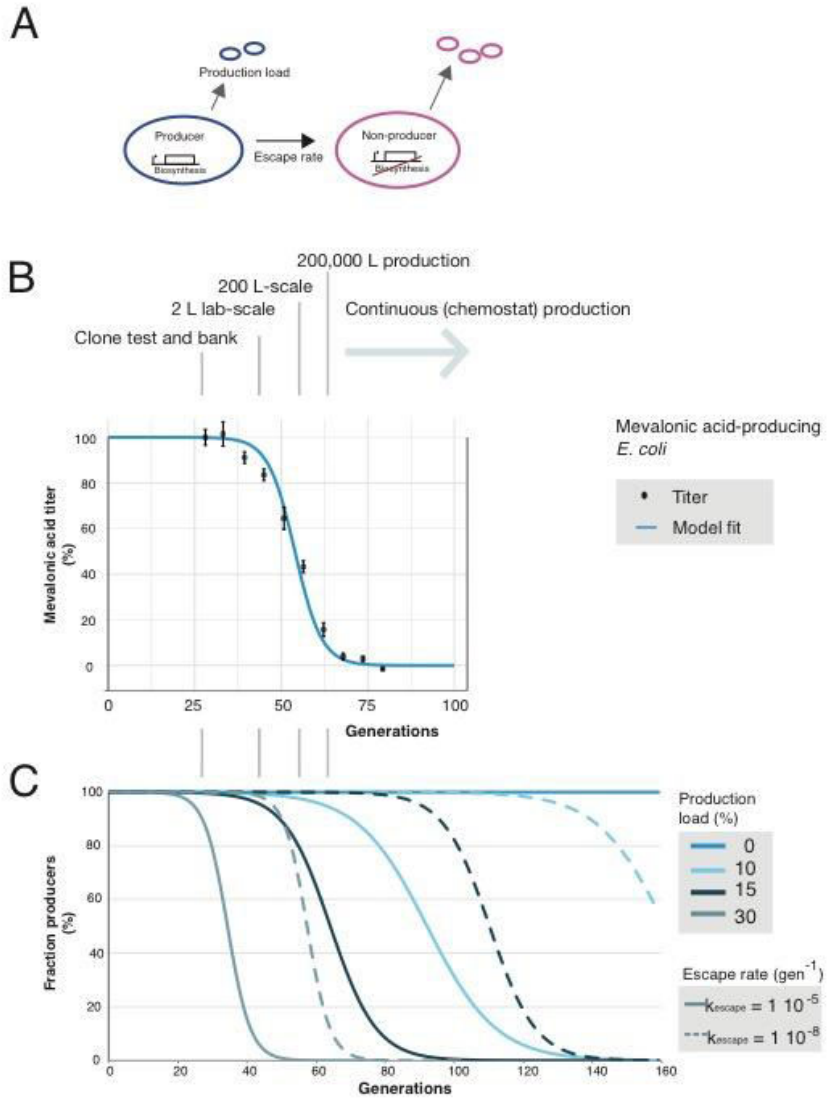


Figure 2. Mathematical modeling is consistent with the observed mevalonic acid production titer in experimentally-simulated fermentations over time. A) Producer cells mutate from the production state at a specific escape rate, thereby alleviating the production load (fitness cost of production) (Supplementary Item 1). B) The best fit of the mathematical model (Supplementary Table 5) to the observed mevalonic acid titer throughout laboratory-simulated mevalonic acid fermentations (relative to earliest time point, Supplementary Table 3). Error bars indicate the standard errors of the means (n = 5). For reference, possible reactor sizes corresponding to particular generation numbers are shown (Supplementary Table 1). C) Modeled fractions of producer cells remaining in the population, in which producer cells irreversibly mutate to non-producers at various escape rates. The magnitude of the production load drives the rate by which spontaneously formed non-producers will enrich the population.

Antibiotic resistance and plasmid loss dynamics have previously been studied using population dynamical models [33–36], but related analysis has not expanded to gene instabilities of metabolically

engineered production organisms. To elucidate evolutionary factors of biotechnological production decline, we developed a simple, two-state deterministic model for the population structure of engineered production strains during fermentation (Fig. 2A, Supplementary Item 1). In our model, a fermentation population contains producing and non-producing cells with different growth rates resulting from the production load. The escape rate describes the transition of producing cells to non-producing cells and represents the combined action of all disruptive mutations that abolish the production load (Fig. 2A). This escape rate depends on numerous factors, including host mutation rate, the size of the genetic targets that abolish production when mutated, and the susceptibility of the genetic targets to recombination or other deleterious genetic events. Owing to production load, the producer cells will be gradually outcompeted by non-producers. The magnitude of the production load determines the rate by which spontaneously formed non-producing cells will enrich in the population. This model offers a simplified description of the population dynamics during fermentation and can be represented with two coupled ordinary differential equations (Supplementary Item 1). Solving these equations yields the respective growth functions of producing and non-producing cells over time, assuming a single producing cell as the starting point, constant escape rate, production load and no nutrient limitation (Supplementary Item 1). To incorporate effects of likely discrete escape events, we also generated a stochastic version of our model (Supplementary Item 4). However, for large populations (>1000 cells), a deterministic model captures an average of the population dynamics and is computationally more efficient due to the existence of an exact analytical solution (Supplementary Table 4) [37,38].

To assess the applicability of our model, we fitted it to the experimentally determined production stability (relative product titers) by non-linear regression to predict the escape rate and production load (Fig. 2B). Using the experimentally determined production load (30 %, Supplementary Fig. 1), the model estimated an effective escape rate of $2.5 \cdot 10^{-8}$ $generation^{-1}$ ($CI_{95\%}$: $\pm 1.2 \cdot 10^{-8}$) (Supplementary Table 5). Such good fit of the production stability data is consistent with our assumption of a genetic basis for production decline.

Our model describes how the fraction of producing cells in the fermentation population will decline sigmoidally over time when a production load is involved in bio-based production (Fig. 2C). Notably, the initial decline determined by the escape rate is low (Fig. 2C) and difficult to detect phenotypically. The production load mainly determines the half-life steepness of this transition, whereas the escape rate largely shifts the timing of the transition. To explore this concept, we calculated the fraction of producer cells over time in specific cases, with production loads of 0-30 %

and escape rates of 10^{-5} – 10^{-8} generation^{-1} (Fig. 2C). We found that slight changes to either parameter have dramatic consequences for the maintenance of producing cells in a fermentation population. These model predictions describe how significant improvements in fermentation end-performance can result from reductions in production load or escape rate.

Production decline originates within pathway genes

Many genome-encoded cell functions are necessary for maintaining biosynthetic production, and any disruption offers an evolutionary trajectory for an engineered strain to re-gain fitness at the expense of production. In a cell factory strain, proteome and genome adaptations might limit metabolic productivity through changes to specific or global transcription factors, protein folding control or precursor fluxes. We therefore wanted to test whether the production plasmid from the evolved populations still conferred mevalonic acid production to a non-evolved host. Plasmid populations were extracted from the five end-points and re-introduced into fresh *E. coli* TOP10 strains. The transformed cultures did not show any detectable mevalonic acid production, demonstrating that the mevalonic acid pathway had been disrupted to incapacitate its biosynthetic potential. Additionally, we found no SNPs in the genomes of nine randomly selected colonies from the end-point populations relative to the ancestral strain (Supplementary Table 6).

Ultra-deep sequencing reveals the dynamics of the pathway disruption landscape

To investigate the genetic basis of production failure at the population level, we ultra-deep-sequenced the heterologous mevalonic acid biosynthetic pathway from three lineages at five sampling points during the experimentally simulated fermentation, and all five at the generation 70 end-point (paired-end 2 x 150bp Illumina sequencing at average depth of 7,200 X, Online Methods).

Prior studies of cell heterogeneity applied SNP analyses to cultured production strains, yet these studies found no evidence of genetic variance [7]. We similarly found a lack of SNPs above 1 % frequencies in our mevalonic acid pathway end-point populations (Supplementary Fig. 4). However, we observed that a declining share of the reads mapped to the production plasmid sequence (Fig. 3A), indicating structural rearrangements of the biosynthetic pathway or other critical parts. This observation prompted us to develop a bioinformatics approach to analyze genetic heterogeneity focusing on structural rearrangements and insertions.

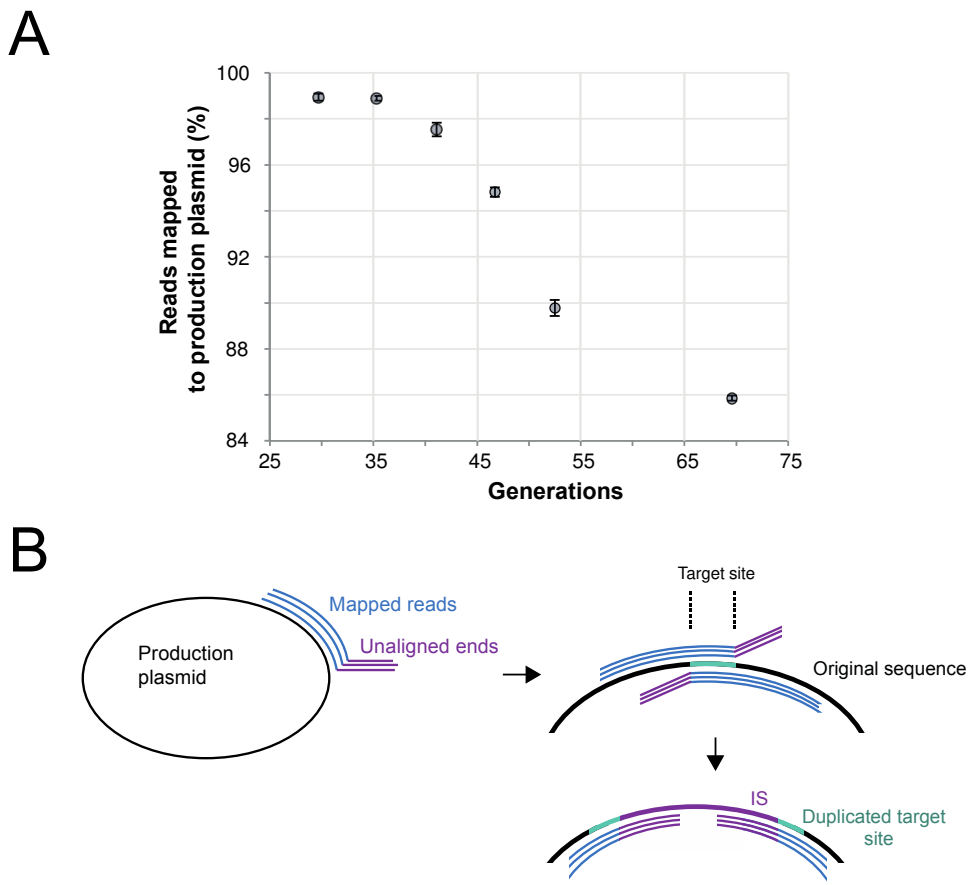


Figure 3. Inferring invasion of deep-sequenced fermentation cell populations by foreign genetic material. A) Percentage of total reads mapped to the production plasmid sequence declined over the course of the fermentation (cell generations) (error bars depict std. error of the mean, n = 3, except generation 70: n = 5). B) Analysis of mapped reads indicated insertion sequence (IS) transposition owing to the presence of crossed-over broken read mappings representing IS insertion and the duplication of target sites, also evident from elevated target site sequencing coverage (Supplementary Fig. 5).

Notably, we observed that reads not mapping to our production pathway frequently exhibited a near perfect partial alignment to our production pathway. We termed such reads broken reads and focused our analysis on these reads (Online Methods). The consensus sequence of several of the unaligned broken read ends showed perfect identity to the termini of insertion sequences (ISs) and tn1000 (also known as gamma-delta), which are mobile elements resident in the *E. coli* TOP10 genome. Accordingly, we speculated that these broken reads result from integration of ISs within the production pathway. Notably, the gap between the right and left breakpoints usually exhibited signature lengths of 3-12 bp, suggesting possible IS target sites (Fig. 3B). At several high-frequency IS target sites, a clear rise in insertion site coverage was observed, likely resulting from duplication of the target insertion region (Supplementary Fig. 5) [39].

To quantify disruption dynamics, we tracked the fraction of position-specific coverage relative to corresponding coverage of non-disrupted reference sequences (Fig. 4A) (Online Methods). We generally detected the position-specific presence of such disruptions in the pathway populations at frequencies down to 0.04 % (three reads).

We found that during the experimentally simulated fermentation, six specific positions in *atoB* and *ERG13* of the metabolic pathway were disrupted by IS10, IS186 and IS5 insertions, jointly constituting > 91 % at the end-points, generation 70 (Fig. 4A). While the *atoB* and *ERG13* pathway genes became increasingly disrupted during the experiment, the final pathway gene *tHMGR* remained free of disruptions (Fig. 4A). This striking degree of preservation of the *tHMGR* gene is probably due to the cytotoxicity of HMG-CoA [19], the substrate of tHMGR. Spontaneous mobile element disruptions of *tHMGR* likely became toxic in cells with active *atoB* and *ERG13*, as these cells would accumulate cytotoxic HMG-CoA concentrations. Because several *atoB* disruptions were also enriched despite the presence of a chromosomal copy, it is very likely that enriched insertions within *atoB* also abolish *ERG13* activity by means of IS-mediated transcriptional termination [40] owing to the operon structure of the mevalonic acid biosynthetic pathway.

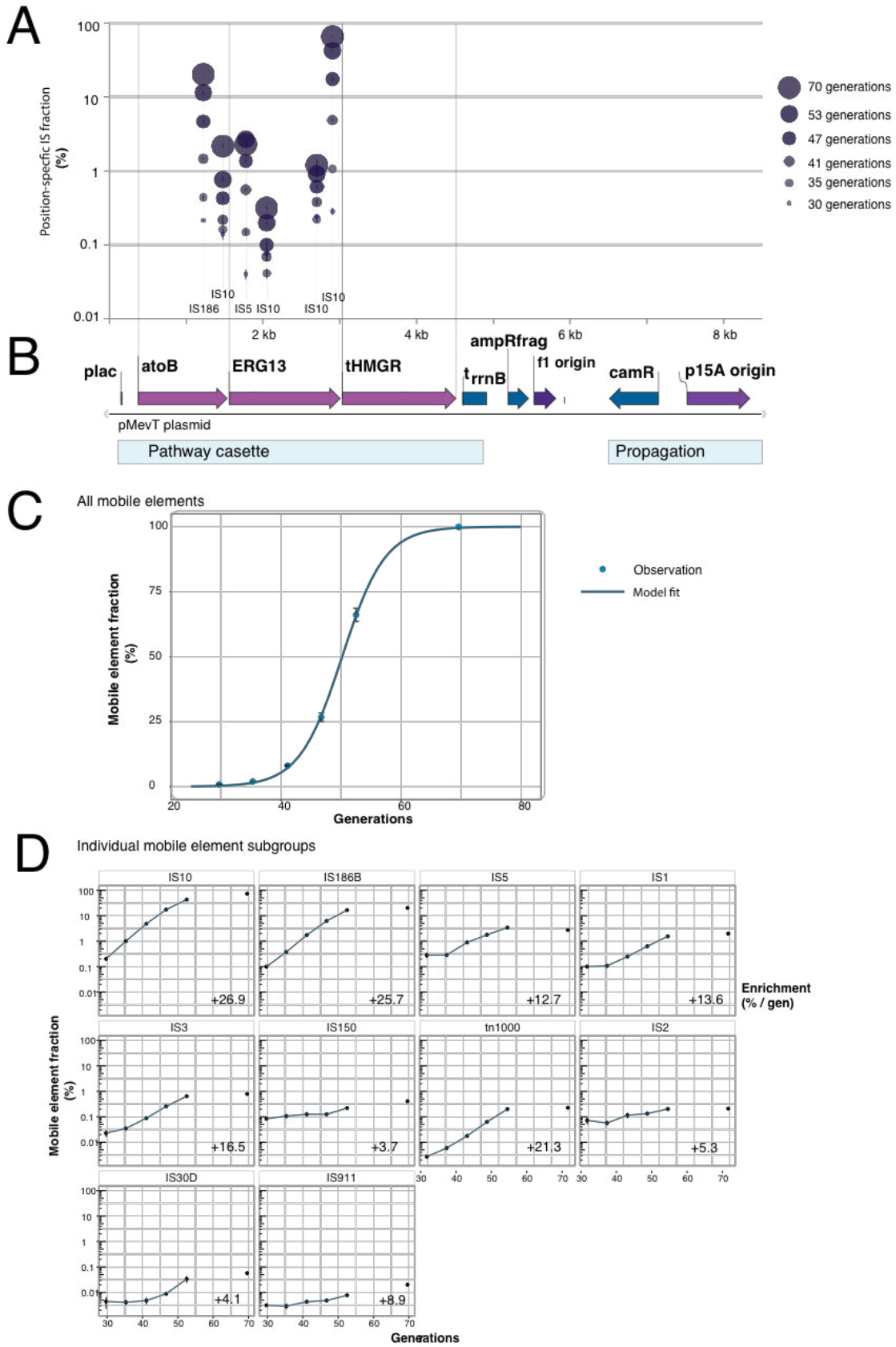


Figure 4. Genetic pathway stability of a mevalonic acid-producing *E. coli* TOP10 clone in parallel lineages. A) Time-lapse high-depth sequencing revealed rising frequencies (population fractions) of mobile element insertions. B) The production plasmid pMevT, which encodes the genes *atoB*, *ERG13* and *tHMGR* in the

mevalonic acid pathway. C) Total enrichment of mobile elements in plasmid populations over the experimentally simulated fermentation period along with the model fit. D) Individual enrichment of host mobile elements in production plasmid populations over the simulated fermentation, and their percent-wise enrichment per generation in the exponential range (generations 30-53) (regression statistics in Supplementary Table 7). For all the graphs, the error bars indicate standard errors (n = 3, except at generation 70: n = 5).

Given that complete production loss was observed in the populations, we speculated that other mobile elements could explain the remaining 9 % fraction. As a strategy to fully resolve the population reads, we mapped all reads to the 24 unique mobile element subgroups in the *E. coli* DH10B genome [41] i.e. not detecting for loci-specific dynamics (Online Methods). We found that joint mobile element coverage relative to the original pMevT approached 99.9 % (std. err. = 1.1 %) at generation 70 (Fig. 4C). A spectrum of ten host mobile element subgroups each transposed to a frequency above 0.01 % in the end-point populations (Fig. 4D).

Using ultra-deep sequencing data to infer population structure is more direct than relative production titers because of prior knowledge of the initial, pure starting point and no requirement for sample re-cultivation. By fitting the time-resolved total mobile element fractions to our production stability model, we improved the confidence of the prediction and estimated an escape rate of $8.7 \cdot 10^{-8}$ generation^{-1} ($CI_{95\%}$: $\pm 0.2 \cdot 10^{-8}$) (Supplementary Table 5). We therefore fitted our data without a pre-determined production load to see how well the model could estimate both parameters freely. From sequencing-based stability data alone, the model very confidently predicted an alleviated production load of 28.1 % ($CI_{95\%}$: ±0.1%) and a revised escape rate of $2.1 \cdot 10^{-7}$ generation^{-1} ($CI_{95\%}$: $\pm 0.1 \cdot 10^{-7}$) (Supplementary Table 5). The predicted 28.1 % production load is notably similar to the experimentally determined 30 % production load. An escape rate of $2.1 \cdot 10^{-7}$ generation^{-1} corresponds well to previously observed mobile element transposition rates into the selectable *cycA* gene in *E. coli* DH10B [41]. The escape rate of our simple model assumes a complete cellular transition from producing to non-producing behavior upon escape. Within each cell, escape begins with a single plasmid mutation, which upon cell divisions is increasingly selected towards a pure non-producing plasmid population (10-15 copies for p15A-origins), potentially giving rise to an intracellular escape heterogeneity. The process towards intracellular escape fixation is driven by uneven plasmid segregation and increasingly selective advantage with each additional pathway escape. Consequently, the effective production escape rate k_{escape} captures the average rate of these combined processes and therefore likely underestimates the single-copy IS insertion rate.

Given the lack of selectable composite elements in ISs such as antibiotic resistance genes, spontaneous insertion rates have traditionally been harder to detect for ISs than for transposons carrying selectable features. Ultra-deep time-lapse sequencing strategies similar to this study should thus be useful to further elucidate the molecular fundamentals of bacterial IS transposition. On a full cell population basis, broad-spectrum enrichment of ten unique mobile element subgroups was detected to final frequencies below 1 % (Fig. 4D). No apparent correlation between genomic IS copy number and the enrichment rate was observed (Fig. 4D), as exemplified by subgroups IS2 and IS5, which enriched slowly although present at 12 and 13 genomic copies, respectively. This behavior is in contrast to the 27 % per generation enrichment of IS10, IS186, which are present at only three and four host copies. IS10 and IS186 thus enriched to 93 % of the end heterologous mevalonic acid pathway population.

Target site selection of mobile elements is influenced by specific consensus sequences and molecular activities, such as on-going transcription [40]. The consensus IS10 target site was present in *tHMGR*, the p15A origin and the f1 origin [39], but insertions were not detected in these loci since disruption of these elements would likely not be advantageous for growth. Instead IS10 insertions were observed in non-canonical sites within the *atoB* and *ERG13* genes. It is surprising that IS10 disrupts the production load through such non-canonical target consensus (Supplementary Fig. 6), and this observation suggests that the spectrum of some ISs is far wider than previously thought.

Tracking broken read dynamics throughout the experimentally simulated fermentation also allowed us to estimate non-mobile element structural variations that were enriched in the populations, and to compare these variations to those without apparent fitness advantages that merely remained at constant frequencies during the experiment. Structural variations detected in the pathway terminator region were not enriched (Supplementary Fig. 7), and thus likely did not influence production. Such variations might have been observed as artificial noise owing to secondary structure formation during sample preparation. We also searched for signs of direct-repeat homologous recombination, but this was not observed, likely owing to the lack of repetitive sequences larger than 18 bp in the pathway plasmid.

The relatively high estimated production escape rate ($2.1 \cdot 10^{-7}$ generation^{-1}) is also consistent with the low presence of enriched pathway SNPs (Supplementary Fig. 5), given the basal SNP formation rate of 10^{-10} bp^{-1} generation^{-1} [42] .

To assess whether the genetic error modes were host-dependent, we repeated the experiment with similar procedures using a cell bank of a different production strain, *E. coli* K-12 strain XL1 (Online Methods). The end-points of five parallel-cultured populations were deep-sequenced to map the error landscape. Comparison of the error modalities in the two different strains revealed both recurrent motifs as well as clone-specific mechanisms (Supplementary Fig. 8).

Model-guided optimization of production stability

As described by the mathematical model, reductions in production load and escape rate represent powerful approaches to improve strain stability. To explore these, we sought to improve production stability by both media and strain engineering to limit the production load and escape rate, respectively.

The production load resulting from HMG-CoA cytotoxicity is speculated to result from a destabilized lipid membrane due to inhibitions in early-step lipid biosynthesis [24], which may be countered by elevated medium osmotic pressure [43]. We therefore quantified the growth rate of mevalonic acid-producing and non-producing populations in a range of NaCl concentrations. At higher osmotic pressure (7.5-12.5 g/L NaCl), we found that the growth rate of pure non-producing cultures was reduced relative to producing cultures, minimizing the original 30 % growth inhibition to 18-28 % (Fig. 5A).

Our model predicts that a reduction of the production load from 30 % to e.g. 22 % would improve the half-life of producing cells from 47 to 63 generations (Fig. 5B). For processes operating in bioreactor-scales >50m^3, such improved stability could provide a crucial enhancement of the product yield and titer. We thus experimentally simulated another long-term fermentation with the same cell bank (*h2m0*) using one such optimized medium composition (10 g/L NaCl). We measured the mevalonic acid production of the population during the fermentation simulation. The production dynamics showed a clear improvement in stability and a slightly less sharp transition of the population (Fig. 5C), matching the reduced selective advantage of production loss. Indeed, the best model fit to this production stability profile (Supplementary Table 5) estimates a reduced production load at 21 % ($CI_{95\%}$: ±0.9 %). This is very close to the measured production load (Fig. 5A) and corresponds to an extension of the production half-life from 54 to 66 generations. The HMG-CoA accumulation of mevalonic acid production is associated with osmotic and oxidative stress[24]. Cross-protection in osmotic and oxidative stress response of *E. coli*[44] may thus also explain why elevated osmotic pressure in part alleviated the HMG-CoA growth inhibition. The observed, reduced basal

mevalonic acid production may also explain the reduced growth inhibition of these cells (Supplementary Table 3).

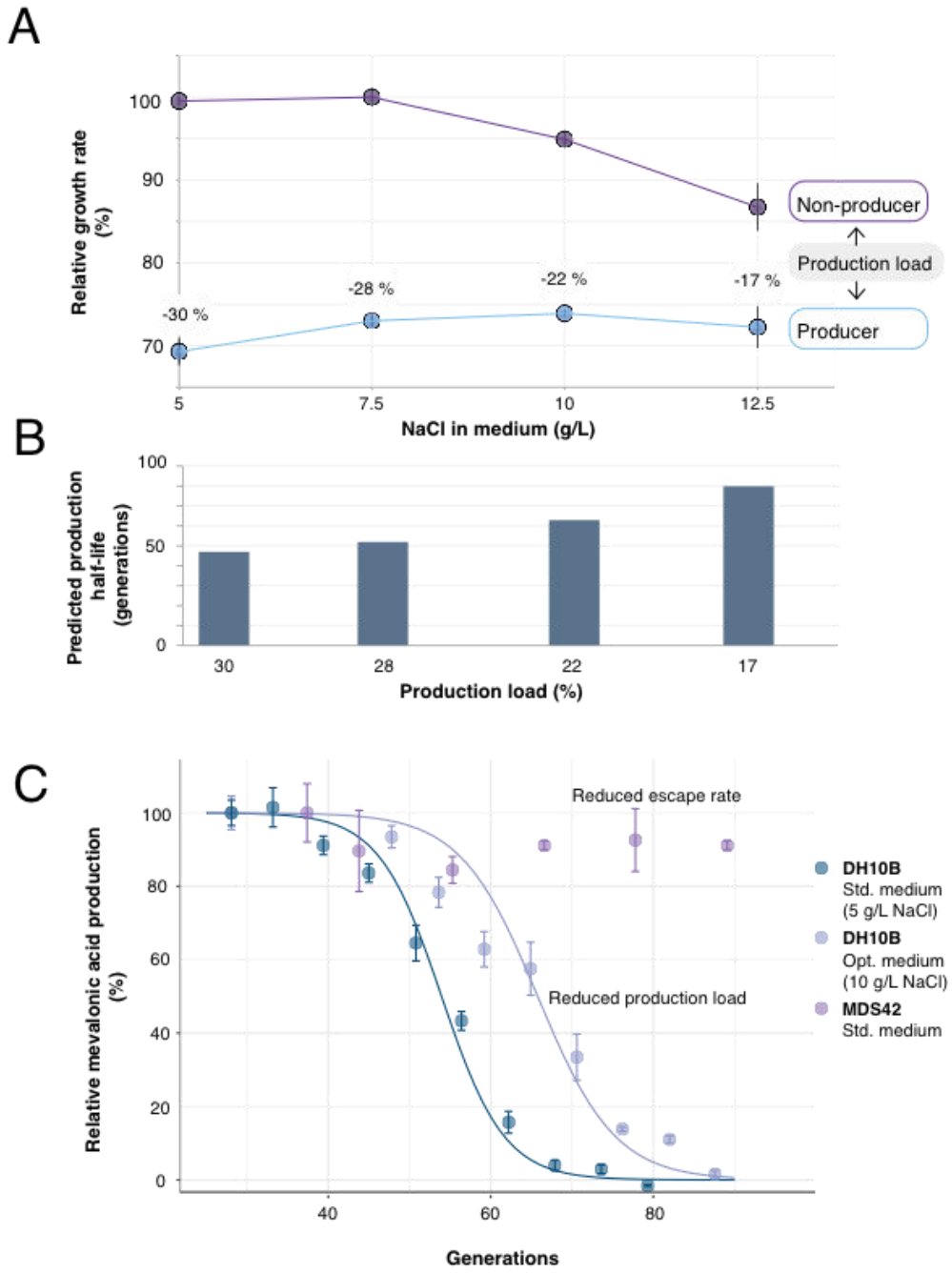


Figure 5. Reducing production decline by optimization of spontaneous escape rate and production load. A) Limiting the production load of mevalonic acid production by supplementing medium with elevated NaCl concentrations (i.e. minimizing the growth-advantage of mutation). This reduction is evident by the growth rate difference of pure producing and non-producing populations. Growth rates depicted relative to fastest growing condition and production load calculated as relative growth rate difference. Error bars depict std. error (n = 6). B) Predicted half-life of a producing cell population with a $2.1 \cdot 10^{-7}$ $generation^{-1}$ escape rate and corresponding production loads estimated from modulated NaCl levels in panel A. C) Utilizing a reduced

production load and escape rate respectively to extend production half-life in experimentally simulated long-term fermentations, evaluated by mevalonic acid production titers (relative to earliest time point) (Supplementary Table 3). Results shown for a TOP10 host (*h2m0*) in standard (n = 5, grey data points) and in optimized medium (n = 4, generation 49: n = 3, blue data points) and for an MDS42 host (*h10m0*) with reduced escape rate (n = 4, orange data points). Error bars depict std. error of the mean. Lines represent best fits to population fraction model at a $2.1 \cdot 10^{-7}$ $generation^{-1}$ escape rate (Supplementary Table 5). Data for TOP10 host in std. medium also shown in Fig. 2B is included for comparison.

Reducing the escape rate offers an alternative strategy that is especially favorable when production loads are difficult to minimize. Due to the observed impact of IS-dominated pathway disruptions, a host strain lacking ISs should prevent this escape mechanism. To benchmark the dynamics in such chassis, we transferred the metabolic pathway (pMevT) to a previously generated, genome-reduced and IS-free *E. coli* K-12 strain MDS42 [45]. We simulated a long-term fermentation with this host strain using our standard medium and quantified the dynamics of mevalonic acid production (Fig. 5C). The marked improvement in production stability supports the hypothesis that lack of host genomic ISs significantly improves production stability, leading to good preservation of production at the end-point of the fermentation simulation (89 generations). However lower escape rate error modes resulting from SNPs and non-homologous recombination may impact stability on a longer term. Further, the fitness-neutral genome reductions guiding the construction of MDS42 [45] may not necessarily be neutral to metabolic production performance, as indicated by the lower mevalonic acid titer in the MDS42 host (Supplementary Table 3).

Discussion

Bio-based production is a central contributor to the transition of our society towards a greener and more sustainable future. However, large-scale bioprocesses are hampered by high yield and productivity requirements, and many new processes cannot be made commercially viable due to declining performance at the scale-up step. Prior work has focused on the phenotypic variance that can contribute to the reduced performance of bio-based processes [48]. While IS elements have been shown to disrupt production phenotypes, the role of evolution and genetic mechanisms in production decline is not well characterized.

In this study, we experimentally simulated the time-scales and population sizes of large-scale bioprocesses in production of the key biochemical building block, mevalonic acid. We introduced a simple framework that captures population dynamics of engineered production strains. We demonstrated in several different host strains that evolution substantially affects population structure over industrial time-scales, with direct ramifications for bio-based process performance.

While ultra-deep sequencing has advanced the understanding of heterogeneities in human disease evolution [49], so far no studies have investigated the potential for biotechnological evolution at population depth, in part possibly due to difficulties in resolving structural variations by short-read sequencing of populations. We observed that pathway error modes are dominated by a broad spectrum of IS insertions in non-canonical target sites. These remove or alleviate the production load, and the error modes differ between strains and clone banks, but appear rapidly in a population.

We find that two key parameters influence the probability and speed by which evolution can impact cell factory stability: the escape rate, which is the rate by which non-producing mutants are generated in a population, and the production load, which manifests a lower fitness of producing cells in direct competition with non-producers. Based on these parameters, a two-state mathematical model accurately describes the essential population dynamics of experimentally simulated fermentations. By de-convoluting stability into its two principal parameters, the model provides a quantitative framework for evaluating the scale-up process, such as the long-term impact of a loaded pathway enzymatic step. This model describes genetic heterogeneity at the population-level and assumes a two-state transition from producer to non-producer cell, which may not be adequate e.g. for pathways operating with several independently loaded biosynthetic genes. Further the model assumes growth without nutrient limitations and a constant escape rate and production load. Average experimental estimations appear to approximate the load well (Supplementary Fig. S1), and may help isolate the production escape rate averages. However, escape rates may be stimulated by different molecular stresses, e.g. in the final phase of bioreactor growth, which are unaccounted for in our simulation. In this study, we have approximated the industrial use of gradually increasing seed train sizes under which most cell divisions occur, by strict passing of cultures in exponential phase throughout the study and shown good fit to a simple model. Thus, integration of dynamic models with time-resolved phenotypic and genotypic data, may help guide an investigator to separate load and mutational effects during production strain and process development to more rationally accommodate evolutionary process limitations.

Our results offer a potential explanation of why lab-scale yields and titers might not accurately predict large-scale fermentation performance following a scale-up procedure, despite vector maintenance. The observed pathway disruptions occur within a plasmid population maintained by selection. We find that these dynamics of pathway disruption are similar to those of segregational plasmid loss [33], although they act at different rate scales and are characterized by a diversity of

mutational error modes (Fig. 4). As an example, our model predicts that an initially pure producing population, with a production load of 30 % and an escape rate of 10^{-7} generation^{-1}, will shift to 96 % non-producing cells over 60 generations, corresponding to a bioreactor of 2 m^3. Remarkably, the same population at lab-scale (age of 37 generations) might appear high performing with less than 3 % non-producers in the population. Because the majority of product is synthesized when the fermentation population reaches the final density in industrial-scale production, it is crucial to investigate the population genotype at this point and not simply extrapolate phenotypic performance from lab-scale experiments. Time-lapse ultra-deep sequencing represents a valuable approach for determining error modes, occurrence rates and their alleviated loads at an early stage. Such ultra-deep sequencing may also be applied to existing scaled-up fermentations, previously thought to be free of heterogeneity.

Common industrial practice employs production strain clone banks stocked as frozen aliquots. Genetic errors in only a few cells might reside in the starting seed although the population appears healthy for a considerable number of divisions. Seeding fermentations from the same cell bank clone therefore generates a highly recurring stability profile. Cell bank aliquots should therefore be evaluated for rare pre-existing mutations that disrupt production and could be selected for during production scale-up. Deep-sequencing of master cell bank aliquots could be applied for this purpose.

Considering that production load improvements by even a few percent can substantially improve stability (Fig. 2C), the specific contributors to production loads must be addressed for each pathway and host cell considering a final large-scale process. For example, reducing the production load from 28 to 23 % should extend the production half-life by 10 generations (assuming a constant escape rate of $2.1 \cdot 10^{-7}$ generation^{-1}).

Practical strategies will be required to reduce factors of production load, including medium optimization, improved balancing of pathway gene expression and the cellular export of toxic by or end products. Poor balancing may favor accumulation of toxic pathway intermediates which carry particularly high potential as a production load. Because intermediates are intracellular, associated toxicities selectively target the producing cells. In this case, adding genes to degrade toxic by-products or to dynamically redirect pathway flux, and the use of specific metabolite- or stress-induced pathway promoters, may be advantageous for limiting production load [50–52]. In the rare cases of growth rate-coupled production, semi-continuous processes have been commercialized to improve productivity, such as R-lactic acid in *Lactobacillus* [53].

In attempts to improve stability, systems for maintaining metabolic pathways through multiple chromosomal integrations are often used[54,55] while stabilizing duplications may also result randomly during optimization[56]. Still, integrated pathways remain subject to intra-pathway disruptions by SNPs, mobile genetic elements and illegitimate recombination, but independence of the individual integration sites means that escapes will not spread in the intracellular pathway population by means of uneven segregation such as some plasmid systems permit. Independently propagated pathway copies may thus provide stabilization in addition to easier antibiotics-free pathway maintenance, which today appears as the major advantage of chromosomal integration. Yet multiple integrations require significantly longer construction protocols, especially at >50 copies and when limited to IP-free engineering. Future studies should investigate the changes in intra-pathway escape dynamics of such multi-copy, chromosomally integrated production pathways. Based on our results, removal of mobile elements from the genomes of microbial production strains [58] is a relevant first step for long-term stabilization and the enabling of toxic and burdened pathway expression. Mechanistically, escape via homologous recombination points can be avoided e.g. by synonymous codons[57]. Such strategies postpone the onset of significant genetic heterogeneity (Fig. 5C).

Dynamic fermentation population models might serve as technical tools to predict the necessary reduction in escape rate or production load for a particular bioreactor size. Knowledge of stability dynamics should ensure a more holistic evaluation of strains by taking into account the potential for rapid performance loss. By characterizing and modeling the interplay between spontaneous genetic errors in-depth and their selection by metabolic burden and pathway toxicity, we have shown the paths for their synergistic impact on pathway stability. Furthermore, we have demonstrated how engineered reductions in both production load and escape rate can improve stability. We expect that the results, methodologies and their implications will open new opportunities for metabolic engineers in the quest to develop sustainable and industrially scalable bioprocesses.

Author Contributions

PR, NMP and MS conceived the study. PR, NMP, AP and KSL conducted the experiments, supervised by MS. All authors analyzed data and wrote the manuscript.

Acknowledgements

We thank Malcolm Rhodes for critical comments to the manuscript. Anna Koza is thanked for help and operation of the Miseq DNA sequencing instrument. Mette Kristensen is thanked for help with HPLC analysis. pMevT was a gift from Jay Keasling (Addgene plasmid # 17815). The research leading to these results has received funding from the Novo Nordisk Foundation, Denmark and from the European Union Seventh Framework Programme (FP7-KBBE-2013-7-single-stage) under Grant agreement no. 613745, Promys.

Online Methods

Strains

E. coli K-12 parental strains below were used to construct the strains analyzed (Table 1) using the specified plasmids (Table 2).

E. coli TOP10, similar to DH10B (Invitrogen)

F– mcrA Δ(mrr-hsdRMS-mcrBC) Φ80lacZΔM15 ΔlacX74 recA1 araD139 Δ(ara, leu) 7697 galU galK rpsL (StrR) endA1 nupG

E. coli XL1 (Agilent)

recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F´ proAB lacI^{q} ZΔM15 Tn10 (Tet^{r})]

E. coli MDS42 ("MDS42 LowMut", Scarab Genomics)

MG1655 genome-reduced for all ISs, *fhuACDB, endA* and more [45].

Standard chemical transformation or electroporation was used for gene introduction.

For test in experimentally simulated fermentations, single colonies were cultured and stored at -80 deg. C to serve as working clone banks (designated by *m* numbers) following standard industrial practice [59].

Table 1. Strains analyzed in this study.

Strain	Plasmid	Parental *E. coli* K12 strain	Chromosomal editing	Clone banks
h2	pMevT	TOP10	-	m0, m1, m2, m3
h8	pMevT4	TOP10	-	m0
XL1-pMevT	pMevT	XL1	-	m0
h10	pMevT	MDS42	-	m0
h9	pMevT4	MDS42	-	m0

Plasmids

pMevT4 was generated by PstI digestion of pMevT followed by re-ligation of the backbone using T4 DNA ligase and standard molecular biology methods, excising the metabolic pathway cassette (*atoB, ERG13, tHMGR*).

PCRs were conducted by standard procedures with Phusion U DNA polymerase (Thermo). Uracil-excision cloning was performed by approx. equimolar mixing of the respective purified PCR products (Table 3) in a 20 µL reaction including 2 µL FastDigest buffer (Thermo), 0.75 µL USER enzyme (NEB) and 0.75 uL FastDigest DpnI (Thermo). The reaction incubated for 60 minutes at 37 deg. C followed by 20 minutes at 25 deg. C, and was subsequently co-transformed with pSIM6 into chemically competent *E. coli* TOP10 cells.

Table 2. Plasmids used to generate strains.

Plasmid	Relevant features	Reference
pMevT	p_{lac}:*atoB-ERG13-tHMGR*:t_{rnnB}, cam^R, p15A	(Martin et al., 2003)
pMevT4	p_{lac}:t_{rrnb}, cam^R, p15A	This study

Media

For all cultivations, standard 2xYT characterization medium was used unless otherwise stated:

2xYT medium: 10 g/L yeast extract (Sigma-Aldrich), 16 g/L tryptone (Bacto), 5 g/L NaCl (pH adjusted to 7.0) supplemented with 30 µg/mL chloramphenicol and 500 µM isopropyl β-D-1-thiogalactopyranoside (IPTG).

For genetic transformations, SOC medium was used. SOC consisted of 5 g/L yeast extract, 20 g/L tryptone (Bacto), 10 mM NaCl, 2.5 mM KCl, 20 mM $MgSO_4$ and 20 mM D-glucose.

Experimentally simulated long-term fermentation by continuous growth of mevalonic acid-producing *E.coli*

Five parallel lineages of the pMevT-harboring TOP10 clone bank *h2m0* were inoculated into aerated cultures (EVO2) (Table 3). Each culture contained 25 mL medium and was grown for 8 hours at 30 °C with horizontal shaking at 250 rpm. After 8 hours, 0.5 mL broth was inoculated into 25 mL fresh medium and incubated under the same conditions for another 8 hours. At each passage, the OD_{600} was recorded to determine the accumulated number of cell divisions (Table

S2). Constant pathway induction (500 µM IPTG) was applied to mimic constitutive promoter designs, as the advantages of late induction (e.g., using p_{lac}) would be unattainable industrially owing to inducer cost [6]. The simulation was repeated (EVO8) . For EVO8, four of five lineages were randomly selected for subsequent analysis.

Table 3 Experimentally simulated long-term fermentations. For growth progressions, see Tables S2 and S9.

EVO no.	Strains	Culture conditions	Clone banks	Lineages	Deep-sequenced seeds
EVO2	*h2*	2xYT, 30 deg. C	m0	c6-c10	All: s9 c6,c8,c10: s2-s6
EVO2B	*XL1-MevT*	2xYT, 37 deg. C	m0	c1-5	s6
EVO8	*h2*	2xYT+5g/L NaCL, 30 deg. C	m0	c1-4	-
	h10	2xYT, 30 deg. C	m0	c6-9	-

Repeated simulated long-term fermentation with different *E. coli* host strain XL1

Five parallel lineages of the pMevT-harboring *E. coli* XL1 clone bank (*XL1-MevT*) were inoculated into aerated cultures (EVO2B) (Table 3). Each culture contained 50 mL medium and was grown for 12 hours at 37 °C with horizontal shaking at 250 rpm to match the slower growth of XL1. After 12 hours, 1 mL broth was inoculated into fresh medium and incubated under the same conditions. At each passage, the OD_{600} absorbance was recorded.

Simulated long-term fermentation with pathway-coupled essential gene expression

Four parallel lineages were inoculated from respectively four *h2* and *kle1#1* master clone banks into aerated cultures and was cultured at 32 deg. C (EVO10) (Table 3), but otherwise following same methods as the first simulated long-term fermentation. Randomly, three lineages of respectively *h2* and *kle1*#1 were selected for subsequent analysis. At each passage, the OD_{600} was recorded to determine the accumulated number of cell divisions (Table S9). Three of four lineages were randomly selected for subsequent analysis.

High-depth DNA sequencing and analysis

Upon each passage to new medium, 1.8 mL of the grown culture was stored at -20 °C. Similarly, 1.8 mL of grown culture was stored at the simulated fermentation end. Production plasmid

populations were subsequently purified from each time point using a standard plasmid purification kit (Macherey-Nagel). The samples were then prepared for Miseq sequencing using the Nextera XT v2 set A kit (Illumina) per manufacturer's instructions with the addition of two 'limited-cycle PCR' cycles.

Sequencing was performed in a pooled run with 150 bp paired-end reading. CLC Genomics Workbench (version 8.5) was used for initial bioinformatics analysis. First, the reads were mapped to the reference pMevT sequence (Addgene #17815). Broken aligned reads were identified using the CLC Genomics Workbench tool Breakpoint analysis to yield a table of the consensus broken unaligned reads and their abundance (maximum three miss-matches allowed in the mapped read region, p-value for the fraction of unaligned reads set to 0.0001) to obtain an initial overview of occurred structural variation. The fraction of mobile element coverage to plasmid coverage was calculated by mapping of reads to the 24 unique *E. coli* DH10B mobile element sequences and the reference plasmid. Subgroups IS10R and IS10L were combined as IS10 and IS1A, B and F as IS1. Position-specific mobile element/reference coverage was calculated by mapping 80 bp sequences consisting of respectively the corresponding 40 bp reference plasmid and 40 bp mobile element sequence or 80 bp reference plasmid sequence. These mappings were performed with the alignment setting 'global' and identity fraction set to '0.6'. Mappings without any reads covering across the mobile element junction were disregarded. Single-nucleotide polymorphisms (SNPs) and short deletions were called using the CLC Genomics Workbench Low Frequency Variant Detection tool with a 1 % required significance level and 0.25 % minimum frequency (unless otherwise specified). The SNP frequencies in the sequenced populations were calculated by division with their respective coverage values.

Five SNPs found in the plasmid backbone at >99% frequencies in the initial seed were regarded as present in the starting plasmid. The deep-sequencing data is available via the ArrayExpress repository (accession no: E-MTAB-5862).

Whole-genome sequencing of single colonies

DNA for whole-genome sequencing from single colonies was extracted from a grown 2 mL culture using a standard DNA extraction kit (Qiagen), but otherwise prepared as above for a pooled Miseq run. For identification of SNPs, reads were mapped to the public available DH10B genome, and detected for using the Low Frequency Variant Detection tool of CLC Genomics with a minimum detection frequency of 80 %.

Measurement of mevalonic acid production by HPLC

Upon each passage to a fresh culture, 900 µL medium was mixed with 900 µL 50 % glycerol and stored at -80 °C. Following the simulated fermentation, each population sample from a 25 µL glycerol stock was used to inoculate 15 mL medium, and the culture was incubated at 30 °C with shaking at 250 rpm for 54-58 hours. Following incubation, 300 µL aliquots were treated with 23 µL 20 % sulfuric acid. Samples were vigorously shaken and then spun down at 13,000 g for 2 minutes. Supernatant (medium) samples were injected into an Ultimate 3000 HPLC running a 5 mM sulfuric acid mobile phase (0.6 mL/min) on an Aminex HPX-87H ion exclusion column (300 mm × 7.8 mm, Bio-Rad Laboratories) at 50 °C. A refractive index detector was used for detection. A standard curve for mevalonic acid was generated with mevalonolactone (Sigma-Aldrich) dissolved in 2xYT medium supernatant of an engineered non-producing strain incubated under same conditions.

Measurement of population growth rates

To measure population growth rates, 1.5 µL aliquots of stationary-phase cultures grown for productivity analysis (as described in the previous section) were used to inoculate 200 µL medium in microtiter plate wells. The microtiter plate was sealed with a Breathe-Easy polyurethane seal (USA Scientific) and was incubated at with "fast" continuous shaking in an ELx808 kinetic plate reader (BioTek), which measured the OD_{630} value every ten minutes.

Background-subtracted OD_{630} values were computed using the measurements from un-inoculated wells. The local growth rates were computed for each background-subtracted OD_{630} value by regressions in rolling windows of five measurement points and background-subtracted OD_{630} values. To represent the growth rates in the actual fermentation simulations, the average was computed of the local growth rates where the background-subtracted OD_{630} was > 0.04 and < 0.4. R script appended (Supplementary Item 2).

Simulation of producer fraction over time and fit to model

The coupled ODE system (Supplementary Item 1) was solved using Maple 2015. Solution growth functions were then combined to yield a function for the fraction of producer cells in time (Supplementary Item 1). Non-linear regression was performed to fit to this model with the nls2 R package (Supplementary Item 3). A stochastic version of the model was constructed by employing the algorithm of Gillespie to simulate discrete mutational events of our system in a stochastic manner [61]. We assume that each event (e.g. cell division and mutation) occurs

according to probabilities scaled with the parameters obtained from the deterministic fit of our data (production load: 30 %, escape rate: $2.1 \cdot 10^{-7}$) (Supplementary Item 4).

Data and code availability

Deep-sequencing data from the study (Figures 3 and 4) have been deposited in ArrayExpress under ID code E-MTAB-5862 .

R scripts used to process raw growth data and fit to model are provided in Supplementary Items 2 and 3. R script for running stochastic version of model is provided in Supplementary Item 4.

References

1. Nielsen, J. & Keasling, J. Engineering Cellular Metabolism. *Cell* **164,** 1185–1197 (2016).
2. Borkowski, O., Ceroni, F., Stan, G. & Ellis, T. Overloaded and stressed: wholecell considerations for bacterial synthetic biology. *Curr. Opin. Microbiol.* **33,** 123130 (2016).
3. Kumar, P. K., Maschke, H. E., Friehs, K. & Schügerl, K. Strategies for improving plasmid stability in genetically modified bacteria in bioreactors. *Trends Biotechnol.* **9,** 279–84 (1991).
4. Ikeda, M. Amino acid production processes. *Adv. Biochem. Eng. Biotechnol.* **79,** 1–35 (2003).
5. Stanbury, P. in *Principles of Fermentation technology* (1995).
6. Lee, S. Y. & Kim, H. U. Systems strategies for developing industrial microbial strains. *Nat. Biotechnol.* **33,** 1061–1072 (2015).
7. Xiao, Y., Bowen, C. H., Liu, D. & Zhang, F. Exploiting nongenetic cell-to-cell variation for enhanced biosynthesis. *Nat. Chem. Biol.* (2016). doi:10.1038/nchembio.2046
8. Müller, S., Harms, H. & Bley, T. Origin and analysis of microbial population heterogeneity in bioprocesses. *Curr. Opin. Biotechnol.* **21,** 100–13 (2010).
9. Carlquist, M. *et al.* Physiological heterogeneities in microbial populations and implications for physical stress tolerance. *Microb. Cell Fact.* **11,** 94 (2012).
10. Heins, A.-L., Lencastre Fernandes, R., Gernaey, K. V. & Lantz, A. E. Experimental and in silico investigation of population heterogeneity in continuous Sachharomyces cerevisiae scale-down fermentation in a two-compartment setup. *J. Chem. Technol. Biotechnol.* **90,** 324–340 (2015).
11. Avery, S. V. Microbial cell individuality and the underlying sources of heterogeneity. *Nat. Rev. Microbiol.* **4,** 577–587 (2006).

12. Mustafi, N. *et al.* Application of a genetically encoded biosensor for live cell imaging of L-valine production in pyruvate dehydrogenase complex-deficient Corynebacterium glutamicum strains. *PLoS One* **9,** (2014).
13. Barbirato, F., Grivet, J. P., Soucaille, P. & Bories, A. 3-Hydroxypropionaldehyde, an inhibitory metabolite of glycerol fermentation to 1,3-propanediol by enterobacterial species. *Appl. Environ. Microbiol.* **62,** 1448–1451 (1996).
14. Berry, A., Dodge, T., Pepsin, M. & Weyler, W. Application of metabolic engineering to improve both the production and use of biotech indigo. *J Ind Microbiol Biotechnol* **28,** 127–133 (2002).
15. Birnbaum, S. & Bailey, J. Plasmid presence changes the relative levels of many host cell proteins and ribosome components in recombinant Escherichia coli. *Biotechnol. Bioeng.* **37,** 736–745 (1991).
16. Glick, B. R. Metabolic load and heterologous gene expression. *Biotechnol. Adv.* **13,** 247–261 (1995).
17. Shachrai, I., Zaslaver, A., Alon, U. & Dekel, E. Cost of Unneeded Proteins in E. coli Is Reduced after Several Generations in Exponential Growth. *Mol. Cell* **38,** 758–767 (2010).
18. Xia, X.-X. *et al.* Native-sized recombinant spider silk protein produced in metabolically engineered Escherichia coli results in a strong fiber. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 14059–63 (2010).
19. Pitera, D. J., Paddon, C. J., Newman, J. D. & Keasling, J. D. Balancing a heterologous mevalonate pathway for improved isoprenoid production in Escherichia coli. *Metab. Eng.* **9,** 193–207Pitera, D. J., Paddon, C. J., Newman, J. D. (2007).
20. Conrad, T. M., Lewis, N. E. & Palsson, B. O. Microbial laboratory evolution in the era of genome-scale science. *Mol. Syst. Biol.* **7,** 509–509 (2011).
21. Dragosits, M. & Mattanovich, D. Adaptive laboratory evolution – principles and applications for biotechnology. *Microb. Cell Fact.* **12,** 64 (2013).
22. Sommer, M. O. a, Church, G. M. & Dantas, G. A functional metagenomic approach for expanding the synthetic biology toolbox for biomass conversion. *Mol. Syst. Biol.* **6,** 360 (2010).
23. Carneiro, S., Ferreira, E. C. & Rocha, I. Metabolic responses to recombinant bioprocesses in Escherichia coli. *J. Biotechnol.* **164,** 396–408 (2013).
24. Kizer, L., Pitera, D. J., Pfleger, B. F. & Keasling, J. D. Application of functional genomics to pathway optimization for increased isoprenoid production. *Appl. Environ. Microbiol.* **74,** 3229–41 (2008).

25. Tyo, K. E. J., Ajikumar, P. K. & Stephanopoulos, G. Stabilized gene duplication enables long-term selection-free heterologous pathway expression. *Nat. Biotechnol.* **27,** 760–765 (2009).
26. Peubez, I. *et al.* Antibiotic-free selection in E. coli: new considerations for optimal design and improved production. *Microb. Cell Fact.* **9,** 65 (2010).
27. Gerdes, K., Rasmussen, P. B. & Molin, S. Unique type of plasmid maintenance function: postsegregational killing of plasmid-free cells. *Proc. Natl. Acad. Sci. U. S. A.* **83,** 3116–3120 (1986).
28. Csorgo, B., Feher, T., Timar, E., Blattner, F. R. & Posfai, G. Low-mutation-rate, reduced-genome Escherichia coli: An improved host for faithful maintenance of engineered genetic constructs. *Microb. Cell Fact.* **11,** 11 (2012).
29. Lee, J. *et al.* Metabolic engineering of a reduced-genome strain of Escherichia coli for L-threonine production. *Microb. Cell Fact.* **8,** 2 (2009).
30. Martin, V. J. J., Pitera, D. J., Withers, S. T., Newman, J. D. & Keasling, J. D. Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat. Biotechnol.* **21,** 796–802 (2003).
31. Xiong, M., Schneiderman, D. K., Bates, F. S., Hillmyer, M. a & Zhang, K. Scalable production of mechanically tunable block polymers from sugar. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 8357–62 (2014).
32. Tabata, K. & Hashimoto, S. I. Production of mevalonate by a metabolically-engineered Escherichia coli. *Biotechnol. Lett.* **26,** 1487–1491 (2004).
33. Proctor, G. N. Mathematics of microbial plasmid instability and subsequent differential growth of plasmid-free and plasmid-containing cells, relevant to the analysis of experimental colony number data. *Plasmid* **32,** 101–130 (1994).
34. Yurtsev, E. A., Chao, H. X., Datta, M. S., Artemova, T. & Gore, J. Bacterial cheating drives the population dynamics of cooperative antibiotic resistance plasmids. *Mol. Syst. Biol.* **9,** 683 (2013).
35. Bentley, W. E. & Quiroga, O. E. Investigation of subpopulation heterogeneity and plasmid stability in recombinant escherichia coli via a simple segregated model. *Biotechnol. Bioeng.* **42,** 222–34 (1993).
36. Caulcott, C. A. *et al.* Investigation of the Effect of Growth Environment on the Stability of Low-copy-number Plasmids in Escherkhiu coli. *J. Gen. Microbiol.* **133,** 1881–1889 (1987).
37. Althaus, C. L. & Bonhoeffer, S. Stochastic interplay between mutation and recombination during the acquisition of drug resistance mutations in human immunodeficiency virus

type 1. *Society* **79,** 13572–13578 (2005).

38. Székely, T. & Burrage, K. Stochastic simulation in systems biology. *Comput. Struct. Biotechnol. J.* **12,** 14–25 (2014).
39. Craig, N. L. Target site selection in transposition. *Annu. Rev. Biochem.* **66,** 437 (1998).
40. Mahillon, J. & Chandler, M. Insertion Sequences. *Microbiol. Mol. Biol. Rev.* **62,** 725–774 (1998).
41. Durfee, T. *et al.* The Complete Genome Sequence of Escherichia coli DH10B: Insights into the Biology of a Laboratory Workhorse. *J. Bacteriol.* **190,** 2597–2606 (2008).
42. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. Rates of Spontaneous Mutation. (1998).
43. Pitera, D. J., Newman, J. D., Kizer, J. L., Keasling, J. D. & Pfleger, B. F. Methods for increasing isoprenoid and isoprenoid precursor production by modulating fatty acid levels. (2012).
44. Gunasekera, T. S., Csonka, L. N. & Paliy, O. Genome-Wide Transcriptional Responses of Escherichia coli K-12 to Continuous Osmotic and Heat Stresses. *J. Bacteriol.* **190,** 3712–3720 (2008).
45. Pósfai, G. *et al.* Emergent properties of reduced-genome Escherichia coli. *Science* **312,** 1044–6 (2006).
46. Doublet, P. & Heijenoort, J. Van. The murI gene of Escherichia coli is an essential gene that encodes a glutamate racemase activity. *J. Bacteriol.* **175,** 2970–2979 (1993).
47. Bonde, M. T. *et al.* Predictable tuning of protein expression in bacteria. *Nat. Methods* **13,** (2016).
48. Binder, D. *et al.* Homogenizing bacterial cell factories: Analysis and engineering of phenotypic heterogeneity. *Metab. Eng.* (2017). doi:10.1016/j.ymben.2017.06.009
49. Beerenwinkel, N. & Zagordi, O. Ultra-deep sequencing for the analysis of viral populations. *Curr. Opin. Virol.* **1,** 413–418 (2011).
50. Zhu, M. M., Skraly, F. a & Cameron, D. C. Accumulation of methylglyoxal in anaerobically grown Escherichia coli and its detoxification by expression of the Pseudomonas putida glyoxalase I gene. *Metab. Eng.* **3,** 218–225 (2001).
51. Dahl, R. H. *et al.* Engineering dynamic pathway regulation using stress-response promoters. *Nat. Biotechnol.* **31,** 1039–46 (2013).
52. Holtz, W. J. & Keasling, J. D. Engineering Static and Dynamic Control of Synthetic Pathways. *Cell* **140,** 19–23 (2010).
53. Zelder, O. & Hauer, B. Environmentally directed mutations and their impact on industrial

biotransformation and fermentation processes. *Curr. Opin. Microbiol.* **3,** 248–251 (2000).

54. Mikkelsen, M. D. *et al.* Microbial production of indolylglucosinolate through engineering of a multi-gene pathway in a versatile yeast expression platform. *Metab. Eng.* **14,** 104–111 (2012).
55. St-Pierre, F. *et al.* One-Step Cloning and Chromosomal Integration of DNA. *ACS Synth. Biol.* **2,** 537–541 (2013).
56. Yanai, K., Murakami, T. & Bibb, M. Amplification of the entire kanamycin biosynthetic gene cluster during empirical strain improvement of Streptomyces kanamyceticus. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 9661–6 (2006).
57. Rugbjerg, P., Knuf, C., Förster, J. & Sommer, M. O. a. Recombination-stable multimeric green fluorescent protein for characterization of weak promoter outputs in Saccharomyces cerevisiae. *FEMS Yeast Res.* **15,** fov085 (2015).
58. Park, M. K. *et al.* Enhancing recombinant protein production with an Escherichia coli host strain lacking insertion sequences. *Appl. Microbiol. Biotechnol.* **98,** 6701–6713 (2014).
59. Molowa, D. T. & Mazanet, R. The state of biopharmaceutical manufacturing. *Biotechnol. Annu. Rev.* **9,** 285–302 (2003).
60. Datta, S., Costantino, N. & Court, D. L. A set of recombineering plasmids for gram-negative bacteria. *Gene* **379,** 109–115 (2006).
61. Gillespie, D. T. Exact Stochastic SImulation of Coupled Chemical Reactions. *J. Phys. Chem.* **93555,** 2340–2361 (1977).
62. Crooks, G., Hon, G., Chandonia, J. & Brenner, S. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).
63. Craig, N. L. Target site selection in transposition. *Annu. Rev. Biochem.* **66,** 437–474 (1997).
64. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2,** 2006.0008 (2006).

Supplementary material

Table S1. Calculated number of cell divisions (generations) needed to occupy the bioreactor type specified at the respective OD_{600} or number of cells.

Bioreactor type	Number of cells	OD_{600}	Volume (L)	Accumulated generations
Strain construction	$2\ 10^{11}$	1	0.01 L	38
Laboratory-scale	$2\ 10^{15}$	100	2 L	51
Industry-scale	$2\ 10^{17}$	100	200 L	57
Industry-scale	$2\ 10^{18}$	100	2.000 L	61
Industry-scale	$1\ 10^{19}$	100	10.000 L	63
Industry-scale	$2\ 10^{20}$	100	200.000 L	67

Table S2. Average determined number of generations undergone by the experimentally simulated fermentation from the mevalonic acid-producing *h2m0* and *h10m0* clone banks (in EVO2 and EVO8), accumulated and at the individual 8-hour passages (seeds) (standard error shown +/-, for EVO2 n = 5, EVO8 n = 4).

Growth tube (seed)	**New generations (-80 stock)**	**Accumulated generations (-80 stock and sequenced sample)**	**Accumulated generations (15 mL re-grown stock for analysis)**	**Experimentally simulated fermentation ID and lineages**	**Strain clone bank**
Bank culture	18.4	18.4	28.1	EVO2 c6-c10	*h2m0*
s1	5.1 +/- 0.02	23.5	33.2	EVO2 c6-c10	*h2m0*
s2	6.2 +/- 0.03	29.7	39.4	EVO2 c6-c10	*h2m0*
s3	5.6 +/- 0.04	35.3	45.0	EVO2 c6-c10	*h2m0*
s4	5.8 +/- 0.03	41.1	50.8	EVO2 c6-c10	*h2m0*
s5	5.6 +/- 0.05	46.7	56.4	EVO2 c6-c10	*h2m0*
s6	5.8 +/- 0.02	52.5	62.3	EVO2 c6-c10	*h2m0*
s7	5.7 +/- 0.00	58.2	67.9	EVO2 c6-c10	*h2m0*
s8	5.7 +/- 0.01	63.9	73.6	EVO2 c6-c10	*h2m0*
s9	5.6 +/- 0.01	69.6	79.3	EVO2 c6-c10	*h2m0*
Bank culture	18.4	18.4	28.1	EVO8 c1-4	*h2m0*
s1	9.8 +/- 0.03	28.2	37.9	EVO8 c1-4	*h2m0*
s2	5.8 +/- 0.04	38.1	47.8	EVO8 c1-4	*h2m0*
s3	5.6 +/- 0.02	43.9	53.6	EVO8 c1-4	*h2m0*
s4	5.7 +/- 0.02	49.5	59.2	EVO8 c1-4	*h2m0*
s5	5.7 +/- 0.02	55.2	64.9	EVO8 c1-4	*h2m0*
s6	5.6 +/- 0.01	60.9	70.6	EVO8 c1-4	*h2m0*
s7	5.8 +/- 0.00	66.5	76.2	EVO8 c1-4	*h2m0*
s8	5.7 +/- 0.01	72.3	82.0	EVO8 c1-4	*h2m0*
s9	5.7 +/- 0.02	77.9	87.6	EVO8 c1-4	*h2m0*
Bank culture	18.4	18.4	28.1	EVO8 c6-c9	*h10m0*
s1	9.3 +/- 0.29	27.7	37.4	EVO8 c6-c9	*h10m0*
s2	6.3 +/- 0.21	34.1	43.8	EVO8 c6-c9	*h10m0*
s3	5.8 +/- 0.07	39.9	49.6	EVO8 c6-c9	*h10m0*

s4	5.7 +/- 0.01	45.6	55.3	EVO8 c6-c9	*h10m0*
s5	5.6 +/- 0.01	51.2	60.9	EVO8 c6-c9	*h10m0*
s6	5.7 +/- 0.01	56.9	66.6	EVO8 c6-c9	*h10m0*
s7	5.6 +/- 0.00	62.5	72.2	EVO8 c6-c9	*h10m0*
s8	5.7 +/- 0.01	68.1	77.8	EVO8 c6-c9	*h10m0*
s9	5.6 +/- 0.01	73.8	83.5	EVO8 c6-c9	*h10m0*
s10	5.7 +/- 0.01	79.4	89.1	EVO8 c6-c9	*h10m0*

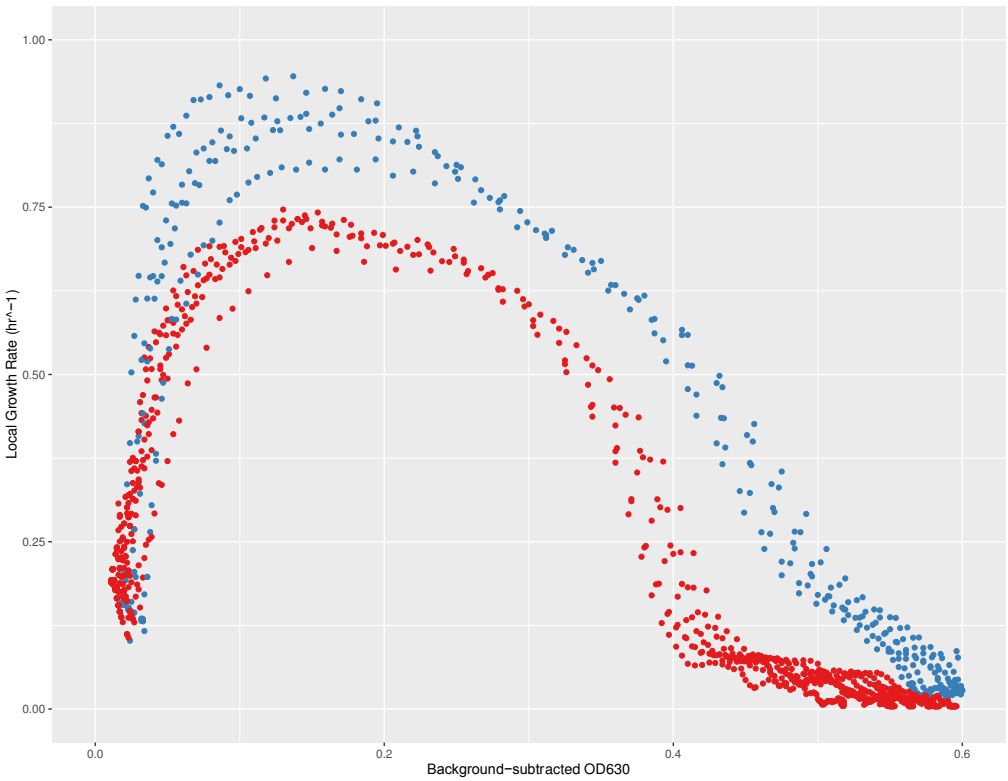


Figure S1. Related to Figure 1. Load of producing mevalonic acid (*h2m0, red*) as measured by comparison to growth of pathway-excised, non-producing control (*h8, blue*). Growth rates were dependent of the phases of growth, and therefore for quantification we use an average in the background-subtracted OD630 region 0.04-0.40, which we quantified to 30 % (n = 8) (Online Methods).

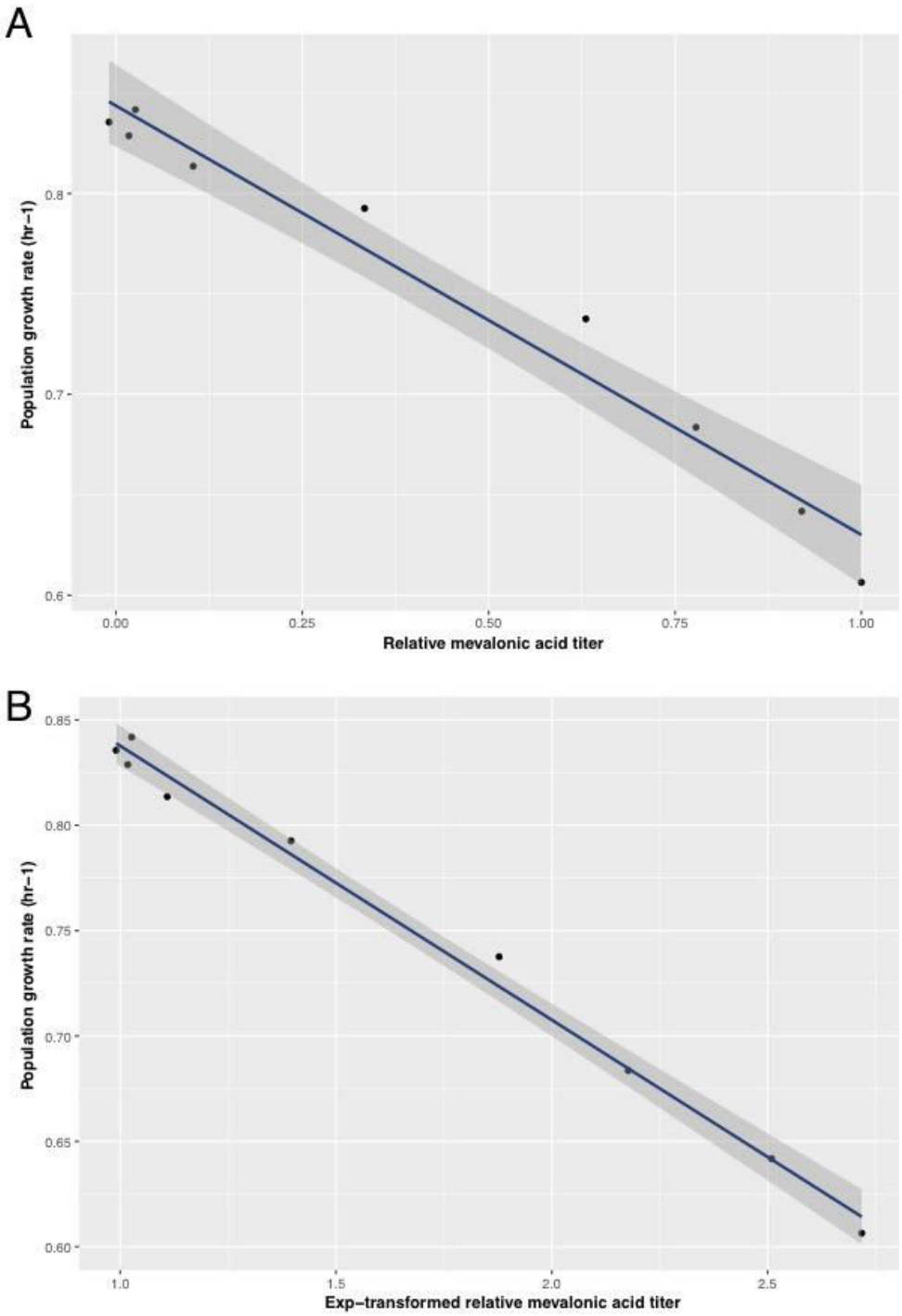


Figure S2. The relation between population growth rate average and mevalonic acid titer of *h2m0* and long-term cultured populations (n = 5) (relative to earliest data point in simulated fermentation "EVO2". Relative mevalonic acid titer is respectively A) non-treated, and B) log-transformed) resulting in the respective regressions shown as lines (grey area depicts 95 % confidence interval of regression): A: $y = -0.21x + 0.84$, $R^2 = 0.97$, $p = 2.0\ 10^{-6}$, and

B: $y = 0.968 - 0.13\ e^x$, $R^2 = 0.99$, $p = 1.3\ 10^{-8}$

Table S3. Related to Figure 1. Measured mevalonic acid titers (g/L) in earliest time point from the indicated simulated fermentation, se: standard error of the mean (EVO2: n = 5, EVO8: n = 4, EVO10: n = 3).

Strain clone bank	**Simulated fermentation ID**	**Cultivation condition**	**Mevalonic acid (g/L) +/- se**
h2 m0	EVO2	Std. medium, 30 deg. C	1.2 +/- 0.04
h2 m0	EVO8	Opt. medium, 30 deg. C	0.9 +/- 0.04
h10 m0	EVO8	Std. medium, 30 deg. C	0.5 +/- 0.05
h2 m1, m2, m3	EVO10	Std. medium, 32 deg. C	1.3 +/- 0.04
kle1#1 m1, m2, m3	EVO10	Std. medium, 32 deg. C	1.0 +/- 0.04
h11 m0, m2, m3, m4	EVO13	Std. medium, 30 deg. C	1.3 +/- 0.05

Table S4. Variation in the cell divisions (generations) till first mutation by four replicate runs of a stochastic version of our population escape model (script in Supplementary Item 3), performed on an 8-core desktop computer (Intel® Core™ i7-4770K CPU @ 3.50GHz × 8). Computation time scales with number of cells and ran >3 days to observe a first mutation (escape) in all four replicates.

Replicate	Generations to first mutation	No. of cells
1	19.42837	705542
2	19.44143	711956
3	20.23096	1230621
4	20.99365	2087940

Table S5. Related to Figure 2. Production loads and escape rates estimated from fermentation simulations by fit to model or by pure-culture measurements.

Parameter	***Strain* and condition**	**Method**	**Data source**	**Estimate value (generation^{-1}) and p-value**
Production load	*h2m0*, relative to *h8* std. medium	Growth rate average	Pure-culture growth curves	30 %
Escape rate	*h2m0*, std. medium	Fit to model using pure-culture determined production load	Production over time	$2.5\ 10^{-8}$ (p-value = 0.0012)
Escape rate	*h2m0*, std. medium	Fit to model using pure-culture determined production load	Total IS fraction over time (from deep-seq)	$8.7\ 10^{-8}$ (p-value < 0.0001)
Production load	*h2m0*, std. medium	Free fit to model	Total mobile element fraction over time (from deep-seq)	28 % (p-value < 0.0001)
Escape rate	*h2m0*, std. medium	Free fit to model	Total mobile element fraction over time (from deep-seq)	$2.1\ 10^{-7}$ (p-value < 0.0001)
Production load	*h2m0*, relative to *h8* opt. medium	Growth rate average	Pure-culture growth curves	23 %
Production load	*h10m0*, relative to *h9*	Growth rate average	Pure-culture growth curves	26 %

	std. medium			
Production load	*h2m0*, std. medium	Fit to model with NGS-determined escape rate	Production over time	26 % ($p < 0.0001$)
Production load	*h2m0*, opt. medium	Fit to model with NGS-determined escape rate	Production over time	21 % ($p < 0.0001$)

Table S6. SNPs found in the chromosomes of colonies picked from streaks from the end-point experimentally simulated fermentations of *h2m0* (std. medium, "EVO2"). SNPs were identified after mapping of reads to the publicly available genome sequence of *E. coli* DH10B (accession CP000948). SNPs were called at minimum coverage of 15, maximum coverage of 1000. Three colonies (k1-3) were randomly picked from lineage c6, c8 and c10 respectively.

Sample colony	**Region**	**Type**	**Reference**	**Allele**	**Coverage**	**Frequency**	**Probability**
TOP10	4272971	Insertion	-	T	36	97.22	1
c6s9k1	4272971	Insertion	-	T	25	100	1
c6s9k2	4272971	Insertion	-	T	24	100	1
c6s9k3	4272971	Insertion	-	T	46	97.83	1
c8s9k1	4272971	Insertion	-	T	35	97.14	1
c8s9k2	4272971	Insertion	-	T	40	95	1
c8s9k3	4272971	Insertion	-	T	42	100	1
c10s9k1	4272971	Insertion	-	T	25	100	1
c10s9k2	4272971	Insertion	-	T	24	100	1
c10s9k3	4272971	Insertion	-	T	46	97.83	1

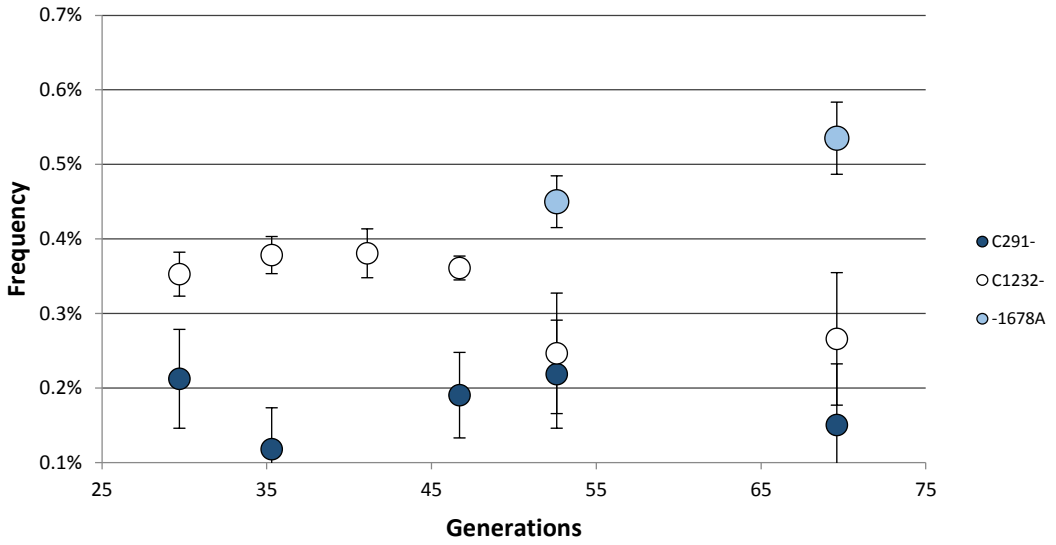


Figure S4. Mean frequencies of three detected single nucleotide variants in the dynamically sampled lineages c6, c8 and c10 of the experimentally simulated long-term fermentation of *E. coli h2 m0,* at a minimum detection level of 0.1 %. Error bars denote standard error (n = 3).

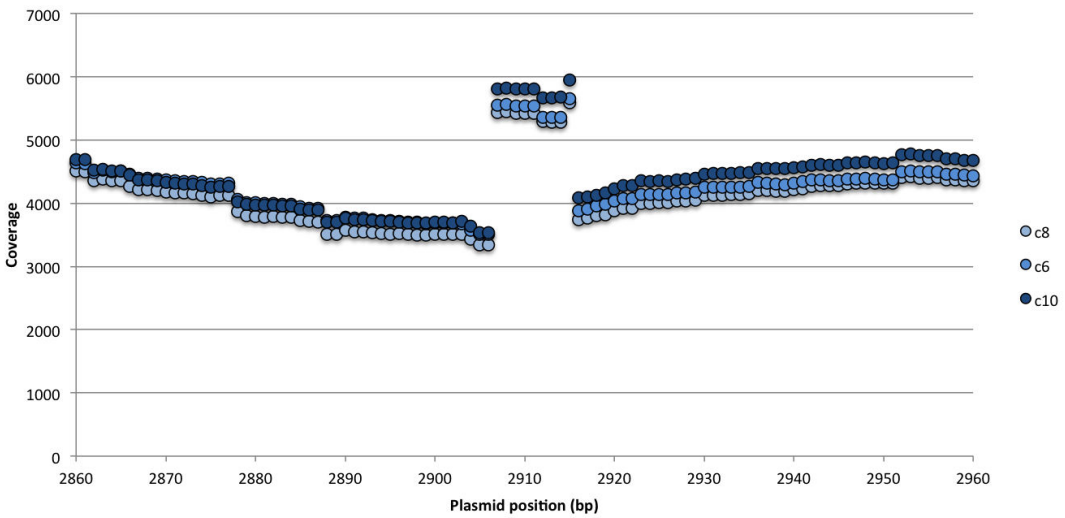


Figure S5. Related to Fig. 3. Number of mapped reads (coverage) at the reference plasmid positions in the proximity of a high-frequency IS10 insertion, likely resulting from duplication of the target recognition region (data from three parallel *h2m0* lineages c6, c8 and c10 shown following seed 9).

Table S7. Related to Fig. 4. Summary statistics of linear regressions of log-transformed mobile element subgroup fractions with cell divisions. Enrichment rates were calculated by linear regression of log10-transformed mobile element frequencies in the exponential phase seeds s2-s6 for the three time-lapse sequenced samples lineages c6, c8 and c10.

IS subgroup	Estimate	Std. Error	t value	Pr(>|t|)	R^2
IS10	0.103524	0.006623	15.63	0.000569	0.98
IS186	0.099346	0.004302	23.09	0.000178	0.99
IS5	0.052057	0.006731	7,734	0.0045	0.95
IS1	0.05529	0.007353	7,519	0.004875	0.95
IS3	0.066145	0.004971	13.31	0.000917	0.98
IS150	0.015975	0.003622	4.41	0.021625	0.86
tn1000	0.083741	0.003768	22.23	0.000199	0.99
IS2	0.022302	0.005604	3.98	0.028387	0.84
IS911	0.01759	0.00377	4,667	0.0186	0.88
IS30D	0.03692	0.01234	2,993	0.05799	0.75

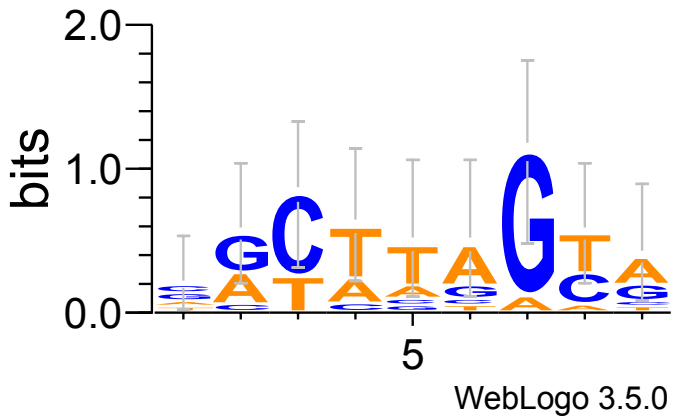


Figure S6. Consensus sequence for ten IS10 insertion sites observed in the deep-sequenced pMevT plasmid populations, as analyzed using WebLogo 3 [62]. This observed consensus deviated from a previously reported consensus target sequence of IS10: NGCTNGACN [63] .

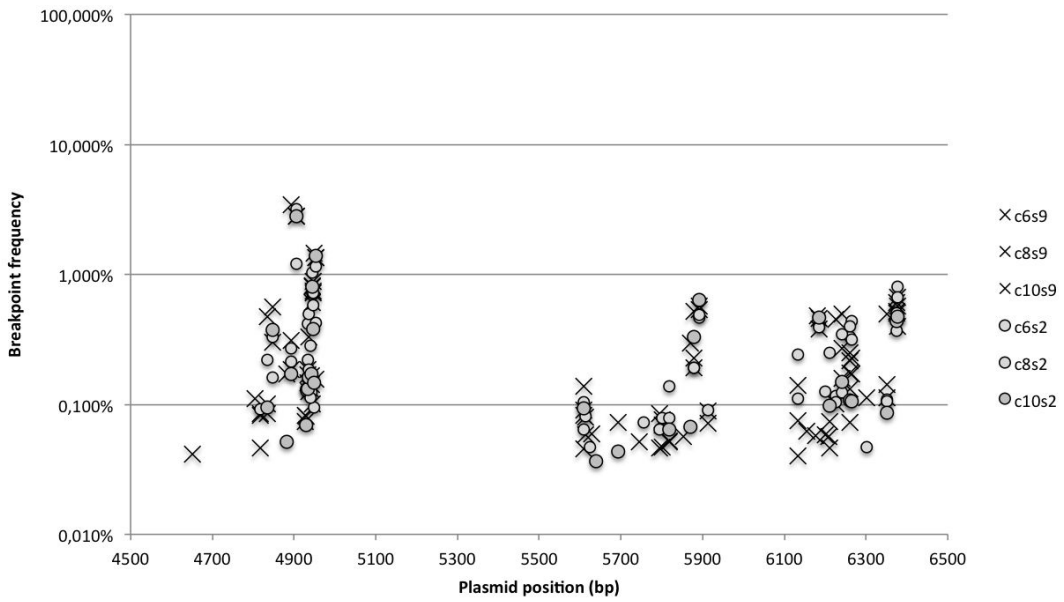


Figure S7. Exemplary structural variation (pMevT plasmid backbone, region 4500-6500 bp featuring the *rnnB* terminator and f1 origin) without enrichment over the course of the experimentally simulated fermentation of *E. coli h2m0* "EVO2" (breakpoint frequency = breakpoint reads per coverage) lineages c6, c8 and c10, sampled after seed 2 (s2) and final seed 9 (s9).

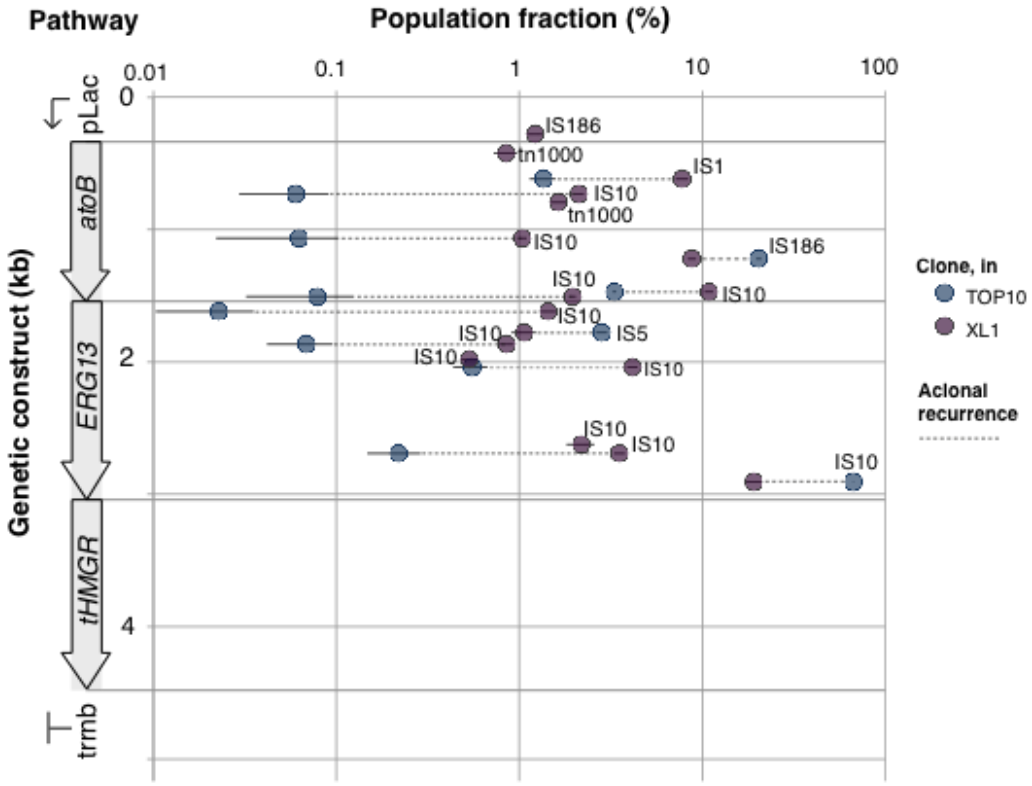


Figure S8. Recurring mobile element disruptions within the load-carrying genetic regions of the mevalonic acid pathway in two *E. coli* production cell banks as assessed by ultra-deep sequencing of pathway populations at the end of experimentally simulated long-term fermentation ("EVO2B") with *E. coli* Top10 (*h2m0*) and XL1 (*XL1-MevTm0*) hosts. Error bars indicate standard errors (n = 5), and dashed lines connect elements recurring across the Top10 and XL1 experimentally simulated fermentations.

Supplementary Item 1

A simple mathematical model for the fraction of producers in time is established by solution of an ordinary differential equation system (eq1 and eq2) with analogy to models of plasmid loss dynamics[33].

p(t): producer cells in time

np(t): non-producer cells in time

μ: specific growth rate

ρ : production load = 1 - $\frac{\mu_p}{\mu_{np}}$

(eq1): $\frac{dp(t)}{dt} = \mu_p \cdot p(t) - k_{escape} \cdot p(t)$

(eq2): $\frac{dnp(t)}{dt} = \mu_{np} \cdot np(t) + k_{escape} \cdot p(t)$

Non-producer cells are converted from producer cells at a rate k_{escape} and such non-producers alleviate a reduction in growth rate due to production load ρ. The solution growth functions (eq3 and eq4) assume a pure initial inoculum of a single, producing cell, and can be found by first solving eq1 and inserting the resulting p(t) in eq2.

(eq3): $p(t) = e^{(\mu_p - k_{escape})\cdot t}$

(eq4): $np(t) = k_{escape} \frac{1-e^{(\mu_p - \mu_{np} - k_{escape})t}}{k_{escape} + \mu_{np} - \mu_p} e^{\mu_{np}\cdot t} = \mathrm{k_{escape}} \frac{1-\mathrm{e}^{-(\cdot\rho\cdot\mu_{np} + \mathrm{k_{escape}})\mathrm{t}}}{\mathrm{k_{escape}} + \rho\cdot\mu_{np}} \mathrm{e}^{\mu_{np}\cdot \mathrm{t}}$

From these, the fraction of producers in time (eq5) can be derived.

(eq5): $\frac{p(t)}{p(t)+np(t)} = \frac{e^{(\mu_p - k_{escape})\cdot t}}{k_{escape}\frac{1-e^{-(\rho\cdot\mu_{np} + k_{escape})t}}{k_{escape} + \rho\cdot\mu_{np}} e^{\mu_{np}\cdot t} + e^{(\mu_p - k_{escape})\cdot t}} = \frac{k_{escape} + \rho\cdot\mu_{np}}{k_{escape}\cdot e^{(k_{escape} + \rho\cdot\mu_{np})\cdot t} + \rho\cdot\mu_{np}}$

Further, by defining u_p relative to u_{np} = 1, the model is only dependent on the relative growth rate (ρ) and the escape rate (k_{escape})

S1 Item 2

```
#Script to analyze growth rate measurement
#Required packages
library(ggplot2)
library(reshape2)
library(scales)
library(plyr)
library(drc)
library(zoo)
library(stringr)


#Define TimeConverter function
TimeConverter <- function(z){
  cTime <- as.character(z[,1])
  Time <- (sapply(strsplit(cTime,":"),
                  function(x) {
                    x <- as.numeric(x)
                    round(x[1]*60+x[2]+x[3]/60, digits = 0)
                  }
  )
  )
  return(Time)
}

#Mac
data <- na.omit(read.delim("~/Google Drive/CFB/Metabolic fitness dynamics (Andreas og Peter)/R scripts/151130 data/c6-c10_s1-s9_Data.csv",
header=T, skip =0, sep = ";"))
map <- read.csv("~/Google Drive/CFB/Metabolic fitness dynamics (Andreas og Peter)/R scripts/151130 data/c6-c10_s1-
s9_Growth_map_EVO2_populations.csv",sep=";", header = T, colClasses = c(rep("character", 2), "numeric"))

#Convert time to min
names(data)[1] = "Time" #Standardize naming
data$Time <- TimeConverter(data)

#AP: Melt the data if not in long format already
x.data <- melt(data, id = "Time")


#AP: Match values from mapping file to data
x.data$Strain<-map[match(x.data$variable, map$Well),2]
x.data$Replicate<-as.factor(map[match(x.data$variable, map$Well),3])
names(x.data)<-c("Time","Well","OD","Strain","Replicate")

x.data$OD<-as.numeric(x.data$OD)

#Set bgOD:
avg.bkg = 0.115

#Subtract background
x.data$bkgOD <- x.data$OD - avg.bkg
#PR: Add ln data
x.data$lnOD <- log(x.data$bkgOD)

#Add time in hours
x.data$Timehr <- x.data$Time/60


#PR: First subset x.data to exclude the BKG wells
x2.data <- subset(x.data, lnOD != "NA" & Strain != "BKG")
x2.data$lnOD <- as.numeric(str_replace_all(x2.data$lnOD, "Inf", "0"))
x2.data$lnOD <- as.numeric(str_replace_all(x2.data$lnOD, "NA", "0"))
x2.data$lnOD <- as.numeric(str_replace_all(x2.data$lnOD, "NaN", "0"))

#AP: Define the window size of rolling regression
WinSize <- 5

#AP: Define time and OD values for regression:
x <- as.numeric(x2.data$Timehr) #E.g. hours or minutes.
y <- as.numeric(x2.data$lnOD)
z <- as.data.frame(cbind(x,y))

#Standardize naming:
names(z)<-c("Time","lnOD")

#Fit data with a
RollFit<-rollapply(zoo(z), width=WinSize,
function(Z) {
reg<-lm(formula=Time~lnOD, data = as.data.frame(Z)) #Apply linear regression to lnOD against time.
cbind(1/coef(reg)[2],summary(reg)$r.squared)}, #calculate 1/slope
```

```
by.column=FALSE, align="right")

#Standardize naming:
names(RollFit)<-c("LocalGrowthRate","Rsquared")

#Attach to existing dataframe *NB: Fit values start at the row corresponding to WinSize!
x3.data <- cbind(x2.data[WinSize:nrow(x2.data),],as.data.frame(RollFit))

xr.data<-subset(x3.data)
```

SI Item 3

R script

```
library(nls2)

#Load total mobile element fractions (NGS-determined)
ISobs <- data.frame(c(29.7,35.3,41.1,46.7,52.5,69.9))
colnames(ISobs)[1] <- "generations"
ISobs$fraction <- c(0.991197834,0.980170492,0.919041917,0.733211699,0.338929329,0.000657726)
z <- nls2(fraction ~ ((u+ro))/((u)*exp((u+ro)*generations)), data = ISobs,
      start = list(u=0.005,ro=0.1), control = list(maxiter = 5000))
summary(z)
```

SI Item 4

R script

```
#Load required libraries
library(ggplot2)
library(foreach)
library(doParallel)

#setup parallel backend to use multiple processors
cl<-makeCluster(4)
registerDoParallel(cl)


###set parameters for growth of producers (gP), non-producers (pNP) and escape rate (k)###
parms=c(gP=0.72,gNP=1,k=2.1E-7)
initial=c(P=29, NP=1)
time.window=c(0, 15)

# define how state variables for cell division and mutational escape
processes <- matrix(0, nrow=3, ncol=2,
                    dimnames=list(c("birth P",
                                    "Escape P",
                                    "birth NP"),
                                  c("P","NP")))

processes[1,1]=1 #If birth of P
processes[2,1]=-1 #If escape of P
processes[2,2]= 1
processes[3,2]=1 #If birth of NP

# process probabilities
probabilities <- function(state){
  P<-state[1] #Define parameters
  NP<-state[2]
  gP<-parms[1]; gNP<-parms[2];k<-parms[3]

  a1<-gP*P #P birth
  a2<-k*P #Escape
  a3<-gNP*NP #NP birth

  a<-c(a1,a2,a3)
  names(a)<-c("a1","a2","a3")
  a
}
```

```
### Initiate parallelized loop: ###
ls <- foreach(n = 1:4) %dopar% {

 # initialize state and time variables and write them into output table
 state <- initial
 time <-  time.window[1]

 # define output dataframe
 output <- data.frame(t=time,
                P=state["P"], NP=state["NP"],
                row.names=1)


 #start timer to get time of  simulation
 strt<-as.numeric(Sys.time())

#Define stop-conditions for each simulation e.g. fraction of producers or absulute time (in seconds)
 while(state["P"]/(state["P"]+state["NP"])>0.5 & as.numeric((Sys.time()-wst))<252000){

  #calculate process probabilities for current state
  a<-probabilities(state)

  #WHEN does the next process happen?
  tau<-rexp(1,rate=sum(a))

  #update time
  time = time+tau

  #WHICH process happens after tau?
  act<-sample(length(a), 1, prob=a)

  #Update states
  state<-processes[act,]+state

  #write into output
  output <- rbind(output,c(time,state))
 }

 #Add info on e.g. replicate and parameters
 output$ER <- "2.1*10^-7" #Escape rate
 output$rep <- n #Replicate
 output$frac <- output$P/(output$P+output$NP) #Producer fraction

 #Write each simulation to a file (optional)
 write.table(output, file=file.path("File path", paste("Filename",".txt", sep="")),
          sep="\t", row.names=F)

 output

}

#Show simulation time
as.numeric((Sys.time()-wst))/60/60

#End parallelization
stopCluster(cl)

#Combine output to a dataframe
SDF<-as.data.frame(ls[[1]])

for( i in 2:length(ls)) {
 SDF <- rbind(SDF,as.data.frame(ls[[i]]))
}

#Calculate generations
SDF$Gen <- log10(SDF$P+SDF$NP)/log10(2)

#Plot output
```

```
ggplot(SDF, aes(x = Gen, y = frac, color = as.factor(rep)))+
  geom_line(size=1)+
  ylab("Producer fraction")+
  xlab("Generations")+
  ggtitle("Stoch. simulation")
```

Manuscript V

Biochemical mechanisms determine the functional compatibility of heterologous genes

Andreas Porse[1], Thea S. Schou[1], Christian Munck[1], Mostafa M. H. Ellabaan[1], Morten O.A. Sommer[1]*

[1]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs. Lyngby, DK-2800, Denmark

*Corresponding author: *msom@bio.dtu.dk*

Abstract

Elucidating the factors governing the functional compatibility of horizontally transferred genes is important to understand bacterial evolution, including the emergence and spread of antibiotic resistance, and to successfully engineer biological systems. *In silico* efforts and work using single-gene libraries have suggested that sequence composition is a strong barrier for the successful integration of heterologous genes. Here we sample 200 diverse genes, representing >80% of sequenced antibiotic resistance genes, to interrogate the factors governing genetic compatibility in new hosts. In contrast to previous work, we find that GC content, codon usage and mRNA-folding energy are of minor importance for the compatibility of mechanistically diverse gene products at moderate expression. Instead, we identify the phylogenetic origin, and the dependence of a resistance mechanism on host physiology, as major factors governing the functionality and fitness of antibiotic resistance genes. These findings emphasize the importance of biochemical mechanism for heterologous gene compatibility, and suggest physiological constraints as a pivotal feature orienting the evolution of antibiotic resistance.

Introduction

A distinct feature of prokaryotes is their ability to share genetic material via horizontal gene transfer[1]. Such open-source evolution provide rapid access to the genetic innovations that continuously shape bacterial genomes[2]. The transfer of genes between bacteria happens primarily via transferrable genetic elements such as plasmids, phages or direct DNA uptake (transformation)[3]. Although such transfer events are believed to occur frequently in nature, the fundamental forces governing the establishment and maintenance of successfully transferred genes are poorly understood[2,4]. Because foreign genes may lose functionality or impose a high biological cost in new hosts, this gap in our understanding also limits current synthetic biology efforts focused on engineering novel functions into biological systems[5,6].

From computational studies of sequenced genomes, the tendency of a gene to be transferred has been inferred to depend mainly on ecological and phylogenetic factors[1,4]. As physical proximity of donor and recipient bacteria is generally required for transfer, ecology has been suggested as a strong dissemination barrier[4]. However, the broad-host range of some transfer mechanisms, and the ubiquitous presence of antibiotic resistance genes across environments, suggest that ecological barriers are largely governed by functional constrains[7,8]. Indeed, numerous studies reveal a functional bias of transferred gene categories, with factors performing largely independent tasks, e.g. those involved in secondary metabolism or virulence, being transferred more often than genes encoding highly interactive proteins involved in transcription and translation[9–12]. Furthermore, metagenomic analyses of DNA from different environments suggest that genes involved in antibiotic resistance are highly confined by phylogeny indicating that genomic factors are important for the acquisition or maintenance of foreign genetic material[4,13,14]. In support of this idea, *in silico* studies have shown a bias in the nucleotide composition of transferred genes in relation to the recipient genome, and suggest that sequence parameters influence successful gene integration[15–17]. Specifically, a role of the codon usage and GC-content on the functional integration of acquired genes has been inferred due to their potential role in gene expression and fitness[18,19]. Upon successful transfer, host compatibility and selection is crucial for the long-term persistence of newly acquired genetic material, as costly genes will eventually be lost through purifying selection[20–23]. Experimental studies characterizing the phenotypic effects of synonymous sequence variation have largely focused on a limited set of phenotypes and most current data is obtained from fluorescent protein expression[24–26]. However, recent work by Kacar and Garmendia *et al.* showed an increased fitness cost of replacing the elongation factor Tu (a highly conserved protein involved in translation) with its distant homologs in *E. coli,* and similar results have been obtained by Lind *et al.* when

replacing ribosomal subunits in *Salmonella typhimurium*[27,28]. While the fitness cost of expressing these core translational genes could not be attributed to differences in sequence composition, Amorós-Moya *et al.* showed that sub-optimal codon usage of a highly-expressed chloramphenicol acetyl transferase resistance gene resulted in lower resistance levels and overall host fitness[29]. Although existing studies provide interesting clues on the relation among sequence composition, gene expression and fitness, the relevance of these findings regarding the functionality of diverse naturally occurring accessory genes involved in antibiotic resistance is not clear.

Antibiotic resistance genes are ubiquitously present on mobile genetic elements that allow their extensive dissemination among bacteria[3]. Given the stagnation in antibiotic discovery, the increasing prevalence of multidrug resistant bacteria constitutes an urgent threat to public health that motivates a deeper understanding of the forces governing the dissemination and long-term maintenance of antibiotic resistance genes[30–32].

Antibiotic resistance is conferred through five major mechanisms: i) enzymatic drug inactivation, ii) active drug efflux, iii) modification of drug target, iv) replacement of the drug target with a resistant variant, and v) regulatory shifts towards a more resistant phenotype[33]. The diversity of mechanisms through which antibiotic resistance is achieved makes antibiotic resistance genes a valuable model system for investigating the factors that may affect the functional compatibility of transferable genes in general.

In this study, we employed a synthetic bottom-up approach by sampling a broad sequence space of 200 diverse open reading frames annotated as antibiotic resistance genes. Via experimental profiling of these genes, we discovered that resistance mechanisms and the phylogenetic relatedness of donor and recipient species act together as important determinants of gene functionality and fitness cost. Consequently, we suggest that these effects dominate the potential transfer barriers imposed by sub-optimal sequence composition of heterologous genes.

Results

By clustering all genes of major publicly available antibiotic resistance gene databases, and selecting the most abundant genotypes, we obtained 200 genes for DNA synthesis (see materials and methods, **Supplementary Fig. 1a** and **Supplementary Table 1**). The selected clusters represented the most abundant genes in resistance gene databases (>80%, **Supplementary Fig. 1a**) and accounted for 98 % of the total dataset homologues in general sequence databases (NCBI NT and Genomes) (**Supplementary Fig. 1b**).

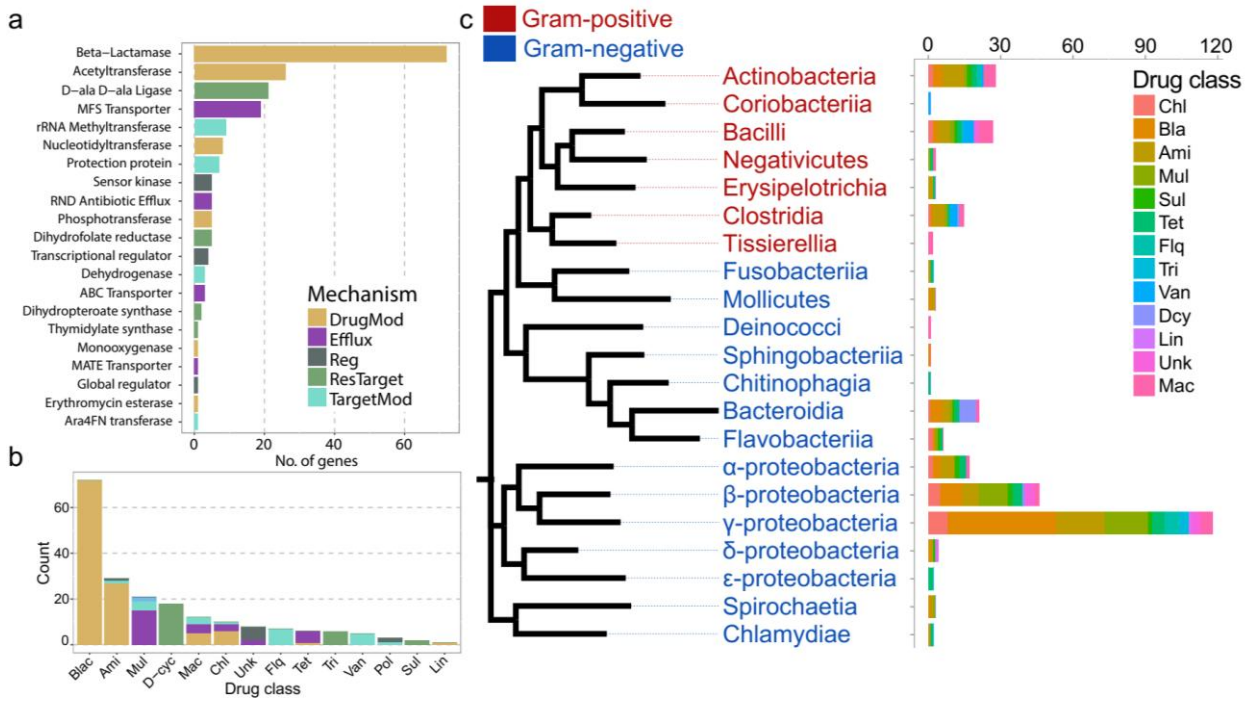


Fig. 1 | Database mechanistic and phylogenetic diversity of 200 synthesized genes. (**a**) Biochemical functions of the included antibiotic resistance genes obtained via Resfams. These genes are grouped into five major mechanistic categories (relative abundance shown in parenthesis): DrugMod = Drug inactivating enzymes (56.5%), Efflux = Efflux pumps (14%), Reg = Regulators (5%), ResTarget = Redundant/resistant target (14.5%), TargetMod = Target modifying/binding proteins (10%). (**b**) Gene counts stratified by resistance class and fractionated by resistance mechanisms. (**c**) Phylogenetic and drug class distributions of the included genes. Genes that could be identified (97% identity) in one or more genomes deposited in RefSeq were quantified for each bacterial phylogenetic class. The colouring of each bar depicts the distribution of the annotated resistance. Drug class aberrations: Chl = Chloramphenicol, Bla = β-lactams, Ami = Aminoglycosides, Mul = Multiple drug classes, Sul = Sulfamethoxazole, Tet = Tetracyclines, Flq = Fluoroquinolones, Unk = Unknown, Tri = Trimethoprim, Van = Vancomycin, Dcy = D-cycloserine, Lin = Lincosamides, Mac = Macrolides.

The 200 selected genes vary widely in resistance mechanisms, targeted drug classes and phylogenetic dissemination (**Fig. 1, Supplementary Table 1**). 64% of the selected genes were associated exclusively with Gram-negative organisms, and of the remaining genes, 13 % were found in Gram-positive organisms, and 10.5 % were found in both Gram-negative and Gram-positive organisms (Fig. 1c). However, 12.5 % of the genes had not been associated with a particular host organism (mostly genes found via metagenomic functional selections). The selected genes were annotated to confer resistance to 11 distinct drug classes via 23 distinct biochemical functions that can be divided into five major mechanistic categories based on their Resfam annotations (**Fig. 1a and**

b)[34]. In addition to genes annotated to confer resistance towards known drug classes, genes annotated to confer antibiotic resistance but without defined antibiotic resistance profiles were included as a consequence of their high abundance in the public antibiotic resistance databases. All of these genes were annotated as having regulatory or efflux functions but were not associated with specific drug classes in the literature (**Supplementary Table 1**).

Functional characterization of putative resistance genes reveals diverse functionality and growth profiles

The 200 selected genes were cloned in a low-expression setup in *E. coli* MG1655 (**Supplementary Fig. 1**). All genes were subjected to 20 phenotypic tests and growth rate measurements. Specifically, the resistance towards 20 antibiotics comprising 12 chemical classes was assessed. A gene was considered a functional resistance gene if it conferred a resistance phenotype of at least a 2-fold increase in the minimal inhibitory concentration (MIC) compared to that of *E. coli* MG1655 carrying the empty vector (**Supplementary Table 2**). Clones displaying a resistance phenotype were subjected to growth rate assessments under non-selective conditions and further antibiotic susceptibility testing in a range from 2- to 30-fold the MIC (**Fig. 2**).

Whereas only 32% of the 200 tested genes could be identified in *E. coli* genomes deposited in NCBI's RefSeq database, 63 % of the genes displayed at least one resistance phenotype in *E. coli*. The resistance phenotypes were distributed unequally among the drug classes, with genes annotated to confer resistance towards tetracyclines, sulfonamides, β-lactams and fluoroquinolones having a high proportion (> 80%) of functional variants. By contrast, D-ala ligases, which confer resistance to D-cycloserine in *E. coli*, as well as genes annotated to confer resistance towards polymyxins or multiple drug classes, showed the lowest proportion of functional genes in our experimental setup (**Fig. 2**, **Supplementary Table 1**). Genes conferring resistance towards β-lactams, aminoglycosides, chloramphenicol and trimethoprim showed the highest average levels of resistance, whereas those conferring resistance towards fluoroquinolones and D-cycloserine displayed the lowest average increase in resistance compared to the susceptible WT (**Fig. 2**). This difference in resistance level between drug classes was statistically significant (ANOVA, $P < 0.001$). We further investigated if the level of resistance could be attributed to differences the codon adaptation index (CAI), GC-content and mRNA-folding energy; however we could not detect any significant correlations between these sequence parameters and the resistance level for the total dataset nor within resistance classes (**Supplementary Fig. 3**).

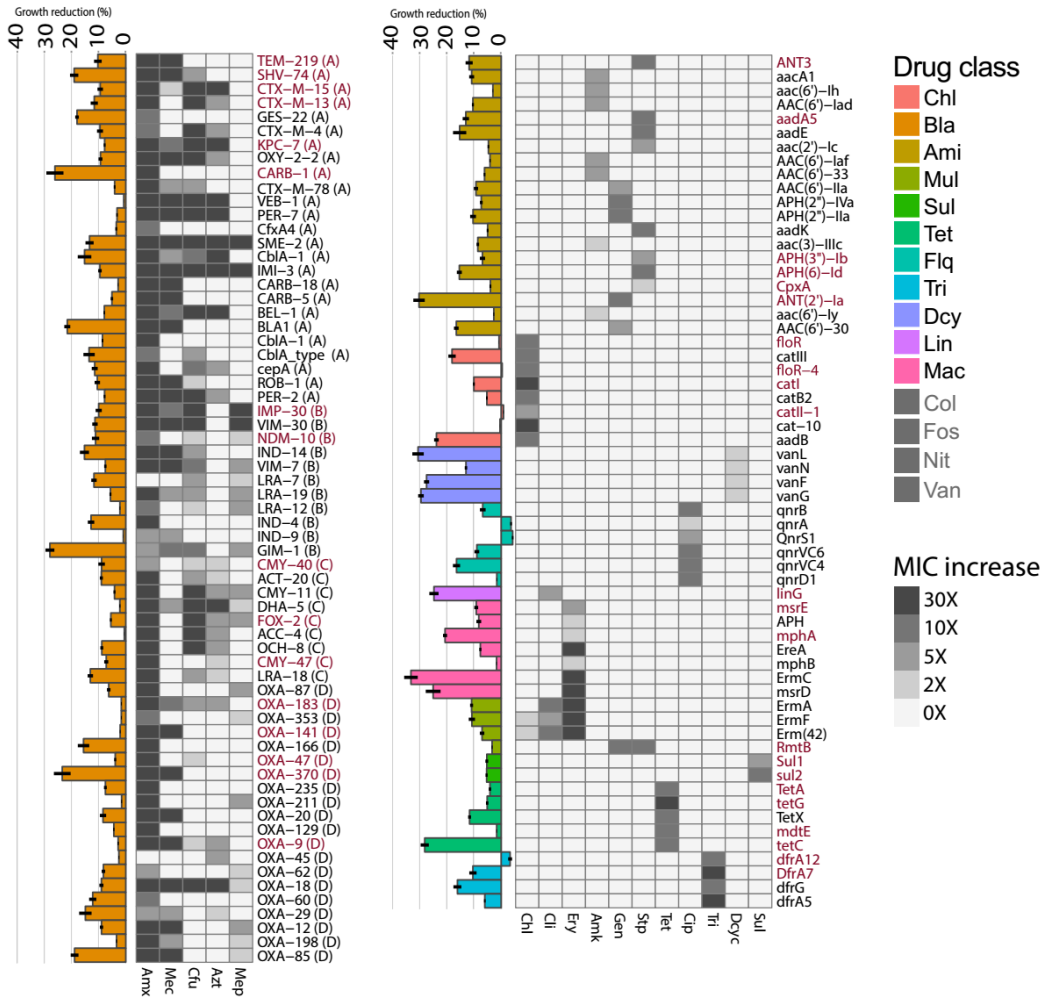


Fig. 2 | The resistance level and relative growth rate of each functional resistance gene in *E. coli*. The heatmaps show the resistance profile (fold change in MIC compared to *E. coli* MG1655 carrying the empty expression backbone) for all functional resistance genes and are grouped by drug class. The β-lactamases (left) are clustered by their molecular (Ambler) classes as shown in parentheses[35]. Bars represent the mean of at least three repeated growth measurements and are normalized to *E. coli* MG1655 carrying the empty expression vector. Error bars show the std-error of the mean. The genes highlighted in red are present in sequenced *E. coli* genomes deposited in NCBI's RefSeq genome database. The drug classes are as follows: Chl = Chloramphenicol, Bla = β-lactams (Amx = Amoxicillin, Mec = Mecillinam, Cfu = Cefotaxime, Azt = Aztreonam, Mep = Meropenem), Ami = Aminoglycosides, Mul = Multiple drug classes, Sul = Sulfamethoxazole, Tet = Tetracyclines, Flq = Fluoroquinolones, Tri = Trimethoprim, Dcy = D-cycloserine, Lin = Lincosamides, Mac = Macrolides, Col = Colistin, Fos = Fosfomycin, Nit = Nitrofurantoin, Van = Vancomycin. Grey coloured drugs were tested, but no genes conferred resistance towards these in *E. coli*.

Sequence composition is not a major compatibility barrier of acquired resistance genes

As 68% of the tested genes from our data set have not yet been identified in *E. coli* genomes, our dataset is well suited for studying factors relevant to the functionality of resistance genes evolved in

a foreign genetic context. The selected genes varied widely in their base composition (**Fig. 3**), which has previously been hypothesized to affect functional expression and successful gene transfer [16,18,19]. The codon usage of an incoming gene might influence its protein expression and it is generally believed that the CAI is important for heterologous gene integration[18,24,36,37]. Surprisingly, we found that the average *E. coli* CAI of functional resistance genes was slightly lower compared to that of the non-functional genes (Mann-Whitney U-test, P = 0.014; **Fig. 3a**). Yet, we found no significant difference in the average GC-content (Mann-Whitney U-test, P = 0.11) (**Fig. 3b**) between the functional and non-functional genes. To investigate whether interactions between these and other parameters would influence the outcome of our analysis, we built a multivariate logistic regression model (**Supplementary Table 3**). The inclusion of GC-content did not significantly change the predictive power of the model compared to CAI alone (P = 0.70), suggesting that the effect of the CAI showed a limited dependence on GC-content (**Fig. 3d**). The folding energy of the N-terminal of a transcript may also influence gene expression[24,26,38], however, N-terminal mRNA-folding energy did not predict functionality, nor resistance level of functional genes, in our logistic regression model (**Fig. 3c**, **Supplementary Fig. 3 and Supplementary Table 3**).

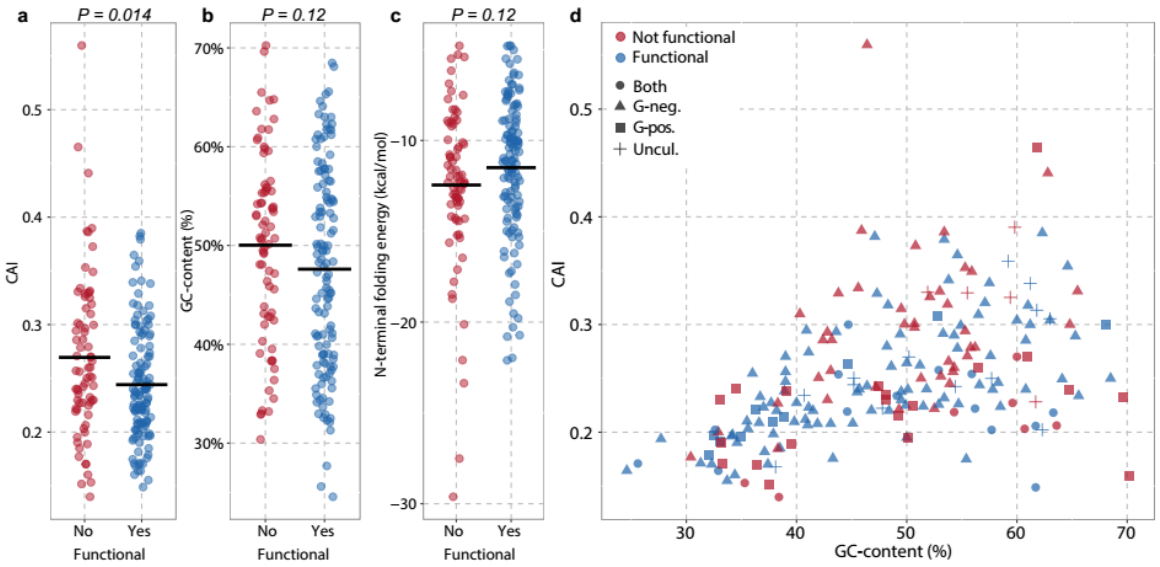


Figure 3 | Divergent sequence composition is not a major barrier for functionality of a foreign antibiotic resistance gene in *E. coli*. a) Comparison of the codon usage between functional and non-functional genes **b)** Comparison of the GC-content between functional and non-functional genes **c)** Comparison of the mRNA N-terminal folding energy between functional and non-functional genes **d)** Functional and non-functional genes are dispersed throughout the sequence space. Point shape indicates whether a gene has been observed exclusively in Gram-negative organisms, Gram-positive organisms, in both Gram-positive and Gram-negative organisms, or has not yet been associated with a sequenced genome (Uncultured).

Although gene functional compatibility exhibited little dependence on sequence composition, it may affect the fitness cost in new hosts. As a proxy for fitness, we measured the growth rate of *E. coli* expressing each of the functional resistance genes (**Fig. 2**). Compared to *E. coli* carrying the empty expression vector, the impact of expressing a resistance gene on the growth rate ranged from a slight increase to more than a 30 % decrease in growth rate (**Fig. 2 and Supplementary Fig. 4**).

The reduced growth rate resulting from expression of a resistance gene differed significantly between the drug classes to which the genes conferred resistance (one-way ANOVA, $P > 0.001$; **Fig. 2**), but the impact on growth did not differ between mechanistic categories (one-way ANOVA, $P = 0.38$; **Supplementary Fig. 5**). The gyrase-protecting *qnr* fluoroquinolone resistance genes showed exceptionally low costs compared to other resistance classes, especially the D-alanine-D-alanine ligases. These confer D-cycloserine resistance in *E. coli* through target replacement, and had a high negative impact on growth rates (**Fig. 2**).

Contrary to current thinking, some antibiotic resistance genes from the *qnr* and *dfr* families were slightly advantageous to *E. coli* in the absence of antibiotic selection (**Fig. 2**). We detected beneficial growth patterns of *qnrA* as well as *qnrS1,* which are found, for example, in *Shigella* and *Salmonella* species that are closely related to *E. coli*, but not for the *qnr* genes found in more distantly related species. Despite a tendency towards a trade-off between resistance level and growth rate for target-modifying ($r = 0.62$, $P = 0.04$), this trend was opposite for resistant target ($r = -0.61$, $P = 0.06$) and efflux mediators ($r = -0.59$, $P = 0.12$), and no correlation was observed between growth reduction and resistance level for the drug modifying mechanistic class ($r = 0.02$, $P = 0.84$; **Supplementary Fig. 6**).

It has previously been recognized that gene level parameters such as the CAI and GC-content of codon-scrambled genes can influence host fitness when expressed in *E. coli*[24,25]. *E. coli* expressing the trimethoprim resistance gene *dfrG* of the dihydrofolate (*dfr*) family, originating from *Staphylococcus aureus,* displayed a lower growth rate compared to *E. coli* expressing *dfr* genes of Gram-negative origin (**Fig. 2**). Indeed, the *dfrG* gene had a low GC-content (32%) and CAI (0.17), which we hypothesized could negatively influence the growth of *E. coli*. To test whether base composition could be optimized to decrease the growth reduction imposed by *dfrG,* we synthesized a codon-optimized variant with a higher GC-content (55%) and CAI (0.66). However, this variant did not show a significant growth improvement compared to the wild-type *dfrG* gene when expressed in *E. coli* (Mann-Whitney U-test, $P = 0.1$; **Supplementary Fig. 7**), suggesting that factors beyond the nucleotide sequence are responsible for the differential cost.

To assess the general trend in our dataset, we investigated the correlations between growth rate effects and the CAI, GC-content, N-terminal mRNA-folding energy and gene length (**Fig. 4**). Although there was a tendency towards higher growth rates in genes with an increasing CAI, in contrast to previous studies on sequence variants of *gfp*[24,25], we found no significant correlation between these individual sequence parameters and the growth rate of *E. coli* for our diverse set of antibiotic resistance genes.

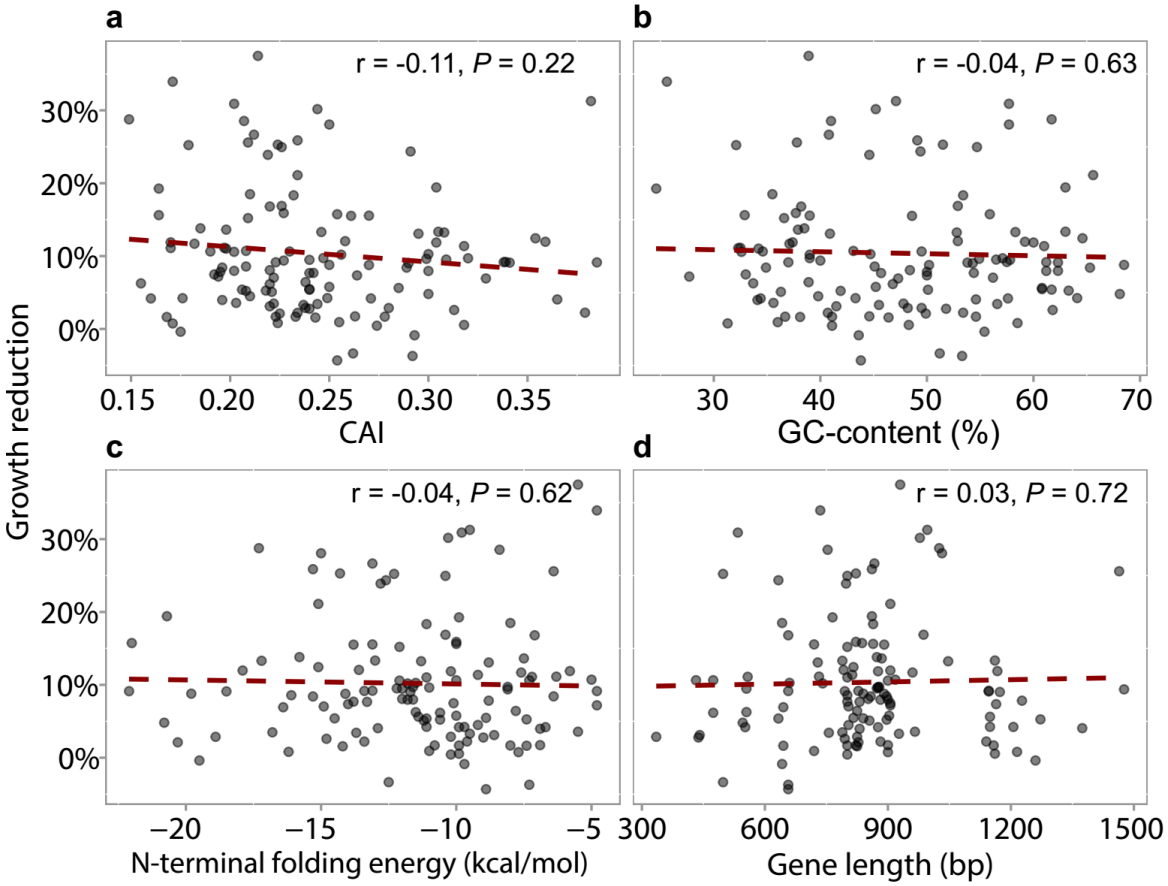


Figure 4 | The cost of resistance gene acquisition shows little or no dependence on sequence composition. **(a)** The CAI showed a small (r = -0.11) but non-significant (*P* = 0.22) correlation with growth reduction. Similarly, the GC-content(r = -0.04) **(b)**, N-terminal mRNA-folding energy (r = -0.04) **(c)** and gene length (r = 0.03) **(d)** did not correlate significantly with the growth reduction experienced by *E. coli* upon receiving the resistance genes in our expression setup (*P* > 0.05).

A multiple linear regression model of our growth data against the CAI, GC-content, N-terminal mRNA-folding energy and gene lengths explained virtually no variations in the growth rate (R^2= -0.01, *P* = 0.6; **Supplementary Table 4**). This result indicated that other factors dominate the potentially minor growth effects imposed by the base and codon composition in our dataset.

Resistance mechanism is a major determinant of gene functional compatibility

The functional compatibility of a gene in a new host may depend on its interaction with specific components of host physiology and metabolism[9]. Antibiotic resistance genes mediate their phenotype through a wide range of cellular interactions, with some mechanisms being dependent on specific host structures, e.g., ribosomal structure or the cell envelope for target modifying and efflux mechanism respectively, whereas drug modifying enzymes act directly on the antibiotic compound[33]. Although we observed a small significant difference in the proportion of functional genes for different targeted drug classes (χ^2, P = 0.031), we also noted that genes annotated to confer resistance to a specific drug class are frequently dominated by specific resistance mechanisms (Fig. 1b)[39]. Accordingly, the mechanistic category of a gene was far better at predicting the functionality of a gene transferred to *E. coli* (χ^2, P < 0.001) than the targeted drug class alone. This result suggested that certain resistance mechanisms are easier to integrate into a novel host physiology than others (**Fig. 5a** and **Supplementary Table 3**). The highest proportion of functional genes was found among the drug-modifying enzymes, including the β-lactamases and aminoglycoside transferases, with most genes conferring high levels of resistance (**Fig. 5a**), and this distribution was not significantly biased by the phylogenetic affiliation of these genes (χ^2, P = 0.08) or whether they had previously been identified in *E. coli* (χ^2, P = 0.77). By contrast, genes conferring resistance through efflux and regulatory mechanisms were least likely to function in *E. coli,* in which only 28.5 % and 10 % of genes displayed resistant phenotypes, respectively. These findings are consistent with the hypothesis that genes involved in limited cellular interactions are more likely to be functionally compatible in a new host[9].

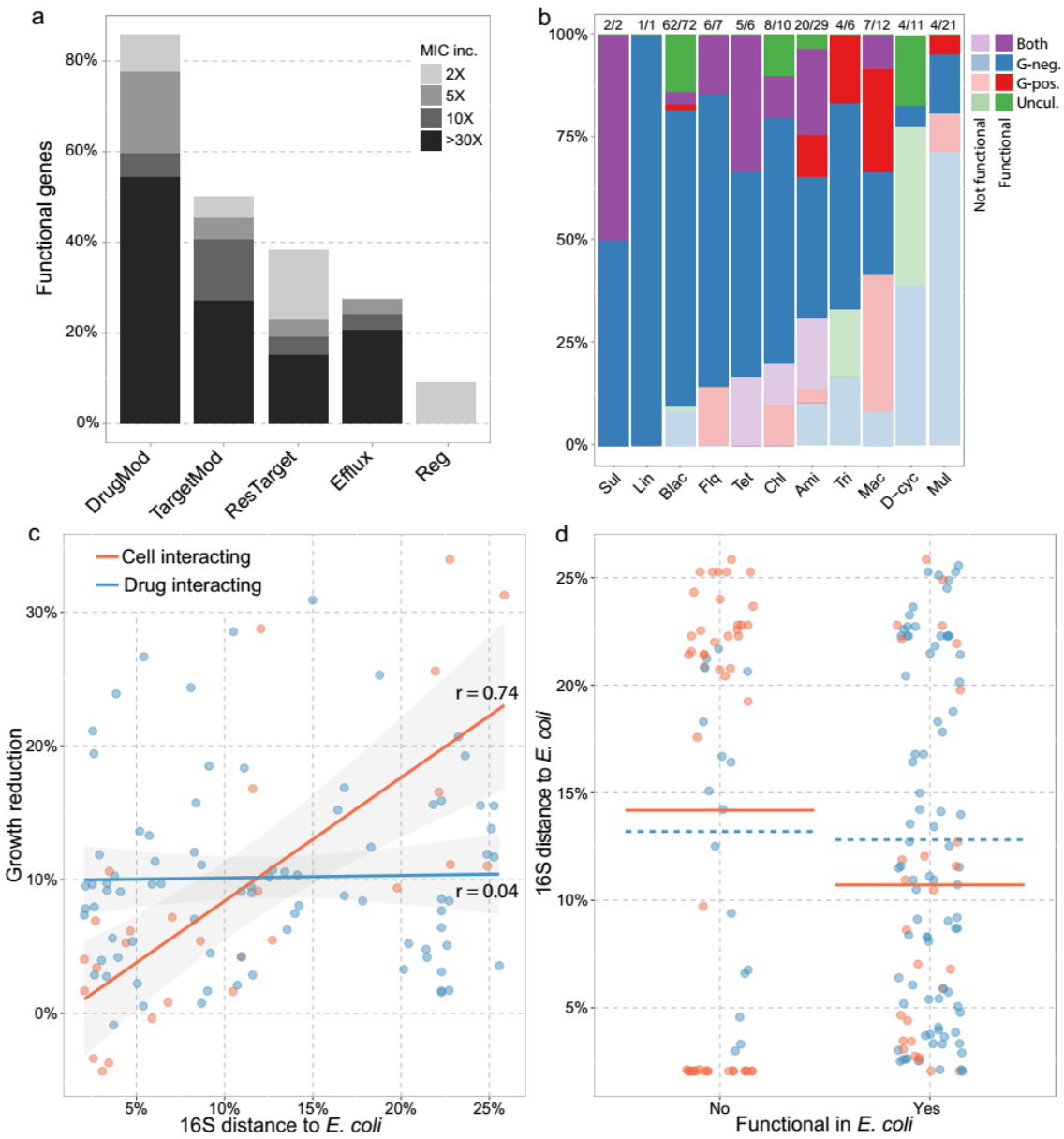


Fig. 5 | Distribution of functional genes on phylogenetic affiliation within drug classes and resistance mechanisms.

(a) Functionality and resistance level of the resistance genes belonging to the different mechanistic categories. DrugMod = Drug inactivating enzymes, TargetMod = Target modifying/binding proteins, ResTarget = Redundant/resistant target, Efflux = Efflux pumps, Reg = Regulators. **(b)** The number of functional genes annotated as conferring resistance towards each drug class are coloured according to their affiliations with Gram-negative organisms (G-neg.); Gram-positive organisms (G-pos.); both Gram-negative and Gram-positive organisms (Both); or none of currently sequenced genomes in RefSeq (Uncul.). The number of functional genes out of the total genes within each class is indicated above the bars. **(c)** The average 16S rRNA distance between *E. coli* and all RefSeq genomes where each gene has been identified correlated with the growth reduction inflicted by each gene. A linear regression model was fitted to the subsets of genes interacting with the drug (n = 74) and cellular components (n = 27), respectively, were phylogenetic data was available. Cell-

interacting mediators include those that conferred resistance via target protection or modification, provision of a resistant target, efflux or regulatory interactions. Drug-interacting genes conferred resistance through modification or breakdown of the antibiotic without interfering directly with host physiology. Shades represent the standard-error of the linear fit. **(d)** The phylogenetic distance from *E. coli* for functional and non-functional genes tested in *E. coli*. The mean is shown for the subsets of drug-modifying (blue dashed line) and cell-interacting (orange solid line) resistance mediators.

Phylogenetic distance affects fitness for cell-interacting resistance mechanisms

If the activity of a gene-product is detrimental for functional genetic integration due to suboptimal physiological interactions, we would expect the phylogenetic relatedness of the donor and recipient species to influence the functional compatibility and fitness cost newly acquired genes. This hypothesis was supported by the fact that the deviation in GC-content from the *E. coli* average (50.8% GC) was a stronger, although still not significant, predictor of the growth reduction resulting from gene expression (r = 0.16, *P* = 0.07; **Supplementary Fig. 8**) compared to the absolute GC-content (**Fig. 4b**). As GC-content varies among phylogenetic groups, this trend could be a proxy for the cellular environment, in which a gene has evolved, rather than a direct effect of sequence composition (**Supplementary Fig. 9**). Indeed, a higher average growth rate was observed for genes affiliated with Gram-negative organisms compared to those that were exclusively associated with in Gram-positive organisms or both (**Supplementary Fig. 10**. Mann-Whitney U-test, *P* = 0.03).

To further investigate the basis of the fitness costs of gene expression, we derived a distance measure based on the average 16S rRNA sequence identity between *E. coli* and the genomes in which the gene has been identified. Using this 16S rRNA based evolutionary distance measure, we found a correlation between the cost of a gene and the relatedness of its genomic context to *E. coli* (r = 0.29, *P* = 0.003). Whereas the relative burden of expressing drug-modifying enzymes was independent of phylogenetic relatedness of the typical hosts of a gene to *E. coli* (r = 0.04, *P* = 0.56; **Fig. 5c**), we found the main drivers of this correlation to be the genes that directly interact with cell factors (r = 0.74, *P* < 0.001; **Fig. 5c**). This correlation was not biased by differences in cell-interacting mechanistic categories (ANOVA, *P* = 0.93; **Supplementary Fig. 11**), targeted drug classes (ANOVA, *P* = 0.57) or phylogenetic groups (ANOVA, *P* = 0.23; **Supplementary Fig. 12**).

These results suggests a trade-off between adaptation towards one host and being broadly functional across phylogeny, which is further supported by the higher cost of genes disseminated across Gram-positives and Gram-negatives (**Supplementary Fig. 10**).

While there was an unequal distribution of functional genes among Gram-classes (χ^2, *P* = 0.002), the

phylogenetic affiliation, measured as the average 16S rRNA-based distance between genomes harbouring the gene, was not significantly correlated with gene functional compatibility when all genes in our dataset were considered (Mann-Whitney U-test, $P = 0.84$). Although a bigger difference was observed when cell-interacting genes were considered in isolation, this difference was still not significant (Mann-Whitney U-test, $P = 0.23$; **Fig. 5d**). However, when excluding native *E. coli* genes (**Supplementary Table 5**), a highly significant dependence of functionality on phylogenetic distance for cell-interacting proteins (Mann-Whitney U-test, $P < 0.001$), but no change for drug-interacting resistance mediators (Mann-Whitney U-test, $P = 0.57$), was observed. For this subset of genes, the effect of phylogenetic differences was dominated by genes conferring resistance through target replacement (Mann-Whitney U-test, $P = 0.031$) and efflux mechanisms (Mann-Whitney U-test, $P = 0.015$). However, while the Gram-class affiliation was not a significant predictor of functionality for efflux, target replacing and regulatory genes (χ^2, $P > 0.05$), genes originating from Gram-negatives were significantly overrepresented in functional target-modifying resistance mediators (χ^2, $P > 0.005$). These observations further support the importance of resistance mechanism and genomic relatedness for gene functional compatibility.

Discussion

To better understand the factors underlying functional compatibility and the fitness cost of genes in new hosts, we assessed a set of diverse antibiotic resistance genes spanning a wide range of sequence compositions and mechanisms. We found that a substantial proportion of genes, not previously observed in *E. coli,* were functional despite huge sequence deviations, in the CAI and GC-content, from the *E. coli* genome average. Interestingly, we found that sequence composition, including the CAI, N-terminal mRNA-folding energy and GC-content, was a poor predictor of functional compatibility and fitness cost (Fig. 3 and 4). Previous studies have shown that the expression level of codon-randomized *gfp* variants is mainly affected by the N-terminal mRNA-folding energy; however, strong negative effects of mRNA folding are expected to be counter-selected in naturally occurring genes, and we found no significant correlation with this parameter and functionality in our diverse gene set[24,38]. Although we did not assess the effect of sequence composition on expression levels directly, our results suggested that gene expression is not trivially linked to phenotypic output as is the case for fluorescent proteins (**Supplementary Fig. 2**)[24]. By virtue of resistance phenotypes, we were able to measure end-point functionality directly to avoid indirect, e.g. protein-level, measurements that might not reflect correct phenotypic integration.

Using antibiotic resistance genes as a model for transferable accessory genes, we experimentally showed that the mechanism of the gene-product is more important for its functional compatibility than the gene sequence composition (**Fig. 5a**). The results obtained for the antibiotic resistance genes included here are likely to apply more broadly to horizontal gene transfer, and our experimental results support the "complexity hypothesis" originally proposed by Jain *et al*., which state that highly interactive gene products are less likely to undergo horizontal transfer [9,40,41]. Although we did not assess protein interactions directly, we observed that genes encoding regulators and efflux pumps, which are dependent on regulatory networks or cell envelope structures, were the least likely to function when expressed in *E. coli* compared to genes encoding enzymes that act directly on the drug (**Fig. 5a**); this result signifies that the extent of physiological decontextualization dictates the likelihood of selection following gene transfer. While the trend towards higher costs of target modifying mediators conferring high-level resistance might be expected, the lack of, and even opposite tendencies, observed for the remaining mechanistic classes, suggests that high resistance and low fitness cost might not be opposing features (**Supplementary Fig. 6)**.

The biological cost of functionally expressing a new gene is an important factor that potentially determines the reversal of antibiotic resistance upon cessation of antibiotic use[22]. Fitness effects are believed to affect the long-term success of gene transfer events as well as the robustness of engineered biological systems, and the origins of these costs have been suggested to stem from suboptimal sequence composition[18,42,43]. Notably, in our setup of moderate expression, our growth measurements did not detect a significant influence of the sequence-level parameters previously suggested to influence the growth rate of *E. coli* expressing heterologous proteins[24,25]. These studies employed libraries of lower or similar sequence-level diversity compared to our dataset, albeit with a narrow mechanistic focus (GFP and ϕ29 DNAP), and observed that the CAI and GC-content affected the growth rate of *E. coli*[24,25]. Yet, in line with our observations, Knöppel and Lind *et al.*[44] were unable to measure a significant fitness effect of parameters such as GC-content and length of random DNA inserts containing a variable number of open reading frames of unknown functions when expressed from a single copy on the chromosome[44]. However, we acknowledge that these effects exist and may be measured by more sensitive fitness assays or at higher expression levels than the ones used here. Compared to the linear influence on the growth rate of *E. coli* expressing *gfp* and ϕ29 DNA polymerase genes within the narrow GC-range (40.4–53.7%) observed by Raghavan *et al.,* our data suggests that the possible effect of GC-content is non-linear, and that the

deviation from the host genome is more important than the absolute GC-content (**Fig. 4b and Supplementary Fig. 8**)[25,44].

While previous studies have found a correlation between fitness and evolutionary distance for homologs of *E. coli* genes involved in translation[27,28], we show that these effects are also evident for accessory genes with non-essential and diverse functionalities, which may contribute to the phylogenetic barriers more generally observed for horizontal gene transfer[1,13,45]. These fitness effects correlated with the relatedness of the donor and recipient species and were independent of sub-mechanistic- and phylogenetic categories (**Supplementary Fig. 11 and 12**). Interestingly, although phylogenetic distance was a much stronger predictor of growth influence than the deviation in GC-content, the correlation between GC–deviation and phylogenetic distance might explain the small deviation in GC-content observed for transferred genes and their recipient genomes (**Supplementary Fig. 9**)[15]. However, by demonstrating the dependence on cellular interactions and showing that GC optimization of the *dfrG* gene did not improve its cost, GC-content may be a confounder rather than the cause of the fitness effects observed here.

We found that sequence-level properties have a limited impact on heterologous gene fitness or function and are therefore unlikely to be the dominant causal factors confining genes within specific phylogenies. Instead, biochemical mechanisms have strong impacts on heterologous gene fitness that are proportional to the phylogenetic distance to its customary host. The notion that the functional compatibility of cell-interacting proteins is dependent on phylogenetic relationships supports a shift in focus away from sequence composition and metabolic constraints as limiting factors in horizontal gene transfer. Interestingly, the positive effects observed for certain members of the *dfr* and *qnr* families, implicated in DNA gyrase protection and folate metabolism, might even enhance their persistence in the absence of antibiotic selection, when exchanged between closely related species.

Importantly, our data suggested that antibiotic resistance genes interacting directly with the drug are more likely to function and be maintained when transferred to a new host. Historically, drug-interacting resistance mechanisms, e.g., aminoglycosides, chloramphenicol and β-lactam resistance, have emerged faster following the introduction of the drug class in question, compared to e.g. macrolide and vancomycin resistance, which are mediated largely through modification of the cellular targets [46]. For aminoglycosides, drug-modifying genes were detected regularly since the late 1940s, whereas ribosome-modifying methyltransferases were first detected in the early 2000s[47].

The detailed phenotypic information on individual resistance genes obtained here will be an important resource for ranking the risk of resistance genes and predicting their evolution against existing and future drugs[32,48]. However, knowledge on antibiotic use, co-selection, regulatory or compensatory interactions in a range of hosts is still needed for accurate predictions[49]. Coupling our findings to factors such as drug usage and ecology of pathogens may allow the construction of predictive models to help guide rational drug usage and development of novel drug classes that are less susceptible to resistance development[32]. Finally, the concepts derived in this study may also guide metabolic engineering or synthetic biology efforts when heterologous proteins are needed to engineer any kind of robust biological system[5,6].

Methods

Database construction and gene synthesis

All resistance gene entries from the ARDB, CARD and Lahey Clinic β-lactamase databases (Accessed Dec, 2014) and functional selection studies were downloaded [34,50–53]. A total of 4253 unique sequences were obtained after clustering at 99% nucleotide identity using CD-HIT software [54], and these sequences were further clustered at 80% identity resulting in 839 gene clusters. These clusters were first sorted according to the mean number of BLAST hits obtained in NCBI to include the most abundant genes and then sorted on cluster size to limit the inclusion of housekeeping genes that are often conserved and occur in small clusters. The 200 largest clusters were selected based on the modal sequence length of each cluster. Trimming was performed using GeneMarkS to only include the longest open reading frame of each sequence [55]. Due to synthesis and cloning limitations, genes longer than 1970 bp were excluded, and *XbaI* and *AscI* sites were synonymously removed by manual sequence inspection. Two primer binding sites for amplification of the entire gene and a unique barcode were added to each of the 200 sequences (**Supplementary Fig. 2)**. Finally, the sequences were ordered as gBlocks® through Integrated DNA Technologies (IDT, Coralville, Iowa, USA). All genes and the data obtained for each functional gene is available in Supplementary Table 1.

Gene dissemination and genomic context analysis

To assess gene association with sequenced genomes, BLAST comparisons were performed at a 95% identity alignment cut-off, and validated through EMBOSS Matcher pairwise, against NCBIs RefSeq database (67,704 entries; last performed October 2016). The rarefaction analysis of nucleotide databases (**Supplementary Fig. 1b**) was performed at a 90% identity and 90% coverage cut-off against the NCBI nucleotide and genomes databases. To quantify the dissemination distance of those antibiotic resistance genes with respect to *E. coli*, we chose one representative 16S rRNA from each genus carrying each ARG in RefSeq. The host distance was then calculated using EMBOSS Matcher pairwise alignment between *E. coli* 16S rRNA and representative 16S rRNA for the ARG-carrying genomes and reported as the percent mismatches in the total alignment of the *E. coli* reference 16S rRNA gene.

Cloning and expression of gene library

The pZAT vector backbone (**Supplementary Fig. 2**) was derived from the medium copy p15A based pZA21 vector [56] by exchanging the resistance marker with the *Sh ble* Zeocin resistance gene (Thermo Fisher Scientific, Waltham, MA, USA) and inserting the low/medium strength BBa_J23110 promoter of the iGEM parts registry (http://parts.igem.org). This backbone was PCR amplified using primers 5'-AATTTGGCGCGCCCATCAAATAAAACGAAAGGC-3' and 5'-AATTTTCTAGATCTCCTCTTTAATGCTCGC-3' to create *XbaI* and *AscI* digestible overhangs. The PCR-product (25 µL) was subsequently treated with 0.5 µL *DpnI* restriction enzyme (Thermo Fisher) overnight (O/N) at 37°C to remove residual template and subsequently PCR purified (NucleoSpin® Gel and PCR Clean-up kit from Macherey-Nagel, standard protocol). In 200 individual reactions, the vector backbone was combined with each synthetic gene block in a reaction containing *XbaI* and *AscI* restriction enzymes (Thermo Fisher). The reactions were incubated at 37°C for 1 h, followed by heat-inactivation of the enzymes at 80°C for 20 min. Subsequently, the 20-µL reactions were ligated by adding 0.5 µL T4 ligase and 3 µL ATP (5 mM)

and incubated at room temperature O/N. The 5-µL ligation product was used to transform 50 µL *E. coli MG1655* chemically competent cells followed by recovery at 37°C for 3 h. Transformed cells were selected on Zeocin-containing plates (40 µg/ml), and correct insertion of gene blocks was verified by colony PCR and subsequent sequencing using primers 5'- TATGCCGATATACTATGC-3' and 5'- AAGCACTTCACTGACACC-3'.

Antibiotic susceptibility testing

The minimal inhibitory concentration of all 20 included antibiotics was measured for all library clones as well as *E. coli* MG1655 carrying the empty vector backbone (pZAT). One colony of each clone was inoculated in 180 µL LB/well (with 40 µg/ml Zeocin added for backbone selection) for overnight (O/N) pre-culturing at 37°C. Ninety-six-well plates were prepared with 100 µL MHB2 medium (Sigma) per well, containing the respective antibiotics, at concentrations 2x, 5x, 10x and 30x the MIC of *E. coli* MG 1655/pZAT (**Supplementary Table 2**). Three replicate plates were inoculated with $5{\cdot}10^5$ cells and incubated for 18 h at 37°C with shaking at 250 rpm (Titramax 1000, Heidolph). Endpoint optical density (OD) was measured at 600 nms (Synergy H1, BioTek®), and the MIC was defined as the highest concentration with less or similar absorbance as the *E. coli* MG1655/pZAT (negative control) subjected to the same antibiotic concentration.

Growth rate measurements

The growth rate of each functional clone was measured as the maximum increase in optical density (OD) over time during exponential growth. Individual colonies were picked and placed onto a pre-inoculation plate and grown for 2-3 h with shaking at 37°C before inoculation of the final measurement plate. Breathe-Easy (Sigma-Aldrich) film was applied to minimize evaporation during measurements. OD measurements were conducted in 96-well plates containing 150 µl LB medium/well by the *ELx808* plate reader (BioTek, USA). OD at 600 nms was measured over 5-minute intervals for a maximum of 16 h, and the plates were incubated at the medium shaking setting at 37°C between measurements.

Sequence parameters and statistical analysis

All statistical analyses were performed using R (version 3.1.1). Sequence composition data were obtained using the native functions of Biopython (version 1.70)[57]. The mRNA-folding energy was calculated for 35 nt up- and downstream the start codon of each gene using the *RNAfold* web server (http://rna.tbi.univie.ac.at/). The non-parametric Mann-Whitney U test was used to compare sample means, and a χ^2 test was used when frequencies were assessed. Spearman's rank correlations were performed to assess the strength of associations between two continuous variables. When the influence of multiple variables was assessed simultaneously, a generalized linear model was fitted to binary response variables (*glm* function in R), and a multiple linear regression was fitted (*lm* and *anova* functions in R) to continuous response data.

Data availability

The authors declare that all the relevant data are provided in this published article and its Supplementary Information files, or are available from the corresponding authors on request.

Acknowledgement

We thank Peter Rugbjerg and Leonie Jahn for critical reading and suggestions to the manuscript. M.O.A.S further acknowledges financial support from the Novo Nordisk Foundation and the Lundbeck Foundation.

Author contributions

AP, MOAS and CM conceived the study. TSS and AP performed the experimental work. AP, TSS and CM analysed the data. AP, CM and MMHE compiled the antibiotic resistance gene database, and MMHE provided genomic association data. AP and MOAS wrote the manuscript with input from CM, TSS and MMHE.

References

1. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* **16,** 472–482 (2015).

2. Ochman, H., Lawrence, J. G. & Groisman, E. a. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405,** 299–304 (2000).

3. Von Wintersdorff, C. J. H. *et al.* Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* **7,** 1–10 (2016).

4. Popa, O. & Dagan, T. Trends and barriers to lateral gene transfer in prokaryotes. *Curr. Opin. Microbiol.* **199,** 615–623 (2011).

5. Marguet, P., Balagadde, F., Tan, C. & You, L. Biology by design: reduction and synthesis of cellular components and behaviour. *J. R. Soc. Interface* **4,** 607–23 (2007).

6. Cardinale, S. & Arkin, A. P. Contextualizing context for synthetic biology - identifying causes of failure of synthetic biological systems. *Biotechnol. J.* **7,** 856–866 (2012).

7. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480,** 241–244 (2011).

8. Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **544,** 124–124 (2017).

9. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 3801–3806 (1999).

10. Aris-Brosou, S. Determinants of adaptive evolution at the molecular level: The extended complexity hypothesis. *Mol. Biol. Evol.* **22,** 200–209 (2005).

11. Puigbo, P., Wolf, Y. I. & Koonin, E. V. The Tree and Net Components of Prokaryote Evolution. *Genome Biol. Evol.* **2,** 745–756 (2010).

12. Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet.* **36,** 760–766 (2004).

13. Forsberg, K. J. *et al.* Bacterial phylogeny structures soil resistomes across habitats. *Nature* **509,** 612–6 (2014).

14. Andam, C. P. & Gogarten, J. P. Biased gene transfer in microbial evolution. *Nat. Rev.*

Microbiol. **9,** 543–555 (2011).

15. Popa, O., Hazkani-Covo, E., Landan, G., Martin, W. & Dagan, T. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res.* **21,** 599–609 (2011).

16. Medrano-Soto, A., Moreno-Hagelsieb, G., Vinuesa, P., Christen, J. A. & Collado-Vides, J. Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. *Mol. Biol. Evol.* **21,** 1884–1894 (2004).

17. Popa, O., Landan, G. & Dagan, T. Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. *ISME J.* 1–12 (2016). doi:10.1038/ismej.2016.116

18. Park, C. & Zhang, J. High expression hampers horizontal gene transfer. *Genome Biol. Evol.* **4,** 523–532 (2012).

19. Tuller, T. *et al.* Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids Res.* **39,** 4743–4755 (2011).

20. De Gelder, L., Ponciano, J. M., Joyce, P. & Top, E. M. Stability of a promiscuous plasmid in different hosts: no guarantee for a long-term relationship. *Microbiology* **153,** 452–63 (2007).

21. Porse, A., Schønning, K., Munck, C. & Sommer, M. O. A. Survival and Evolution of a Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts. *Mol. Biol. Evol.* **33,** 2860–2873 (2016).

22. Johnsen, P. J. *et al.* Factors affecting the reversal of antimicrobial-drug resistance. *Lancet Infect. Dis.* **9,** 357–64 (2009).

23. Vogwill, T. & MacLean, R. C. The genetic basis of the fitness costs of antimicrobial resistance: a meta-analysis approach. *Evol. Appl.* **8,** 284–295 (2015).

24. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science (80-.).* **324,** 255–258 (2009).

25. Raghavan, R., Kelkar, Y. D. & Ochman, H. A selective force favoring increased G+C content in bacterial genes. *Proc. Natl. Acad. Sci.* **109,** 14504–14507 (2012).

26. Goodman, D. B., Church, G. M. & Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science (80-.).* **342,** 475–479 (2013).

27. Lind, P. A., Tobin, C., Berg, O. G., Kurland, C. G. & Andersson, D. I. Compensatory gene amplification restores fitness after inter-species gene replacements. *Mol. Microbiol.* **75,** 1078–1089 (2010).

28. Kacar, B., Garmendia, E., Tuncbag, N., Andersson, D. I. & Hughes, D. Functional constraints on replacing an essential gene with its ancient and modern homologs. *MBio* **8,** 1–30 (2017).

29. Amorós-Moya, D., Bedhomme, S., Hermann, M. & Bravo, I. G. Evolution in regulatory regions rapidly compensates the cost of nonoptimal codon usage. *Mol. Biol. Evol.* **27,** 2141–2151 (2010).

30. Hughes, D. & Andersson, D. I. Environmental and genetic modulation of the phenotypic expression of antibiotic resistance. *FEMS Microbiol. Rev.* 1–18 (2017). doi:10.1093/femsre/fux004

31. WHO. *Antimicrobial resistance. Bulletin of the World Health Organization* **61,** (2014).

32. Sommer, M. O. A., Munck, C., Toft-Kehler, R. V. & Andersson, D. I. Prediction of antibiotic resistance: time for a new preclinical paradigm? *Nat. Rev. Microbiol.* **15,** 689–696 (2017).

33. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O. & Piddock, L. J. V. Molecular mechanisms of antibiotic resistance. *Nat. Rev. Microbiol.* **13,** 42–51 (2015).

34. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9,** 1–10 (2014).

35. Bush, K., Jacoby, G. A. & Medeiros, A. A. A functional classification scheme for beta-lactamases and its correlation with molecular structure. *Antimicrob. Agents Chemother.* **39,** 1211–33 (1995).

36. Sharpl, P. M. & Li, W. The codon Adaptation Index - A measure of directional synonymous codon usage bias, and its possible applications. *Nucleic Acids Res.* **15,** 1281–1295 (1987).

37. Welch, M., Villalobos, A., Gustafsson, C. & Minshull, J. You're one in a googol: optimizing genes for protein expression. *J. R. Soc. Interface* **6,** S467–S476 (2009).

38. Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci.* **107,** 3645–3650 (2010).

39. Munck, C., Ellabaan, M., Klausen, M. S. & Sommer, M. O. A. The Resistome Of Important Human Pathogens. *bioRxiv* (2017). doi:http://dx.doi.org/10.1101/140194

40. Lercher, M. J. & Pál, C. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol. Biol. Evol.* **25,** 559–567 (2008).

41. Cohen, O., Gophna, U. & Pupko, T. The complexity hypothesis revisited: Connectivity Rather Than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* **28,** 1481–1489 (2011).

42. Bull, J. J. & Barrick, J. E. Arresting Evolution. *Trends Genet.* **xx,** 1–11 (2017).

43. Baltrus, D. a. Exploring the costs of horizontal gene transfer. *Trends Ecol. Evol.* **28,** 489–95 (2013).

44. Knöppel, A., Lind, P. a, Lustig, U., Näsvall, J. & Andersson, D. I. Minor fitness costs in an experimental model of horizontal gene transfer in bacteria. *Mol. Biol. Evol.* **31,** 1220–7 (2014).

45. Hu, Y. *et al.* The Bacterial Mobile Resistome Transfer Network Connecting the Animal and Human Microbiomes. **82,** 6672–6681 (2016).

46. Palumbi, S. R. Humans as the World's Greatest Evolutionary Force. *Science (80-.).* **293,** 1786–1790 (2010).

47. Waglechner, N. & Wright, G. D. Antibiotic resistance: it's bad, but why isn't it worse? *BMC Biol.* **15,** 84 (2017).

48. Martínez, J. L., Coque, T. M. & Baquero, F. What is a resistance gene? Ranking risk in resistomes. *Nat. Publ. Gr.* **13,** 116–123 (2014).

49. Hughes, D. & Andersson, D. I. Evolutionary Trajectories to Antibiotic Resistance. 579–596

(2017). doi:10.1146/annurev-micro-090816

50. Sommer, M. O. a, Dantas, G. & Church, G. M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* **325,** 1128–1131 (2009).

51. Bush, K. & Jacoby, G. A. Updated functional classification of β-lactamases. *Antimicrob. Agents Chemother.* **54,** 969–976 (2010).

52. McArthur, A. G. *et al.* The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* **57,** 3348–3357 (2013).

53. Liu, B. & Pop, M. ARDB - Antibiotic resistance genes database. *Nucleic Acids Res.* **37,** 443–447 (2009).

54. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).

55. Besemer, J. 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 (2001).

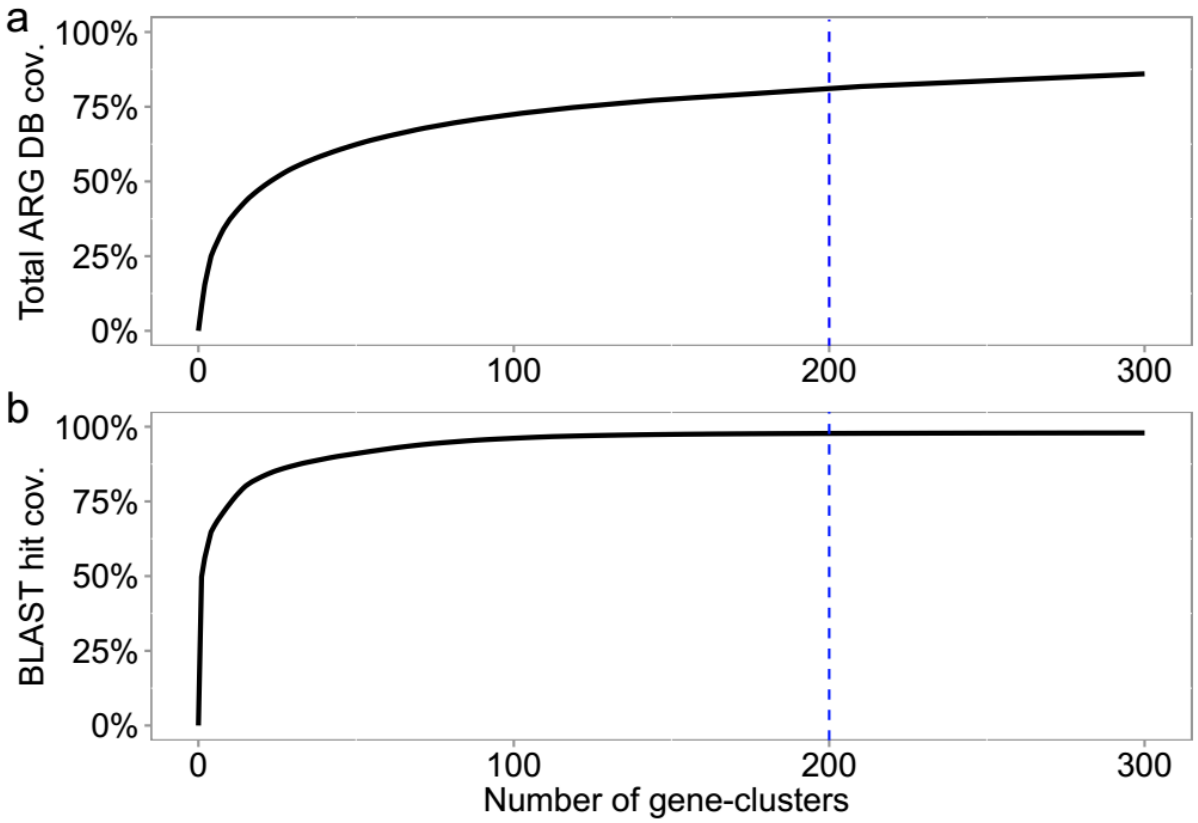
56. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res.* **25,** 1203–1210 (1997).

57. Cock, P. J. A. *et al.* Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).

Supplementary data

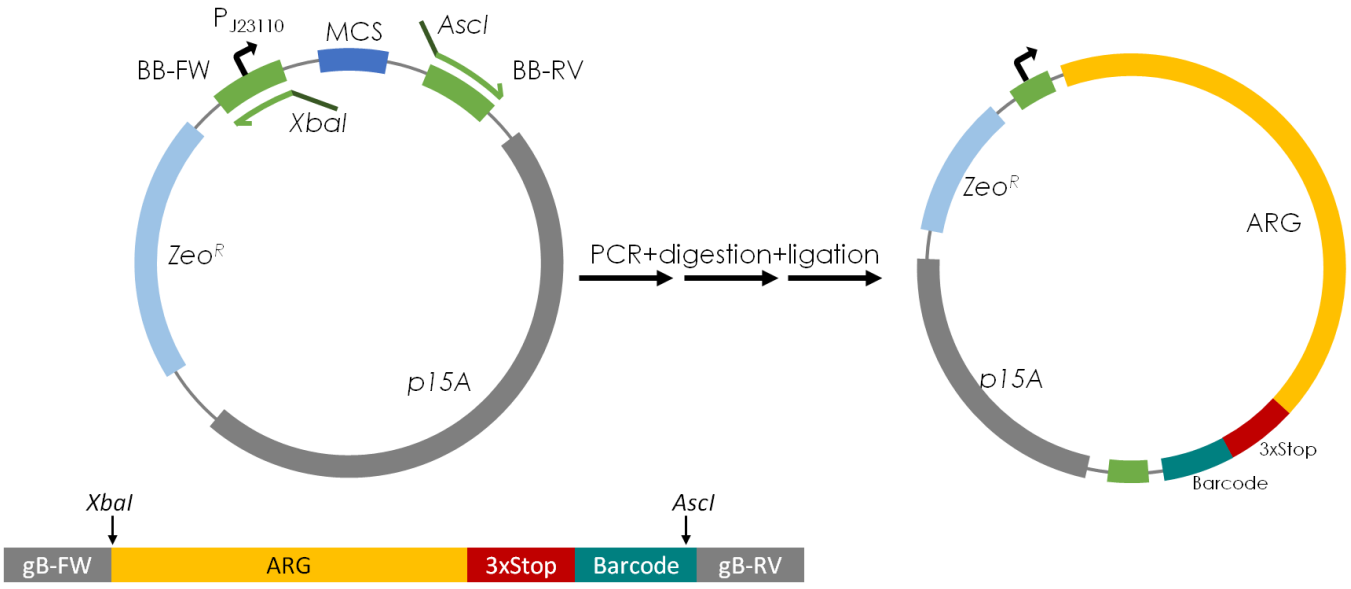
Supplementary Figure 1:

(**a**) Rarefaction plots depicting the cumulative coverage of the included resistance gene databases by the included ARG clusters (**b**) Cumulative coverage of NCBI's NT and Genomes sequence databases as a function of added clusters. The 200 included clusters are indicated by the blue dotted line.



Supplementary Figure 2:

Cloning procedure. The vector backbone was amplified with restriction sites and digested by XbaI and AscI to allow ligation of the final library construct. The ordered genes were designed with primer binding sides (gB-FW and gB-RV) to allow amplification. 3 stop codons were inserted downstream each cloned gene to ensure proper translational termination and a 7nt barcode was incorporated for easy sequence validation.

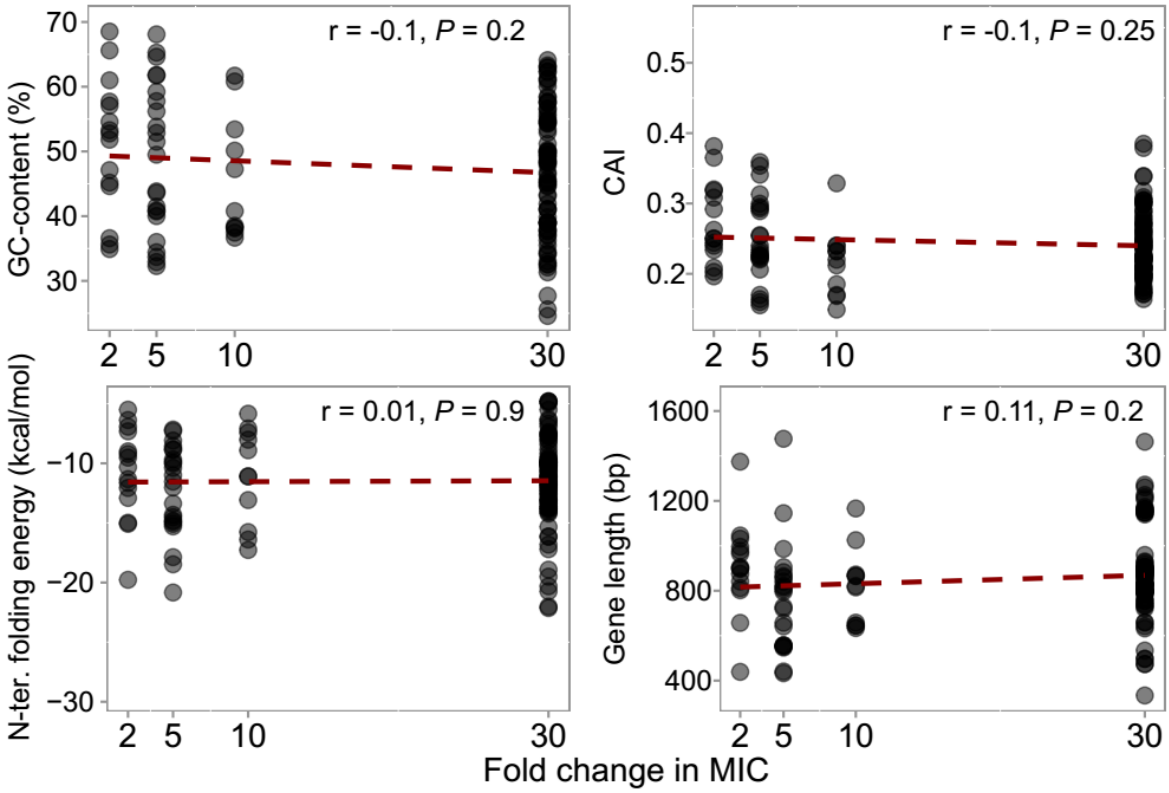


Supplementary Table 2 – MG1655/pZAT MICs:

Antibiotic	MIC (μg/ml)
Amikacin	8
Amoxicillin	8
Aztreonam	0.1
Cefuroxime	4
Chloramphenicol	4
Ciprofloxacin	0.0075
Clindamycin	26.67
Colistin	0.33
D-cycloserine	80
Erythromycin	16
Fosfomycin	0.42
Gentamicin	2
Mecillinam	0.16
Meropenem	0.12
Nitrofurantoin	16
Sulfamethoxazole	128
Streptomycin	4
Tetracycline	1
Trimethoprim	0.25
Vancomycin	128

Supplementary Figure 3

Correlations of sequence parameters and resistance level.



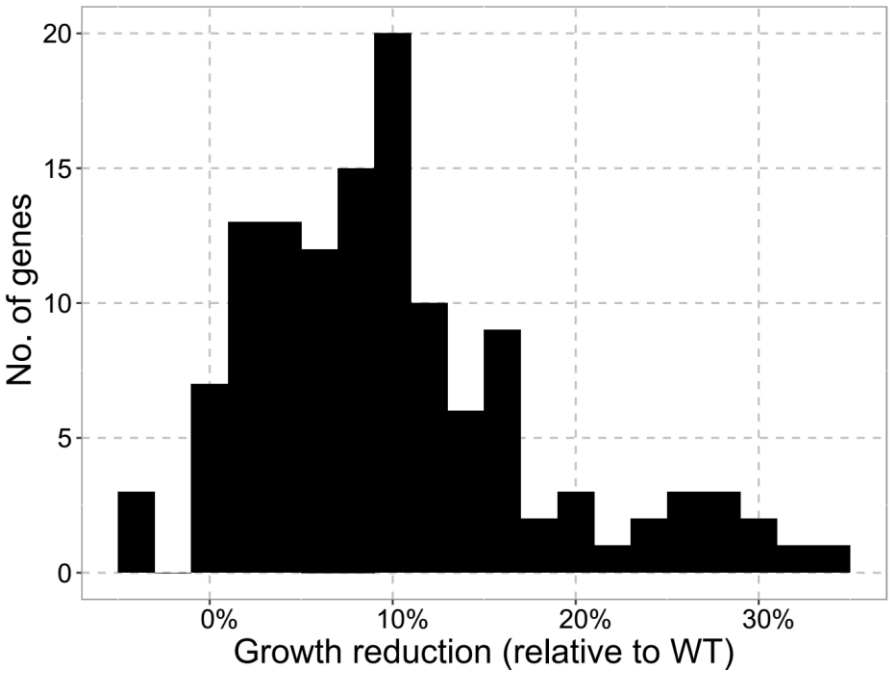
Supplementary Table 3 - GLM:

Sequential fitting of a logistic regression model with absolute functionality in *E. coli* as dependent variable. A model was fitted using the *glm* function in R. Each expanded model was assessed as a chi-squared statistics of differences in deviance/error compared to the previous model (*P*-value added factor) and the null-model (*P*-value all). dGmRNA: N-terminal mRNA folding energy; GramClass: whether the a gene matches Gram-positives only, Gram-negatives only or Both, when blasted against RefSeq.

Model components	**Deviance**	**DF**	***P*-value added factor**	***P*-value all**
null model	263.58	199	NA	NA
GC	260.87	198	0.09	0.09
GC+CAI	255.8	197	0.025	0.02
GC+CAI+dGmRNA	254.96	196	0.36	0.034
GC+CAI+dGmRNA+Mechanism	190.99	192	4.24E-13	4.42E-13
GC+CAI+dGmRNA+Mechanism +DrugClass	150.98	180	7.16E-05	2.66E-15
GC+CAI+dGmRNA+Mechanism +DrugClass+GramClass	146.56	177	0.21	6.11E-15

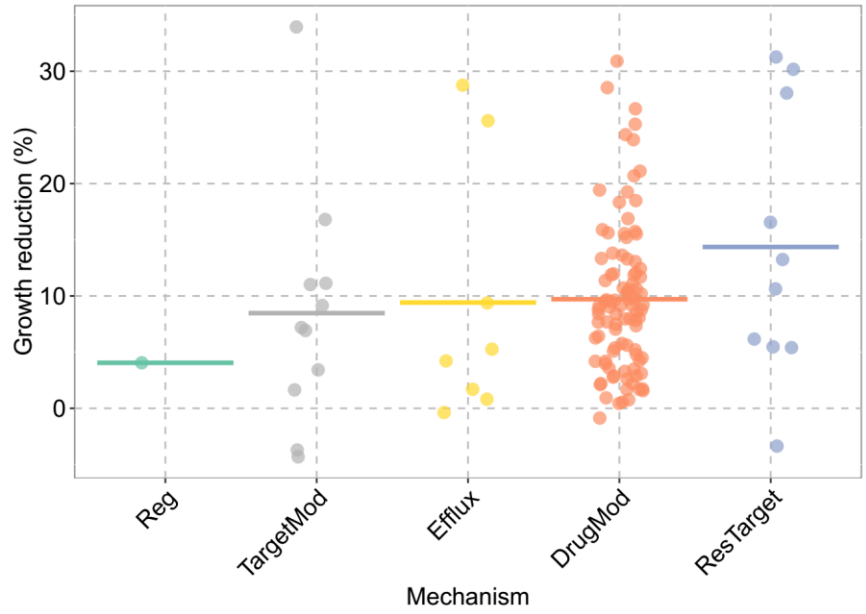
Supplementary Figure 4

Growth rate distribution of all functional genes tested.



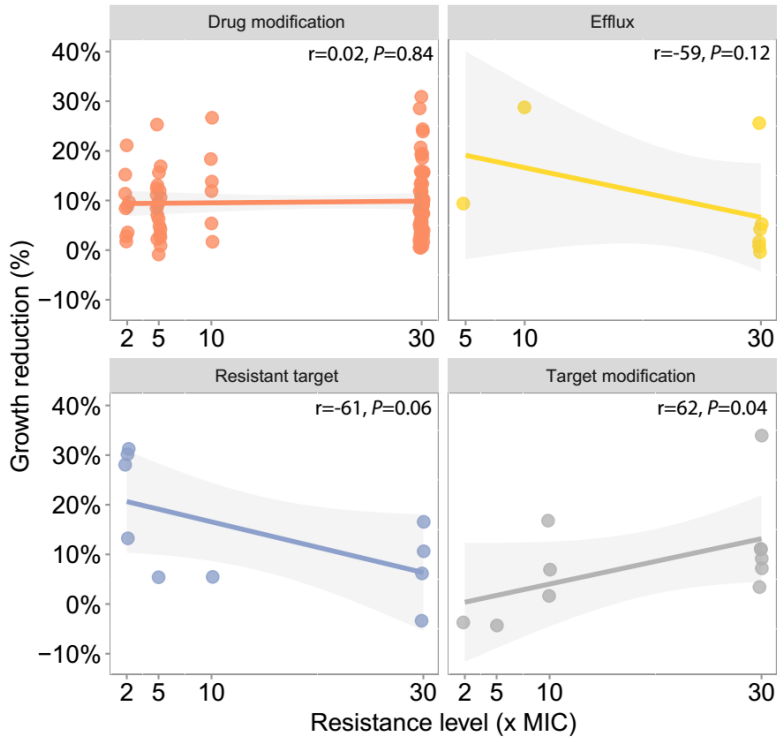
Supplementary Figure 5

The impact on the growth rate imposed by ARGs on *E. coli* stratified on resistance mechanism.



Supplementary Figure 6

Correlation between MIC and growth rates for the different mechanistic classes. Regulators were omitted due to the small sample size (n=1; see above).



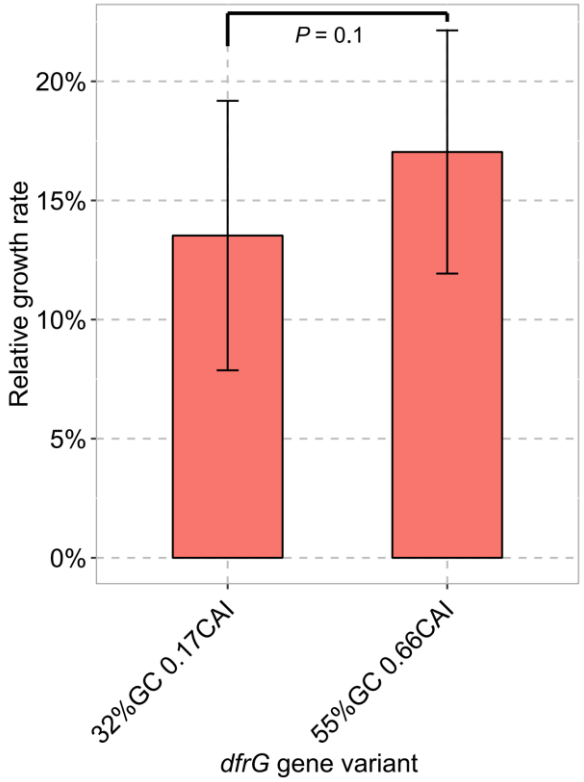
Supplementary Figure 7

A codon optimized *dfrG* gene variant was synthesized to test the influence of gene composition on growth rate in *E. coli*.

Optimized gene sequence:

”ATGAAAGTTTCTTTGATTGCTGCGATGGATAAAAACCGCGTGATCGGCAAAGAGAACGACATCCCGTGGCGCATCCCGAAAGACTGGGAGTACGTGAA
AAACACCACCAAAGGCCACCCGATCATCCTGGGCCGCAAAAACCTGGAGAGCATCGGCCGCGCGCTGCCGGACCGCCGCAACATCATCCTGACCCGCGA
CAAAGGCTTCACCTTCAACGGCTGCGAGATCGTGCACAGCATCGAGGACGTGTTCGAGCTGTGCAAGAACGAGGAGGAGATCTTCATCTTCGGCGGCGA
GCAGATCTACAACCTGTTCTTCCCGTACGTGGAGAAGATGTACATCACCAAAATCCACCACGAGTTCGAGGGCGACACCTTCTTCCCGGAGGTGAACTACG
AGGAGTGGAACGAGGTGTTCGCGCAGAAAGGCATCAAGAACGACAAAAACCCGTACAACTACTACTTCCACGTGTACGAGCGCAAGAACCTGCTGAGCT
GA”

The first 10 codons were kept unchanged to avoid changes in mRNA folding from affecting expression. The growth rate was measured for the optimized (55% GC and 0.68 CAI) and the wild type (32% GC and 0.17 CAI) variant. No significant difference in growth was observed (Mann-Whitney U-test, P = 0.1). Error-bars show standard deviation of 16 biological replicates.



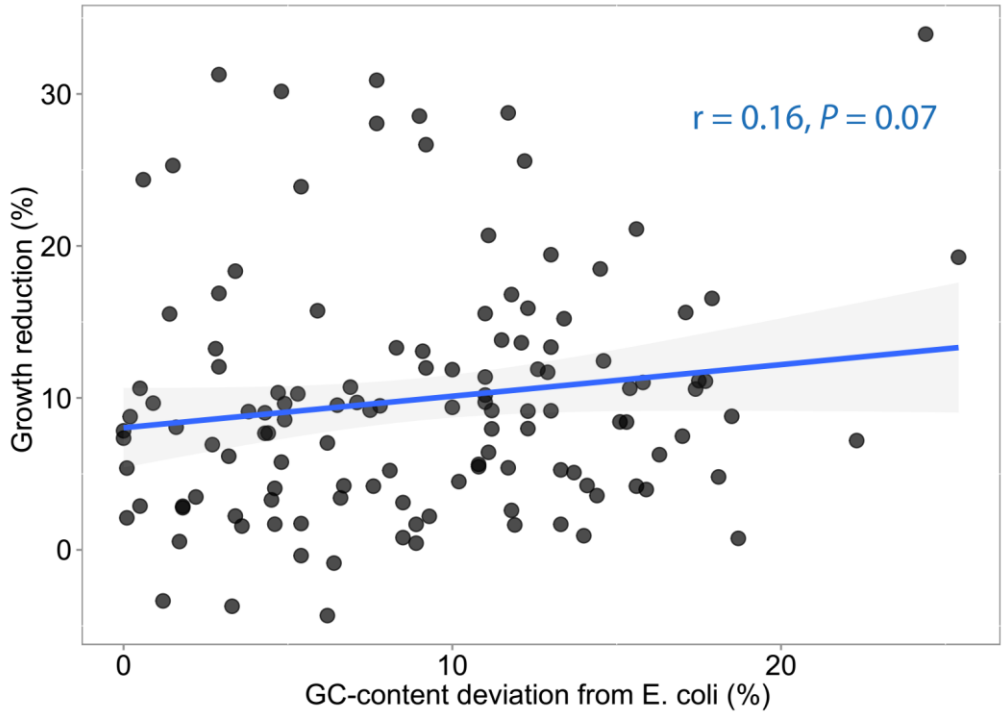
Supplementary Table 4 – MLR

Sequential fitting of a multiple linear regression model to the change in growth rate imposed by functional resistance genes in *E. coli*. A model was fitted using the *lm* function in R. Each expanded model was assessed by the F-statistics of explained variance compared to the model without the added factor (*P*-value added factor) and the null-model (*P*-value all). dGmRNA: N-terminal mRNA folding energy; GramClass: whether the a gene matches Gram-positives only, Gram-negatives only or Both, when blasted against RefSeq.

Model components	**Adj. R^2**	**DF**	***P*-value added factor**	***P*-value all**
GC	-0.002	124	0.4	0.4
GC+CAI	-0.002	123	0.3	0.4
GC+CAI+dGmRNA	-0.007	122	0.5	0.5
GC+CAI+dGmRNA+GeneLength	-0.01	121	0.58	0.6
GC+CAI+dGmRNA+GeneLength +DrugClass	0.16	111	3.00E-04	1.00E-03
GC+CAI+dGmRNA+GeneLength+DrugClass +GramClass	0.2	108	0.038	4.00E-04
GC+CAI+dGmRNA+Mechanism+DrugClass +GramClass+Mechanism	0.21	105	0.34	6.00E-04
GC+CAI+dGmRNA+Mechanism+DrugClass +GramClass+Mechanism+MaxMIC	0.2	104	0.59	0.001

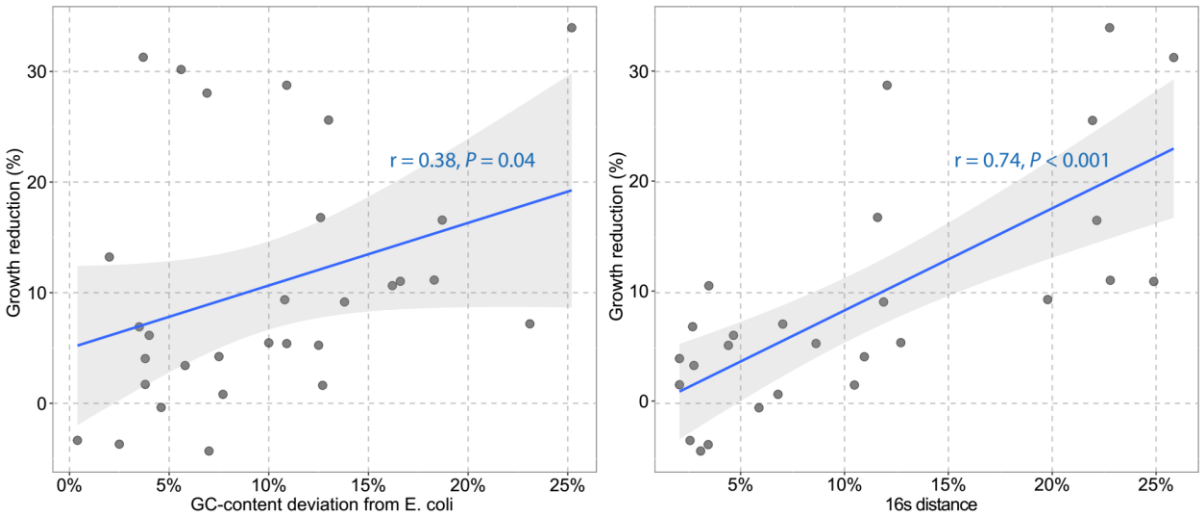
Supplementary Figure 8

Correlation between relative GC-content and growth rate.



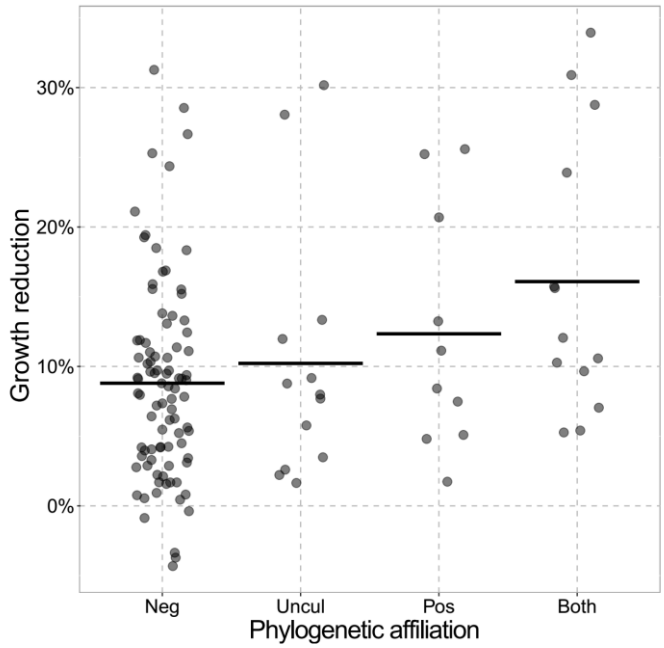
Supplementary Figure 9

Correlation of growth rate with GC-content and 16s distance for cell-interacting genes.



Supplementary Figure 10

The impact on the growth rate of *E. coli* stratified on phylogenetic origin at the Gram-class affiliation. Gram-negative organisms (Neg); Gram-positive organisms (Pos); both Gram-negative and Gram-positive organisms (Both); or none of currently sequenced genomes in RefSeq (Uncul).



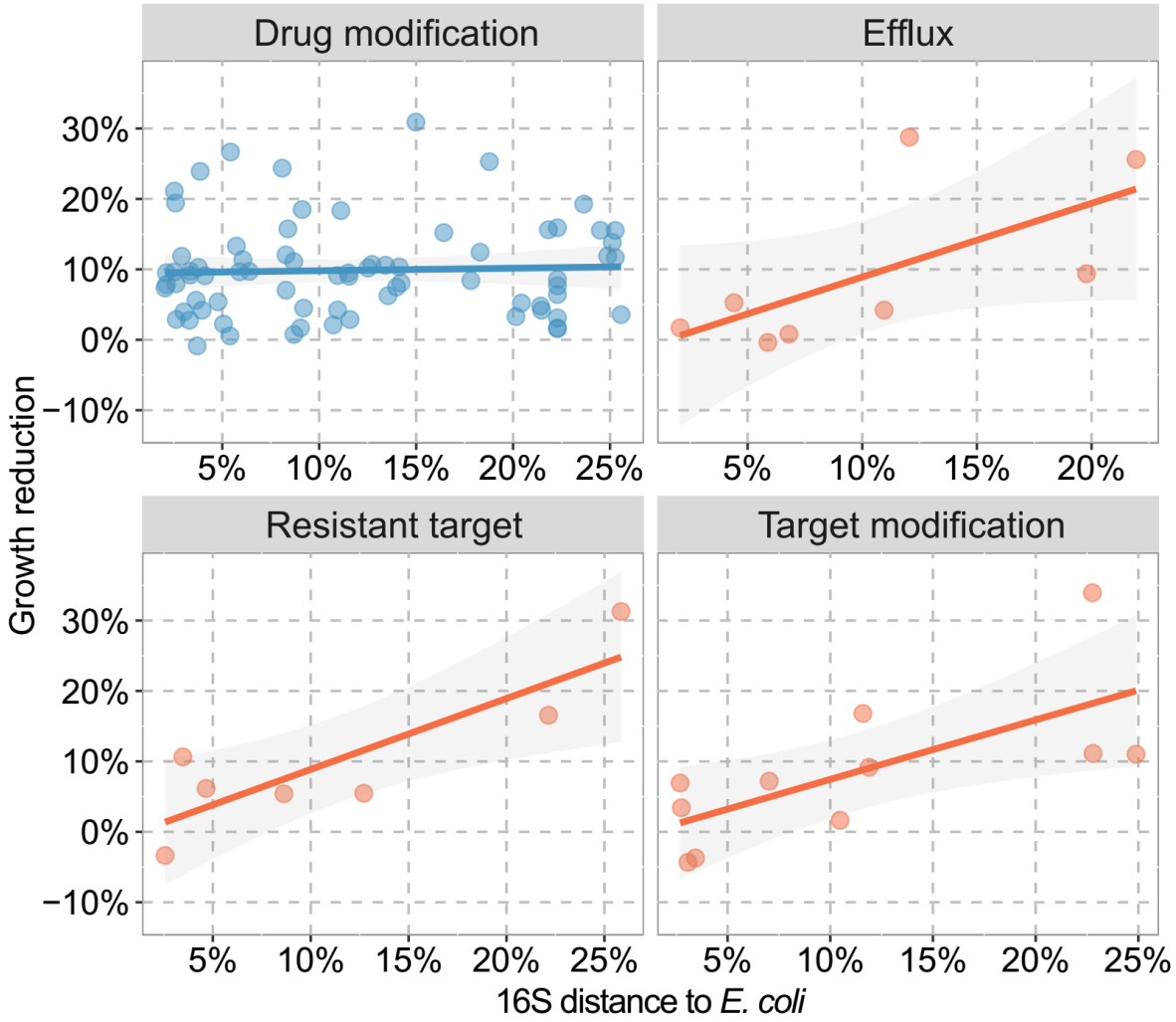
Supplementary Table 5 – Native *E. coli* genes

Due to their high abundance in resistance gene databases a number of native *E. coli* genes vaguely associated with antibiotic resistance were included in the synthetic gene selection. For the full list of gene information, see Supplementary data Table 1.

Cluster	GeneName	Mechanism	Drug class	GC (%)	CAI
77	*pmrC*	Reg	Unk	49.5	0.315
89	*emrB*	Efflux	Mul	56.2	0.279
104	*mdtP*	Efflux	Mul	55.8	0.279
108	*phoQ*	Reg	Mul	51.3	0.25
121	*mdtD*	Efflux	Mul	55.5	0.271
124	*baeS*	Reg	Unk	53.7	0.319
136	*hmrM*	Efflux	Mul	53	0.331
157	*mdtA*	Efflux	Unk	55.3	0.353
164	*mdtM*	Efflux	Mul	53.8	0.266
172	*mdtG*	Efflux	Mul	52.5	0.222
184	*MdtH*	Efflux	Mul	55.3	0.294
213	*mdtL*	Efflux	Mul	53.1	0.252
216	*EmrA*	Efflux	Mul	53.4	0.386
257	*BasS/PmrB*	Reg	Pol	54.2	0.256
292	*mdtN*	Efflux	Mul	55.9	0.349
326	*arnC*	TargetMod	Pol	49.8	0.3
791	*HNS*	Reg	Unk	46.4	0.56
228	*AcrE*	Efflux	Mul	52.1	0.32
221	*mdtA*	Efflux	Unk	55.3	0.353
135	*cpxA*	Reg	Ami	54.6	0.365

Supplementary Figure 11

The effect of the average 16S distance on the growth reduction imposed by genes of different mechanistic categories. There was no significant effect of the different cell-interacting categories (orange) in a multiple linear regression (ANOVA, $P = 0.93$). Regulators were excluded from the analysis because this mechanistic category only contained one functional gene (*cpxA*).



Supplementary Figure 12

The effect of the average 16S distance on the growth reduction imposed by genes detected in different phylogenetic classes. Only taxonomical classes with more than two genes in each category were included. There was no significant effect of the phylogenetic class affiliation on the correlation of growth rate and average 16S distance (across affiliated genomes) in a multiple linear regression model including these factors (ANOVA, $P = 0.23$).