



**HAL**  
open science

# Plasticity of the genome of the course of a long term evolution experiment in *Escherichia coli*

Colin Raeside

► **To cite this version:**

Colin Raeside. Plasticity of the genome of the course of a long term evolution experiment in *Escherichia coli*. Virology. Université de Grenoble, 2014. English. NNT : 2014GRENV070 . tel-01555835

**HAL Id: tel-01555835**

**<https://theses.hal.science/tel-01555835>**

Submitted on 4 Jul 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## THÈSE

Pour obtenir le grade de

### DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Virologie, Microbiologie et Immunologie**

Arrêté ministériel : 7 août 2006

Présentée par

**Colin RAESIDE**

Thèse dirigée par **Joël GAFFE** et **Dominique SCHNEIDER**

préparée au sein du **Laboratoire Adaptation et Pathogénie des Microorganismes** (CNRS UMR5163)

dans l'École Doctorale Chimie et Sciences du Vivant

## PLASTICITE DU GENOME AU COURS D'UNE EXPERIENCE D'EVOLUTION AU LONG TERME CHEZ *ESCHERICHIA COLI*

Thèse soutenue publiquement le **17/10/2014**, devant le jury composé de :

**Mr Guillaume BESLON**

PU, Lyon, Rapporteur.

**Mme Amel LATIFI**

PU, Marseille, Rapporteur.

**Mme Patricia RENESTO**

DR, Grenoble, Examineur.

**Mr Olivier TENAILLON**

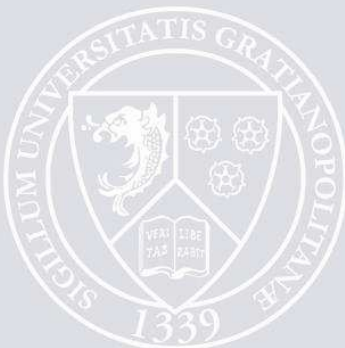
DR, Paris, Examineur.

**Mr Joël GAFFE**

MCU, Grenoble, Directeur de thèse.

**Mr Dominique SCHNEIDER**

PU, Grenoble, Directeur de thèse.



# Acknowledgements

I would first like to thank the Ministry of Higher Education and Research for the attribution of the PhD grant that allowed me to pursue my doctorate.

I also wish to thank Guillaume Beslon, Amel Latifi, Patricia Renesto and Olivier Tenailon for accepting to be part of my defence committee.

I would like to thank to my PhD advisors, Professor Dominique Schneider and Assistant Professor Joel Gaffe for the supporting me throughout the three years of the PhD and the two years of Masters before. They have both been the constant source of energy with their dedication and passion for the subject. Without both their support and their astute supervision I would not be able to present this work.

Thank you for Olivier Tenailon and his laboratory for hosting me and for the help accorded to me on numerous occasions. The time spent in his lab was invaluable to me and to the PhD.

Thank you to Thomas Hindré for first receiving me for my Master 1 project and introducing me to the laboratory and for your help throughout the PhD.

Thank you to Antonia Suau-Pernet for your insights into PCR techniques and for the general good humour you brought to the lab.

Thank you to the other numerous members of the team GEM that I have had the chance to meet and interact with over the years. Without all of your input this work would not have been completed.

A special thank you to the technicians of the team; Evelyne Coursange, Léa Baume and Amandine Gaude for all your good humour and for helping lending me a hand during the most time consuming experiments.

Thank you to my family for all the support me throughout the time and for allowing me to pursue my five years in France.

Thank you to the various flatmates that helped me relax after a day at the lab, for the weekends passed and for helping me discover the region.

# Table of Contents

Table of Figures .....	1
Abstract .....	2
Résumé.....	5
Objectives .....	8
General Introduction .....	11
I. Preface.....	12
Generation of Diversity .....	13
Single nucleotide polymorphisms .....	14
II. Rearrangements.....	14
Mechanisms for rearrangements. ....	15
Types of rearrangements and their effects .....	17
Horizontal gene transfer (Incorporation of foreign DNA).....	18
Rearrangement of existing genome .....	21
Identification of rearrangements .....	21
Limits to genome rearrangements .....	27
Rearrangements in Evolution Experiments.....	27
III. Mobile genetic elements .....	29
IS identification and transposase.....	29
Effect of IS transpositions on gene expression .....	33
Insertion specificity of IS elements .....	33
Control of IS transposition .....	34
IV. The Long Term Evolution Experiment (LTEE) .....	38
Experiment set-up .....	38
Common evolution between the populations .....	39
Genetic analysis of the twelve populations.....	40
Diversity between the populations.....	41
Evolution of polymorphism in Ara-2 .....	42
Evolution of a new catabolic pathway Ara-3 .....	43
Results.....	47
Results Part I.....	48

Feedback from the Editor of <i>mBio</i> and two anonymous reviewers .....	113
Results Part II:.....	123
General Discussion .....	167
References.....	181

# Table of Figures

<u>Figure 1 Mechanisms by which recombinations via RecA can occur</u>	17
<u>Figure 2. The three major types of homologous rearrangement</u>	21
<u>Figure 3. Explanation of Whole Genome Mapping™ (i.e. optical mapping)</u>	22
<u>Figure 4. Stability of the duplication and phenotypes of the different derived genotypes</u>	29
<u>Figure 5. Organization of a typical IS</u>	29
<u>Figure 6. Circular copy-paste IS transposition mechanism common to the IS3 family</u>	31
<u>Figure 7. Method for the long term evolution experiment (LTEE)</u>	38
<u>Figure 8. Average fitness distribution of the 12 evolving populations</u>	40
<u>Figure 9. Insertion of an IS150 element into <i>nadR</i> in population Ara + 1, and a subsequent inversion involving that new element.</u>	42
<u>Figure 10. Divergence of the population Ara-2.</u>	43
<u>Figure 11. Phylogeny of Ara-3.</u>	44
<u>Figure 12. Duplication of citrate transport genes in Ara-3</u>	46

# Abstract

Large-scale DNA rearrangements, including inversions, amplifications, duplications, deletions, insertions, and transposition of mobile genetic elements, are major drivers of evolution and strongly impact on chromosome organization and expression, thereby altering organismal phenotypes. However, their long-term evolutionary dynamics and effects on organismal fitness are often unknown. We addressed these questions by using the longest-running evolution experiment, during which twelve independent populations are propagated from a common *E. coli* ancestor in a glucose-limited environment for now over 60,000 generations (26 years). Most past studies have focused on point mutations and small InDels. Using evolved clones sampled over time in all 12 populations, we characterized all large-scale DNA rearrangements by using whole genome sequences and Whole Genome Mapping<sup>TM</sup> (*i.e.* optical mapping). After 40,000 generations, we identified a total of 110 rearrangements including 82 deletions, 19 inversions and 9 duplications. Many chromosomal regions were repeatedly affected by similar rearrangements and, at least in one population, they occurred early in evolution when fitness increase was strong. Therefore, many rearrangements may be under positive selection. At the very least, these rearrangements strongly affected the structure of the chromosome during evolution.

At the molecular level, we showed that  $\sim 70\%$  of all rearrangements occurred by recombination between Insertion Sequence (IS) elements, illustrating their importance in mediating genome plasticity. We therefore investigated the distribution and temporal dynamics of these small mobile genetic elements in all 12 populations. We showed that IS elements were strong contributors of the total mutations after 40,000 generations. In one population, they even represented about half of the total mutations and one IS type, *IS150*, revealed a strong 6-fold increase in copy number, accounting for the production of most of the rearrangements detected in this population. We showed that *IS150* revealed a dynamic temporal behavior with a strong expansion followed by domestication by the host. By testing



three evolutionary scenarios, we demonstrated that the *IS150* expansion was related to a strong fitness increase conferred by the initial transposition events that occurred before 2000 generations. Later, between 20,000 and 40,000 generations, we measured a decreased transposition frequency, likely owing to a down regulation imposed by the host. Finally, and for the first time, we developed an evolution model of IS dynamics confirming that the IS expansion was related to a threshold number of initial IS beneficial insertions. All of our data showed that large-scale chromosomal rearrangements and IS elements have played an active role in the evolutionary outcomes after 40,000 generations of bacterial evolution.

# Résumé

Les réarrangements d'ADN à grande échelle, tels que inversions, amplifications, duplications, délétions, insertions et transposition des éléments génétiques mobiles, sont des acteurs essentiels de l'évolution. Ils ont une forte incidence sur l'organisation et l'expression des chromosomes, ce qui affecte le phénotype des organismes. Toutefois, la dynamique de ces réarrangements au cours de l'évolution et leurs effets sur l'adaptation des organismes sont souvent inconnus. Nous avons abordé ces questions en utilisant la plus longue expérience d'évolution en cours. A partir d'un ancêtre commun d'*Escherichia coli*, douze populations indépendantes sont cultivées dans un milieu limité en glucose depuis plus de 60 000 générations, soit 26 ans. La plupart des études antérieures ont porté sur les mutations ponctuelles et les petites insertions et délétions (InDels). En utilisant des clones isolés au cours du temps dans ces 12 populations, nous avons caractérisé les réarrangements d'ADN à grande échelle à la fois par l'analyse des séquences de génomes et par cartographie optique. A 40 000 générations, nous avons identifié 110 réarrangements parmi lesquelles 82 délétions, 19 inversions et 9 duplications. Plusieurs régions du chromosome ont été touchées à plusieurs reprises par le même type de réarrangements dans des populations indépendantes. Dans une des populations au moins, les réarrangements se sont produits au début de l'expérience d'évolution, au moment où l'augmentation de la valeur sélective est la plus élevée. Par conséquent, certains de ces réarrangements pourraient être bénéfiques dans ces conditions. Même dans le cas contraire, nous avons montré que ces réarrangements affectaient fortement la structure du chromosome au cours de l'expérience d'évolution.

Au niveau moléculaire, nous avons montré que ~ 70% des réarrangements se produisent par recombinaison entre séquences d'insertion (IS), ce qui illustre l'importance de ces dernières dans la plasticité du génome. Nous avons donc caractérisé la distribution et la dynamique de ces petits éléments génétiques mobiles dans l'ensemble des 12 populations. Nous avons montré que les éléments IS ont fortement contribué à l'ensemble des mutations après 40 000

génération. Dans une population, les IS représentent même la moitié des mutations, et un des types d'IS, *IS150*, présente une forte prolifération avec 6 fois plus de copies à 40 000 générations, intervenant dans la plupart des réarrangements détectés dans cette population. Nous avons montré une forte dynamique temporelle d'*IS150*, avec une forte expansion suivie d'une domestication par l'hôte. En testant trois scénarii évolutifs, nous avons démontré que l'expansion d'*IS150* était liée à une forte augmentation de la valeur sélective conférée par les événements initiaux de transposition ayant eu lieu avant 2000 générations. Plus tard, entre 20 000 et 40 000 générations, nous avons mesuré une diminution de la fréquence de transposition, probablement en raison d'une régulation négative de la transposition imposée par l'hôte. Enfin, et pour la première fois, nous avons développé un modèle d'évolution de la dynamique des IS, qui confirme que leur expansion est liée à un nombre seuil d'insertions bénéfiques initiales. Ces résultats montrent que les réarrangements chromosomiques à grande échelle et les éléments IS ont joué un rôle actif dans la trajectoire évolutive au cours de 40 000 générations d'évolution bactérienne.

# Objectives

As we strive to understand the world around us, the most fundamental question of how we came to be is often evoked. To elucidate the evolutionary processes that may have been involved during the geological ages, many experiments have been designed in laboratories with the aim of seeing evolution in action. Most of these experiments used organisms that reproduce quickly, are easily to manipulate and can be preserved as living fossils. One classical organism used for over 100 years in laboratories is *Escherichia coli*. We can manipulate it in many ways and a wealth of information is available about its biological properties. Therefore, it is not surprising that *E. coli* is used in many evolution experiments in which an ancestral strain of *E. coli* is propagated under defined conditions for various time lengths. Professor Richard Lenski initiated in 1988, 12 independent populations of *E. coli*, founded from a common ancestor, and propagated them since then by daily transfers in a specified environment (See Introduction: Fig 7). The evolution experiment now represents over 60,000 generations of evolution. Many studies have been performed with these populations to characterize the pace of evolution including both phenotypic and genomic changes. All the populations have evolved higher fitness, larger cell sizes and faster division rates. Most genetic studies have focused on specific mutations and on their fitness effect by moving them in ancestral or evolved genetic backgrounds. However, except for few cases, large-scale rearrangements, (deletions, duplications/amplifications, inversions, transposition of Insertion Sequence (IS) elements) have not been exhaustively investigated.

The objective of my Ph.D thesis was to investigate the evolutionary dynamics of chromosomal rearrangements during the long-term evolution experiment of Richard Lenski. This manuscript will be divided in three parts: first, an Introduction will give an overview of the current knowledge about DNA rearrangements and IS elements, and of the evolution experiment. The second part will give the main results of the thesis, and the third part will discuss them. The results of my work are included in two publications, one in revision in

*mBio* and one that will be submitted in the next months. First, I focused on the extensive analysis of all the chromosomal rearrangements that occurred during 40,000 generations of the long-term evolution experiment. To reach this goal, we took advantage of the whole genome sequences and optical maps of evolved clones sampled at 40,000 generations from all 12 populations. The borders of some rearrangements were precisely checked with a PCR strategy. Many rearrangements have been identified, including their molecular mechanisms, and two lines of evidence (parallelism, occurrence early during evolution) suggested that some may be under positive selection.

The second part of the results will investigate the evolutionary dynamics of the main drivers of the identified chromosomal rearrangements: Insertion Sequence (IS) elements. These mobile genetic elements are involved in chromosomal reorganisation and we analysed their distribution and activity in all 12 populations. The IS dynamics is based on IS fingerprints, analysis of genome sequences and *in vivo* measurements of transposition and recombination frequencies. We also, and for the first time, developed an evolution model of IS dynamics. Both the experiments and the model revealed that the dynamics of IS elements, including a strong expansion in one population, was driven by Darwinian selection whereupon the initial insertions of one type of IS elements in one population were strongly beneficial for the bacterial cells.

# General Introduction



## I. Preface

Life is incredibly diverse. There exists life in every environment that humans have looked at, from clams, mussels, and vestimentiferan worms thriving on chemosynthetic microbial living in the deepest oceans near hydrothermal vents (Grassle 1985) to over 17 different bacterial taxa found in the troposphere ( 8-15km) above the earth's surface (DeLeon-Rodriguez, Latham et al. 2013). Not only have ecological niches been occupied but there is also a huge variety in the size and complexity of the forms organisms have incarnated. The smallest free living organism is the prokaryote, *Mycoplasma genitalium*, which has a genome of 580,070 base pairs and a total of 470 predicted coding regions (Fraser, Gocayne et al. 1995). On the other hand, the largest genome known belongs to the eukaryote, Loblolly Pine and spans 23.2 Gbp with a predicted 50,172 genes present (Neale, Wegrzyn et al. 2014). Not only does the size of genomes of organisms vary vastly but so does their biological complexity (Adami 2002). The large variations give the ability for life to colonise every space shows the huge capability organisms have to adapt and to use a variety of available resources.

While we are more perceptible to large, easily observable organisms, it is in fact prokaryotes, comprising of the families Archea and Bacteria, which make up the majority of the individuals alive on the planet. A conservative estimate of prokaryote numbers puts them at  $4-6 \times 10^{30}$  cells in total (Whitman, Coleman et al. 1998) but the actual count is likely to be much higher. The number of species this represents is almost impossible to say as the species boundaries are not clearly defined for prokaryotes nor do we have the capacity to identify all the species. But a rough estimate gives approximately  $2 \times 10^6$  species present in the sea, while a ton of soil could contain up to  $4 \times 10^6$  different taxa (Curtis, Sloan et al. 2002). This huge number of individuals represents a large pool from which there can be adaptation and colonisation of new niches.

Not only do prokaryotes account for the highest number of cells but they are also the most rapidly replicating organisms found on earth. The fastest replicating bacteria can replicate once every 10 minutes given the right conditions (Elsgaard 2011). This is close to the fastest possible duplication within the realms of thermodynamics (England 2013). On the other hand, eukaryotic organisms have a much slower duplication time, as fast dividing human cell may divided in about 24 hours, this may be due to an increase in eukaryotic physical complexity. With the number of cells combined with the capacity for fast growth, prokaryotes have a large potential to create genetic diversity and therefore be capable of colonising new ecological niches.

### **Generation of Diversity**

According to classical theories all life descends from a common ancestor. When we observe the world around us there is, however, a huge variety in the forms that have evolved. Therefore, to understand how these organisms have appeared we must understand the creation of the diversity. The creation and the selection of difference within a population allow it to adapt to new environments (Nevo 2001; Sahney, Benton et al. 2010), to overcome both external and internal stresses and to be more robust to changes. This process occurs due to the selection of individuals within the population that are more adapted to a specific environment, therefore more likely to produce offspring. The process is best described by Darwin “As many more individuals of each species are born than can possibly survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be *naturally selected*. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form.”(Darwin 1859). The creation of diversity comes about

by changes to the genetic makeup of individuals within a population so to truly understand the apparition of diversity we must understand the mechanisms by which they appear.

### **Single nucleotide polymorphisms**

The most common way to create genetic diversity is through point mutations generated by imperfect DNA replication. This happens at a rate of between  $10^{-7}$  and  $10^{-8}$  mutations per base pair duplicated in *Escherichia coli* (Schaaper 1993). This means that on average every two division cycles of *E. coli* will produce one mutation. Estimates for mutation rates in humans give numbers in the same order of magnitude (Crow 1993), however due to larger genome ( $6.4 \times 10^9$ ) this translates to 64 mutations per zygote. Given this high rate of occurrence and ease of identification a lot of attention has been focused on SNP production and effects. However, , we know SNPs are not the only way by which genetic diversity can occur and the advent of modern biological techniques allows to illustrate more in depth these variations. Other less studied mutations in bacteria include chromosome rearrangements and insertion sequence (IS) movements.

## **II. Rearrangements**

Large-scale genomic rearrangements play a major role in the evolution of species, both eukaryote and prokaryote (Putnam, Butts et al. 2008; Toussaint and Chandler 2012) as they can generate effects not producible by SNP including gene and regulatory sequence reassortment, elimination of DNA, creation of duplications and evolution of new genes (Moore 2012). Large-scale DNA rearrangements occur spontaneously at low frequencies. For instance, duplication events arise at frequencies comprised between  $10^{-2}$  and  $10^{-5}$  duplication per cell division depending on their chromosomal location (Anderson and Roth 1981). The

relatively low frequency of chromosome rearrangements can be explained by several factors. First, culturing of strains may induce the loss of unstable DNA rearrangements like duplications, resulting in underestimation of their frequency (Maharjan, Gaffe et al. 2013). Second, these previous estimates have all been performed based on observable phenotypes that were produced after the rearrangements occurred (Zieg and Kushner 1977; Albertini, Hofer et al. 1982; Segall, Mahan et al. 1988; Bierne, Seigneur et al. 1997). Third duplications are inherently unstable and they are prone to reversion that limits the estimation of their rate (Anderson and Roth 1981).

### **Mechanisms for rearrangements.**

Chromosomal rearrangements will occasionally occur during the DNA replication or repair, and will involve DNA recombination. There are several documented methods of genome recombination. The most frequent being homologous recombination between two DNA segments sharing homology. Even though bacteria do not have many homologous genes, recombination occurs between ribosomal operons (Anderson and Roth 1981) and mobile genetic elements like transposons (Cui, Neoh et al. 2012), insertion sequence (IS) elements (Daveran-Mingot, Campo et al. 1998), and prophage sequences (Iguchi 2006). Moreover there are cases of atypical rearrangements which involve homologous sequences (Roth and Wilson 1985; Weller, Kysela et al. 2002). In all cases, homologous recombination requires a double strand break in the DNA and reparation of this break in a new conformation. Two main pathways have been described for bacterial homologous recombination. In both case, the main actor is RecA, a 38kDa protein, ubiquitous in all class of life. RecA is a DNA dependent ATPase and will bind to single strand DNA as a nucleoprotein filament (Patel, Jiang et al. 2010). In *E. coli*, RecA is involved in 2 cellular functions; one is recombinational DNA repair with the binding of single strand DNA and the strand invasion of the homologous DNA molecule. RecA is vital for cells as not only does it contribute to

homologous recombination but it plays a vital role in restarting collapsed DNA replication forks (Cox, Goodman et al. 2000). The second main function is binding to and promoting the self-cleavage of LexA, UmuD and polV to promote the SOS response. (Michel, Boubakri et al. 2007).

Depending of the nature of the DNA degradation, a single strand break (ssDNA) or a double strand break (dsDNA) molecular complexes RecFOR with the ssDNA or RecBCD with dsDNA will bind to the DNA partially degrade a part of the damaged molecule, load the RecA protein on the resulting single strand DNA molecule. The RecA filament will invade the homologous dsDNA and the resulting Holliday junction will migrate under the catalysis of RuvAB and RecG (Michel, Boubakri et al. 2007) (Fig 1). Non homologous end joining also occurs in bacteria and is RecA independent. In some bacteria as *Mycobacteria* and *Bacillus subtilis*, the enzymes involved in the rudimentary bacterial NHEJ machinery consists of two proteins, namely the multifunctional ATP-dependent ligase-D and Ku (Della, Palmos et al. 2004). However, in *E coli*, Ku-like and ligase-D like homologue have not been detected and an alternative End-Joining mechanism involving the multiprotein complex RecBCD had been described (Chayot, Montagne et al. 2010). There are other types of recombination that are mediated by other methods than the main RecA pathway. Transposition of specific sequences such as IS elements are a prime example and are discussed in length later on.

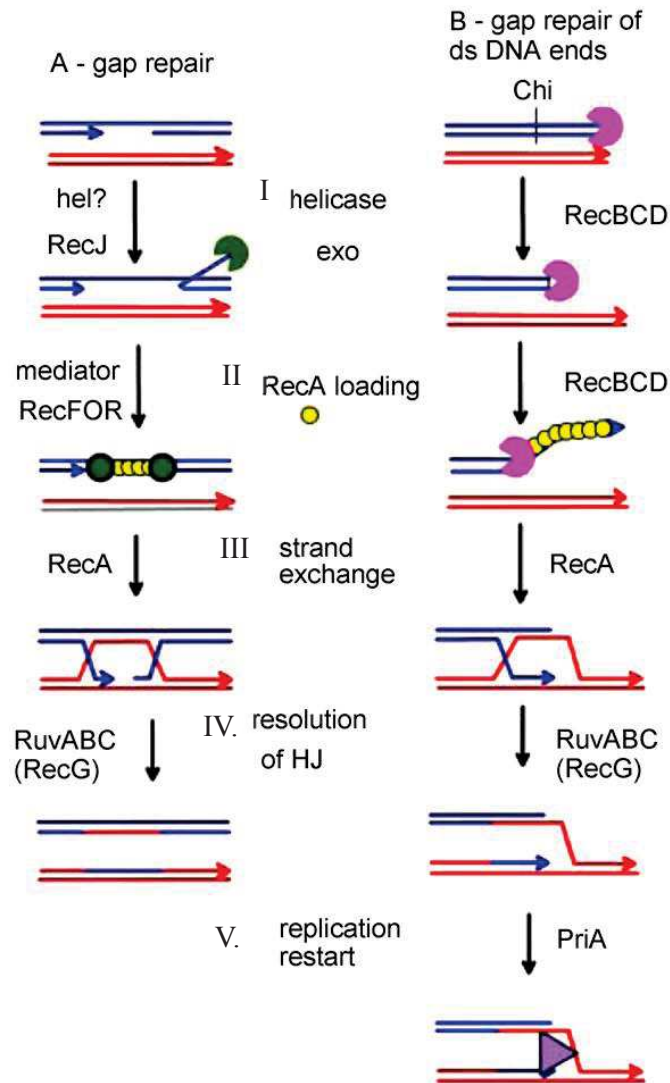


Figure 1: **Mechanisms by which recombinations via RecA can occur.** The principal activity of RecA is to fill gaps in replication and to rescue replication fork collapse which has a secondary effect leading to recombination between the two DNA strands. A. Example of DNA repair mechanism of RecA. B. Example of a gap repair. Both cases require I: unravelling of coiled DNA. II: binding of RecA machinery. III: Strand exchange due to homologous sequences. IV: resolution of Holliday Junction and V: restart of replication fork (adapted from Michel 2007).

### Types of rearrangements and their effects

Two classes of rearrangements are described I) Incorporation of foreign DNA acquired by either horizontal transfer or through DNA scavenging and II) rearrangements of existing genes.

### **Horizontal gene transfer (Incorporation of foreign DNA)**

Incorporation of foreign DNA is a major pathway for new gene acquisition in bacteria. This is especially important in the dissemination of antibacterial resistance genes throughout an ecosystem (Skippington and Ragan 2011). There are three main ways new DNA can be incorporated into the cell (Halary, Leigh et al. 2010), the first and most common is through conjugative transfer where there is a physical contact between two cells and DNA is exchanged through the gap that is created through both cell membranes. The main type of genetic material to be exchanged is F-plasmids that contain all the genes necessary for its own replication and sometimes can also contain gene coding for antibacterial properties. The donor and recipients are generally closely related species or found within the same order, but it can also occur between widely different species, for example between bacteria and fungi (Zhang, Pereira et al. 2014). This has a large impact on the evolution of both organisms as it gives them access to genes not found within their own pangenome (Zhang, Pereira et al. 2014). This method of genome rearrangement is very well studied due to its prevalence in evolution and dissemination of antibacterial resistance genes that can be carried by pathogenic islands on the transferred genetic material (for review see Schmidt and Hensel 2004). Examples of resistance acquisition due to conjugation is frequently described (for reviews see Alekshun and Levy 2007; Acar and Moulin 2012) and occur in almost any environment where there is selection for resistance including in many animals (Ahmed, Clegg et al. 2010; Pan, Yuan et al. 2014). Samples taken throughout nature also reveal that resistance carrying plasmids are widely spread in most environments and therefore constitute a very easily accessible pool of genes (for review see Allen, Donato et al. 2010). Conjugation is also involved in the formation of social structures in biofilms and spreading of co-operative traits (Ghigo 2001; Nogueira, Rankin et al. 2009).

A second type of horizontal transfer is transformation, where a recipient can take up free DNA from the environment. There is some debate as to how the trait evolved. There is a high relative fitness cost for competent bacteria, due to associated protein synthesis and DNA length augmentation, which might explain why not all bacteria are competent (Peterson, Cline et al. 2000). Therefore, in naturally competent cells there is a strict and intricate regulation of the many genes required (over 40 in *Bacillus subtilis*) (Hamoen, Venema et al. 2003). In general, DNA uptake in competent bacteria requires the presence of type IV pili or pseudo pili (Chen, Provvedi et al. 2006). The large amount of DNA found in the environment could confer a number of new functions to bacteria. These include the acquisition of genes for resistance, biofilm and other beneficial aspects. Competence is also a way for acquisition of undamaged nucleotides. This in some part mitigates the effect of DNA damaging antimicrobials, especially antimicrobials creating double strand breaks. *S. pneumoniae* transformation protects against the bactericidal effect of mitomycin C and streptomycin (Engelmoer and Rozen 2011). It is shown that most antibacterial treatments that produce double strand breaks in bacteria can induce competence in bacteria that are lacking the SOS-RecA response mechanism; *E.coli* has this mechanism and is therefore not naturally competent (Charpentier, Kay et al. 2011; Charpentier, Polard et al. 2012). Giving argument that competence is an adaptation to reduce stress on the genome.

A third type of rearrangement involving foreign DNA is transduction. This is the incorporation of a viral genome into the host's. The viral DNA might carry additional genes that confer an advantage to the host. This form of horizontal transfer has the advantage that DNA is protected from DNAses in the environment; however it has a big cost to the cell as it is infected by a phage which has more potential to damage the cell. The most common example that is routinely used in laboratory settings is the lambda phage of *E.coli*. While it is a lytic bacteriophage, lambda phage can undergo a lysogenic phase wherein the phage DNA



is incorporated into the host genome and then virulence factors are down regulated, leading to the phage maintaining itself in the genome and are subsequently included in DNA replication. The insertion is site specific and the location of the site suggests that there has been co-evolution between lambda phage and *E.coli* to facilitate this interaction (Tal, Arbel-Goren et al. 2014). While most of the literature focuses on lytic phages there are a large number of examples of temperate phages (for review see Fortier and Sekulovic 2013) Lysogenic phages can bring in new genes to the genome including virulence factors and can convert a non-pathogenic bacteria to a pathogenic one (Waldor and Mekalanos 1996). In addition to the transfer of virulence genes, prophages allow bacteria to adapt to rapidly changing environments. For example, they are important for gene expression in the various stages in the complicated life cycle of *Bacillus anthracis* (Schuch and Fischetti 2009), for the biofilm formation (Stanley and Lazazzera 2004) and can be responsible for spore formation in *Bacillus subtilis* (Silver-Mysliwiec and Bramucci 1990). Since prophages are principally involved in adaptation to new environments, one can theorise that in a stable conditions, there might be a loss or attenuation of the dissemination of prophages.

Even though horizontal transfer is an important means of DNA rearrangements, there are organisms that are not naturally competent and cannot incorporate foreign DNA into their genomes. Many do this to limit the uptake of pathogens and DNA is broken down within the cells. However, organisms are able to increase their genetic diversity by the modification of their existing chromosome, rearrangement of genetic material has been documented in organisms representing eukaryotic, bacterial and archean domains of life.

## Rearrangement of existing genome

Rearrangements of the existing genome play a major role in evolution. This can be in the form of deletions, inversions, duplications and amplifications of parts of the genome (Fig 2).

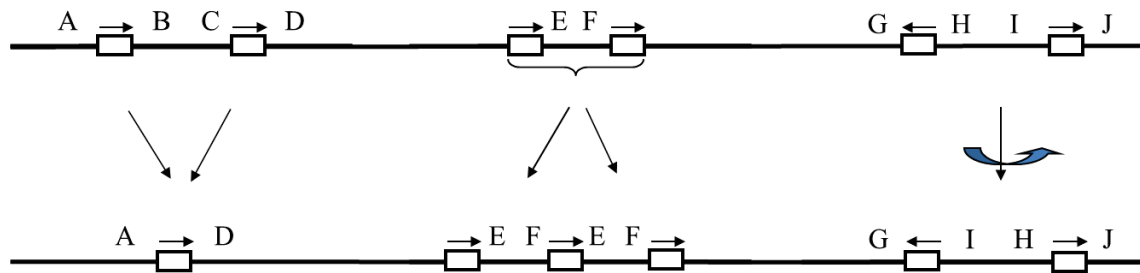


Figure 2: **The three major types of homologous rearrangement.** The three major chromosomal rearrangements, deletions, duplications and inversions are represented here. Thick black lines indicate the chromosome, boxes indicate homologous sequences and the arrows above them indicate the direction of their homology. Letters indicate various chromosomal locations.

## Identification of rearrangements

Chromosomal rearrangements are usually investigated by using methods such as pulsed-field gel electrophoresis, microarray-based hybridization experiments, and whole genome sequencing. Unfortunately, typical whole genome sequencing technologies that rely on sequencing short DNA fragment libraries cannot detect all large-scale DNA rearrangements, in particular inversions and other rearrangements involving long sequence repeats. Optical mapping techniques have been developed that produce a high-resolution ordered restriction map of bacterial genomes that can complement sequencing data (Fig. 3) (Schwartz, Li et al. 1993; Kotewicz, Jackson et al. 2007; Shukla, Kislow et al. 2009; Turner, Yomano et al. 2012).

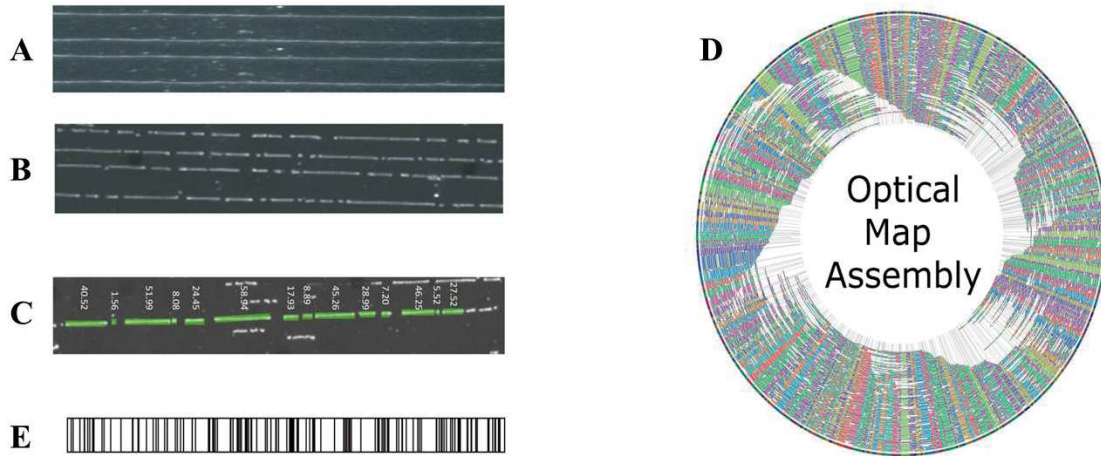


Figure 3: **Explanation of Whole Genome Mapping™ (i.e. optical mapping).** Methodology used by OpGen© to construct optical maps from purified DNA. A. DNA molecules are stretched and fixed to a glass plate by charge. B. Fixed DNA is washed with a restriction enzyme. C. Restriction profiles and fragment lengths are taken. D. Multiple fragments are used to create a whole genome profile. E. Genome profiles are compared to an *in silico* reference (adapted from Shukla, Kislw et al. 2009)

### Deletions

Deletions are the most commonly observed rearrangements observed in evolving populations and perform a versatile role in the evolution and the adaptation of populations to their environment (Mira, Ochman et al. 2001). They may cause massive gene loss and are important for the evacuation of superfluous genes that are not expressed in certain conditions (Andersson and Andersson 1999).

The extra genes are therefore a cost to the organism as it increases the time and energy required for duplication and there may be residual transcription of unneeded proteins. Obligate endosymbiont are a particular example as they show that with a massive streamlining of the genome they evacuate genes. Up to now, the smallest known genome is attributed to *Nasuia deltocephalinicola* an endosymbionte from sap feeding insect, with a 112 kb genome carrying only 137 protein encoding genes (Bennett and Moran 2013). This represents an extreme in genome reduction. Genome reduction is also observed in the adaptation to any stable environment. In cystic fibrosis patients, deletions play an important

part in the long-term evolution of *P. aeruginosa*, with some isolates losing up to 8% of its genome by homologous and non-legitimate deletions (Rau, Marvig et al. 2012).

*Salmonella enterica* cultured in rich media over 1000 generations showed significant levels of selection driven gene loss. Rearrangements occurred between *Tn10* elements interspersed over the genome with some regions being more targeted. This further suggests that deletions are a way of evacuating superfluous DNA (Koskiniemi, Sun et al. 2012). Deletions of specific gene targets are also a common feature during evolution including the emergence of resistance to antimicrobial peptides in *E.coli* where a deletion of a nonessential gene renders the bacteria resistant to a multi target peptide (Narayanan, Modak et al. 2014). Deletions can sometimes even be development-regulators, by creating chimeric genes involved in sporulation (Abe, Yoshinari et al. 2013). In this case, a 48 kb element of the genome is excised from *Bacillus subtilis* which leads to the formation of a functional regulator *sigK* (Abe, Yoshinari et al. 2013).

### **Inversions**

Inversions occur when a region of the chromosome is cut out of the chromosome by double strand breaks and then reinserted with the opposite orientation. Comparisons of genomes from closely related bacterial species have detected a high level of DNA inversions among chromosomal rearrangements (Eisen, Heidelberg et al. 2000; Tillier and Collins 2000; Zivanovic, Lopez et al. 2002). Inversions play several roles in the genome. First it might change gene expression by either breaking the genes at the end points of the inversion, or by changing the orientation of genes in relation to the replication fork. In effect, genes that are transcribed in greater quantities are in the same direction as the replication fork on the leading strand, as this minimises the incidences of collisions between the RNA and the DNA polymerases (Rocha and Danchin 2003). Second inversions change the genome conformation by displacing secondary structures and by displacing genome features involved with physical

chromosome conformation such as the *parS* site (Umbarger, Toro et al. 2011). However, some genome features are not largely impacted since a selection bias exists leading to the over representation of inversions around the replication origin in bacteria that maintains the symmetry of the genome (Eisen, Heidelberg et al. 2000; Darling, Miklos et al. 2008).

The actual benefit of inversions is hard to quantify but they are present in a large number of species and can happen very easily. After sub culturing *E. coli* O157:H7 for only 50 passages on new media, one third of the isolated clones were found to have novel inversions (Iguchi A 2006). The comparison of *Vibrio cholerae*, *Streptococcus pneumoniae* and *Mycobacterium tuberculosis* to those of closely related species – *E. coli*, *Streptococcus pyogenes* and *Mycobacterium leprae* reveals that inversions (up to 10% of the genome) happen on a large scale during evolution, especially around the origin and terminus of replication (Eisen, Heidelberg et al. 2000). A similar pattern can be found in *Pseudomonas stutzeri* (Ginard, Lalucat et al. 1997), *Pseudomonas aeruginosa* (Kresse, Dinesh et al. 2003), *Francisella tularensis* (Rohmer, Fong et al. 2007), *Neisseria gonorrhoeae* (Spencer-Smith, Varkey et al. 2012), *Lactococcus lactis* (Daveran-Mingot, Campo et al. 1998), and *Staphylococcus aureus* (Cui, Neoh et al. 2012) when they are all compared within their respective pangenomes. On average 20 % of the genome is inverted in these cases.

### **Duplications and Amplifications**

Duplications and amplification events are among the most important rearrangements to occur in asexual bacteria. Duplications is defined by a copying of existing genome while an amplification can result in more than two copies of a stretch of genome being present. The advent of mass genome sequencing of bacterial genomes reveals that the majority of duplications are small scale duplications such as tandem and operon duplications (Gevers, Vandepoele et al. 2004)

A variable gene dosage caused by duplications allows organisms in overcoming limitations linked to inefficient gene expression as it increases the gene product. For example, in helping an evolving population to overcome limited expressions of specific functions such as to increase lactose uptake in lactose limited environments in *E.coli* (Novick and Horiuchi 1961), escape from certain stringent conditions by amplification of specific genes followed by mutations in defective genes (Pranting and Andersson 2011), modification of gene expression (Kugelberg, Kofoid et al. 2006) or even contributes to antibiotic or metal resistance (Sun, Berg et al. 2009, von Rozycki et al. 2009). The temporary nature of duplications allows organisms to revert to their original state once the selection pressure is released (for review see Kondrashov 2012). Duplications can also have negative effects on the organism due to superfluous gene expression. This is particularly the case in eukaryote cells and is linked to diseases such as defective eye formation in *Drosophila*, (Bridges 1936), causes unbalance in protein complexes that are toxic in *Saccharomyces cerevisiae* (Papp, Pal et al. 2003) or even is linked to autism in humans (Sebat, Lakshmi et al. 2007). In bacterial genomes even mildly deleterious effects of gene duplication are rapidly purged from the population.

Not only do duplications have an immediate impact on cells through altered gene dosage but there are many long term effects associated. Duplicated genes can act as a redundancy mechanism whereby if one gene is affected by a deleterious mutation there is always another functional copy (Clark 1994). The long term fate for the duplicates is not fixed and it largely depends on the fitness change brought by the duplication. The most common fate for duplicated genes is to accumulate a number of mutations that inactivates them and creates pseudo genes from the duplicates (Lynch and Conery 2000). This fate, while common in eukaryotes, is less common in bacteria due to the cost increase in duplicating non-coding DNA. A second fate for gene duplication is for the distribution of the gene function over several genes. Through observation of the genomes of eukaryote organism it is clear that this

model explains the phenomena of complementary of gene functions (Force, Lynch et al. 1999). This method of gene dissemination is also present in most sequenced bacteria and is true for around 89% of all homologous gene families with the other 11% accounted for by horizontal gene transfer (Pushker, Mira et al. 2004). Finally, one of the duplicates can be subject to a rare beneficial mutation that endows it with a novel gene function while the remaining copy retains its ancestral function (Nasvall, Sun et al. 2012). This method of *de novo* gene creation is especially relevant in asexual species where there is no exchange of DNA. As many genes have secondary gene functions, duplication of a gene can favour this secondary gene product in a specific environment and therefore the duplication is maintained due to the secondary function rather than the primary function. A mutation might occur in one of the gene copies that might amplify the effect of a secondary gene function or the loss of the primary gene function and through gradual genetic drift the copies might diverge to the point where both genes have different functions (Lynch and Force 2000; Ward and Durrett 2004; Elliot 2012). Another, less appreciated function of duplications is gene shuffling, similar to shuffling produced by inversions, where regulatory networks of genes can be changed. This can be due to promoter capture to increase gene expression where a copy is placed under the control of a downstream promoter resulting in the expression of a hybrid protein (Whoriskey, Nghiem et al. 1987) or the expression of the intact ORF under the control of another promoter (Blount, Barrick et al. 2012). Classical models for gene duplication assume that there is no cost for the cell; newer models have shown that gene duplication has a substantial cost to cell fitness (Reams, Kofoid et al. 2010; Adler, Anjum et al. 2014). Therefore, due to the cost, tandem duplications tend to be unstable within a population and are often resolved by recombination before fixation (Reams, Kofoid et al. 2010). To achieve fixation within a population, duplications need to achieve equilibrium between benefit and cost.

## **Limits to genome rearrangements**

While the general consensus is that rearrangements occur randomly, there are several constraints that may affect the occurrence of DNA rearrangements.

First, selective forces preserve a certain amount of symmetry in the size of the two replichores of a circular chromosome between the origin and terminus of replication (Roth JR 1996; Eisen, Heidelberg et al. 2000). This may partly explain the strong conservation of gene order (synteny) between *E. coli* and *Salmonella*, although large DNA inversions can be observed in laboratory conditions where those selective pressures may be relaxed.

Second, the structural organisation of the *E. coli* chromosome has been shown to affect the probability of DNA rearrangements (Esnault, Valens et al. 2007). Specifically, the genome is physically organised into distinct macrodomains (Boccard, Esnault et al. 2005), and rearrangements affecting the replication origin or terminus domain and DNA inversions between the left and right macrodomains have been shown to be detrimental due to their effects on replication fork progression (Esnault, Valens et al. 2007).

## **Rearrangements in Evolution Experiments**

Chromosomal rearrangements have also been identified in evolution experiments whereupon bacterial strains are propagated for various times under controlled conditions. It can be theorised that this is due to the fact that many evolution experiments are a constant environment, therefore radical optimisation of the genome is highly beneficial in these circumstances. Deletions, duplications, large-scale inversions have all been detected in *E. coli* populations that have been propagated for various time lengths under different conditions. It is important to keep in mind that what is observed in these populations does not constitute the whole population as they are samples taken from each experiment. Two types of experiments are routinely used, batch cultures where populations routinely go through different growth



phases and a dilution phase and continuous cultures where populations are in a constant environment without a dilution phase throughout the experiment. Batch cultures provide the largest amount of data as they have been more studies. During these experiments, there have been multiple deletions and amplification events that have conferred fitness advantages and even phenotypic innovation (Schneider, Duperchy et al. 2000; Cooper, Schneider et al. 2001; Blount, Barrick et al. 2012). In 3 out of 6 lines adapting to thermal stress there was evidence of duplications and deletions. 3 of the duplications affected the same chromosomal region containing 4 candidate genes for increased fitness in the conditions. The parallelism indicates that they were beneficial, even though it was not quantified (Riehle, Bennett et al. 2001).

Continuous cultures of *E.coli* evolving in nutritional limitation chemostats have adapted to their environment including fixing a duplication containing the sigma factor *rpoS*. One of the copies then had a deletion of a part of the duplication. When clones were cultured outside its evolution environment the duplication was readily resolved to either the ancestral or the new (containing deletion) version (Fig 4). This shows that rearrangements can be specific to given environments and that rearrangements not previously seen can be observed in specific conditions. The exact beneficial advantages brought by this rearrangement are not known but could represent a second order process conferring a higher evolutionary potential (Gaffe, McKenzie et al. 2011; Maharjan, Gaffe et al. 2013). A recent study (Sun, Ke et al. 2012) based on whole genome sequencing of a bacterial population of *Salmonella enterica* var. Typhimurium that was propagated for 240 generations in a chemostat revealed that duplications, inversions, and small deletions were detectable in  $\geq 20\%$  of the population after only 50 generations of growth.

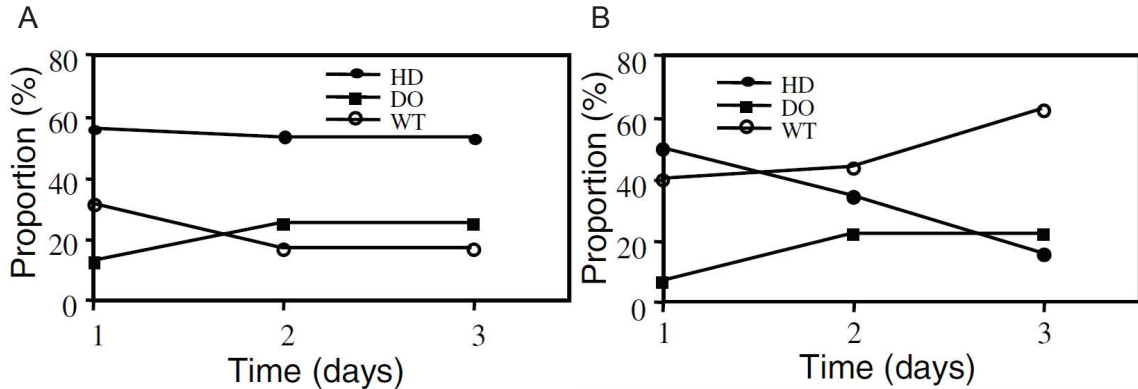


Figure 4: **Stability of the duplication and phenotypes of the different derived genotypes.** The proportion of clones carrying duplications or derived copies in different conditions after three days of culture. HD stands for clones with the duplication WT and DO stand for clones that have either one wild type copy or one evolved copy of the duplication. A. In the evolution conditions (chemostat). B. In another condition (LB 24hr batch cultures) (adapted from Maharjan, Gaffe et al. 2013).

### III. Mobile genetic elements

Mobile genetic elements can be split into two groups, those able to transfer horizontally between chromosomes (plasmids and bacteriophages) and those that are only mobile between genetic material from within the cell or necessitate a vector for horizontal transfer (transposons and insertion sequences (IS)).



Figure 5: **Organization of a typical IS.** The IS is represented as an open box in which the terminal IRs are shown as blue arrows labelled IRL (left inverted repeat) and IRR (right inverted repeat). A single open reading frame encoding the transposase is indicated as a green arrow. XYZ enclosed in a pointed box flanking the IS represents short DR (Direct Repeat) sequences generated in the target DNA as a consequence of insertion.

#### IS identification and transposase

Insertion sequences (IS) are among the smallest transposable element present in the prokaryote genome (between 700bp and 2,5kbp (Mahillon and Chandler 1998)). These elements are ubiquitous in almost every bacterial genome and all the information relevant to IS are stored in a specialised database called ISbiofinder (Kichenaradja, Siguier et al. 2010).

They are classified into families depending on (1) their primary sequence and their gene products (Siguier, Goubeyre et al. 2014); (2) the length and sequence of the short imperfect terminal inverted repeat sequences (IRs) varying from 10-40bp carried by many ISs at their ends; (3) the length and sequence of the short flanking direct target DNA repeats (DRs) often generated on insertion up to 10bp; (4) the organisation of their open reading frames or (5) the target sequences into which they insert (Fig 5). Up to now over 2000 IS have been identified (Mahillon and Chandler 1998), organised into some 20 major families. Some of these IS contain additional functions encoded by partial translation of ORFs or on separate ORFs that regulate the levels of transposition and some contain genes of unknown function (for reviews see Nagy and Chandler 2004; Gueguen, Rousseau et al. 2006)

IS elements may act as simple transposable elements capable of replicating and/or translocating to different sites within the genome or onto plasmids in the bacterium, but IS elements are also responsible for the formation of composite transposons, by flanking on either side, unrelated genes causing the transposition of everything in between both IS as well as the IS (Rosner and Guyer 1980). Upon insertion IS elements often create a direct duplication of the target site which varies depending on the element (Mahillon and Chandler 1998). This can be up to 10bp creating a substantial effect even if the IS is later excised as the direct repeat is maintained in the genome. This is also variable for some IS elements (Iida, Hiestand-Nauer et al. 1985). There are three main families of IS transposases, the classical and most abundant type being the DDE transposase (present in the major families IS1, IS3, IS4 and IS5), named after the conservation of two Asp and a Glu at the active site. It can be encoded by one or two ORF and are closely related to the retroviral integrase catalytic core (IN) by the spacing of the DDE triad and by the appearance of additional conserved residues. The Holliday junction resolvase RvuC is also related. A structurally related transposase type only found in the IS family IS110 is DEDD (three Asp and a Glu); however, they do not

create inverted repeats nor direct repeats upon insertion as observed with DDE transposase. The second family encodes for a tyrosine transposase (or HUH transposase) (Ronning and Guynet 2005), largely present in the family IS91, IS2000 and IS605. The transposase contains no inverted repeats nor does it cause a direct repeat upon insertion. Transposition occurs through the formation of a hairpin secondary Structures coded for by subterminal sequences. The third major family of transposase are serine transposases carried by the families IS607. They are closely related to serine resolves carried by Tn3 elements. Interestingly they are the only bacterial transposase family that are present in several eukaryote genomes, but are mostly likely present due a capture of a longer bacterial DNA through infection rather than capture via transposition (Gilbert and Cordaux 2013).

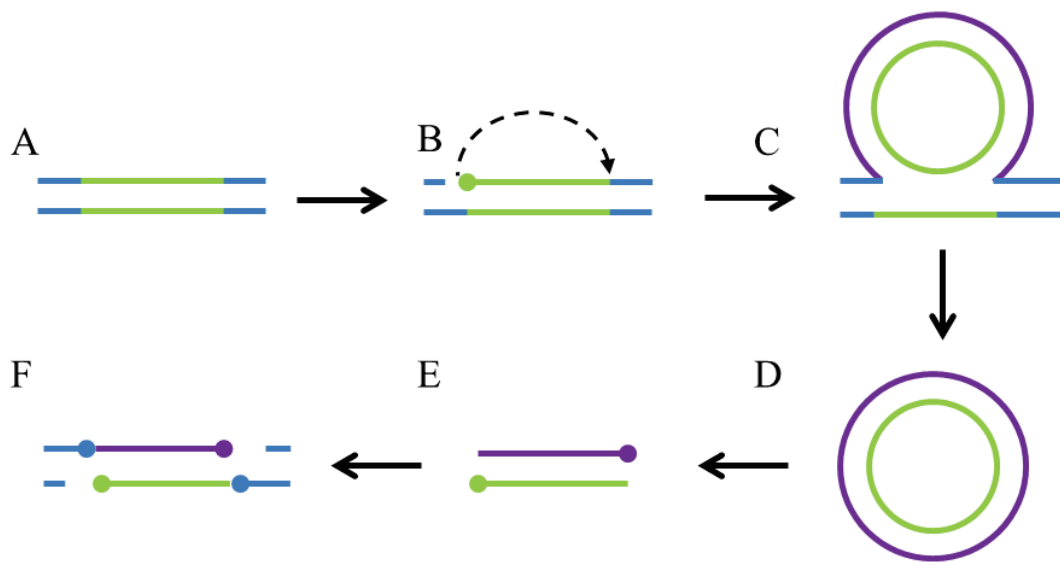


Figure 6: **Circular copy-paste IS transposition mechanism common to the IS3 family.** A. Representation of a dsDNA (blue line) containing a IS element (green line) B. A nick is generated at one 3' end. The resulting 3' OH attacks the same strand immediately outside the transposon (dotted line with arrow). C. Replication is primed from the nick at the transposon junction resolves this intermediate into D. an excised transposon circle. E. A second nick occurs at the ends of the transposase to create a linear product. F. The transposases then inserts the transposon between two staggered nicks in the target which is repaired by host enzymes

Transposition of IS elements can occur via 2 mechanisms depending on the IS. First, conservative transposition through a cut paste method, where the overall number of copies of IS elements is conserved. Examples of this type of replication are IS21 and IS30. The IS physically transpose from one region of the genome to another (Leonard and Mahillon 1998) by the following events, a double strand break occurs on both sides of the IS and a staggered break at the target site. The IS is then ligated into the target site. Even though there is no trace of the IS element at the original site there is still the direct repeat caused by the insertion of the element present. The second method of transposition is by replicative transposition. This mechanism allows for the increase in copy numbers of the IS element in the cell. This is due to the IS forming a complex intermediate form, for example circular (IS150) or a figure of eight (IS911) (Haas and Rak 2002). In the case of IS150, transposition occurs by circular intermediates. The IS150 transposase is regulated by a frameshift that occurs at a heptanucleotide AAAAAAG between two ORFs *insA* and *insB*. It is also regulated by the gene product of *insA*. To transpose, the DDE transposase catalyses the cleavage of one DNA strand to form a 3'OH at one end of the IS. This then serves to attack several nucleotides exterior to the second end to generate a single-strand bridge, with a three nucleotides between the two ends, leaving a free 3'OH on the IS flank. The 3'OH can act as a replication primer. IS replication would regenerate an intact copy reconstituting the donor plasmid and produce a double-strand circular DNA intermediate. Due to low basal transposase levels, this initial step may occur in a stochastic manner. However, formation of the circular intermediate results in the assembly of a transient strong promoter composed of a -35 promoter element in the right IS end oriented outwards and a -10 promoter element in the left end oriented inwards (Ton-Hoang et al. 1997). This promoter serves to drive transposase synthesis and consequent integration and disassembly of the promoter. Thus, the circular intermediate once generated is committed to terminate transposition (Fig 6).

Some IS elements, such as *IS1* are capable of undergoing both types of transposition (Biel and Berg 1984). *IS1* undergoes replicative transposition in 3/4 cases and conservative transposition is due to an imperfect replication.

### **Effect of IS transpositions on gene expression**

IS elements have strong effects on the regions that they inserted into. The most common effect is gene inactivation due to the IS disrupting the coding sequence (Kumar, Grover et al. 2014). This is one of the most efficient ways the genome has of disrupting potentially superfluous gene expression (Nakamura and Inouye 1981; Park, Lee. et al. 2014). Due to highly mobile nature of IS elements, this gene inactivation is potentially reversible if the direct repeat left by the IS element is not disruptive. This is of particular concern in pathogens as IS elements play a regular role in the activation of pathogenic genes which can be selected for in the right environments (Ziebuhr, Krimmer et al. 1999). This can lead to activation or increased expression of antibacterial resistance genes especially for multidrug efflux systems, creating a major health problem (Jellen-Ritter and Kern 2001; Aubert, Naas et al. 2006). Gene expression can also be modified by insertion of IS elements into promoter regions. This can either reduce expression but can also increase expression levels if the IS is inserted into a repressor sequences. IS can also have an impact on gene expression because several have outward facing promoter regions (for example -35 boxes (Prentki, Teter et al. 1986; Flechard 2014), when inserted in phase and in the correct context with a gene they can activate its transcription (Camarena, Poggio et al. 1998; Kallastu, Horak et al. 1998).

### **Insertion specificity of IS elements**

Most of the IS target sites are thought to be randomly distributed, however some IS elements have a preference for certain sequence characteristics, potentially biasing the genes that are affected by IS transposition. For example the transposase of *IS903* is able to form dimers or

multimers and recognise specific sequences inducing transposition to specific areas (Hu and Derbyshire 1998). The more common *IS1* has a preference for AT rich sequences through interaction with the transposase (Zerbib, Gamas et al. 1985). Some IS show a preference for certain DNA secondary structures such as repeated extragenic palindromes (Clement, Wilde et al. 1999; Tobes and Pareja 2006), integrons (Tetu and Holmes 2008) or other even IS elements (Hallet, Rezsöházy et al. 1991).

### **Control of IS transposition**

The transposition of IS elements is strictly controlled by a multitude of mechanisms and is generally kept at low levels. This is probably to offset the deleterious effects of excessive transposition as most mutations are associated with negative effects on bacteria (Doolittle, Kirkwood et al. 1984). IS elements auto regulate their transposition activity in several ways that can be complementary, these are the most common mechanisms with several other mechanisms are specific to a small number of IS elements. I) The most common IS families (families *IS1*, *IS3*, *IS4* and *IS5*) have an ORF that codes for small transcriptional repressors of the transposase as either a separate ORF or more commonly as a truncated version of the transposase (Zerbib, Polard et al. 1990; Escoubas, Prere et al. 1991). II) Several IS families (families *IS1*, *IS3*, *IS5* and *IS630*) require a programmed ribosomal frameshift during translation of the transposase RNA (Escoubas, Prere et al. 1991; Vogele, Schwartz et al. 1991). This commonly leads to 3 gene products being expressed, *InsAB* – the transposase, *insA*- a regulatory protein and *insB* of unknown function (Vogele, Schwartz et al. 1991; Haas and Rak 2002). III) transposase is unstable. The transposase of *IS903* and *IS50* are sensitive to degradation by the Lon protease (Derbyshire, Kramer et al. 1990; Koonin and Ilyina 1993) and the transposase of *IS911* is sensitive to exposure to 42°C (Haren, Betermier et al. 1997). IV) many IS elements, through the formation of secondary structures in mRNA initiated from outside the IS, protect themselves from impinging transcription (Davis, Simons et al. 1985;

Krebs and Reznikoff 1986).V) small residual IS elements are capable of producing truncated transposases that bind with functional transposases to regulate transposition (Gueguen, Rousseau et al. 2006). VI) certain IS elements lack transcriptional stop sites and rely on insertion next to stop codons to have a functioning transcription of the transposase (De Meirsmen, Van Soom et al. 1990). VII) Antisense RNA regulates the translation of *IS10* transposase and impairs the binding of the ribosome to the RBS (Jain 1997).

IS elements generally exist in low copy numbers in the bacterial genome and their transposition is highly limited (Mahillon and Chandler 1998). It is hypothesised that they are a recent addition to the bacterial genome as they are highly conserved between species on the other hand their functionality stems from their highly conserved sequence. The origin and capture of IS elements in prokaryotes is not well understood and none have been catalogued. However, DDE transposases, specifically from the IS3 family, are closely related to the retroviral integrase catalytic core (IN) which could indicate an evolutionary link (Haren, Ton-Hoang et al. 1999). The persistence of IS elements in the genome is a topic of debate and there emerges two hypotheses. The first hypothesis is that movements of IS elements have mostly negative effects on genomes by breaking genes and their functions (Charlesworth, Sniegowski et al. 1994). Therefore they are maintained in the cell by a high level of replication to additional genomic sites and onto plasmids, this coupled with horizontal transfer allows for the persistence of an element even if it's deleterious to the cell. The second hypothesis is that even though the majority of IS related events are deleterious they are capable of creating enough beneficial mutations, through insertion, to allow for their fixation within their host (Schneider and Lenski 2004).

Evidences exist for the beneficial effects of IS elements. IS elements are capable of rapid expansion in copy numbers present on a genome. This leads to massive amounts of gene inactivation, genome rearrangements and genome reduction which in certain contexts is a



beneficial adaptation. While the exact nature of control over this phenomena is not known, there is a correlation with high levels of IS elements and the recent passage of a bacteria to a facultative endosymbiont (Bordenstein and Reznikoff 2005; Touchon and Rocha 2007; Gil, Belda et al. 2008). One explanation could be that generally IS movement is negatively selected for, as it has strong polar effects, until the host finds itself in an stable environment where it renders genes redundant due to non-expression or even constitutive expression that is not beneficial (i.e. the passage to facultative endosymbiont) and therefore IS movements won't have such large effects as there are more targets with neutral effects and they will not be counter selected (Touchon and Rocha 2007). The inactivation of genes would even represent a fitness gain for the host in these circumstances, so much so that IS elements can account for up to 23% of the genome of newly formed endosymbionts (Gil, Belda et al. 2008; Schmitz-Esser, Penz et al. 2011). Several different IS elements are able to sweep through populations in these case creating a genome that has a high level of homologous sequences (Cerveau, Leclercq et al. 2011). This represents the step before the massive genome reduction that is more commonly associated with endosymbionts. However, the exact mechanisms for IS increase is not understood, while there is evidence that higher transposition rates increased levels of IS is thought to be more of a stochastic process (Kleiner, Young et al. 2013). This process could also apply to any evolving population that is kept in a steady environment over a long period of time. *E.coli* was cultured for up to 480 generations in a glucose limited chemostat. This represents a stable environment where there can clearly be redundancy in the genome as genes are not expressed. As such, IS elements largely participated in the adaptation of several populations. Changes included insertion of IS1 into the stationary phase sigma factor *rpoS* and insertion of IS5 into a gene controlling mutation rates (Gaffe, McKenzie et al. 2011). Significant IS movements were also categorised in other evolution experiments. IS10 is one of the major drivers in the evolution of *E.coli* to osmotic pressure

with a beneficial mutation linked to *IS10* present in all ten evolved populations (Stoebel and Dorman 2010).

With the increase in IS elements over the whole genome there is an increase in homologous targets. This allows for virtually any rearrangement to potentially occur. For example in the, new insertions of *IS150* create the opportunity for the deletion of ribose operon in evolving populations (Cooper, Schneider et al. 2001). *IS150* is equally important for evolution to freeze thaw cycles in *E.coli* (Sleight, Orlic et al. 2008). IS elements are also important for the adaptation of *Lactococcus lactis* to cycles of growth/starvation (Sleight, Orlic et al. 2008) and to allow the utilisation of cobalt in *Methylobacterium extorquens* (Chou, Berthet et al. 2009).

It is therefore of crucial importance to understand the evolutionary role of IS elements and in particular their relative involvement in the overall bacterial mutagenesis. There have been two methods to approach this question until now. The first is experimental measures of IS transposition with tools that are linked to particular genetic loci. In one experiment the insertion of IS was determined in the LacZ loci, this is an especially interesting tool to quantify IS movement rates (Huisman and Kleckner 1987). Another approach uses a system where a reporter gene lacking transcription and translational start points is flanked by *IS1* IRs. The plasmid also contains the *IS1* transposase without the ribosomal frameshift which forces an under regulation of transposase and a resulting increase in transposition. The GFP is only produced if it transposes in the correct location. This gives an estimate for *IS1* transposition (Saito, Chibazakura et al. 2010). However, these approaches have its drawbacks as it doesn't give a global view of IS movements in the genome as some IS have specific hotspots which could be located outside the target loci and there are different IS present on the genome. The second drawback is that this approach does not quantify the link between natural selection and IS activity as the bacteria are not under a well-defined selective

pressure. The second method used to quantify the involvement of IS in bacterial mutagenesis involves looking at the distribution of IS elements in closely related bacteria through RFLP or genome sequencing. This approach also has various limitations as it doesn't allow the analysis of the dynamics of IS movements throughout its evolution (i.e. did IS distribution occur as a burst or as a stochastic method?) as all that is analysed is essentially bacteria that have diverged a long (evolutionary) time ago. To overcome this one evolution experiment involving 50 replicates of *E.coli* were analysed for IS movements over 1610 generations IS transposition rates were estimated at  $1.15 \times 10^{-5}$  per element per generation, unfortunately this approach is limited to a small timescale and doesn't have the resolution over a longer period of time (Sousa, Bourgard et al. 2013).

#### IV. The Long Term Evolution Experiment (LTEE)

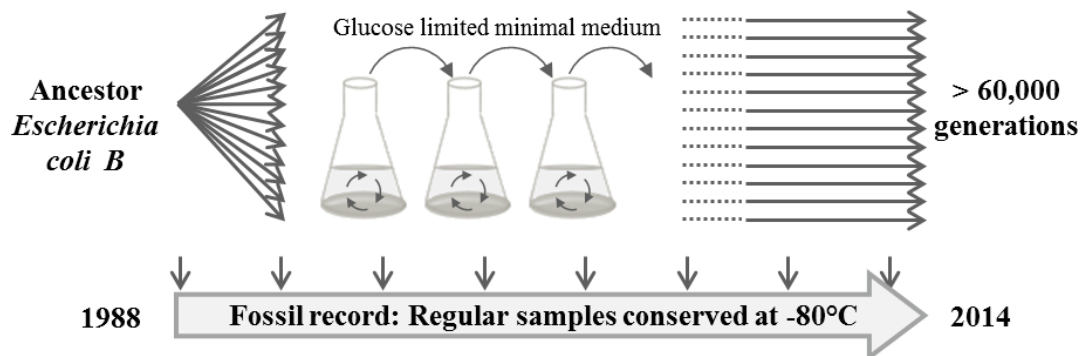


Figure 7: **Method for the Long Term Evolution Experiment (LTEE).** Twelve independent cultures founded from a common ancestor have been propagated by daily serial transfer in glucose limited minimal media (Davis media). Over 60,000 generations have been obtained at this day and regular samples have been taken every 500 – 1000 generations and conserved at -80°C.

##### Experiment set-up

The longest ongoing evolution experiment was initiated by Richard Lenski in 1988. Twelve populations were founded from an *E.coli* B strain acting as a common ancestor (Lenski 2014). They are daily propagated after a 1/100 dilution in 10 ml Davies Minimal media supplemented with 25mg/L of glucose creating an environment where cells alternate between

glucose abundance to glucose starvation and are selected through a bottleneck every 24 hours. There is roughly 6.64 generations per 24 hour cycle with the final cell count arriving at  $\sim 5 \times 10^7$  cells per ml (Lenski 1991) from  $10^5$  cells per ml with the ancestor. Arabinose is used as a marker for competition experiments and is shown to be neutral under the evolution experiment conditions (Lenski 1991). Every 500 generations a sample is taken from each population and is frozen. The frozen sample can be later revived and therefore constitutes a frozen fossil record of the evolving populations, allowing for comparison between the ancestor and evolved clones. The ancestor is not naturally competent and is asexual; therefore any evolution that occurs is due to genetic drift and natural selection (Lenski 1991; Lenski 2004). The populations were named Ara+1 to Ara+6 and Ara-1 to Ara-6, alluding to their capacity to utilise arabinose as a carbon source for growth. The populations have now been evolving for over 60,000 generations in these conditions

### **Common evolution between the populations**

The twelve populations rapidly adapted to their environments, showing an increase in fitness compared to their ancestor (Cooper and Lenski 2000; Wisser, Ribbeck et al. 2013). The fitness increased by almost 50% in the first 5000 generations and then the increase was less marked, increasing by around 25% more over the following 15,000 generations (Cooper and Lenski 2000; Philippe, Crozat et al. 2007) (Fig 8).

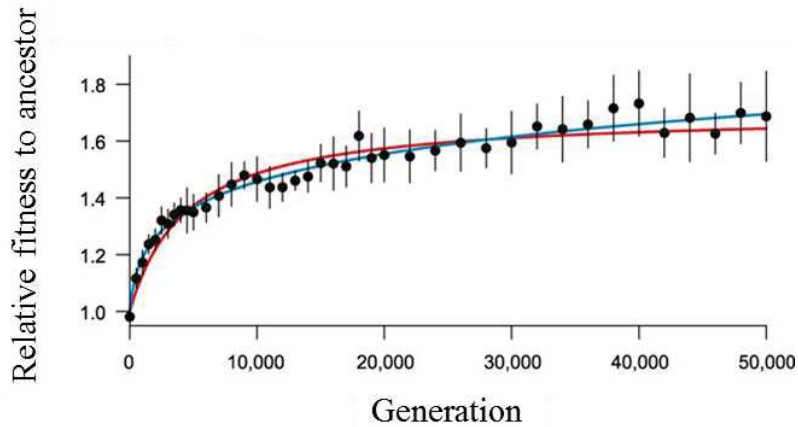


Figure 8: **Average fitness distribution of the 12 evolving populations.** Hyperbolic (red) and power-law (blue) models fit to the set of mean fitness values (black symbols) from all 12 populations (adapted from Wisner, Ribeck et al. 2013).

Not only did they increase in fitness in relations to the media, they also underwent many other phenotypic changes. They all had an increase in cell size, different cell morphology, a decrease in lag phase and an increase in maximal growth rates (Lenski and Travisano 1994; Lenski 2004; Philippe, Crozat et al. 2007; Philippe, Pelosi et al. 2009). They all presented an increase in their capacity to utilise glucose and lost the capability to grow on certain carbon sources, most notably, all twelve populations had a deletion in the ribose operon (Cooper, Schneider et al. 2001) and they all presented deficiencies in the operon for the use of maltose (Pelosi, Kuhn et al. 2006).

### **Genetic analysis of the twelve populations.**

The twelve populations show large changes in the regulation networks (Hindre, Knibbe et al. 2012). The rewiring of the global regulatory networks was mainly due to mutations in *spoT* in all twelve populations (Cooper, Rozen et al. 2003). This was not the only mutation common to many populations. Mutations in *fis*, *dusB* and *topA* were found to increase the levels of supercoiling of the DNA of the majority of the twelve populations (Crozat, Philippe et al. 2005; Crozat, Winkworth et al. 2010). Two other genes, *pykF* and *nadR* were found to be mutated in every population. It is noteworthy that the mutations were not identical and could

affect different parts of the gene (Woods, Schneider et al. 2006). This high level of parallelism for these genes between the populations indicates that they are part of the evolution and adaptation to the evolution experiment conditions. This is confirmed when the mutations were reconstructed in ancestral genetic backgrounds, conferring to them an advantage. Analysis of early mutations in one population shows that there is a high level of negative epistasis between positive mutations with the more beneficial the mutation, the stronger the effect of negative epistasis (Khan, Dinh et al. 2011). The only parallel rearrangement that has been shown to be beneficial is the deletion of the ribose operon in all twelve populations. The deletion occurred between an ancestral *IS150* copy and a new *IS150* copy that inserted into the operon downstream. In all twelve populations the new *IS150* copy inserted into different positions but they were all within the operon (Cooper, Schneider et al. 2001). However there are many more evolved alleles that are common in several populations that have either no detectable or even slightly deleterious effects on fitness (Croizat, Winkworth et al. 2010).

#### **Diversity between the populations.**

Even though there is a high level of parallelism between the populations each population has evolved different characteristics. Strikingly, there are 6 populations that have evolved higher mutation rates than the ancestor through mutations in their DNA repair mechanisms (Sniegowski, Gerrish et al. 1997; Wielgoss, Barrick et al. 2013). In the population Ara-1 a mutations in *mutT* which increase mutation rates by ~150 fold (Wielgoss, Barrick et al. 2013) before a subsequent mutation in *mutY* that reduces the mutation rate by ~40-60% (Wielgoss, Barrick et al. 2013). High levels of IS movements have also been observed for one population (Ara+1) in particular, mimicking higher SNP rates (Papadopoulos, Schneider et al. 1999; Schneider, Duperchy et al. 2000). Several populations were also analysed for large chromosomal rearrangements at the same time as being analysed for IS elements (Schneider,

Duperchy et al. 2000). At 2,000 generations one clone in the population Ara+1 had an inversion of over 1/3 of its genome. This clone is then established in the population. Interestingly, this is the population where there is most movement of IS elements. In Ara-1, analysis of the distribution of IS elements and whole-genome sequencing detected four large deletions with sizes ranging from ~8 to ~23 kbp and an inversion of one-third of the chromosome that is different from the Ara+1 inversion (Schneider, Duperchy et al. 2000; Barrick, Yu et al. 2009) (Fig 9). The effect of rearrangements in the populations is difficult to estimate as there are no isogenic clones without the rearrangement and to reconstruct any rearrangement would necessitate complex genomic manipulation with a strong possibility of introducing new mutations. At the most it is possible to estimate that a given rearrangement is beneficial through parallelism (Woods, Schneider et al. 2006). So while the populations have been subject to parallel evolution, certain populations have taken unique evolutionary trajectories. This is especially true for two populations, Ara-2 and Ara-3.

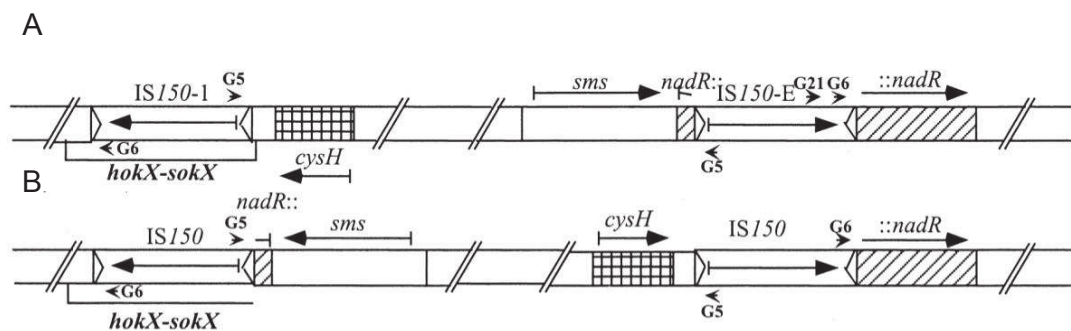


Figure 9: **Insertion of an IS150 element into *nadR* in population Ara + 1, and a subsequent inversion involving that new element.** Example of an inversion detected in the evolving populations. A. A new IS150 copy inserted into the gene *nadR*. B. subsequent rearrangement between newly inserted IS and an ancestral IS located in the intergenic region of *hokX-sokX* (adapted from Schneider, Duperchy et al. 2000).

### Evolution of polymorphism in Ara-2

The population Ara-2 evolved along a unique evolutionary path. The population split in to two ecotypes called S and L, one (S) characterised by small colony formers and small cells,

the other (L) by large colonies and large cell sizes (Rozen and Lenski 2000; Rozen, Schneider et al. 2005). This was unexpected because the experiment was designed to minimise ecological niches by firstly limiting metabolite production by glucose utilisation (principally acetate) and secondly by limitation of cell numbers. First detected at 6500 generations, the clones co-exist in a dynamic relationship for over 50,000 generations and are still present today (Rozen and Lenski 2000; Le Gac, Plucain et al. 2012; Plucain, Hindre et al. 2014). Not only were they present in co-existence, based on IS profiles, there was no genetic exchange between the two populations, which based on classical markers indicates that they two separate populations (Rozen, Schneider et al. 2005). The emergence of the populations can be traced back to essentially 3 mutations. Before the divergence the gene *spoT* is mutated and is found in both sub-populations. Mutations in *arcA* and *gntR* then occurred in the S lineage but are not present in the L. These three mutations together are responsible for S phenotypes and the capabilities for S to maintain itself in early emergence (Plucain, Hindre et al. 2014) (Fig 10).

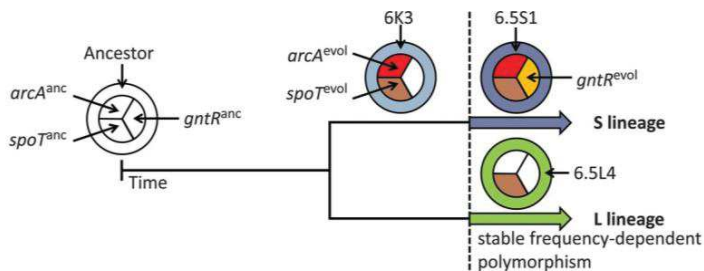


Figure 10: **Divergence of the population Ara-2.** The population Ara-2 diverged into two co-existing ecotypes after 6,000 generations. Three mutations, in *arcA*, *spoT*, and *gntR* have been shown to cause the S phenotype. Evolved genotypes are shown in circles (adapted from Plucain, Hindre et al. 2014).

### Evolution of a new catabolic pathway Ara-3

The population Ara-3 presents a very special case of evolution. Historically the ancestor is not capable of internalising citrate and therefore using it as a carbon source for growth (Koser



1924). However in the evolution media there is citrate present along with the glucose, the principal carbon source. There exist several cases where citrate utilisation (*cit+*) in *E. coli* is acquired through either horizontal transfer of genes or several mutations to a residual citrate transport system (Ishiguro, Oka et al. 1979; Hall 1982). At around 33,000 generations clones from the population Ara-3 had evolved to utilise the citrate as a carbon source (Blount, Borland et al. 2008).

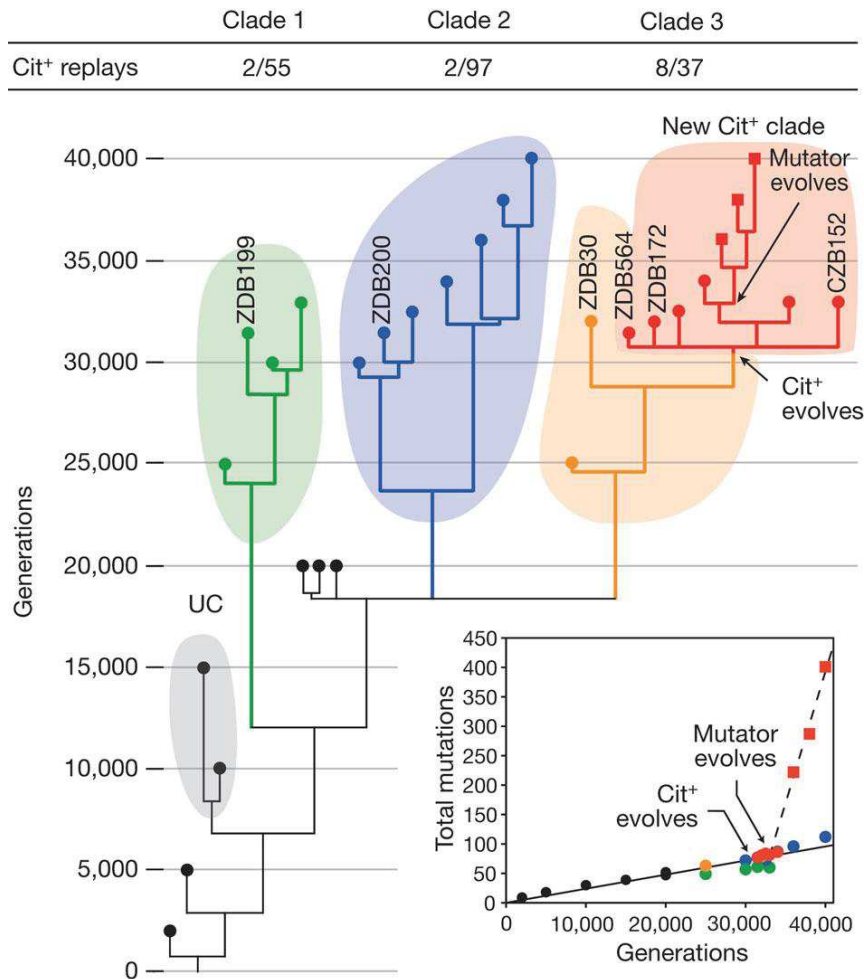


Figure 11: **Phylogeny pf Ara-3.** Symbols at branch tips mark 29 sequenced clones; labels are shown for clones mentioned in main text and figures. Shaded areas and coloured symbols identify major clades. Fractions above the tree show the number of clones belonging to the clade that yielded Cit<sup>+</sup> mutants during replay experiments (numerator) and the corresponding total used in those experiments (denominator). Inset shows number of mutations relative to the ancestor. The solid line is the least-squares linear regression of mutations in non-mutator genomes; the dashed line is the corresponding regression for mutator genomes. (adapted from Blount, Z. D., J. E. Barrick, et al. (2012))

This has an important fitness increase for the clone as it is able to achieve higher cell numbers in the population. The appearance of this trait is highly dependent on the genetic background of the clone. When evolution was replayed, clones from later generations were more likely to produce *cit*<sup>+</sup> progeny than the ancestor, earlier clones or contemporary clones sampled from other populations. In effect, the genetic backgrounds of the evolved clones potentiated the clone for the evolution of citrate utilisation. When genome sequences were available for the population, the genetic history was exposed (Fig 11). There were in fact 3 potentiated clades that coexisted over 10,000 generations that were capable in producing *cit*<sup>+</sup> progeny in replay experiments. The clade that is capable of using citrate had a tandem duplication event of the gene *citT*, the principal gene for citrate transport (Blount, Z. D., J. E. Barrick, et al. 2012). The duplication event placed the gene behind an aerobically expressed promoter (Fig 12). This allowed for the transport protein to be expressed in the experimental conditions. The first clone sequenced with the duplication is in fact at 31,500 generations and has a poor utilisation of citrate that would represent an actualisation step. Further modifications to regulatory networks explain the increased use of citrate. To confirm the potentiated nature of the clone the evolution experiment was replayed and citrate<sup>+</sup> clones were analysed to see if by what mechanism they had evolved to utilise citrate. Out of 19 re-evolved clones only 8 had new duplications. They had the same effect as the first duplication, placing the citrate transport genes behind a strong promoter. In six clones an IS3 inserted just before the gene. IS3 has an outward facing promoter sequence and was thus able to jump start gene expression (Charlier, Piette et al. 1982). Two mutants have large duplications encompassing all or part of the *cit* operon. One mutant has a large inversion that places most of that operon downstream of the promoter for the fimbria regulatory gene *fimB*, and another has a deletion in *citG* that presumably formed a new promoter. Also, most of these mutants have stronger phenotypes than the earliest Cit<sup>+</sup> clones in the main experiment.

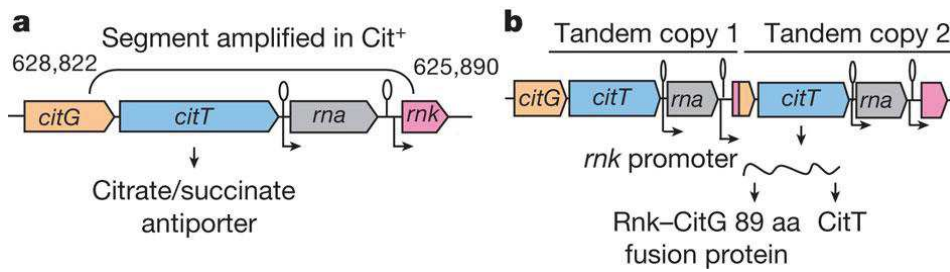


Figure 12: **Duplication of citrate transport genes in Ara-3.** Duplication bordered by IS3 explains the evolution of citrate utilisation in the population Ara-3. The genes *citT* and *rna* and a part of *citG* are duplicated behind a strong promoter of *rnk*. a: the ancestral genome b: Evolved genome (adapted from Blount, Z. D., J. E. Barrick, et al. 2012).

As such we know of only very little of the rearrangements present during the evolution experiment. We know of only examples that have had a dramatic consequence in the fitness of populations, the deletion of the ribose operon and the duplication of *citT*. The only general study was done on the populations Ara+1 and Ara-1 up to 20,000 generations. Even though there are a large number of sequences available at different time points for the twelve populations, chromosomal rearrangements such as inversion and duplications could be missed due to sequencing nature. In this study I will identify large chromosomal rearrangements that have evolved and also characterise a duplication present in several populations. The second part of my thesis I will investigate the role of IS elements in the populations, with particular focus on the population Ara+1 which presented an unusually high number of IS elements at 20,000 generations. Together these two aspects will help further characterise the populations in understudied areas.

# Results

# Results Part I

**Article 1:** Large Chromosomal Rearrangements during  
Long-Term Evolution in *Escherichia coli*

**State:** Under revision in *mBio*

Large Chromosomal Rearrangements during a Long-Term Evolution Experiment with  
*Escherichia coli*

Colin Raeside,<sup>a,b</sup> Joël Gaffé,<sup>a,b</sup> Daniel E. Deatherage,<sup>c</sup> Olivier Tenaillon,<sup>d,e</sup> Adam M. Briska,<sup>ff</sup>  
Ryan N. Ptashkin,<sup>fx</sup> Stéphane Cruveiller,<sup>g,h</sup> Claudine Médigue,<sup>g,h</sup> Richard E. Lenski,<sup>i,j</sup> Jeffrey  
E. Barrick,<sup>c,j</sup> Dominique Schneider<sup>a,b\*</sup>

Univ. Grenoble Alpes, Laboratoire Adaptation et Pathogénie des Microorganismes (LAPM),  
F-38000 Grenoble, France<sup>a</sup>; Centre National de la Recherche Scientifique (CNRS), LAPM,  
F-38000 Grenoble, France<sup>b</sup>; Department of Molecular Biosciences, Institute for Cellular and  
Molecular Biology, Center for Systems and Synthetic Biology, The University of Texas at  
Austin, Austin, TX 78712<sup>c</sup>; IAME, UMR 1137, INSERM, F-75018 Paris, France<sup>d</sup>; IAME,  
UMR 1137, Université Paris Diderot, Sorbonne Paris Cité, F-75018 Paris, France<sup>e</sup>; OpGen,  
Inc., Gaithersburg, MD 20878<sup>f</sup>; Direction des Sciences du Vivant, Commissariat à l'Energie  
Atomique et aux Energies Alternatives (CEA), Institut de Génomique, Genoscope & CNRS-  
UMR8030, Evry, France<sup>g</sup>; Laboratoire d'Analyses Bioinformatiques en Génomique et  
Métabolisme (LABGeM), Evry, France<sup>h</sup>; Department of Microbiology and Molecular  
Genetics, Michigan State University, East Lansing, MI 48824, USA<sup>i</sup>; BEACON Center for  
the Study of Evolution in Action, Michigan State University, East Lansing, MI 48824, USA<sup>j</sup>

Running Head: Rearrangements during experimental evolution

<sup>†</sup>Present address: DNASTAR, Inc., Madison, WI 53705, USA.

‡Present address: Department of Computational Medicine & Bioinformatics, University of Michigan Medical School, Ann Arbor, MI 48109, USA.

\*Address correspondence to Dominique Schneider, dominique.schneider@ujf-grenoble.fr.

**Word count for abstract: 243 words**

**Word count for text: 7062 words.**

**ABSTRACT** Large-scale DNA rearrangements may be important drivers of evolution because they can alter chromosome organization and gene expression in ways not possible through point mutations. In a long-term evolution experiment, twelve *E. coli* populations have been propagated in a glucose-limited environment for over 25 years. We used Whole Genome Mapping<sup>TM</sup> (*i.e.* optical mapping) combined with genome sequencing and PCR analysis to identify the large-scale chromosomal rearrangements in clones from each population after 40,000 generations. A total of 110 independent rearrangement events were detected, including 82 deletions, 19 inversions and 9 duplications, with different lineages having between 5 and 20 such events. In three populations, successive rearrangements affected particular genomic regions. In five populations, the rearrangements affected over a third of the chromosome. Most rearrangements (~70%) involved recombination between IS elements, illustrating their importance in mediating genome plasticity. Two lines of evidence suggest that many of these rearrangements may confer higher fitness. First, parallel changes were observed across the independently evolving populations, with ~65% of the rearrangements affecting the same loci in at least two populations. For example, the ribose-utilization operon and the *manB-cpsG* region were deleted in 12 and 10 populations, respectively, suggesting positive selection, and this inference was previously confirmed for the former case. Second, optical maps from additional clones sampled over time from one population showed that most rearrangements occurred early in the experiment, when fitness was increasing most rapidly. Therefore, large-scale genomic rearrangements evidently affected evolutionary outcomes in these populations.

**IMPORTANCE** Bacterial chromosomes are dynamic structures shaped by long histories of evolution. Among genomic changes, large-scale DNA rearrangements can have large effects on the presence, order, and expression of genes. Whole-genome sequencing that relies on



short DNA reads cannot identify all large-scale rearrangements. Therefore, deciphering changes in the overall organization of genomes requires the use of alternative methods, such as optical mapping. We analyzed the longest-running microbial evolution experiment (more than 25 years of evolution in the laboratory) by optical mapping, genome sequencing, and PCR analyses. We found multiple large genome rearrangements in all 12 independently evolving populations. In most cases, it is unclear whether these changes were beneficial themselves or, alternatively, they hitchhiked to fixation with other beneficial mutations. In either case, many genome rearrangements accumulated over decades of evolution, providing these populations with genetic plasticity reminiscent of that observed in some pathogenic bacteria.

## Introduction

Large-scale chromosomal rearrangements have played important roles in long-term organismal evolution (1, 2), including in processes of speciation (3, 4) and genome reduction (5). On shorter time scales, even single rearrangement events such as duplications, amplifications, inversions, deletions and translocations can have profound effects on organismal phenotypes, typically by altering gene regulation or disrupting genes. In bacteria, some genome rearrangements have led to traits important for virulence (6), and rearrangements are sometimes even developmentally regulated (7).

Large-scale rearrangements have been identified in diverse bacteria—Gram-negative, Gram-positive, pathogenic, and non-pathogenic—including *Escherichia coli* (8), *Salmonella typhi*, *Yersinia pestis*, *Helicobacter pylori*, *Mycobacterium leprae* (9, 10), *Pseudomonas stutzeri* (11), *Pseudomonas aeruginosa* (12), *Francisella tularensis* (13), *Neisseria gonorrhoeae* (14), *Lactococcus lactis* (15), and *Staphylococcus aureus* (16). Besides their effects on chromosome structure, bacterial DNA rearrangements may cause phenotypic changes by the incorporation of foreign DNA into host genomes through horizontal gene transfer (17, 18), by changes in gene expression (19), or by genome reduction through the loss of non-essential genes (20). The major mechanisms producing chromosomal rearrangements are recombinational exchanges between homologous sequences including ribosomal operons (21) as well as mobile genetic elements such as transposons (16), insertion sequence (IS) elements (15), and prophages (8). Comparisons of related genomes often reveal numerous DNA inversions among other rearrangements (9, 22, 23).

Some constraints may influence the occurrence of DNA rearrangements. First, selection may preserve symmetry in the size of the two replichores of a circular chromosome between the origin and terminus of replication (9, 24). This selection may help explain the

strong conservation of gene order (synteny) between *E. coli* and *Salmonella*, although large inversions have been observed under laboratory conditions where such selection may be relaxed. Second, the structural organization of the *E. coli* chromosome can affect rearrangements (25). Specifically, the genome is organized into distinct macrodomains (26), and rearrangements affecting the replication origin or terminus domain and inversions between the left and right macrodomains have been shown to be detrimental owing to their effects on replication-fork progression (25).

Large-scale rearrangements occur spontaneously at measurable frequencies, although the rates at which they occur are uncertain. In an older study, duplications were reported to arise at frequencies between  $10^{-2}$  and  $10^{-5}$  per cell division, depending on their chromosomal location (21). A more recent whole-genome sequencing study (27) of a population of *Salmonella enterica* var. Typhimurium that was propagated in a chemostat found duplications, inversions, and small deletions in  $\geq 20\%$  of the cells after only 50 generations. There are several possible explanations for these differences. First, it is generally difficult to disentangle underlying mutation rates from the effects of selection that may cause some mutants to replicate faster or slower than non-mutant cells. Second, the earlier study involved plating to isolate clonal genotypes, which may induce the loss of unstable rearrangements including duplications, thereby underestimating their true frequency (28). Third, earlier studies were based on observable phenotypes that were produced after the rearrangements occurred and might have missed many other events (29–32).

Chromosomal rearrangements have been discovered in many evolution experiments in which bacterial populations are propagated under various laboratory conditions. Deletions, duplications, and large-scale inversions have been detected in *E. coli* populations propagated in batch (33–35) and chemostat (28, 36) cultures as well as under stressful conditions (37). Rearrangements have also been found to occur in *P. aeruginosa* populations evolving in

cystic fibrosis patients (38). In some cases, specific rearrangements have been shown to confer increased fitness or phenotypic innovations in these evolution studies (28, 34, 35, 37). To date, however, no study has attempted to provide an exhaustive analysis of the multiple, large-scale chromosomal rearrangements that have arisen during a long evolution experiment.

Rearrangements have usually been investigated using methods such as pulsed-field gel electrophoresis, microarray-based hybridization experiments, and whole-genome sequencing. Despite the name, even “whole-genome sequencing” typically relies on sequencing short DNA fragment libraries and therefore cannot detect certain large-scale rearrangements including inversions and other events involving long sequence repeats. To complement sequencing data, Whole Genome Mapping (hereafter, optical mapping) techniques have been developed that produce a high-resolution, physically ordered restriction map of bacterial genomes (39–41). We produced optical chromosomal maps for clones isolated from a long-term evolution experiment (LTEE) with *E. coli*. In this on-going experiment, 12 populations have been independently propagated from a common ancestor in the same glucose-limited minimal medium for more than 25 years and 50,000 cell generations. These evolving populations have adapted to the experimental environment and have increased in competitive fitness relative to the ancestor by more than 70%, on average (42).

Some chromosomal rearrangements have been previously detected in the LTEE populations by using other techniques. However, with the exception of deletions involving the ribose operon found in all 12 populations (34), few rearrangements have been analyzed in detail. The ribose deletions were shown to occur at a high rate as well as to confer a slight fitness benefit in the LTEE conditions. In addition, other rearrangements have been detected in three populations (designated Ara+1, Ara-1 and Ara-3). In Ara+1, an analysis of the distribution of IS elements by Southern blotting revealed an inversion of about one-third of

the chromosome by recombination between two copies of *IS150* (33). A substantial increase in the *IS150* copy number also occurred in this population (43). In Ara-1, analysis of the distribution of IS elements and whole-genome sequencing found four large deletions (ranging from ~8 to ~23 kbp) and an inversion of one-third of the chromosome that is different from the inversion in Ara+1 (33, 44). The fitness consequences of these rearrangements are unknown. In Ara-3, numerous deletions, duplications and amplifications were detected in evolved clones by using genome sequencing data, including a specific tandem duplication and further amplification events involved in the production of a novel Cit<sup>+</sup> phenotype (35).

In an effort to obtain a more complete picture of the number and types of rearrangements that were substituted over time in all 12 LTEE populations, we combined optical mapping, genome sequencing, and targeted PCR and Sanger sequencing to analyze a total of 19 clones including a single clone sampled at 40,000 generations from each population as well as additional clones sampled at each of 7 other time points from population Ara-1. The resolution of optical maps cannot reliably detect rearrangements smaller than ~5 kbp (including, for example, new insertions of IS elements) unless they alter restriction sites. Nevertheless, we found that all 12 populations experienced large-scale chromosomal rearrangements and that most of these involved IS elements or other repeated sequences. Moreover, we saw many cases of parallel evolution across the populations in the genes that were affected by these rearrangements. Also, three populations had undergone complex rearrangements that involved successive inversion events.

## Results

**Chromosomal rearrangements in the twelve populations of the long-term evolution experiment.** We combined optical mapping and genome sequence analyses to identify the precise location and borders of all large-scale chromosomal rearrangements that occurred in one clone sampled at 40,000 generations from each of the twelve *E. coli* populations of the LTEE (Fig.1, Table S1). Combining these approaches allowed us to resolve rearrangements between large repeated elements, which is difficult or impossible with genome sequencing data alone, and to map the borders of the rearrangements with single-nucleotide resolution, which is impossible with optical mapping data alone. We also verified the rearrangement borders for two evolved clones from populations Ara+1 and Ara+2 in which we detected more complex rearrangements using PCR and Sanger sequencing. Primer pairs were designed adjacent to repeat sequences including IS elements and rRNA-encoding genes for these assays. The results agreed with our predictions in all cases, giving us confidence that the events we inferred in other clones are also accurate. Note, however, that optical mapping cannot detect IS insertion events because most rearrangements smaller than ~5 kbp are too small to resolve unless they affect restriction sites.

We identified a total of 110 rearrangement events in the twelve 40,000-generation clones, including 82 deletions, 19 inversions, and 9 duplications (Fig. 1, Table S1). Among the inversions, nine were involved in successive series of events that occurred over time in three populations (see next section). Among the duplications, three apparently involved successive events in which a typical tandem duplication was followed by deletion of the junction between the two duplicated copies, thereby resulting in an imperfect duplication (see next section).

Large deletions were the most frequent type of rearrangement, and they were found in all twelve populations, ranging in size up to ~55 kbp. Prophage remnants were often affected

by these deletions; 30 of the 82 large deletions (36.6%) resulted in the loss of prophage DNA, although these regions cover only ~4% of the ancestral genome. This over-representation of prophage DNA is highly significant (binomial test  $P < 2.2e-16$ ). The 19 inversions were found in nine populations, ranging in size from ~164 kbp to ~1.8 Mbp (see Fig. 1 legend for explanation of inversion sizes). In seven cases (one each in populations Ara+1, Ara-1, and Ara-5, and two each in Ara+2 and Ara-3), more than a quarter of the chromosome was affected by the inversions (Fig. 1). Successive inversions were inferred in some cases, and they were confirmed by examining multiple evolved clones from different generations (see next section). Nine duplications, ranging in size from ~3 kbp to ~180 kbp, were found in clones from four populations, including three with further deletions of the copy junctions (see next section).

The total number of rearrangements after 40,000 generations ranged from 5 in population Ara+3 to 20 in Ara-3, with an average of 9 rearrangements per population (Fig. S1). By that time, six of the 12 populations (Ara+3, Ara+6, Ara-1, Ara-2, Ara-3, and Ara-4) had evolved hypermutable phenotypes that caused greatly elevated point mutation rates compared to the ancestor (42). We did not characterize the chronology of the rearrangements in relationship to the hypermutator status of each population, but there is no suggestion of any difference in the overall rate at which rearrangements accumulated in mutator versus non-mutator populations (Poisson test  $P = 1$ ).

Most rearrangements occurred by recombination between repeated sequences, including 76 between homologous IS copies, 7 between the *manB* and *cpsG* genes (which share 96% sequence identity), and 5 between rRNA-encoding operons (Table S1, Fig. S1). The other 22 rearrangements occurred by unknown mechanisms not involving any repeated sequences, and 11 of these changes resulted in the loss of prophage remnants. Thus, IS elements were the main drivers of large-scale chromosomal rearrangements in these

populations. In fact, IS-mediated events accounted for at least half of the rearrangements in the 40,000-generation clones from every population, including all of them in Ara+1 (Fig. S1). More generally, there were no obvious differences in the distribution of rearrangement types among the twelve populations.

**Complex rearrangements.** In several cases, we observed genomic rearrangements that appear to have involved multiple successive events. A total of nine inversion events could have generated the complex rearrangements seen in the 40,000-generation clones from populations Ara+1, Ara-3 and Ara-6. A total of three duplications in the clones from populations Ara+2 and Ara+5 were imperfect, with the junctures between the duplicate copies apparently having been deleted following the duplication events (Fig. 1). To evaluate these hypotheses more thoroughly, we analyzed the genome sequences of the corresponding regions in these clones as well as other evolved clones sampled from earlier generations, with a particular focus on the rearrangements observed in population Ara+1.

Optical mapping suggested that the complex rearrangements in population Ara+1 resulted from four successive inversions (Fig. 2). We used PCR to analyze an additional 18 clones from this population including 3 clones sampled at each of 2000, 15,000, 20,000, 30,000, and 35,000 generations, 1 clone sampled at 25,000 generations, and 2 clones (in addition to the one used in the original analysis) sampled at 40,000 generations (Table 1, Table S2). We PCR-amplified each locus affected by each inversion using primer pairs designed for the two borders of the inverted region, both before and after each inversion (Fig. 2, Table S3). We then scored the presence or absence of each inversion in the ancestor and evolved clones based on the presence or absence of the expected PCR products (Fig. 2, Table 1, Table S3). We also sequenced the PCR products to confirm the expected rearrangement boundaries, and these data supported the scenario of four successive inversions, which we will call Inversions 1 through 4 in order of their appearance. Inversion 1 was present in 1 of 3



clones from 2000 generations, in 2 of 3 clones from 15,000 generations, and in all sampled clones from later generations. This inversion was previously reported in a study of changes in the distribution of IS elements in some LTEE populations (33). Inversion 2 extends for ~600 kbp and overlaps ~500 kbp of Inversion 1. Inversion 2 was seen in all of the clones sampled at generations 20,000, 25,000, and 30,000, and in one of the 3 clones from generation 40,000. All three clones at 35,000 generations and the other two clones at 40,000 generations carried Inversion 1 together with a second rearrangement, likely also an inversion, different from Inversion 2. Thus, PCR assays using the primer pair +1Inv2RF and +1Inv2RR detected the right border of Inversion 2, but no product was obtained using the primer pair +1Inv2LF and +1Inv2LR designed to detect the left border of Inversion 2. None of the evolved clones that we tested had the first three inversions without having all four, with one clone each from generations 30,000 and 40,000 carrying all four inversions. Inversions 3 and 4 have sizes of ~470 and ~260 kbp, respectively. These data together support the scenario in Fig. 2 that outlines the likely chronology of the inversions in population Ara+1. In a similar vein, the optical maps of the clones sampled at 40,000 generations from populations Ara-3 and Ara-6 imply two and three successive inversions, respectively (Fig. S2).

Optical maps suggested three imperfect duplications in the 40,000-generation samples, one in the clone from population Ara+2 and two in the clone from Ara+5 (Fig. 1), with deletions of the conjoined regions between the duplicated copies. The sequence of that region in the Ara+2 clone showed the absence of the junction sequence between the two copies and the presence instead of an *IS1* element (Table S1), suggesting the scenario depicted in Fig. 3. Two copies of *IS1* likely inserted, in the same orientation, at the end and start of the first and second tandem copies of the duplication, respectively. A subsequent recombination event between these *IS1* elements resulted in the deletion of the intervening region, generating the junction seen with the imperfect duplication (Fig. 3, Table S1). The

two imperfect duplications in population Ara+5 seem to have occurred by the same mechanism, with deletions of the duplication junctions again being associated with a new *IS1* element (Table S1).

**Effects of rearrangements on genome size and structure.** We analyzed both the optical maps and the genome sequences to estimate the genome size of each evolved clone sampled at 40,000 generations. Genome size was reduced in 10 of the 12 clones by amounts ranging from 0.9% to 3.5% of the ancestral genome size (Fig. 1). However, the clones from populations Ara+4 and Ara+5 showed slight increases in genome size of 0.8% and 0.3%, respectively, that resulted from duplications, including one encompassing ~4% of the genome (~180 kbp) in Ara+4. The overall tendency toward reduced genome size reflected the fact that large deletions were much more common than large duplications.

Two types of structural constraints have been hypothesized to influence genome structure. One hypothesizes a requirement for symmetry between the origin and terminus of replication in a circular chromosome (9, 24), and the other is based on the organization of the chromosome into distinct macrodomains (25). Imbalances of less than ~10% in the lengths of the two replichores have been reported to have little or no effect on *E. coli* growth (25). By contrast, inversions that disrupt the replication terminus macrodomain such that the replication forks meet far from the ancestral replicore junction have negative effects on growth, as do inversions between the replication origin and right macrodomains. The inversion events that we identified in the LTEE populations affect the symmetry of the evolved genomes to various extents (Fig. 4A). Five populations (Ara+3, Ara+4, Ara+5, Ara+6, and Ara-6) show little change; four populations (Ara-2, Ara-3, Ara-4, and Ara-5) show moderate changes ranging from ~3 to ~5 min of the leading replication branch; and three populations (Ara+1, Ara+2, and Ara-1) show larger changes of ~8 to ~10 min. In Ara+2, one inversion affected both the replication terminus and left macrodomains, while

another affected the replication origin and left macrodomains (Fig. 4B). Inversions in Ara-1 and Ara-5 also affected the replication terminus and left macrodomains, while inversions in Ara-2 and Ara-4 affected the replication origin and the adjacent non-structured domains. The four successive inversions in Ara+1 had the most dramatic effects on macrodomain organization, spanning the right replication terminus and left macrodomains. All of the LTEE populations became much more fit than their common ancestor based on competition assays (42), and thus none of these inversions had any highly deleterious effects; however, we do not know whether the inversions were beneficial mutations or, alternatively, were selectively neutral or even weakly deleterious mutations that hitchhiked with other beneficial mutations.

**Parallel rearrangements across populations.** The optical maps revealed a high level of parallel evolution – that is, similar large-scale rearrangements – across the twelve populations (Fig. 1). We define parallel rearrangements as those chromosomal regions that were affected by the same type of rearrangement event (i.e., deletion, inversion, or duplication) in at least two populations. Based on this criterion, nine distinct chromosomal regions were repeatedly affected by deletions (numbered  $\Delta 1$  to  $\Delta 9$  from left to right in Fig. 1), three by inversions (called Intervals I1 to I3 in Fig. 1), and one by duplications (numbered D1 in Fig. 1).

Chromosomal region  $\Delta 8$  was deleted in all twelve populations (Fig. 1, Tables S1, Table S4), causing the loss of part or all of the *rbs* operon, which encodes proteins required for growth on ribose. These deletions have been described previously and shown to contribute a small but consistent fitness benefit in the glucose-limited environment of the LTEE (34). Region  $\Delta 4$ , which encompasses the DNA between the *manB* and *cpsG* genes, was deleted in the clones from ten populations (Fig. 1, Table S1, Table S4). In seven of them (Ara+3, Ara+4, Ara-1, Ara-2, Ara-3, Ara-4 and Ara-5), the deletions occurred by recombination directly between these two genes, which share 96% nucleotide identity,

resulting in the loss of all 21 intervening genes. In the three other populations (Ara+1, Ara+2, and Ara-6), the deletions occurred by recombination of *IS1* elements located between *manB* and *cpsG*. In Ara+1 and Ara-6, the deletions involved the ancestral *IS1* element present at genome position 2,034,326 and new *IS1* copies that had inserted between *manB* and *cpsG*, leading to the loss of 18 and 12 genes, respectively; in Ara+2, the deletion occurred between two new copies of *IS1* and caused 14 genes to be lost (Table S1, Table S4). Overall, a common set of 12 genes associated with O antigen biosynthesis was lost in all 10 populations affected by these deletions, and a set of six additional genes associated with colanic acid biosynthesis was eliminated in 8 of these populations.

Four of the other repeatedly deleted chromosomal regions  $\Delta 1$ ,  $\Delta 3$ ,  $\Delta 5$ , and  $\Delta 7$  contained prophage remnants (Fig. 1, Table S1, Table S4). The DLP12-like (region  $\Delta 1$ ) and prophage 2 (region  $\Delta 5$ ) loci were each lost in 10 populations, the Qin-like locus (region  $\Delta 3$ ) was deleted in 6 populations, and the CP-44-like locus (region  $\Delta 7$ ) was lost in 3 populations. Eleven populations lost either two or three of these prophage regions. Most of these deleted genes have unknown functions. In population Ara-4, the deletion including the DLP12-like prophage was larger and overlapped a deletion found in Ara-2 (region  $\Delta 2$  in Fig. 1, Table S4). Region  $\Delta 2$  is also affected in population Ara+1.

Region  $\Delta 2$  contains several genes that encode proteins involved with the production and regulation of enterobactin, an iron-scavenging siderophore; it was deleted in populations Ara+1, Ara-2, and Ara-4. Region  $\Delta 6$ , deleted in populations Ara+1, Ara+2, and Ara-6, spans 17 genes including 10 having unknown functions; the other 7 genes are annotated as phage proteins, which suggests that region  $\Delta 6$  also corresponds to a phage remnant, although it has not been annotated as such. Region  $\Delta 9$  was deleted in populations Ara-3 and Ara-6, resulting in the loss of the *hsdSM* genes that encode a type 1 restriction-modification complex.

For inversions, we view as parallel changes the chromosomal intervals containing genes that were inverted in at least two populations, regardless of the overall length of the inversions. Three intervals fulfilled this rule (Fig. 1, Table S1, Table S4): I1 in populations Ara+1, Ara+2, Ara-1, and Ara-5; I2 in Ara+2 and Ara-3; and I3 in Ara+2, Ara-2, Ara-4, and Ara-6. Within the I1 interval, a sub-region denoted as 1.1 was affected by smaller inversions in populations Ara+3, Ara-2, Ara-3, and Ara-4; this sub-region was thus inverted in eight of the twelve populations. Sub-region 1.1 contains 148 genes, and this number increases to 1366 over the entire I1 interval. The I2 and I3 intervals contain 345 and 670 genes, respectively. The inversions spanning I1 were mediated by *IS1* elements in Ara+2 and Ara-1 and by *IS150* elements in Ara+1 and Ara-5, while all of the inversions spanning sub-region 1.1 were mediated by *IS3* elements. The inversions spanning I2 were mediated by *IS150* elements in Ara-3 and by recombination between rRNA-encoding operons in Ara+2. All of the inversions spanning I3 were mediated by recombination between rRNA-encoding operons.

A single chromosomal region, D1, underwent parallel duplications in three populations, Ara+2, Ara+4, and Ara+5 (Fig. 1, Table S1, Table S4). These three duplications ranged in size from ~11 to ~60 kbp, but they share ~11 kbp and 14 genes (Table S1, Table S4). The *rpoS* gene that encodes an alternative sigma factor is present in all three duplications, as are *pcm* and *surE*, both essential for survival in stationary phase, and the *cysDNC* operon involved in sulfur metabolism.

**Temporal dynamics of rearrangements in population Ara-1.** We analyzed the optical maps and genome sequences of eight clones sampled from population Ara-1 at 2000, 5000, 10,000, 15,000, 20,000, 30,000, 40,000, and 50,000 generations. We detected 6 deletion events ranging from ~7 to ~23 kbp in size (Table 2), three of which went to fixation including one that arose before 2000 generations and two that occurred between generations

5000 and 10,000. The ~1.5 Mbp inversion that we detected in the 40,000-generation clone (Fig. 1, Table S1) occurred between 5000 and 10,000 generations and was then fixed in the population. Two translocation events were present in the 5000-generation clone but were not seen in any of the later samples. Of the five rearrangements detected at 50,000 generations, four were already present at 10,000 generations (Table 2). Assuming a uniform rate of  $5/50,000$ , one is unlikely to observe 4 or more mutations by generation 10,000 (one-tailed Poisson test,  $P = 0.019$ ), suggesting heterogeneity in the evolution of these large-scale rearrangements over time. This heterogeneity might reflect a change in the underlying mutational processes that generate these rearrangements. Alternatively, the rate of fitness improvement and the corresponding rate at which beneficial mutations went to fixation were much higher early in the experiment than later on (42, 44), and this difference may explain the greater number of rearrangements fixed in the early generations.

## Discussion

We combined optical mapping and genome sequencing to identify chromosomal rearrangements that occurred in each of twelve populations over the course of 40,000 generations of experimental evolution. We detected a total of 110 rearrangements, of which 75% were deletions, 17% were inversions, and 8% were duplications; the resolution of the optical mapping did not allow the identification of new IS insertion events. Some of the complex rearrangements were shown to involve a succession of events including multiple inversions as well as duplications followed by deletions overlapping the junction of the two copies. Most rearrangements (~70%) occurred by recombination between IS elements, and many chromosomal regions were repeatedly affected by similar rearrangements in two or more populations. In most populations, the overall chromosomal organization was maintained without a large imbalance of the symmetry between the origin and terminus of replication or

any major disruption of the chromosomal macrodomains. However, three populations evolved rather substantial asymmetry between the replication origin and terminus. The dynamics of the rearrangements over time were examined in one population, and they showed that most of the rearrangements were substituted in the early generations of the experiment, when the rate of fitness increase was fastest, suggesting that these rearrangements contributed to the genetic adaptation of these populations.

The twelve experimental populations have been evolving in and adapting to the same environment for tens of thousands of generations. Therefore, one might expect them to lose unused functions and evolve smaller genomes, as observed for bacteria adapting to stable environments, such as endosymbionts adapting to their hosts (45, 46). As predicted, deletions were indeed the predominant rearrangements detected in the LTEE. We also observed a high level of parallelism, with nine chromosomal regions deleted in at least two populations (Fig. 1, Table S4). These parallel deletions removed genes from the *rbs* operon, genes involved in O antigen and colanic acid biosynthesis, and prophage-related genes. In a previous study (34), we demonstrated that *rbs* deletions occurred at a very high frequency owing to the presence of an *IS150* element adjacent to the operon and that they conferred a small but significant fitness increase. More generally, the deleted genes have functions that are not used under the conditions prevailing during the LTEE. These deletions might have conferred higher fitness by eliminating unnecessary and costly gene expression (47), or they might have been effectively neutral if the affected genes were already not expressed. The involvement of IS elements in producing many of these deletions may reflect increased local mutagenesis caused by homologous recombination between two identical elements, similar to the process demonstrated for the *rbs* operon (34). Deletions have also been reported in evolution experiments with other bacteria and environments (27, 38, 48).

In the LTEE, the genome size of the *E. coli* declined after 40,000 generations in 10 of the 12 populations, with the reductions ranging from 0.9% to 3.5% relative to the ancestor; two populations showed slight increases in genome size of 0.3% and 0.8%. Over all 12 populations, we recorded 70 reductive events with sizes of 1 kbp or larger, with an average of ~17 kbp and a median of ~11 kbp. These values are far below the reductions inferred for pathogenic bacteria including *Mycobacterium leprae*, *Yersinia pestis*, and *Mycoplasma ulcerans* (49); of course, the time periods over which these pathogens have evolved their reduced genomes are much longer. On a time scale more commensurate with the LTEE, a study of *P. aeruginosa* adapting to the lungs of human patients with cystic fibrosis found that up to 8% of the ancestral genome was lost over the course of 35 years (38); in that study, the average and median deletion sizes were 44.5 kbp and 26.6 kbp, respectively, for 27 deletion events with sizes of at least 1 kbp. Most deletions in the *P. aeruginosa* study occurred through illegitimate and homologous recombination events, but IS elements were not involved. By contrast with that dramatic reduction, an analysis of 11 natural isolates of *E. coli* O157:H7 found only very limited genome reductions of up to 3.7 kbp, or ~0.1%, and slight increases in chromosome size were detected in several of those isolates (39). Note, however, that these estimates exclude the effect of new insertions of IS elements, which are not detected by optical mapping.

Three evolutionary factors have often been suggested as drivers of reduced genomes in pathogenic and endosymbiotic bacteria: severe population bottlenecks, the absence of horizontal gene transfer, and the elimination of selection for various functions owing to the availability of nutrients and other services provided by the host (5, 45, 46). Bottlenecks are not severe in the LTEE, with more than  $10^6$  cells transferred each day to fresh medium. There is no horizontal transfer in the LTEE, as plasmids and functional phage are absent and *E. coli* does not undergo natural transformation. The nutritional environment of the LTEE consists of



a minimal medium with glucose and ammonium providing carbon and nitrogen, respectively. Owing to this simple environment, certain functions cannot be lost including, for example, the production of amino-acids. However, some functions are dispensable including those involved with using alternative resources (e.g., the loss of the ability to grow on ribose) and those necessary to thrive in natural environments (e.g., loss of genes involved with O antigen and colanic acid biosynthesis). Thus, the simple flask environment – like a host organism – provides environmental constancy and protection that allow certain functions to be discarded. In doing so, the cells may save energy, thereby providing a competitive advantage; even without that benefit, any unused functions will tend to decay or be deleted by on-going mutations (20, 50, 51). Another factor that contributed to genome reductions in the LTEE is homologous recombination, especially that mediated by IS elements. The majority of rearrangements detected in this study involved IS elements, and these elements often flank non-core genes that were acquired by horizontal gene transfer in the distant past, *i.e.*, prior to *E. coli* B being brought into the laboratory (52). These horizontally acquired genes would thus be both dispensable and prone to deletion.

Besides deletions, we detected two other types of large-scale rearrangements, namely duplications and inversions. As for deletions, some of these other rearrangements affected the same chromosomal regions in multiple populations, with that parallel evolution providing indirect evidence that these mutations were beneficial in the context of the LTEE (34, 44, 53, 54). Duplications were rare, however, with only nine such events detected among the 12 populations after 40,000 generations. The paucity of duplications compared to deletions probably reflects the intrinsic instability of duplications, which readily collapse back to a single copy when cells are propagated under conditions that do not favor having multiple copies of the relevant genes (28). The nine duplications we detected range in size from ~3 to ~180 kbp. Owing to the number and diversity of the genes found in the duplicated regions, it

is difficult to know whether and how the duplications affected the fitness of the bacteria. However, one region was duplicated in three populations, and it spans an ~11-kbp region containing 14 genes including *rpoS*, which encodes the alternative sigma factor involved in the transition into stationary phase (55). Previous work has shown that the LTEE populations underwent changes in the regulatory networks involved in the transitions between exponential and stationary phases, which they experience on a daily basis (56, 57). The parallel duplications may thus affect the expression of *rpoS* and thereby confer a competitive advantage during these transitions.

Inversions were the second most common type of large-scale rearrangement seen in the LTEE populations. We detected a total of 19 inversions, of which seven affected more than a quarter of the chromosome, and multiple successive inversions occurred in several populations. Three chromosomal regions were inverted in multiple populations, whereas only one region (spanning genome positions 4,453,625 to 146,102) was not affected by any of the inversions. Owing to the large number of genes in these inversions, it is difficult to predict their effects, if any, on the fitness or other phenotypes of the evolved cells. Chromosomal inversions have been found in natural isolates of many bacteria, including *E. coli* (39), *Staphylococcus aureus* (16, 40), *Enterococcus faecium* (58), *Francisella tularensis* (13), and *Bacillus anthracis* (59). Some of these inversions have been related to phenotypic changes, including colony morphology (16) and virulence (59), but in most cases their effects are unknown. Experiments have shown that some inversions adversely affect cell growth because they substantially reduce the symmetry between the origin and terminus of replication (24), or because they disrupt the overall organization of the chromosome into macrodomains including especially the structure of the terminus (25). The inversions and other rearrangements that we observed in the evolved genomes of the LTEE had variable effects on the symmetry between the origin and terminus of replication: five populations showed almost

no change in symmetry, four had increased asymmetry leading to imbalances of the replication arms of a few percent, and three evolved imbalances of ~8% to ~10% (following the numerical scheme for calculating imbalance used in ref. 25). None of the clones we studied, however, had imbalances as great as the 15% imbalances previously shown to impair cell growth (25). In any case, the variations in chromosomal organization produced by the rearrangements we detected are clearly well tolerated under the conditions of the LTEE. The strong conservation of gene order between *E. coli* and *Salmonella*, which began to diverge from one another many tens of millions of years ago, has led to the suggestion that selection places important constraints on gene order (29). In that respect, the extent of inversions observed in the LTEE over two decades is surprising. One possible explanation for this discrepancy is that horizontal gene transfer—in particular, its importance for adaptation to changing environments—generates the constraint. In the LTEE, there is no gene transfer and the experimental environment does not change, relieving the constraint and thus allowing gene order to vary more freely.

An alternative hypothesis is that these inversions may confer higher fitness, perhaps by changing the distribution of genes on the leading and lagging strands. During replication, the DNA and RNA polymerase complexes move along the same DNA molecule, and their physical interactions depend on gene orientation (60). Collisions between DNA polymerase and RNA polymerase transcribing genes from the lagging strand occur with a higher probability, thereby causing the replication machinery to stall and potentially also generating truncated transcripts. By contrast, when genes are located on and transcribed from the leading strand, such collisions merely slow down the replication complex and the transcript is released after completion. Thus, essential (as well as highly expressed) genes tend to be located on the leading strand (61). The LTEE ancestral strain has 170 genes that are defined as essential and conserved in 80% of the sequenced bacterial genomes (62), and 74% of them

are indeed on the leading strand. Using the optical mapping data, we can determine whether the proportion of these genes located on the leading strand changed as a result of inversions in the 40,000-generation clones. Many inversions included either the origin or terminus of replication, and hence they did not affect whether genes were on the leading or lagging strand. However, inversions in three populations (Ara+1, Ara-3, and Ara-6) affected a total of 33 genes, with 27 moving from the leading to the lagging strand and only 6 moving from the lagging to the leading strand. Therefore, these results do not support the hypothesis that the inversions improved fitness by reducing collisions between the DNA and RNA polymerase complexes. However, this negative result may reflect the minimal medium used in the LTEE environment, which limits growth rates and might reduce the importance of these collisions. Also, we cannot exclude the possibility that particular genes might have been subject to this effect, as some essential genes did move from the lagging to the leading strand.

About 70% of the large-scale rearrangements we detected in the evolved clones occurred by homologous recombination between IS elements, and that proportion does not include new insertions of IS elements (because optical mapping cannot resolve such events). These elements have previously been shown to contribute to evolution in the LTEE in three ways. First, IS elements have generated some beneficial mutations including the deletions of the *rbs* operon (34). Second, some new IS insertions occurred in genes that were mutated in many or all of the LTEE populations (33, 54), and that genetic parallelism strongly suggests that these insertions were also beneficial. Third, population Ara+1 has undergone a striking increase in *IS150* copy number (43), including some insertions that can be inferred to be beneficial based on the previous criterion. Here, we have further shown that IS elements, by providing a substrate for homologous recombination, have played the major role in large-scale rearrangements that have restructured the genomes during this long-term experiment. IS elements have also been shown to contribute to genomic plasticity in other studies where they

were investigated, including both evolution experiments in the laboratory (36, 37, 63) and analyses of natural isolates (13, 64, 65).

In summary, we used optical mapping to find large-scale chromosomal rearrangements that occurred during a long-term evolution experiment with *E. coli*. The many rearrangements thus discovered had substantial effects on the size and structure of the chromosome, demonstrating the impressive plasticity of bacterial genomes. Several lines of evidence, including parallel changes observed in independently evolving populations, indicate that at least some of the rearrangements conferred higher fitness in the experimental environment. IS elements mediated most of the large-scale rearrangements by providing a substrate for recombination. While new sequencing technologies make it increasingly easy to find point mutations and other small changes in the genomes of experimentally evolving populations, our results demonstrate the value of also analyzing large-scale chromosomal rearrangements in these studies.

## **Materials and Methods**

**Bacterial strains.** All strains came from the *E. coli* long-term evolution experiment (42, 66). Twelve populations, named Ara+1 to Ara+6 and Ara−1 to Ara−6, were founded from the same ancestral strains, REL606 and REL607 (a spontaneous Ara<sup>+</sup> mutant of REL606). The populations have been propagated by daily transfers in Davis minimal medium containing 25 µg/ml glucose (DM25) as a limiting carbon source (66). Samples from each population have been taken at 500-generation intervals and stored at −80°C. For optical mapping, we used one clone isolated from each population at 40,000 generations, as well as one clone from population Ara−1 at each of 2000, 5000, 10,000, 15,000, 20,000, and 50,000 generations.

Additional clones were sampled at several time points from Ara+1 to investigate specific rearrangements. All strains used in this study are listed in Table S2.

**Optical Mapping.** The optical mapping procedure was performed by OpGen (Gaithersburg, Maryland), as described elsewhere (40). Clones were revived from stocks kept at  $-80^{\circ}\text{C}$  in 15% glycerol by overnight growth in LB medium. Genomic DNA was extracted using the OpGen Sample Preparation Kit (OpGen, Inc., MD) and Agencourt Genfind v2 Kit (Beckman Coulter, FL). Single DNA molecules were captured on an Argus<sup>®</sup> surface within a MapCard, digested with the *NcoI* restriction enzyme, and stained with JOJO-1 on the Argus MapCard Processor. They were analyzed by automated fluorescent microscopy using the Argus Optical Mapper. This software records the size and order of restriction fragments for each DNA molecule. Collections of single-molecule restriction maps for each genome were assembled according to overlapping fragment patterns to produce a whole-genome optical map assembly. The consensus optical map assemblies for each evolved clone were then compared to the predicted restriction map of the ancestral strain's genome (67) to identify large-scale rearrangements using the MapSolver software. Rearrangements smaller than  $\sim 5$  kbp are too small to resolve. Thus, IS insertion events were not detected in this study.

**Characterization of rearrangement borders.** The precise locations of the rearrangement borders were identified (Table S1) by analyzing the genome sequences of the evolved clones. The genomes of population Ara-1 clones from generations 2000 to 40,000 were previously sequenced (44, 68), as were those of the 40,000-generation clones from populations Ara-3, Ara-5, Ara-6, Ara+1, Ara+2, Ara+4 and Ara+5 (69). The additional genomes analyzed in this study were the 40,000-generation clones from populations Ara-2, Ara-4, Ara+3, and Ara+6, and they were sequenced on the Illumina Genome Analyzer platform at the Centre National de Séquençage, Genoscope, with one lane of single-end 36-bp reads per genome. Sequence reads were compared to the genome of the ancestral strain

REL606 (67), using both *breseq*, a computational pipeline for analyzing resequenced bacterial genomes (35, 44, 68), and a customized pipeline (70). The four new genome sequences have been deposited in the European Nucleotide Archive Sequence Read Archive (accession nos. XXXX). The borders of the rearrangements detected by optical mapping were further checked by PCR experiments for two populations, Ara+1 and Ara+2 (Table S1). PCR was performed using 1x reaction buffer, 3 mM MgCl<sub>2</sub>, 0.2 mM dNTP, 0.2 mM of each primer, 50 ng of genomic DNA, and 1.25 unit of Taq DNA polymerase (Invitrogen, Carlsbad, CA) in a 25- $\mu$ l reaction volume. Reaction mixtures were heated at 95°C for 2 min, then subjected to 32 cycles of 30 s at 94°C, 90 s at 55°C, and 3 min at 72°C, before a final step of 10 min at 72°C. Table S3 lists the primers used to amplify the borders of specific rearrangements. The PCR products were separated by agarose (0.8%) gel electrophoresis in 1x TAE buffer, purified using the Qiagen Gel purification kit, and sequenced (GATC-Biotech, Germany) with the same primers used for PCR assays.

**Sequence analysis.** Sequences of PCR products containing the borders of rearrangements were analyzed with BioEdit (version 7.0.9.0), and the resulting FASTA files were analyzed using the CLC Sequence Viewer software (v7.0.2, CLC Bio). All sequences were compared to the ancestral genome sequence (67) and checked to confirm the rearrangements deduced from the optical maps. A Python script was written to construct FASTA files containing the reconstructed genome sequences of the evolved clones; we have deposited the script at the Dryad Digital Repository (doi:XXXX). Point mutations and deletions detected in the evolved clones were automatically inserted; other rearrangements were added by hand in CLC Sequence Viewer. The leading strand branch length was calculated from the origin of replication *oriC* (71) to the middle of the terminus region defined by the *dif* locus (72). For the evolved clones, the branch-length measurements were based on the rearranged genomes using the new locations of these two loci.

## Supplemental Material

Supplemental material for this article may be found at [XXXX](#)

Figure S1 Number of total rearrangements in one clone sampled in each of the twelve evolving populations at 40,000 generations, Word file, 36 Ko.

Figure S2 Successive inversions in populations Ara-3 and Ara-6 at 40,000 generations, Word file, 260 Ko.

Table S1, Rearrangements detected in one clone sampled at 40,000 generations from each of the twelve populations of the LTEE, Word file, 54 Ko.

Table S2 Clones from the LTEE used in this study, Word file, 18 Ko.

Table S3 Primers used in this study, Word file, 22 Ko.

Table S4 Rearrangement parallelism during the LTEE, Word file, 38 Ko.

## Acknowledgments

This work was supported by the European Union program FP7-ICT-2013-10 project EvoEvo grant 610427 to D.S., the Université Grenoble Alpes to D.S., the Centre National de la Recherche Scientifique to D.S., the U.S. National Institutes of Health grant R00-GM087550 to J.E.B., the U.S. National Science Foundation grant DEB-1019989 to R.E.L., and the U.S. National Science Foundation support DBI-0939454 for the BEACON Center for the Study of Evolution in Action. C.R. was supported by a fellowship from the French Ministry of Education and Research. D.E.D was supported by a traineeship from the Cancer Research Prevention Institute of Texas.



## References

1. **Ohno S.** 1970. Evolution by gene duplication. Springer, New York.
2. **Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, Benito-Gutiérrez EL, Dubchak I, Garcia-Fernández J, Gibson-Brown JJ, Grigoriev IV, Horton AC, de Jong PJ, Jurka J, Kapitonov VV, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin-I T, Toyoda A, Bronner-Fraser M, Fujiyama A, Holland LZ, Holland PW, Satoh N, Rokhsar DS.** 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**:1064–1071.
3. **Delneri D, Colson I, Grammenoudi S, Roberts IN, Louis EJ, Oliver SG.** 2003. Engineering evolution to study speciation in yeasts. *Nature* **422**:68–72.
4. **Navarro A, Barton NH.** 2007. Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution* **57**:447–459.
5. **Sloan DB, Moran NA.** 2013. The evolution of genomic instability in the obligate endosymbionts of whiteflies. *Genome Biol. Evol.* **5**:783–793.
6. **Morrow CA, Lee IR, Chow EWL, Ormerod KL, Goldinger A, Byrnes III EJ, Nielsen K, Heitman J, Schirra HJ, Fraser JA.** 2012. A unique chromosomal rearrangement in the *Cryptococcus neoformans* var. *grubii* type strain enhances key phenotypes associated with virulence. *mBio* **3**:e00310–11.
7. **Abe K, Yoshinari A, Aoyaqi T, Hirota Y, Iwamoto K, Sato T.** 2013. Regulated DNA rearrangement during sporulation in *Bacillus weihenstephanensis* KBAB4. *Mol. Microbiol.* **90**:415–427.

8. **Iguchi A, Iyoda S, Terajima J, Watanabe H, Osawa R.** 2006. Spontaneous recombination between homologous prophage regions causes large-scale inversions within the *Escherichia coli* O157:H7 chromosome. *Gene* **372**:199–207.
9. **Eisen JA, Heidelberg JF, White O, Salzberg SL.** 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.* **1(6)**:research0011.1–0011.9.
10. **Darling AE, Miklos I, Ragan MA.** 2008. Dynamics of genome rearrangement in bacterial populations. *PLoS Genet.* **4(7)**:e1000128.
11. **Ginard M, Lalucat J, Tummeler B, Romling U.** 1997. Genome organization of *Pseudomonas stutzeri* and resulting taxonomic and evolutionary considerations. *Int. J. Syst. Bacteriol.* **47**:132–143.
12. **Kresse AU, Dinesh SD, Larbig K, Romling U.** 2003. Impact of large chromosomal inversions on the adaptation and evolution of *Pseudomonas aeruginosa* chronically colonizing cystic fibrosis lungs. *Mol. Microbiol.* **47**:145–158.
13. **Rohmer L, Fong C, Abmayr S, Wasnick M, Larson Freeman TJ, Radey M, Guina T, Svensson K, Hayden HS, Jacobs M, Gallagher LA, Manoil C, Ernst RK, Drees B, Buckley D, Haugen E, Bovee D, Zhou Y, Chang J, Levy R, Lim R, Gillett W, Guentherer D, Kang A, Shaffer SA, Taylor G, Chen J, Gallis B, D'Argenio DA, Forsman M, Olson MV, Goodlett DR, Kaul R, Miller SI, Brittnacher MJ.** 2007. Comparison of *Francisella tularensis* genomes reveals evolutionary events associated with the emergence of human pathogenic strains. *Genome Biol.* **8**:R102.
14. **Spencer-Smith R, Varke EM, Fielde MD, Snyder LA.** 2012. Sequence features contributing to chromosomal rearrangements in *Neisseria gonorrhoeae*. *PLoS One* **7**:e46023.

15. **Daveran-Mingot ML, Campo N, Ritzenthaler P, Le Bourgeois P.** 1997. A natural large chromosomal inversion in *Lactococcus lactis* is mediated by homologous recombination between two insertion sequences. *J. Bacteriol.* **180**:4834–4842.
16. **Cui L, Neoh HM, Iwamoto A, Hiramatsu K.** 2012. Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria. *Proc. Natl Acad. Sci. USA* **109**:1647–1656.
17. **Skippington E, Ragan MA.** 2011. Lateral genetic transfer and the construction of genetic exchange communities. *FEMS Microbiol. Rev.* **35**:707–735.
18. **Janssen PJ, Van Houdt R, Moors H, Monsieurs P, Morin N, Michaux A, Benotmane MA, Leys N, Vallaeyts T, Lapidus A, Monchy S, Médigue C, Taghavi S, McCorkle S, Dunn J, van der Lelie D, Mergeay M.** 2010. The complete genome sequence of *Cupriavidus metallidurans* strain CH34, a master survivalist in harsh and anthropogenic environments. *PLoS One* **5**:e10433.
19. **Kugelberg E, Kofoed E, Reams AB, Andersson DI, Roth JR.** 2006. Multiple pathways of selected gene amplification during adaptive mutation. *Proc. Natl Acad. Sci. USA* **103**:17319–17324.
20. **Lee MC, Marx CJ.** 2012. Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet.* **8**:e1002651.
21. **Anderson P, Roth J.** 1981. Spontaneous tandem genetic duplications in *Salmonella typhimurium* arise by unequal recombination between rRNA (*rrn*) cistrons. *Proc. Natl Acad. Sci. USA* **78**:3113–3117.
22. **Tillier ER, Collins RA.** 2000. Genome rearrangement by replication-directed translocation. *Nat. Genet.* **26**:195–197.
23. **Zivanovic Y, Lopez P, Philippe H, Forterre P.** 2002. *Pyrococcus* genome comparison evidences chromosome shuffling-driven evolution. *Nucleic Acids Res.* **30**:1902–1910.

24. **Roth JR, Benson N, Galitski T, Haack K, Lawrence JG, Miesel L.** 1996. Rearrangements of the bacterial chromosome: formation and applications, p 2256–2276. *In* Neidhardt FC (ed), *Escherichia coli* and *Salmonella*: cellular and molecular biology. ASM Press, Washington, DC.
25. **Esnault E, Valens M, Espéli O, Boccard F.** 2007. Chromosome structuring limits genome plasticity in *Escherichia coli*. *PLoS Genet.* **3(12):e226.**
26. **Boccard F, Esnault E, Valens M.** 2005. Spatial arrangement and macrodomain organization of bacterial chromosomes. *Mol. Microbiol.* **57:9–16.**
27. **Sun S, Ke R, Hugues D, Nilsson M, Andersson DI.** 2012. Genome-wide detection of spontaneous chromosomal rearrangements in bacteria. *PLoS One* **7(8):e42639.**
28. **Maharjan RP, Gaffé J, Plucain J, Schliep M, Wang L, Feng L, Tenailon O, Ferenci T, Schneider D.** 2013. A case of adaptation through a mutation in a tandem duplication during experimental evolution in *Escherichia coli*. *BMC Genomics* **14:441–452.**
29. **Segall A, Mahan MJ, Roth JR.** 1988. Rearrangement of the bacterial chromosome: forbidden inversions. *Science* **241:1314–1318.**
30. **Zieg J, Kushner SR.** 1977. Analysis of genetic recombination between two partially deleted lactose operons of *Escherichia coli* K-12. *J. Bacteriol.* **131:123–132.**
31. **Bierne H, Seigneur M, Ehrlich SD, Michel B.** 1997. *uvrD* mutations enhance tandem repeat deletion in the *Escherichia coli* chromosome via SOS induction of the RecF recombination pathway. *Mol. Microbiol.* **26:557–567.**
32. **Albertini AM, Hofer M, Calos MP, Miller JH.** 1982. On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. *Cell* **29:319–328.**

33. **Schneider D, Duperchy E, Coursange E, Lenski RE, Blot M.** 2000. Long-term experimental evolution in *Escherichia coli*. IX. Characterization of insertion-sequence mediated mutations and rearrangements. *Genetics* **156**:477–488.
34. **Cooper VS, Schneider D, Blot M, Lenski RE.** 2001. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J. Bacteriol.* **183**:2834–2841.
35. **Blount ZD, Barrick JE, Davidson CJ, Lenski RE.** 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* **489**:513–518.
36. **Gaffé J, McKenzie C, Maharjan RP, Coursange E, Ferenci T, Schneider D.** 2011. Insertion sequence-driven evolution of *Escherichia coli* in chemostats. *J. Mol. Evol.* **72**:398–412.
37. **Riehle MM, Bennett AF, Long AD.** 2001. Genetic architecture of thermal adaptation in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **98**:525–530.
38. **Rau MH, Marvig RL, Ehrlich GD, Molin S, Jelsbak L.** 2012. Deletion and acquisition of genomic content during early stage adaptation of *Pseudomonas aeruginosa* to a human host environment. *Environ. Microbiol.* **14**:2200–2211.
39. **Kotewicz ML, Jackson SA, LeClerc JE, Cebula TA.** 2007. Optical maps distinguish individual strains of *Escherichia coli* O157 : H7. *Microbiology* **153**:1720–1733.
40. **Shukla SK, Kislow J, Briska A, Henkhaus J, Dykes C.** 2009. Optical mapping reveals a large genetic inversion between two methicillin-resistant *Staphylococcus aureus* strains. *J. Bacteriol.* **191**:5717–5723.
41. **Turner PC, Yomano LP, Jarboe LR, York SW, Baggett CL, Moritz BE, Zentz EB, Shanmugam KT, Ingram LO.** 2012. Optical mapping and sequencing of the *Escherichia coli* KO11 genome reveal extensive chromosomal rearrangements, and multiple tandem

- copies of the *Zymomonas mobilis* *pdc* and *adhB* genes. J. Ind. Microbiol. Biotechnol. **39**:629–639.
42. **Wiser MJ, Ribeck N, Lenski RE.** 2013. Long-term dynamics of adaptation in asexual populations. Science **342**:1364–1367.
43. **Papadopoulos D, Schneider D, Meier-Eiss J, Arber W, Lenski RE, Blot M.** 1999. Genomic evolution during a 10,000-generation experiment with bacteria. Proc. Natl Acad. Sci. USA **96**:3807–3812.
44. **Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF.** 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature **461**:1243–1247.
45. **Ochman H, Moran NA.** 2001. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. Science **292**:1096–1099.
46. **Stinear TP, Seemann T, Pidot S, Frigui W, Reysset G, Garnier T, Meurice G, Simon D, Bouchier C, Ma L, Tichit M, Porter JL, Ryan J, Johnson PDR, Davies JK, Jenkin GA, Small PLC, Jones LM, Tekaia F, Laval F, Daffé M, Parkhill J, Cole ST.** 2007. Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. Genome Res. **17**:192–200.
47. **Dekel E, Alon U.** 2005. Optimality and evolutionary tuning of the expression level of a protein. Nature **436**:588–592.
48. **Nilsson AI, Koskiniemi S, Eriksson S, Kugelberg E, Hinton JC, Andersson DI.** 2005. Bacterial genome size reduction by experimental evolution. Proc. Natl Acad. Sci. USA **102**:12112–12116.
49. **Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MT, Churcher CM, Bentley SD, Mungall KL, Cerdeño-Tárraga AM, Temple L, James K, Harris B, Quail MA, Achtman M, Atkin R, Baker S, Basham D, Bason N, Cherevach**

- I, Chillingworth T, Collins M, Cronin A, Davis P, Doggett J, Feltwell T, Goble A, Hamlin N, Hauser H, Holroyd S, Jagels K, Leather S, Moule S, Norberczak H, O'Neil S, Ormond D, Price C, Rabbinowitsch E, Rutter S, Sanders M, Saunders D, Seeger K, Sharp S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Unwin L, Whitehead S, Barrell BG, Maskell DJ.** 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nature Genet.* **35**:32-40.
50. **Cooper VS, Lenski RE.** 2000. The population genetics of ecological specialization in evolving *Escherichia coli* populations. *Nature* **407**:736-739.
51. **Leiby N, Marx CJ.** 2014. Metabolic erosion primarily through mutation accumulation, and not tradeoffs, drives limited evolution of substrate specificity in *Escherichia coli*. *PLoS Biol.* **12**:e1001789.
52. **Studier FW, Daegelen P, Lenski RE, Maslov S, Kim JF.** 2009. Understanding the differences between genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3) and comparison of the *E. coli* B and K-12 genomes. *J. Mol. Biol.* **394**:653-680.
53. **Crozat E, Philippe N, Lenski RE, Geiselmann J, Schneider D.** 2005. Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. *Genetics* **169**:523–532.
54. **Woods R, Schneider D, Winkworth CL, Riley MA, Lenski RE.** 2006. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **103**:9107-9112.
55. **Hengge-Aronis R.** 1993. Survival of hunger and stress: the role of *rpoS* in early stationary phase gene regulation in *E. coli*. *Cell* **72**:165–168.
56. **Philippe N, Crozat E, Lenski RE, Schneider D.** 2007. Evolution of global regulatory

- networks during a long-term experiment with *Escherichia coli*. *BioEssays* **29**:846–860.
57. **Hindré T, Knibbe C, Beslon G, Schneider D.** 2012. New insights into bacterial adaptation through *in vivo* and *in silico* experimental evolution. *Nature Rev. Microbiol.* **10**:352–365.
58. **Lam MM, Seemann T, Bulach DM, Gladman SL, Chen H, Haring V, Moore RJ, Ballard S, Grayson ML, Johnson PD, Howden BP, Stinear TP.** 2012. Comparative analysis of the first complete *Enterococcus faecium* genome. *J. Bacteriol.* **194**:2334–2341.
59. **Okinaka RT, Price EP, Wolken SR, Gruendike JM, Chung WK, Pearson T, Xie G, Munk C, Hill KK, Challacombe J, Ivins BE, Schupp JM, Beckstrom-Sternberg SM, Friedlander A, Keim P.** 2011. An attenuated strain of *Bacillus anthracis* (CDC 684) has a large chromosomal inversion and altered growth kinetics. *BMC Genomics* **12**:477.
60. **Rocha EPC.** 2004. Order and disorder in bacterial genomes. *Current Op. Microbiol.* **7**:519–527.
61. **Rocha EPC, Danchin A.** 2003. Essentiality, not expressiveness, drives gene strand bias in Bacteria. *Nature Genet.* **34**:377–378.
62. **Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D’Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabasi AL, Oltvai ZN, Osterman AL.** 2003. Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J Bacteriol.* **14**:5673–5684.
63. **Zinser ER, Schneider D, Blot M, Kolter R.** 2003. Bacterial evolution through the selective loss of beneficial genes. Trade-offs in expression involving two loci. *Genetics* **164**:1271–1277.



64. **Tanaka KH, Dallaire-Dufresne S, Daher RK, Frenette M, Charette SJ.** 2012. An insertion sequence-dependent plasmid rearrangement in *Aeromonas salmonicida* causes the loss of the type three secretion system. *PLoS One* **7**:e33725.
65. **Mijnendonckx K, Provoost A, Monsieurs P, Leys N, Mergeay M, Mahillon J, Van Houdt R.** 2011. Insertion sequence elements in *Cupriavidus metallidurans* CH34: distribution and role in adaptation. *Plasmid* **65**:193–203.
66. **Lenski RE, Rose MR, Simpson SC, Tadler SC.** 1991. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *Am. Nat.* **138**:1315–1341.
67. **Jeong H, Barbe V, Lee CH, Vallenet D, Yu DS, Choi SH, Couloux A, Lee SW, Yoon SH, Cattolico L, Hur CG, Park HS, Ségurens B, Kim SC, Oh TK, Lenski RE, Studier FW, Daegelen P, Kim JF.** 2009. Genome sequences of *Escherichia coli* B strains REL606 and BL21 (DE3). *J. Mol. Biol.* **394**:644–652.
68. **Wielgoss S, Barrick JE, Tenaillon O, Wiser MJ, Dittmar WJ, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D.** 2013. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc. Natl Acad. Sci. USA* **110**:222–227.
69. **Wielgoss S, Barrick JE, Tenaillon O, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D.** 2011. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3 (Bethesda)* **1**:183–186.
70. **Vallenet D, Belda E, Calteau A, Cruveiller S, Engelen S, Lajus A, Le Fèvre F, Longin C, Mornico D, Roche D, Rouy Z, Salvignol G, Scarpelli C, Thil Smith AA, Weiman M, Médigue C.** 2013. MicroScope--an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucl. Acids Res.* **41(Database issue)**:D636–D647.

71. **Kaguni JM.** 2011. Replication initiation at the *Escherichia coli* chromosomal origin. *Curr. Op. Chem. Biol.* **15**:606–613.
72. **Ip SC, Bregu M, Barre FX, Sherratt DJ.** 2003. Decatenation of DNA circles by FtsK-dependent Xer site-specific recombination. *EMBO J.* **22**:6399–6407.

## Figure Legends

**FIG 1** Large-scale chromosomal rearrangements in evolved clones sampled after 40,000 generations from each of the 12 populations of the long-term evolution experiment. Each clone is indicated by the name of the population from which it was sampled. Asterisks mark clones that evolved higher point-mutation rates than the ancestor. The percentage shown below each clone indicates the change in its genome size relative to the ancestor. The optical map of the ancestral strain, computed from its genome sequence (67), is shown on the top, with the vertical blue lines showing the locations of the *NcoI* restriction sites used for this procedure. The locations of the replication origin and terminus are shown on the ancestral map, together with the *manB-cpsG* region that was affected by deletions in ten evolved clones. The chromosomal macrodomains (26) are indicated below the ancestral map. All large rearrangements are shown, relative to the ancestral genome for easier comparison, using the color key below the figure. New IS-element insertions cannot be detected by optical mapping because they generally produce rearrangements too small to be resolved by this method. Vertical lines labelled by  $\Delta 1$  to  $\Delta 9$  indicate regions affected repeatedly (in two or more populations) by deletions; the D1 vertical line indicates a region affected repeatedly by duplication events. Three chromosomal intervals, shown as I1 to I3, and a sub-region 1.1 within I1, were affected repeatedly by inversions. We describe the boundaries of all chromosomal rearrangements according to the ancestral map shown here. As a consequence, three inversions (inversion 1 in Ara+1, inversions 1 and 2 in Ara-3) have sizes larger than one half of the chromosome (Table S1). These inversions could have been described alternatively as inversions of the other part of the chromosome, with their sizes then being smaller than one half of the chromosome. For example, we describe inversion 1 as being ~2.8 Mbp, whereas according to the alternative its size would

be ~1.8 Mbp. We use the coordinates according to the ancestral map for internal consistency, and this choice does not affect any conclusions.

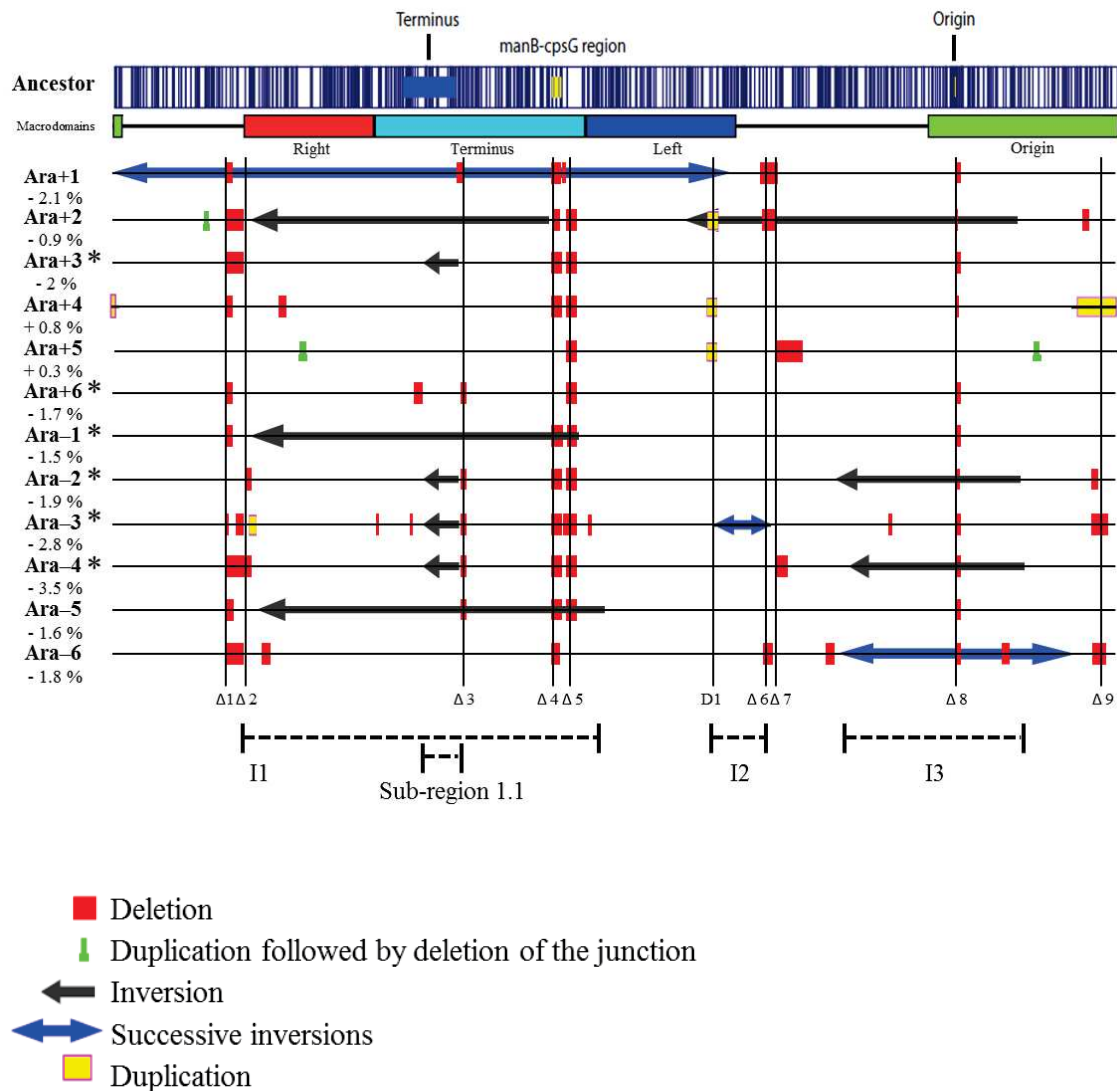
**FIG 2** Successive inversions in population Ara+1. (A) Optical map of the genome of the evolved clone sampled at 40,000 generations from population Ara+1 compared to the ancestor. Dark blue lines indicate *NcoI* restriction sites. White boxes show discrepant regions. The variously colored arrows indicate homologous regions of the two genomes with their corresponding locations in the two chromosomes. Red boxes indicate deletions. Black lines connecting the two genomes show alignment. (B) Chronology of the four inversions that occurred over evolutionary time in population Ara+1. Time points (2K, 20K and 30K for 2000, 20,000 and 30,000 generations, respectively) indicate the earliest detected occurrence of each inversion. IND indicates an inversion that was not detected in any of the evolved clones that were analyzed, but which represents one of the two possible intermediate steps leading to the genome observed at 40,000 generations. The variously colored arrows are the same as in (A). Black lines indicate the inversions, with the names and locations of primers used during the PCR experiments also shown.

**FIG 3** Hypothetical mechanism for the imperfect duplication seen in the evolved clone from population Ara+2. The horizontal line represents a section of the genome. The letters x and y show the borders of the duplication; a and b are the future insertion sites of two *IS1* elements. The two *IS1* insertions are indicated by arrows. The *IS1*-mediated deletion event is indicated by the large cone, with only one *IS1* copy remaining in the 40,000-generation clone.

**FIG 4** Changes in genome symmetry. (A) Changes in genome symmetry are shown in minutes between the *oriC* and *dif* loci along the leading replication branch for the evolved clones sampled after 40,000 generations from each of the 12 populations. (B) Circular maps of the chromosome show the near-perfect symmetry of the ancestor (left) and the imbalance in the evolved clone from population Ara+2 (right). The curved black arrow inside the circle corresponds to the leading strand, and the colored arrows show the two large inversions. The positions of *oriC* and *dif* are shown in minutes outside the circle; the length of the leading strand and the genome size (bp) are shown inside the circles.

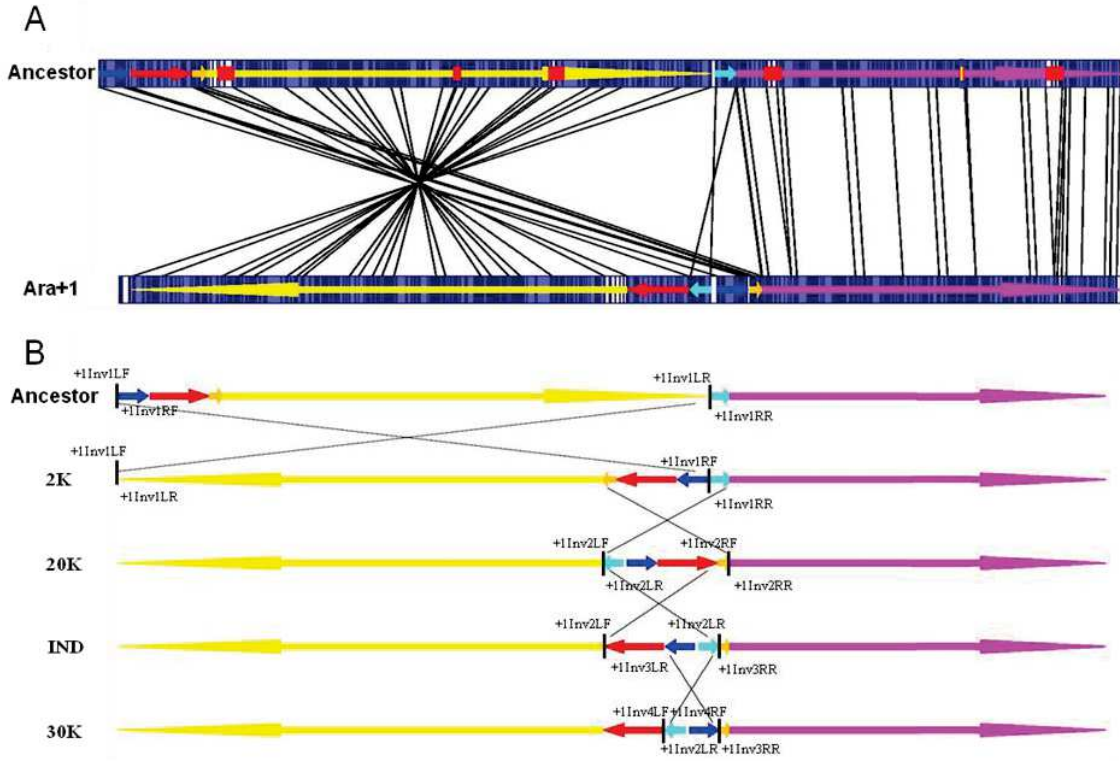
**FIG S1** Number of rearrangements in clones sampled from each of the 12 populations at 40,000 generations. Colors indicate the different types of rearrangements according to the genetic mechanism of production (see text).

**FIG S2** Successive inversions leading to genomes observed in populations Ara-3 and Ara-6 at 40,000 generations. Optical maps and hypothetical chronologies of the two and three successive inversions found in populations Ara-3 (A) and Ara-6 (B), respectively, compared to the ancestor genome. Symbols are identical to Fig. 2.



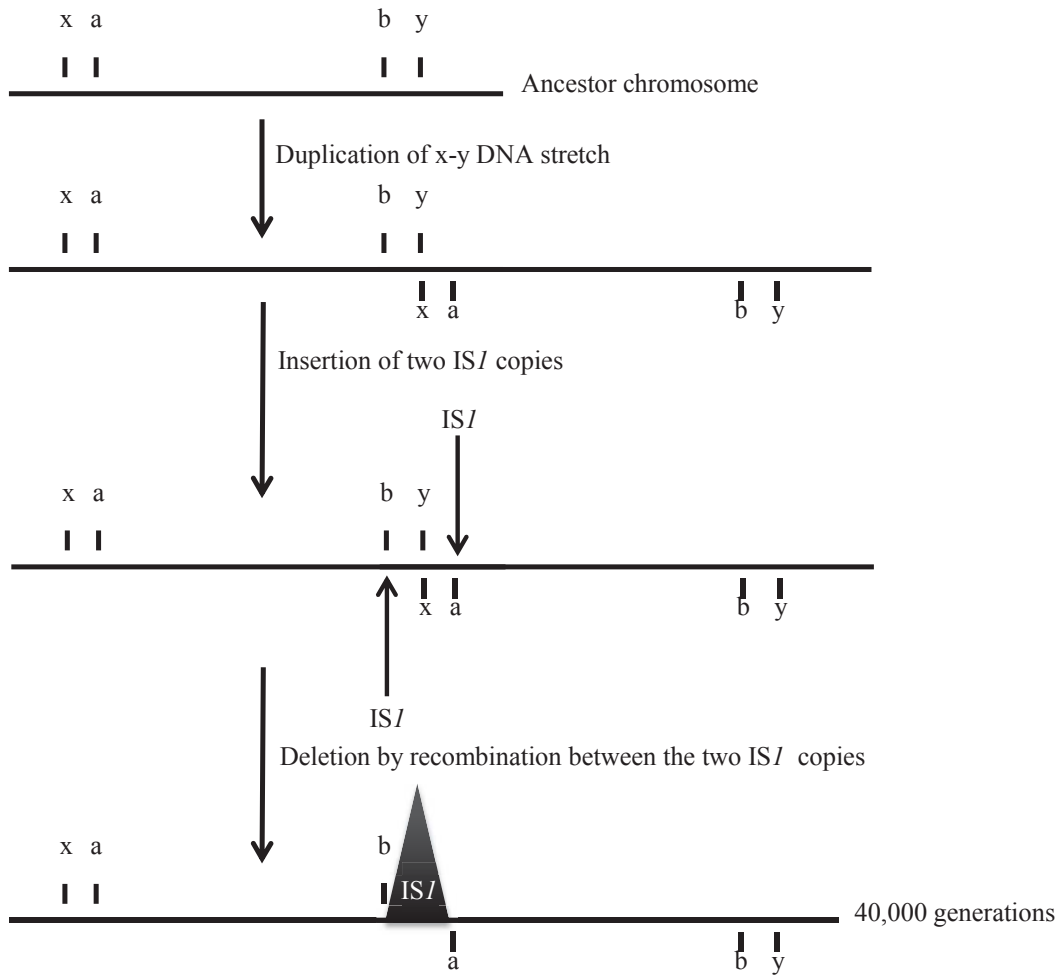
**FIG 1** Large-scale chromosomal rearrangements in evolved clones sampled after 40,000 generations from each of the 12 populations of the long-term evolution experiment. Each clone is indicated by the name of the population from which it was sampled. Asterisks mark clones that evolved higher point-mutation rates than the ancestor. The percentage shown below each clone indicates the change in its genome size relative to the ancestor. The optical map of the ancestral strain, computed from its genome sequence (67), is shown on the top, with the vertical blue lines showing the locations of the *NcoI* restriction sites used for this procedure. The locations of the replication origin and terminus are shown on the ancestral

map, together with the *manB-cpsG* region that was affected by deletions in ten evolved clones. The chromosomal macrodomains (26) are indicated below the ancestral map. All large rearrangements are shown, relative to the ancestral genome for easier comparison, using the color key below the figure. New IS-element insertions cannot be detected by optical mapping because they generally produce rearrangements too small to be resolved by this method. Vertical lines labelled by  $\Delta 1$  to  $\Delta 9$  indicate regions affected repeatedly (in two or more populations) by deletions; the D1 vertical line indicates a region affected repeatedly by duplication events. Three chromosomal intervals, shown as I1 to I3, and a sub-region 1.1 within I1, were affected repeatedly by inversions. We describe the boundaries of all chromosomal rearrangements according to the ancestral map shown here. As a consequence, three inversions (inversion 1 in Ara+1, inversions 1 and 2 in Ara-3) have sizes larger than one half of the chromosome (Table S1). These inversions could have been described alternatively as inversions of the other part of the chromosome, with their sizes then being smaller than one half of the chromosome. For example, we describe inversion 1 as being ~2.8 Mbp, whereas according to the alternative its size would be ~1.8 Mbp. We use the coordinates according to the ancestral map for internal consistency, and this choice does not affect any conclusions.

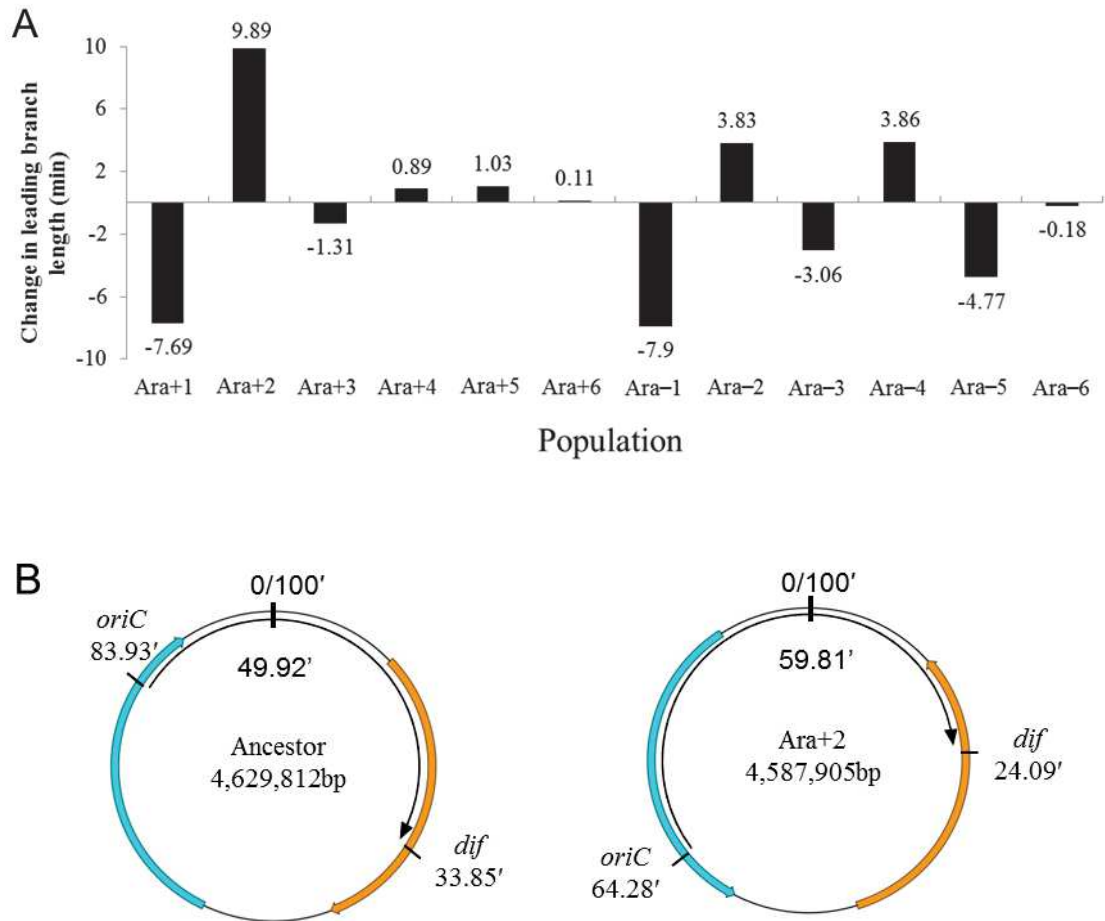


**FIG 2** Successive inversions in population Ara+1. (A) Optical map of the genome of the evolved clone sampled at 40,000 generations from population Ara+1 compared to the ancestor. Dark blue lines indicate *NcoI* restriction sites. White boxes show discrepant regions. The variously colored arrows indicate homologous regions of the two genomes with their corresponding locations in the two chromosomes. Red boxes indicate deletions. Black lines connecting the two genomes show alignment. (B) Chronology of the four inversions that occurred over evolutionary time in population Ara+1. Time points (2K, 20K and 30K for 2000, 20,000 and 30,000 generations, respectively) indicate the earliest detected occurrence of each inversion. IND indicates an inversion that was not detected in any of the evolved clones that were analyzed, but which represents one of the two possible intermediate steps leading to the genome observed at 40,000 generations. The variously colored arrows are the same as in (A). Black lines indicate the inversions, with the names and locations of primers used during the PCR experiments also shown.

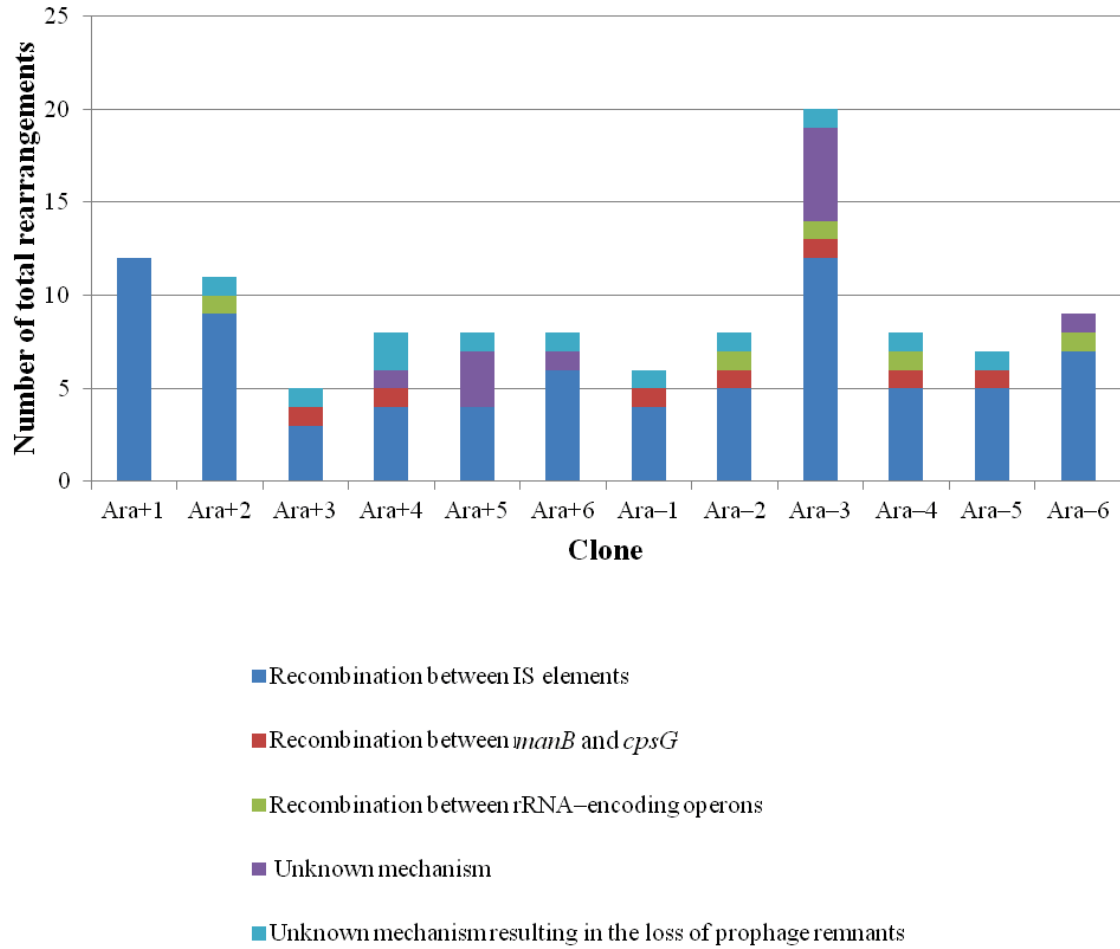




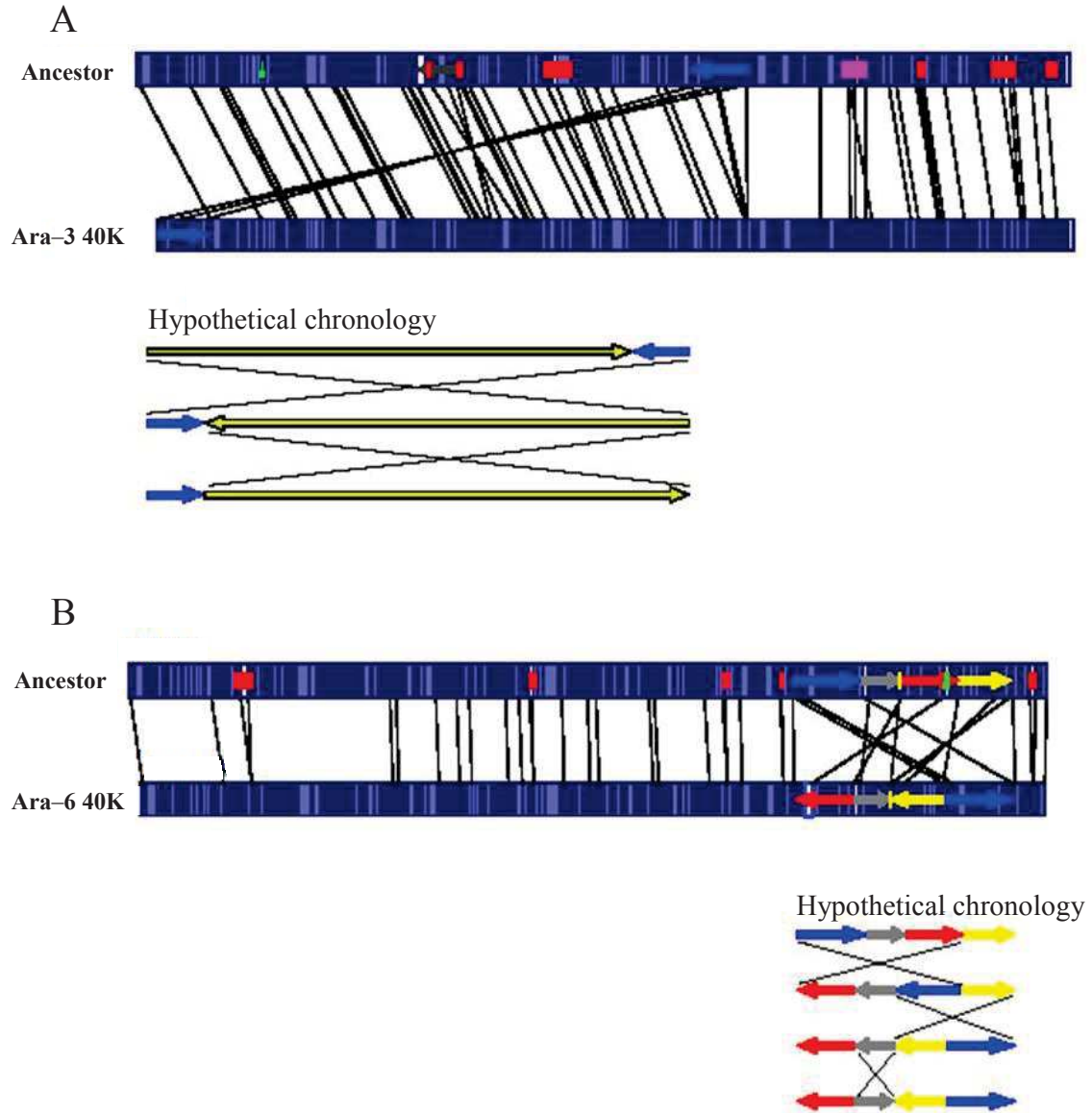
**FIG 3** Hypothetical mechanism for the imperfect duplication seen in the evolved clone from population Ara+2. The horizontal line represents a section of the genome. The letters x and y show the borders of the duplication; a and b are the future insertion sites of two IS1 elements. The two IS1 insertions are indicated by arrows. The IS1-mediated deletion event is indicated by the large cone, with only one IS1 copy remaining in the 40,000-generation clone.



**FIG 4** Changes in genome symmetry. (A) Changes in genome symmetry are shown in minutes between the *oriC* and *dif* loci along the leading replication branch for the evolved clones sampled after 40,000 generations from each of the 12 populations. (B) Circular maps of the chromosome show the near-perfect symmetry of the ancestor (left) and the imbalance in the evolved clone from population Ara+2 (right). The curved black arrow inside the circle corresponds to the leading strand, and the colored arrows show the two large inversions. The positions of *oriC* and *dif* are shown in minutes outside the circle; the length of the leading strand and the genome size (bp) are shown inside the circles.



**FIG S1** Number of rearrangements in clones sampled from each of the 12 populations at 40,000 generations. Colors indicate the different types of rearrangements according to the genetic mechanism of production (see text).



**FIG S2** Successive inversions leading to genomes observed in populations Ara-3 and Ara-6 at 40,000 generations. Optical maps and hypothetical chronologies of the two and three successive inversions found in populations Ara-3 (A) and Ara-6 (B), respectively, compared to the ancestor genome. Symbols are identical to Fig. 2.

**TABLE S1** Rearrangements detected in 40,000-generation clones sampled from each of the 12 populations of the LTEE

<b>Population Ara+1</b>						
Type	Start <sup>a</sup>	End <sup>a</sup>	Size (bp)	Mechanism <sup>b</sup>	Genes altered <sup>c</sup>	Parallelism <sup>d</sup>
Inversion 1 <sup>e</sup>	4,615,673	2,774,435	2,788,574	RRE-IS150	<i>nadR-insK-3</i> , 2661 genes	11
Inversion 2 <sup>e</sup>	490,480	2,877,312	607,496	RRE-IS150	<i>nadR-ybbN</i> and <i>cysH-kduD</i> , 549 genes	
Inversion 3 <sup>e</sup>	353,643	2,877,312	470,659	RRE-IS150	<i>nadR-yaiO</i> and <i>cysH-kduD</i> , 418 genes	
Inversion 4 <sup>e</sup>	146,102	2,877,312	263,118	RRE-IS150	<i>nadR-yadG</i> and <i>cysH-kduD</i> , 224 genes	
Deletion	547,701	556,319	8618	RRE-IS1	DLP12-like locus: <i>ycbR</i> , <i>ycbS</i> , <i>ycbT</i> , <i>ycbU</i> , <i>ECB_00510</i> , <i>nohB</i> , <i>ECB_00512</i> , <i>ECB_00513</i>	□1
Deletion	590,445	601,096	10,651	RRE-IS150	<i>hokE</i> , <i>IS 186</i> , <i>entD</i> , <i>fepA</i> , <i>fes</i> , <i>ybdZ</i> , <i>entF</i> , <i>fepE</i>	□2
Deletion	1,353,847	1,354,425	578	RRE-IS150	<i>sapA</i>	
Deletion	1,609,176	1,615,468	6292	RRE-IS3	Qin-like locus: <i>ycbW</i> , <i>ynsB</i> , <i>ynfN</i> , <i>ECB_01516</i> , <i>cspI</i> , <i>ydfP</i> , <i>ydfQ-2</i> , <i>ydfR</i> , <i>essQ</i> , <i>ECB_01522</i> , <i>ECB_01523</i>	□3
Deletion	2,034,326	2,053,851	19,525	RRE-IS1	<i>wbbD</i> , <i>wbbC</i> , <i>wzy</i> , <i>wbbB</i> , <i>wbbA</i> , <i>vioB</i> , <i>vioA</i> , <i>wzx</i> , <i>rmlC</i> , <i>rfbA</i> , <i>rfbD</i> , <i>rfbB</i> , <i>galF</i> , <i>wcaM</i> , <i>wcaL</i> , <i>wcaK</i> , <i>wxC</i> , <i>wcaJ</i>	□4
Deletion	2,086,534	2,093,027	6093	RRE-IS150	<i>yegM</i> , <i>yegN</i> , <i>yegO</i>	
Deletion	2,999,596	3,048,488	48,892	RRE-IS150	<i>flu-yghK</i> , 35 genes including CP-44-like locus	□6, □7
Deletion	3,894,996	3,901,134	6138	RRE-IS150	<i>rbsD</i> , <i>rbsA</i> , <i>rbsC</i> , <i>rbsB</i> , <i>rbsK</i> , <i>rbsR</i>	□8
<b>Population Ara+2</b>						
Deletion	547,701	588,493	40,792	RRE-IS1	DLP12-like locus: <i>ycbR-ybdK</i> , 37 genes	□1
Inversion	590,322	2,034,326	1,444,004	RRE-IS1	<i>rne-manB</i> , 1447 genes	11
Imperfect duplication <sup>f</sup>	1,734,709	1,748,420	13,711	Unknown	<i>sufE</i> , <i>lpp</i> , <i>ynhG</i> , <i>sufS</i> , <i>sufD</i> , <i>sufC</i> , <i>sufB</i> , <i>sufA</i> , <i>ydiH</i> , <i>ydiI</i> , <i>ydiJ</i> , <i>ydiK</i> , <i>ydiL</i> , <i>ydiM</i>	

Deletion of duplication junction				RRE-IS1		
Deletion	2,037,724	2,048,103	10,379	RRE-IS1	<i>wbbD, wbbC, wzy, wbbB, wbbA, vioB, vioA, wzx, rmlC, rfbA, rfbD, rfbB, galF, wcaM</i>	□4
Deletion	2,100,286	2,122,432	22,146	Unknown	Prophage 2 locus: <i>ogrK-ECB_02012</i> , 26 genes	□5
Inversion	2,651,000	4,189,422	1,538,422	RRE-ribosomal operon	<i>rrsG-rrsE</i> , 1489 genes	I2, I3
Duplication	2,713,428	2,774,197	60,769	RRE-IS186	<i>serV-iap</i> , 59 genes	D1
Deletion	2,996,588	3,023,945	27,357	RRE-IS1	<i>ECB_02797-ECB_02822</i> , 26 genes	□6
Deletion	3,894,996	3,895,833	837	RRE-IS150	<i>rbsD, rbsA</i>	□8
Deletion	4,506,111	4,521,569	15,458	RRE-IS1	<i>sgcR, sgcE, sgcA, sgcQ, sgcC, sgcB, sgcX, yjhP, yjhQ, ECB_04174, yjhR, yjhS, yjhT, yjhA</i>	

### Population Ara+3

Deletion	547,701	588,493	40,792	RRE-IS1	DLP12-like locus: <i>ybcR-ybdK</i> , 37 genes	1
Inversion	1,443,893	1,615,474	171,581	RRE-IS3	<i>ydbC-ECB_01523</i> , 161 genes	Sub-region 1.1
Deletion	2,031,432	2,054,725	23,293	RRE- <i>manB-cpsG</i>	<i>manB, manC, insB-14, insA-14, wbbD, wbbC, wzy, wbbB, wbbA, vioB, vioA, wzx, rmlC, rfbA, rfbD, rfbB, galF, wcaM, wcaL, wcaK, wzxC, wcaJ</i>	□4
Deletion	2,100,286	2,122,432	22,146	Unknown	Prophage 2 locus: <i>ogrK-ECB_02012</i> , 26 genes	□5
Deletion	3,894,996	3,901,160	6164	RRE-IS150	<i>rbsD, rbsA, rbsC, rbsB, rbsK, rbsR, yieO</i>	□8

### Population Ara+4

Deletion	547,701	559,507	11,806	RRE-ISI	DLP12-like locus: <i>ycbR</i> , <i>ycbS</i> , <i>ycbT</i> , <i>ycbU</i> , <i>ECB_00510</i> , <i>nohB</i> , <i>ECB_00512</i> , <i>ECB_00513</i> , <i>ECB_00514</i> , <i>ECB_00515</i> , <i>ECB_00516</i> , <i>ECB_00517</i> , <i>appY</i>	□1
Deletion	787,866	799,956	12,090	Unknown	Prophage 434 locus: <i>ECB_00726</i> , <i>ECB_00727</i> , <i>ECB_00728</i> , <i>ECB_00729</i> , <i>ECB_00730</i> , <i>ECB_00731</i> , <i>ECB_00732</i> , <i>ECB_00733</i> , <i>ECB_00734</i> , <i>ECB_00735</i> , <i>ECB_00736</i> , <i>ECB_00737</i> , <i>ECB_00738</i> , <i>ECB_00739</i>	
Deletion	2,031,432	2,054,725	23,293	RRE- <i>manB-cpsG</i>	<i>manB</i> , <i>manC</i> , <i>insB-14</i> , <i>insA-14</i> , <i>wbbD</i> , <i>wbbC</i> , <i>wzy</i> , <i>wbbB</i> , <i>wbbA</i> , <i>vioB</i> , <i>vioA</i> , <i>wzx</i> , <i>rmlC</i> , <i>rfaA</i> , <i>rfaD</i> , <i>rfaB</i> , <i>galF</i> , <i>wcaM</i> , <i>wcaL</i> , <i>wcaK</i> , <i>wzcC</i> , <i>wcaJ</i>	□4
Deletion	2,100,286	2,122,432	22,146	Unknown	Prophage 2 locus: <i>ogrK-ECB_02012</i> , 26 genes	□5
Deletion	2,129,367	2,143,064	13,501	RRE-ISI	<i>gatY</i> , <i>fbaB</i> , <i>yegT</i> , <i>yegU</i> , <i>yegV</i> , <i>yegW</i> , <i>yegX</i> , <i>thiD</i> , <i>thiM</i> , <i>ECB_02033</i> , <i>ECB_02034</i> , <i>ECB_02035</i> , <i>ECB_02036</i> , <i>ECB_02037</i> , <i>ECB_02038</i> , <i>ECB_02039</i>	
Duplication	2,760,570	2,771,648	11,078	RRE-ISI	<i>ygbN</i> , <i>rpoS</i> , <i>nlpD</i> , <i>pcm</i> , <i>surE</i> , <i>truD</i> , <i>ispF</i> , <i>ispD</i> , <i>ftsB</i> , <i>ygbE</i> , <i>cysC</i> , <i>cysN</i> , <i>cysD</i> , <i>iap</i>	D1
Deletion	3,894,996	3,898,944	3948	RRE-IS150	<i>rbsD</i> , <i>rbsA</i> , <i>rbsC</i> , <i>rbsB</i>	□8
Duplication	4,456,320	6316	179,808	Unknown	<i>thrL-lasT</i> , 161 genes	

### Population Ara+5

Imperfect duplication <sup>f</sup>	887,915	914,970	27,055	Unknown	<i>ECB_00825-ybjO</i> , 39 genes	
Deletion of duplication junction				RRE-ISI		
Deletion	2,100,286	2,122,432	22,146	Unknown	Prophage 2 locus: <i>ogrK-ECB_02012</i> , 26 genes	□5
Duplication	2,734,828	2,774,455	39,627	RRE-ISI	<i>ascF-cysD</i> , 38 genes	D1

Deletion	3,024,346	3,080,112	55,766	RRE-IS150	CP-44-like locus: <i>insB-22-hybO</i> , 51 genes	□7
Deletion	3,894,996	3,895,279	283	RRE-IS150	<i>rbsD</i>	□8
Imperfect duplication <sup>f</sup>	4,261,613	4,290,705	29,092	Unknown	<i>yjcF, actP, yjcH, acs, nrfA, nrfB, nrfC, nrfD, nrfE, nrfF, nrfG, gltP, yjcO, fdhF, yjcP, yjcQ, yjcR, yjcS, alsK, alsE, alsC, alsA, alsB</i>	
Deletion of duplication junction				RRE-IS1		

### Population Ara+6

Deletion	547,701	556,319	8618	RRE-IS1	DLP12-like locus: <i>ybcR, ybcS, ybcT, ybcU, ECB_00510, nohB, ECB_00512, ECB_00513</i>	□1
Deletion	1,433,349	1,442,639	9290	RRE-IS3	<i>hslJ, ldhA, ydbH, ynbE, ydbL, feaR, feaB, tynA</i>	
Deletion	1,609,176	1,615,468	6292	RRE-IS3	Qin-like locus: <i>ybcW, gnsB, ynfN, ECB_01516, cspI, ydfP, ydfQ-2, ydfR, essQ, ECB_01522, ECB_01523</i>	□3
Deletion	1,974,042	1,976,274	2232	RRE-IS1	<i>yedU, yedV, yedW</i>	
Deletion	2,100,286	2,122,432	22,146	Unknown	Prophage 2 locus: <i>ogrK-ECB_02012</i> , 26 genes	□5
Deletion	3,024,713	3,025,120	407	RRE-IS1	<i>ECB_02825</i>	
Deletion	3,775,351	3,796,244	20,893	Unknown	<i>selC, ECB_03516, ECB_03517, ECB_03518, ECB_03519, ECB_03520, ECB_03521, ECB_03522, ECB_03523, ECB_03524, ECB_03525, ECB_03526, ECB_03527, ECB_03528, ECB_03530, ECB_03531, ECB_03532, ECB_03533, ECB_03534, ykgN, insI, ECB_03537, ECB_03538, ECB_03539, ECB_03540</i>	
Deletion	3,894,996	3,900,156	5160	RRE-IS150	<i>rbsD, rbsA, rbsC, rbsB, rbsK, rbsR</i>	□8

### Population Ara-1



Deletion	547,701	555,877	8176	RRE-IS1	DLP12-like locus: <i>ybcR, ybcS, ybcT, ybcU, ECB_00510, nohB, ECB_00512, ECB_00513</i>	□1
Inversion	634,745	2,128,599	1,493,854	RRE-IS1	<i>ycjW-gatA</i> , 1455 genes	11
Deletion	2,031,703	2,054,996	23,293	RRE- <i>manB-cpsG</i>	<i>manB, manC, insB-14, insA-14, wbbD, wbbC, wzy, wbbB, wbbA, vioB, vioA, wzx, rmlC, rfbA, rfbD, rfbB, galF, wcaM, wcaL, wcaK, wzxC, wcaJ</i>	□4
Deletion	2,100,286	2,122,432	22,146	Unknown	Prophage 2 locus: <i>ogrK-ECB_02012</i> , 26 genes	□5
Deletion	2,129,369	2,137,411	8042	RRE-IS1	<i>gatY, fbaB, yegT, yegU, yegV, yegW, yegX</i>	
Deletion	3,894,996	3,901,921	6927	RRE-IS150	<i>rbsD, rbsA, rbsC, rbsB, rbsK, rbsR</i>	□8

### Population Ara-2

Deletion	590,472	619,116	28,644	RRE-IS1	<i>hokE, insL-3, entD, fepA, fes, ybdZ, entF, fepE, fepC, fepG, fepD, ybdA, fepB, entC, entE, entB, entA, ybdB, cstA, ybdD, ybdH, ybdL, ybdM, ybdN</i>	□2
Inversion	1,443,893	1,607,920	164,027	RRE-IS3	<i>ybdC-ECB_01510</i> , 148 genes	Sub-region 1.1
Deletion	1,609,176	1,615,468	6292	RRE-IS3	Qin-like locus: <i>ybcW, ynsB, ynfN, ECB_01516, cspI, ydfP, ydfQ-2, ydfR, essQ, ECB_01522, ECB_01523</i>	□3
Deletion	2,031,432	2,054,725	23,293	RRE- <i>manB-cpsG</i>	<i>manB, manC, insB-14, insA-14, wbbD, wbbC, wzy, wbbB, wbbA, vioB, vioA, wzx, rmlC, rfbA, rfbD, rfbB, galF, wcaM, wcaL, wcaK, wzxC, wcaJ</i>	□4
Deletion	2,100,286	2,122,432	22,146	Unknown	Prophage 2 locus: <i>ogrK-ECB_02012</i> , 26 genes	□5
Inversion	3,270,936	4,066,255	795,319	RRE-ribosomal operon	<i>yhgA-frvR</i> , 750 genes	13
Deletion	3,894,996	3,899,897	4901	RRE-IS150	<i>rbsD, rbsA, rbsC, rbsB, rbsK</i>	□8

Deletion	4,547,206	4,550,677	3471	RRE-IS1	<i>yjiN, yjiO, yjiPQ</i>	
<b>Population Ara-3</b>						
Deletion	547,701	550,351	2650	RRE-IS1	DLP12-like locus: <i>ycbR, ycbS, ycbT, ycbU</i>	□1
Deletion	581,861	588,495	6631	RRE-IS150	<i>cusA, pheP, ybdG, nfnB, ybdF, ybdJ, ybdK</i>	
Duplication	625,890	628,823	2933	Unknown		
Deletion	1,270,143	1,270,569	426	Unknown	<i>idrC</i>	
Deletion	1,424,369	1,426,343	1984	Unknown	<i>ECB_01344, pinR, ynaE</i>	
Inversion	1,420,707	1,607,920	187,213	RRE-IS3	<i>ECB_01321-ECB_01510</i> , 168 genes	Sub-region 1.1
Deletion	1,451,972	1,462,318	10,346	RRE-IS150	<i>acpD, hrpA, ydcF, aldA, gapC, insA-12, insB-12, cybB, ydcA, hokB, mokB</i>	
Deletion	1,604,719	1,605,335	616	Unknown	<i>stfR</i>	
Deletion	1,609,176	1,615,468	6292	RRE-IS3	Qin-like locus: <i>ycbW, gnsB, ynfN, ECB_01516, cspI, ydfP, ydfQ-2, ydfR, essQ, ECB_01522, ECB_01523</i>	□3
Deletion	1,729,054	1,731,495	2441	RRE-IS150	<i>ydhV, ydhY</i>	
Deletion	2,032,711	2,056,011	23,300	RRE- <i>manB-cpsG</i>	<i>manB, manC, insB-14, insA-14, wbbD, wbbC, wzy, wbbB, wbbA, vioB, vioA, wzx, rmlC, rfbA, rfbD, rfbB, galF, wcaM, wcaL, wcaK, wzc, wcaJ</i>	□4
Deletion	2,086,611	2,122,432	35,821	Unknown	Prophage 2 locus: <i>yegM-ECB_02012</i> , 34 genes	□5
Inversion 1 <sup>e</sup>	16,972	3,015,762	2,998,790	RRE-IS150	<i>nhaA-ECB_02816</i> , 2847 genes	I2
Inversion 2 <sup>e</sup>	16,972	2,775,877	2,758,905	RRE-IS150	<i>nhaA-insJ-3</i> , 2633 genes	
Duplication	3,517,305	3,625,448	108,143	RRE-IS186	<i>ggt-yhjQ</i> , 87 genes	
Deletion	3,549,957	3,553,255	3268	Unknown	<i>rhsB</i>	
Deletion	3,741,969	3,742,144	175	RRE-IS1	Intergenic	
Deletion	3,894,996	3,901,457	6461	RRE-IS150	<i>rbsD, rbsA, rbsC, rbsB, rbsK, rbsR</i>	□8

Deletion	4,017,756	4,018,101	345	RRE-ribosomal operon	<i>rrlB</i>	
Deletion	4,522,340	4,561,283	38,943	RRE-ISI	<i>fimE-hsdR</i> , 35 genes	□9

#### Population Ara-4

Deletion	547,701	619,884	71,413	RRE-ISI	DLP12-like locus: <i>ybcR-ybdN</i> , 63 genes	□1, □2
Inversion	1,443,893	1,607,920	164,027	RRE-IS3	<i>ybdC-ECB_01510</i> , 148 genes	Sub-region 1.1
Deletion	1,609,176	1,615,468	6292	RRE-IS3	Qin-like locus: <i>ybcW, ynsB, ynfN, ECB_01516, cspI, ydfP, ydfQ-2, ydfR, essQ, ECB_01522, ECB_01523</i>	□3
Deletion	2,031,432	2,054,725	23,293	RRE- <i>manB-cpsG</i>	<i>manB, manC, insB-14, insA-14, wbbD, wbbC, wzy, wbbB, wbbA, vioB, vioA, wzx, rmlC, rfbA, rfbD, rfbB, galF, wcaM, wcaL, wcaK, wzxC, wcaJ</i>	□4
Deletion	2,100,286	2,122,432	22,146	Unknown	Prophage 2 locus: <i>ogrK-ECB_02012</i> , 26 genes	□5
Deletion	3,024,712	3,063,026	38,314	RRE-ISI	CP-44-like locus: <i>ECB_02825-ECB_02856</i> , 32 genes	□7
Inversion	3,354,888	4,189,422	834,534	RRE-ribosomal operon	<i>rrsD-rrsE</i> , 811 genes	I3
Deletion	3,894,996	3,901,405	6409	RRE-IS150	<i>rbsD, rbsA, rbsC, rbsB, rbsK, rbsR</i>	□8

#### Population Ara-5

Deletion	16,972	17,043	71	RRE-IS150	Intergenic	
Deletion	547,701	558,574	10,873	RRE-ISI	DLP12-like locus: <i>ybcR, ybcS, ybcT, ybcU, ECB_00510, nohB, ECB_00512, ECB_00513, ECB_00514, ECB_00515, ECB_00516</i>	□1

Inversion	666,130	2,283,472	1,617,342	RRE-IS150	<i>ydcM-yfaA</i> , 1560 genes	11
Deletion	1,609,176	1,615,468	6292	RRE-IS3	Qin-like locus: <i>ybcW</i> , <i>ynsB</i> , <i>ynfN</i> , <i>ECB_01516</i> , <i>cspI</i> , <i>ydfP</i> , <i>ydfQ-2</i> , <i>ydfR</i> , <i>essQ</i> , <i>ECB_01522</i> , <i>ECB_01523</i>	□3
Deletion	2,031,432	2,054,725	23,293	RRE- <i>manB-cpsG</i>	<i>manB</i> , <i>manC</i> , <i>insB-14</i> , <i>insA-14</i> , <i>wbbD</i> , <i>wbbC</i> , <i>wzy</i> , <i>wbbB</i> , <i>wbbA</i> , <i>vioB</i> , <i>vioA</i> , <i>wzx</i> , <i>rmlC</i> , <i>rfbA</i> , <i>rfbD</i> , <i>rfbB</i> , <i>galF</i> , <i>wcaM</i> , <i>wcaL</i> , <i>wcaK</i> , <i>wzc</i> , <i>wcaJ</i>	□4
Deletion	2,100,286	2,122,432	22,146	Unknown	Prophage 2 locus: <i>ogrK-ECB_02012</i> , 26 genes	□5
Deletion	3,894,996	3,900,623	5624	RRE-IS150	<i>rbsD</i> , <i>rbsA</i> , <i>rbsC</i> , <i>rbsB</i> , <i>rbsK</i> , <i>rbsR</i>	□8

#### Population Ara-6

Deletion	547,701	589,555	41,854	RRE-IS1	DLP12-like locus: <i>ycbR-ybdK</i> , 38 genes	□1
Deletion	2,034,326	2,045,407	11,081	RRE-IS1	<i>wbbD</i> , <i>wbbC</i> , <i>wzy</i> , <i>wbbB</i> , <i>wbbA</i> , <i>vioB</i> , <i>vioA</i> , <i>wzx</i> , <i>rmlC</i> , <i>rfbA</i> , <i>rfbD</i> , <i>rfbB</i>	□4
Deletion	3,001,956	3,015,762	13,806	RRE-IS150	<i>flu</i> , <i>yeeR</i> , <i>ECB_02802</i> , <i>yafZ</i> , <i>ECB_02804</i> , <i>yeeS</i> , <i>yeeT</i> , <i>yeeU</i> , <i>yeeV</i> , <i>yeeW</i> , <i>ECB_02810</i> , <i>ECB_02811</i> , <i>ECB_02812</i> , <i>ECB_02813</i> , <i>ECB_02814</i> , <i>ECB_02815</i> , <i>ECB_02816</i>	□6
Deletion	3,289,781	3,297,620	7839	Unknown	<i>gltB</i> , <i>gltD</i> , <i>yhcG</i> , <i>ECB_03080</i> , <i>yhcH</i> , <i>nanK</i> ,	
Deletion	3,894,996	3,901,703	6707	RRE-IS150	<i>rbsD</i> , <i>rbsA</i> , <i>rbsC</i> , <i>rbsB</i> , <i>rbsK</i> , <i>rbsR</i>	□8
Inversion 1 <sup>e</sup>	3,356,670	4,187,735	831,065	RRE-ribosomal operon	<i>yhgA-rrsE</i> , 777 genes	13
Inversion 2 <sup>e</sup>	3,713,694	4,453,625	739,931	RRE-IS150	<i>yihU-mgtA</i> , 679 genes	
Inversion 3 <sup>e</sup>	3,713,694	3,901,703	188,009	RRE-IS150	<i>yihU-yieO</i> , 181 genes	
Deletion	4,551,448	4,573,236	21,788	RRE-IS1	<i>yjiV</i> , <i>mcrC</i> , <i>mcrB</i> , <i>yjiW</i> , <i>hsdS</i> , <i>hsdM</i> , <i>hsdR</i> , <i>mrr</i> , <i>yjiA</i> , <i>yjiX</i> , <i>yjiY</i> , <i>hpaC</i> , <i>hpaB</i> , <i>hpaA</i> , <i>hpaX</i>	□9

<sup>a</sup> All positions are given according to the genomic coordinates of the ancestral strain (67). IS insertion events are not reported because optical mapping cannot resolve them.

<sup>b</sup> RRE, recombination between repeated elements, with the identity of the repeated element indicated after the hyphen.

<sup>c</sup> For rearrangements involving more than 25 genes, only the first and last genes altered by the rearrangement are shown, along with the total number of genes affected. Prophage regions are also indicated.

<sup>d</sup> Rearrangements involving chromosomal regions that evolved in parallel (at least two populations) are indicated based on the designations shown in Fig. 1.

<sup>e</sup> These rearrangements involved multiple successive events. The first event in a series gives the location of the altered region relative to the ancestral genome. For the later events in the series, the indicated coordinates reflect the altered region compared to the preceding state in that series.

<sup>f</sup> Specific type of successive events, in which a duplication was followed by a deletion of the junction between the duplicated copies. The indicated size of the duplication is the one that existed before the deletion occurred.

**TABLE S2** Clones from the LTEE used in this study

Population	Generation	Clone name
Ancestor	0	REL606
Ara+1	2000	REL1158A
Ara+1	2000	REL1158B
Ara+1	2000	REL1158C
Ara+1	15,000	REL7183A
Ara+1	15,000	REL7183B
Ara+1	15,000	REL7183C
Ara+1	20,000	REL9282A
Ara+1	20,000	REL9282B
Ara+1	20,000	REL9282C
Ara+1	25,000	REL10241
Ara+1	30,000	REL10450
Ara+1	30,000	REL10451
Ara+1	30,000	REL10452
Ara+1	35,000	REL10796
Ara+1	35,000	REL10797
Ara+1	35,000	REL10798
Ara+1 <sup>a</sup>	40,000	REL11008
Ara+1	40,000	REL11009
Ara+1	40,000	REL11010
Ara+2 <sup>a</sup>	40,000	REL10950
Ara+3 <sup>a</sup>	40,000	REL10953
Ara+4 <sup>a</sup>	40,000	REL10956
Ara+5 <sup>a</sup>	40,000	REL10982
Ara+6 <sup>a</sup>	40,000	REL10985
Ara-1 <sup>a</sup>	2000	REL1164A
Ara-1 <sup>a</sup>	5000	REL2179A
Ara-1 <sup>a</sup>	10,000	REL4536A
Ara-1 <sup>a</sup>	15,000	REL7177A
Ara-1 <sup>a</sup>	20,000	REL8593A
Ara-1 <sup>a</sup>	30,000	REL10391
Ara-1 <sup>a</sup>	40,000	REL10938
Ara-1 <sup>a</sup>	50,000	REL11330
Ara-2 <sup>a</sup>	40,000	REL11041
Ara-3 <sup>a</sup>	40,000	REL10979
Ara-4 <sup>a</sup>	40,000	REL10944
Ara-5 <sup>a</sup>	40,000	REL10947
Ara-6 <sup>a</sup>	40,000	REL11005

<sup>a</sup> Clones used for optical mapping.

**TABLE S3** Primers used in this study

Rearrangement <sup>a</sup>	Name	Sequence (5' to 3')	Expected PCR product size (bp)
Ara+1 Deletion 1	+1Del1F +1Del1R	ccttttgactgaaggtaagca tcaacagacacactaccgt	2858
Ara+1 Deletion 2	+1Del2F +1Del2R	ttaaccacctgcattccag cagttcgtcgatttcgcct	2175
Ara+1 Deletion 3	+1Del3F +1Del3R	atccccacggggatgccaaca gtcctgggatgaacgtcgct	2456
Ara+1 Deletion 4	+1Del4F +1Del4R	gagccagatgggtttccct gtgatgaagccagaaaggca	3044
Ara+1 Deletion 5	+1Del5F +1Del5R	accggaataagaggtagct gagcttccgtaatcagccgtgg	2678
Ara+1 Deletion 6	+1Del6F +1Del6R	gtctatcgctgacttgcgga gcctgtaaagccggtgacat	4129
Ara+1 Deletion 7	+1Del7F +1Del7R	ttctcttgcgtgactgcctt ccgatattcacaatgtggcg	3561
Ara+1 Deletion 8	+1Del8F +1Del8R	aactctgcgcaccgaagacg tgcaaaatcgatggttaccca	3592
Ara+1 Inversion 1 - Right	+1Inv1RF +1Inv1RR	gcggtatgaccaaagggtat gcgggaaatagctggcatgac	2168
Ara+1 Inversion 1 - Left	+1Inv1LF +1Inv1LR	taccgaataccacaccaatg acggcaaaactgtgggaacagg	2099
Ara+1 Inversion 2 - Right	+1Inv2RF +1Inv2RR	ccgctttacgttgggaccgg tggaatgcggtcgaggattt	1660
Ara+1 Inversion 2 - Left	+1Inv2LF +1Inv2LR	aggctttccagatcggggt gttatgggcgtgacgcgatt	1696
Ara+1 Inversion 3 - Right	+1Inv2RR +1Inv3RR	tggaatgcggtcgaggattt tgcagtacgcaccagcactt	3379
Ara+1	+1Inv2RF	ccgctttacgttgggaccgg	4239

Inversion 3 - Left	+1Inv3LR	cctgagtttattggctcgtct	
Ara+1	+1Inv4RF	agagttcgccccggctcaaga	2736
Inversion 4 - Right	+1Inv3RR	tcagtagcgcaccagcactt	
Ara+1	+1Inv4LF	ctgcttctccaggtagtgt	2214
Inversion 4 - Left	+1Inv2RR	tggaatgcggtcgaggattt	
Ara+2	+2Del1F	ccttttgactgaaggtaaagca	2195
Deletion 1	+2Del1R	gcagtattgcgaccagatgg	
Ara+2	+2Del2F	ttcgaagcgagaacagcct	2442
Deletion 2	+2Del2R	taacatcacctggatgtgc	
Ara+2	+2Del3F	gcctgacaaaaatgggcgt	2261
Deletion 3	+2Del3R	tagccggaacctgtgggagca	
Ara+2	+2Del4F	tatcaacggacctccacgga	2749
Deletion 4	+2Del4R	tcagcgttacaaggcttga	
Ara+2	+2Del5F	aactctgcgcaccgaagacg	3721
Deletion 5	+2Del5R	agccatcggctttgaactgg	
Ara+2	+2Del6F	ggggagaaagatgcacagtg	2440
Deletion 6	+2Del6R	taggaaacctcaatcggtca	
Ara+2	+2-Inv1RF	atthtgcgtcgtaaagcgggagc	960
Inversion 1 - Right	+2-Inv1RR	gthtgcctgatcagaaaacgct	
Ara+2	+2-Inv1LF	accggaataagaggtgagct	1053
Inversion 1 - Left	+2-Inv1LR	acagcaaaaagacgaggtgg	
Ara+2	+2-Inv2RF	tcaaacatcacccgaagatg	2438
Inversion 2 - Right	+2-Inv2RR	aaacgtctggaagaactgg	
Ara+2	+2-Inv2LF	gtagacagcagctccacacc	2297
Inversion 2 - Left	+2-Inv2LR	tgagcactgcaaagtacgct	

<sup>a</sup> The rearrangements are numbered according to their location from left to right in Fig. 1 and, for the four successive inversions in Ara+1, in their chronological order (Fig. 2).



**TABLE S4** Parallel rearrangements in the LTEE

Region <sup>a</sup>	Population	Start <sup>b</sup>	End <sup>b</sup>	Mechanism <sup>c</sup>	Genes affected in parallel <sup>d</sup>
<b>547,701-550,351:</b>					
□1	Ara+1	547,701	556,319	RRE-ISI	DLP12-like locus: <i>ycbR</i> , <i>ycbS</i> , <i>ycbT</i> , <i>ycbU</i> , <i>ECB_00510</i> , <i>nohB</i> , <i>ECB_00512</i> , <i>ECB_00513</i>
□1	Ara+2	547,701	588,493	RRE-ISI	
□1	Ara+3	547,701	588,493	RRE-ISI	
□1	Ara+4	547,701	559,507	RRE-ISI	
□1	Ara+6	547,701	556,319	RRE-ISI	
□1	Ara-1	547,701	555,877	RRE-ISI	
□1	Ara-3	547,701	550,351	RRE-ISI	
□1	Ara-4	547,701	619,884	RRE-ISI	
□1	Ara-5	547,701	558,574	RRE-ISI	
□1	Ara-6	547,701	589,555	RRE-ISI	
<b>590,472-601,096:</b>					
□2	Ara+1	590,445	601,096	RRE-IS150	<i>hokE</i> , <i>insL-3</i> , <i>entD</i> , <i>fepA</i> , <i>fes</i> , <i>ybdZ</i> , <i>entF</i> , <i>fepE</i>
□2	Ara-2	590,472	619,116	RRE-ISI	
□2	Ara-4	547,701	619,884	RRE-ISI	
<b>1,609,176-1,615,468:</b>					
□3	Ara+1	1,609,176	1,615,468	RRE-IS3	Qin-like locus: <i>ycbW</i> , <i>ynsB</i> , <i>ynfN</i> , <i>ECB_01516</i> , <i>cspI</i> , <i>ydfP</i> , <i>ydfQ-2</i> , <i>ydfR</i> , <i>essQ</i> , <i>ECB_01522</i> , <i>ECB_01523</i>
□3	Ara+6	1,609,176	1,615,468	RRE-IS3	
□3	Ara-2	1,609,176	1,615,468	RRE-IS3	
□3	Ara-3	1,609,176	1,615,468	RRE-IS3	
□3	Ara-4	1,609,176	1,615,468	RRE-IS3	
□3	Ara-5	1,609,176	1,615,468	RRE-IS3	

**2,037,724-2,045,407:**

□4	Ara+1	2,034,326	2,053,851	RRE-ISI	<i>wbbD, wbbC, wzy, wbbB, wbbA, vioB, vioA, wzx,</i>
□4	Ara+2	2,037,724	2,048,103	RRE-ISI	<i>rmlC, rfbA, rfbD, rfbB</i>
□4	Ara+3	2,031,432	2,054,725	RRE- <i>manB-cpsG</i>	
□4	Ara+4	2,031,432	2,054,725	RRE- <i>manB-cpsG</i>	
□4	Ara-1	2,031,703	2,054,996	RRE- <i>manB-cpsG</i>	
□4	Ara-2	2,031,432	2,054,725	RRE- <i>manB-cpsG</i>	
□4	Ara-3	2,032,711	2,056,011	RRE- <i>manB-cpsG</i>	
□4	Ara-4	2,031,432	2,054,725	RRE- <i>manB-cpsG</i>	
□4	Ara-5	2,031,432	2,054,725	RRE- <i>manB-cpsG</i>	
□4	Ara-6	2,034,326	2,045,407	RRE-ISI	

**2,100,286-2,122,432:**

□5	Ara+2	2,100,286	2,122,432	Unknown	Prophage 2 locus: <i>ogrK, yegZ, ECB_01989,</i>
□5	Ara+3	2,100,286	2,122,432	Unknown	<i>ECB_01990, ECB_01991, ECB_01992, ECB_01993,</i>
□5	Ara+4	2,100,286	2,122,432	Unknown	<i>ECB_01994, ECB_01995, ECB_01996, ECB_01997,</i>
□5	Ara+5	2,100,286	2,122,432	Unknown	<i>ECB_01998, ECB_01999, ECB_02000, ECB_02001,</i>
□5	Ara+6	2,100,286	2,122,432	Unknown	<i>ECB_02002, ECB_02003, ECB_02004, ECB_02005,</i>
□5	Ara-1	2,100,286	2,122,432	Unknown	<i>ECB_02006, ECB_02007, ECB_02008, ECB_02009,</i>
□5	Ara-2	2,100,286	2,122,432	Unknown	<i>ECB_02010, ECB_02011, ECB_02012</i>
□5	Ara-3	2,086,611	2,122,432	Unknown	
□5	Ara-4	2,100,286	2,122,432	Unknown	
□5	Ara-5	2,100,286	2,122,432	Unknown	

**3,001,956-3,015,762:**

□6	Ara+1	2,999,596	3,048,488	RRE-IS150	<i>flu, yeeR, ECB_02802, yafZ, ECB_02804, yeeS, yeeT,</i>
□6	Ara+2	2,996,588	3,023,945	RRE-ISI	<i>yeeU, yeeV, yeeW, ECB_02810, ECB_02811,</i>
□6	Ara-6	3,001,956	3,015,762	RRE-IS150	<i>ECB_02812, ECB_02813, ECB_02814, ECB_02815,</i>
					<i>ECB_02816</i>

**3,024,712-3,048,488:**

□7	Ara+1	2,999,596	3,048,488	RRE-IS150	CP-44-like locus: <i>ECB_02825</i> , <i>ECB_02826</i> , <i>ECB_02827</i> , <i>ECB_02828</i> , <i>yghD</i> , <i>yghE</i> , <i>ECB_02831</i> , <i>ECB_02832</i> , <i>ECB_02833</i> , <i>ECB_02834</i> , <i>ECB_02835</i> , <i>ECB_02836</i> , <i>ECB_02837</i> , <i>ECB_02838</i> , <i>yghF</i> , <i>yghG</i> , <i>pppA</i> , <i>yghJ</i> , <i>yghK</i>
□7	Ara+5	3,024,346	3,080,112	RRE-IS150	
□7	Ara-4	3,024,712	3,063,026	RRE-IS1	

**3,894,996-3,895,279:**

□8	Ara+1	3,894,996	3,901,134	RRE-IS150	<i>rbs</i> operon (34) <sup>e</sup>
□8	Ara+2	3,894,996	3,895,833	RRE-IS150	
□8	Ara+3	3,894,996	3,901,160	RRE-IS150	
□8	Ara+4	3,894,996	3,898,944	RRE-IS150	
□8	Ara+5	3,894,996	3,895,279	RRE-IS150	
□8	Ara+6	3,894,996	3,900,156	RRE-IS150	
□8	Ara-1	3,894,996	3,901,921	RRE-IS150	
□8	Ara-2	3,894,996	3,899,897	RRE-IS150	
□8	Ara-3	3,894,996	3,901,457	RRE-IS150	
□8	Ara-4	3,894,996	3,901,405	RRE-IS150	
□8	Ara-5	3,894,996	3,900,623	RRE-IS150	
□8	Ara-6	3,894,996	3,901,703	RRE-IS150	

**4,551,448-4,561,283:**

□9	Ara-3	4,522,340	4,561,283	RRE-IS1	<i>yjiV</i> , <i>mcrC</i> , <i>mcrB</i> , <i>yjiW</i> , <i>hsdS</i> , <i>hsdM</i> , <i>hsdR</i>
□9	Ara-6	4,551,448	4,573,236	RRE-IS1	

**2,760,570-2,771,648:**

D1	Ara+2	2,713,428	2,774,197	RRE-IS186	<i>ygbN</i> , <i>rpoS</i> , <i>nlpD</i> , <i>pcm</i> , <i>surE</i> , <i>truD</i> , <i>ispF</i> , <i>ispD</i> , <i>ftsB</i> ,
----	-------	-----------	-----------	-----------	--

D1	Ara+4	2,760,570	2,771,648	RRE-IS1	<i>ygbE, cysC, cysN, cysD, iap</i>
D1	Ara+5	2,734,828	2,774,455	RRE-IS1	
<b>666,130-2,034,326:</b>					
I1	Ara+1	4,615,673	2,774,435	RRE-IS150	<i>gltK-manC</i> , 1366 genes
I1	Ara+2	590,322	2,034,326	RRE-IS1	
I1	Ara-1	634,745	2,128,599	RRE-IS1	
I1	Ara-5	666,130	2,283,472	RRE-IS150	
<b>1,443,893-1,607,920:</b>					
Sub-region 1.1	Ara+3	1,443,893	1,615,474	RRE-IS3	<i>ydbC-ECB_01510</i> , 148 genes
Sub-region 1.1	Ara-2	1,443,893	1,607,920	RRE-IS3	
Sub-region 1.1	Ara-3	1,420,707	1,607,920	RRE-IS3	
Sub-region 1.1	Ara-4	1,443,893	1,607,920	RRE-IS3	
<b>2,651,000-3,015,762:</b>					
I2	Ara+2	2,651,000	4,189,422	RRE-ribosomal operon	<i>rrsG-ECB_02816</i> , 345 genes
I2	Ara-3	16,972	3,015,762	RRE-IS150	
<b>3,356,670-4,066,255:</b>					
I3	Ara+2	2,651,000	4,189,422	RRE-ribosomal operon	<i>rrsD-yiiG</i> , 667 genes
I3	Ara-2	3,270,936	4,066,255	RRE-ribosomal operon	
I3	Ara-4	3,354,888	4,189,422	RRE-ribosomal operon	
I3	Ara-6	3,356,670	4,187,735	RRE-ribosomal operon	

---

<sup>a</sup> The regions are numbered according to Fig. 1. The range of genomic coordinates for each region are shown in bold font, with positions given according to the genomic coordinates of the ancestral strain (67).

<sup>b</sup> All positions are given according to the genomic coordinates of the ancestral strain the ancestor genome (67).

<sup>c</sup> RRE, recombination between repeated elements, with the identity of the repeated element indicated after the hyphen.

<sup>d</sup> For parallel inversions, we give the name of the first and last genes affected followed by the total number of genes included in each inversion. For multiple inversions, we use the same rules to describe their coordinates as used in Table S1.

<sup>e</sup> The sizes of these deletions vary among the populations, and only some of them eliminate the entire operon (34).

**Feedback from the Editor of *mBio* and two**  
**anonymous reviewers**

Re: mBio01377-14 (Large Chromosomal Rearrangements during a Long-Term Evolution Experiment with *Escherichia coli*)

Dear Dr. Schneider:

We have completed our review of your manuscript, and I am pleased to inform you that, in principle, we expect to accept it for publication in *mBio*. However, acceptance will not be final until you have addressed the reviewer comments

The editor who handled your manuscript thinks that the modifications required are minor and that the work can be accepted if the reviewer comments are adequately addressed

Thank you for the privilege of reviewing your work. Below you will find instructions from the *mBio* editorial office and comments generated during the review.

Sincerely,

Arturo Casadevall

Editor in Chief, *mBio*

mBioEditorInChief@asmusa.org

(Signing for the editors)

Editor comments:

The mBio Editor in charge fully agrees with the decision of the Invited Editor for this manuscript. He would like to comment few points to be considered by the authors to produce the final version:

1. A main message of the manuscript is that a large number of rearrangements occurs along the long-term experimental evolution experiment, influencing the size and structure of the chromosome, "demonstrating the impressive plasticity of bacterial genomes" (line 21). However, in many microorganisms, bacterial clones isolated along many years in different places, probably resulting from very, very long-term natural evolution processes, show a remarkable chromosomal stability, in spite of exposure to different environments that could eventually select for particular conformations. In the author's classic experiment, the conditions of experimental propagation are fairly constant, and, unexpectedly, heavy plasticity is found. Might be the authors would like to include a comment this paradox, helping the reader to understand the possible generality of their findings. Additionally, such a comment would certainly satisfy Reviewer 1, who is asking for better explanation about the "beneficial effects" of evolved chromosomal changes.

2. Line 197: "Six of the 12 populations had evolved hypermutable phenotypes", however

(line 200-202) "there is no suggestion of any difference in the overall rate at which rearrangements accumulated in mutator versus non-mutator populations". To the knowledge of this Editor, a recent paper (Turrientes MC et al. 2013. Normal Mutation Rate Variants Arise in a Mutator (Mut S) Escherichia coli Population. PLoS ONE 8(9): e72963. doi:10.1371/journal.pone.0072963) found in a clinical E. coli mutator clinical strain submitted to serial passages, a high frequency of transposition events, and different inversion-patterns of chromosomal fragments were detected in evolved variants.

3. Lines 300-303: "Region  $\Delta 4$ , which encompasses the DNA between the manB and cpsG genes", "deletions occurred by recombination directly between these two genes, which share 96% nucleotide identity, resulting in the loss of all intervening genes". In occasions manB is confused with cspG (<http://www.uniprot.org/uniprot/P24175>), and there is "another" cpsG (cold shock protein G) gene unrelated with manB. Clarify.

4. Line 311. Probably "colanic", not "colonic" acid.

5. In relation with the above point 1, the authors should consider to mention (in the discussion) the paper from Moore JM. et al. 2012. Ann N Y Acad Sci. 2012 Sep;1267:103-9. doi: 10.1111/j.1749-6632.2012.06587.x. Gross chromosomal rearrangement mediated by DNA replication in stressed cells: evidence from Escherichia coli". They underline the possible effect of RecA, which is not mentioned in the manuscript under revision.

6. Also in relation with the above point 1, the authors might consider the paper from Lin D. et al., 2011. Global chromosomal structural instability in a subpopulation of starving



Escherichia coli cells. PLoS Genet. 2011 Aug;7(8):e1002223. doi: 10.1371/journal.pgen.1002223. A sentence in the discussion about the relation between stress and genomic stability could be appropriate -is the "long-term evolution experiment" performed by the authors a stressful experience for E. coli?

Invited Editor comments:

Your manuscript on "Large chromosomal rearrangements..." has now been reviewed by two experts in the field and myself. There is a general agreement that the results presented in your manuscript have general interest and are important as reference for evolutionary experiments, but as commented by reviewer 1 the text could be shortened considerable, which would help focusing on the parts with general information to the readers. Both reviewers point out important things which you should address carefully.

Reviewer #1 (Comments for the Author)

The authors describe the use of whole-genome optical mapping to identify large-scale chromosomal rearrangements in the genomes of E. coli that are part of a long-term evolution experiment.

The experiment is an important one, and these results are likely to be of interest, however the manuscript is very long and dry, with the minutiae of the rearrangements spelt out in great detail. I believe that the overview and general import of the results are likely to be of broad interest, but the fine details are very specific to this set of experimental conditions, and are likely to be of interest only to those who are invested in this particular experiment.

Specific points:

There is a tendency to see, or imply, selection and functional effects underlying each change, or at least to put this forward as the first explanation.

For example: I have some concerns about the discussion of inversions, which starts out with quite strong statements, but is heavily caveated towards the end of the discussion. At line 190, the authors state "In seven cases ... more than a quarter of the chromosome was affected ...". Can you really say that all these genes are "affected" by the inversion in any meaningful way? As the authors themselves admit later on, the vast majority of these genes will have no meaningful change in local or global context (as most are reciprocal around the origin or terminus), and are unlikely to be affected in any way. Therefore the fact that they are affected in multiple strains may be simply chance (especially given the size of the inversions). I think this needs to be re-worded in a more neutral way.

In lines 477-481, the authors describe the apparent lack of constraint on gene order in the LTEE, suggesting that horizontal transfer may generate this constraint in the wild. Another

possibility is that selection on gene order is weak, and only effective over very long timescales. A similar example would be selection for codon usage - this is clearly apparent over very long timescales, but there is no evidence it is detectable on the timescale of the LTEE.

Lines 316-317, discussing prophage deletions: "Most of these genes have unknown functions". Given that they are within prophage, it's highly likely that most of these genes have no function at all for the E. coli.

Lines 338. Seeing repeated inversions between rRNA genes is not surprising - these are frequent in nature (certainly in Salmonella), and given the size of the perfect repeats, they may just happen more frequently than other rearrangements, and thus be seen more often.

All these points belie the statement in line 439 (which recurs throughout the manuscript), that events affecting multiple locations provide "indirect evidence that these mutations were beneficial". This may not be true - they may just be very large and neutral, or smaller and simply occur more frequently, and this needs to be discussed more rigorously.

Reviewer #2 (Comments for the Author):

The paper provides a detailed analysis of all large chromosomal rearrangements that have

appeared in 12 independently evolving populations of *E. coli* propagated in shake-flasks for ~40,000 generations.

The investigations on the 12 populations have already led to numerous papers addressing questions in evolutionary biology, and as such the experiment constitutes an important point of reference within the field of experimental microbial evolution. While genome sequences of clones from eight of the 12 populations have previously been analyzed by Barrick et al. (*Nature* 461(7268), 2009) and Wielgoss et al. (*G3* 1(3), 2011), this is the first whole-genome comparison that include clones from all 12 populations. Furthermore, as the authors choose to focus their analysis on large chromosomal rearrangements, they use optical mapping on genomes from all 12 populations to complement the short-read whole-genome sequence data, and in particular the optimal mapping enables the identification of inversions involving long sequence repeats.

All together, this is an immense amount of genomic data, but the authors manage to present it in a clear and efficient manner. Accordingly, the study is successful in giving a general overview of large chromosomal rearrangements that have accumulated in clones from the 12 populations after 40,000 generations of growth. Also, genomic information of longitudinal collected isolates from some of the populations is analyzed to determine the succession and dynamics of rearrangements over time. Finally, the authors identify a high level of parallel evolution across the populations, and this suggests that these rearrangements have been positively selected. Some of the findings have already been reported previously (for example parallel loss of the *rbs* operon), but the authors are clear and transparent about this.

Overall, the study gives an overview of large chromosomal rearrangements occurring in a

long-term evolution experiment, and its findings may become of relevance as more studies employ similar experimental and analytical techniques to explore genetic variation beyond identification of point mutations. However, the study is very descriptive in its nature, and many of the hypotheses are not substantiated by further experiments.

The text reads nicely, and evolutionary theory is well presented.

Additional suggestions and comments:

(1) Line 175: It is not possible to identify IS element insertions using optical mapping. However, why are the short Illumina reads not used to detect such events? Barrick et al. uses the exact same type of data (36-nt single-end reads) to identify 10 IS element insertions in the Ara-1 population (Nature 461(7268)).

(2) Line 177: The study concerns large deletions and insertions. However, it is not stated what is the lower limit of "large" deletions and insertions? For example, would a 300-bp deletion be reported in this study?

(3) Line 248: There are three cases of imperfect duplications with deletions of the conjoined regions between duplicated copies. This is an interesting observation, could authors speculate about possible selective advantages from such genetic constructions? In line 441 authors discuss the intrinsic instability of duplications - could it be that the imperfect duplications are more stably maintained in the chromosome?

(4) Line 458: It is understandable that it is difficult to predict the fitness of the inversions due to the large number of affected genes. Nonetheless, in some cases it might be relevant to focus on the genes at the flanks of the inversions in cases where the promoters (or other regulatory elements) of genes are replaced. It could be interesting if it could be predicted that some of the inversions changed transcriptional regulation of genes?

Staff comments:

Please submit your modified manuscript through the following link: <http://mbio.msubmit.net/cgi-bin/main.plex?el=A2FE7chE4A6GqE3I3A9ftdq059VOGUEiNXKL8iZu8udgZ>. You should include point-by-point responses to the issues raised during review in a file labeled "Response to Reviewer Comments."

Please return the manuscript within 30 days; if you cannot complete the modification within this time period, please contact the editorial office at [mbio@asmusa.org](mailto:mbio@asmusa.org). If you do not wish to modify the manuscript and prefer to submit it to another journal, please notify the editorial office of your decision immediately so that the manuscript may be withdrawn from consideration by mBio.

Since your manuscript is likely to be accepted, please be sure to upload all of your native files

(i.e., not a single PDF) in production-ready formats. You can check your files using the Rapid Inspector tool at <http://rapidinspector.cadmus.com/RapidInspector/zmw/index.jsp>, and you can direct questions to our digital art specialists at [figures@asmusa.org](mailto:figures@asmusa.org). This will help us avoid processing delays and move quickly to online publication of your important work.

Please contact us if you have any questions.

Maisha Miles

Managing Editor, mBio

Rob Arthur

Assistant Managing Editor, mBio

[mBio@asmusa.org](mailto:mBio@asmusa.org)

American Society for Microbiology

1752 N St. NW

Washington, DC 20036

(202) 737-3600

(202) 942-9355, fax

# Results Part II:

**Article:** Darwinian selection drives the dynamics of IS elements during bacterial long term evolution

**State:** To be submitted to *PNAS*



BIOLOGICAL SCIENCES: Evolution

## **Darwinian selection drives the dynamics of IS elements during bacterial long term evolution**

Colin Raeside<sup>a,b</sup>, Joël Gaffé<sup>a,b</sup>, Richard E. Lenski<sup>c,d</sup>, Olivier Tenaillon<sup>e,f</sup>, and Dominique Schneider<sup>a,b,1</sup>

<sup>a</sup>Univ. Grenoble Alpes, Laboratoire Adaptation et Pathogénie des Microorganismes (LAPM), F-38000 Grenoble, France; <sup>b</sup>Centre National de la Recherche Scientifique (CNRS), LAPM, F-38000 Grenoble, France; <sup>c</sup>Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA; <sup>d</sup>BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, MI 48824, USA; <sup>e</sup>IAME, UMR 1137, INSERM, F-75018 Paris, France; <sup>f</sup>IAME, UMR 1137, Université Paris Diderot, Sorbonne Paris Cité, F-75018 Paris, France

Short title: Long-term dynamics of bacterial IS elements.

To whom correspondence may be addressed. dominique.[schneider@ujf-grenoble.fr](mailto:schneider@ujf-grenoble.fr)

## Abstract

Insertion Sequences (IS) are mobile genetic elements ubiquitous in bacterial genomes where they have profound impact on both structure and expression. Like point mutations, they produce harmful, neutral and beneficial effects on organismal fitness. In contrast however to mutation rates, whether and how the balance between these opposing effects affect the long-term dynamics of IS elements in bacterial genomes is unknown. Moreover, their distribution and copy number are highly variable, for unclear reasons. Therefore, it is still controversial whether IS elements are genomic parasites maintained through high transposition rates and genetic exchange or major evolutionary drivers under positive selection. Here, we investigated the long-term dynamics of IS elements and their contribution to the genomic mutagenesis during the longest-running evolution experiment, where twelve populations are propagated from an *Escherichia coli* ancestor for over 60,000 generations. We demonstrated that IS elements accounted for 15 to 50% of the total mutations substituted during 40,000 generations of evolution. In one population, we observed a substantial ~6-fold increase in the copy number of one IS type, *IS150*. By experimentally checking three evolutionary scenarios and using an evolutionary model, we show that this genomic expansion of *IS150* was most likely associated with early positive effects on fitness. Later during evolution, when bacterial cells were adapted to their environment, this expansion was compensated by a down regulation of *IS150* transposition activity, suggesting domestication. Therefore, the long-term dynamics of IS elements, as for mutation rates, involves a balance between robustness and evolvability.

**Keywords:** Insertion Sequences | experimental evolution | *Escherichia coli* | copy number dynamics

## Significance Statement

Insertion Sequence (IS) elements are bacterial mobile genetic elements that generate diversity by either moving within genomes or being substrates for homologous recombination. Their distribution is highly variable in bacterial species and IS-related mutations confer detrimental to beneficial fitness effects. Invasion of genomes by IS elements is driven by evolutionary forces that are mostly unknown, but likely involving a balance between these contrasting effects. We investigated the evolutionary dynamics of IS elements during long-term evolution with *Escherichia coli*. Genome invasion by the IS150 type of element is driven by its fitness-enhancing effects, followed later during evolution by a reduction of transposition activity, most likely owing to a higher proportion of deleterious mutations when cells are well adapted to their environment.

Microbial organisms evolved several mechanisms, including the ability to produce mutations, the global regulation of gene expression, the activity of mobile genetic elements and genetic exchange, to both promote phenotypic and genetic diversity and adapt to diverse environments. The pace of evolutionary processes involves, both at short and longer time scales, a balance between evolvability and robustness (1–4) that manifests itself in the fine-tuned dynamics of the molecular pathways underlying these mechanisms. While this balance has been deeply investigated in the case of mutation rates and regulatory networks, it is less well understood for the dynamics of mobile genetic elements. For instance, the forces driving the ability of bacterial Insertion Sequence (IS) elements to colonize genomes over long evolutionary times are still mostly unknown.

Both theory and experiments have shown that mutation rates are highly dynamic. A high proportion of hypermutator strains, owing to defects in DNA repair genes, evolved both in clinical settings and during propagation in the laboratory (5–10). By producing a higher number of beneficial mutations, hypermutators increase in frequency in populations subject to changing environments (11, 12). However, they also increase the genetic load by producing more deleterious mutations. Hence, after a first rise of mutation rates, compensatory processes can occur to decrease mutation rates resulting in higher fitness once populations are adapted (4, 13).

Global regulatory and metabolic networks have highly dynamic topologies that allow exquisite and fast adjustment of genome expression after environmental challenges (14). Both microbial evolution experiments and natural isolates revealed that cellular networks are evolvable at a longer time scale, being rewired after the occurrence of mutations that increased organismal fitness (15). Many theoretical and empirical studies highlighted the pleiotropy, epistasis, plasticity and modularity of these networks. In particular, modularity

can provide robustness when pleiotropic effects are deleterious (16). Therefore, both mutation rates and cellular networks result in a balance between stability and evolvability.

IS elements are small (~0.7 to ~2.5 kbp) and highly diverse mobile genetic elements, present in almost all bacterial genomes (17). All IS elements encode functions devoted only to their transposition and its regulation, including one to three Open Reading Frames producing the transposase enzyme and in most cases terminal inverted repeats (IRs). Upon insertion, they can generate a direct duplication of a few base pairs of their target site. Both their activity and presence produce major effects on bacterial fitness because they impact on genome structure and expression (18, 19). Indeed, IS transposition can inactivate gene expression but also result in gene deregulation owing to the presence at their extremities of outwardly directed promoters or transcriptional regulatory elements. In addition, homologous recombination between IS copies can produce large chromosomal rearrangements (inversions, deletions, duplications). IS elements are characterized by a very high variability in their type and frequency, as well as in the distribution of different IS families and the copy number of individual IS elements, among bacterial genomes (20). They can have two conflicting effects on bacterial fitness, thereby resulting in debates about their evolutionary significance. First, transposition of IS elements can have detrimental effects and are considered as genomic parasites, maintained mainly through high rates of transposition and acquisition by horizontal transfer (21). As a consequence, transposition is highly regulated by both the element itself and host factors (18). Second, IS elements may generate genetic diversity on which natural selection can act. During laboratory evolution experiments, IS elements have been shown to produce fitness-enhancing mutations under diverse environmental conditions (22–27), and may therefore be under positive selection. Moreover, IS elements have been suggested to be involved in reductive genome evolution (28). However, it is mostly unknown whether and how the tension between the detrimental and

beneficial effects of IS elements can impact on the dynamics of their copy number distribution inside genomes of single species. One reason is that the long-term dynamics of IS copy number during organismal evolution is difficult to analyze.

To address these questions, we used the longest-ongoing evolution experiment in which 12 independent populations have been founded from a common *E. coli* ancestor and propagated for over 60,000 generations in a defined environment with a limited amount of glucose as a carbon source (29). Two populations were previously analyzed for the IS distribution after 10,000 generations, revealing an unexplained strong increase in copy number of the *IS150* element in one of them (30). Here, we investigated the distribution, dynamics, regulation and overall mutagenesis contribution of IS elements in the 12 populations during 40,000 generations. We found a specific and strong increase in *IS150* copy number in one population, called Ara+1, in which IS elements accounted for ~50% of all genomic mutations. We investigated three scenarios for this high copy number increase: mutational alteration of *IS150* itself, deregulation of transposition, and fitness-related effects. Both experiments and a simulation model revealed that Darwinian selection accounted for the copy number dynamics, followed by changes in transposition regulation most likely due to a higher number of deleterious events later on when bacterial cells are adapted to their environment.

## **Results and Discussion**

**Distribution of IS elements after 40,000 generations of evolution.** We previously examined the distribution of the seven types of IS elements present in the ancestral genome (*IS1*, *IS2*, *IS3*, *IS4*, *IS30*, *IS150* and *IS186*) in two populations (Ara-1 and Ara+1) after 10,000 generations (30). Here, we compared the IS fingerprints of 28 evolved clones sampled

in all 12 populations at 40,000 generations (Table S1) to the ancestor, revealing a total of 210 unique IS copies (Table S2). *IS1* and *IS150* were active in all 12 populations while changes related to *IS3*, *IS4* and *IS186* were detected in 8, 3 and 11 populations, respectively (Table S3). No changes were detected for *IS2* and *IS30*. The number of IS-related changes was different in each population and unrelated to their ancestral distribution ( $X^2 = 1.09 \times 10^{-57}$ , Fig. S1). The phylogenetic tree, based on these IS fingerprints, revealed a high genetic diversity allowing to distinguish each individual evolved clone (Fig. S2). However, clones from identical populations grouped together with two noticeable exceptions: Ara-2 which contains two ecotypes (31), and Ara-6 which therefore may also contain at least two lineages. A previous study indeed showed that co-existing lineages may be present in several populations (32).

**Contribution of IS elements to total mutations during 40,000 generations.** We next scrutinized the genome sequences of the same 28 evolved clones to compute the proportion of IS changes relative to all mutations. The total number of mutations was inferred from the genome sequences and the number of IS-related changes from both the IS fingerprints and genome data for comparative purposes (Fig. 1). Similar results were obtained with both datasets except for Ara-5 where 41 and 90 IS-related changes were detected for the 3 analyzed clones by genome sequencing and IS fingerprinting, respectively. We distinguished three groups of populations (Fig. 1): first, populations Ara+3, Ara+6, Ara-1, Ara-2 and Ara-4 showed a low contribution (< 5%) of IS-related changes, reflecting their mutator status and therefore a high proportion of SNPs. Second, in populations Ara+2, Ara+4, Ara+5, Ara-3, Ara-5 and Ara-6, IS-related changes accounted for ~10 to ~30% of all mutations. Third, population Ara+1 was unique in that IS-related changes accounted for almost 50% of all mutations. The proportion of IS-related changes had a median value of 0.28 in the non-

mutator populations, and only Ara+1 caused a distortion toward a higher ratio (Fig. S3). These data may suggest differential activities of IS elements in Ara+1.

**Dynamics of IS-related changes in Ara+1 over time.** The distribution of IS-related changes in Ara+1 was almost similar when comparing the genome sequences and IS fingerprints (Fig. 1). We therefore scrutinized the genome sequences of 20 evolved clones sampled over time from population Ara+1 (Table S1) to determine the number of IS-related changes and their contribution relative to the total mutations (Fig. 2). The number of IS-related changes increased almost linearly over time, with most changes attributed to *IS150* (Fig. 2A). The proportion of IS-related changes relative to total mutations was very high and dynamic, increasing strongly to over 50% during the first 5000 generations and then stabilizing (Fig. 2B). Again, most of this dynamics involved *IS150*.

Three non-exclusive hypotheses may explain this strong increase in *IS150* copies during evolution: changes in the *IS150* sequence may produce an over-active copy; changes in *IS150* regulation owing to mutations substituted during Ara+1 evolution may increase its transposition frequency; initial *IS150* transposition events may enhance cell fitness resulting in subsequent increased copy number and thereby transposase amount.

**Sequence and transposition frequencies of *IS150* during evolution of Ara+1.** We scrutinized the genome sequences of the 40,000-generation clones from population Ara+1 (Table S1) and found neither point mutations nor indels inside *IS150*. Moreover, we PCR-amplified and sequenced 19 of the 29 new *IS150* copies from the evolved clone 11008 using primers complementary to their flanking regions (Table S4). We found no changes compared to the ancestral copy.

We measured transposition frequencies of *IS150* by introducing the reporter plasmid pFDX2339 (33) in the ancestor and five evolved clones from population Ara+1 (Table S1):



one at 2000 (1158C), one at 20,000 (9282B), and three at 40,000 generations (11008, 11009, and 11010). This thermosensitive plasmid carries a modified *IS150* element and allows to estimate the frequency of both recombination between the plasmid-borne *IS150* and any chromosomal copy, which we used as a proxy to account for the increase in *IS150* copy number, and *IS150* transposition from the plasmid into the chromosome. As expected, recombination frequencies increased over time with *IS150* chromosomal copies (Fig. 3A). Although transposition frequencies increased by ~50-fold up to 20,000 generations, it likely reflected the copy number increase, and therefore the higher transposase amount, since the proportion of transposition events relative to recombination events was stable (Fig. 3B). Therefore, we detected no measurable changes in the regulation of *IS150* transposition during 20,000 generations of evolution in Ara+1 that may explain the strong increase in *IS150* copy number. The mutations that were substituted during 20,000 generations did therefore not interfere with *IS150* transposition.

In striking contrast, *IS150* transposition frequencies relative to recombination decreased by 2 to 10 fold in all three 40,000-generation clones (Fig. 3B). Therefore, after the initial strong increase in *IS150* copy number, transposition decreased later on, implying that mutations substituted between 20,000 and 40,000 generations may affect transposition and likely reflecting detrimental effects of new *IS150* copies. Hence, after 20,000 generations, the *IS150* copy number is very high and cells are adapted to their environment. Any new transposition events may therefore produce deleterious mutations. This transposition dynamics is reminiscent to what was observed for point mutation rates (4, 13).

**Fitness effect of early *IS150* dynamics in Ara+1.** No changes in either *IS150* sequence or transposition frequency can explain the initial strong copy number increase. The genome sequence of the 2000-generation clone 1158C, which was shown to be on the line of descent of population Ara+1, revealed already the presence of three new *IS150* insertions and two

IS150-related deletions. The three new insertions accounted for one third of all mutations (Table S5). One of the two deletions removed *rhs* genes involved in ribose consumption and was shown to be beneficial in these conditions (25). (A similar situation was found in another 2000-generation clone called 1158A, Table S5). The new insertion IS150::*ybeB* was previously shown to confer a fitness increase of 1.176 (34,  $H_0 = 1$ ; *t*-test,  $t_s = 18.0391$ ;  $P = 0.0001$ ). The two other IS150 insertions occurred within *nadR* and *hokB*. We constructed a deletion of each of these genes within the Ara<sup>-</sup> ancestral chromosome, and competed each mutant strain against an Ara<sup>+</sup> derivative of the ancestor. These competition assays showed that inactivation of *nadR* conferred a fitness benefit of 1.14 relative to the ancestral allele ( $H_0 = 1$ , *t*-test,  $t_s = 4.9880$ ,  $P = 0.0076$ ), while inactivation of *hokB* was neutral ( $H_0 = 1$ , *t*-test,  $t_s = 0.4090$ ,  $P = 0.6920$ ). The sum of the fitness effects of both *nadR* and *pbpA* evolved alleles accounted for the estimated fitness of the clone 1158C (34, 1.243,  $H_0 = 1$ ; *t*-test,  $t_s = 15.0853$ ;  $P < 0.0001$ ). Thus, the initial new IS150 copies conferred very high fitness to the evolved clones, and the initial copy number increase from 5 in the ancestor to 8 after 2000 generations was likely driving the later increase. The initial copy number increase may produce IS150 transposase levels high enough to further increase transposition activity (33).

**Simulation of the dynamics of IS elements.** We constructed an evolution model that includes both point mutations and IS-related changes. One hundred replicate populations were propagated for 20,000 generations under each of 18 parameter sets: three mutation rates ( $\mu^m$ ) similar to those in the long-term evolution experiment (4), three rates of IS-related changes ( $\mu^{IS}$ ) for each mutation rate, and impact or not of IS elements on their transposition rates in each case. Without impact of IS elements on transposition, their proportion relative to point mutations after 20,000 generations were as expected from their initial transposition rates (Fig. 4). By including an impact, no significant contribution of IS-related changes relative to point mutations was detected at the lowest  $\mu^{IS}$  (0.2) whereas a significant

difference was observed at higher  $\mu^{IS}$  (0.3) for one of the three  $\mu^m$  and for all three  $\mu^m$  at the highest transposition rate (0.5), thereby only at high initial  $\mu^{IS}$ .

We analyzed into deeper details the simulation with  $\mu^m = 2 \times 10^{-5}$  and  $\mu^{IS} = 0.3$  (Fig. S4). Whether IS elements impact or not on  $\mu^{IS}$ , all 100 replicate populations revealed similar fitness trajectories, mimicking the 12 long-term *E. coli* populations (29). Three of the 100 replicate populations revealed both a higher number of IS-related changes including a gradual increase of these changes and a higher proportion of IS-related changes relative to total mutations after 20,000 generations that started early during evolution (Fig. S5,  $P = 0.002$ ), reminiscent to the IS trajectories in population Ara+1 (Fig. 2). We investigated the fitness effect of the five initial mutations, including 1 point mutation and 4 IS-related changes for each of the three populations. Not only were all IS-related changes beneficial, but they also conferred over the three populations 85% of the fitness increase after the first 5 mutations (and 82% after 20,000 generations), for an average of 36% over all replicates as expected for  $\mu^{IS} = 0.3$ . In 84 other populations, the 5 initial mutations also included IS insertions, but less than four in all cases. One single exception was observed with one replicate population showing 4 IS-related changes among the 5 initial mutations without a later massive IS expansion. However, IS-related changes accounted for 60% of the mutations in this replicate. In this case, IS-related changes conferred 74% of the fitness increase after the first 5 mutations (and 70% after 20,000 generations).

Although simple, this model suggests that an initial increase of IS elements, selected owing to their positive effect on fitness, can drive a highly dynamic IS trajectory with a strong expansion during later evolution without any changes in either the IS nucleotide sequences or their regulation. It also suggests a threshold level in both the initial copy number increase and fitness effect above which the later expansion can occur.

**Conclusions.** While IS elements have been considered as genomic parasites (21) with harmful effects, evolution experiments provided evidence that IS elements can persist within bacterial genomes by generating beneficial mutations (22–27). However, the dynamics of these opposing effects during long-term evolution is mostly unknown. In particular, how IS elements can invade bacterial genomes without extinction of their host is unclear.

Here, we investigated the dynamics of IS elements in 12 bacterial populations that evolved over tens of thousands of generations. IS elements generated a high level of genetic diversity and their contribution to total mutations was high from 15 to as much as 50% in non-mutator populations. One population experienced a large expansion of one type of IS elements, *IS150*. This expansion was not related to evolutionary changes in either the sequence of *IS150* itself or its transposition regulation, but rather to the initial new copies that conferred most of the fitness increase during the first 2000 generations. This initial increase in *IS150* copy numbers, associated with positive selection, likely drove the subsequent boost of transposition of *IS150* by increasing the intracellular transposase concentration. Expansion of IS elements has been commonly reported for bacteria with host-restricted lifestyles or pathogenic bacteria, and has been suggested to be involved in the first steps of genome reduction and in the improvement of bacterial ability to fight host defences (18, 28). We also recently showed that the expansion of *IS150* was associated with profound restructuring of the chromosome in population Ara+1 including successive large chromosomal inversions by recombination between *IS150* copies (mBio manuscript).

The *IS150* expansion was followed, after 40,000 generations and a 6-fold increase in copy number, by a reduced frequency of *IS150* transposition, which might indicate *IS150* domestication by these bacterial cells (18). This behaviour is reminiscent to the dynamics of mutation rates during long-term evolution, including a first strong increase followed by compensation to reduce genetic load (4, 13). Therefore, the expansion of IS elements during

the initial evolution steps was likely driven by the fitness-enhancing effect of the first IS copies, followed much later by a reduced transposition activity to compensate for the deleterious effects of further transposition events. We hypothesize that this transposition compensation may result from changes in transposition regulation owing to mutations substituted during the evolution of the Ara+1 population. Whether these mutations were selected for their direct effects on transposition or other effects, with transposition reduction being merely an associated trait, is unknown. The population Ara+1, although retaining the low ancestral mutation rate, can be considered as a mutator population owing to the dynamics of IS150.

The long-term evolution experiment provided enough time to observe the dynamics of IS elements, and in particular how they can initially invade bacterial genomes and be domesticated later on.

## **Materials and Methods**

**Bacterial strains and culture conditions.** Twelve populations, called Ara+1 to Ara+6 and Ara-1 to Ara-6, were founded from a common ancestor of *E. coli*, REL606 (35), and serially propagated for over 60,000 generations in a minimal medium (DM25) containing 25 µg/mL glucose (36). Samples from each population were collected at 500-generation intervals and stored at -80°C. Individual clones can be isolated directly from these frozen fossil samples after plating on solid medium. Here, we used the ancestor, two evolved clones sampled from each population after 40,000 generations, two additional clones from Ara-2, one additional clone from Ara+1, Ara-1 and Ara-5, and two clones from Ara+1 at each of the time points 500, 1000, 1500, 2000, 5000, 10,000, 15,000, 20,000, and 30,000 generations (Table S1). All strains were grown in LB liquid cultures or onto LB-agar (12 g/L) plates supplemented when

relevant with kanamycin (Kan, 30 µg/mL) and/or chloramphenicol (Cam, 40 µg/mL). Selection for sucrose-resistant clones during allelic exchange experiments was performed by plating cells onto sucrose plates in which 5% sucrose but no NaCl was added to LB-agar. To distinguish Ara<sup>-</sup> from Ara<sup>+</sup> clones, cells were spread onto tetrazolium-arabinose indicator plates (34).

**IS fingerprinting.** IS fingerprints were obtained for a total of 29 strains: the ancestor REL606 and 28 evolved clones that were sampled from the 12 populations at 40,000 generations (Table S1). Strains were grown in LB medium for 12 h at 37 °C. Their genomic DNA was extracted by standard methods and quantified spectrophotometrically. Genomic DNA (5 µg) was digested by restriction enzymes that do not cut within the DNA fragments used as probes for each IS element: *EcoRV* for the fingerprints of IS1, IS2, IS3, IS30 and IS186, and *HincII* and *PvuII* in the case of IS4 and IS150, respectively. The resulting fragments were separated by agarose gel electrophoresis and transferred onto nylon membrane. Southern blot hybridizations were performed at high stringency (68 °C) using internal fragments of IS elements as probes. Probes were produced by PCR of REL606 genomic DNA using primer pairs specific to each IS element (37), and then cold-labelled using the DIG system (Roche) according to the manufacturer's protocol. IS fingerprints were obtained after immuno-detection using an anti-DIG antibody, washing steps and exposition of the membranes to autoradiography films. Each IS copy appears as a hybridizing-band and the 29 strains were scored for the presence or absence of each copy. We detected a total of 210 unique IS copies yielding therefore a matrix of 29 strains and 210 fragments over the seven IS elements (Table S2).

**Phylogenetic analysis.** The phylogeny of all evolved clones was constructed as a Neighbour-Joining tree (38), with the root being the ancestor, and plotted using the DRAWGRAM program (39). Genetic distances were computed as the total number of differences between

each evolved clone and their ancestor. Losses and gains of IS copies have equal weights and simultaneous loss and gain of hybridizing bands from single mutational events were not accounted for.

**PCR experiments for probe design and sequencing of new *IS150* copies.** Polymerase chain reaction (PCR) was performed in a 25- $\mu$ L reaction buffer containing 3mM MgCl<sub>2</sub>, 0.2mM dNTP, 0.2mM of each primer and 1.25 unit of Taq polymerase. The PCR products were separated by agarose gel electrophoresis, purified using the Qiagen Gel purification kit (Qiagen) and sequenced (GATC Biotech) with the same primers. We used primer pairs to construct the probes specific to all IS elements (37) or sequence new copies of *IS150* (Table S4). All sequences were compared to the *IS150* sequence present in the genome of the ancestor REL606 (35).

**Measure of the transposition frequencies of *IS150*.** Cells from the ancestor and five evolved clones from population Ara+1 (1158C, 9282B, 11008, 11009 and 11010, Table S1) were electro-transformed with the pFDX2339 reporter plasmid (33) using the GenePulser II equipment (BioRad) according to standard procedures. This plasmid carries a thermosensitive replication origin, allowing its maintenance at low but not high temperatures, a gene conferring Kan resistance for selection of plasmid-carrying cells, and a modified *IS150* element bearing a gene conferring Cam resistance (33). At high temperature, three scenarios may happen: first, the plasmid and its modified *IS150* may be lost; second, the entire plasmid may integrate in the chromosome by recombination between its own *IS150* copy and one chromosomal *IS150*; third, the plasmid-borne *IS150* may transpose into the chromosome while the rest of the plasmid is lost. Each alternative can be distinguished by plating cells onto different selective media.

Each plasmid-carrying strain was grown in 2 mL LB-Kan at 28 °C for 48 h, diluted and plated onto two sets of agar plates, LB-Kan-Cam and LB-Cam, which were then incubated at 28 °C and 42 °C, respectively. We scored the total number of cells carrying pFDX2339 ( $N_T$ ) as the number of colonies growing on the first set of plates, and the total number of cells that integrated into the chromosome either the entire plasmid by recombination ( $N_R$ ) or the modified *IS150* copy by transposition ( $N_{IS}$ ) as the number of colonies growing on the second set of plates. To distinguish between the two last types of events, colonies that grew at 42 °C on LB-Cam plates were streaked onto both LB-Kan-Cam and LB-Cam plates and incubated overnight at 42 °C. Growth onto both plates gave ( $N_R + N_{IS}$ ) while growth onto only LB-Cam gave  $N_{IS}$ . Finally, the proportion of cells that experienced *IS150*-mediated recombination and transposition was calculated as the ratios  $N_R/N_T$  and  $N_{IS}/N_T$ , respectively.

**Contribution of IS elements relative to total genomic changes.** The proportion of IS-mediated changes was calculated as the ratio (number of IS-mediated mutational events) / (total number of mutations). The mean of the values obtained for the evolved clones from the same population was then calculated. The number of IS-mediated changes was derived from both the IS fingerprints and the genome sequences. The total number of mutations sums the number of SNPs identified in its genome sequence and the number of IS-mediated changes. We gave equal weight to losses and gains of IS copies and we did not adjust for simultaneous loss and gain of hybridizing bands from single mutational events.

**Construction of isogenic mutant strains and competition experiments.** We constructed an in-frame deletion of each of the *nadR* and *hokB* loci in the ancestral chromosome by allelic exchange. All gene replacement experiments were performed by using suicide plasmid pKO3 as described (34). To construct *in vitro* the 300-bp *hokB* in-frame deletion, we PCR-amplified the *hokB* upstream and downstream regions using primer pairs G125 (5'-



accggtcgattctgataagc-3')/G236 (5'-cagaaagtcgacaggcacccg-3') and G126 (5'-agtcgtggcatgaaacgctg-3')/G237 (5'-cgcgagtcgactagcagattc-3'), respectively, and fuse them to generate the deleted allele. For *nadR*, we PCR-amplified *nadR* with its adjacent regions with primer pair G122 (5'-gtctggcaatcctgctggct-3')/G153 (5'-ggtcacccgtgcagggtt-3'), followed by a *NruI/HincII* digestion which generates a 522-bp *nadR* in-frame deletion.

Pairwise competition experiments were performed to estimate the fitness effects of the deleted *nadR* and *hokB* alleles. In all cases, the competing strains also differed by the neutral Ara marker, which allowed them to be distinguished by colony color on tetrazolium-arabinose plates. To ascertain the effect of the deleted alleles, the Ara<sup>+</sup> variant of the ancestral strain, REL607, was used in competition experiments against each of the constructed strain. The fitness of one competitor relative to the other was calculated as the ratio of their net growth rates during competition (36). We used *t*-tests to evaluate whether the mean fitness differed from the null hypothetical expectation of 1 for each set of assays.

**Model of evolution and IS dynamics.** The code of the model that we developed has been deposited in the Dryad Digital Repository (doi: deposition pending acceptance). The model includes an initial simulated population of  $10^7$  digital organisms with each individual having a fitness of 1. The population experienced seasonal cycles consisting in a 100-fold expansion before going through a bottleneck to be reduced back to  $10^7$ . During each cycle, the fitness ( $F$ ) of each individual can be altered by substitution of a mutation among a pool of 20,000 with effects that are governed by random selection (Table 1). Organisms with higher  $F$  have a faster doubling time and can invade the population. We included in our model point mutations occurring at a rate  $\mu^m$  and IS insertions occurring at a transposition rate  $\mu^{IS}$  divided by the number of initial IS elements that was set to 5 which corresponds to the ancestral copy number of IS/50. Values of mutation and transposition rates were varied:  $\mu^m$  was set to either  $1.5 \times 10^{-5}$ ,  $2 \times 10^{-5}$  or  $3 \times 10^{-5}$  and  $\mu^{IS}$  to ratios of 0.2, 0.3 and 0.5 for each  $\mu^m$  value, resulting

in a total of 9 different parameter sets. In addition, the model allows, or not, IS elements to increase their transposition rate ( $\mu^{IS} / n$ ), where  $n$  is the IS initial copy number. One hundred replicate populations were propagated with each parameter set, leading therefore to a total of 1800 populations (100 x 9 parameters x 2 IS insertion effects). Each population was sampled at 100-generation intervals for 20,000 generations and, for both the entire population and the dominant sub-population, we scored the number of mutations, number of new IS copies, mutation order, and  $F$ .

**Acknowledgments.** This work was supported by grants from the European Union program FP7-ICT-2013-10 project EvoEvo grant 610427 to D.S., the Université Grenoble Alpes to D.S., the Centre National de la Recherche Scientifique to D.S., the XXXX. C.R. thanks the French Ministry of Education and Research for a doctoral fellowship.

**Author Contributions.** D.S. designed the research; C.R., J.G., and O.T. performed the research; J.G., and O.T. contributed analytic tools; C.R., J.G., R.E.L., O.T., and D.S. analyzed the data; O.T., R.E.L., and D.S. wrote the paper.

The authors declare no conflict of interest.

This article contains supporting information online at [www.pnas.org](http://www.pnas.org)

Data deposition: The code for the evolution and IS dynamics model has been deposited at the Dryad Digital Repository (doi: deposition pending acceptance). Genome sequence data have been deposited in the ENA Sequence Read Archive (XXXX).

1. Kirschner M, Gerhart J (1998) Evolvability. *Proc Natl Acad Sci USA* 95(15):8420–8427.
2. Earl DJ, Deem MW (2004) Evolvability is a selectable trait. *Proc Natl Acad Sci USA* 101(32):11531–11536.
3. Wagner A. (2005) *Robustness and Evolvability in Living Systems* (Princeton: Princeton University Press).
4. Wielgoss S, et al. (2013) Mutation rate dynamics in an *Escherichia coli* population reflect tension between adaptation and genetic load. *Proc Natl Acad Sci USA* 110(1):222–227.
5. Sniegowski PD, Gerrish PJ, Lenski RE (1997) Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* 387(6634):703–705.
6. Notley-McRobb L, Seeto S, Ferenci T (2002) Enrichment and elimination of *mutY* mutators in *Escherichia coli* populations. *Genetics* 162(3):1055–1062.
7. Chao L, Cox EC (1983) Competition between high and low mutating strains of *Escherichia coli*. *Evolution* 37(1):125–134.
8. Giraud A, et al. (2001) Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science* 291(5513):2606–2608.
9. Oliver A, Cantón R, Campo P, Baquero F, Blázquez J (2000) High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science* 288(5469):1251–1253.
10. Matic I, et al. (1997) Highly variable mutation rates in commensal and pathogenic *Escherichia coli*. *Science* 277(5333):1833–1834.
11. Taddei F, et al. (1997) Role of mutator alleles in adaptive evolution. *Nature* 387(6634):700–702.
12. Loh E, Salk JJ, Loeb LA (2010) Optimization of DNA polymerase mutation rates during bacterial evolution. *Proc Natl Acad Sci USA* 107(3):1154–1159.

13. McDonald MJ, Hsieh Y-Y, Yu Y-H, Chang S-L, Leu J-Y (2012) The evolution of low mutation rates in experimental mutator populations of *Saccharomyces cerevisiae*. *Curr Biol* 22(13):1235–1240.
14. Seshasayee AS, Bertone P, Fraser GM, Luscombe NM (2006) Transcriptional regulatory networks in bacteria: from input signals to output responses. *Curr Opin Microbiol* 9(5):511–519.
15. Hindré T, Knibbe C, Beslon G, Schneider D (2012) New insights into bacterial adaptation through *in vivo* and *in silico* experimental evolution. *Nature Rev Microbiol* 10(5):352–365.
16. Wagner GP, Pavlicev M, Cheverud JM (2007) The road to modularity. *Nature Rev Genet* 8(12):921–931.
17. Kichenaradja P, Siguier P, Perochon J, Chandler M (2010) ISbrowser: an extension of ISfinder for visualizing insertion sequences in prokaryotic genomes. *Nucleic Acids Res* 38(Database issue):D62–68.
18. Siguier P, Gourbeyre E, Chandler M (2014) Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol Rev* DOI: 10.1111/1574-6976.12067.
19. Schneider D, Lenski RE (2004) Dynamics of insertion sequence elements during experimental evolution of bacteria. *Res Microbiol* 155(5):319–327.
20. Touchon M, Rocha EPC (2007) Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol* 24(4):969–981.
21. Hartl DL, Sawyer SA (1988) Why do unrelated insertion sequences occur together in the genome of *Escherichia coli*? *Genetics* 118(3):537–541.
22. Treves DS, Manning S, Adams J (1998) Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of *Escherichia coli*. *Mol Biol Evol* 15(7):789–797.

23. Stoebel DM, Dorman CJ (2010) The effect of mobile element *IS10* on experimental regulatory evolution in *Escherichia coli*. *Mol Biol Evol* 27(9):2105–2112.
24. Sleight SC, Orlic C, Schneider D, Lenski RE (2008) Genetic basis of evolutionary adaptation by *Escherichia coli* to stressful cycles of freezing, thawing and growth. *Genetics* 180(1):431–443.
25. Cooper VS, Schneider D, Blot M, Lenski RE (2001) Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J Bacteriol* 183(9):2834–2841.
26. de Visser JA, Akkermans AD, Hoekstra RF, de Vos WM (2004) Insertion-sequence-mediated mutations isolated during adaptation to growth and starvation in *Lactococcus lactis*. *Genetics* 168(3):1145–1157.
27. Chou HH, Berthet J, Marx CJ (2009) Fast growth increases the selective advantage of a mutation arising recurrently during evolution under metal limitation. *PLoS Genet* 5(9):e1000652.
28. Song H, et al. (2010) The early stage of bacterial genome-reductive evolution in the host. *PLoS Pathog* 6(5):e1000922.
29. Wisner MJ, Ribeck N, Lenski RE (2013) Long-term dynamics of adaptation in asexual populations. *Science* 342(6164):1364–1367.
30. Papadopoulos D, et al. (1999) Genomic evolution during a 10,000-generation experiment with bacteria. *Proc Natl Acad Sci USA* 96(7):3807–3812.
31. Plucain J, et al. (2014) Epistasis and allele specificity in the emergence of a stable polymorphism in *Escherichia coli*. *Science* 343(6177):1366–1369.
32. Elena SF, Lenski RE (1997) Long-term experimental evolution in *Escherichia coli*. VII. Mechanisms maintaining genetic variability within populations. *Evolution* 51:1058–1067.

33. Haas M, Rak B (2002) *Escherichia coli* insertion sequence IS150: transposition via circular and linear intermediates. *J Bacteriol* 184(21):5833–5841.
34. Philippe N, Pelosi L, Lenski RE, Schneider D (2009) Evolution of penicillin-binding protein 2 concentration and cell shape during a long-term experiment with *Escherichia coli*. *J Bacteriol* 191(3):909–921.
35. Jeong H, et al. (2009) Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J Mol Biol* 394(4):644–652.
36. Lenski RE, Rose MR, Simpson SC, Tadler SC (1991) Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *Am Nat* 138(6):1315–1341.
37. Gaffé J, et al. (2011) Insertion sequence-driven evolution of *Escherichia coli* in chemostats. *J Mol Evol* 72(4):398–412.
38. Saitou N, Nei M (1987) The Neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4(4):406–425.
39. Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5:164–166.

## Figure Legends

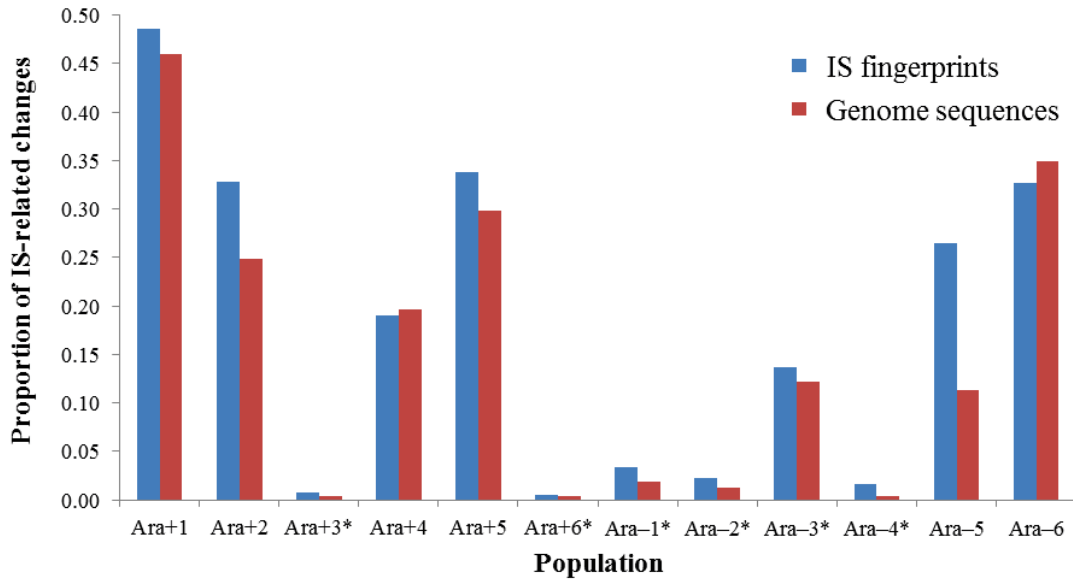
**Fig. 1.** Proportion of IS-related changes relative to total mutations after 40,000 generations. The number of IS-related changes was inferred from either the IS fingerprints (blue) or the genome sequences (red). The ratios to the total number of mutations were calculated and the mean for the evolved clones sampled from identical populations is shown. The 12 populations are shown by their names with stars indicating those that evolved increased mutation rates compared to the ancestor.

**Fig. 2.** Dynamics of IS-related changes and of their contribution to total mutations in population Ara+1. (A) Number of IS-related changes in evolved clones sampled over time. The mean of the number of changes in the two evolved clones sampled at each time point is shown. The black and blue lines indicate the numbers of all IS-related changes and of the *IS150*-related changes, respectively. (B) Proportion of IS-related changes relative to the total mutations in the same evolved clones. The ratios are calculated as the mean of the number of either all IS-related changes (black) or *IS150*-related changes (blue) to the mean of the number of all mutations from the two clones sampled at each time point.

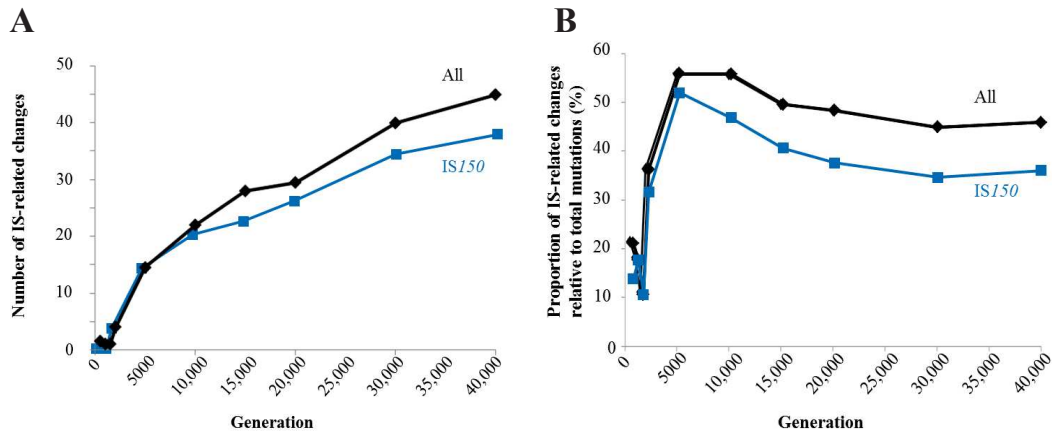
**Fig. 3.** Recombination and transposition frequencies of *IS150* in the ancestor and five evolved clones from population Ara+1. (A) Frequencies of *IS150*-mediated recombination (red) and transposition (blue) events measured using the reporter plasmid pFDX2339. Note the scale change in the *y*-axis. (B) Proportion of transposition events relative to recombination events.

**Fig. 4.** Simulation of IS dynamics and contribution relative to point mutations. Boxplots represent the distribution of the proportions of IS-related changes relative to point mutations for each of the 18 parameter sets. Each boxplot represents 100 measures per condition after 20,000 generations. Each box indicates the upper and lower quartiles, the heavy line the median, and the whiskers the 95% confidence interval.  $\mu^m$  represents the point mutation rate and  $\mu^{IS}$  the transposition rate set as three proportions of IS-related changes relative to point mutation rates. Each condition was run by integrating or not an effect of IS elements on their transposition (grey and white, respectively). \* Student's *t*-test  $0.05 < P < 0.005$ , \*\* Student's *t*-test  $P < 0.005$ . Small circles indicate replicate populations that were outliers.

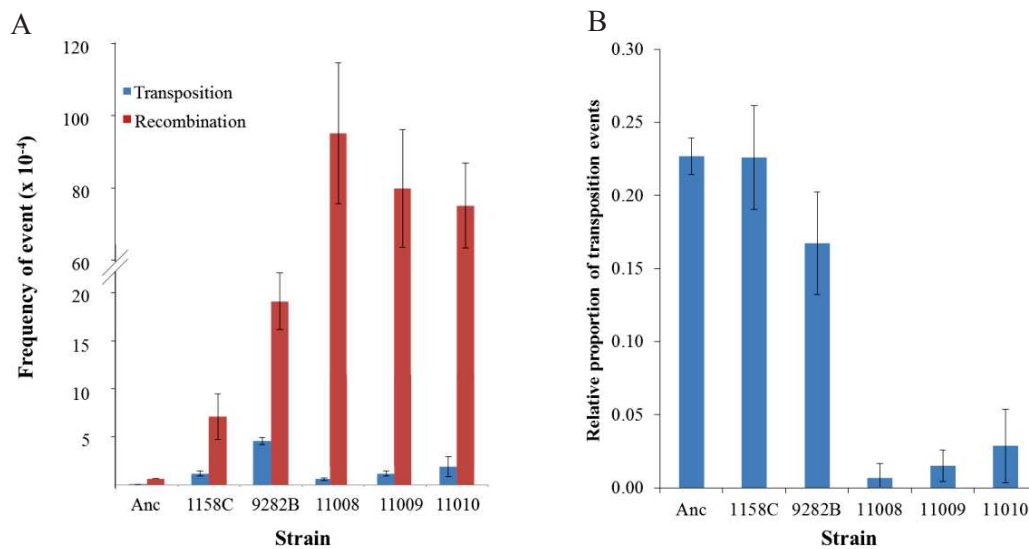




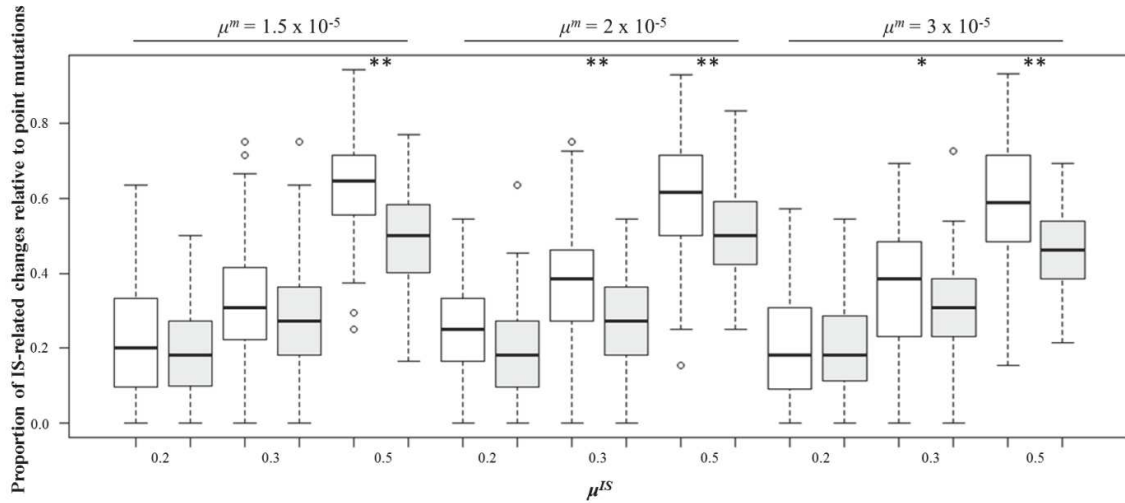
**Fig. 1.** Proportion of IS-related changes relative to total mutations after 40,000 generations. The number of IS-related changes was inferred from either the IS fingerprints (blue) or the genome sequences (red). The ratios to the total number of mutations were calculated and the mean for the evolved clones sampled from identical populations is shown. The 12 populations are shown by their names with stars indicating those that evolved increased mutation rates compared to the ancestor.



**Fig. 2.** Dynamics of IS-related changes and of their contribution to total mutations in population Ara+1. (A) Number of IS-related changes in evolved clones sampled over time. The mean of the number of changes in the two evolved clones sampled at each time point is shown. The black and blue lines indicate the numbers of all IS-related changes and of the *IS150*-related changes, respectively. (B) Proportion of IS-related changes relative to the total mutations in the same evolved clones. The ratios are calculated as the mean of the number of either all IS-related changes (black) or *IS150*-related changes (blue) to the mean of the number of all mutations from the two clones sampled at each time point.



**Fig. 3.** Recombination and transposition frequencies of *IS150* in the ancestor and five evolved clones from population Ara+1. (A) Frequencies of *IS150*-mediated recombination (red) and transposition (blue) events measured using the reporter plasmid pFDX2339. Note the scale change in the y-axis. (B) Proportion of transposition events relative to recombination events.



**Fig. 4.** Simulation of IS dynamics and contribution relative to point mutations. Boxplots represent the distribution of the proportions of IS-related changes relative to point mutations for each of the 18 parameter sets. Each boxplot represents 100 measures per condition after 20,000 generations. Each box indicates the upper and lower quartiles, the heavy line the median, and the whiskers the 95% confidence interval.  $\mu^m$  represents the point mutation rate and  $\mu^{IS}$  the transposition rate set as three proportions of IS-related changes relative to point mutation rates. Each condition was run by integrating or not an effect of IS elements on their transposition (grey and white, respectively). \* Student's  $t$ -test  $0.05 < P < 0.005$ , \*\* Student's  $t$ -test  $P < 0.005$ . Small circles indicate replicate populations that were outliers.

**Table S1. *E. coli* strains used in this study**

Strain	Population	Generation	Accession
REL606		Ancestor	
768A	Ara+1	500	
768B	Ara+1	500	
958A	Ara+1	1000	
958B	Ara+1	1000	
1062A	Ara+1	1500	
1062B	Ara+1	1500	
1158A	Ara+1	2000	
1158C	Ara+1	2000	
2173A	Ara+1	5000	
2173B	Ara+1	5000	
4530A	Ara+1	10,000	
4530B	Ara+1	10,000	
7183A	Ara+1	15,000	
7183B	Ara+1	15,000	
9282A	Ara+1	20,000	
9282B	Ara+1	20,000	
10450	Ara+1	30,000	
10451	Ara+1	30,000	
11008	Ara+1	40,000	
11009	Ara+1	40,000	
11010*	Ara+1	40,000	

10950	Ara+2	40,000	
10951	Ara+2	40,000	
10953	Ara+3	40,000	
10954	Ara+3	40,000	
10956	Ara+4	40,000	
10957	Ara+4	40,000	
10982	Ara+5	40,000	
10983	Ara+5	40,000	
10985	Ara+6	40,000	
10986	Ara+6	40,000	
10938	Ara-1	40,000	SRS007219 <sup>†</sup>
10939	Ara-1	40,000	ERS068526 <sup>†</sup>
10940	Ara-1	40,000	ERS068529 <sup>†</sup>
11035	Ara-2	40,000	
10941	Ara-2	40,000	
11036	Ara-2	40,000	
11214	Ara-2	40,000	

10979	Ara-3	40,000	SRA026813 <sup>‡</sup>
10988	Ara-3	40,000	SRA026813 <sup>‡</sup>
10944	Ara-4	40,000	
10945	Ara-4	40,000	
10947	Ara-5	40,000	
10948	Ara-5	40,000	
10949	Ara-5	40,000	
11005	Ara-6	40,000	
11006	Ara-6	40,000	

---

All of the strains are evolved clones derived from the REL606 strain (35) and sampled from the 12 populations of the long-term evolution experiment.

\*This particular 40,000-generation clone was used only to measure the frequency of IS150 transposition whereas all other clones from the same time point were used in all the analyses.

†These genome sequences are from (4).

‡These genome sequences are from (ref. 1).

1. Blount ZD, Barrick JE, Davidson CJ, Lenski RE (2012) Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* 489(7417):513–518.

**Table S3. Number of copies of the five active IS elements in the ancestor and of IS-related changes in each of the 28 evolved clones sampled at 40,000 generations.**

Population	Clone	IS1	IS3	IS4	IS150	IS186
	REL606	24	5	2	5	5
Ara+1	11008	11	3	0	29	7
	11009	11	3	0	29	7
Ara+2	10950	9	0	0	8	7
	10951	9	0	0	8	5
Ara+3	10953	2	0	2	4	4
	10954	2	0	1	4	4
Ara+4	10956	5	0	3	3	0
	10957	6	0	2	3	0
Ara+5	10982	6	0	0	14	3
	10983	8	0	0	14	3
Ara+6	10985	1	3	0	5	2
	10986	1	3	0	6	1
Ara-1	10938	12	0	0	7	5
	10939	13	3	0	8	4
	10940	13	3	0	8	4
Ara-2	11035	7	4	0	8	1
	10941	8	4	0	8	2
	11036	8	3	0	11	6
	11214	7	3	0	11	4
Ara-3	10979	10	4	0	17	8



	10988	11	4	0	23	8
Ara-4	10944	8	3	0	5	1
	10945	6	3	0	5	1
Ara-5	10947	6	3	0	14	6
	10948	5	3	1	15	6
	10949	6	3	1	15	6
Ara-6	11005	8	1	0	12	6
	11006	6	1	0	10	2

---

Data are from the IS fingerprints. For the ancestor, the copy number is given. For the evolved clones, the number of changes in IS-hybridizing bands is given, including therefore the loss and gain of IS copies. The data for IS2 and IS30 (1 copy each in the ancestor) are not shown since no change was detected.

**Table S4. Primers used in this study to PCR-amplify 19 new IS150 copies in evolved clones**

Primer	Sequence 5'-3'	Position *	Gene <sup>†</sup>	Position of IS insertion*
IS150-3-F	tgatgttgaactggaagtcg	335,981	<i>lacZ</i>	336,055
IS150-3-R	tgttttgaccgctgggatct	336,100		
IS150-5-F	ccgattatcctctggcgtcg	601,005	<i>fepE</i>	601,096
IS150-5-R	gctgatctctttcatcctgc	601,126		
IS150-6-F	cccctgggtgaatcaaatag	651,527	<i>ybeB(177)</i>	651,703
IS150-6-R	gccagacacgactttgtaga	651,777		
IS150-7-F	cccgtgccgttcacctct	711,881	<i>rhcC</i>	711,955
IS150-7-R	gcggatatccagattcgtg	712,001		
IS150-8-F	ggataacgggtcgtggtg	896,391	ECB_00838	896,463
IS150-8-R	ttcccgaacgccgggttat	896,546		
IS150-10-F	ctttcaggtcacgttcgatg	969,768	<i>pflB</i>	969,836
IS150-10-R	accaacgctcaggaagctat	969,881		
IS150-11-F	attttgccccggtctacgac	1,101,953	<i>phoH</i>	1,102,032
IS150-11-R	cacattctgcgcctcgtcaa	1,102,097		
IS150-12-F	caagtatgcatcgctgatgg	1,376,305	<i>ycjT</i>	1,376,381
IS150-12-R	accagttcgttgattcacc	1,376,429		
IS150-13-F	atggcgggtaagaggctaag	1,462,199	<i>hokB(313)</i>	1,462,266
IS150-13-R	ggtgcctgagactttctgtt	1,462,306		
IS150-14-F	tagcctgcttctccttgct	1,519,763	<i>hokD(544)</i>	1,519,832
IS150-14-R	tgaacaccaacggcagatag	1,519,885		
IS150-15-F	atggctttcggattagtcag	1,858,946	<i>yeaS</i>	1,859,000

IS150-15-R	ttgctctatctggggagta	1,859,061		
IS150-16-F	ggcaacattttagaccaga	2,019,729	<i>yeeZ</i> (196)	2,019,802
IS150-16-R	ggcctggagtgactacgta	2,019,849		
IS150-19-F	gccctaatacgcacaaaagcc	2,283,403	<i>yfaA</i> (21)	2,283,472
IS150-19-R	tgaaatcgctccggaagtgg	2,283,522		
IS150-20-F	gtttggtatgcgagtgggta	2,397,339	<i>yfcT</i>	2,397,529
IS150-20-R	cgaacatgcaacgtaccgga	2,397,641		
IS150-21-F	gcggtaatcacttcccataa	2,591,569	<i>hcaC</i>	2,591,547
IS150-21-R	taaaactcgccccaacgtt	2,591,594		
IS150-22-F	cgtaccggatcgtcttgatt	2,655,684	<i>yfiH</i>	2,655,779
IS150-22-R	ctggcggatatttatctgct	2,655,853		
IS150-23-F	tggaatgcggtcgaggattt	2,877,238	<i>kduD</i>	2,877,312
IS150-23-R	gttatgggcgtgacgcgatt	2,877,368		
IS150-26-F	gggtataaggctgtagcgcc	3,493,736	ECB_03279	3,493,822
IS150-26-R	ccagtaacgccgtgcctaata	3,493,884		
IS150-28-F	gctggtaggtacaacagctgc	4,415,599	<i>cycA</i>	4,415,710
IS150-28-R	gcagtaccagcgaaaggct	4,415,886		

---

\*Positions are given according to the ancestral genome sequence (35).

†If the insertion of a new IS150 copy occurred within a gene, its name is given. If the insertion occurred in an intergenic region, the name of the upstream gene is given together with the distance in basepairs, in brackets, between the last base of its stop codon and the IS150 insertion.

**Table S5. Mutations detected in two clones sampled from population Ara+1 at 2000 generations**

Population	Clone	Position*	Mutation Type <sup>†</sup>		Affected Gene <sup>‡</sup>	Effect
Ara+1	REL1158A	70,867	NS-SNP	T->C	<i>araA</i>	D92G
Ara+1	REL1158A	1,236,764	IS insertion	IS150	<i>ycgB</i>	
Ara+1	REL1158A	1,406,224	SNP	G->A	<i>ydaN</i> (254)	
Ara+1	REL1158A	1,462,266	IS insertion	IS150	<i>hokB</i> (313)	
Ara+1	REL1158A	1,733,865	NS-SNP	G->T	<i>pykF</i>	A301S
Ara+1	REL1158A	2,847,052	NS-SNP	A->G	<i>recD</i>	V10A
Ara+1	REL1158A	3,251,604	NS-SNP	T->G	<i>nusA</i>	Y386S
Ara+1	REL1158A	3,761,696	NS-SNP	C->A	<i>spoT</i>	H314N
Ara+1	REL1158A	3,894,998	IS150-mediated deletion	6133 bp	<i>rbs</i> operon	
Ara+1	REL1158A	4,615,673	IS insertion	IS150	<i>nadR</i>	
Ara+1	REL1158C	70,867	NS-SNP	T->C	<i>araA</i>	D92G
Ara+1	REL1158C	651,703	IS insertion	IS150	<i>ybeB</i> (177)	
Ara+1	REL1158C	664,687	IS150-mediated deletion	1444 bp	<i>insJ, insK</i>	
Ara+1	REL1158C	1,462,266	IS insertion	IS150	<i>hokB</i> (313)	
Ara+1	REL1158C	1,733,865	NS-SNP	G->T	<i>pykF</i>	A301S
Ara+1	REL1158C	2,847,052	NS-SNP	A->G	<i>recD</i>	V10A
Ara+1	REL1158C	3,102,720	Insertion	1 bp	<i>mqsR</i> (201)	
Ara+1	REL1158C	3,440,702	NS-SNP	C->A	<i>yhfZ</i>	A109S
Ara+1	REL1158C	3,894,998	IS150-mediated deletion	6133 bp	<i>rbs</i> operon	
Ara+1	REL1158C	4,615,673	IS insertion	IS150	<i>nadR</i>	

\*All positions are given according to the ancestral genome sequence (35).

<sup>†</sup>SNP : single-nucleotide polymorphism ; NS : non-synonymous.

<sup>‡</sup>For intergenic mutations, the upstream gene is given together with, in brackets, the distance in basepairs from the last base of the stop codon to the mutation.

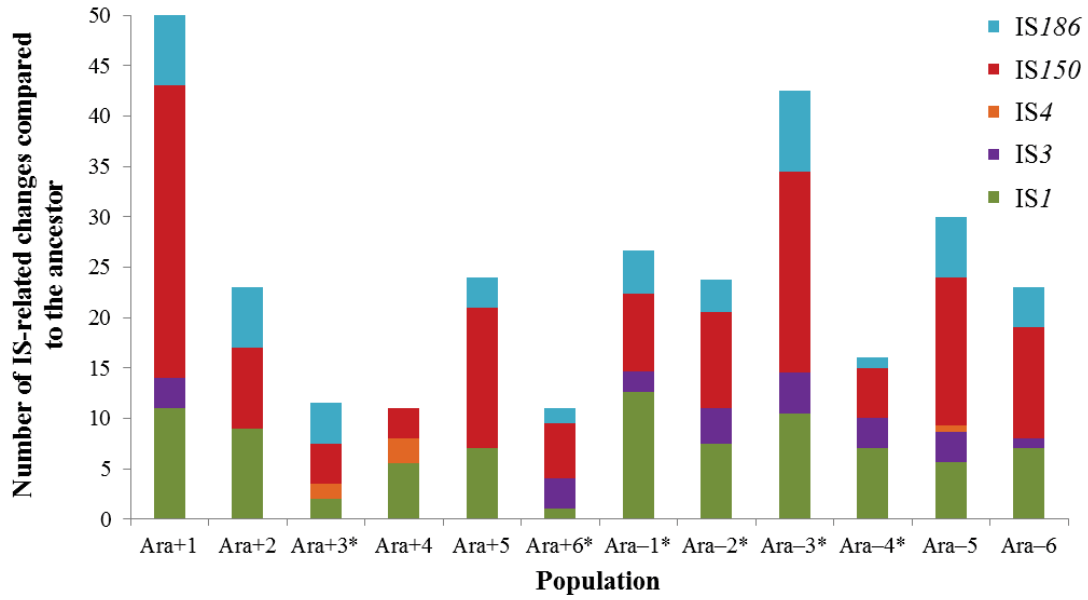
**Table S6. Data used to calculate IS150 recombination and transposition frequencies.**

Clone (Replicate)	UFC/mL *	Dilution Cm <sup>R†</sup>	N <sub>T</sub>	N <sub>T</sub> per UFC	N <sub>R</sub>	N <sub>IS</sub>	N <sub>R</sub> + N <sub>IS</sub>	N <sub>R</sub> /(N <sub>R</sub> + N <sub>IS</sub> )	N <sub>IS</sub> /(N <sub>R</sub> + N <sub>IS</sub> )	N <sub>T</sub> x N <sub>R</sub> / (N <sub>R</sub> + N <sub>IS</sub> )	N <sub>T</sub> x N <sub>IS</sub> / (N <sub>R</sub> + N <sub>IS</sub> )
Ancestor (1)	9.00 x 10 <sup>8</sup>	1000	72	8.00 x 10 <sup>-5</sup>	83	18	101	0.822	0.178		
Ancestor (2)			44	4.89 x 10 <sup>-5</sup>	83	19	102	0.814	0.186		
Ancestor (3)			87	9.67 x 10 <sup>-5</sup>	77	19	96	0.802	0.198		
Ancestor (4)			78	8.67 x 10 <sup>-5</sup>	79	17	96	0.823	0.177		
Ancestor (5)			74	8.22 x 10 <sup>-5</sup>	79	18	97	0.814	0.186		
Ancestor (Total)	9.00 x 10 <sup>8</sup>	1000	71	7.89 x 10 <sup>-5</sup>	401	91	492	0.815	0.185	6.43 x 10 <sup>-5</sup>	1.46 x 10 <sup>-5</sup>
1158C (1)	2.19 x 10 <sup>8</sup>	1000	180	8.23 x 10 <sup>-4</sup>	61	12	73	0.836	0.164		
1158C (2)			133	6.08 x 10 <sup>-4</sup>	64	14	78	0.821	0.179		
1158C (3)			197	9.00 x 10 <sup>-4</sup>	53	12	65	0.815	0.185		
1158C (4)			152	6.95 x 10 <sup>-4</sup>	62	13	75	0.827	0.173		
1158C (5)			170	7.77 x 10 <sup>-4</sup>	52	15	67	0.776	0.224		
1158C (Total)	2.19 x 10 <sup>8</sup>	1000	166.4	7.61 x 10 <sup>-4</sup>	292	66	358	0.816	0.184	6.20 x 10 <sup>-4</sup>	1.40 x 10 <sup>-4</sup>
9282B (1)	4.12 x 10 <sup>8</sup>	1000	112	2.72 x 10 <sup>-4</sup>	56	9	65	0.862	0.138		
9282B (2)			97	2.35 x 10 <sup>-4</sup>	49	11	60	0.817	0.183		
9282B (3)			106	2.57 x 10 <sup>-4</sup>	55	8	63	0.873	0.127		
9282B (4)			125	3.03 x 10 <sup>-4</sup>	59	8	67	0.881	0.119		
9282B (5)			124	3.01 x 10 <sup>-4</sup>	50	9	59	0.847	0.153		
9282B (Total)	4.12 x 10 <sup>8</sup>	1000	112.8	2.74 x 10 <sup>-3</sup>	269	45	314	0.857	0.143	2.35 x 10 <sup>-3</sup>	3.93 x 10 <sup>-4</sup>
11008 (1)	3.08 x 10 <sup>8</sup>	10000	288	9.34 x 10 <sup>-3</sup>	59	1	60	0.983	0.017		
11008 (2)			306	9.92 x 10 <sup>-3</sup>	59	0	59	1.000	0.000		
11008 (3)			292	9.47 x 10 <sup>-3</sup>	63	0	63	1.000	0.000		
11008 (4)			285	9.24 x 10 <sup>-3</sup>	55	1	56	0.982	0.018		
11008 (5)			267	8.66 x 10 <sup>-3</sup>	61	0	61	1.000	0.000		
11008 (Total)	3.08 x 10 <sup>8</sup>	10000	287.6	9.33 x 10 <sup>-3</sup>	297	2	299	0.993	0.007	9.26 x 10 <sup>-3</sup>	6.24 x 10 <sup>-5</sup>

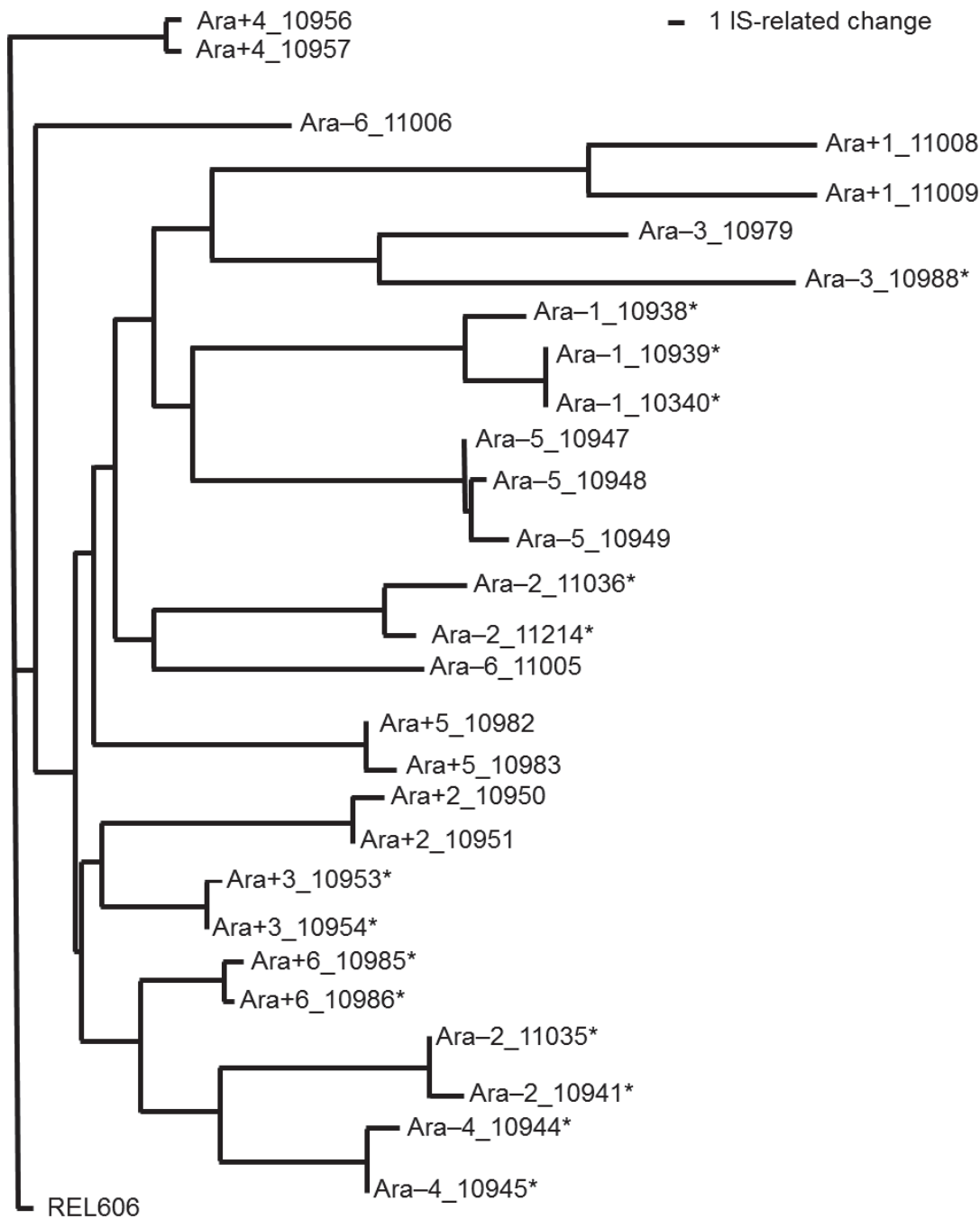
11009 (1)	3.84 x 10 <sup>8</sup>	20000	168	8.75 x 10 <sup>-3</sup>	64	1	65	0.985	0.015		
11009 (2)			171	8.90 x 10 <sup>-3</sup>	66	1	67	0.985	0.015		
11009 (3)			209	1.09 x 10 <sup>-2</sup>	66	2	68	0.971	0.029		
11009 (4)			124	6.45 x 10 <sup>-3</sup>	62	0	62	1.000	0.000		
11009 (5)			153	7.96 x 10 <sup>-3</sup>	73	1	74	0.986	0.014		
11009 (Total)	3.84 x 10 <sup>8</sup>	20000	165	8.59 x 10 <sup>-3</sup>	331	5	336	0.985	0.015	8.46 x 10 <sup>-3</sup>	1.28 x 10 <sup>-4</sup>
11010 (1)	3.39 x 10 <sup>8</sup>	10000	230	6.79 x 10 <sup>-3</sup>	49	0	49	1.000	0.000		
11010 (2)			256	7.56 x 10 <sup>-3</sup>	40	2	42	0.952	0.048		
11010 (3)			255	7.53 x 10 <sup>-3</sup>	52	1	53	0.981	0.019		
11010 (4)			262	7.74 x 10 <sup>-3</sup>	55	1	56	0.982	0.018		
11010 (5)			249	7.36 x 10 <sup>-3</sup>	49	3	52	0.942	0.058		
11010 (Total)	3.39 x 10 <sup>8</sup>	10000	250.4	7.40 x 10 <sup>-3</sup>	245	7	252	0.972	0.028	7.19 x 10 <sup>-3</sup>	2.05 x 10 <sup>-4</sup>

\*Number of UFC/mL after 48 h of growth in LB-Kan at 24°C.

†Dilution factor of culture after 48 h of growth in LB-Kan at 24°C before plating onto LB-Kan-Cam and LB-Cam.

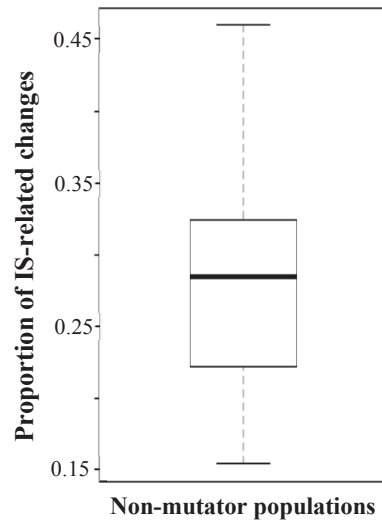


**Fig. S1.** Number of IS-related changes compared to the ancestor. The mean of the number of IS-related changes was calculated for the evolved clones sampled from the same population (Table S1), and is given for each IS element according to the indicated colour code. The 12 populations are shown by their names (Ara+1 to Ara+6 and Ara-1 to Ara-6) with stars indicating those that evolved increased mutation rates compared to their ancestor. Two populations (Ara+1 and Ara-3) revealed a strong increase ( $\geq 30$ ) in IS-related changes, seven (Ara+2, Ara+5, Ara-1, Ara-2, Ara-4, Ara-5 and Ara-6) a moderate increase (15 to 30), and three (Ara+3, Ara+4 and Ara+6) a lower increase ( $< 15$ ).

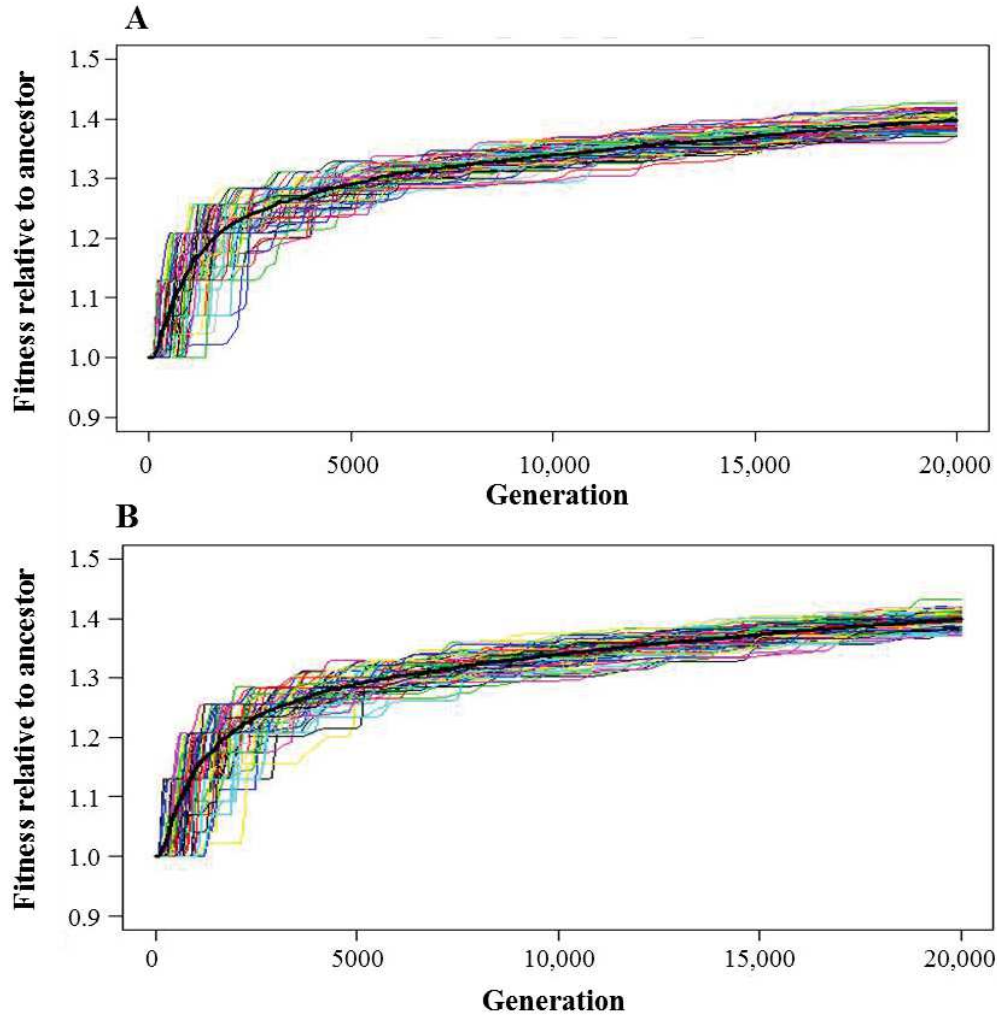


**Fig. S2.** Phylogenies for 28 evolved clones sampled from the 12 populations of *E. coli* at 40,000 generations, rooted by using the ancestor genotype. Phylogenies were inferred from the IS fingerprints. The names of the clones are given according to the nomenclature of Table S1, preceded by the name of the population from which they were sampled. Stars denote clones that evolved a mutator phenotype. The scale for one IS-related change is indicated.

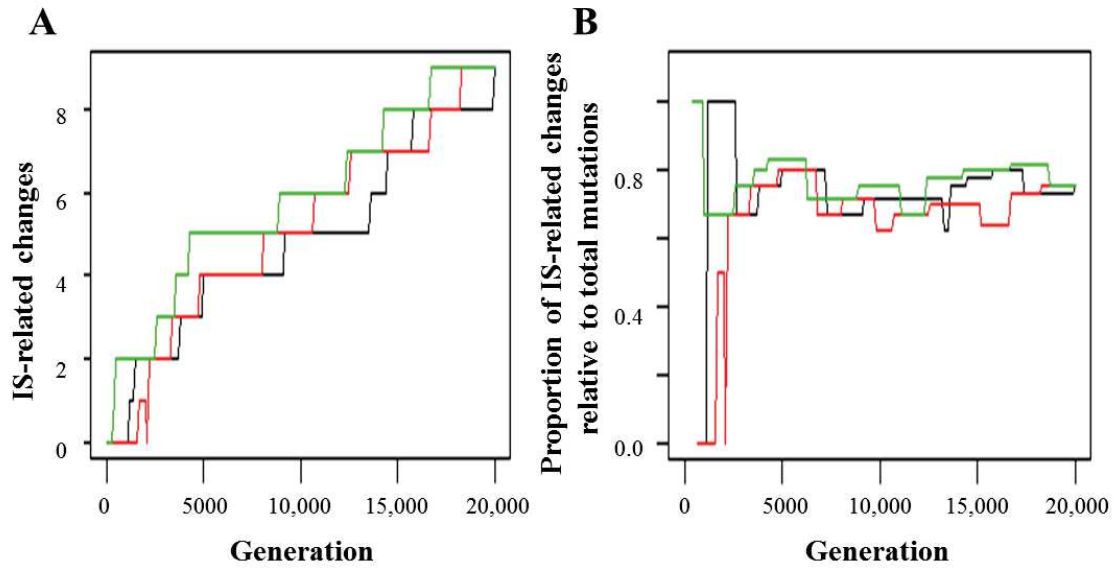




**Fig. S3.** Proportion of IS-related changes in the non-mutator populations. The boxplot represents the distribution of the proportion of IS-related changes in the evolved clones sampled from populations Ara+1, Ara+2, Ara+4, Ara+5, Ara-5 and Ara-6 at 40,000 generations. The box indicates the upper and lower quartiles, the heavy line the median, and the whiskers the 95% confidence interval.



**Fig. S4.** Fitness values of 100 evolved replicate populations relative to their common ancestor over 20,000 generations. These simulations were run at  $\mu^m = 2 \times 10^{-5}$  and  $\mu^{IS} = 0.3$ , with (A) or without (B) impact of IS elements on  $\mu^{IS}$ . The fitness trajectory of each replicate population is represented by a different coloured line, and the thick black line the mean fitness trajectory.



**Fig. S5.** Dynamics of IS elements in three of 100 replicate populations ( $\mu^m = 2 \times 10^{-5}$ ,  $\mu^{IS} = 0.3$ ) where both the number of IS-related changes (*A*) and the proportion of IS-related changes relative to total mutations (*B*) were higher. Simulation measures were performed at each 100-generation interval and the graphs represent the dominant lineage in each population. In the 97 other replicate populations after 20,000 generations, the number of IS-related changes were between 1 and 7, and their proportion between 0.1 and 0.69.

# General Discussion

This work has shown the importance of chromosomal rearrangement events and insertion sequence (IS) movements during the adaptive processes of bacteria to a new environment. Moreover, these results illustrate the power of using experimental evolution experiments to perform in depth analysis of evolution. Our tool, the Long Term Evolution Experiment (LTEE), allows for an exhaustive approach as we can retrace almost every step undertaken by evolution due to easy access to a revivable fossil record. Classical studies of evolution are based on the observation of clones that have diverged for a long evolutionary time, focusing only on the surviving and then successful mutants and using a highly incomplete and imperfect fossil record.

“I endeavoured, also, to show that intermediate varieties, from existing in lesser numbers than the forms which they connect, will generally be beaten out and exterminated during the course of further modification and improvement. The main cause, however, of innumerable intermediate links not now occurring everywhere throughout nature depends on the very process of natural selection, through which new varieties continually take the places of and exterminate their parent-forms. But just in proportion as this process of extermination has acted on an enormous scale, so must the number of intermediate varieties, which have formerly existed on the earth, be truly enormous. Why then is not every geological formation and every stratum full of such intermediate links? Geology assuredly does not reveal any such finely graduated organic chain; and this, perhaps, is the most obvious and gravest objection which can be urged against my theory. The explanation lies, as I believe, in the extreme imperfection of the geological record. “

C Darwin "On the Origin of Species" 1859 chapter IX: "On the Imperfection of the Geological Record"

The utilisation of the longest ongoing experiment gives access to previous analysis and allows us to complement what has already been observed in these lines. The levels of single nucleotide polymorphisms (Wielgoss, Barrick et al. 2013) , changes in global regulator activity (Cooper, Remold et al. 2008) or regulatory genes (Plucain, Hindre et al. 2014) expression of cryptic metabolic pathways (Blount, Barrick et al. 2012) have been extensively studied in the 12 populations in up to 40,000 generations. This provides a large understanding of evolved phenotypes and their beneficial effect to the cell. However, chromosomal rearrangements and movements of IS elements have been less studied (Papadopoulos, Schneider et al. 1999; Schneider, Duperchy et al. 2000) but their involvement in genetic innovation is crucial as illustrated by the tandem duplication involved in the cit<sup>+</sup> phenotype (Blount, Barrick et al. 2012). We only know of the IS dynamics up to 10,000 generations for two populations (Papadopoulos, Schneider et al. 1999) and the rearrangement profiles for the same two populations up to 2000 generations.

**Rearrangements and Insertion Sequences are complimentary and both important actors in adaptation.**

Part of this work was to extend the existing knowledge of rearrangements in the LTEE up to 40,000 generations in the twelve populations. In total we found 110 rearrangements, deletions, inversions and duplications all included in 1 strain issued from each of the twelve populations. Deletions were the most numerous, accounting for 75% of all rearrangements, inversions for 17% and duplications for 8%. IS elements are heavily implicated and account for 73% of the rearrangements. This could be expected as IS sequences are in multiple homologous copies distributed randomly throughout the genome and genomes with large

numbers of IS copies can undergo more homologous recombination (Rohmer, Fong et al. 2007) and previously described rearrangements in the twelve populations were all attributed to IS elements. Indeed, in all twelve populations IS elements have increased in copy numbers with a total of 209 IS signals detected by RFLP over 29 clones selected from the twelve populations at 40,000 generations. They were distributed randomly across the genome with no clear pattern, facilitating the potential for homologous recombination. However, even though they have a random insertion profile, parallel events occurred when IS sequences inserted in the *rbs* operon in all twelve populations (Cooper, Schneider et al. 2001). In another evolution experiments on continuous cultures, IS elements were principal actors in rearrangements, 4% of the ancestor genome and delimited by *IS1* element had been observed (Maharjan, Gaffe et al. 2013), and IS were involved in the disruption of *rpoS*, deletion involving *mutY*, *yegS* and *yegR* (Gaffe, McKenzie et al. 2011). Even though the reason for the presence and conservation of IS elements in bacterial genomes is under debate, it is recognise that IS elements play an essential role in the shaping of bacterial genomes by acting as the major source for rearrangements (Siguier et al 2014). This is especially true for the passage of a bacteria to an obligate intracellular pathogen where genome reduction is preceded by an explosion in the number of IS elements prior to massive reduction in chromosome size (Bordenstein and Reznikoff 2005). As such in LTEE the increase in IS elements might favour an increase in rearrangement rates (Papadopoulos, Schneider et al. 1999). A previous study at 10,000 generations in the evolved population Ara-1 and Ara+1, shows that there are already large inversions present due to IS elements (Papadopoulos, Schneider et al. 1999). However, the present work indicates that IS elements are not the only participants in recombination; other factors also participate in rearrangements in the evolving populations and in natural isolates (Rau, Marvig et al. 2012). The second most common event cause was the deletion of prophage remnants. There were no homologous regions at either end of the majority of

prophage deletions so it can be theorised that this type of recombination could be recombination independent. Prophage elimination is a common feature where populations have evolved in environments where constitutively expressed prophage genes are not advantageous and they are often described as dispensable regions of DNA (Varani, Monteiro-Vitorello et al. 2013). Similar to IS elements, prophage sequences can play a role in homologous recombination (Darling, Miklos et al. 2008) even though none of this type was observed in the LTEE .

DNA coding for rRNA is also a source for rearrangements. There are 7 rRNA operons in the ancestor spanning over 5000bp with a very high degree of homology (Jeong, Barbe et al. 2009). There were only four inversion involving rRNA sequences and even though their effect is not known. Rearrangements and IS elements participate in a major way in the populations adaptation to the experimental evolution as evidenced by their relatively high frequencies. Even though the exact effect of the majority of these events are not known, there is strong argument that they could be beneficial due to the high levels of parallelism observed between the different populations.

### **Parallelism in the Long Term evolution Experiment**

The evolution of the 12 populations is characterized by a strong phenotypic parallelism: fitness and cell volume increase (Lenski and Travisano 1994; Cooper and Lenski 2000; Lenski 2004; Wisser, Ribick et al. 2013). Globally the cells increase their ability to use glucose as a source of carbon which is the main carbon source in the evolution environment whereas the ability to use other sources of carbon not present in the evolution media such as ribose or maltose is reduced (Cooper and Lenski 2000; Cooper, Schneider et al. 2001; Pelosi, Kuhn et al. 2006). Simultaneously, the expression of many genes is affected in parallel in the populations (Cooper, Rozen et al. 2003; Pelosi, Kuhn et al. 2006) by modifications in the



activity of global regulators such as CRP (Cooper, Remold et al. 2008). In ten populations, the chromosome superhelicity is increased (Croizat, Winkworth et al. 2010) through major modifications to *fis* expression (Croizat, Philippe et al. 2005).

The optical mapping and the whole genome sequencing of clones from the 12 populations allowed us to observe a high level of parallelism in the chromosomal rearrangements. There were a total of nine regions where deletions were parallel in two or more populations, including one previously analysed deletion of the operon containing the genes necessary for the utilisation of ribose in all twelve populations (Cooper, Schneider et al. 2001). This deletion occurred between an ancestral *IS150* and newly inserted *IS150* that inserted in the same operon in the twelve populations but at different locations (Cooper, Schneider et al. 2001). It has become fixed in all twelve populations due to the fitness benefit associated with it. IS insertion signatures tended to be specific in each population and even between each strain within populations studied. This shows the largely stochastic method of IS dissemination. The fitness benefits of the other deleted regions are not known, however, since they mostly involve cryptic prophage remnants that do not encode for any vital genes, it can be supposed that their elimination will be beneficial. The prophage remnants are defective and do not code for any virulence genes however,  $\lambda$ -prophages such as DLP12 and Rac, encode proteins which might cooperate in the host physiology (Canchaya, Proux et al. 2003). These prophages encode for the *bor* gene whose product is an outer membrane lipoprotein (Barondess and Beckwith 1995). The protein was detected in various *E. coli* lysogens including many clinical isolates and plays a direct role in the pathogeny of *E. coli* by significantly and specifically increasing the survival of *E. coli* in animal sera (Barondess and Beckwith 1990). Therefore in the evolution media where constitutive expression of certain proteins is not essential for growth in minimal media the presence of prophages and their proteins would represent a cost to fitness to the cell. This could explain why the elimination

of prophages is commonly observed. Another region that was deleted in 10 out of 12 populations was between two homologous genes *manB* and *cpsG* and also represented the loss of non-essential genes in the evolution environment. The strong parallelism between the populations suggests; even though there is no quantitative data, that all these deletions confer a fitness advantage in the evolution experiment.

A high level of parallelism was also detected for a duplication of a stretch of genome containing *rpoS* the sigma factor involved in stationary phase gene expression and stress response (Hengge-Aronis 1993). Our analysis indicates that the duplication occurred in 10 out of the 12 populations, involved various IS elements as *IS1*, *IS150*, *IS186*, depending of the population and six different sizes for these duplications have been noted. Heterology of duplications could even be observed within two populations. This could be due to duplications being bound to the stochastic profile of IS insertion. The complete parallel duplicated region comprised *rpoS*, *nlpD*, *pcm*, *surE*, *truD*, *ispF*, *ispD*, *ygbE*, *ftsB*, *cysC*, *cysN*, *cysD* and *iap*. Along with *rpoS* the genes *nlpD*, *pcm* and *surE* are all essential for stationary phase survival and might be the targets of the duplication as the populations spend most of their time in stationary phase during the evolution experiment. The other operons are involved in the pathway for the biosynthesis of isoprenoids (*ispF* and *ispD*) (Campbell and Brown 2002) and for the sulphate activation pathway of ATP-sulphurylase (*cysC*, *cysN* and *cysD*) (Leyh, Taylor et al. 1988) and two genes (*ftsB* and *iap*) involved in cell division (Buddelmeijer and Beckwith 2004) and an alkaline phosphatase isozyme conversion protein, respectively. While there is no information about the contribution of each duplicate gene on the bacterial fitness, the most probable target for duplication remains the global regulator *rpoS*. It is also a target in other evolution experiments highlighting its importance (Ferenci. 2008).

In the LTEE another duplication event gave rise to a clone capable of using citrate as a carbon source in the population Ara-3 (Blount, Borland et al. 2008, Blount, Barrick 2012 ). The tandem duplication of a 2.9 kb DNA fragment containing *citT* encoding for a citrate/succinate antiport resulted in the gene being placed under the control of the aerobically expressed promoter of *rnk*. However, when the evolution experiment on precursor clones was replayed from a clone that had not yet evolved the duplication, out of the 270 attempts only 17 gave the Cit<sup>+</sup> phenotype. They were not all due to a duplication and they included 6 results from a cryptic promoter located at the terminal part of a newly transposed IS3 element increased the expression of the *citT* gene (Blount, Barrick et al. 2012). Duplications have been identified in other evolution experiments, however, due to their inherent instability their fitness effects are elusive (Maharjan, Gaffe et al. 2013), but, similar to deletions, the level of parallelism suggests that it is beneficial. There were three regions of the genome that were affected by inversions between two or more populations. They varied in size from 364,762bp to 1,368,196bp and involved a large number of genes. In one shared inversion we noticed a stretch of 164,027bp that is inverted in eight of the twelve populations covering a large number of genes but also the *dif* region where the sister chromosomes are segregated after duplication (Stouf, Meile et al. 2013). Therefore gene orientation is maintained in these cases. It is nearly impossible to know the effect of these large inversions as they occur in already well adapted populations where there are many other mutations that have large effects on cell fitness. It can be theorized that these inversions are beneficial to the populations since they occur in the majority of populations.

### **Diversity of events and potential effects**

One common result was that there were large differences in the extent that each population was affected by rearrangements or IS elements. For rearrangements all the twelve populations but one sustained more than five rearrangements and four of these populations more than ten. The average number of rearrangement is 7.5. The extent of these rearrangements is variable as some population (Ara+1 and Ara+2) sustained reorganization of more than the half of the chromosome whereas some other as Ara+4, Ara+5, Ara+6 show small changes.

There is no clear link between the number and the extent of chromosome rearrangement. Population Ara+5 sustains the lowest number and extent of rearrangement, on the other end the population Ara-3 with 20 rearrangement is the most heavily affected but with a limited impact on the overall organization of the chromosome. Similarly IS elements have varying impact on the populations. Two populations, Ara+1 and Ara-3, had high levels of IS participation in the rearrangement events while other populations varied between 30 and 12 changes. This variety in IS copy numbers and in the IS distribution can be explained by the random method of IS dissemination throughout the genome. IS insertions account for ~50% of all mutations in the population Ara+1 while in other non-mutator populations it varies between 35% and 20%. Large differences in contribution of IS changes have been previously observed in experimental evolution, with batch cultures (Rhiele, Bennett et al. 2001 ) and chemostat cultures (Maharjan, Gaffe et al. 2013) or in natural isolates (Kresse, Dinesh et al. 2003, Rhomer, Fong et al. 2007). The IS insertions detected are generally those that have been fixed in a population and may represent insertions that have either a beneficial or neutral effect on the cell. There was no correlation between the number of IS element changes and the number of rearrangements. Even though it can be supposed that the number of IS elements would increase the number of available targets for recombination. Major phenotype changes occurred in several populations leading to the cit<sup>+</sup> phenotype in Ara-3 (Blount,

Barrick et al. 2012) and the ability to grow on cell culture by product leading to two co-existing subpopulations in Ara-2 (Plucain, Hindre et al. 2014). We show that, based on IS - RFLP profiles, there might also be two co-evolving subpopulations in the population Ara-6.

### **Benefits to evolution of rearrangements and IS elements**

This study identifies a large number of rearrangements and IS insertions. Unfortunately the fitness effects of many of these events are unknown due to their size and their complexity. There has only been two rearrangement events that have been characterised, the ribose operon deletion in all twelve strains increased the fitness by 1 to 2% compared to the ancestor (Cooper, Schneider et al. 2001), the duplication leading to citrate utilization in Ara-3 gave the population a 1% advantage in relation to a clone without a duplication; however, more importantly this duplication was potentiating step for more mutations (Blount, Borland et al. 2008, Blount, Barrick et al. 2012). Changes on fitness brought by IS are even less studied, excepted for one clone in the Ara+1 population at 2000 generations where IS elements account for almost all its fitness increase compared to the ancestor. From the presented results, it is difficult to comment on specific rearrangements and IS movements, however when compared to other observations, it can be hypothesized that the majority of these events are beneficial, in particular in the case of parallel events. IS elements are thought to maintain themselves in the genome via two opposing mechanisms, either as parasites with a high rate of transposition and through dissemination via horizontal transfer or as beneficial actors that create enough beneficial mutations that they are maintained in the cell (Siguier, Gorbeyre et al. 2014). Using the LTEE we show that early IS insertions are beneficial to the cell and can be responsible for the majority of its fitness increase. Transposition then increase linearly in relation to the number of IS copies on the chromosome. We are able to use this data to

propose a simple model for IS invasion. The model shows that if enough beneficial IS insertions occur then they are able to maintain themselves through parasitic behavior. This shows that both previous theories of IS dissemination are essentially correct.

### **Rate of events and implication in natural isolates**

The levels of rearrangements and IS changes,  $2.3 \times 10^{-4}$  and  $6.2 \times 10^{-4}$  per generation, respectively (all types added together) were within previously estimated levels for rearrangements  $10^{-3} \sim 10^{-5}$  and IS  $3.5 \times 10^{-4}$  (in *E. coli* K12 batch cultures) (Sousa, Bourgard et al. 2013). These rates are mostly inferred from experimental evolution experiments and are not entirely representative of what occurs in natural isolates. Compared to point mutations, descriptions of reorganization events in clinical or natural isolate of *E. coli* are relatively rare while IS movements are largely underreported. This suggests that either rearrangements are not beneficial in a complex environment or counter selected during the horizontal transfers maintaining the chromosomal integrity. A more trivial alternative might be that these rearrangements are not easily detected, as already report during experimental evolution (Maharjan, Gaffe et al. 2013). IS movements have been largely described in relation to the passage of bacteria to endosymbionts, where the transposition rate has not been described, only that IS elements eventually account for up to 23% of the genome (Gil, Belda et al. 2008). In other conditions, IS changes have been largely understated. This could be due to difficulties in detection by sequencing or a difficulty in understanding their impact on the genome. To overcome these limitations new technologies have been developed and are now routinely used to investigate chromosomal changes in clinical isolates even though the rate of these events are not usually given as there is an unknown number of generations between the compared strains. The use of optical mapping on *E. coli* O157:H7 natural isolates from a

2006 outbreak demonstrated that large chromosomal inversion (up to 1.6 Mbp) are observed in 2 out of the 11 populations analysed and eventually involved complex events resulting from sequential inversions (Iguchi terajima et al. 2006). Chromosomal rearrangements have been detected in clinical isolates of *Samonella enterica* (Matthews, Rabsch et al. 2011), *Bordetella pertussis*, *Neisseria meningitidis* or *N. gonorrhoeae* where an inversion corresponding to more than one third of the chromosome had been found, (Spencer-Smith, Varkey et al. 2012). In *P. aeruginosas* large chromosomal inversions are induced by transposition and recombination between insertion sequences which in parallel with gene inactivation may improve the fitness of the bacteria through a trial and error process (Rau, Marvig et al. 2012). In most of these cases IS elements are the source for the rearrangements.

#### **Limitations of events.**

In general there are more conditions that have been described limiting the scale of rearrangements but not of IS movements. The main limiting factor is the change in fitness brought by each event. There are limits on rearrangements due to the structure of the genome. *E. coli* genome is organized in 4 macrodomains corresponding to structured segments of the chromosome and two inter macrodomain regions that correspond to unstructured segments of the chromosome. The overall function and structure of these macrodomains is still unclear (Boccard, Esnault et al. 2005). It is generally thought that within these macrodomains rearrangements are possible but between macrodomains they are deleterious (Esnault, Valens et al. 2007). This means that certain inversions are not permitted and cannot be observed. All observed inversions in this study and in previous studies of *E.coli* fell within these boundaries.

The main effect of chromosomal imbalance caused by inversions is a change in chromosomal replication times as seen in other evolution experiments. In a short evolution experiment performed in *E. coli* O157:H7, several chromosome inversions occurred in the vicinity of the Ter region (Iguchi, Terajima et al. 2006) most of these inversion affect the genome balance of 30 to 83 kb. However, one inversion induced a significant unbalance of the replicore, close to 7.5% of the genome (Iguchi, Terajima et al. 2006). However cells dividing in the a minimum media depleted in glucose conditions used during the LTEE, , have a generation time of over 3.12hours and allow an off-balance of 10% of the total genome to be still viable.

### **Conclusions and Perspectives**

These data shows that rearrangements and insertion sequences have an important role in the adaptation of an organism to its environment due to the large range of genomic flexibility accorded. Both event types can create a varied mutation spectre including large ranging effects over entire operons. While the actual fitness gain of each rearrangement is unknown it can be speculated that, for at least, the parallel events give an advantage to the cell. The participation of IS elements is important as it shows that they have a more dynamic role in the cell than just gene interactions. These results also show that the populations are not clonal but rather heterogeneous.

This work focused on the evolution of rearrangements and insertion sequences (IS) in the longest ongoing evolution experiment (LTEE). Both rearrangements and IS play vital roles in the evolution of the populations. To continue this work it is essential to elucidate the exact benefits that are brought by the mutations. Some mutations have already been studied in previous works like the deletion of the ribose operon. Other deletions such as the rearrangement between *manB-cpsG* are candidates for investigation as it is deleted in ten out



of the twelve populations. A duplication worth investigating is the duplication containing *rpoS* found in two of the twelve populations by optical mapping. Since this study has mainly focused on IS movements in the population Ara+1, the dynamics of IS elements is worth investigating in other populations. On a longer term it is possible to deconstruct the rearrangements and the mutations in some evolved clones, focusing on Ara+1 at 2000 generations, and to replay the evolution experiment to see if the evolutionary trajectory is the same.

# References

- Abe, K., A. Yoshinari, et al. (2013). "Regulated DNA rearrangement during sporulation in *Bacillus weihenstephanensis* KBAB4." Mol Microbiol **90**(2): 415-427.
- Acar, J. F. and G. Moulin (2012). "Antimicrobial resistance: a complex issue." Rev Sci Tech **31**(1): 23-31.
- Adami, C. (2002). "What is complexity?" Bioessays **24**(12): 1085-1094.
- Adler, M., M. Anjum, et al. (2014). "High Fitness Costs and Instability of Gene Duplications Reduce Rates of Evolution of New Genes by Duplication-Divergence Mechanisms." Mol Biol Evol **31**(6): 1526-35
- Ahmed, M. O., P. D. Clegg, et al. (2010). "Antimicrobial resistance in equine faecal *Escherichia coli* isolates from North West England." Ann Clin Microbiol Antimicrob **9**: 12.
- Albertini, A. M., M. Hofer, et al. (1982). "On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions." Cell **29**(2): 319-328.
- Alekshun, M. N. and S. B. Levy (2007). "Molecular mechanisms of antibacterial multidrug resistance." Cell **128**(6): 1037-1050.
- Allen, H. K., J. Donato, et al. (2010). "Call of the wild: antibiotic resistance genes in natural environments." Nat Rev Microbiol **8**(4): 251-259.
- Anderson, P. and J. Roth (1981). "Spontaneous tandem genetic duplications in *Salmonella typhimurium* arise by unequal recombination between rRNA (rm) cistrons." Proc Natl Acad Sci U S A **78**(5): 3113-3117.
- Andersson, J. O. and S. G. Andersson (1999). "Genome degradation is an ongoing process in *Rickettsia*." Mol Biol Evol **16**(9): 1178-1191.

- Aubert, D., T. Naas, et al. (2006). "Functional characterization of IS1999, an IS4 family element involved in mobilization and expression of beta-lactam resistance genes." J Bacteriol **188**(18): 6506-6514.
- Barondess, J. J. and J. Beckwith (1990). "A bacterial virulence determinant encoded by lysogenic coliphage lambda." Nature **346**(6287): 871-874.
- Barondess, J. J. and J. Beckwith (1995). "bor gene of phage lambda, involved in serum resistance, encodes a widely conserved outer membrane lipoprotein." J Bacteriol **177**(5): 1247-1253.
- Barrick, J. E., D. S. Yu, et al. (2009). "Genome evolution and adaptation in a long-term experiment with Escherichia coli." Nature **461**(7268): 1243-1247.
- Bennett, G. M. and N. A. Moran (2013). "Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect." Genome Biol Evol **5**(9): 1675-1688.
- Biel, S. W. and D. E. Berg (1984). "Mechanism of IS1 transposition in E. coli: choice between simple insertion and cointegration." Genetics **108**(2): 319-330.
- Bierne, H., M. Seigneur, et al. (1997). "uvrD mutations enhance tandem repeat deletion in the Escherichia coli chromosome via SOS induction of the RecF recombination pathway." Mol Microbiol **26**(3): 557-567.
- Blount, Z. D., J. E. Barrick, et al. (2012). "Genomic analysis of a key innovation in an experimental Escherichia coli population." Nature **489**(7417): 513-518.
- Blount, Z. D., C. Z. Borland, et al. (2008). "Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli." Proc Natl Acad Sci U S A **105**(23): 7899-7906.
- Boccard, F., E. Esnault, et al. (2005). "Spatial arrangement and macrodomain organization of bacterial chromosomes." Mol Microbiol **57**(1): 9-16.

- Bordenstein, S. R. and W. S. Reznikoff (2005). "Mobile DNA in obligate intracellular bacteria." Nat Rev Microbiol **3**(9): 688-699.
- Bridges, C. B. (1936). "The Bar "Gene" a Duplication." Science **83**(2148): 210-211.
- Buddelmeijer, N. and J. Beckwith (2004). "A complex of the Escherichia coli cell division proteins FtsL, FtsB and FtsQ forms independently of its localization to the septal region." Mol Microbiol **52**(5): 1315-1327.
- Camarena, L., S. Poggio, et al. (1998). "An IS4 insertion at the glnA control region of Escherichia coli creates a new promoter by providing the -35 region of its 3'-end." Plasmid **39**(1): 41-47.
- Campbell, T. L. and E. D. Brown (2002). "Characterization of the depletion of 2-C-methyl-D-erythritol-2,4-cyclodiphosphate synthase in Escherichia coli and Bacillus subtilis." J Bacteriol **184**(20): 5609-5618.
- Canchaya, C., C. Proux, et al. (2003). "Prophage genomics." Microbiol Mol Biol Rev **67**(2): 238-276.
- Cerveau, N., S. Leclercq, et al. (2011). "Short- and long-term evolutionary dynamics of bacterial insertion sequences: insights from Wolbachia endosymbionts." Genome Biol Evol **3**: 1175-1186.
- Charlesworth, B., P. Sniegowski, et al. (1994). "The evolutionary dynamics of repetitive DNA in eukaryotes." Nature **371**(6494): 215-220.
- Charlier, D., J. Piette, et al. (1982). "IS3 can function as a mobile promoter in E. coli." Nucleic Acids Res **10**(19): 5935-5948.
- Charpentier, X., E. Kay, et al. (2011). "Antibiotics and UV radiation induce competence for natural transformation in Legionella pneumophila." J Bacteriol **193**(5): 1114-1121.

- Charpentier, X., P. Polard, et al. (2012). "Induction of competence for genetic transformation by antibiotics: convergent evolution of stress responses in distant bacterial species lacking SOS?" Curr Opin Microbiol **15**(5): 570-576.
- Chayot, R., B. Montagne, et al. (2010). "An end-joining repair mechanism in Escherichia coli." Proc Natl Acad Sci U S A **107**(5): 2141-2146.
- Chen, I., R. Provvedi, et al. (2006). "A macromolecular complex formed by a pilin-like protein in competent Bacillus subtilis." J Biol Chem **281**(31): 21720-21727.
- Chou, H. H., J. Berthet, et al. (2009). "Fast growth increases the selective advantage of a mutation arising recurrently during evolution under metal limitation." PLoS Genet **5**(9): e1000652.
- Clark, A. G. (1994). "Invasion and maintenance of a gene duplication." Proc Natl Acad Sci U S A **91**(8): 2950-2954.
- Clement, J. M., C. Wilde, et al. (1999). "IS1397 is active for transposition into the chromosome of Escherichia coli K-12 and inserts specifically into palindromic units of bacterial interspersed mosaic elements." J Bacteriol **181**(22): 6929-6936.
- Cooper, T. F., S. K. Remold, et al. (2008). "Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in Escherichia coli." PLoS Genet **4**(2): e35.
- Cooper, T. F., D. E. Rozen, et al. (2003). "Parallel changes in gene expression after 20,000 generations of evolution in Escherichiacoli." Proc Natl Acad Sci U S A **100**(3): 1072-1077.
- Cooper, V. S. and R. E. Lenski (2000). "The population genetics of ecological specialization in evolving Escherichia coli populations." Nature **407**(6805): 736-739.

- Cooper, V. S., D. Schneider, et al. (2001). "Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B." J Bacteriol **183**(9): 2834-2841.
- Cox, M. M., M. F. Goodman, et al. (2000). "The importance of repairing stalled replication forks." Nature **404**(6773): 37-41.
- Crow, J. F. (1993). "How much do we know about spontaneous human mutation rates?" Environ Mol Mutagen **21**(2): 122-129.
- Crozat, E., N. Philippe, et al. (2005). "Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection." Genetics **169**(2): 523-532.
- Crozat, E., C. Winkworth, et al. (2010). "Parallel genetic and phenotypic evolution of DNA superhelicity in experimental populations of *Escherichia coli*." Mol Biol Evol **27**(9): 2113-2128.
- Cui, L., H. M. Neoh, et al. (2012). "Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria." Proc Natl Acad Sci U S A **109**(25): E1647-1656.
- Curtis, T. P., W. T. Sloan, et al. (2002). "Estimating prokaryotic diversity and its limits." Proc Natl Acad Sci U S A **99**(16): 10494-10499.
- Darling, A. E., I. Miklos, et al. (2008). "Dynamics of genome rearrangement in bacterial populations." PLoS Genet **4**(7): e1000128.
- Darwin, C. (1859). On the Origin of Species.
- Daveran-Mingot, M. L., N. Campo, et al. (1998). "A natural large chromosomal inversion in *Lactococcus lactis* is mediated by homologous recombination between two insertion sequences." J Bacteriol **180**(18): 4834-4842.
- Davis, M. A., R. W. Simons, et al. (1985). "Tn10 protects itself at two levels from fortuitous activation by external promoters." Cell **43**(1): 379-387.

- De Meirman, C., C. Van Soom, et al. (1990). "Nucleotide sequence analysis of IS427 and its target sites in *Agrobacterium tumefaciens* T37." Plasmid **24**(3): 227-234.
- DeLeon-Rodriguez, N., T. L. Lathem, et al. (2013). "Microbiome of the upper troposphere: species composition and prevalence, effects of tropical storms, and atmospheric implications." Proc Natl Acad Sci U S A **110**(7): 2575-2580.
- Della, M., P. L. Palmbo, et al. (2004). "Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine." Science **306**(5696): 683-685.
- Derbyshire, K. M., M. Kramer, et al. (1990). "Role of instability in the cis action of the insertion sequence IS903 transposase." Proc Natl Acad Sci U S A **87**(11): 4048-4052.
- Doolittle, W. F., T. B. Kirkwood, et al. (1984). "Selfish DNAs with self-restraint." Nature **307**(5951): 501-502.
- Eisen, J. A., J. F. Heidelberg, et al. (2000). "Evidence for symmetric chromosomal inversions around the replication origin in bacteria." Genome Biol **1**(6): RESEARCH0011.
- Elsgaard, L. P., D (2011). "Hydrothermal vents in Lake Tanganyika harbor spore-forming thermophiles with extremely rapid growth." Journal of Great Lakes Research **37**(1): 203-206.
- Engelmoer, D. J. and D. E. Rozen (2011). "Competence increases survival during stress in *Streptococcus pneumoniae*." Evolution **65**(12): 3475-3485.
- England, J. L. (2013). "Statistical physics of self-replication." J Chem Phys **139**(12): 121923.
- Escoubas, J. M., M. F. Prere, et al. (1991). "Translational control of transposition activity of the bacterial insertion sequence IS1." EMBO J **10**(3): 705-712.
- Esnault, E., M. Valens, et al. (2007). "Chromosome structuring limits genome plasticity in *Escherichia coli*." PLoS Genet **3**(12): e226.



- Ferenci T. (2008) "The spread of a beneficial mutation in experimental bacterial populations: the influence of the environment and genotype on the fixation of rpoS mutations." *Heredity*; **100**(5): 446-452.
- Fléchar M, Gilot P. (2014), "Physiological impact of transposable elements encoding DDEtransposases in the environmental adaptation of *Streptococcus agalactiae*. *Microbiology*" **160**(7): 1298-1315
- Force, A., M. Lynch, et al. (1999). "Preservation of duplicate genes by complementary, degenerative mutations." *Genetics* **151**(4): 1531-1545.
- Fortier, L. C. and O. Sekulovic (2013). "Importance of prophages to evolution and virulence of bacterial pathogens." *Virulence* **4**(5): 354-365.
- Fraser, C. M., J. D. Gocayne, et al. (1995). "The minimal gene complement of *Mycoplasma genitalium*." *Science* **270**(5235): 397-403.
- Gaffe, J., C. McKenzie, et al. (2011). "Insertion sequence-driven evolution of *Escherichia coli* in chemostats." *J Mol Evol* **72**(4): 398-412.
- Gevers, D., K. Vandepoele, et al. (2004). "Gene duplication and biased functional retention of paralogs in bacterial genomes." *Trends Microbiol* **12**(4): 148-154.
- Ghigo, J. M. (2001). "Natural conjugative plasmids induce bacterial biofilm development." *Nature* **412**(6845): 442-445.
- Gil, R., E. Belda, et al. (2008). "Massive presence of insertion sequences in the genome of SOPE, the primary endosymbiont of the rice weevil *Sitophilus oryzae*." *Int Microbiol* **11**(1): 41-48.
- Gilbert, C. and R. Cordaux (2013). "Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes." *Genome Biol Evol* **5**(5): 822-832.

- Ginard, M., J. Lalucat, et al. (1997). "Genome organization of *Pseudomonas stutzeri* and resulting taxonomic and evolutionary considerations." Int J Syst Bacteriol **47**(1): 132-143.
- Grassle, J. F. (1985). "Hydrothermal vent animals: distribution and biology." Science **229**(4715): 713-717.
- Gueguen, E., P. Rousseau, et al. (2006). "Truncated forms of IS911 transposase downregulate transposition." Mol Microbiol **62**(4): 1102-1116.
- Haas, M. and B. Rak (2002). "Escherichia coli insertion sequence IS150: transposition via circular and linear intermediates." J Bacteriol **184**(21): 5833-5841.
- Halary, S., J. W. Leigh, et al. (2010). "Network analyses structure genetic diversity in independent genetic worlds." Proc Natl Acad Sci U S A **107**(1): 127-132.
- Hall, B. G. (1982). "Chromosomal mutation for citrate utilization by *Escherichia coli* K-12." J Bacteriol **151**(1): 269-273.
- Hallet, B., R. Rezsöházy, et al. (1991). "IS231A from *Bacillus thuringiensis* is functional in *Escherichia coli*: transposition and insertion specificity." J Bacteriol **173**(14): 4526-4529.
- Hamoen, L. W., G. Venema, et al. (2003). "Controlling competence in *Bacillus subtilis*: shared use of regulators." Microbiology **149**(Pt 1): 9-17.
- Haren, L., M. Betermier, et al. (1997). "IS911-mediated intramolecular transposition is naturally temperature sensitive." Mol Microbiol **25**(3): 531-540.
- Haren, L., B Ton-Hoang, et al. (1999). Integrating DNA: transposases and retroviral integrases. Annu Rev Microbiol **53**: 245-281.
- Hengge-Aronis, R. (1993). "Survival of hunger and stress: the role of rpoS in early stationary phase gene regulation in *E. coli*." Cell **72**(2): 165-168.

- Hindre, T., C. Knibbe, et al. (2012). "New insights into bacterial adaptation through in vivo and in silico experimental evolution." Nat Rev Microbiol **10**(5): 352-365.
- Hu, W. Y. and K. M. Derbyshire (1998). "Target choice and orientation preference of the insertion sequence IS903." J Bacteriol **180**(12): 3039-3048.
- Huisman, O. and N. Kleckner (1987). "A new generalizable test for detection of mutations affecting Tn10 transposition." Genetics **116**(2): 185-189.
- Iguchi, A., IS Terajima et al. (2006). "Spontaneous recombination between homologous prophage regions causes large-scale inversions within the Escherichia coli O157:H7 chromosome." Gene **372**: 199-207.
- Iida, S., R. Hiestand-Nauer, et al. (1985). "Transposable element IS1 intrinsically generates target duplications of variable length." Proc Natl Acad Sci U S A **82**(3): 839-843.
- Ishiguro, N., C. Oka, et al. (1979). "Plasmids in Escherichia coli controlling citrate-utilizing ability." Appl Environ Microbiol **38**(5): 956-964.
- Jain C. (1997). "Models for pairing of IS10 encoded antisense RNAs in vivo." J Theor Biol **186**(4): 431-439.
- Jellen-Ritter, A. S. and W. V. Kern (2001). "Enhanced expression of the multidrug efflux pumps AcrAB and AcrEF associated with insertion element transposition in Escherichia coli mutants Selected with a fluoroquinolone." Antimicrob Agents Chemother **45**(5): 1467-1472.
- Kallastu, A., R. Horak, et al. (1998). "Identification and characterization of IS1411, a new insertion sequence which causes transcriptional activation of the phenol degradation genes in Pseudomonas putida." J Bacteriol **180**(20): 5306-5312.
- Khan, A. I., D. M. Dinh, et al. (2011). "Negative epistasis between beneficial mutations in an evolving bacterial population." Science **332**(6034): 1193-1196.

- Kichenaradja, P., P. Siguier, et al. (2010). "ISbrowser: an extension of ISfinder for visualizing insertion sequences in prokaryotic genomes." Nucleic acids research **38**(Database issue): D62-68.
- Kleiner, M., J. C. Young, et al. (2013). "Metaproteomics reveals abundant transposase expression in mutualistic endosymbionts." MBio **4**(3): e00223-00213.
- Kondrashov, F. A. (2012). "Gene duplication as a mechanism of genomic adaptation to a changing environment." Proc Biol Sci **279**(1749): 5048-5057.
- Koonin, E. V. and T. V. Ilyina (1993). "Computer-assisted dissection of rolling circle DNA replication." Biosystems **30**(1-3): 241-268.
- Koser, S. A. (1924). "Correlation of Citrate Utilization by Members of the Colon-Aerogenes Group with Other Differential Characteristics and with Habitat." J Bacteriol **9**(1): 59-77.
- Koskiniemi, S., S. Sun, et al. (2012). "Selection-driven gene loss in bacteria." PLoS Genet **8**(6): e1002787.
- Kotewicz, M. L., S. A. Jackson, et al. (2007). "Optical maps distinguish individual strains of Escherichia coli O157 : H7." Microbiology **153**(Pt 6): 1720-1733.
- Krebs, M. P. and W. S. Reznikoff (1986). "Transcriptional and translational initiation sites of IS50. Control of transposase and inhibitor expression." J Mol Biol **192**(4): 781-791.
- Kresse, A. U., S. D. Dinesh, et al. (2003). "Impact of large chromosomal inversions on the adaptation and evolution of Pseudomonas aeruginosa chronically colonizing cystic fibrosis lungs." Mol Microbiol **47**(1): 145-158.
- Kugelberg, E., E. Kofoed, et al. (2006). "Multiple pathways of selected gene amplification during adaptive mutation." Proc Natl Acad Sci U S A **103**(46): 17319-17324.

- Kumar, R., S. Grover, et al. (2014). "IS30-related transposon mediated insertional inactivation of bile salt hydrolase (bsh1) gene of *Lactobacillus plantarum* strain Lp20." Microbiol Res **169**(7-8): 553-560.
- Le Gac, M., J. Plucain, et al. (2012). "Ecological and evolutionary dynamics of coexisting lineages during a long-term experiment with *Escherichia coli*." Proc Natl Acad Sci U S A **109**(24): 9487-9492.
- Lenski, R. E. (2004). Phenotypic and Genomic Evolution during a 20,000-Generation Experiment with the Bacterium *Escherichia coli*. Oxford, UK, John Wiley & Sons, Inc.
- Lenski, R. E. (2014). "The *E. coli* long-term experimental evolution project site <http://myxo.css.msu.edu/ecoli/strainsource.html>."
- Lenski, R. E., M. Rose, et al. (1991). "Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations." The American Naturalist **138**(6): 1315-1341.
- Lenski, R. E. and M. Travisano (1994). "Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations." Proc Natl Acad Sci U S A **91**(15): 6808-6814.
- Leonard, C. and J. Mahillon (1998). "IS231A transposition: conservative versus replicative pathway." Res Microbiol **149**(8): 549-555.
- Leyh, T. S., J. C. Taylor, et al. (1988). "The sulfate activation locus of *Escherichia coli* K12: cloning, genetic, and enzymatic characterization." J Biol Chem **263**(5): 2409-2416.
- Lynch, M. and J. S. Conery (2000). "The evolutionary fate and consequences of duplicate genes." Science **290**(5494): 1151-1155.
- Lynch, M. and A. Force (2000). "The probability of duplicate gene preservation by subfunctionalization." Genetics **154**(1): 459-473.

- Maharjan, R. P., J. Gaffe, et al. (2013). "A case of adaptation through a mutation in a tandem duplication during experimental evolution in *Escherichia coli*." *BMC Genomics* **14**: 441.
- Mahillon, J. and M. Chandler (1998). "Insertion sequences." *Microbiol Mol Biol Rev* **62**(3): 725-774.
- Matthews, T. D., W. Rabsch, et al. (2011). "Chromosomal rearrangements in *Salmonella enterica* serovar Typhi strains isolated from asymptomatic human carriers." *MBio* **2**(3): e00060-00011.
- Michel, B., H. Boubakri, et al. (2007). "Recombination proteins and rescue of arrested replication forks." *DNA Repair (Amst)* **6**(7): 967-980.
- Mira, A., H. Ochman, et al. (2001). "Deletional bias and the evolution of bacterial genomes." *Trends Genet* **17**(10): 589-596.
- Nagy, Z. and M. Chandler (2004). "Regulation of transposition in bacteria." *Res Microbiol* **155**(5): 387-398.
- Nakamura, K. and M. Inouye (1981). "Inactivation of the *Serratia marcescens* gene for the lipoprotein in *Escherichia coli* by insertion sequences, *IS1* and *IS5*; sequence analysis of junction points." *Mol Gen Genet* **183**(1): 107-114.
- Narayanan, S., J. K. Modak, et al. (2014). "Mechanism of *Escherichia coli* Resistance to Pyrrolic acid." *Antimicrob Agents Chemother* **58**(5): 2754-2762.
- Nasvall, J., L. Sun, et al. (2012). "Real-time evolution of new genes by innovation, amplification, and divergence." *Science* **338**(6105): 384-387.
- Neale, D. B., J. L. Wegrzyn, et al. (2014). "Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies." *Genome Biol* **15**(3): R59.
- Nevo, E. (2001). "Evolution of genome-phenome diversity under environmental stress." *Proc Natl Acad Sci U S A* **98**(11): 6233-6240.

- Nogueira, T., D. J. Rankin, et al. (2009). "Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence." Curr Biol **19**(20): 1683-1691.
- Novick, A. and T. Horiuchi (1961). "Hyper-production of beta-galactosidase by Escherichia coli bacteria." Cold Spring Harb Symp Quant Biol **26**: 239-245.
- Pan, Y. S., L. Yuan, et al. (2014). "A multidrug-resistance region containing blaCTX-M-65, fosA3 and rmtB on conjugative IncFII plasmids in Escherichia coli ST117 isolates from chicken." J Med Microbiol **63**(Pt 3): 485-488.
- Papadopoulos, D., D. Schneider, et al. (1999). "Genomic evolution during a 10,000-generation experiment with bacteria." Proc Natl Acad Sci U S A **96**(7): 3807-3812.
- Papp, B., C. Pal, et al. (2003). "Dosage sensitivity and the evolution of gene families in yeast." Nature **424**(6945): 194-197.
- Park, M. K., S. H. Lee, et al. (2014). "Enhancing recombinant protein production with an Escherichia coli host strain lacking insertion sequences." Appl Microbiol Biotechnol.
- Patel, M., Q. Jiang, et al. (2010). "A new model for SOS-induced mutagenesis: how RecA protein activates DNA polymerase V." Crit Rev Biochem Mol Biol **45**(3): 171-184.
- Pelosi, L., L. Kuhn, et al. (2006). "Parallel changes in global protein profiles during long-term experimental evolution in Escherichia coli." Genetics **173**(4): 1851-1869.
- Peterson, S., R. T. Cline, et al. (2000). "Gene expression analysis of the Streptococcus pneumoniae competence regulons by use of DNA microarrays." J Bacteriol **182**(21): 6192-6202.
- Philippe, N., E. Crozat, et al. (2007). "Evolution of global regulatory networks during a long-term experiment with Escherichia coli." Bioessays **29**(9): 846-860.
- Philippe, N., L. Pelosi, et al. (2009). "Evolution of penicillin-binding protein 2 concentration and cell shape during a long-term experiment with Escherichia coli." J Bacteriol **191**(3): 909-921.

- Plucain, J., T. Hindre, et al. (2014). "Epistasis and allele specificity in the emergence of a stable polymorphism in *Escherichia coli*." Science **343**(6177): 1366-1369.
- Pos, K. M., P. Dimroth, et al. (1998). "The *Escherichia coli* citrate carrier CitT: a member of a novel eubacterial transporter family related to the 2-oxoglutarate/malate translocator from spinach chloroplasts." J Bacteriol **180**(16): 4160-4165.
- Pranting, M. and D. I. Andersson (2011). "Escape from growth restriction in small colony variants of *Salmonella typhimurium* by gene amplification and mutation." Mol Microbiol **79**(2): 305-315.
- Prentki, P., B. Teter, et al. (1986). "Functional promoters created by the insertion of transposable element IS1." J Mol Biol **191**(3): 383-393.
- Pushker, R., A. Mira, et al. (2004). "Comparative genomics of gene-family size in closely related bacteria." Genome Biol **5**(4): R27.
- Putnam, N. H., T. Butts, et al. (2008). "The amphioxus genome and the evolution of the chordate karyotype." Nature **453**(7198): 1064-1071.
- Rau, M. H., R. L. Marvig, et al. (2012). "Deletion and acquisition of genomic content during early stage adaptation of *Pseudomonas aeruginosa* to a human host environment." Environ Microbiol **14**(8): 2200-2211.
- Reams, A. B., E. Kofoed, et al. (2010). "Duplication frequency in a population of *Salmonella enterica* rapidly approaches steady state with or without recombination." Genetics **184**(4): 1077-1094.
- Riehle, M. M., A. F. Bennett, et al. (2001). "Genetic architecture of thermal adaptation in *Escherichia coli*." Proc Natl Acad Sci U S A **98**(2): 525-530.
- Rocha, E. P. and A. Danchin (2003). "Gene essentiality determines chromosome organisation in bacteria." Nucleic Acids Res **31**(22): 6570-6577.



- Rohmer, L., C. Fong, et al. (2007). "Comparison of *Francisella tularensis* genomes reveals evolutionary events associated with the emergence of human pathogenic strains." Genome Biol **8**(6): R102.
- Rosner, J. L. and M. S. Guyer (1980). "Transposition of *IS1*- $\lambda$ BIO-*IS1* from a bacteriophage lambda derivative carrying the *IS1*-cat-*IS1* transposon (Tn9)." Molecular & general genetics : MGG **178**(1): 111-120.
- Roth, D. B. and J. H. Wilson (1985). "Relative rates of homologous and nonhomologous recombination in transfected DNA." Proc Natl Acad Sci U S A **82**(10): 3355-3359.
- Roth JR, B. N., Galitski T, Haack K, Lawrence JG, Miesel L (1996). Rearrangements of the bacterial chromosome: formation and applications. Washington D.C, ASM Press.
- Rozen, D. E. and R. E. Lenski (2000). "Long-Term Experimental Evolution in *Escherichia coli*. VIII. Dynamics of a Balanced Polymorphism." Am Nat **155**(1): 24-35.
- Rozen, D. E., D. Schneider, et al. (2005). "Long-term experimental evolution in *Escherichia coli*. XIII. Phylogenetic history of a balanced polymorphism." J Mol Evol **61**(2): 171-180.
- Sahney, S., M. J. Benton, et al. (2010). "Links between global taxonomic diversity, ecological diversity and the expansion of vertebrates on land." Biol Lett **6**(4): 544-547.
- Saito, T., T. Chibazakura, et al. (2010). "Measurements of transposition frequency of insertion sequence *IS1* by GFP hop-on assay." J Gen Appl Microbiol **56**(3): 187-192.
- Schaaper, R. M. (1993). "Base selection, proofreading, and mismatch repair during DNA replication in *Escherichia coli*." J Biol Chem **268**(32): 23762-23765.
- Schmidt, H. and M. Hensel (2004). "Pathogenicity islands in bacterial pathogenesis." Clin Microbiol Rev **17**(1): 14-56.

- Schmitz-Esser, S., T. Penz, et al. (2011). "A bacterial genome in transition--an exceptional enrichment of IS elements but lack of evidence for recent transposition in the symbiont *Amoebophilus asiaticus*." BMC Evol Biol **11**: 270.
- Schneider, D., E. Duperchy, et al. (2000). "Long-term experimental evolution in *Escherichia coli*. IX. Characterization of insertion sequence-mediated mutations and rearrangements." Genetics **156**(2): 477-488.
- Schneider, D. and R. E. Lenski (2004). "Dynamics of insertion sequence elements during experimental evolution of bacteria." Res Microbiol **155**(5): 319-327.
- Schuch, R. and V. A. Fischetti (2009). "The secret life of the anthrax agent *Bacillus anthracis*: bacteriophage-mediated ecological adaptations." PLoS One **4**(8): e6532.
- Schwartz, D. C., X. Li, et al. (1993). "Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping." Science **262**(5130): 110-114.
- Sebat, J., B. Lakshmi, et al. (2007). "Strong association of de novo copy number mutations with autism." Science **316**(5823): 445-449.
- Segall, A., M. J. Mahan, et al. (1988). "Rearrangement of the bacterial chromosome: forbidden inversions." Science **241**(4871): 1314-1318.
- Shukla, S. K., J. Kislow, et al. (2009). "Optical mapping reveals a large genetic inversion between two methicillin-resistant *Staphylococcus aureus* strains." J Bacteriol **191**(18): 5717-5723.
- Siguiet, P., E. Gourbeyre, et al. (2014). "Bacterial insertion sequences: their genomic impact and diversity." FEMS Microbiol Rev.
- Silver-Mysliwiec, T. H. and M. G. Bramucci (1990). "Bacteriophage-enhanced sporulation: comparison of spore-converting bacteriophages PMB12 and SP10." J Bacteriol **172**(4): 1948-1953.

- Skippington, E. and M. A. Ragan (2011). "Lateral genetic transfer and the construction of genetic exchange communities." FEMS Microbiol Rev **35**(5): 707-735.
- Sleight, S. C., C. Orlic, et al. (2008). "Genetic basis of evolutionary adaptation by *Escherichia coli* to stressful cycles of freezing, thawing and growth." Genetics **180**(1): 431-443.
- Sniegowski, P. D., P. J. Gerrish, et al. (1997). "Evolution of high mutation rates in experimental populations of *E. coli*." Nature **387**(6634): 703-705.
- Sousa, A., C. Bourgard, et al. (2013). "Rates of transposition in *Escherichia coli*." Biol Lett **9**(6): 20130838.
- Spencer-Smith, R., E. M. Varkey, et al. (2012). "Sequence features contributing to chromosomal rearrangements in *Neisseria gonorrhoeae*." PLoS One **7**(9): e46023.
- Stanley, N. R. and B. A. Lazazzera (2004). "Environmental signals and regulatory pathways that influence biofilm formation." Mol Microbiol **52**(4): 917-924.
- Stoebel, D. M. and C. J. Dorman (2010). "The effect of mobile element *IS10* on experimental regulatory evolution in *Escherichia coli*." Mol Biol Evol **27**(9): 2105-2112.
- Stouf, M., J. C. Meile, et al. (2013). "FtsK actively segregates sister chromosomes in *Escherichia coli*." Proc Natl Acad Sci U S A **110**(27): 11157-11162.
- Sun, S., R. Ke, et al. (2012). "Genome-wide detection of spontaneous chromosomal rearrangements in bacteria." PLoS One **7**(8): e42639.
- Tal, A., R. Arbel-Goren, et al. (2014). "Location of the unique integration site on an *Escherichia coli* chromosome by bacteriophage lambda DNA in vivo." Proc Natl Acad Sci U S A **111**(20): 7308-12
- Tetu, S. G. and A. J. Holmes (2008). "A family of insertion sequences that impacts integrons by specific targeting of gene cassette recombination sites, the *IS1111*-attC Group." J Bacteriol **190**(14): 4959-4970.

- Tillier, E. R. and R. A. Collins (2000). "Genome rearrangement by replication-directed translocation." Nat Genet **26**(2): 195-197.
- Tobes, R. and E. Pareja (2006). "Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements." BMC Genomics **7**: 62.
- Touchon, M. and E. P. Rocha (2007). "Causes of insertion sequences abundance in prokaryotic genomes." Mol Biol Evol **24**(4): 969-981.
- Toussaint, A. and M. Chandler (2012). "Prokaryote genome fluidity: toward a system approach of the mobilome." Methods Mol Biol **804**: 57-80.
- Turner, P. C., L. P. Yomano, et al. (2012). "Optical mapping and sequencing of the Escherichia coli KO11 genome reveal extensive chromosomal rearrangements, and multiple tandem copies of the Zymomonas mobilis pdc and adhB genes." J Ind Microbiol Biotechnol **39**(4): 629-639.
- Umbarger, M. A., E. Toro, et al. (2011). "The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation." Mol Cell **44**(2): 252-264.
- Varani, A. M., C. B. Monteiro-Vitorello, et al. (2013). "The role of prophage in plant-pathogenic bacteria." Annu Rev Phytopathol **51**: 429-451.
- Vogele, K., E. Schwartz, et al. (1991). "High-level ribosomal frameshifting directs the synthesis of IS150 gene products." Nucleic Acids Res **19**(16): 4377-4385.
- Waldor, M. K. and J. J. Mekalanos (1996). "Lysogenic conversion by a filamentous phage encoding cholera toxin." Science **272**(5270): 1910-1914.
- Ward, R. and R. Durrett (2004). "Subfunctionalization: How often does it occur? How long does it take?" Theor Popul Biol **66**(2): 93-100.
- Weller, G. R., B. Kysela, et al. (2002). "Identification of a DNA nonhomologous end-joining complex in bacteria." Science **297**(5587): 1686-1689.

- Whitman, W. B., D. C. Coleman, et al. (1998). "Prokaryotes: the unseen majority." Proc Natl Acad Sci U S A **95**(12): 6578-6583.
- Whoriskey, S. K., V. H. Nghiem, et al. (1987). "Genetic rearrangements and gene amplification in Escherichia coli: DNA sequences at the junctures of amplified gene fusions." Genes Dev **1**(3): 227-237.
- Wielgoss, S., J. E. Barrick, et al. (2013). "Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load." Proc Natl Acad Sci U S A **110**(1): 222-227.
- Wiser, M. J., N. Ribeck, et al. (2013). "Long-term dynamics of adaptation in asexual populations." Science **342**(6164): 1364-1367.
- Woods, R., D. Schneider, et al. (2006). "Tests of parallel molecular evolution in a long-term experiment with Escherichia coli." Proc Natl Acad Sci U S A **103**(24): 9107-9112.
- Zerbib, D., P. Gamas, et al. (1985). "Specificity of insertion of IS1." J Mol Biol **185**(3): 517-524.
- Zerbib, D., P. Polard, et al. (1990). "The regulatory role of the IS1-encoded InsA protein in transposition." Mol Microbiol **4**(3): 471-477.
- Zhang, M., E. S. M. Pereira, et al. (2014). "The mycosphere constitutes an arena for horizontal gene transfer with strong evolutionary implications for bacterial-fungal interactions." FEMS Microbiol Ecol [Epub ahead of print]
- Ziebuhr, W., V. Krimmer, et al. (1999). "A novel mechanism of phase variation of virulence in Staphylococcus epidermidis: evidence for control of the polysaccharide intercellular adhesin synthesis by alternating insertion and excision of the insertion sequence element IS256." Mol Microbiol **32**(2): 345-356.
- Zieg, J. and S. R. Kushner (1977). "Analysis of genetic recombination between two partially deleted lactose operons of Escherichia coli K-12." J Bacteriol **131**(1): 123-132.

Zivanovic, Y., P. Lopez, et al. (2002). "Pyrococcus genome comparison evidences chromosome shuffling-driven evolution." Nucleic Acids Res **30**(9): 1902-1910.

## Résumé

Les réarrangements d'ADN à grande échelle, tels que inversions, amplifications, duplications, délétions, insertions et transposition des éléments génétiques mobiles, sont des acteurs essentiels de l'évolution. Ils ont une forte incidence sur l'organisation et l'expression des chromosomes, ce qui affecte le phénotype des organismes. Toutefois, la dynamique de ces réarrangements au cours de l'évolution et leurs effets sur l'adaptation des organismes sont souvent inconnus. Nous avons abordé ces questions en utilisant la plus longue expérience d'évolution en cours. A partir d'un ancêtre commun d'*Escherichia coli*, douze populations indépendantes sont cultivées dans un milieu limité en glucose depuis plus de 60 000 générations, soit 26 ans. La plupart des études antérieures ont porté sur les mutations ponctuelles et les petites insertions et délétions (InDels). En utilisant des clones isolés au cours du temps dans ces 12 populations, nous avons caractérisé les réarrangements d'ADN à grande échelle à la fois par l'analyse des séquences de génomes et par cartographie optique. A 40 000 générations, nous avons identifié 110 réarrangements parmi lesquelles 82 délétions, 19 inversions et 9 duplications. Plusieurs régions du chromosome ont été touchées à plusieurs reprises par le même type de réarrangements dans des populations indépendantes. Dans une des populations au moins, les réarrangements se sont produits au début de l'expérience d'évolution, au moment où l'augmentation de la valeur sélective est la plus élevée. Par conséquent, certains de ces réarrangements pourraient être bénéfiques dans ces conditions. Même dans le cas contraire, nous avons montré que ces réarrangements affectaient fortement la structure du chromosome au cours de l'expérience d'évolution.

Au niveau moléculaire, nous avons montré que ~ 70% des réarrangements se produisent par recombinaison entre séquences d'insertion (IS), ce qui illustre l'importance de ces dernières dans la plasticité du génome. Nous avons donc caractérisé la distribution et la dynamique de ces petits éléments génétiques mobiles dans l'ensemble des 12 populations. Nous avons montré que les éléments IS ont fortement contribué à l'ensemble des mutations après 40 000 générations. Dans une population, les IS représentent même la moitié des mutations, et un des types d'IS, *IS150*, présente une forte prolifération avec 6 fois plus de copies à 40 000 générations, intervenant dans la plupart des réarrangements détectés dans cette population. Nous avons montré une forte dynamique temporelle d'*IS150*, avec une forte expansion suivie d'une domestication par l'hôte. En testant trois scénarii évolutifs, nous avons démontré que l'expansion d'*IS150* était liée à une forte augmentation de la valeur sélective conférée par les événements initiaux de transposition ayant eu lieu avant 2000 générations. Plus tard, entre 20 000 et 40 000 générations, nous avons mesuré une diminution de la fréquence de transposition, probablement en raison d'une régulation négative de la transposition imposée par l'hôte. Enfin, et pour la première fois, nous avons développé un modèle d'évolution de la dynamique des IS, qui confirme que leur expansion est liée à un nombre seuil d'insertions bénéfiques initiales. Ces résultats montrent que les réarrangements chromosomiques à grande échelle et les éléments IS ont joué un rôle actif dans la trajectoire évolutive au cours de 40 000 générations d'évolution bactérienne.