

## The role of recombination in evolutionary adaptation of *Escherichia coli* to a novel nutrient

H.-Y. CHU\*, K. SPROUFFSKE\* & A. WAGNER\*†‡

\*Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

†The Swiss Institute of Bioinformatics, Quartier Sorge, Batiment Genopode, Lausanne, Switzerland

‡The Santa Fe Institute, Santa Fe, NM, USA

### Keywords:

adaptation;  
experimental evolution;  
genomics.

### Abstract

The benefits and detriments of recombination for adaptive evolution have been studied both theoretically and experimentally, with conflicting predictions and observations. Most pertinent experiments examine recombination's effects in an unchanging environment and do not study its genomewide effects. Here, we evolved six replicate populations of either highly recombining  $R^+$  or lowly recombining  $R^-$  *E. coli* strains in a changing environment, by introducing the novel nutrients L-arabinose or indole into the environment. The experiment's ancestral strains are not viable on these nutrients, but 130 generations of adaptive evolution were sufficient to render them viable. Recombination conferred a more pronounced advantage to populations adapting to indole. To study the genomic changes associated with this advantage, we sequenced the genomes of 384 clones isolated from selected replicates at the end of the experiment. These genomes harbour complex changes that range from point mutations to large-scale DNA amplifications. Among several candidate adaptive mutations, those in the tryptophanase regulator *tnaC* stand out, because the *tna* operon in which it resides has a known role in indole metabolism. One of the highly recombining populations also shows a significant excess of large-scale segmental DNA amplifications that include the *tna* operon. This lineage also shows a unique and potentially adaptive combination of point mutations and DNA amplifications that may have originated independently from one another, to be joined later by recombination. Our data illustrate that the advantages of recombination for adaptive evolution strongly depend on the environment and that they can be associated with complex genomic changes.

### Introduction

The molecular mechanisms that mediate recombination may have originated to repair DNA damage and maintain genome integrity on short, physiological time scales (Bernstein, 1977; Michod *et al.*, 1988, 2008; Takeuchi *et al.*, 2014; Croucher *et al.*, 2016). However,

recombination can also affect organisms on longer, evolutionary time scales. Especially the potential benefits of recombination for adaptive evolution have been of long-standing interest to evolutionary biologists (reviewed in Weismann, 1889; Fisher, 1930a,b; Muller, 1932; Hill & Robertson, 1966; Felsenstein, 1974). They have been studied both theoretically and experimentally (reviewed in Barton & Charlesworth, 1998; Otto & Lenormand, 2002; Michod *et al.*, 2008; Vos, 2009; Hartfield & Keightley, 2012; Vos *et al.*, 2015; Pesce *et al.*, 2016; Ambur *et al.*, 2016). The prevailing hypotheses to explain the benefits of recombination involve its ability to avoid Hill–Robertson interference (Hill & Robertson, 1966), in which tight linkage between two alleles can interfere with adaptive evolution. By changing allelic associations,

*Correspondence:* Andreas Wagner, Department of Evolutionary Biology and Environmental Studies, Y27-J-54, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland.  
Tel.: +41 44 635 61 41; fax: +41 44 635 6144;  
e-mail: andreas.wagner@ieu.uzh.ch  
Whole genome sequencing data will be deposited at the European Nucleotide Archive (Project accession: PRJEB18430).

recombination can aid selection, either helping to bring two beneficial alleles into the same genome (Hill & Robertson, 1966; Felsenstein, 1974), or removing deleterious alleles from an otherwise well-adapted genome (Muller, 1964). Conversely, recombination can also have detrimental effects on adaptive evolution, breaking apart beneficial allele combinations and thus reducing fitness (Agrawal *et al.*, 2005; Ram & Hadany, 2016). Whether recombination has a net beneficial or detrimental effect depends on many genetic factors, including epistasis (Kouyos *et al.*, 2007), population size, mutation rate, recombination rate and environmental changes (Bretscher *et al.*, 2004; Nagaraja *et al.*, 2016; Whitlock *et al.*, 2016). Organisms may also modulate their recombination rate, increasing recombination in novel, stressful environments so as to bring together beneficial alleles, and decreasing it in well-adapted lineages to preserve combinations of beneficial alleles (Hadany & Beker, 2003; Agrawal *et al.*, 2005; Griffiths & Bonser, 2013; Ram & Hadany, 2016). The contradictory predictions of theoretical studies about the benefits or detriments of recombination make it challenging to explain the high prevalence of sex and recombination in nature (Otto & Lenormand, 2002; Lehtonen *et al.*, 2012; Baltrus, 2013). This limitation of theory calls for experiments to study the effect of recombination on adaptive evolution under controlled laboratory conditions.

Past laboratory evolution experiments have studied the effect of recombination in several types of laboratory conditions. In one such experiment, *E. coli* populations pre-adapted for 2000 generations to a glucose minimal medium were allowed to evolve for an additional 1000 generations in the same environment, but subject to recombination (Souza *et al.*, 1997). Under these conditions, recombination increased genetic variation but did not accelerate further adaptive evolution. In contrast, a more recent study under similar conditions, but in the yeast *Saccharomyces cerevisiae* (McDonald *et al.*, 2016) showed that recombination can accelerate adaptation by purging deleterious mutations. More generally, studies that expose evolving populations to nutrient-limiting conditions, unusual carbon sources, or stressful environments (e.g. containing antibiotics) tend to show that sexually recombining populations can outperform asexual populations (Zeyl & Bell, 1997; Goddard *et al.*, 2005; Cooper, 2007; Gray & Goddard, 2012; Winkler & Kao, 2012; Peabody *et al.*, 2016). Removal of deleterious alleles may play an important role in such beneficial recombination (Zeyl & Bell, 1997; Cooper, 2007; Peabody *et al.*, 2016). Quantitatively and qualitatively different effects of recombination on adaptation in such studies may have various causes, including the kinds of adaptation considered, the number of loci underlying an adaptive trait and the fitness effects of individual mutations. For instance, one study (Zeyl & Bell, 1997) found that recombination did not accelerate adaptation to a novel

environment (galactose), if adaptation required only a few beneficial mutations of large effect, and if recombination did not affect the removal of deleterious mutations substantially.

With one exception (McDonald *et al.*, 2016), previous experimental evolution studies did not examine the genomic basis of adaptation subject to recombination. Here, we ask whether and how recombination can accelerate the evolution of novel adaptive traits by evolving two *E. coli* strains with high ( $R^+$ ) and low ( $R^-$ ) recombination rates (Table S1) towards viability on novel nutrients, and examine the basis of the resulting adaptations by genome sequencing.

Bacteria have highly changeable genomes, continuously acquiring and losing genes via horizontal gene transfer (Ochman *et al.*, 2000; Vos *et al.*, 2015; Koonin, 2016). Such transfer can be mediated by three major mechanisms (reviewed in Thomas & Nielsen, 2005; Arber, 2014). First, 'competent' bacteria can import naked DNA from the environment via transformation. Second, cells can exchange genetic material (e.g. integrative and conjugative elements) via bacterial conjugation. Third, they can receive DNA from defective phages acting as gene transfer agents through viral transduction. Internalized DNA molecules that cannot replicate may integrate into the host genome via several homology-based recombination mechanisms that minimally require 4 to 25 bp of sequence homology (Shen & Huang, 1986; Thomas & Nielsen, 2005; Hastings *et al.*, 2009a; Kingston *et al.*, 2015). Once integrated into the recipient genome, genes in newly acquired DNA can be expressed and may bring about novel traits that can allow rapid adaptation (Ochman *et al.*, 2000; Vos *et al.*, 2015). Recombination in *E. coli* may occur at least as frequently as spontaneous mutation (Touchon *et al.*, 2009; Didelot *et al.*, 2012; Rodríguez-Beltrán *et al.*, 2015; Lin & Kussell, 2016). Individual recombination events can have manifold outcomes. They include gene conversion (Didelot *et al.*, 2012), gene gain (Shen *et al.*, 2013), gene loss (Croucher *et al.*, 2016), large-scale deletions and segmental amplifications (Hastings *et al.*, 2009b). The last two kinds of changes are especially important, because they can have much greater phenotypic effects than single point mutations (Bobay *et al.*, 2015). The incidence of segmental amplifications may be increased by stressors (Hastings *et al.*, 2009a,b), which can also lead to hypermutation (Rosenberg *et al.*, 2012).

Building on previous work (Souza *et al.*, 1997; Cooper, 2007), we focus here on the kind of recombination that occurs when a cell receives foreign DNA from a bacterial donor via conjugation (Fig. 1). In deviation from this work, however, we ask whether recombination confers an advantage in the early stages of adaptive evolution to a novel and challenging nutrient environment, in which the starting *E. coli* strains are not viable (Fig. S1). These starting strains differ strongly in their recombination rate

but have similar mutation rates (Winkler & Kao, 2012). Specifically, we conducted two parallel experiments, each of them with six replicate  $R^+$  (highly recombining) and six replicate  $R^-$  (lowly recombining) populations, and each of them lasting for 40 days ( $\sim 132$  generations). In the first experiment, we gradually shifted the carbon source from glycerol, on which both ancestral strains are viable, to L-arabinose, on which neither ancestral strain is viable (Fig. 1b). In the second experiment, we analogously changed the growth medium from one containing L-tryptophan, which both ancestors require for growth, to the tryptophan precursor indole, which cannot serve as a substitute for tryptophan to either ancestor. To find out if recombination speeds up adaptation, we compared the fraction of individuals viable on agar plates containing the novel nutrient (Fig. 1b) between the  $R^+$  and  $R^-$  replicate populations during the experiment. In addition, at the end of the experiment, we compared the fitness of multiple clones from both kinds of populations in liquid medium. Furthermore, we characterized the genomic changes during adaptive evolution by sequencing 382 genomes of clones isolated from selected replicate populations, and from the  $R^+$  and  $R^-$  ancestors.

## Materials and methods

### Strains

We evolved replicate populations of two *E. coli* strains towards new metabolic phenotypes (L-arabinose and indole prototrophy (Text S1)). We designate the starting (ancestral) strains for this experiment the  $R^+$  and  $R^-$  strains. The  $R^+$  strain has a 25-fold higher conjugation-induced recombination rate than the  $R^-$  strain (Winkler & Kao, 2012). Both strains are BW25113 derivatives provided by the Kao laboratory (Table S1) (Winkler & Kao, 2012). They are unable to utilize L-arabinose due to a  $\Delta araB-D$  deletion in the BW25113 background. Both strains also harbour a conjugative F plasmid and three origins of transfer sequences (*OriT*) in their genome (Fig. 1a) (Typas *et al.*, 2008; Winkler & Kao, 2012). The conjugative F plasmid is inserted into the *trp* operon, rendering both strains auxotrophic for tryptophan. The integrate encodes the *tra* genes, which mediate bacterial conjugation, and one *OriT* sequence. In addition, both strains contain two *OriT* sequences in the pseudogenes *mbhA* and *hyfC*. *OriT* is the genomic sequence recognized by the conjugative proteins that initiate the transfer of DNA to the bacterial recipient at this sequence. Recombination between the recipient genome and the transferred chromosomal DNA fragment can occur at homologous genomic regions after the DNA transfer. The *traST* genes, encoding proteins that influence conjugation specificity, have been inactivated in the  $R^+$  strains and account for the 25-fold increase in recombination rate in the  $R^+$  strain (Winkler

& Kao, 2012). The  $R^+$  and  $R^-$  strains have similar (maximum likelihood estimated) mutation rates, which are  $2.13 \times 10^{-8}$  and  $1.92 \times 10^{-8}$  per nucleotide and generation, respectively (Winkler & Kao, 2012). Because of its construction, the  $R^+$  strain is resistant to gentamycin, kanamycin and chloramphenicol, whereas the  $R^-$  strain is resistant to gentamycin only.

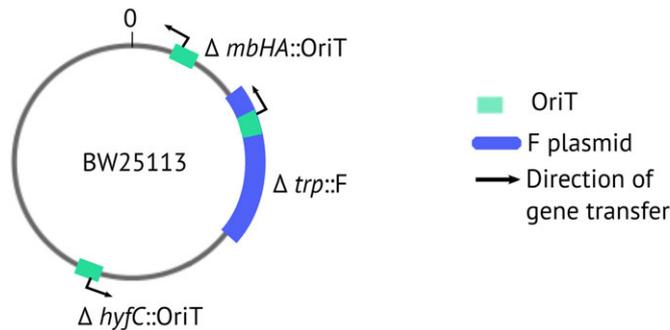
### Growth media and conditions

We performed all experiments in Davis minimal (DM) medium (Sigma-Aldrich 93753; Sigma-Aldrich Chemie GmbH, Buchs, Switzerland) supplemented with either glucose (Sigma-Aldrich G7021), glycerol (Sigma-Aldrich G2025), L-arabinose (Sigma-Aldrich A3256), L-tryptophan (Sigma-Aldrich T8941) or indole (Sigma-Aldrich I3408) according to the evolution protocol specified below. Thiamine hydrochloride (vitamin B12, Sigma-Aldrich T4625, 0.001%) is essential for *E. coli* growth, and we thus added it to all DM media. We grew all cultures at 37 °C in a shaking incubator (Edmund Bühler TH30; Edmund Bühler GmbH, Hechingen, Germany) at 100 rpm. We used lysogeny broth (LB, also known as Luria-Bertani broth) (Becton, Dickinson and Company, Eysins, Switzerland Difco 244610) for preparing solid agar in the experiment. We set up all overnight cultures in DM medium supplemented with 0.2% glucose and 50  $\mu\text{g mL}^{-1}$  tryptophan. We used three antibiotics in the experiments: chloramphenicol (Sigma-Aldrich C0378) (25  $\mu\text{g mL}^{-1}$ ), gentamicin sulphate (AppliChem A14920010; AXON LAB AG, Baden-Dättwil, Switzerland) (30  $\mu\text{g mL}^{-1}$ ) and kanamycin sulphate (Sigma-Aldrich K1377) (50  $\mu\text{g mL}^{-1}$ ) as described below.

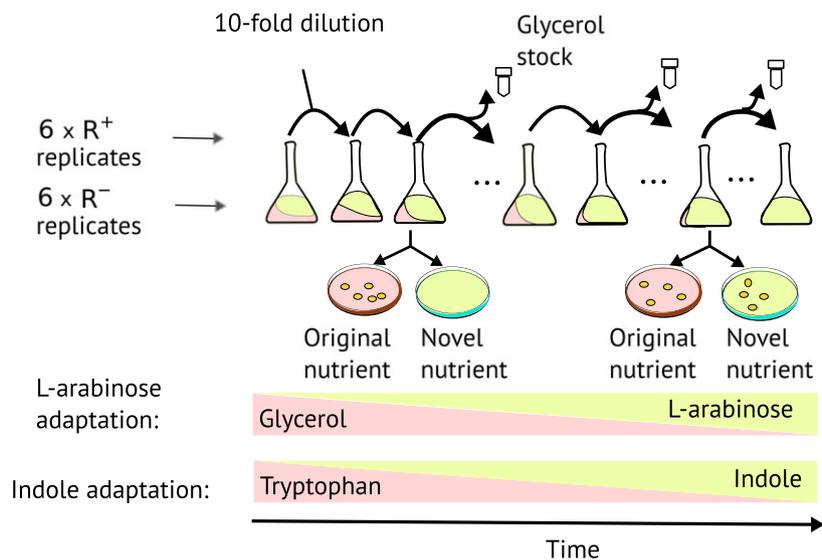
### Growth characteristics of ancestral $R^+$ and $R^-$ strains

To confirm that the ancestral  $R^+$  and  $R^-$  strains cannot grow with only L-arabinose as a carbon source and in the absence of tryptophan (Fig. S1), we first grew two replicates of 5 mL overnight cultures of the  $R^+$ , the  $R^-$ , and the *E. coli* K12 MG1655 strain (as a positive control) in appropriate media. Subsequently, we washed 1 mL of the overnight culture twice in DM medium without added nutrients, resuspended cells in 1 mL DM medium, added 2  $\mu\text{L}$  of cell suspension to 198  $\mu\text{L}$  DM medium supplemented with various concentrations of glycerol/L-arabinose or tryptophan/indole (Table S2), and let the resulting cultures grow for 24 h. We set the combined concentration of glycerol and L-arabinose to 0.2%, and the combined concentration of tryptophan and indole to 50  $\mu\text{g mL}^{-1}$  in all growth conditions, as previously described (Hall, 1993; Winkler & Kao, 2012). We kept the total combined nutrient concentration constant for all conditions to ensure that a lack of growth in a given condition was not simply due to starvation, but to the inability to metabolize a nutrient.

## (a) Ancestral strain genotype



## (b) Experimental evolution



**Fig. 1** Ancestral strain genotypes and experimental design. (a)  $R^+$  and  $R^-$  strain genotypes. Both  $R^+$  and  $R^-$  strains are derived from BW25113. An F plasmid is inserted into the *trp* operon, disrupting the operon and leading to tryptophan auxotrophy of both ancestral strains. Conjugative transfer of chromosomal DNA can originate from the internal origin of transfer (*OriT*) sequence within the F plasmid. In addition, *OriT* sequences have been inserted at the pseudogenes *mbhA* (5.29 min) and *hyfC* (56.08 min), where chromosomal DNA transfer can also be initiated. (b) Design of the evolution experiment. We seeded six replicates each from the  $R^+$  and  $R^-$  strains and adapted each of them to two novel nutrients in independent experiments (6 replicates  $\times$  2 strains  $\times$  2 nutrients = 24 independently evolving populations). Every 24 h, we transferred a sample of the evolving population into fresh growth medium (10-fold dilution). In the L-arabinose adaptation experiment, we slowly replaced glycerol with L-arabinose, and in the indole-adaptation experiment, we slowly replaced tryptophan with indole. Every five days, we archived samples of the evolving populations as glycerol stocks and monitored the fraction of cells able to metabolize the novel nutrient by plating. At the end of the experiment, we estimated the fitness of each evolved population as the fraction of cells able to grow on novel nutrient plates among all cells able to grow on nonselective plates.

We then repeated this 100-fold dilution and 24-h growth procedure twice, for a total of three dilution and growth cycles. We assessed the optical density (OD) at 600 nm as a proxy for population growth at the beginning and end of each transfer cycle. All strains grew well when supplemented with glycerol and tryptophan (Fig. S1). As expected, reduced levels of glycerol and tryptophan caused reduced growth for the  $R^+$  and  $R^-$  strains. Furthermore, we confirmed that DM media

supplemented with only tryptophan, indole or thiamine (Fig. S1) could not sustain growth, because they do not yield any OD readings above the background (Fig. S1, condition 12). We identified the minimum glycerol and tryptophan conditions that support growth of our strains at the end of the last growth period as 0.05% glycerol and  $10 \mu\text{g mL}^{-1}$  of tryptophan (Fig. S1). These served as the starting conditions for the evolution experiments.

## Experimental evolution

We conducted two evolution experiments, each lasting 40 days ( $\sim 132$  generations) (Fig. 1b). In the first, we gradually replaced glycerol with L-arabinose to assess whether the  $R^+$  and  $R^-$  strains could evolve to use L-arabinose as the main carbon source. We started the experiment with 0.05% of glycerol and decreased the glycerol concentration every five days according to the schedule outlined in Table S3. Each decrease in glycerol was matched by an increase in L-arabinose until glycerol had been entirely replaced by L-arabinose. In the second experiment, we gradually replaced tryptophan with indole in the same fashion to assess whether the  $R^+$  and  $R^-$  populations could restore tryptophan biosynthesis. Specifically, we started the experiment with  $10 \mu\text{g mL}^{-1}$  of tryptophan and decreased this concentration every five days. We compensated each decrease in tryptophan by an equal increase in indole until the tryptophan had been completely replaced by indole (Table S3). After complete replacement of the old nutrient with the novel nutrient, we evolved the replicates for ten further days to ensure that adaptation and not just persistence had occurred.

We initialized our experiments with a 5 mL overnight culture of each  $R^+$  and  $R^-$  strain. We washed 2 mL of each culture twice and resuspended cells in 2 mL of DM medium not supplemented with glycerol or tryptophan. We added 200  $\mu\text{L}$  of this cell suspension to 1800  $\mu\text{L}$  of growth medium (specified in Table S3) to seed six independent replicates from both  $R^+$  and  $R^-$  strains in 12 wells of a 24-well round-bottom plate (Sigma-Aldrich WHA77015102). We diluted these cultures 10-fold into fresh medium every 24 h ( $\sim 3.3$  generations) and prepared glycerol stocks (15% v/v final glycerol) every five days. This dilution protocol amounts to a periodic population bottlenecking with post-bottleneck population sizes between  $1 \times 10^4$  and  $1 \times 10^7$  cells, depending on how fast the replicate populations grow. To reduce the possibility of cross-contamination, we arranged the populations in the plates in a checker-board pattern, where wells containing cell culture are interspersed with wells containing only medium. We also checked for potential contamination of  $R^-$  populations with  $R^+$  cells by plating cells on plates containing gentamycin, kanamycin and chloramphenicol every five days. We detected no such cross-contamination during the experiment.

### Phenotypic characterization of evolving replicates

We monitored the adaptation of the evolving populations to the novel nutrients every five days by estimating the fraction of cells that could metabolize the novel nutrient. To do so, we divided the cell density of the nutrient-adapted subpopulation by the total cell density (Fig. 1b). To estimate total cell density, we counted the number of cells after plating 100  $\mu\text{L}$

of  $10^6$ -fold and  $10^8$ -fold dilutions on solid DM agar plates supplemented with 0.2% glucose and  $50 \mu\text{g mL}^{-1}$  tryptophan. (We had reason to expect that all cells in the evolving population could grow and form colonies on these plates, because the ancestral strain could grow in glucose and tryptophan Winkler & Kao, 2012.) We estimated the cell density of the nutrient-adapted clones in the L-arabinose adaptation experiment by plating 100  $\mu\text{L}$  of  $10^2$ -fold and  $10^4$ -fold dilutions on DM agar plates supplemented with 0.2% L-arabinose and  $50 \mu\text{g mL}^{-1}$  tryptophan, and in the indole-adaptation experiment by plating 100  $\mu\text{L}$  of  $10^2$ -fold and  $10^4$ -fold dilutions on DM agar plates supplemented with 0.2% glucose and  $50 \mu\text{g mL}^{-1}$  indole. We used plating here, because growth on solid agar imposes stringent selection for adaptation. The reason is that diffusion of nutrients limits growth of immobilized cells. Cells thus have to metabolize a novel nutrient well to form visible colonies on a plate and would otherwise form microcolonies (Jeanson *et al.*, 2016).

We used glucose as the main carbon source, because *E. coli* grows faster in glucose than in glycerol, which reduces the incubation time and thus the risk of spontaneously emerging adaptive mutations during growth on the plate. We counted colonies on the total cell density plates after 48 h, and on the novel nutrient plates after five days of incubation. The reason is that bacteria grow more slowly on the novel nutrient plates, requiring extended incubation time before counting.

### Phenotypic characterization of evolved populations

We measured population growth curves over 24 h to find out whether some of our replicate populations grow significantly faster than others (Fig. S4). To this end, we established a 1 mL overnight culture of each evolved replicate population from  $\sim 10 \mu\text{L}$  of glycerol stock to generate a representative population sample. Subsequently, we established a 50-fold dilution of each replicate population in 200  $\mu\text{L}$  of a test medium containing either the ancestral or the novel nutrient condition (0.2% glycerol and  $50 \mu\text{g mL}^{-1}$  tryptophan, as well as 0.2% L-arabinose and  $50 \mu\text{g mL}^{-1}$  tryptophan for the L-arabinose-adapted replicates; 0.2% glycerol and  $50 \mu\text{g mL}^{-1}$  tryptophan, as well as 0.2% glycerol and  $50 \mu\text{g mL}^{-1}$  indole for the indole-adapted replicates) in a flat-bottom TPP 96-well microplate (Sigma-Aldrich Z797910). We also tested one further medium condition, 0.2% glucose and  $50 \mu\text{g mL}^{-1}$  indole, for the replicates evolved in the indole-adaptation experiment to assess if there are profound differences between growth on glucose and glycerol. We performed three technical replicates for each replicate population per nutrient condition, yielding [3 replicates  $\times$  12 populations  $\times$  3 indole adaptation-related conditions + (3 replicates  $\times$  12 populations  $\times$  2 arabinose adaptation-related conditions = )] 180 24-h growth curves. We

read the optical density at 600 nm on a Tecan Pro200 plate reader every 10 min for 24 h at 37 °C with shaking at an orbital amplitude of 2 mm. We estimated growth parameters using the R package `GROWTHCURVER` v0.2.1 (Sprouffske & Wagner, 2016). Because the length of the lag phase, the growth rate and the carrying capacity vary widely between populations, we used the total area under the growth curve as our primary fitness proxy, which includes contributions of all relevant growth parameters (Sprouffske & Wagner, 2016). We determined the fitness of each replicate population as the area under the curve and computed the mean fitness of each biological replicate (population) from three technical replicates. Area under the curve measurements have smaller coefficients of variation than other fitness proxies and are thus most reproducible for technical replicates in novel nutrient conditions (data not shown).

We assessed clonal growth only for replicate populations evolved in indole, because these replicates showed clear growth differences between  $R^+$  and  $R^-$  populations (Figs 2c and S2B). We did so in two consecutive sets of experiments. In the first, we measured 24-h growth curves in indole of 94 indole-adapted clones isolated from each selected replicate. We focused our efforts on the three  $R^+$  and the three  $R^-$  replicate populations ( $R_1^+$ ,  $R_2^+$ ,  $R_3^+$ ,  $R_1^-$ ,  $R_2^-$  and  $R_3^-$ ) evolved on indole that showed the highest cell density when plated on indole (Fig. S2B). We plated these replicates on 0.2% glucose DM agar plates supplemented with the original nutrient (50  $\mu\text{g mL}^{-1}$  tryptophan) or the novel nutrient (50  $\mu\text{g mL}^{-1}$  indole) and picked 94 colonies from each plate to seed 188 clonal 500  $\mu\text{L}$  overnight cultures per replicate (thus (94 clones + 2 ancestors)  $\times$  (1+1) groups of clones  $\times$  (3+3) replicate population = 1152 cultures). We prepared glycerol stocks from each culture of the evolved clones and then measured growth curves for each clone by diluting an overnight culture from the glycerol stock 100-fold into 200  $\mu\text{L}$  total of DM medium supplemented with 0.2% glucose and 50  $\mu\text{g mL}^{-1}$  indole. We included the ancestors as negative controls of the growth curve measurements with each group of 94 clones in a 96-well plate. Because of the more than 1000 required measurements, we determined the fitness (area under the growth curve) only once for each clone. We then compared fitness between the groups of 94 clones for different replicate populations.

In the second set of experiments, we performed further phenotypic analyses on clones isolated from the top two  $R^+$  and two  $R^-$  replicates with the highest average fitness on indole (Fig. S3). Specifically, we determined whether the evolved replicates continued to be unable to grow without tryptophan, or whether they showed reduced growth in tryptophan compared to the ancestor. To this end, we grew two replicates of each of the 94 evolved clones and 94 indole-adapted

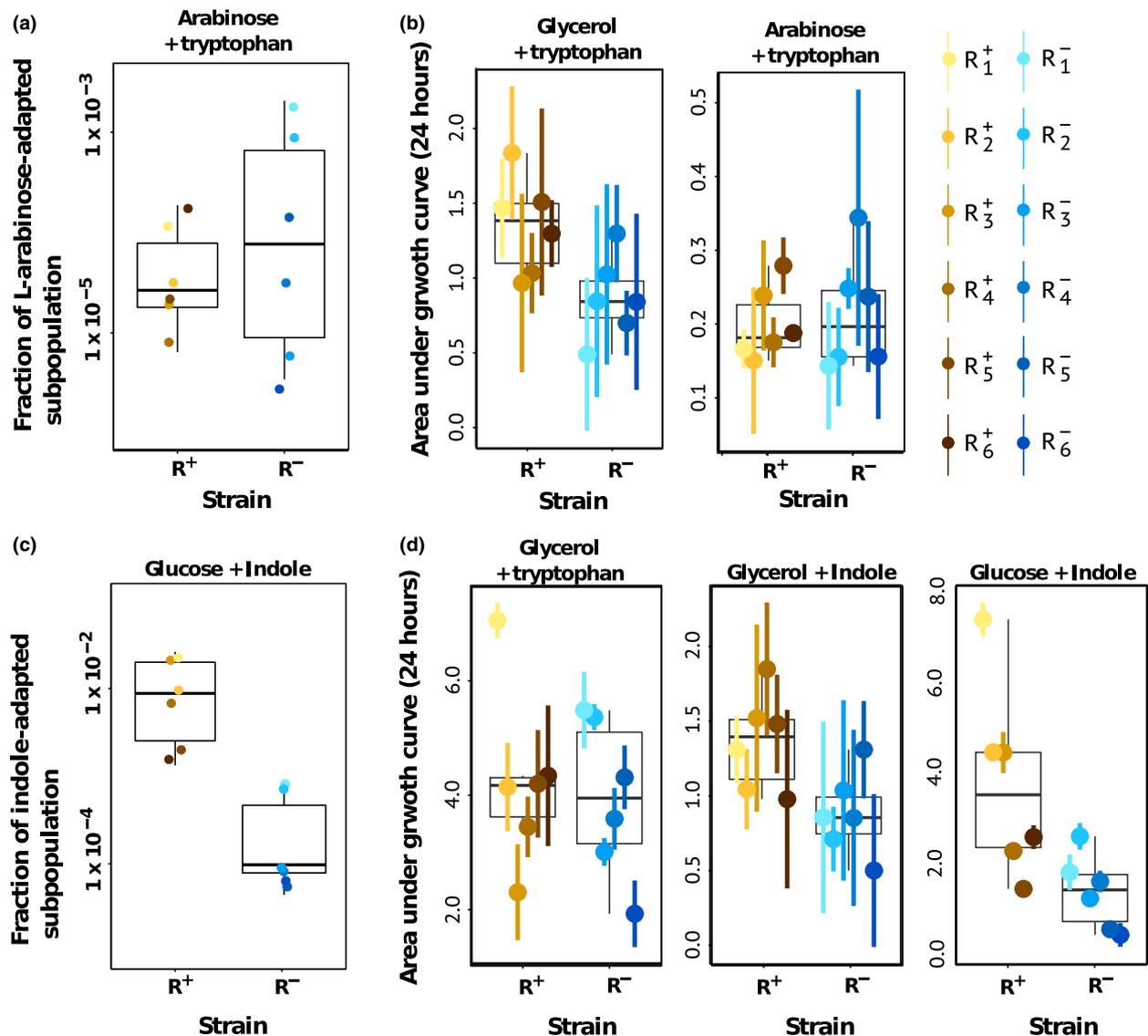
clones for 24 h in DM medium with 0.2% glucose supplemented with either tryptophan (50  $\mu\text{g mL}^{-1}$ ), indole (50  $\mu\text{g mL}^{-1}$ ) or neither nutrient. Specifically, we diluted the overnight cultures of the clones 100-fold into 200  $\mu\text{L}$  final volume of appropriately supplemented fresh medium, incubated them in a Stuart SI500 shaking incubator for 48 h at 37 °C (400 rpm) and measured the optical density at 600 nm after 24 h. The resulting large number of 2256 total growth assays made it infeasible to measure complete growth curves. We thus estimated the fitness for each clone as the final optical density after 24 h of growth. Whereas the final optical density measurement and the estimated carrying capacity from `Growthcurver` (Sprouffske & Wagner, 2016) differed slightly among the clones, their ranked fitness values were consistent between the two experiments.

### Library construction for whole-genome sequencing of clones

We sequenced the genomes of the ancestor  $R^+$  and  $R^-$  strains. We also sequenced between 46 and 48 randomly selected evolved clones from the total and indole-adapted subpopulation, and did so for the two  $R^+$  and two  $R^-$  replicate populations with the highest gain in fitness (Tables S4 and S5; Figs S3–S5). In total, we sequenced 384 clones. For each clone, we isolated genomic DNA from 1 mL of overnight culture using the QiaAmpDNA Mini kit (Qiagen 51304; QIAGEN AG, Hombrechtikon, Switzerland). We used the Nextera XT genomic DNA library prep kit (Illumina FC-121-1031; Illumina Cambridge Ltd., Saffron Walden Essex, UK), and the Nextera XT 96 indices kit (Illumina FC-121-1012) and prepared four sets of 96-clone libraries according to the Nextera DNA sample preparation guide (Illumina 15027987 Rev. B). We pooled each set of 96-clone libraries for  $2 \times 250$  bp paired-end read sequencing on the Illumina MiSeq platform hosted by the Functional Genomic Center Zurich. We sequenced 85% of clones (327/384) to at least fivefold genomewide sequence coverage, and 59% (227/384) to at least 10-fold genomewide sequence coverage (Fig. S5). We sequenced the  $R^+$  and  $R^-$  ancestors to 22- and 27-fold genomewide sequence coverage, respectively.

### Mutation identification and filtering

We used `Trimmomatic 0.27` (Bolger *et al.*, 2014) to remove Illumina-specific sequencing adaptors or bases of quality below 10 at the ends of reads. We eliminated reads shorter than 40 bp and aligned the remaining reads to the *E. coli* MG1655K12 reference genome NC\_000913.3 (Riley *et al.*, 2006) using `Bowtie2` (Langmead & Salzberg, 2012). We refined the resulting alignments using `Picard tools` (<http://picard.sourceforge.net>), `GATK RealignerTargetCreator` and `IndelRealigner`



**Fig. 2** Adaptation of  $R^+$  and  $R^-$  populations. The fraction of cells adapted to novel nutrients in  $R^+$  (yellow) and  $R^-$  (blue) replicate evolving populations at the end of the experiments ( $\sim 132$  generations), where (a) L-arabinose had replaced glycerol, or (c) indole had replaced tryptophan in the growth medium (see colour legend for population labels). (b) Fitness of each L-arabinose evolved population, measured via 24-h growth curves in liquid medium containing glycerol and tryptophan (pre-evolution condition), as well as L-arabinose and tryptophan (post-evolution condition). (d) Analogously, fitness of each indole-evolved population, measured via 24-h growth curves in liquid medium containing glycerol and tryptophan (pre-evolution condition), glycerol and indole (post-evolution condition), and glucose and indole (plating condition). The fitness values of each replicate population in (b) and (d) correspond to the mean area under the growth curve (circles) averaged over three (technical) replicate growth curves. Error bars indicate one standard deviation above and below the mean. The  $R^+$  and  $R^-$  ancestors reached a fitness of 0.47 and 0.51, respectively, in the pre-evolution condition and were not viable in post-evolution and plating conditions. The uncertainties of our fitness measures in different conditions, calculated as the coefficient of variation (for three technical replicate measurements), averaged over all replicate populations are 0.119 (glucose and indole), 0.193 (glycerol and indole), 0.34 (glycerol and tryptophan) and 0.375 (L-arabinose and tryptophan). In each box and whisker plot, the center bar corresponds to the median, the box extends to the first and third quartiles, and whiskers mark nonoutliers within the 95% confidence interval of the mean fitness among replicate populations of the same strain.

(McKenna *et al.*, 2010) and identified point mutations and indels smaller than 5 bp using GATK v3.14 HaplotypeCaller in haploid mode (Auwerwa *et al.*, 2013). We

identified genes or regions sharing an extremely high level of sequence similarity with a BLASTn (Altschul *et al.*, 1997) search against the *E. coli* reference genome.

We eliminated putative mutations located in these regions (Table S6) from further analysis, because they are likely to result from spurious Bowtie2 alignment (Langmead & Salzberg, 2012). For any one replicate population, we called all genotypes of clones from either the indole-adapted subpopulation, from the total population, or from the pooled set of clones (total and indole-adapted), using GATK's GenotypeGVCFs (McKenna *et al.*, 2010), and estimated each clone's genomewide and gene-specific coverage using GATK's DepthOfCoverage (McKenna *et al.*, 2010). For the genomic sequence analyses described in the sections below, we filtered the number of clones analysed based on specific criteria mentioned in these sections (Table S5).

We labelled putative genotypes with accuracy lower than 99.999999% (genotype likelihood < 80) as 'missing' and then eliminated clones having 'missing' genotypes at more than 35% of the sites in the genome. This filtering step left us with 336 clones for mutation discovery (Table S5). In addition, we eliminated sites with 'missing' genotypes in more than 70% of clones, because GATK may have encountered systematic problems in accurate genotype calling at such sites. Furthermore, we removed sites with 'missing' ancestral genotypes. After these filtering steps, we compared at each site the genotype of each evolved clone to that of its ancestor. We called the site 'ancestral' if it had not changed during laboratory evolution, 'derived' if it had changed, or 'missing' if its genotype was missing in the evolved clone. We inferred the likely phenotypic effects of mutations using SnpEff (Cingolani *et al.*, 2012).

### Mapping the F plasmid

We checked for the continued integration of the F plasmid in the genomes of  $R^+$  and  $R^-$  clones after evolution, by comparing the fraction of sequence reads alignable to the F plasmid sequence between evolved clones and the ancestral genome. In this analysis, we restricted ourselves to clones with at least 10-fold genomewide sequence coverage (Table S5), because for these clones we could confidently distinguish segmental chromosomal deletion from insufficient sequencing coverage. We aligned the reads of each clone to the F plasmid reference sequence (Plasmid pECC-1470\_100 NCBI CP010325.1) using Bowtie2 (Langmead & Salzberg, 2012). We verified integration by inspecting the read alignments to the *E. coli* K12 reference sequence spanning the integration site at genomic coordinate 1 317 300–1 317 320 bp, within the *trpB* coding sequence (Fig. S6A). When the F plasmid is integrated into the genome, reads that span the integration site are trimmed by Bowtie2 (Langmead & Salzberg, 2012) when aligned to the *E. coli* reference genome. We confirmed proper F plasmid integration

by identifying these partially aligned reads. In addition, we compared the relative fraction of sequence reads that could align to the F plasmid reference sequence between each clone and its ancestor (Fig. S6B). We considered that a clone still had the F plasmid inserted if partially aligned reads occurred in the genome sequence, and if the above ratio deviated from the expected value of one by no more than 0.2.

### Computing population diversity

We computed the derived allele frequency (DAF), Shannon Index ( $H$ ), Shannon Equitability Index ( $E_H$ ) and various measures of linkage disequilibrium ( $D$ ,  $|D'|$ ,  $r^2$  and  $r'^2$ ) (see Text S2) in two  $R^+$  and two  $R^-$  replicate populations at all polymorphic nucleotide sites, as described below.

We calculated derived allele frequency as the frequency of a nonancestral allele found in clones within the indole-adapted subpopulation or the total population:

$$\text{DAF} = D/N, \quad (1)$$

where  $D$  is the number of derived genotypes and  $N$  is the number of observed genotypes. For each replicate population, we computed the DAF for the set of evolved clones we had sequenced from (i) the total population and (ii) the indole-adapted subpopulation.

In addition, for each replicate population, we computed the Shannon Index ( $H$ ) from the genomic haplotypes of both the indole-adapted subpopulation and the total population as follows:

$$H = - \sum_{i=1}^S p_i \log_2 p_i, \quad (2)$$

where  $p_i$  is the frequency of the  $i$ th haplotype and  $S$  is the number of all haplotypes observed in each replicate population (Shannon, 1948). We identified haplotypes by concatenating each clone's nucleotide sequence for sites in which 95% of clones had been accurately genotyped. We only considered clones with nonmissing genotypes across all of these sites in this procedure (Table S5). Because of our strict data quality filters and our definition of haplotype, the number of sites and clones entering the calculation of  $H$  may differ across replicate populations. We calculated each haplotype's frequency by dividing the number of clones with this haplotype by the total number of haplotypes within a given group of clones. Furthermore, we computed the Shannon Equitability Index ( $E_H$ ) (Magurran, 1988) for each replicate population as

$$E_H = \frac{H}{\ln S}, \quad (3)$$

where  $S$  is the total number of haplotypes determined by sequencing. We included the Shannon Equitability Index ( $E_H$ ) (Magurran, 1988) to correct

for the tendency of the Shannon Diversity Index to become inflated with an increasing number of haplotypes. The Shannon Equitability Index is bound between 0 and 1.

### Detecting structural variation

Bacterial recombination can cause genome rearrangements (Hastings *et al.*, 2000; Darmon & Leach, 2014). In analysing such rearrangements, we focused on DNA amplifications. More specifically, we estimated the copy number of each gene within each clone isolated and sequenced from our evolving populations. We did not screen for segmental deletion, because most of our sequenced clones did not have sufficient genomewide sequence coverage ( $< 30\times$ ) to confidently detect such deletions (Deatherage *et al.*, 2015). We focused on clones with higher than  $10\times$  genomewide coverage for this analysis: 56 clones for the  $R_1^+$  population, 70 clones for  $R_2^+$ , 61 clones for  $R_1^-$  and 58 clones for  $R_2^-$  (Table S5). We then used GATK v3.14 DepthofCoverage (McKenna *et al.*, 2010) and event-wise testing (Yoon *et al.*, 2009) together with breseq v0.27.0 (Deatherage & Barrick, 2014) to predict the likely copy number for each gene. Specifically, we used GATK DepthofCoverage for gene copy number estimation, breseq v0.27.0 to identify the segmental amplification and their breakpoints, and event-wise testing for both, as described below.

#### *Estimating gene copy numbers with GATK DepthofCoverage*

We estimated the sequence coverage of all genes in each clone using GATK DepthofCoverage v3.14 (McKenna *et al.*, 2010), correcting the gene-specific sequence coverage for G + C content, because Illumina sequencing technology exhibits coverage bias at regions of extreme G + C content (Yoon *et al.*, 2009). To this end, we first binned genes into ten groups (with G + C content ranging from 0 to 1 at intervals of 0.1) and calculated the median coverage across all genes within the same G + C content bin. We then computed a gene's putative copy number in a given clone by dividing the gene-specific sequence coverage to the median coverage of a given G + C content bin, and rounded the result to the nearest integer.

#### *Estimating gene copy numbers with event-wise testing (EWT)*

Independently from the above analysis, we also used the event-wise testing (EWT) method (Yoon *et al.*, 2009) to detect segmental duplications. To do so, we first computed sequence coverage in 100 bp nonoverlapping sliding windows across the genome and corrected for G + C content bias (Yoon *et al.*, 2009). Specifically, we corrected the sequence coverage of each window by multiplying it by the following gene G

+ C content correction factor ( $k_i$ ):

$$k_i = \frac{\text{global coverage median}}{\text{median coverage of bin } i}, \quad (4)$$

where  $i$  indexes one of the ten G + C content bins (with G + C content ranging from 0 to 1 at intervals of 0.1). We then screened all clones for regions with at least 10 kb of consistently increased sequence coverage by implementing the event-wise testing (EWT) method (Yoon *et al.*, 2009) using custom Python v2.7 scripts. We validated the EWT-based predicted segmental amplification by obtaining the structural variation breakpoints from breseq v0.27.0 (Deatherage *et al.*, 2015) with default settings. We estimated the copy number of genes spanning the regions of structural variation by rounding the quotient of the median gene-specific sequence coverage and the genomewide coverage to the nearest integer. We only report a gene as having undergone an amplification if it occurs in a region of putative structural variation according to both event-wise testing and breseq v0.27.0, and if the difference of its copy number estimates from GATK DepthofCoverage and from the EWT method are no greater than one. On simulated data, our procedure reached 95.5% accuracy in detecting amplified genes and reached a minimum of 84% accuracy in inferring gene copy number (Text S3 and Table S8).

## Results

### Evolutionary benefits of recombination depend on the nutrient environment

The  $R^+$  and  $R^-$  strains we used in our evolution experiments differed approximately 25-fold in their recombination rate. During experimental evolution, we monitored how well cells in each replicate population of these strains adapted to the novel nutrient by measuring the fraction of nutrient-adapted cells, that is, the fraction of cells able to grow on solid agar supplemented with only the novel nutrient (Fig. 1b). We used plating to monitor adaptation, because growth on solid agar imposes more stringent selection than liquid culture. The reason is that diffusion of nutrients limits growth of immobilized cells, which have to metabolize a novel nutrient especially well to form visible colonies on a plate (Jeanson *et al.*, 2016). In other words, 'adapted' cells that can form visible colonies are likely to harbour mutations that greatly improve growth, more so than cells that can survive on the nutrient in liquid culture. The fraction of such adapted cells increased in all replicate populations, and on both L-arabinose and indole (Figs 2 and Fig. S2). In the arabinose adaptation experiment, this fraction was less variable in  $R^+$  populations than in  $R^-$  populations (Fig. 2a), but overall, the  $R^+$  replicates did not give rise to a significantly larger fraction of arabinose-adapted cells

(Table S9 test 1,  $W_{12} = 15.5$ ,  $P = 0.73$ ). In contrast,  $R^+$  populations harboured a significantly higher fraction of indole-adapted cells (Fig. 2c, Table S9 test 2,  $W_{12} = 36$ ,  $P = 0.001$ ). In other words, based on this plating assay,  $R^+$  populations adapted better to indole than  $R^-$  populations.

We also measured mean population fitness in liquid medium at the end of the experiment.  $R^+$  populations that had evolved on L-arabinose outperformed  $R^-$  population in the ancestral (pre-evolution) medium (Fig. 2d, Table S9 test 6,  $W_{12} = 32$ ,  $P = 0.013$ ) but not in L-arabinose (Fig. 2d, Table S9 test 7,  $W_{12} = 19$ ,  $P = 0.47$ ). In contrast, evolved  $R^+$  replicates gained higher fitness in the novel nutrient indole than  $R^-$  replicates (Fig. 2d, Table S9 test 3,  $W_{11} = 13$ ,  $P = 0.48$ ; 4,  $W_{12} = 31$ ,  $P = 0.021$ ; and 5,  $W_{12} = 32$ ,  $P = 0.013$ ). These observations suggest that high recombination is beneficial for adaptation to indole but not to L-arabinose.

### Selecting a subset of replicate populations for genomic analysis

We focused all our subsequent phenotypic and genomic analyses on the indole-adaptation experiments, because  $R^+$  populations displayed a higher fraction of indole-adapted cells (Fig. 2c), as well as a growth advantage in liquid culture relative to  $R^-$  populations (Fig. 2d). In addition, we restricted ourselves to the three  $R^+$  and three  $R^-$  replicates with the highest density of indole-adapted cells at the end of the experiment ( $R_1^+$ ,  $R_2^+$ ,  $R_3^+$ ,  $R_1^-$ ,  $R_2^-$  and  $R_3^-$ ) to assay growth of isolated clones (Figs S3 and S4). We isolated 94 clones that grew on a novel nutrient (indole) plate from each of the six remaining replicate populations. We refer to such a collection of 94 clones as an *indole-adapted subpopulation* sample and reasoned that it would be enriched for mutations beneficial for growth on indole. In addition, we isolated from each population 94 clones derived from cells plated on DM medium agar plates supplemented with tryptophan, which we refer as a *total population* sample. (Because the fraction of indole-adapted cells did not exceed 3% of the total population (Table S4), even in the best-adapted population, the vast majority of these cells would not come from the indole-adapted subpopulation.) We next measured the growth curves of all these clones ((94 clones + 2 ancestors)  $\times$  (1 indole-adapted subpopulation + 1 total population)  $\times$  (1+1) strains  $\times$  3 replicate populations  $\times$  1 technical replicate = 1152 assays) in indole-and glucose-supplemented liquid culture (see Methods). Specifically, we measured the fitness of each clone in the novel nutrient (indole) as the area under the clone's growth curve, measured during 24 h of growth. These measurements showed that clones of both the total and indole-adapted subpopulation from the  $R^+$  replicates generally grew better than those from the  $R^-$  replicates (Table S9 test

8,  $W_{1128} = 192931$ ,  $P = 2.94 \times 10^{-10}$ , Table S4 and Fig. S3).

We focused our subsequent genome sequencing and all further analyses on the clones from  $R_1^-$ ,  $R_2^-$ ,  $R_1^+$  and  $R_2^+$  replicate populations (Fig. S4), because their clones grew noticeably better than those from replicates  $R_3^+$  and  $R_3^-$  (Fig. S3). Specifically, we aimed to identify beneficial mutations for growth in indole and to find out whether recombination had facilitated their emergence and spread. In three of the four selected replicates ( $R_2^+$ ,  $R_1^-$  and  $R_2^-$ ), indole-adapted clones grew better in indole than those in the total population (Fig. S3, Table S8 test 10,  $W_{188} = 497$ ,  $P < 2.2 \times 10^{-16}$ ; 11,  $W_{188} = 3407$ ,  $P = 0.0034$ ; 12,  $W_{188} = 1137$ ,  $P < 2.2 \times 10^{-16}$ ), as one might expect. The only exception was population  $R_1^+$ , where clones of the indole-adapted subpopulation (Methods) were less fit than clones from the total population (Table S8 test 9,  $W_{188} = 5333$ ,  $P = 0.0071$ ). The  $R_1^+$  replicate also proved to be exceptional in another way, as we found out when we asked whether the evolved clones had regained the ability to synthesize tryptophan in some other indole-independent manner. We did this by growing the clones in glucose minimal medium supplemented with tryptophan, indole or neither of the two nutrients (Table S4, Fig. S7). Whereas clones from replicates  $R_2^+$ ,  $R_1^-$  and  $R_2^-$  required either indole or tryptophan for growth, most clones from the  $R_1^+$  replicate grew well, even without tryptophan or indole. Thus, the  $R_1^+$  replicates had restored tryptophan prototrophy.

### $R^+$ and $R^-$ populations acquired both unique and at least one parallel point mutations

We next studied the genomic basis of the growth differences between the two  $R^+$  and the two  $R^-$  replicates, and of the growth differences between the total population and the indole-adapted subpopulations. To this end, we sequenced the genomes of at least 46 clones from both the adapted subpopulation and the total population for the  $R_1^+$ ,  $R_2^+$ ,  $R_1^-$  and  $R_2^-$  replicates from the endpoint (day 40) of the evolution experiment (Table S5). We found that the  $R^+$  and  $R^-$  replicates had accrued a total of 85 base substitutions and 25 indels (Table 1) across all sequenced clones. Among these mutations, only two had risen to a frequency of 100%, one of them in an indole-adapted subpopulation and the other in a total population sample (Table 1). Most of these mutations occurred at a low derived allele frequency (DAF  $< 0.05$  for 74.00% of  $R^+$  and 67.57% of  $R^-$  populations) and were located in gene coding regions. The majority of these mutations were nonsynonymous (73.19% in  $R^+$  and 53.85% in  $R^-$  replicates). We asked whether coding regions contained a larger fraction of nonsynonymous mutations than the fraction of 0.67 (Wielgoss *et al.*, 2011) expected by chance alone, which would indicate that selection has

preferentially retained some nonsynonymous mutations (Tenaillon *et al.*, 2012a). However, this was not the case (Table S9 test 13,  $B(n = 46, k = 34, P = 0.67), P = 0.28$ ; 14,  $B(n = 266, k = 14, P = 0.67), P = 0.21$ ).

Most genetic changes present were unique to the indole-adapted subpopulations or the total populations (Fig. 3a,b). Only a minority of mutations occurred in both groups of clones ( $R_1^+$ : 21.42%,  $R_2^+$ : 11.41%,  $R_1^-$ : 10.53% and  $R_2^-$ : 33.33% of the total number of mutations).

We also identified five genes in which mutant alleles had increased to a derived allele frequency exceeding 10% at the end of the experiment (Table 2). Because such an increase is not expected by genetic drift alone, given the large populations in our experiment, the observed changes in these genes are potentially beneficial, either for adaptation to indole or to the experimental environment. The most intriguing among the affected genes is *tnaC*, which acquired mutations in three of four replicate populations. The same single nucleotide insertion occurred in both  $R^-$  replicates, whereas a deletion occurred in two neighbouring amino acids in replicate  $R_2^+$  (Table 2 and Fig. 3c). Derived *tnaC* alleles were present in at least 60% of the sequenced clones of replicate  $R_2^+$  and  $R_1^-$ . They are especially highly enriched in the total population samples of these replicates (Fig. 3c). The *tnaC* mutation was fixed in the  $R_2^-$  total population sample and had a frequency of 98% in its indole-adapted subpopulation. The *tnaC* gene encodes a regulator for the tryptophan metabolism genes *tnaA* and *tnaB*. The three genes form the *tna* operon.

The only replicate that did not show a mutation in the *tnaC* gene was replicate  $R_1^+$ , which is consistent with this replicate's unique mechanism for restoring tryptophan biosynthesis. Indeed, when we analysed the genomic changes in this replicate, we found that the genomic F plasmid integrate had been lost in the  $R_1^+$  replicate (Fig. S6B). Specifically, 96% of the  $R_1^+$  clones contained no genomic sequences mapping to the F plasmid, and the *trp* operon did not harbour any

evidence for foreign sequence insertions or other mutations. In other words, excision of the F plasmid had restored the *trp* operon. This excision may have been facilitated by the inverted repeats of a *Tn10* transposable element, into which the F plasmid had been integrated in the first place (Typas *et al.*, 2008).

### $R^+$ populations exhibit higher genetic diversity

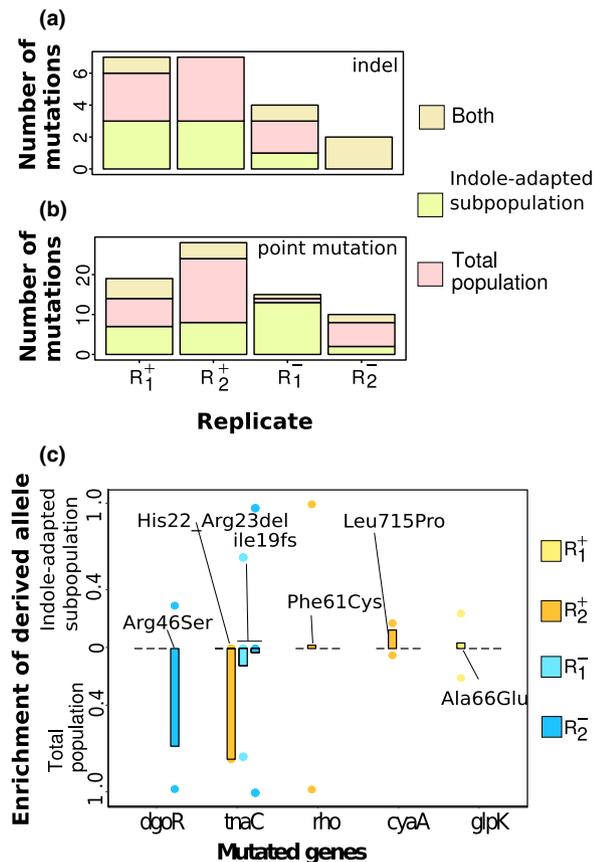
Recombination may facilitate adaptation by increasing the genetic diversity within a population, because it can 'shuffle' alleles among individuals (Vos, 2009). To quantify a population's diversity, we computed the Shannon Index and its normalized version, the Shannon Equitability Index, which is constrained between 0 and 1, from the frequencies of alleles observed in our populations. The Shannon Index increases as more alleles are observed, and both indices are maximized when the frequencies of the haplotypes within a population are similar. Using both metrics, we found that the  $R^+$  replicates exhibited a significantly higher population diversity than the  $R^-$  replicates (Table S6, Table S9 test 15,  $W_8 = 15, P = 0.023$ ), which suggests a positive influence on diversity caused by recombination. A greater number of recombination events in the  $R^+$  populations might also be reflected in reduced amounts of linkage disequilibrium. However, an extensive analysis of various measures of linkage disequilibrium, such as  $r^2$  (VanLiere & Rosenberg, 2008), showed that our sequence data do not have sufficient resolving power to detect such a difference (Text S2, Figs S8 and S9). The reasons lie in the small number of point mutations (Table 1) and in the low frequency of the mutant alleles we observed.

### Both $R^+$ and $R^-$ populations experienced segmental duplications

We next studied structural genomic variation, an indirect measure of recombination's effects, reasoning that frequent recombination might bring about increased

**Table 1** Summary of point mutations and indels discovered in the  $R^+$  and  $R^-$  indole-evolved populations.

Strain	Total/ Adapted	Replicate	Number of mutations	Number of	Number of	Number of	Number of	Number of	Number of
				point mutations (percentage of all mutations)	indels (percentage of all mutations)	mutations within coding regions	synonymous mutations	nonsynonymous mutations	indels within coding regions
$R^+$	Adapted	1	17	13 (75%)	4 (25%)	15	0	12	3
		2	15	12 (80%)	3 (20%)	12	2	8	2
	Total	1	17	12 (70.5%)	5 (28.5%)	13	0	12	1
		2	24	20 (83.3%)	4 (16.7%)	20	2	13	5
$R^-$	Adapted	1	16	14 (87.5%)	2 (12.5%)	14	6	7	1
		2	6	4 (66.7%)	2 (33.3%)	6	2	2	2
	Total	1	5	2 (40%)	3 (60%)	4	0	1	3
		2	10	8 (80%)	2 (20%)	9	1	5	3



**Fig. 3** Summary of mutations revealed by genome sequencing. The number of (a) indels and (b) base substitutions found uniquely in clones from the total population (light pink), in clones from the adapted subpopulation (light green), and in clones from both subpopulations (brown) of each sequenced replicate population. (c) Differential enrichment of derived alleles in either the indole-adapted subpopulation (positive *y*-axis) and the total population (negative *y*-axis), for alleles located in five genes that may have been subject to adaptive evolution, because each allele harbours a nonsynonymous change with a derived allele frequency (DAF) exceeding 10% (Methods) in either the indole-adapted subpopulation or the total population (see also Table 2). Each circle depicts the DAF of a putatively adaptive mutation in either an indole-adapted subpopulation (positive *y*-axis) or a total population (negative *y*-axis). The differential enrichment ( $DAF_{\text{Indole-adapted subpopulation}} - DAF_{\text{Total population}}$ ) of these mutations is indicated by bar height. The mutations in questions are either amino acid changing mutations (denoted as XNY for a change from amino acid X to Y at position N of an encoded protein), frameshift mutations (denoted as XNfs for a mutation at position N, where the frameshift occurred at encoded amino acid X), or deletion mutations (denoted as XN\_YMdel for a deletion spanning amino acid X at position N and extending to amino acid Y at position M).

structural variation (Hastings *et al.*, 2009b; Lobkovsky *et al.*, 2016), some of which may be adaptive. In analysing such variation, we focused on segmental

duplication and copy number variation in individual genes, which we identified using a computational method that proved to be reliable in a benchmark analysis involving simulated structural variation and Illumina sequencing data (see Methods). We used this method to identify genes with copy number changes relative to the ancestor in those 227 clones that had at least 10-fold genomewide sequence coverage (Table S5) (see Methods).

Among the tested clones, we found that 12.5% of the  $R_1^+$  clones, 61% of the  $R_2^+$  clones, 82% of the  $R_1^-$  clones and 7% of the  $R_2^-$  clones carried segmental amplifications in their genomes (Table S7, Fig. 4). Overall, the incidence of clones harbouring amplifications was similar between  $R^+$  and  $R^-$  replicates (Table S9 test 16,  $\chi_1^2 = 2.82$ ,  $P = 0.090$ ). The indole-adapted subpopulations harboured significantly higher frequencies of clones with amplifications than the total population in replicates  $R_2^+$  and  $R_1^-$  but not  $R_1^+$  or  $R_2^-$  (Fig. 4, Table S9 test 17,  $B(n = 33, k = 33, P = 0.22)$ ,  $P < 2.2 \times 10^{-16}$ ; 18,  $B(n = 33, k = 33, P = 0.60)$ ,  $P = 8 \times 10^{-8}$ ; 19,  $B(n = 22, k = 3, P = 0.12)$ ,  $P = 0.74$ ; 20,  $B(n = 28, k = 1, P = 0.03)$ ,  $P = 0.66$ ).

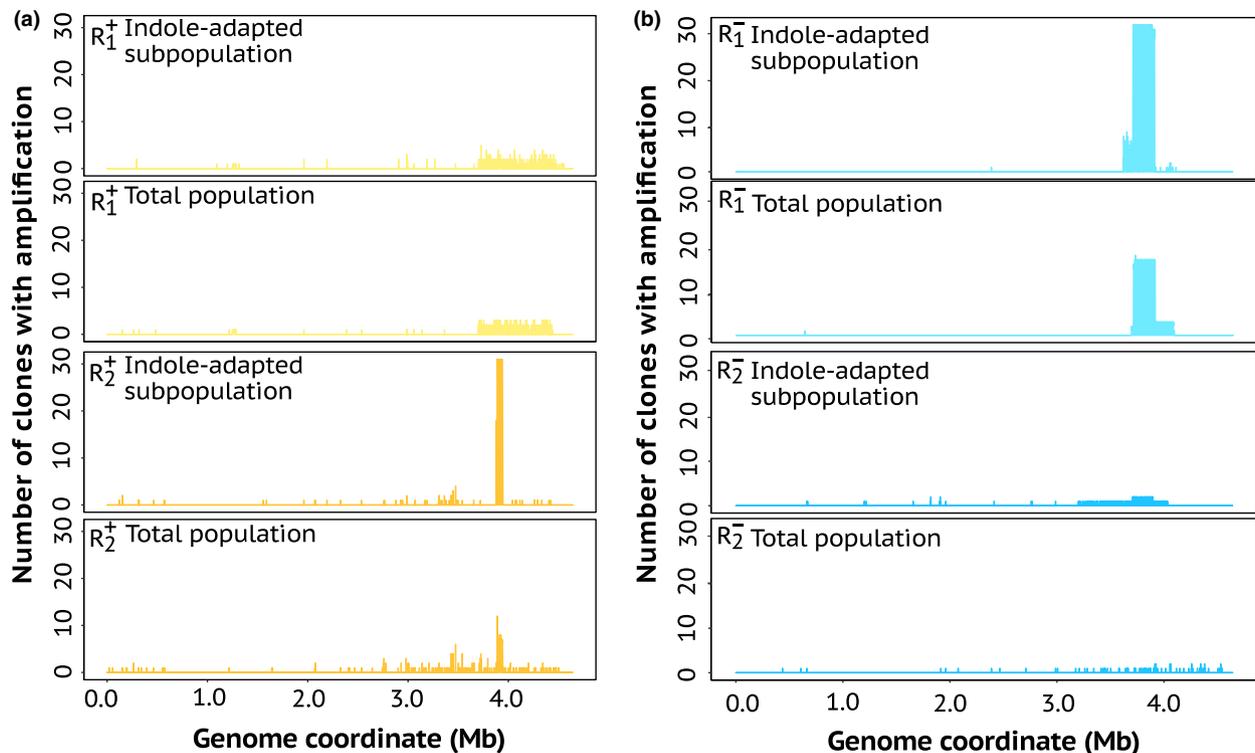
Next, we analysed gene copy numbers in the amplified regions in more detail, focusing on replicate  $R_2^+$  and  $R_1^-$ , because their indole-adapted population samples harboured more than 30 clones with amplifications. We found that in the  $R_2^+$  population, amplified genes had significantly higher copy numbers in the indole-adapted population than in the total population (Fig. 5, Table S9 test 21,  $W_{3175} = 1835372$ ,  $P < 2.2 \times 10^{-16}$ ). These large copy number increases may be mediated by multiple recombination events. We did not observe the same pattern for the  $R_1^-$  population (Table S9 test 22,  $W_{9611} = 10\,680\,759$ ,  $P = 0.15$ ).

Most amplified genes clustered in a small genomic region spanning 200 kb (Fig. 4). Because this region was overrepresented in terms of both amplified genes and increased gene copy number in the  $R_2^+$  and  $R_1^-$  populations, we reasoned that some of its gene amplifications may be adaptive. We thus focused on this region when further characterizing the amplifications.

The segmental amplifications in this region may have originated more than once, because they extend through different coordinates in different clones. For example, in  $R_1^-$ , one dominant amplification (93.47% of clones) spans genomic coordinates 3 710 249 to 3 914 582 bp, whereas the remainder of the amplifications have two other start and end points (3 617 769–3 890 228 bp or 3 691 190–4 083 916 bp). Furthermore, genes within the segmentally amplified regions vary in their copy numbers for a subset of clones. Despite these heterogeneities, there are commonalities between the amplified regions. Specifically, a subregion of ca. 32 kb (genomic coordinates 3.88 and 3.92 Mb) is amplified in the indole-adapted subpopulations of both the  $R_2^+$  and  $R_1^-$  populations (Fig. 4). We suspected that it may be

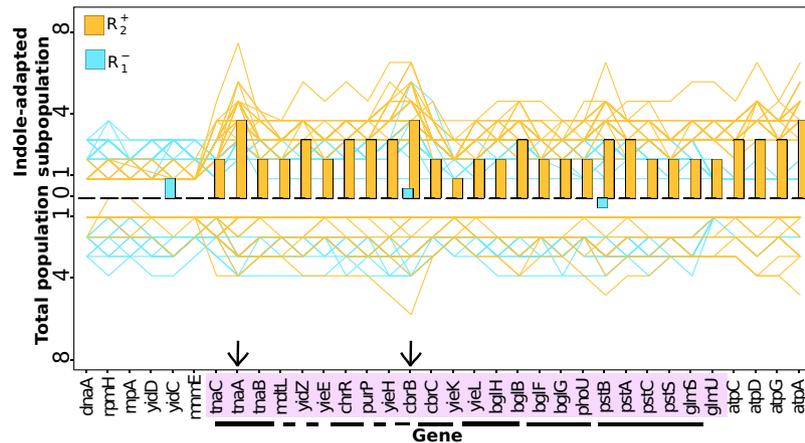
**Table 2** Summary of genes with nonsynonymous mutations that increased in allele frequency to at least 10% in either the total population or the indole-adapted subpopulation.

Mutated gene	Gene function	Mutation	Genomic location (bp)	Replicate	Derived allele frequency in indole-adapted subpopulation	Derived allele frequency in total population
<i>cyaA</i>	Adenylate cyclase	Leu715Pro	3993296	$R_2^+$	0.175	0.045
<i>dgoR</i>	Putative DNA-binding transcriptional regulator	Arg46Ser	3875025	$R_2^-$	0.30	1.00
<i>glpK</i>	Glycerol kinase	Ala66Glu	4117026	$R_1^+$	0.24	0.20
<i>rho</i>	Modulator of Rho-dependent transcription termination	Phe62Cys	3966601	$R_2^+$	1.00	0.98
<i>tnaC</i>	Tryptophanase leader peptide	Ile19fs	3888485	$R_1^-$	0.63	0.75
		Ile19fs	3888485	$R_2^-$	0.98	1.00
		His22_Arg23del	3888496	$R_2^+$	0.00	0.77

**Fig. 4** Genes with increased copy number due to segmental amplifications. The number of clones ( $y$ -axes, see also Table S5) with a detected amplified gene (arranged according to genomic coordinates in mega base pairs (Mb) along the  $x$ -axes) of the indole-adapted subpopulation and the total population of (a)  $R_1^+$  and  $R_2^+$ , and (b)  $R_1^-$  and  $R_2^-$  replicate populations. The  $R_2^+$  and  $R_1^-$  replicates show a cluster of amplified genes between 3.0 and 4.0 Mb.

associated with the fitness advantage experienced by indole-adapted clones and thus examined in more detail the 23 genes in this region, which fall into nine operons (Riley *et al.*, 2006) (Fig. 5). These genes showed significantly increased copy numbers in clones of the  $R_1^-$  (Fig. 5, Table S9 test 23,  $W_{1128} = 120\,169$ ,  $P = 1.63 \times 10^{-5}$ ) and the  $R_2^+$  indole-adapted

subpopulations that harboured segmental amplifications, compared to the total populations (Fig. 5, Table S9 test 24,  $W_{889} = 102\,035.5$ ,  $P < 2.2 \times 10^{-16}$ ). In addition, in both replicates, eight of nine operons composed of these 23 genes (Table S10) were enriched in amplified genes in the indole-adapted subpopulation relative to the total population.



**Fig. 5** Gene copy number amplification in the indole-adapted subpopulation. Genes indicated on the horizontal axis include those occurring in a region amplified in both  $R_2^+$  and  $R_1^-$  populations (genomic coordinate 3.88–3.92 Mb, highlighted in purple), as well as several genes flanking this region. The vertical axis shows the estimated copy number of each gene, for each clone tested for amplification (see also Table S5). Three genes (*rpmH*, *yidD* and *tnaC*) within this region are too short (< 200 bp) (Methods) to have their copy number accurately inferred. Each coloured line (blue for  $R_1^-$ , dark yellow for  $R_2^+$ ) indicates the estimated copy number of a specific gene in a sequenced clone from either the total population (below the origin) or the indole-adapted subpopulation (above the origin). Note that the copy numbers of genes within the amplified region of each clone varied, suggesting complex amplification events. Differences in the median copy number of each gene of the indole-adapted subpopulation from that of the total population are indicated by the bars (blue for  $R_1^-$ , dark yellow for  $R_2^+$ ). For example, a dark yellow bar above the origin indicates that the indole-adapted clones from the  $R_2^+$  population have a higher median gene copy number than the total population. Only three genes show a difference in copy number in the  $R_1^-$  population (*yidC* and *cbrB* have higher median copy number in the indole-adapted subpopulation, whereas *pstB* has higher median copy number in the total population). In contrast, in the  $R_2^+$  population almost all genes in the amplified region show a higher copy number in the indole-adapted population. The highest copy numbers are reached by *tnaA* (eight copies) and *cbrB* (eight copies) in the  $R_2^+$  indole-adapted subpopulation (marked by arrows).

We further compared the copy numbers of these 23 genes in clones harbouring segmental amplification between the  $R_2^+$  and  $R_1^-$  replicate populations. We observed that  $R_2^+$  clones with amplifications reached higher copy numbers of these genes compared to the  $R_1^-$  clones with amplifications (Fig. 5, Table S9 test 25,  $W_{1993} = 751\,966$ ,  $P < 2.2 \times 10^{-16}$ ). This suggested that the incidence of gene amplification may be increased in highly recombining strains. Finally, genes that reached the highest copy number do so in the indole-adapted subpopulation of the  $R_2^+$  clones (Fig. 5). These genes include *tnaA* and *cbrB*. Each of them has a maximum number of eight copies, suggesting that both genes are strongly selected for increased gene dosage (Fig. 5). *tnaA* encodes a tryptophanase that converts tryptophan to indole, but that can catalyse this reaction in both forward and reverse directions equally efficiently (Watanabe & Snell, 1972). CbrBC promotes bacterial growth in minimal medium and causes the colicin E2-tolerant phenotype (Avison *et al.*, 2001; Cariss *et al.*, 2010).

### Better adapted genotype combinations arise in one $R^+$ population

To test for signs of the Hill–Robertson effect (Hill & Robertson, 1966), we next asked whether some allelic

combinations, possibly assembled via recombination, had led to increased growth on indole. In doing so, we focused on base substitutions that are likely beneficial because of their high frequency (Table 2), and on gene amplifications. We found only one likely beneficial allelic combination in our data, which involves a *tna* operon amplification and mutations in *rho*, which occurred in the  $R_2^+$  indole-adapted subpopulation. Specifically, the  $R_2^+$  clones we had sequenced varied mainly in three ways, that is, in the presence or absence of (i) an indel in *tnaC*, (ii) a nonsynonymous mutation in *rho* and (iii) an amplification of the *tna* operon. The two most common combinations of these variants in the total  $R_2^+$  population (i) were the *rho* and *tnaC* double mutations (70.27%) and (ii) the *rho* mutation and amplification of the *tna* operon (16.22%) (Table S11). The fitness of the latter in indole-containing selective medium was higher than that of the former (Table S9 test 26,  $W_{33} = 173$ ,  $P = 2.2 \times 10^{-5}$ ). In addition, clones isolated from the indole-selective plates only harboured the second, higher-fitness variant, also suggesting that it may be involved in the adaptation to indole. The individual mutations in the derived alleles may have originated independently from one another, because one of the clones in the  $R_2^+$  population harboured only the mutant *tnaC* allele, and four other

clones harboured only the *rho* allele. Thus, one or more recombination events may have brought these alleles together. Alternatively, the double mutant may have arisen on the same chromosome, and recombination later broke it apart. In the two  $R^-$  replicates, we also found differences in the incidence of allelic combinations between the total population and the indole-adapted subpopulation, but the allelic combinations in question do not convey higher fitness on indole (Text S4).

## Discussion

Our principal goal was to understand if and how conjugation-mediated bacterial recombination influences evolutionary adaptation in a challenging novel environment. To do so, we evolved  $R^+$  and  $R^-$  replicate populations for 40 days ( $\sim 132$  generations) on two novel nutrients. We then examined the resulting genomic and phenotypic changes in hundreds of bacterial clones isolated from the evolved populations. We compared these changes between the  $R^+$  and  $R^-$  populations, and between a sample of each total population and a nutrient-adapted subpopulation. Below, we relate our observations to prominent ideas about recombination's effects on adaptive evolution and to past laboratory evolution experiments.

Previous experimental evolution studies showed that recombination can accelerate adaptation in stressful environments (Goddard *et al.*, 2005; Gray & Goddard, 2012; Winkler & Kao, 2012). Some of our observations agree with these studies, because our highly recombining populations adapted faster to the novel nutrient indole. However, recombination did not provide a similarly pronounced advantage during adaptation to L-arabinose. This observation illustrates that the advantage of recombination depends not only on genetic factors, but also on the environment in which evolution takes place. One limitation of our experiment was that we could at best predict the likely phenotypic changes (Text S1, Fig. S10) but not the precise genetic changes required for adaptation to these novel nutrients. Thus, we cannot explain rigorously why recombination is more effective driving indole adaptation than driving L-arabinose adaptation. A possible reason may be that evolution had to overcome deletions of three genes in the L-arabinose pathway, which might have involved at least three alternative sugar catabolic operons (*xyl*, *fuc*, *yiaKLMNO-lyx-sgbHUE* and *rhaBAD*) (Text S1, Fig. S10A). If multiple, individually weakly adaptive mutations need to be joined by recombination to provide a strongly adaptive genotype, these mutations may simply have occurred at too low a frequency during the short time ( $\sim 132$  generations) of our experiment. In contrast, the adaptive challenge in our indole evolution experiment was smaller. It required cells to bypass only a defective *trpB* gene, encoding tryptophan synthase. This could be minimally accomplished by a single

change, the overexpression of the *tna* operon (Cruz-Vera & Yanofsky, 2014) (Text S1, Fig. S10B). However, the varied genetic changes that occurred in our  $R^+$  and  $R^-$  populations suggest that this adaptive challenge can be met in multiple ways.

These observations illustrate a limitation of our experiment that it shares with many other laboratory evolution experiments where complex genetic changes occur: actual adaptive changes are more diverse and varied than likely adaptive mutations predicted from previous biochemical knowledge (Text S1, Fig. S10). And in the absence of a full mechanistic understanding of such changes, it is difficult to explain why recombination is more effective during adaptation to indole than to L-arabinose. For example, we cannot distinguish whether genetic factors, such as different forms of epistasis (Kouyos *et al.*, 2007; Moradigaravand *et al.*, 2014), different biochemical properties of nutrient metabolism, or both, are responsible for differences in the adaptive role of recombination.

Because recombination provided a benefit for adaptation to indole, we focused our genomic analyses on indole-adapted populations. Specifically, we sequenced the genomes of at least 46 unique clones from each of the two highest-fitness  $R^+$  and  $R^-$  populations evolved on indole. We first studied population-level changes that can result from recombination, such as decreased linkage disequilibrium (Barton & Charlesworth, 1998; Abecasis *et al.*, 2001), and increased genetic diversity (Souza *et al.*, 1997; Hellmann *et al.*, 2003; Hanage *et al.*, 2006). Linkage disequilibrium quantifies how often two alleles occur on the same haplotype. It is expected to be low in highly recombining populations, where Hill–Robertson interference is predicted to be weak (McVean & Charlesworth, 2000), because every recombination event between two loci separates two co-occurring alleles. Unfortunately, our populations harboured too few point mutations, which occurred at a frequency too low to reliably estimate differences in linkage disequilibrium between  $R^+$  and  $R^-$  populations. In contrast,  $R^+$  populations indeed harboured more genetic diversity than  $R^-$  populations. This is consistent with the previous experimental finding of (Souza *et al.*, 1997) that recombination increases population diversity. We do not know the fitness consequences of this diversity, but note that a recombination-induced increase in diversity can facilitate adaptation, if the loci involved affect fitness (Charlesworth, 1993).

Our evolved genomes harboured between six and 24 small-scale genetic changes (base substitutions and indels, Table 1). We focused on those changes that are potentially adaptive, because they had increased to a frequency exceeding 10% (Table 2). Among them is a nonsynonymous mutation in the DNA-binding region of *dgoR* (Rigali *et al.*, 2002) in the  $R_2^-$  population (Table 2). *dgoR* mutations enhance *E. coli* growth and D-galactonate metabolism in the mouse gut (Lescat

*et al.*, 2016), but do not have a known relationship with indole metabolism. We also identified a high frequency mutant allele in the adenylate cyclase (*cyaA*) gene, which occurred in the  $R_2^+$  population. *cya*<sup>-</sup> mutants are resistant to UV, thermal stress and antibiotics but divide more slowly than wild-type cells (Kumar, 1976). We speculate that these *cya* mutants may have arisen in response to toxins present in the growth medium, because *E. coli* toxin genes are induced by high intracellular indole (Domka *et al.*, 2006; Hu *et al.*, 2010).

The most intriguing small-scale change involves the *tna* operon, which can affect a cell's ability to utilize indole. Three of our four replicates harboured mutations in the *tnaC* gene that had risen to a frequency > 10% (Table 2, Fig. 3) in multiple replicate populations. All of these mutations occur in a small region of *tnaC* previously associated with the constitutive expression of the entire operon (Konan & Yanofsky, 2000; Gong & Yanofsky, 2003; Trabuco *et al.*, 2010). The operon consists of the three genes *tnaC*, which encodes a cis-regulatory leader peptide (Gish & Yanofsky, 1995), *tnaA* which encodes the enzyme tryptophanase (Deeley & Yanofsky, 1981), and *tnaB*, which encodes a tryptophan permease (Edwards & Yudkin, 1982). Tryptophanase catalyses a variety of reactions that involve tryptophan and indole, which has not only metabolic but also signalling functions (Hu *et al.*, 2010). Most importantly, tryptophanase is an enzyme that also catalyses the reaction creating tryptophan from indole (Watanabe & Snell, 1972), and can thus relieve the tryptophan auxotrophy of our ancestral strains. The expression of the *tna* operon is repressed in the absence of tryptophan through Rho-dependent premature transcriptional termination (Stewart *et al.*, 1986; Konan & Yanofsky, 1999, 2000). That is, the Rho termination factor can bind to the nascent *tna* mRNA and terminate transcription before *tnaA* and *tnaB* are expressed (Cruz-Vera & Yanofsky, 2014), if translation of the *tnaC* leader peptide is enabled by the presence of tryptophan. In the course of our experiment, where tryptophan is gradually replaced by indole, the wild-type *tna* operon would become increasingly repressed. Following up on the *tnaC* mutation, we observed that the  $R_2^+$  population additionally harboured a mutation in *rho*, which directly regulates *tnaA* transcription. It is a nonsynonymous mutation in the RNA-binding domain of the Rho protein (Dombroski *et al.*, 1988). More generally, *rho* is also a common selection target driving adaptation to stressful environments (Tenaillon *et al.*, 2012b). This *tnaC*-*rho* allele combination may be an instance of the Hill–Robertson effect (Hill & Robertson, 1966), because it may have emerged through independent mutations that have been brought together by recombination.

We next turned our attention to segmental and gene amplifications, which are important classes of structural changes that are caused by recombination (reviewed in

Hastings *et al.*, 2009b; Reams & Roth, 2015). Such amplifications have been observed in multiple evolution experiments and are a recurring adaptive strategy (Riehle *et al.*, 2001; Dunham *et al.*, 2002; Kugelberg *et al.*, 2006; Andersson & Hughes, 2009; Sun *et al.*, 2009; Blount *et al.*, 2012; Maharjan *et al.*, 2013; Raeside *et al.*, 2014; Toll-Riera *et al.*, 2016). Gene amplification can be highly beneficial if increased gene expression is adaptive (Riehle *et al.*, 2001; Dunham *et al.*, 2002; Kugelberg *et al.*, 2006). However, amplified genes may be maintained only transiently, because high gene expression can also carry substantial fitness costs. The same holds for segmental duplications, which carry the added pleiotropic cost of overexpressing multiple genes (Kugelberg *et al.*, 2006; Chang *et al.*, 2013). Adaptive amplification frequently involves ectopic recombination anchored at transposons and their DNA repeats (Dunham *et al.*, 2002; Kugelberg *et al.*, 2006; Chang *et al.*, 2013).

We found evidence for multiple segmental and gene amplifications in our sequenced clones. The nature of our sequence data did not allow us to reconstruct the multiple recombination events that gave rise to these amplifications.  $R^+$  populations did not show a greater overall incidence of such amplifications than  $R^-$  populations (Table S9 test 16,  $P = 0.09$ ). We focused our in-depth analysis on one  $R^-$  population ( $R_1^-$ ) and one  $R^+$  population ( $R_2^+$ ), because in these populations the indole-adapted subpopulation harboured significantly more clones with amplifications than the total population. Focusing further on a 32 kb region with 23 genes, in which amplification occurred at especially high frequency in these two populations, we found that the affected genes had significantly greater copy number in the  $R_2^+$  compared to the  $R_1^-$  lineage (Table S9 test 25).

Remarkably, the amplified region encompasses the three genes of the *tna* operon, where our previous analysis had identified point mutations of potential adaptive significance. We also found clones in the  $R_2^+$  population that combine these two kinds of mutations (Text S4). Especially noteworthy is a combination of a point mutation in the Rho termination factor of *tna*, and an amplification of this operon, which is especially prevalent in the indole-adapted subpopulation of  $R_2^+$ , and confers high fitness in liquid culture (Table S8 test 25  $P < 1.1 \times 10^{-16}$ ). Some clones harbour only one, but not the other mutation.

One limitation of our observations is that they do not allow us to disentangle genomic changes due to differences in recombination rate from replicate-specific changes that may be caused by historically contingent genetic events, even though both kinds of changes likely occurred. For example, historically contingent mutation or recombination events explain why a specific, putatively beneficial allele combination only arose in population  $R_2^+$ . Conversely, the consistently greater incidence of putatively adaptive amplifications in

indole-adapted samples of this population shows that recombination-specific effects also exist. This and other limitations of our experiments could be alleviated in obvious ways, by evolving many more replicate populations and for many more generations, as well as by sequencing many more genomes in a highly time-resolved manner. However, considering the laborious nature of our short-term experiment, with thousands of fitness measurements and hundreds of sequenced genomes, such extensions will remain challenging in the foreseeable future.

In sum, our experiments make three core observations. First, the environment can be crucial in determining whether recombination is beneficial for adaptive evolution. The reason is that the two parallel experiments we conducted started with the same  $R^+$  and  $R^-$  strain pairs but showed different benefits of recombination, depending on which of the two nutrients we used. Second, adaptation is not strictly limited by the recombination rate in our experiments, because even our  $R^-$  strains, where this rate is 25-fold lower, adapted to some extent to the new nutrient. Third, where recombination facilitates adaptation in our experiments, it is likely to do so in two ways. It causes segmental amplifications, for example, of operons like *tna*, and it may bring adaptive alleles together on the same chromosome. The relevant changes are complex and range from single nucleotide alterations to multigene amplifications.

## Acknowledgments

We would like to thank James Winkler and Katy Kao for providing us with the ancestral strains. This work was supported by the Swiss National Science Foundation (31003A\_146137) and SystemsX.ch (EpiphysX RTD). We also thank Olivier Tenaillon for valuable comments on a previous version of this manuscript. We acknowledge Dr. Heidi Lischer for her advice on data analyses and Dr. Lucy Poveda Mozolowski at the Functional Genomics Center Zurich (FGCZ) for her assistance on library preparations and sequencing. We also acknowledge support by the University Priority Research Program in Evolutionary Biology at the University of Zurich.

## References

- Abecasis, G., Noguchi, E., Heinzmann, A., Traherne, J., Bhatnagaryya, S., Leaves, N. *et al.* 2001. Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68**: 191–197.
- Agrawal, A.F., Hadany, L. & Otto, S.P. 2005. The evolution of plastic recombination. *Genetics* **171**: 803–812.
- Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W. *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402.
- Ambur, O., Engelstädter, J., Johnsen, P., Miller, E. & Rozen, D. 2016. Steady at the wheel: conservative sex and the benefits of bacterial transformation. *Phil. Trans. R. Soc. B* **371**: 20150528.
- Andersson, D. & Hughes, D. 2009. Gene amplification and adaptive evolution in bacteria. *Annu. Rev. Genet.* **43**: 167–195.
- Arber, W. 2014. Horizontal gene transfer among bacteria and its role in biological evolution. *Life* **4**: 217–224.
- Auwer, G., Carneiro, M., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A. *et al.* 2013. From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**: 11.10.1–11.10.33.
- Avison, M., Horton, R., Walsh, T. & Bennett, P. 2001. *Escherichia coli* CreBC is a global regulator of gene expression that responds to growth in minimal media. *J. Biol. Chem.* **276**: 26955–26961.
- Baltrus, D.A. 2013. Exploring the costs of horizontal gene transfer. *Trends Ecol. Evol.* **28**: 489–495.
- Barton, N. & Charlesworth, B. 1998. Why sex and recombination? *Science* **281**: 1986–1990.
- Bernstein, H. 1977. Germline recombination may be primarily a manifestation of dna repair processes. *J. Theor. Biol.* **69**: 371–380.
- Blount, Z., Barrick, J., Davidson, C. & Lenski, R. 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* **489**: 513–518.
- Bobay, L.M., Traverse, C.C. & Ochman, H. 2015. Impermanence of bacterial clones. *Proc. Natl. Acad. Sci. USA* **112**: 8893–8900.
- Bolger, A.M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.
- Bretscher, M.T., Althaus, C.L., Müller, V. & Bonhoeffer, S. 2004. Recombination in HIV and the evolution of drug resistance: for better or for worse? *BioEssays* **26**: 180–188.
- Cariss, S.J.L., Constantinidou, C., Patel, M.D., Takebayashi, Y., Hobman, J.L., Penn, C.W. *et al.* 2010. YieJ (CbrC) mediates CreBC-dependent colicin E2 tolerance in *Escherichia coli*. *J. Bacteriol.* **192**: 3329–3336.
- Chang, S.L., Lai, H.Y., Tung, S.Y. & Leu, J.Y. 2013. Dynamic large-scale chromosomal rearrangements fuel rapid adaptation in yeast populations. *PLoS Genet.* **9**: e1003232.
- Charlesworth, B. 1993. Directional selection and the evolution of sex and recombination. *Genet. Res.* **61**: 205–224.
- Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L. *et al.* 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: Snps in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**: 80–92.
- Cooper, T.F. 2007. Recombination speeds adaptation by reducing competition between beneficial mutations in populations of *Escherichia coli*. *PLoS Biol.* **5**: e225.
- Croucher, N.J., Mostowy, R., Wymant, C., Turner, P., Bentley, S.D. & Fraser, C. 2016. Horizontal dna transfer mechanisms of bacteria as weapons of intragenomic conflict. *PLoS Biol.* **14**: e1002394.
- Cruz-Vera, L.R. & Yanofsky, C. 2014. Instructing the translating ribosome to sense L-Tryptophan during synthesis of the

- TnaC nascent regulatory peptide. In: *Regulatory Nascent Polypeptides* (K. Ito, ed.), pp. 151–163. Springer-Verlag, New York.
- Darmon, E. & Leach, D.R.F. 2014. Bacterial genome instability. *Microbiol. Mol. Biol. Rev.* **78**: 1–39.
- Deatherage, D.E. & Barrick, J.E. 2014. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. In: *Engineering and Analyzing Multicellular Systems: Methods and Protocols* (L. Sun, W. Shou, eds), pp. 165–188. Springer-Verlag, New York.
- Deatherage, D.E., Traverse, C.C., Wolf, L.N. & Barrick, J.E. 2015. Detecting rare structural variation in evolving microbial populations from new sequence junctions using breseq. *Front. Genet.* **5**: 468.
- Deeley, M.C. & Yanofsky, C. 1981. Nucleotide sequence of the structural gene for tryptophanase of *Escherichia coli* K-12. *J. Bacteriol.* **147**: 787–796.
- Didelot, X., Méric, G., Falush, D. & Darling, A.E. 2012. Impact of homologous and nonhomologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genom.* **13**: 1–15.
- Dombroski, A.J., Brennan, C.A., Spear, P. & Platt, T. 1988. Site-directed alterations in the ATP-binding domain of Rho protein affect its activities as a termination factor. *J. Biol. Chem.* **263**: 18802–18809.
- Domka, J., Lee, J. & Wood, T.K. 2006. YliH (BssR) and YceP (BssS) regulate *Escherichia coli* K-12 biofilm formation by influencing cell signaling. *Appl. Environ. Microbiol.* **72**: 2449–2459.
- Dunham, M.J., Badrane, H., Ferea, T., Adams, J., Brown, P.O., Rosenzweig, F. *et al.* 2002. Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* **99**: 16144–16149.
- Edwards, R.M. & Yudkin, M.D. 1982. Location of the gene for the low-affinity tryptophan-specific permease of *Escherichia coli*. *Biochem. J.* **204**: 617–619.
- Felsenstein, J. 1974. The evolutionary advantage of recombination. *Genetics* **78**: 737–756.
- Fisher, R.A. 1930a. The fundamental theorem of natural selection. In: *The Genetical Theory of Natural Selection* (H. Bennett, ed.), pp. 22–47. The Clarendon Press, Oxford.
- Fisher, R.A. 1930b. Sexual reproduction and sexual selection. In: *The Genetical Theory of Natural Selection* (H. Bennett, ed.), pp. 121–145. The Clarendon Press, Oxford.
- Gish, K. & Yanofsky, C. 1995. Evidence suggesting cis action by the TnaC leader peptide in regulating transcription attenuation in the tryptophanase operon of *Escherichia coli*. *J. Bacteriol.* **177**: 7245–7254.
- Goddard, M.R., Godfray, H.C.J. & Burt, A. 2005. Sex increases the efficacy of natural selection in experimental yeast populations. *Nature* **434**: 636–640.
- Gong, F. & Yanofsky, C. 2003. A transcriptional pause synchronizes translation with transcription in the tryptophanase operon leader region. *J. Bacteriol.* **185**: 6472–6476.
- Gray, J.C. & Goddard, M.R. 2012. Sex enhances adaptation by unlinking beneficial from detrimental mutations in experimental yeast populations. *BMC Evol. Biol.* **12**: 1.
- Griffiths, J.G. & Bonser, S.P. 2013. Is sex advantageous in adverse environments? A test of the abandon-ship hypothesis. *Am. Nat.* **182**: 718–725.
- Hadany, L. & Beker, T. 2003. On the evolutionary advantage of fitness-associated recombination. *Genetics* **165**: 2167–2179.
- Hall, B.G. 1993. The role of single-mutant intermediates in the generation of *trpAB* double revertants during prolonged selection. *J. Bacteriol.* **175**: 6411–6414.
- Hanage, W.P., Fraser, C. & Spratt, B.G. 2006. The impact of homologous recombination on the generation of diversity in bacteria. *J. Theor. Biol.* **239**: 210–219.
- Hartfield, M. & Keightley, P.D. 2012. Current hypotheses for the evolution of sex and recombination. *Integr. Zool.* **7**: 192–209.
- Hastings, P.J., Bull, H.J., Klump, J.R. & Rosenberg, S.M. 2000. Adaptive amplification: an inducible chromosomal instability mechanism. *Cell* **103**: 723–731.
- Hastings, P.J., Ira, G. & Lupski, J.R. 2009a. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* **5**: e1000327.
- Hastings, P.J., Lupski, J.R., Rosenberg, S.M. & Ira, G. 2009b. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10**: 551–564.
- Hellmann, I., Ebersberger, I., Ptak, S.E., Pääbo, S. & Przeworski, M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72**: 1527–1535.
- Hill, W.G. & Robertson, A. 1966. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**: 269–294.
- Hu, M., Zhang, C., Mu, Y., Shen, Q. & Feng, Y. 2010. Indole affects biofilm formation in bacteria. *Indian J. Microbiol.* **50**: 362–368.
- Jeanson, S., Floury, J., Gagnaire, V., Lortal, S. & Thierry, A. 2016. Bacterial colonies in solid media and foods: a review on their growth and interactions with the micro-environment. *Front. Microbiol.* **6**: 6.
- Kingston, A.W., Roussel-Rossin, C., Dupont, C. & Raleigh, E.A. 2015. Novel reca-independent horizontal gene transfer in *Escherichia coli* K-12. *PLoS ONE* **10**: e0130813.
- Konan, K.V. & Yanofsky, C. 1999. Role of ribosome release in regulation of *tna* operon expression in *Escherichia coli*. *J. Bacteriol.* **181**: 1530–1536.
- Konan, K.V. & Yanofsky, C. 2000. Rho-dependent transcription termination in the *tna* operon of *Escherichia coli*: roles of the boxA sequence and the rut site. *J. Bacteriol.* **182**: 3981–3988.
- Koonin, E.V. 2016. Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions [version 1; referees: 2 approved]. *F1000Research* **5**: 1805. <https://doi.org/10.12688/f1000research.8737.1>.
- Kouyos, R.D., Silander, O.K. & Bonhoeffer, S. 2007. Epistasis between deleterious mutations and the evolution of recombination. *Trends Ecol. Evol.* **22**: 308–315.
- Kugelberg, E., Kofoid, E., Reams, A.B., Andersson, D.I. & Roth, J.R. 2006. Multiple pathways of selected gene amplification during adaptive mutation. *Proc. Natl. Acad. Sci. USA* **103**: 17319–17324.
- Kumar, S. 1976. Properties of adenylyl cyclase and cyclic adenosine 3', 5'-monophosphate receptor protein-deficient mutants of *Escherichia coli*. *J. Bacteriol.* **125**: 545–555.
- Langmead, B. & Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**: 357–359.
- Lehtonen, J., Jennions, M.D. & Kokko, H. 2012. The many costs of sex. *Trends Ecol. Evol.* **27**: 172–178.
- Lescat, M., Launay, A., Ghalayini, M., Magnan, M., Glodt, J., Pintard, C. *et al.* 2016. Using long-term experimental evolution to uncover the patterns and determinants of molecular

- evolution of an *Escherichia coli* natural isolate in the streptomycin-treated mouse gut. *Mol. Ecol.* **26**: 1802–1817.
- Lin, M. & Kussell, E. 2016. Correlated mutations and homologous recombination within bacterial populations. *Genetics* **205**: 891–917.
- Lobkovsky, A.E., Wolf, Y.I. & Koonin, E.V. 2016. Evolvability of an optimal recombination rate. *Genome Biol. Evol.* **8**: 70–77.
- Magurran, A.E. 1988. *Ecological Diversity and its Measurement*. Princeton University Press, Princeton.
- Maharjan, R.P., Gaffé, J., Plucain, J., Schliep, M., Wang, L., Feng, L. *et al.* 2013. A case of adaptation through a mutation in a tandem duplication during experimental evolution in *Escherichia coli*. *BMC Genom.* **14**: 1–12.
- McDonald, M.J., Rice, D.P. & Desai, M.M. 2016. Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature* **531**: 233–236.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytzky, A. *et al.* 2010. The genome analysis toolkit: a map reduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**: 1297–1303.
- McVean, G.A. & Charlesworth, B. 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* **155**: 929–944.
- Michod, R.E., Wojciechowski, M.F. & Hoelzer, M.A. 1988. DNA repair and the evolution of transformation in the bacterium *Bacillus subtilis*. *Genetics* **118**: 31–39.
- Michod, R.E., Bernstein, H. & Nedelcu, A.M. 2008. Adaptive value of sex in microbial pathogens. *Infect. Genet. Evol.* **8**: 267–285.
- Moradigaravand, D., Kouyos, R., Hinkley, T., Haddad, M., Petropoulos, C.J., Engelstädter, J. *et al.* 2014. Recombination accelerates adaptation on a large-scale empirical fitness landscape in HIV-1. *PLoS Genet.* **10**: e1004439.
- Muller, H.J. 1932. Some genetic aspects of sex. *Am. Nat.* **66**: 118–138.
- Muller, H.J. 1964. The relation of recombination to mutational advance. *Mutat. Res.* **1**: 2–9.
- Nagaraja, P., Alexander, H.K., Bonhoeffer, S. & Dixit, N.M. 2016. Influence of recombination on acquisition and reversion of immune escape and compensatory mutations in HIV-1. **14**: 11–25.
- Ochman, H., Lawrence, J.G. & Groisman, E.A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299–304.
- Otto, S. & Lenormand, T. 2002. Resolving the paradox of sex and recombination. *Nat. Rev. Genet.* **3**: 252–261.
- Peabody, G., Winkler, J., Fountain, W., Castro, D.A., Leiva-Aravena, E. & Kao, K.C. 2016. Benefits of a Recombination-Proficient *Escherichia coli* System for Adaptive Laboratory Evolution. *Appl. Environ. Microbiol.* **82**: 6736–6747.
- Pesce, D., Lehman, N. & de Visser, J.A.G. 2016. Sex in a test tube: testing the benefits of *in vitro* recombination. *Phil. Trans. R. Soc. B* **371**: 20150529.
- Raeside, C., Gaffé, J., Deatherage, D.E., Tenaille, O., Briska, A.M., Ptashkin, R.N. *et al.* 2014. Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. *MBio* **5**: e01377-14.
- Ram, Y. & Hadany, L. 2016. Condition-dependent sex: who does it, when and why? *Phil. Trans. R. Soc. B* **371**: 20150539.
- Reams, A.B. & Roth, J.R. 2015. Mechanisms of gene duplication and amplification. *Cold Spring Harb. Perspect. Biol.* **7**: a016592.
- Riehle, M.M., Bennett, A.F. & Long, A.D. 2001. Genetic architecture of thermal adaptation in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **98**: 525–530.
- Rigali, S., Derouaux, A., Giannotta, F. & Dusart, J. 2002. Subdivision of the helix-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. *J. Biol. Chem.* **277**: 12507–12515.
- Riley, M., Abe, T., Arnaud, M., Berlyn, M., Blattner, F., Chaudhuri, R. *et al.* 2006. *Escherichia coli* K-12: a cooperatively developed annotation snapshot – 2005. *Nucleic Acids Res.* **34**: 1–9.
- Rodríguez-Beltrán, J., Turret, J., Tenaille, O., López, E., Bourdelier, E., Costas, C. *et al.* 2015. High recombinant frequency in extra intestinal pathogenic *Escherichia coli* strains. *Mol. Biol. Evol.* **32**: 1708–1716.
- Rosenberg, S.M., Shee, C., Frisch, R.L. & Hastings, P. 2012. Stress-induced mutation via DNA breaks in *Escherichia coli*: a molecular mechanism with implications for evolution and medicine. *BioEssays* **34**: 885–892.
- Shannon, C. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**: 379–423.
- Shen, P. & Huang, H.V. 1986. Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. *Genetics* **112**: 441–457.
- Shen, D., Liu, T., Ye, W., Liu, L., Liu, P., Wu, Y. *et al.* 2013. Gene duplication and fragment recombination drive functional diversification of a super family of cytoplasmic effectors in *Phytophthora sojae*. *PLoS ONE* **8**: e70036.
- Souza, V., Turner, P.E. & Lenski, R.E. 1997. Long-term experimental evolution in *Escherichia coli* V. effects of recombination with immigrant genotypes on the rate of bacterial evolution. *J. Evol. Biol.* **10**: 743–769.
- Sprouffske, K. & Wagner, A. 2016. Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves. *BMC Bioinformatics* **17**: 1–4.
- Stewart, V., Landick, R. & Yanofsky, C. 1986. Rho-dependent transcription termination in the tryptophanase operon leader region of *Escherichia coli* K-12. *J. Bacteriol.* **166**: 217–223.
- Sun, S., Berg, O., Roth, J. & Andersson, D. 2009. Contribution of gene amplification to evolution of increased antibiotic resistance in *Salmonella typhimurium*. *Genetics* **182**: 1183–1195.
- Takeuchi, N., Kaneko, K. & Koonin, E.V. 2014. Horizontal gene transfer can rescue prokaryotes from Muller's ratchet: benefit of DNA from dead cells and population subdivision. *G3* **4**: 325–339.
- Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R.L., McDonald, P., Bennett, A.F., Long, A.D. *et al.* 2012a. The molecular diversity of adaptive convergence. *Science* **335**: 457–461.
- Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R.L., McDonald, P., Bennett, A.F., Long, A.D. *et al.* 2012b. The molecular diversity of adaptive convergence. *Science* **335**: 457–461.
- Thomas, C.M. & Nielsen, K.M. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**: 711–721.
- Toll-Riera, M., SanMillan, A., Wagner, A. & MacLean, R.C. 2016. The genomic basis of evolutionary innovation in *Pseudomonas aeruginosa*. *PLoS Genet.* **12**: e1006005.

- Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P. *et al.* 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* **5**: e1000344.
- Trabuco, L.G., Harrison, C.B., Schreiner, E. & Schulten, K. 2010. Recognition of the regulatory nascent chain TnaC by the ribosome. *Structure* **18**: 627–637.
- Typas, A., Nichols, R., Siegele, D., Shales, M., Collins, S., Lim, B. *et al.* 2008. High-throughput, quantitative analyses of genetic interactions in *E. coli*. *Nat. Methods* **5**: 781–787.
- VanLiere, J.M. & Rosenberg, N.A. 2008. Mathematical properties of the  $r^2$  measure of linkage disequilibrium. *Theor. Popul. Biol.* **74**: 130–137.
- Vos, M. 2009. Why do bacteria engage in homologous recombination? *Trends Microbiol.* **17**: 226–232.
- Vos, M., Hesselman, M.C., te Beek, T.A., van Passel, M.W.J. & Eyre-Walker, A. 2015. Rates of lateral gene transfer in prokaryotes: high but why? *Trends Microbiol.* **23**: 598–605.
- Watanabe, T. & Snell, E.E. 1972. Reversibility of the tryptophanase reaction: synthesis of tryptophan from indole, pyruvate, and ammonia. *Proc. Natl. Acad. Sci. USA* **69**: 1086–1090.
- Weismann, A. 1889. The significance of sexual reproduction in the theory of natural selection. In: *Essays upon Heredity and Kindred Biological Problems* (E.B. Poulton, S. Schonland & A.E. Shipley, eds), vol. **1**, pp. 255–332. The Clarendon Press, Oxford.
- Whitlock, A.O., Peck, K.M., Azevedo, R.B. & Burch, C.L. 2016. An evolving genetic architecture interacts with Hill–Robertson interference to determine the benefit of sex. *Genetics* **203**: 923–936.
- Wielgoss, S., Barrick, J.E., Tenaillon, O., Cruveiller, S., Chane-Woon-Ming, B., Médigue, C. *et al.* 2011. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3* **1**: 183–186.
- Winkler, J. & Kao, K.C. 2012. Harnessing recombination to speed adaptive evolution in *Escherichia coli*. *Metab. Eng.* **14**: 487–495.
- Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* **19**: 1586–1592.
- Zeyl, C. & Bell, G. 1997. The advantage of sex in evolving yeast populations. *Nature* **388**: 465–468.

## Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article:

**Text S1** Potential adaptive routes.

**Text S2** Linkage disequilibrium (LD) analysis.

**Text S3** The accuracy of our gene copy number detection procedure.

**Text S4** Novel genotype combinations found in the replicate populations.

**Table S1** Summary of bacterial genotypes.

**Table S2** Nutrient concentrations for growth characterization of ancestral  $R^+$  and  $R^-$  strains.

**Table S3** Nutrient concentrations used in experimental evolution.

**Table S4** Fitness summary of clones of the indole-adapted subpopulation and total population of selected evolved populations.

**Table S5** Summary of the number of clones from the  $R^+$  and  $R^-$  indole-evolved populations used in various analyses.

**Table S6** Summary of genomic regions excluded in the data analysis because of their likely involvement in sequencing artefacts.

**Table S7** Summary of population diversity and the number of segmental amplified clones in the  $R^+$  and  $R^-$  populations.

**Table S8** Summary of the accuracy of our structural variation detection pipeline using simulated data.

**Table S9** Statistical tests conducted in this study.

**Table S10** Tests for number of copies of operons.

**Table S11** Fitness of clones with various allelic combinations.

**Figure S1** Growth of *E. coli* MG1655, and ancestral  $R^+$  and  $R^-$  strains.

**Figure S2** Cell density of evolving replicates.

**Figure S3** Fitness of clones isolated from selected evolved replicate populations.

**Figure S4** Flow chart illustrating filtering steps for selecting the indole-evolved replicate populations for genomic analysis.

**Figure S5** Distributions of genomewide sequence coverage of 384 sequenced clones.

**Figure S6** F plasmid integration into the genome validated by read alignment and mapping of reads to the F plasmid sequence.

**Figure S7** Growth improvement of clones from  $R_1^+$ ,  $R_2^+$ ,  $R_1^-$ , and  $R_2^-$  replicate populations in three different conditions.

**Figure S8** Possible genotype frequency categories for biallelic sites, how they can arise, and the resulting LD patterns.

**Figure S9** Insufficient polymorphism data to infer  $r^2$ -decay.

**Figure S10** Schematic diagram of the original and alternative L-arabinose catabolic pathway and L-tryptophan biosynthetic pathway.

Received 1 March 2017; revised 11 May 2017; accepted 5 June 2017