Control regions for chromosome replication are conserved with respect to sequence and location among Escherichia coli strains

Jakob Frimodt-Møller1,2, Godefroid Charbon1, Karen A. Krogfelt2 and Anders Løbner-Olesen1*

1 Department of Biology, Section for Functional Genomics and Center for Bacterial Stress Response and Persistence, University of Copenhagen, Copenhagen, Denmark, 2 Department of Microbiology and Infection Control, Statens Serum Institut, Copenhagen, Denmark

In Escherichia coli, chromosome replication is initiated from oriC by the DnaA initiator protein associated with ATP. Three non-coding regions contribute to the activity of DnaA. The datA locus is instrumental in conversion of DnaAATP to DnaAADP (datA dependent DnaAATP hydrolysis) whereas DnaA rejuvenation sequences 1 and 2 (DARS1 and DARS2) re activate DnaAADP to DnaAATP. The structural organization of oriC, datA, DARS1, and DARS2 were found conserved among 59 fully sequenced E. coli genomes, with differences primarily in the non-functional spacer regions between key protein binding sites. The relative distances from oriC to datA, DARS1, and DARS2, respectively, was also conserved despite of large variations in genome size, suggesting that the gene dosage of either region is important for bacterial growth. Yet all three regions could be deleted alone or in combination without loss of viability. Competition experiments during balanced growth in rich medium and during mouse colonization indicated roles of datA, DARS1, and DARS2 for bacterial fitness although the relative contribution of each region differed between growth conditions. We suggest that this fitness advantage has contributed to conservation of both sequence and chromosomal location for datA, DARS1, and DARS2.

Keywords: E. coli, oriC, non-coding replication control regions, datA, DARS1 and DARS2, fitness in vitro and in vivo

Introduction

In Escherichia coli chromosome replication is initiated from a single origin, oriC, and proceeds bi-directionally until the two replication forks meet at terminus of replication (terC). The initiator protein DnaA belongs to the AAA+ (ATPases Associated with diverse Activities) proteins. DnaA can bind ATP and ADP with similar high affinities (Skarstad and Katayama, 2013), but only DnaA bound to ATP is able to initiate DNA replication (Sekimizu et al., 1987). Different recognition sites for DnaA has been identified in oriC; three high to medium affinity sites (R1, R4, and R2) that binds both DnaAATP and DnaAADP (Fuller et al., 1984), and multiple lower affinity sites (R3, R5/M, I1, I2, I3, C1, C2, C3, r1, and r2) (McGarry et al., 2004; Kawakami et al., 2005; Rozgaja et al., 2011) (Figure 1). Only DnaAATP is capable to bind to low-affinity sites (McGarry et al., 2004; Kawakami et al., 2005; Rozgaja et al., 2011), and single stranded DnaA boxes (Speck and Messer, 2001). Binding of the Fis protein to oriC is reported to both inhibit initiation of
Initiation of replication is a highly regulated process in *E. coli*. Replication begins essentially simultaneously at all cellular origins (Skarstad et al., 1986), i.e., in synchrony and only once per cell cycle. The tight control is primarily ensured by the oscillation of DnaATP, which has a temporal increase around the time of initiation, and decreases rapidly thereafter (Kurokawa et al., 1999). Following initiation, oriC is temporarily inactivated by the binding of SeqA to hemi-methylated GATC-sites (Campbell and Kleckner, 1990; Lu et al., 1994). This sequestration lasts for about 1/3 of the doubling time and provides a time period for RIDA (Regulatory Inactivation of DnaA) and DDAH (datA-dependent DnaATP hydrolysis) to hydrolyse DnaATP to DnaADP. In RIDA, the Hda protein, in association with the DNA-loaded β-clamp (DnaN), activates the intrinsic ATPase activity of DnaA, which converts DnaATP into DnaADP (Kurokawa et al., 1999; Kato and Katayama, 2001). DDAH is an IFH dependent hydrolysis of DnaATP to DnaADP, which takes place at the datA locus (Kasho and Katayama, 2013). *datA* contains five DnaA boxes as well as an IFH-binding site (Nozaki et al., 2009; Kasho and Katayama, 2013) (Figure 1). Common for both RIDA and DDAH is that both processes lower the DnaATP/DnaADP ratio to counter unwanted re-initiation of replication. At later stages in the cell cycle the DnaATP level must increase past a critical level for a new round of initiation of replication. This is done by rejuvenation of DnaADP to DnaATP at the DARS1 and DARS2 loci, where rejuvenation at the DARS2 locus is dependent on IHF and Fis (Kasho et al., 2014). In addition *de novo* synthesis of DnaA, which by and large will be ATP bound because ATP is more abundant than ADP within the cell, will also contribute to the increase in DnaATP (Kurokawa et al., 1999). DARS1 and DARS2 contain a core of three DnaA boxes (Figure 1). In addition, DARS1 needs a specific DNA region flanking the core for stimulation of ADP dissociation from DnaA (Fujimitsu et al., 2009), while DARS2 contains three additional DnaA boxes and requires both Fis binding sites (FBS) 2 and 3, and IHF binding.
to IHF binding sites (IBS) 1 and 2 be active (Kasho et al., 2014) (Figure 1).

Termination of replication occurs in terC, a poorly-defined region approximately 180° away from oriC (Hill et al., 1987). If an uneven number of homologous recombination events between daughter chromosomes have taken place during replication, the end result will be a chromosome dimer (Sherratt et al., 2004). Resolution takes place at a 28 bp site dif, located in terC (Sherratt et al., 2004) in a process involving two tyrosine recombinases, XerC and XerD. The XerCD recombinase is activated and delivered at dif by the PtsK translocase (Bigot et al., 2005). Numerous forces seem to shape the organization of bacterial chromosomes, and the pattern of these forces on the chromosome is evident at different levels. In both the Gram-negative bacteria E. coli (Bergthorsson and Ochman, 1998) and Salmonella enterica (Liu and Sanderson, 1995a,b), as well as the Gram-positive bacterium Lactococcus lactis (Campo et al., 2004), selective pressure maintains a global architecture of the chromosome, which preserves two replication arms of nearly equal length. In addition to chromosome symmetry further chromosomal constrains are observed in E. coli. Four insulated macrodomains (MD) and two less constrained regions called non-structured (NS) regions has been uncovered (Niki et al., 2000; Valens et al., 2004). MDs are defined as regions where DNA interactions occur preferentially, while DNA interactions between the different MDs are highly restricted. NS regions can however interact with both its flaking MDs (Valens et al., 2004). oriC and dataA are contained within the Ori MD, while the Ter MD contains dif. The Ori MD is flanked by NSRight and NSLeft (where DARS2 is found) whereas the Ter MD is flanked by the Left MD and the Right MD which contain DARS1 (Valens et al., 2004) (Figure 2). Several observations indicates that the MDs and NS plays a part in the segregation of sister chromatids and the mobility of chromosomal DNA. The Ori MD is centered on a centromere-like 25 bp sequence designated migS, which affects oriC positioning during chromosome segregation (Yamaichi and Niki, 2004; Fekete and Chattoraj, 2005). Furthermore, movement of the Ter MD is maintained by several factors including the MatP/matS system (Mercier et al., 2008), and ZapA, ZapB, and PtsZ (Espeli et al., 2012; Buss et al., 2015).

Despite of the restraints on the E. coli K-12 chromosome, the size of genomes of other E. coli species varies from 4.6 to 5.7 Mb, indicating that horizontal gene transfer and genome reductions frequently takes place (Leimbach et al., 2013). A very dynamic genome structure underlies the metabolic and phenotypic diversity of E. coli. The genome of a bacterial species can be grouped into two categories. The core genome contains genes present in all strains, while the flexible genome comprises genes that are present in only a few strains or unique to single isolates (Medini et al., 2005). The pan-genome of a bacterial species is the combination of the core genome and the flexible genome (Medini et al., 2005). A typical E. coli genome has approximately 5000 genes, where roughly 2200 genes represent the core genome (Rasko et al., 2008). E. coli has a very large pan-genome (~18,000 genes), which grows for each new genome sequenced (Medini et al., 2005; Rasko et al., 2008). This indicates that there is a great diversity in gene content between E. coli species. Nevertheless, comparison of bacterial

![FIGURE 2 | Location of dataA, DARS1, DARS2, and dif regions on the E. coli genomes. The range (the relative distance ± the standard deviation in centisomes) in the chromosomal position of dataA (red), DARS1 (blue), DARS2 (green), and dif (gray) are shown in the outer circle. For each region the average relative distance with the standard deviation (see Materials and Methods), the median, as well as the chromosomal position in MG1655 (MG1655) are indicated. All distances are given in centisomes. The inner circle schematically shows the location of the different MD- and NS-regions as indicated by Esnault et al. (2007).](image-url)
chromosomes from related genera revealed a conservation of organization (Eisen et al., 2000). For instance, even though E. coli and Salmonella typhimurium diverged from a common ancestor about 140 million years ago their genetic maps are extensively superimposable (Groisman and Ochman, 1997).

Here we report a conserved chromosomal position of the non-coding regions dataA, DARS1, and DARS2 relative to oriC in E. coli. In addition, we report that the structural organization of oriC, dataA, DARS1, and DARS2 regions are conserved in all E. coli strains analyzed. Furthermore, we demonstrate that even though the loss of dataA, DARS1, or DARS2 did not result in a measurable reduction in growth rate, the mutant cells had a lower fitness than wild-type when tested under laboratory conditions or in mice.

Materials and Methods

Growth Conditions

Cells were grown in Luria–Bertani Broth (LB) or AB minimal medium supplemented with 0.2% glucose, 10 µg/ml thiamine, and when indicated 0.5% casamino acids. MacConkey agar was used for mice experiments. All cells were cultured at 37°C. When necessary, antibiotics were added to the following concentrations: kanamycin, 50 µg/ml; streptomycin, 100 µg/ml; chloramphenicol, 20 µg/ml; ampicillin, 150 µg/ml.

Bacterial Strains and Plasmids

A spontaneous streptomycin resistance mutant of MG1655 (ALO1825) was obtained by plating an overnight culture on streptomycin-plates, resulting in MG1655 StrR (ALO4292) (see Table S1 for used strains).

The DARS2 region containing DnaA box I-III was replaced with the cat gene in MG1655 by the lambda red procedure (Datsenko and Wanner, 2000), resulting in the ΔDARS2:cat mutant (ALO4254). Briefly, DNA fragments were PCR amplified using modified primers MutH-9 (5′-TCACAGTTATGTGCAAGTTATAAAAACAGAGGAAGGTGATAGCCTTTCATATTATCAAAAGCAGCAG-3′) and MutH-10 (5′-CTACGAGATTTACTACGGGAAAACCCGGAGCGATTTCTGATTTGCAAAGGCTTACACC-3′), where the underlined sequence will anneal to pKD3. Each deletion was verified by PCR. The DARS2 deletion was moved from ALO4254 to ALO4292 by P1 transduction using established procedures (Miller, 1972) and by selection for chloramphenicol resistance, resulting in ALO4310. The cat gene was removed from ALO4618 and ALO4619 by pCP20, before transformation of pALO75 (Løbner-Olesen et al., 1987) for investigation of β-galactosidase synthesis from the mioC promoter. Strain RB210 (MC1000 carries a dnaA-lacZ translational fusion on phage λRB1 integrated at attL) (Braun et al., 1985). λRB1 was transduced from RB210 to ALO1257, resulting in ALO1265. The deletion of DARS2 or dataA was done as described above and resulted in strains ALO4626 and ALO4627, respectively, for investigation of β-galactosidase synthesis from the dnaA promoter.

Flow Cytometry

Flow cytometry was performed as described previously (Løbner-Olesen et al., 1989) using an Apogee A10 instrument. For each sample, a minimum of 30,000 cells were analyzed. Numbers of origins per cell and relative cell mass were determined as described previously (Løbner-Olesen et al., 1989).

The distribution of origins per cell was measured after treating exponentially growing cells with rifampicin and cephalin for 4 h. Rifampicin block initiation of replication, while cephalin will block cell division. The average number of chromosomes per cell will therefore be equivalent to the number of oriC3's present in the cell at the time the drugs were added.

Relative Distance

The relative distance between oriC and DARS1, DARS2, dataA, and dif were calculated in centisomes (see equation below). Each E. coli chromosome is by definition 100 centisomes. The relative distance is set as the distance in base pairs between oriC and the region of interest divided by the size of the genome in base pairs of the investigated E. coli strain. There are two distances to dif, one for each replication arm. In this study only the shortest replication arm is presented, while the distance of the longest replication arm by default is the sum of the shortest replication arm subtracted from 100.

Relative distance in centisomes = \left(\frac{\text{Position of oriC} - \text{Position of region of interest}}{\text{Chromosome size}}\right) \times 100

To calculate the relative distance the chromosome needs to be fully assembled. For this the sequence of 70 fully assembled E. coli chromosomes from The European Nucleotide Archive (http://www.ebi.ac.uk/genomes/bacteria.html) were obtained and analyzed. Two times two strains were uploaded under the same
strain name, i.e., W and ST540. We denote them Wa (uploaded under sequence CP002185), Wb (uploaded under sequence CP002967), ST540a (uploaded under sequence CP002185), and ST540b (uploaded under sequence CP002967). Wa are identical to Wb, why Wa was used, while ST540b was used as ST540a was excluded (see below). BL21-DE3 and BL21-Gold were excluded for being deviates of B str. REL606, while KO11, KO11FL, and LY180 were excluded for being deviates of W.

Six E. coli genomes were found to have relative distances, which were two times the standard deviation or more away from the average (Table S2). Of these W3110, MC4100, and strain ST540a were excluded. ST540a and MC4100 has 20% or more imbalance between the length of the two replication arms, which have been shown to give abnormal cells that was dependent on the RecBC-dependent homologous recombination for viability (Essault et al., 2007), why they were excluded. W3110 is disqualified due to a known inversion around oriC (Hayashi et al., 2006), which explains the altered relative distances compared to the E. coli average.

The final dataset comprised of 59 fully assembled E. coli genomes (Table S3). The position and sequence of oriC, DARS1, DARS2, datA, and dif are known in MG1655, but not annotated in the dataset. We therefore choose to annotate the regions in 58 remaining E. coli genomes. The sequence of each of the regions (see Supplementary Data) from MG1655 where therefore aligned with the chosen E. coli genomes to obtain the chromosomal position and sequence of the region in each individual E. coli genomes.

**Mutation Frequency**

The mutation frequency was estimated for intergenic regions. Regions between protein-coding genes of more than 300 bp were selected in MG1655. These regions were trimmed for 100 bps on each side, to avoid conserved promoter-regions, and blasted against the 58 remaining E. coli genomes. If not present in the entire dataset the intergenic region was discharged, resulting in 109 regions. Of these 13 intergenic regions contained known conserved sRNA or tRNAs, why they were removed, resulting in 96 intergenic regions. Each intergenic region was aligned and number of nucleotides that were not present in every genome was calculated to give a mutation frequency.

**Competition Experiment in LB**

The fitness of ΔDARS1, ΔDARS2, ΔDARS1 ΔDARS2, and ΔdatA compared to the wild-type were investigated during direct competition in LB medium. The competing strains were inoculated pairwise at an approximate concentration of (10^7 CFU/mL) each. The populations were propagated by continuously transfers in LB medium. Samples from each population were taken at 10-generation intervals. Each sample was diluted in 0.9% NaCl and plated on LB plates with appropriate antibiotics. To distinguish the various E. coli strains, dilutions were plated on LB plates containing no antibiotic, kanamycin, or chloramphenicol. All plates were incubated for 18–24 h at 37°C prior to counting. When necessary to distinguish strains, 100 colonies from plates containing no antibiotic were toothpicked onto LB plates containing kanamycin or LB plates containing chloramphenicol.

**Mouse Colonization Experiments**

The specifics of the streptomycin-treated mouse model used to compare the large intestine colonizing abilities of E. coli strains in mice have been described previously (Leatham et al., 2005; Leatham-Jensen et al., 2012). Briefly, Six-to-eight-week-old, outbreed female CD-1 (Charles River Laboratories, Netherlands) mice were given drinking water containing streptomycin sulfate (5 g/l) for 24 h to eliminate resident facultative anaerobic bacteria (Miller and Bohnhoff, 1963). Mice were orally fed 100 µL of 20% (wt/vol) sucrose containing 10^6 CFU LB grown E. coli strains. The number of E. coli colonizing the mouse large intestine is reflected in the mouse feces, which is why fecal counts are used to estimate the various E. coli strains' ability to colonize the mouse intestine (Leatham-Jensen et al., 2012). After ingesting the bacterial suspension; feces was collected after 24 h, and as indicated. The mice were caged in groups of three mice, and cages were changed weekly. Mice were marked so they could be isolated and fecal pellets could be collected from each individual mouse. Mice were given fresh drinking water containing streptomycin sulfate (5 g/l) each day. Each fecal sample was homogenized in 1% Bacto tryptone (Difco Laboratories, NJ, USA), diluted in the same medium, and plated on MacConkey agar plates with appropriate antibiotics. When appropriate, 1 ml of a fecal homogenate (sampled after the feces had settled) was centrifuged at 12,000 X g, resuspended in 100 µL of 1% Bacto tryptone, and plated on a MacConkey agar plate with the appropriate antibiotics. This procedure increases the sensitivity of the assay from 10^2 CFU/gram of feces to 10 CFU/ per g of feces. To distinguish the various E. coli strains in feces, dilutions were plated on lactose MacConkey agar containing either streptomycin, streptomycin and kanamycin, or streptomycin and chloramphenicol. All plates were incubated for 18–24 h at 37°C prior to counting. When necessary to distinguish strains, 100 colonies from plates containing streptomycin were toothpicked onto MacConkey agar plates containing streptomycin and kanamycin or onto MacConkey agar plates containing streptomycin and chloramphenicol. Ethics approval statement; 2007/561-1430.

**β-Galactosidase Assays**

Cells were grown exponentially at 37°C in AB minimal medium supplemented with casamino acids, and β-galactosidase activities were measured as described by Miller (1972).

**Results**

**Conserved Relative Distance from oriC to datA, DARS1, DARS2, and dif in E. coli**

Bergthorsson and Ochman (1998) suggested that there is an evolutionary pressure on keeping the E. coli chromosome symmetric, so an approximately equal length of the two replication arms are maintained. The non-coding regions DARS1, DARS2, and datA are all indirectly involved in initiation of replication at oriC as they modulate the activity and for datA
also the amount of DnaA available for initiation in a dosage dependent manner. The replication-associated gene dosage of each region relative to oriC changes with growth rate and is given by the formula \( N_c/N_{oriC} = 2^{[(C - (1-x)) + D]/t} \) where \( x \) is the relative distance from oriC, \( C \) is the replication period, \( D \) is the time following termination of replication until cell division, and \( t \) is the doubling time (Bremer and Churchward, 1977). We therefore decided to investigate if there was any evolutionary pressure on their chromosomal position relative to oriC. The genome size of \( E. coli \) varies from 4.6 to 5.7 Mb (Leimbach et al., 2013). Thus, to compare chromosomal positions between genomes with up to 1 Mb difference we calculated a relative distance from oriC (see Materials and Methods) while using MG1655 as reference strain. The replication terminus is not as well defined as the origin of replication, this is why \( dif \) was chosen to represent \( terC \) (Hendrickson and Lawrence, 2007).

Although the study was limited to 59 “closed” \( E. coli \) genomes, the dataset includes a wide variety of different \( E. coli \) (see Table S3). Pathogenic \( E. coli \) strains are categorized into pathotypes (Kaper et al., 2004). The dataset includes four pathotypes, which are associated with diarrhea, namely shiga toxin-producing \( E. coli \) (STEC)/enterohemorrhagic \( E. coli \) (EHEC), enterotoxigenic \( E. coli \) (ETEC), enteropathogenic \( E. coli \) (EPEC), and enteroaggregative \( E. coli \) (EAEC). In addition to the intestinal pathogens two \( E. coli \) associated with the inflammatory bowel disease Crohn's disease were also included. In contrast to intestinal pathogenic \( E. coli \) (IPEC), which are obligate pathogens, extraintestinal pathogenic \( E. coli \) (ExPEC) are facultative pathogens which belong to the normal gut flora of a certain fraction of the healthy population where they live as commensals (Köhler and Dobrindt, 2011). The dataset contains ExPEC associated with neonatal-meningitis, asymptomatic bacteriuria, acute cystitis, the multidrug resistant ST131, as well as several uropathogenic \( E. coli \) (UPEC). In addition to human pathogenic \( E. coli \) strains, several \( E. coli \) strains isolated from the feces of healthy individuals (human commensals) are included. Apart from numerous common \( E. coli \) laboratory strains the dataset is concluded by three \( E. coli \) strains shown to be pathogenic in animals (avian pathogenic \( E. coli \) (APEC), and porcine enterotoxigenic \( E. coli \) (Pec) as well as an \( E. coli \) isolated from a toxic-metal contaminated site (for references see Table S3).

The chosen \( E. coli \) genome dataset had a median genome size of 5,095,204 bp, spanning from 3,976,195 bp (MDS42) to 5,697,240 bp (O26:H11 str. 11368). MDS42 is a “man-made” reduced \( E. coli \) K-12 genome derived from MG1655, which was constructed to identify non-essential genes (Pósfai et al., 2006). The smallest non-lab constructed \( E. coli \) chromosome was BW2952 with 4,578,159 bp. Due to the great diversity in both origin of isolation and genome size we believe that the dataset will be representative of \( E. coli \) as a whole.

Despite the large differences in genome size between \( E. coli \) strains, we found approximately the same relative distance from oriC to dif, datA, DARS1, and DARS2, respectively (Figure 2). This observation points to a conserved chromosomal organization. This organization is further conserved at the repliophile level as, DARS2 was always found on one repliophile, while datA and DARS1 were always found on the other repliophile. 

**E. coli Chromosome Symmetry**

The conserved position of DARS1, DARS2, datA, and dif relative to oriC in the tested 59 \( E. coli \) genomes, suggests that new DNA obtained by horizontal gene transfer has been equally distributed between the two replication arms, but also between the different regions on each of the replication arms. Strain O157:H7 EDL933 that has a genome size of 5.53 Mb, i.e., about 0.9 Mb bigger than the laboratory strain MG1655, exemplifies this (Figure 3). MG1655 and O157:H7 EDL933 share a common 4.1 Mb backbone, which is co-linear except for one 422-kilobase inversion spanning the replication terminus (Perna et al., 2001). The differences between the two genomes are reflected in K-islands (0.53 Mb), which is the DNA present only in MG1655 and O-islands (1.34 Mb), which is unique to O157:H7 EDL933 (Perna et al., 2001). When a circular genome map of O157:H7 EDL933 is compared to MG1655 the 1.34 Mb DNA unique to O157:H7 EDL933 is not only distributed between the two replication arms but as expected also among the cis-acting regions for regulation of initiation of replication (gray boxes; see Figure 3).

The \( E. coli \) strain MDS42 (Pósfai et al., 2006) contains a 14.0% reduced genome relative to it's the parental MG1655. However, it maintained a similar relative distance from oriC to DARS1, DARS2, datA, and dif as the parental strain (see Table S3), i.e., the non-essential DNA lost from MG1655 was distributed between the different non-coding cis-acting regions.
Conservation of oriC, DARS1, DARS2 and datA regions

Only a few genomes showed 100% sequence identity of the oriC, DARS1, DARS2, and datA-regions to those of MG1655. The comparison between the nucleotide sequences from the 59 different E. coli genomes is found in the Supplementary Material (Supplementary Figures S12–S15).

In order to estimate the mutation pressure on oriC, datA, DARS1, and DARS2 we calculated the mutation frequency for intergenic regions in E. coli (see Materials and Methods). It was found to be 6 mutations per 100 ± 3 bp. Neither of the oriC, datA, DARS1, or DARS2 regions differed significantly from this average frequency (Not shown). However, the vast majority of changes observed were found in spacer regions whereas binding sites for key proteins were conserved among all genomes (Table 1) which underlines their important role for cell cycle control.

oriC

In oriC both of the AT-rich 13-mer regions L and M were identical among the 59 strains, whereas the R 13-mer varied in the two outer positions (Supplementary Figure S1). Three of the six 6-mer sites present in the AT-rich region (Supplementary Figures S3, S4, S6) that specifies binding of DnaAATP when in the single stranded configuration (Figure 1) were identical to the same regions in MG1655. However, Supplementary Figures S1, S2, S5 carries single nucleotide changes relative to MG1655 in a subset of strains. In five strains a nucleotide alteration was found in the single stranded DnaA-ATP box 2 that demolished a GATC-site, which is the substrate for Dam methyltransferase (Supplementary Figure S2).

The majority of DnaA binding sites in oriC (R1, R2, R5, τ1, τ2, I3, C1, and C2) were completely conserved, whereas R3, C3, I1, I2, and R4 binding sites carried differences to the corresponding MG1655 sequences in some strains (Supplementary Figure S12). Controversy about which of DnaA Box R3 and DnaA Box C3 are functional during unwinding of oriC exists (Kaur et al., 2014). The DnaA Box R3 overlaps with both DnaA Box C2 and C3 (Figure 1; Supplementary Figure S12). Nonetheless, all three DnaA boxes are included in the present study.

The consensus sequence for the R-box is TTWTNCACA (W is dA or dT and N is any nucleotide) (Schaper and Messer, 1995). In MG1655 DnaA Box R1 differs from the R-box consensus sequence by three nucleotides, while DnaA Box C3 and I2 differ by four nucleotides (Grimwade et al., 2000; Ryan et al., 2002; Rozgaja et al., 2011). We only identified sequence alterations that resulted in an altered identity to the R-box consensus sequences for DnaA binding sites, I2, R3, and R4. The I2 binding site from strain ED1a was found to fit better to the R-box consensus sequence compared to the DnaA box I2 sequence from MG1655 (Supplementary Figure S4). The MG1655 DnaA Box R3 differs from the R-box consensus sequence by one nucleotide, while the 10 E. coli strains deviating from the MG1655 DnaA Box R3 sequence deviate from the R-box consensus sequence by two nucleotides (Supplementary Figure S5). Strain O127:H16 E2348/69 deviates from the R-box consensus sequence by a nucleotide in DnaA Box R4 (Supplementary Figure S6). The change from a dT to a dC in nucleotide position number 4 diminishes the identity to the R-box consensus sequence.

The IHF binding site consensus sequence is WATCAANNNTTR [W is dA or dT, R is dA or dG, and N is any nucleotide (Hales et al., 1994)]. Three strains were found to differ with respect to the oriC IBS (Supplementary Figure S7).

**Table 1** | Sequence deviations in conserved regions of oriC, DARS1, DARS2 and datA.

<table>
<thead>
<tr>
<th>Region</th>
<th>Identity</th>
<th>Functional sitea</th>
<th>No. of sequences deviating from MG1655b</th>
<th>See Figures</th>
</tr>
</thead>
<tbody>
<tr>
<td>oriC</td>
<td>Conserved</td>
<td>13-mer L</td>
<td>6/59 S1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13-mer M</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box R1</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box R2</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box R5</td>
<td>5/59 S9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box R1 τ1</td>
<td>1/59 S3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box τ2</td>
<td>1/59 S3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box I3</td>
<td>1/59 S4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box C1</td>
<td>10/59 S5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box C2</td>
<td>1/59 S6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SSDA Box 3</td>
<td>3/59 S7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SSDA Box 4</td>
<td>3/59 S7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SSDA Box 6</td>
<td>3/59 S7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Not conserved</td>
<td>13-mer R</td>
<td>6/59 S1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SSDA Box 1</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SSDA Box 2</td>
<td>5/59 S9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SSDA Box 5</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box I1</td>
<td>1/59 S4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box I2</td>
<td>1/59 S4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box C3/R3</td>
<td>10/59 S5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box R4</td>
<td>1/59 S6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IBS</td>
<td>3/59 S7</td>
<td></td>
</tr>
<tr>
<td>DARS1</td>
<td>Conserved</td>
<td>DnaA Box 1</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box 2</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box 3</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td>DARS2</td>
<td>Conserved</td>
<td>DnaA Box 1</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box 2</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box 3</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box 4</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box 5</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Not conserved</td>
<td>FBS-2</td>
<td>3/59 S8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FBS-3</td>
<td>9/59 S9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IBS-2</td>
<td>4/59 S10</td>
<td></td>
</tr>
<tr>
<td>datA</td>
<td>Conserved</td>
<td>DnaA Box 1</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box 2</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box 3</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box 4</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DnaA Box 5</td>
<td>1/59 S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Not conserved</td>
<td>IBS</td>
<td>12/59 S11</td>
<td></td>
</tr>
</tbody>
</table>

aFBS, Fis Binding Site; IBS, IHF Binding Site; SSDA, Single stranded binding site for DnaAATP.
bNumber of isolates within the dataset which deviates from the region in MG1655.
O26:H11 strain 11368 was found to have a diminished identity to the IHF consensus sequence compared to MG1655, while both O145:H28 strain RM12761 and O145:H28 strain RM13516 was found to have a better fit. The Fis binding site in oriC was found completely conserved in all strains (Supplementary Figure S12).

The nucleotide distances between the protein binding regions were highly conserved between strains. Only O26:H11 strain 11368 and O111:H- strain 11228 lacked a nucleotide in the spacer region between the AT-rich 13-mer termed R and the DnaA Box R1 in oriC compared to MG1655 (Supplementary Figure S12).

Based on this analysis it is hard to deduce a hierarchy of the importance of the different DnaA binding sites in oriC. Ten strains had changes in the R3/C3 boxes. Whereas the alterations resulted in a R3 box with poorer resemblance to the R-box consensus, this was not the case for C3. Therefore, it is likely that C3 represents the functional DnaA binding site in the replication origins.

**DARS1, DARS2, and datA**

All DnaA binding sites in datA, DARS1, and DARS2 were completely conserved between the strains analyzed (Table 1) (Supplementary Figures S13–S15). FBS-2 of DARS2 differed from that of MG1655 in three strains. However, since Fis has the consensus sequences GNHYANNNNNNTRNNC (Y is dC or dT, R is dA or dG, and N is any nucleotide) (Finkel and Johnson, 1992) none of the observed differences resulted in a reduced similarity to the Fis consensus sequence (Supplementary Figure S8). For FBS-3 of DARS2, four _E. coli_ strains differed from MG1655 and had a reduced identity the Fis consensus sequence (Supplementary Figure S9).

The IHF binding site IBS-1 of DARS2 was identical in all 59 genomes. The sequence variations of IBS-2 of DARS2 (Supplementary Figure S10) or the IBS in datA (Supplementary Figure S11), relative to MG1655 did not change the identity to the IHF consensus sequences. The datA region of strains HS and O103:H2 str. 12009 DNA lacked four nucleotides between DnaA Box 1 and DnaA Box 2, while strain 536 lacked a nucleotide between the IBS and DnaA Box 3.

Altogether, these observations suggest that there is a strong selection pressure on maintaining the sequence and spacing of protein binding sites, and thereby functionality of oriC, DARS1, DARS2, and datA. The majority of the nucleotide differences observed was located in the non-functional spacer regions between the different protein binding sites. The majority of differences found within protein binding sites, did not reduce the identity to the investigated consensus sequence.

**Importance of datA, DARS1, and DARS2 for Cell Cycle Control**

Despite of the conservation of the chromosomal positions of DARS1, DARS2, and datA relative to oriC, neither is essential (Kitagawa et al., 1998; Fujimitsu et al., 2009). Cells with and without datA were also previously found to have similar doubling times (Kitagawa et al., 1998). We created cells with deletions of datA, DARS1, and DARS2 individually and in various combinations. These cells were viable no matter which combination of DARS1, DARS2, and datA we deleted. The cellular doubling time was not affected by individual deletions but increased when combinations of DARS1, DARS2, and datA were deleted (Table 2). Cells carrying ΔDARS1 ΔDARS2 and ΔDARS1 ΔDARS2 ΔdatA were found to have the longest doubling time in minimal medium supplemented with glucose and casamino acids, while ΔDARS1 ΔdatA and ΔDARS2 ΔdatA cells were found to have the longest doubling time in the same medium without casamino acids (Table 2).

We used lacZ fusions of _dnaA_ and _mioC_ promoters to assess the effect of _datA_, _DARS1_, and _DARS2_ loss on the cellular DnaA<sub>ATP</sub>/DnaA<sub>ADP</sub> ratio. The _dnaA_ gene is transcribed from two upstream promoters, termed _dnaA1p_ and _dnaA2p_. Four DnaA boxes are located between the two promoters, with only one of them containing the stringent consensus sequence (Hansen et al., 1982, 2007; Armengod et al., 1988). Both _dnaA_ promoters are negatively regulated by the DnaA protein (Hansen et al., 2007), with DnaA<sub>ATP</sub> being most efficient in repressing _dnaA_ expression (Speck et al., 1999). DnaA<sub>ATP</sub> also represses the _mioC_ promoter located upstream of oriC prior to initiation by binding to five DnaA boxes located within and/or close to the promoter. Of the five DnaA boxes only one contains the stringent consensus sequence (Figure 1) (Ogawa and Okazaki, 1994; Bogan and Helmstetter, 1997; Hansen et al., 2007). Loss of _datA_ resulted in a slight repression of _dnaA_ (Table 3). This is in agreement with an increase in the DnaA<sub>ATP</sub>/DnaA<sub>ADP</sub> ratio.

<table>
<thead>
<tr>
<th>Strain/Genotype&lt;sup&gt;a&lt;/sup&gt;</th>
<th>ABTG + CAA&lt;sup&gt;b&lt;/sup&gt; (Min)</th>
<th>Origin/&lt;Cell&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Cell mass&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Origins/&lt;Mass&lt;sup&gt;e&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild-type</td>
<td>40</td>
<td>5.2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ΔdatA</td>
<td>41</td>
<td>7.9</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>ΔDARS1</td>
<td>40</td>
<td>4.7</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>ΔDARS2</td>
<td>41</td>
<td>4.4</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>ΔDARS1 ΔDARS2</td>
<td>45</td>
<td>3.1</td>
<td>1.1</td>
<td>0.6</td>
</tr>
<tr>
<td>ΔDARS1 ΔdatA</td>
<td>41</td>
<td>7.2</td>
<td>1</td>
<td>1.4</td>
</tr>
<tr>
<td>ΔDARS2 ΔdatA</td>
<td>42</td>
<td>6.1</td>
<td>1.1</td>
<td>1.2</td>
</tr>
<tr>
<td>ΔDARS1 ΔDARS2 ΔdatA</td>
<td>49</td>
<td>4</td>
<td>1</td>
<td>0.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ABTG&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild-type</td>
</tr>
<tr>
<td>ΔdatA</td>
</tr>
<tr>
<td>ΔDARS1</td>
</tr>
<tr>
<td>ΔDARS2</td>
</tr>
<tr>
<td>ΔDARS1 ΔDARS2</td>
</tr>
<tr>
<td>ΔDARS1 ΔdatA</td>
</tr>
<tr>
<td>ΔDARS2 ΔdatA</td>
</tr>
<tr>
<td>ΔDARS1 ΔDARS2 ΔdatA</td>
</tr>
</tbody>
</table>

<sup>a</sup>Wild-type is MG1655.<br><sup>b</sup>Doubling time in minimal medium supplemented with glucose and casamino acids or minimal medium supplemented with glucose. Each strain has a standard deviation of ±2 min for growth in ABTG + CAA and ±3 min for growth in ABTG.<br><sup>c</sup>Determined from flow cytometric analysis.<br><sup>d</sup>Determined as average light scatter from flow cytometric analysis. Numbers are normalized to 1 for wild-type.<br><sup>e</sup>Average fluorescence/average light scatter. Numbers are normalized to 1 for wild-type.
and the dnaA promoter being repressed by DnaA\textsuperscript{ATP} (Kitagawa et al., 1998; Speck et al., 1999; Kasho and Katayama, 2013). Loss of DARS1 led to an increased expression of the mioC gene while loss of DARS2 led to an increased expression of both the dnaA and mioC genes (Table 3), since both promoters are subject to negative transcriptional control by DnaA\textsuperscript{ATP} (Speck et al., 1999; Hansen et al., 2007). This agrees with DARS1 and DARS2 being instrumental in increasing the cellular DnaA\textsuperscript{ATP} level, and that DARS2 is more efficient than DARS1 (Fujimitsu et al., 2009).

We proceeded to analyze the cell cycle characteristics by flow cytometry (Table 2; Figure 4). Wild-type cells exhibited the expected synchronous initiation pattern with the majority of cells containing 2, 4, or 8 replication origins (Figure 4A) (Skarstad et al., 1986). datA deficient cells had an increased origin concentration (origins/mass) (Kitagawa et al., 1998), which resulted both from an increase number of origins per cell and a decreased cell mass (during slow growth only) (Table 2). A high degree of initiation asynchrony was observed for ΔdatA cells (Figure 4E) (Kitagawa et al., 1998). Cells deficient in DARS1, DARS2 or both regions had a reduced origin concentration relative to wild-type cells (Table 2; Figures 4B–D) (Fujimitsu et al., 2009). Compared to wild-type, all DARS mutant cells had an increased cell mass during slow growth whereas only the ΔDARS1 ΔDARS2 double mutant had increased cell mass during fast growth. Asynchrony of initiation was observed for ΔDARS2 (Figure 4C) and ΔDARS1 ΔDARS2 (Figure 4D) cells, but not for cells carrying the ΔDARS1 mutation alone (Figure 4B) (Fujimitsu et al., 2009).

Because the datA region promotes inactivation of DnaA\textsuperscript{ATP} to DnaA\textsuperscript{ADP}, and the DARS regions promote the opposite, i.e., DnaA reactivation, we decided to see whether loss of DARS1, DARS2 or both could suppress the initiation defect of datA cells. Deletion of DARS1 in ΔdatA cells only marginally lowered the origin per mass (from 1.5 to 1.4; Table 2). A similar but larger effect was observed when ΔDARS2 was deleted suggesting that ΔDARS2 was more efficient than ΔDARS1 for DnaA rejuvenation. Deleting both DARS1 and DARS2 in ΔdatA cells lowered the origin concentration below wild-type level (Table 2) and also partly restored initiation synchrony (Figure 4H). Overall, these experiments show that loss of rejuvenation activity overcompensates for loss of DDAH. This may be explained by the RIDA process, which being active in the triple mutant so that DnaA\textsuperscript{ATP} to DnaA\textsuperscript{ADP} conversion is still ongoing.

DARS1 and DARS2 are Required for Mouse Colonization

In order to examine the fitness cost of losing DARS or datA activity we performed two different competition experiments: continued growth in LB medium and during mouse colonization, where the streptomycin-treated mouse was chosen as the in vivo model. For both competition experiments strains were introduced pairwise at approximately equal numbers (Figure 5). If they have the same fitness they would also be recovered in equal numbers.

During growth in LB the wild-type was more fit than the cells deficient in either DARS1, DARS2, both DARS1 and DARS2, or datA (Figure 5). The biggest fitness cost resulted from loss of both DARS1 and DARS2 (Figure 5C) followed by loss of DARS2 (Figure 5B) loss of DARS1 (Figure 5A) which was similar to loss of datA (Figure 5D).

The same order of fitness was not observed when evaluated in mice. Following colonization, the number of wild-type E. coli increased for about 3 days until stabilizing around 10^9 cfu

**Table 3** Expression of the dnaA and mioC genes.

<table>
<thead>
<tr>
<th>Strain</th>
<th>dnaA expression (SD)</th>
<th>mioC expression (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild-type</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>ΔDARS1</td>
<td>ND</td>
<td>108 (± 5)</td>
</tr>
<tr>
<td>ΔDARS2</td>
<td>116 (± 1)</td>
<td>122 (± 4)</td>
</tr>
<tr>
<td>ΔdatA</td>
<td>95 (± 3)</td>
<td>ND</td>
</tr>
</tbody>
</table>

\*Measured in MG1655 using the dnaA-lacZ translational fusion carried on λR1 (Brown et al., 1985) in strain MG1655 lacZ::Tn5. Numbers are given relative to wild-type expression of 100% corresponding to 46 Miller units. ND, Not determined; SD, Standard deviation.

\*Measured in MG1655 lacZ::Tn5 using the mioC-lacZ transcriptional fusion carried on plasmid pALO75 (Lister-Olesen et al., 1987). Numbers are given relative to wild-type expression of 100% corresponding to 302 Miller units. ND, Not determined; SD, Standard deviation.

---

**Figure 4** Cellular origin distribution. Prior to flow cytometric analysis exponentially growing wild-type and mutant cells was treated with rifampicin and cephalaxin. Cells were grown in AB minimal medium supplemented with 0.2% glucose, 10 μg/ml thiamine, and 0.5% casamino acids at 37°C. Wild-type is MG1655 (A); relevant mutations are indicated in individual panels (B–H).
per gram of mouse feces (Figures 5E–H). Cells deficient in DARS1 increased in number to peak at about $10^7$ cfu/gram feces at day 3 followed by a rapid decline in number over the next days to end around $10^2$ cfu/gram feces at day 14 (Figure 5E) suggesting that these cells were rapidly out-competed by wild-type cells. Cells deficient in both DARS1 and DARS2 (Figure 5G) were outcompeted at a slightly faster rate than cells deficient in only DARS1, suggesting that DARS2 plays a
minor role to DARS1 in fitness during mouse colonization. In agreement with this, DARS2 mutant cells were able to coexist in the mouse along with wild-type cells albeit at a lower number (Figure 5F). Loss of datA was similar to the loss of DARS2. Following co-infection in mice both wild-type and datA cells increased in numbers to level at $10^9$ and $10^7$ cfu per gram feces, respectively, and remained at these levels for the duration of the experiment (Figure 5H). Therefore, datA and DARS2 deficient cells were poor at establishing colonization relative to the wild-type, but once established cells were not outcompeted with time.

On day 14 post-feeding, wild-type and a ΔDARS1, ΔDARS2, ΔDARS1ΔDARS2, or ΔdatA cells (depending on the competition experiment) were isolated from the feces of each mouse for further study. The origin per mass and asynchrony index score were determined for each strain isolated post-infection and found to be similar to the initial strains fed to each mouse [data not shown] showing that secondary mutations were not likely to have been selected during growth in the mouse.

Overall these experiments indicate that different factors determine fitness of cells dependent on growth conditions. During continued growth in LB medium, both promotion and prevention of DnaA<sub>ATP</sub> to DnaA<sub>ADP</sub> conversion resulted in a fitness cost. On the other hand, overinitiation resulting from DnaA<sub>ATP</sub> accumulation in datA cells did not seem to affect mouse colonization to the same extent as loss of rejuvenation ability, especially promoted by DARS1.

**Discussion**

In this study we found conservation in distances from oriC to the non-coding regions DARS1, DARS2, datA, and dif in *E. coli*. DARS1 and datA were always found on the same replichore, while DARS2 were found on the other replichore. The oriC, DARS1, DARS2, and datA regions were found to be structurally similar among the tested *E. coli*, with most of the sequence differences found to be in the non-functional spacer regions between key protein binding sites. Cells deficient in DARS1, DARS2, or datA were viable and had doubling times similar to wild-type. However, replication initiation was perturbed. Cells deficient in datA were found to initiate asynchronously, and this could not be counteracted by further deletions of either DARS1 or DARS2. Cells deficient in DARS1, DARS2, DARS1, and DARS2, or datA were found to be less fit than the wild-type in both LB medium and during mouse colonization.

**Conservation of oriC**

In the chromosomal context initiation of replication can be initiated from a mutant oriC without DnaA box R2, R3, R4, or R5, the IBS or the FBS (Weigel et al., 2001), as well as DnaA boxes I1, I2, or I3 (Riber et al., 2009). It is also possible to invert the direction of R4, add 14 bp between DnaA Box R3 and DnaA Box R4, or delete the right half of oriC (from position 275 to 352) (Weigel et al., 2001). Although DnaA Box R1 was originally found to be essential (Weigel et al., 2001), a more efficient recombining technique demonstrated that DnaA Box R1 is also dispensable (Kaur et al., 2014). Surprisingly, only deletion of DnaA Box R3, R4, and the right half of oriC (from position 275 to 352) was reported to result in slow growth relative to wild-type cells (Weigel et al., 2001). Asynchrony, a sensitive measure for perturbations of the initiation process, was observed with the deletion of the IBS, DnaA Box R2, R4, R5, extending the spacer region between R3 and R4 (Weigel et al., 2001), and the deletion of DnaA Box R1 (Kaur et al., 2014). These studies demonstrate that initiation from oriC is very robust and that only major changes in the origin results in loss of function altogether. On the other hand mutant origins fail to compete with their wild-type counterparts as shown by the inability to establish minichromosomes carrying oriC mutations in cells with a wild-type chromosomal copy of oriC (Weigel et al., 2001). It is conceivable that a similar competition takes place between cells in a population and that even small changes in important regions of oriC, that does not affect viability, may result in replication perturbation, loss of fitness and inability to co-exist with wild-type cells and that this explain the high degree of oriC conservation observed.

**Chromosomal Position of datA, DARS1, and DARS2**

The relative chromosomal locations of the *datA*, DARS1, and DARS2 regions are conserved among *E. coli* strains. This is somewhat surprising as *E. coli* genomes are highly fluidic, i.e., they frequently mutate, change size, and rearrange. The frequency of genome rearrangement, measured between *rrn* sites, is about $10^5$–$10^4$ changes/(generation, genome) (Hill and Gray, 1988). For example, the 1.34 Mb DNA unique to O157:H7 EDL933, is inserted compared to MG1655 in such a way that the chromosomal location of *datA*, DARS1, and DARS2 relative to oriC, remain unchanged. Similarly, the non-essential DNA that was removed from MG1655 to create MSD42 (Pósfai et al., 2006), was also dispersed between regions so that MSD42 has the same relative chromosomal location of *datA*, DARS1, and DARS2. There may be at least three reasons for the conserved location of the three regions. First, chromosome asymmetry, i.e., different lengths of the two replication arms leads to slow growth (Hill and Gray, 1988). Second, chromosomal rearrangements resulting in a mixture of different macromdomains have deleterious effects of cell growth (Esnault et al., 2007) which may explain why *datA*, DARS1 and DARS2 regions located within the Ori MD, the Right MD and NS<sub>4</sub>EF, respectively (Valens et al., 2004), are always found on the same replichore. Third, the correct chromosomal location of DARS1, DARS2, and *datA* may be important for proper function and cell cycle progression (see below). As the activity of these regions in modulating DnaA binding to ATP or ADP is dependent on their copy number the proper distance to oriC becomes important for function. The relative copy number of DARS1, DARS2, and *datA* (replication-associated gene dosage) decreases with distance from oriC (Bremer and Churchward, 1977; Couturier and Rocha, 2006). *datA* was always found close to oriC. It is therefore conceivable that, *datA* is duplicated while oriC is still sequestered (Kitagawa et al., 1998; Kasho and Katayama, 2013), in all strains. The *datA*
site promotes DnaA\textsuperscript{ATP} to DnaA\textsuperscript{ADP} conversion, to prevent re-initiation when the concentration of DnaA\textsuperscript{ATP} is high, i.e., just after sequestration ends and this may provide the evolutionary pressure that has resulted in a conserved chromosomal location. In agreement with this relocation of datA to a chromosomal position close to terC resulted in high asynchrony in initiation of replication while other positions closer to oriC resulted in a near wild-type phenotype (Kitagawa et al., 1998). It is likely that the chromosomal positions of DARS1 and DARS2 are also important for cell cycle control as they serve to re-activate the DnaA initiator protein in time for the next initiation. Relocation of DARS sequences has not been experimentally pursued. It is however tempting to speculate that the genomic arrangement of DARS1 and DARS2 will ensure that rejuvenation of DnaA\textsuperscript{ATP} to DnaA\textsuperscript{ADP} will be accelerated during later stages of the replication cycle and following duplication of these regions. This rejuvenation is important for increasing the DnaA\textsuperscript{ATP} level for the following round of initiations.

**The datA, DARS1, and DARS2 Regions Are Important for Fitness**

In the DARS1 and DARS2 regions, all DnaA binding boxes as well as spacer distances were conserved among *E. coli* species. Especially the DARS1 region had a very low mutation frequency which is consistent with reports that all three DnaA boxes along with the region flanking the last DnaA Box (42 bp spanning from base number 198–239) is required for full ADP-releasing activity of *in vitro* (Fujimitsu et al., 2009). Similarly DnaA Box 1 and DnaA Box 2 are crucial for ADP-releasing activity of DARS2, while DnaA Box 3 is required for full ADP-releasing activity *in vitro* (Fujimitsu et al., 2009). IBS1–2 and FBS2–3 are required for DnaA\textsuperscript{ATP} regeneration *in vivo* (Kasho et al., 2014). In four *E. coli* isolates, FBS-3 has a weaker sequence identity to the Fis consensus sequence than in MG1655. Fis-binding sites are difficult to define due to the lack of an obvious consensus sequence (Finkel and Johnson, 1992), and the effect of the observed weaker identity in FBS-3 is hard to interpret.

**DARS2** was previously reported to be more efficient in rejuvenation of DnaA\textsuperscript{ADP} to DnaA\textsuperscript{ATP} than DARS1 (Fujimitsu et al., 2009) and both the fitness experiment performed in LB medium (compare Figure 5A and Figure 5B) and the dnaA and mioC expression studies agreed with this. The situation was reversed in the mouse model where loss of DARS1 was associated with the biggest fitness cost (compare Figure 5E and Figure 5F). DARS2 is activated by the binding of both IHF and Fis, whereas no protein factors are required for the function of DARS1 (Fujimitsu et al., 2009; Kasho et al., 2014). While IHF is abundant in the cell during every growth phase, although the concentration is highest in the stationary phase (Azam and Ishihama, 1999), the concentration of Fis is dependent on the growth phase; i.e., it is highly abundant (10,000–50,000 molecules/cell) in early exponential phase, but decrease to <100 molecules/cell from late exponential phase to stationary phase. The level of Fis also varies during steady state growth; i.e., it is low during slow growth and high during fast growth (Nilsson et al., 1992; Flätten and Skarstad, 2013). Limited data are available on the growth of *E. coli* in mouse intestines but overall slow growth was reported with doubling times between 80 and 125 min (Rang et al., 1999). It also seems reasonable that cells under these conditions never reaches exponential growth but grows whenever food becomes available, i.e., with relative short growth phases and frequent entries into stationary phase. Therefore, the bacterial Fis level during intestinal colonization may be significantly lower than during fast exponential growth in rich medium. The relative contribution of DARS2 to DnaA rejuvenation may therefore be low in the mouse, and explain the bigger fitness cost associated with loss of DARS1 under these conditions. In agreement with this we only observed a minor further fitness cost associated with deletion of DARS2 in DARS1 deficient cells. Such ΔDARS1 ΔDARS2 cells rely on *de novo* synthesis of DnaA or the speculated DARS3 to produce DnaA\textsuperscript{ATP} during colonization of a mouse (Kasho et al., 2014). The fitness cost associated with loss of DARS1 or DARS2 may readily explain why mutations are rarely observed in these regions, but not the conserved distance to oriC. This needs to be further elucidated by relocation to other chromosomal positions.

The activity of datA is absolutely dependent on DnaA Boxes 2 and 3 along with the IBS (Nozaki et al., 2009; Kasho and Katayama, 2013). Also the spacing between DnaA Box 2 and IBS, as well as the spacing between the IBS and DnaA Box 3 has been shown to be important for datA function (Nozaki et al., 2009; Kasho and Katayama, 2013). In accordance with this we found all DnaA boxes as well as the identities to the IHF consensus sequence were conserved. Changes were only observed in the length of the spacer region between DnaA Box 1 and DnaA Box 2 in strain HS and O103:H2 str. 12009 and between the IBS and DnaA Box 3 in strain 536. The effect of the altered spacing is hard to interpret although the latter may lead to a lower efficiency in converting DnaA\textsuperscript{ATP} to DnaA\textsuperscript{ADP} compared to MG1665 (Nozaki et al., 2009). Cells deficient in datA only had a 5% decrease in dnaA expression correlating with previous reports showing that a datA deletion slightly (i.e., 5–10%) increased the DnaA\textsuperscript{ATP} level (Katayama et al., 2001). Loss of datA accompanied overinitiation only resulted in modest fitness cost during fast growth in LB and during colonization. The DDH and RIDA (Hda dependent) pathways both contribute to convert DnaA\textsuperscript{ATP} to DnaA\textsuperscript{ADP} in *E. coli*, but where loss of RIDA is associated with severe overinitiation and inviability unless second site suppressor mutations arise (Riber et al., 2006); loss of DDH is tolerated. Therefore, DDH plays a minor role to RIDA and this may explain the limited fitness cost of datA cells. The limited fitness cost of losing datA relates poorly to the high degree of conservation observed between species. We do not have a good explanation for this observation but it may relate to the DDH process being important during growth conditions other than those employed by us.

Of interest the Gram-positive bacteria *Bacillus subtilis* and *Streptomyces coelicolor* contain DnaA box clusters close to oriC that can repress untimely initiation (Smulczyk-Krawczyszyn et al., 2006; Okumura et al., 2012), i.e., a function similar to that of datA in *E. coli*. In addition, several *E. coli* related bacterial species contains DARS1-like sequence and DARS2-like sequences in a genomic position similar to that of *E. coli* (Fujimitsu et al., 2009; Kasho et al., 2014). These observations indicates that both datA and DARSs mechanism, and genomic positions, maybe common
to many bacterial species whose genomes contain DNA box clusters.

**Author Contributions**

JF and AL planned the experiments. JF performed the experiments. JF, GC, KK, and AL analyzed data. JF and AL wrote the manuscript.

**Funding**

This work was supported by grant PIRG05-GA-2009-247241 from the European Union, by grant 09-064250/FNU from the Danish Research Council for Natural sciences, by grant 09-067075 from the Danish Strategic Research Council and by grants from the Lundbeck Foundation and the Novo Nordisk Foundation.

**Acknowledgments**

We thank Shiraz Shah for help with the bioinformatics analysis and Michaela Lederer for technical assistance.

**Supplementary Material**

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.01011

**References**


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

Copyright © 2015 Frimodt-Møller, Charbon, Krogfelt and Løbner-Olesen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.