RESEARCH ARTICLE

Complete Genome Sequence of ER2796, a DNA Methyltransferase-Deficient Strain of *Escherichia coli* K-12

Brian P. Anton[1], Emmanuel F. Mongodin[2], Sonia Agrawal[2], Alexey Fomenkov[1], Devon R. Byrd[1¤], Richard J. Roberts[1], Elisabeth A. Raleigh[1]*

1 New England Biolabs, Inc., Ipswich, Massachusetts, United States of America, **2** Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, United States of America

¤ Current address: Devon R. Byrd, O'Gara Group Training and Services, Montross, Virginia, United States of America.
* raleigh@neb.com.

Citation: Anton BP, Mongodin EF, Agrawal S, Fomenkov A, Byrd DR, Roberts RJ, et al. (2015) Complete Genome Sequence of ER2796, a DNA Methyltransferase-Deficient Strain of *Escherichia coli* K-12. PLoS ONE 10(5): e0127446. doi:10.1371/journal.pone.0127446

Academic Editor: John R Battista, Louisiana State University and A & M College, UNITED STATES

Received: December 19, 2014

Accepted: April 14, 2015

Published: May 26, 2015

Data Availability Statement: The genome sequences of ER2796 and ER3413 are available from GenBank with the accession numbers CP009644 and CP009789, respectively. Both strains are available from New England Biolabs and the Coli Genetic Stock Center (CGSC).

Funding: The funder provided support in the form of salaries for authors BA, EAR, DB, AF, and RJR, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of

Abstract

We report the complete sequence of ER2796, a laboratory strain of *Escherichia coli* K-12 that is completely defective in DNA methylation. Because of its lack of any native methylation, it is extremely useful as a host into which heterologous DNA methyltransferase genes can be cloned and the recognition sequences of their products deduced by Pacific Biosciences Single-Molecule Real Time (SMRT) sequencing. The genome was itself sequenced from a long-insert library using the SMRT platform, resulting in a single closed contig devoid of methylated bases. Comparison with K-12 MG1655, the first *E. coli* K-12 strain to be sequenced, shows an essentially co-linear relationship with no major rearrangements despite many generations of laboratory manipulation. The comparison revealed a total of 41 insertions and deletions, and 228 single base pair substitutions. In addition, the long-read approach facilitated the surprising discovery of four gene conversion events, three involving rRNA operons and one between two cryptic prophages. Such events thus contribute both to genomic homogenization and to bacteriophage diversification. As one of relatively few laboratory strains of *E. coli* to be sequenced, the genome also reveals the sequence changes underlying a number of classical mutant alleles including those affecting the various native DNA methylation systems.

Introduction

The Gram-negative bacterium *Escherichia coli* has been foundational to our understanding of bacterial genetics since 1946, when Lederberg and Tatum first demonstrated bacterial conjugation [1]. Between that time and the start of the genome sequence era one half century later, a wealth of *E. coli* genotypic and phenotypic information was generated through laboratory manipulation of strains. Currently, there are finished genome sequences available for more than 75 *E. coli* strains, but the vast majority of these are wild type isolates, both pathogenic and

these authors are articulated in the 'author contributions' section.

Competing Interests: Authors BA, EAR, DB, AF, and RJR are employed by New England Biolabs, a company that supplies competent E. coli cells as well as reagents that may be used in the construction of sequencing libraries. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

commensal. The few laboratory strains that have been completely sequenced include nine strains derived from K-12 (Table 1 and Fig 1). Two of these strains (MG1655 and W3110) resulted from only minimal genetic manipulation of the wild type K-12 isolate [2, 3] and are highly similar to one another [3]. The other seven strains have been more extensively manipulated, and provide some insight into the nature of both classical mutant alleles and previously unidentified lineage-specific mutations [4, 5].

Some important genetic markers not present in any of the sequenced strains above relate to DNA methylation. *E. coli* K-12 encodes four DNA methyltransferases (MTases), one of which forms part of a Type I restriction-modification (R-M) system and three of which are solitary (Table 2). Methylation of GATC sites by M.EcoKDam has been well studied and is known to serve several important functions including directing mismatch repair to the nascent strand, regulating the timing of chromosome replication, and controlling the expression of certain genes (reviewed in [6]). Methylation of CCWGG by M.EcoKDcm is less well understood, but the *dcm* gene partially overlaps *vsr*, which encodes the very short patch (VSP) repair endonuclease (ENase), suggesting the two gene products function in a common process. Indeed, the VSP repair system fixes T:G mismatches that arise from the deamination of 5-methylcytosine (m^5C), the product of Dcm methylation [7]. The biological function of Dcm methylation remains unclear, but recent evidence suggests it plays a role in controlling the expression of certain genes in stationary phase [8]. In particular, Dcm methylation appears to downregulate the expression of ribosomal genes [9], possibly by regulating *rpoS* expression [8]. The MTase M. EcoKII, encoded by the gene *yhdJ*, has been shown to methylate the site ATGCAT at the second A residue when overexpressed from a plasmid copy, conferring protection from the restriction enzyme NsiI [10]. However, this activity has yet to be observed in wild type cells, suggesting the gene is silent under all growth conditions tested to date. Its biological function remains unknown despite its wide conservation in almost all sequenced *E. coli* genomes. The MTase encoded by *hsdM* forms part of the Type I R-M system EcoKI, and methylates both strands of the asymmetric sequence AACNNNNNNGTGC. In Type I R-M systems, specific DNA recognition by the MTase (M) is not intrinsic, but rather is conferred by a specificity subunit (S), with the active methylation complex having the stoichiometry M_2S [11].

Table 1. Laboratory strains of *E. coli* with finished (ungapped) genome sequences in GenBank.

Strain	Ancestor	RefSeq or *INSDC* Accession (chromosome)	Reference
MG1655	K-12	NC_000913	[2, 64]
W3110	K-12	NC_007779	[3]
DH1	K-12	NC_017625, NC_017638	[65]
DH10B	K-12	NC_010473	[4]
BW2952 [MC4100(MuLac)]	K-12	NC_012759	[5]
MDS42	K-12	NC_020518	unpublished
MC4100	K-12	*HG738867*	[63]
BW25113	K-12	*CP009273*	unpublished
KLY	K-12	*CP008801*	[66]
ER2796	K-12	*CP009644*	this work
ER3413	K-12	*CP009789*	this work
REL606	B	NC_012967	[67]
BL21(DE3)	B	NC_012971, NC_012892	[67]
BL21-Gold	B	NC_012947	unpublished
W (ATCC 9637)	W	NC_017635, NC_017664	[68, 69]
KO11	W	NC_017660, NC_016902	[69]
LY180	W	NC_022364	unpublished
ATCC 8739	Crooks	NC_010468	unpublished

doi:10.1371/journal.pone.0127446.t001

None of the four MTases are essential for viability. R-M systems are exchanged primarily by horizontal gene transfer, and the EcoKI system lies in a highly plastic region of the genome referred to as the immigration control region (ICR) [12]. The general variability of this region between *E. coli* strains [13] suggests this region can be removed without significant consequence, and consistent with this, the sequenced strain DH10B shows the Δ(*mrr-hsdRMS-mcrBC*) allele to be a deletion of a block of 45 genes including the entire ICR as well as flanking regions [4]. The gene *yhdJ* has also been deleted with no detectable phenotype, which is not surprising given that it appears to be silent under normal conditions [10]. Inactivating mutations of *dcm* also have no visible phenotype, although VSP repair is lost in at least one allele [14] and there are clearly alterations in the expression patterns of certain genes [8, 9]. *E. coli dam* mutant strains are viable, but show a pleiotropic effect with most phenotypes attributable to loss of strand discrimination during mismatch repair. These include increased rates of mutation [15] and recombination [16], increased sensitivity to certain cytotoxic agents such as cisplatin [17] and alkylating agents [18], widespread alteration of gene expression patterns [19] including

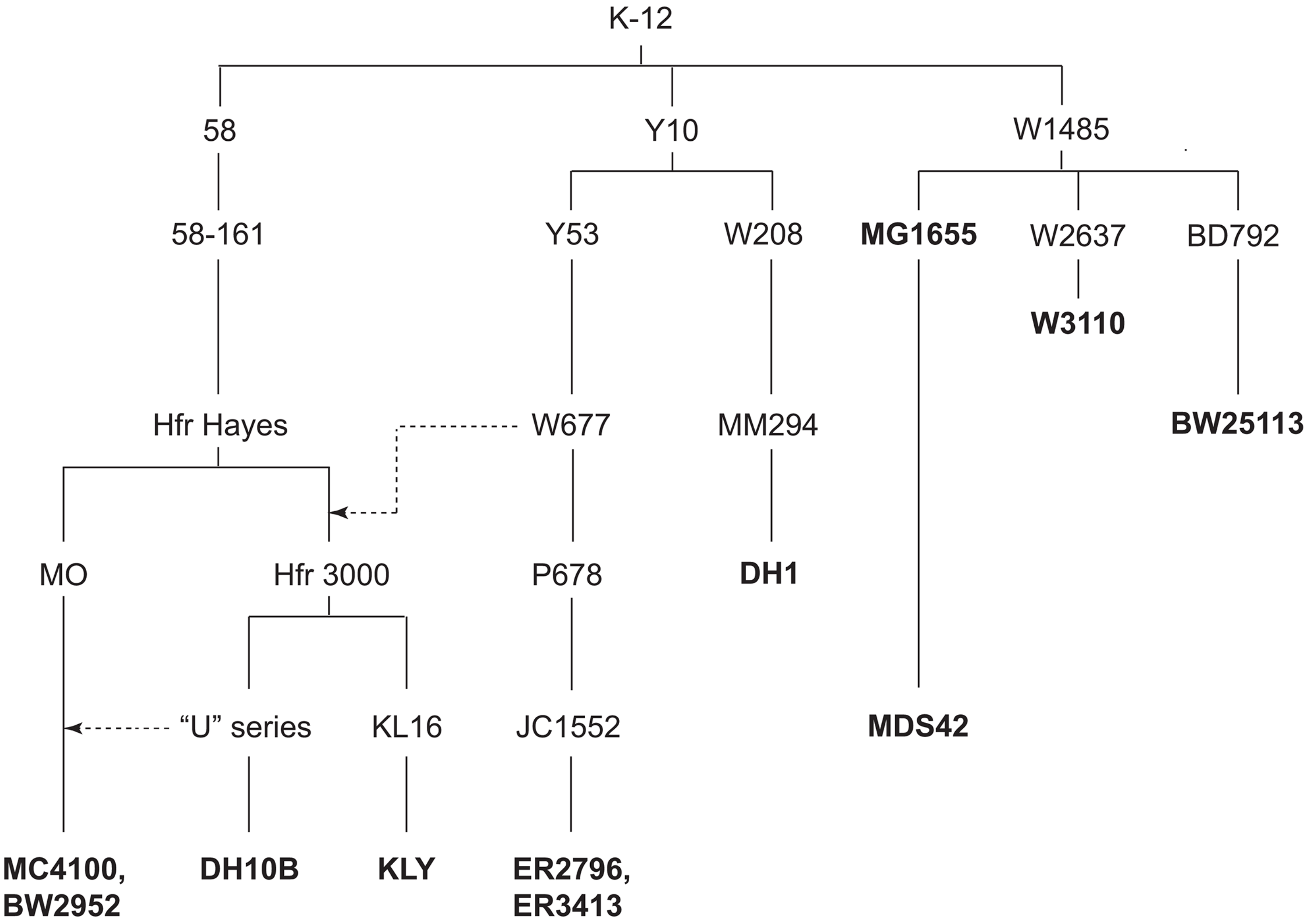


Fig 1. Relationship of ER2796 and ER3413 to the nine other completely sequenced *E. coli* K-12 strains. Completely sequenced strains are shown in bold type, and selected ancestral strains in Roman type. Most of the tree has been abstracted from Bachmann [28], except for the ancestries of MC4100 and DH10B back to Hfr Hayes, which are based on Laehnemann [63] and Durfee [4], respectively. Selected additional contributions of genetic material via crosses are shown by dotted lines. It appears based on genotype that Hfr 3000 U482 is the "U series" ancestor of DH10B, while Hfr 3000 U169 contributed genetic material in the ancestry of MC4100.

doi:10.1371/journal.pone.0127446.g001

Table 2. DNA restriction-modification genes in *E. coli* K-12 MG1655.[a]

	Gene	Product	Activity
DNA MTases			
	dam	orphan MTase M.EcoKDam	$G^{m6}A\underline{T}C$
	dcm	orphan MTase M.EcoKDcm	$C^{m5}CW\underline{G}G$
	yhdJ	orphan MTase M.EcoKII	$A\underline{T}GC^{m6}AT$ (silent)
	hsdM	Type I MTase M.EcoKI	$A^{m6}ACNNNNNNG\underline{T}GC$
Other Genes			
	hsdR	Type I restriction ENase R.EcoKI	
	hsdS	Type I specificity subunit S.EcoKI	
	mcrA	Type IV restriction ENase	
	mcrBC	Type IV restriction ENase	
	mrr	Type IV restriction ENase	

[a] All of these genes have been deleted or otherwise inactivated in ER2796 except for *yhdJ*, which is additionally inactivated in ER3413.

doi:10.1371/journal.pone.0127446.t002

activation of genes in the SOS regulon [20], and an absolute requirement for recombination (*recA*, *recB*, *recC*, *recG*, *ruvA*, *ruvB*, and *ruvC* genes) due to accumulation of double-strand DNA breaks [20–22].

Strains deficient in one or more of these MTases, particularly the *dam* and *dcm* functions, have found important uses in the molecular biology laboratory [14]. Since *E. coli* is the primary host for propagating plasmid DNA in the laboratory, and most *E. coli* strains are Dam^+/Dcm^+, plasmids typically bear methylation at GATC and CCWGG sites. This methylation can interfere with digestion by REases whose recognition sites overlap with or contain these patterns, and so propagation in a Dam^-/Dcm^- strain prior to digestion by such enzymes is required in these cases. In addition, methylation can trigger cleavage of DNA by Type IV (methyl-dependent) R-M systems, greatly reducing transformation efficiency of methylated plasmids in bacteria containing such systems [23]. Propagation in a Dam^-/Dcm^- strain to erase methylation patterns prior to transformation can alleviate this problem [24]. Older methodologies such as site-directed mutagenesis and Maxam-Gilbert DNA sequencing also benefited from Dam^-/Dcm^- strains [14], but these have largely been supplanted by newer technologies.

One emerging technology that benefits from methyl-deficient *E. coli* strains is Single-Molecule Real-Time (SMRT) DNA sequencing. Recent studies have shown that this sequencing technology can easily detect DNA methylation sites on both plasmid clones [25] and whole chromosomes [26], and has thus enabled the facile determination of recognition sites of DNA MTases. These patterns are most obvious when not obscured by *dam* and *dcm* methylation, and so the *E. coli* strain used for cloning DNA MTase genes in both of these studies was ER2796 (also called DB24 [27]), in which all three active, endogenous MTases have been inactivated (*dam dcm hsdM*), to ensure that the cloned heterologous MTase is solely responsible for any methylation observed. In this work we describe the complete genome sequence of ER2796.

Materials and Methods

Construction of ER2796

ER2796 is derived from JC1552, whose lineage from K-12 has been described previously (reference [28], with a condensed version shown in Fig 1). The construction of JC1552 involved, at various stages, treatment with X-ray and UV radiation, nitrogen mustard, and ethyl methanesulfonate, as well as a single conjugative cross with another K-12 derivative. Fig 2 shows the

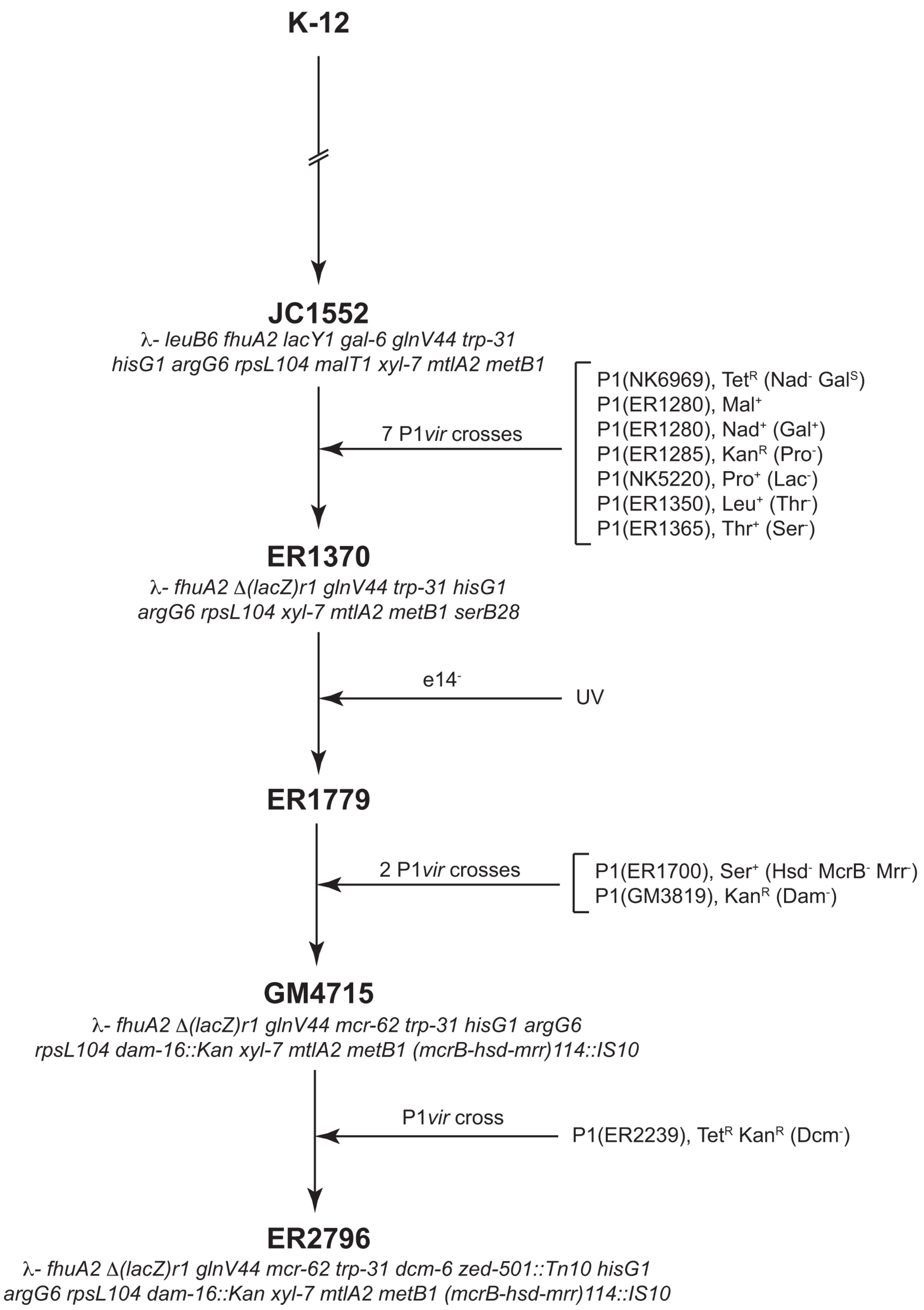

K-12
JC1552
λ- *leuB6 fhuA2 lacY1 gal-6 glnV44 trp-31*
hisG1 argG6 rpsL104 malT1 xyl-7 mtlA2 metB1
7 P1*vir* crosses
P1(NK6969), Tet^R (Nad$^-$ Gal^S)
P1(ER1280), Mal$^+$
P1(ER1280), Nad$^+$ (Gal$^+$)
P1(ER1285), Kan^R (Pro$^-$)
P1(NK5220), Pro$^+$ (Lac$^-$)
P1(ER1350), Leu$^+$ (Thr$^-$)
P1(ER1365), Thr$^+$ (Ser$^-$)
ER1370
λ- *fhuA2* Δ*(lacZ)r1 glnV44 trp-31 hisG1*
argG6 rpsL104 xyl-7 mtlA2 metB1 serB28
e14$^-$
UV
ER1779
2 P1*vir* crosses
P1(ER1700), Ser$^+$ (Hsd$^-$ McrB$^-$ Mrr$^-$)
P1(GM3819), Kan^R (Dam$^-$)
GM4715
λ- *fhuA2* Δ*(lacZ)r1 glnV44 mcr-62 trp-31 hisG1 argG6*
rpsL104 dam-16::Kan xyl-7 mtlA2 metB1 (mcrB-hsd-mrr)114::IS10
P1*vir* cross
P1(ER2239), Tet^R Kan^R (Dcm$^-$)
ER2796
λ- *fhuA2* Δ*(lacZ)r1 glnV44 mcr-62 trp-31 dcm-6 zed-501::Tn10 hisG1*
argG6 rpsL104 dam-16::Kan xyl-7 mtlA2 metB1 (mcrB-hsd-mrr)114::IS10

Fig 2. Lineage showing the construction of ER2796 from JC1552. Selected intermediate genotypes are shown. Markers that were selected are shown, followed by those that were screened for in parentheses. The lineage from K-12 to JC1552 has been described previously [28]. Genotypes are shown here using the historic allele names, but we suggest an updated nomenclature for some of these in Table 3 based on the genome sequence.

doi:10.1371/journal.pone.0127446.g002

derivation of ER2796 from JC1552. In brief, strain ER1370 was derived from JC1552 by a series of P1*vir* crosses; strain ER1779 was derived from ER1370 by UV treatment, which inactivated *mcrA* restriction, retrospectively attributed to loss of the e*14* prophage; and ER2796 was derived from ER1779 again by P1*vir* crosses, which deleted the restriction cluster (including *hsdM*) and introduced inactivating alleles of *dam*, and *dcm*. The intermediate strain GM4715, and the strain GM3819 used for the *dcm*-inactivating P1*vir* cross, have been described previously [14].

Genomic DNA and library preparation

100 mL of an overnight culture of ER2796, grown in LB medium [29], was resuspended in 10 mL [50 mM Tris (pH 8.0), 1 mM EDTA, 25% sucrose]. To this was added 8 mL 10 mg/mL chicken egg white lysozyme (Sigma-Aldrich, St. Louis, MO) dissolved in [250 mM Tris (pH 8.0), 250 mM EDTA]. Cells were incubated with the lysozyme at 37°C for 2 hrs, followed by two freeze-thaw cycles in dry ice/ethanol to facilitate cell breakage. To this was added 12 mL [50 mM Tris (pH 8.0), 62.5 mM EDTA, 1% Triton X-100] to complete breakage. Lysed cells were extracted once with 30 mL Tris-buffered phenol and once with 30 mL methylene chloride. Roughly 17 mL of the top layer was recovered. DNA was precipitated by addition of 0.1 volumes of 5 M NaCl and 0.7 volumes isopropanol, washed twice with 70% ethanol, and resuspended in a total of 0.8 mL buffer TE (Qiagen, Germantown, MD). Recovery was 93 μg as measured on a Qubit fluorometer (Invitrogen, Carlsbad, CA). All mixing was performed by gentle inversion to minimize DNA breakage.

To remove RNA, 15 μg ER2796 genomic DNA was incubated with 100 units of RNase If (New England Biolabs, Ipswich, MA) at 37°C for 1 hr in a 150 μL volume in the manufacturer's recommended buffer. DNA was sheared using a g-TUBE (Covarys, Woburn, MA) following the manufacturer's recommendations for 20 kb fragments (one 60 s pass at 5800 rpm in an Eppendorf 5415 microcentrifuge). DNA was purified using the PowerClean DNA Cleanup Kit (MoBio, Carlsbad, CA) and resuspended in a total of 75 μL manufacturer's buffer 7. Recovery was 5.7 μg as measured on a Qubit fluorometer. Analysis on a Bioanalyzer 2100 (Agilent Technologies, Lexington, MA) using a DNA-12000 chip showed a median fragment size of 10.8 kb.

A sequencing library was prepared using the DNA Template Prep Kit 2.0 (3–10 kb) (Pacific Biosciences, Menlo Park, CA) according to the manufacturer's protocol for 20 kb template preparation with BluePippin size-selection. Input was 66 μL (5 μg) sheared DNA, and the final elution step was in 31 μL of the manufacturer's Elution Buffer. Recovery was 1.7 μg DNA as measured on a Nanodrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE). The entire library (30 μL) was size-selected using BluePippin (Sage Science, Beverly, MA) with a 4000 bp start. Eluate (40 μL) was collected and the well was washed with 0.1% Tween20 buffer (Sage Science), which was combined with the eluate. Size-selected DNA (total 100 μL) was purified by one 1x AMpure PB magnetic bead step (Pacific Biosciences) and eluted in 31 μL Elution Buffer (Pacific Biosciences). Recovery of the library was 730 ng at 23.7 ng/μL, as measured by Nanodrop spectrometry.

DNA sequencing and assembly

Genome sequencing was carried out on the PacBioRS2 (Pacific Biosciences) using the DNA/Polymerase Binding Kit P4, MagBead Loading Kit, and Sequencing Kit 2.0 (all Pacific

Biosciences). Data from 4 SMRT cells was used, with one 180 min movie per cell. Sequencing reads were assembled using the HGAP 2.0 program in Pacific Biosciences' SMRTAnalysis pipeline. A mean coverage of 325x was achieved.

The sequence resolved into a single contig of 4,572,343 bp, of which roughly 13,700 bp was duplicated on the ends. Errors in the ends were analyzed by reassembling all reads against this duplicated region using the RS_Resequencing.1 program, and it was confirmed that all errors were confined to the 5' end of the overlap at the start of the contig, and the 3' end at the end of the contig, where coverage is lowest. The final non-redundant sequence was extracted to avoid the error-containing regions and rotated to bring the start in line with that of *E. coli* MG1655 (GenBank NC_000913.2).

Genome annotation

Genome annotation of the 4,558,663 bp chromosome of ER2796 was performed using the Institute for Genome Sciences (IGS) Prokaryotic Annotation Pipeline (http://ae.igs.umaryland.edu/cgi/intro_info.cgi). Briefly, an initial set of open reading frames (ORFs) likely to encode proteins was identified by GLIMMER (http://cbcb.umd.edu/software/glimmer/), overlapping ORFs were removed, and the resulting set of ORFs was searched against a database of non-redundant protein sequences, nr (composed of non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF, excluding those in env_nr), downloaded locally on the IGS systems. Two sets of hidden Markov models (HMMs) were used to determine ORF membership in families and superfamilies. These included 14,831 HMMs from PFAM version 27.0 (http://pfam.sanger.ac.uk/) and 4,284 HMMs from TIGRFam version 13.0 (http://www.jcvi.org/cgi-bin/tigrfams/index.cgi). TOPPRED was used to identify membrane-spanning domains in proteins.

Identification of candidate gene conversion events

The list of single nucleotide polymorphisms (SNPs) was inspected for clusters of changes from MG1655 (>2/100 nt). Four such clusters were identified, all within repeated sequences. These segments, with flanking sequences, were used as probes to BLAST the NCBI Genomes TaxID 83333 (*E. coli* K-12 sequences NC_000913.3 [MG1655], NC_020518.1 [MDS42], NC_007779.1 [W3110], NZ_CM000960.1 [MG1655star], NC_012759.1 [BW2952], and NC_010473.1 [DH10B]), looking for 100% match. For each such cluster, at least one matching sequence was identified at a distant locus, as discussed below.

Growth curves

Flasks of 25 mL Rich medium [29] with 100 μg/mL ampicillin were inoculated 1:250 with overnight cultures of MG1655 or ER2796 and grown at 37°C shaking at 225 rpm. At various time points, 200 μL of culture was withdrawn, diluted to 2 mL with Rich medium, and the OD_{600} of the diluted culture measured with a Biowave Cell Density Meter CO8000 (Biochrom Ltd., Cambridge, UK). Three replicate flasks were grown for each strain, and the mean of the readings was plotted. Growth rate constants were determined for the logarithmic growth phases (*t* = 50–159 min for MG1655, and *t* = 81–228 min for ER2796).

Inactivation of YhdJ (M.EcoKII)

The wild type *yhdJ* (encoding M.EcoKII) was amplified by PCR from *E. coli* ER2796 using the forward primer TAGTTGCGAGCTCTTAAGGTTAACATATGAGAACAGGATGTGAACCGAC and reverse primer TTATTAGCATGCTTACTTTGTAATGAGATCGGGGTC. The resulting 885 bp product was digested with SacI and SphI (sites underlined) and subcloned into the respective sites of pUC19. The *yhdJ* ORF in the resulting clone was inactivated by digesting it at an internal AgeI site, filling in the 4-base overhangs using the Klenow fragment of DNA polymerase I, and religating the resulting blunt ends. We confirmed its inactivation by transforming *E. coli* ER2796 with the wild type and disrupted plasmids and digesting the genomic DNA with NsiI. As expected [10], the wild type clone conferred protection from NsiI digestion, while the inactivated clone did not.

We inactivated the chromosomal copy of *yhdJ* in ER2796 by allelic exchange using the suicide vector pRE112 (ATCC 87692) [30]. This plasmid contains a conditional R6K origin of replication, positive selectable marker encoding chloramphenicol resistance, and a derivative of the *Bacillus subtilis sacB* gene for counterselection on sucrose-containing media. We first subcloned the inactivated allele of *yhdJ*, described above, into pRE112 at the SacI and SphI sites and transformed the mobilizing donor strain *E. coli* S17-1 (a gift of the late Saul Roseman) with the resulting plasmid, called pRE112:M.EcoKIIΔAgeI. Transconjugation was performed by mixing S17-1 [pRE112:M.EcoKIIΔAgeI] with the original 2796 strain on an LB plate for 24 hours at 37°C. The accepting ER2796 strain is resistant to kanamycin (Km) and tetracycline (Tc), and the donor S17-1[pRE112:M.EcoKIIΔAgeI] strain only carries chloramphenicol (Cm) resistance from the pRE112-derived plasmid. Therefore, recombinant clones were selected on LB with Km, Tc and Cm. The pRE112:M.EcoKIIΔAgeI plasmid cannot replicate in ER2796, so the plasmid must integrate into the chromosome to confer Cm resistance. Site-specific recombinants were identified by colony PCR with *yhdJ* flanking primers TTAGTTGCTCTAGATT AAGGTTAACATATGTTCGAACAACGCGTAAATTCTGAC and TTATTAGCATGCATG GCAAAAAGAACCAAAGCCG.

Among 20 screened clones, only two (#9 and #18) demonstrated the correct site-specific insertion into the *yhdJ* locus (S1B Fig).

Through a second round of recombination, we selected for clones that had lost the vector backbone and replaced the wild type *yhdJ* allele with the inactivated one by growth on LB medium containing Km, Tc, and 5% sucrose. Cm-sensitive and sucrose tolerant clones were screened by PCR amplification with the *yhdJ* flanking primers and digestion with AgeI. All 10 of the screened clones were resistant to AgeI digestion (S1C Fig). Sequencing of the *yhdJ* region of one such clone confirmed the disruption of the ORF as the expected 4 bp insertion, and this strain was designated ER3413.

Nucleotide accession numbers

The genome sequences of ER2796 and ER3413 are available from GenBank with the accession numbers CP009644 and CP009789, respectively. Both strains are available from New England Biolabs and the Coli Genetic Stock Center (CGSC).

Results

Inactivation of MTases

The three active, native *E. coli* MTases were all inactivated relatively recently in the lineage of ER2796, and all by P1*vir* crosses to introduce the mutant alleles (Fig 2). The *hsdM* gene, part of the EcoKI Type I RM system, is found in a cluster of restriction enzyme genes called the immigration control region (ICR). A deletion mutation (*Δ(fimB-opgB)114*::*IS10*) removing all of the restriction activities of this cluster was isolated following selection for loss of the ICR as in reference [31]. The deletion resulted from the action of the IS10 elements comprising the Tn10 insertion in *opgB* (formerly designated *zjj202*::*Tn10*). The IS10 action yielded an inversion/

deletion event that removed adjacent DNA and left a tandem IS10 repeat at the position of the original insertion [32].

The inactivation of *dam* has been described previously [33]. Briefly, a 0.5 kb region of a plasmid copy of the gene was removed by digestion at unique EcoRV and HpaI sites, and a 1.1 kb fragment containing the Kan^R marker from Tn903 was inserted in its place. This defective copy, which conferred no detectable Dam methylation activity *in vitro* or *in vivo*, was then recombined onto the chromosome. Only the first 54 residues of the protein remain encoded by ER2796.

The inactivation of *dcm* has also been described [34]. Mutagenesis was accomplished by treatment of cultures with *N*-methyl-*N*-nitro-*N*-nitrosoguanidine, and mutants were screened for the ability to accept radiolabeled methyl groups using wild type crude extracts. It had been shown previously that the *dcm-6* null allele (which also results in loss of *vsr* activity) resulted from a nonsense mutation at the 45th codon [35], and our sequencing results confirm this in strain ER2796.

The disruption or loss of *dam*, *dcm*, and *hsdM* in ER2796 together render the strain free of DNA methylation under all known conditions. The remaining DNA MTase, *yhdJ*, is functional when cloned and overexpressed [10], but is silent in its native context for unknown reasons. Unlike most DNA MTase genes, which are horizontally acquired and quickly lost, *yhdJ* is well conserved among enteric bacteria, exhibiting a similar host range to *dcm*. Thus it seems likely that it has a biological role to play, and may be activated under specific conditions not yet identified. To guard against this possibility, we constructed the derivative strain ER3413 in which *yhdJ* was permanently inactivated by a 4 bp insertion in the coding sequence. Either of these two strains should provide a methylation-free background in which to observe the activities of exogenously introduced DNA MTases by SMRT sequencing.

Sequencing and annotation of ER2796

We sequenced ER2796 using the Pacific Biosciences SMRT DNA sequencing platform from a large-insert library (see Materials and Methods), which resulted in a single linear contig. The closed circular genome is 4,558,663 base pairs in length and contains 50.8% G+C. Overall, the ER2796 genome shows high co-linearity with that of MG1655 (Fig 3). While ER2796 has undergone several insertions and deletions, there are no major rearrangements or inversions relative to MG1655.

Using the MG1655 annotation (NC_000913.2) as a template, we annotated a total of 4,083 protein-coding genes in ER2796. Genes annotated as "pseudogenes" in MG1655 were not reannotated in ER2796. However, 31 genes intact in MG1655 are disrupted in some way (by frameshift, insertion, deletion, or nonsense mutation) in ER2796 and were annotated as new "pseudogenes" and included in the 4,083 total. (Not counted among these 31 are *hisG*, which suffered a small in-frame deletion, and *rph*, which suffered a frameshift that restored an ancestral sequence.) Another 86 genes intact in MG1655 are missing altogether from ER2796, caused by 6 independent deletion events. Twenty-three genes in ER2796 are not present in MG1655 and were introduced primarily by insertion sequences. Finally, 109 genes in ER2796 suffered one or more missense mutations relative to their MG1655 orthologs. These mutations are a subset of the 249 SNPs identified relative to MG1655.

We similarly annotated a total of 174 RNA-coding genes in ER2796, including 89 tRNA, 22 rRNA, 2 tmRNA, and 61 ncRNA genes. Two RNA genes from MG1655 are missing from ER2796, *arcZ* (a ncRNA that acts as a positive antisense regulator of *rpoS*) and *symR* (a ncRNA that destabilizes the mRNA of *symE*, which is also missing from ER2796). Automated

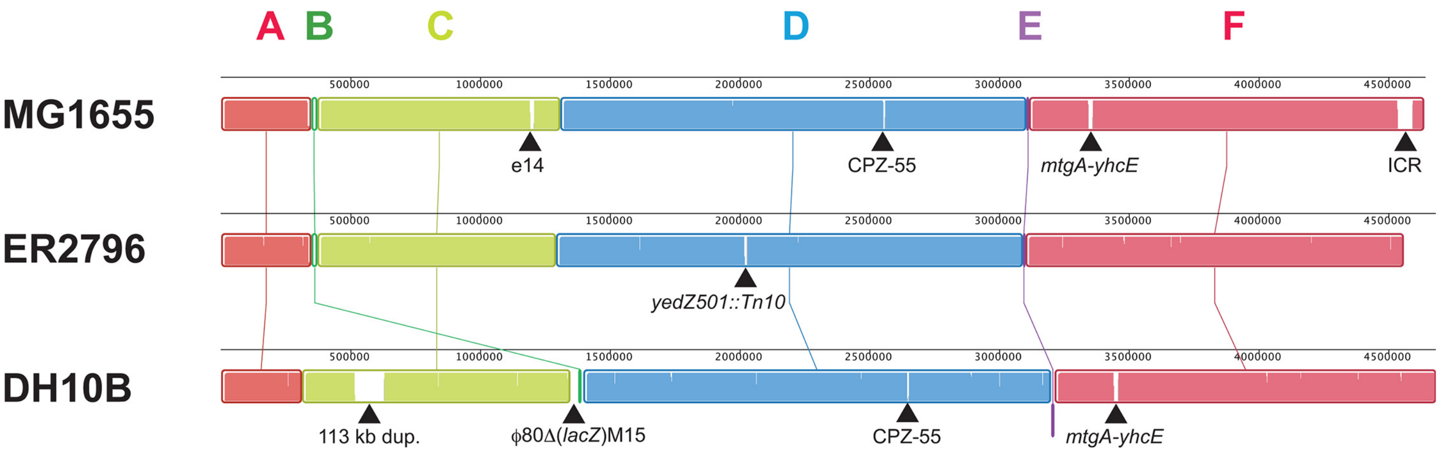


Fig 3. Alignment of the MG1655 genome with ER2796 and DH10B, conducted with Progressive-Mauve. Boundaries of the major contiguous blocks of sequence, labeled with capital letters, are formed by two major events specific to the DH10B lineage: block B results from deletion of a 34.6 kb region of MG1655 followed by partial restoration as part of a φ80Δ*(lacZ)M15* mosaic prophage insertion in DH10B; and block E results from the IS10-mediated inversion of an 11 kb segment of MG1655, again in DH10B [4]. The following larger indels visible in the figure are labeled: prophage e14 lost in both ER2796 and DH10B; prophage CPZ-55 lost in ER2796; the 16 kb *mtgA-yhcE* region lost in ER2796 through IS5-mediated deletion; the ICR region deleted in both ER2796 and DH10B; Tn10 insertion at *yedZ* in ER2796; tandem duplication of a 113 kb region in DH10B, presumably IS5-mediated; the φ80Δ*(lacZ)M15* mosaic prophage insertion in DH10B, including the *lacZ* region (part of block B).

doi:10.1371/journal.pone.0127446.g003

annotation of ER2796 also identified a cryptic tRNA gene at 347,232–347,312 that is also present but not annotated in MG1655; this is presumably a pseudogene and was not annotated in ER2796.

A complete list of all mutations relative to MG1655 is shown in S1 Table, and a complete list of all genes affected by these mutations is shown in S2 Table. Strain ER3413 was also completely sequenced, and in addition to the engineered disruption of *yhdJ*, we observed seven single base changes, one single base insertion, and a likely gene conversion event relative to ER2796 (S3 Table).

Analysis of DNA methylation in ER2796

We used the program RS_Modification_and_Motif_Analysis.1 (Pacific Biosciences) to confirm the absence of methylation in the ER2796 sequence. As expected, no methylated motifs were identified at modification quality value (QV) thresholds of 30 or 20. To look for possible sparse methylation of ATGCAT sites indicative of M.EcoKII activity, we examined each of the 1644 instances of the site (on both strands) in the reference for possible methylation at the second A residue. Although 162 of these residues had mean IPD ratios with QV values greater than 20, none were identified by the program as having a characteristic m^{6}A kinetic signature. In addition, we repeated the analysis using a negative control sequence, TTGCAA, and obtained a comparable result (166/2004 with QV > 20, compared to 162/1644). Thus, despite the presence of an intact *yhdJ* gene in this strain, there is no evidence of activity in the sample sequenced here.

Mutations underlying the ER2796 genotype

The ER2796 genotype comprises 18 genetic markers with known phenotypes, including 15 historical markers based on its lineage and 3 additional markers revealed by sequencing (Table 3). These last 3 are changes from MG1655 found by others in various K-12 strains and shown to have phenotypic consequences: *rpoS396* [36], *luxS11* [37] and *rph* WT [38]. EcoGene [39], EcoCyc [40] and the resources of the *E. coli* Genetic Stock Center (CGSC) were relied on for

Table 3. Genotype markers in ER2796 and underlying sequence features.

Allele	Old Allele Name (if changed)	Alteration	Genes Affected	ER2796 Sequence	MG1655 Sequence	Amino Acid Changes
fhuA2::*IS2*	*fhuA2*	*IS2* disruption	*fhuA* (b0150)	167920–169255 (169251–169255 is target site duplication)	between 167919–167920	ER2796_149 (aa 1–145 + 13 aa), and ER2796_151 (aa 158–747)
ΔlacZ4826	*Δ(lacZ)r1*	deletion	*lacZ* (b0344)	between 365092–365093	362419–364862	Δ223–1024; adds 40 aa extension overlapping *lacY*
glnX44[a]	*glnV44*	tRNA transition	*glnX* (b0664)	693302 (T)	695693 (C)	(DNA nt) G34A
e14$^-$ (McrA$^-$)	*mcr-62*	excision	*ymfDE*, *lit*, *intE*, *xisE*, *ymfIJ*, *cohE*, *croE*, *ymfLM*, *oweE*, *aaaE*, *ymfR*, *bee*, *jayE*, *ymfQ*, *stfP*, *tfaPE*, *stfE*, *pinE*, *mcrA* (b1137-b1141, b1143-b1148, b4692-b4693, b1150-b1159, respectively)	between 1193386–1193387	1195598–1210801	null; associated changes in *icd* sequence
trpE31	*trp-31*	missense	*trpE* (b1264)	1302017 (T)	1319610 (C)	G454D (ER2796_1284)
dcm-6		silent	*dcm* (b1961)	2012149 (T)	2029184 (C)	E386
		nonsense (TGA)[b]	*dcm* (b1961)	2013172 (T)	2030207 (C)	W45stop (ER2796_2013; also ER2796_2012 from internal start at aa 111)
yedZ501:: *Tn10*(Tet^{R})	*zed-501*:: *Tn10*	*Tn10* insertion	*yedZ* (b1972)	2021730–2030885 (2030877–2030885 is target site duplication)	between 2038764–2038765	Δ87–211; ER2796_2025 (aa 1–86 + 15 aa), and ER2796_2035 (from internal start at aa 102)
Δ(hisG)1	*hisG1*(Fs)	deletion, in-frame	*hisG* (b2019)	between 2080789–2080790	2088669–2088704	Δ152–163 (ER2796_2082)
luxS11[c]		–1 frameshift	*luxS* (b2687)	between 2798558–2798559	2812480 (A)	Δ92–171; ER2796_2763 (aa 1–90 + 20 aa), and ER2796_2762 (from internal start at aa 108)
		silent	*luxS* (b2687)	2798561 (C)	2812483 (T)	L91 (ER2796_2762)
rpoS396 (Am)[d]		nonsense (TAG)	*rpoS* (b2741)	2851555 (A)	2865477 (G)	E33stop ER2796_2821 (from internal start at aa 40)
argG6(Fs)	*argG6*	–1 frameshift	*argG* (b3172)	between 3304561–3304562	3317286 (C)	Δ210–447 ER2796_3265 (aa 1–209 + 13 aa), and ER2796_3266 (from internal start at aa 221)
rpsL104		missense	*rpsL* (b3342)	3443238 (G)	3472313 (T)	K88Q
		missense	*rpsL* (b3342)	3443372 (G)	3472447 (T)	K43T (ER2796_3428)
Δdam-16:: *Kan^{R}*		deletion + 1266 bp Kan^{R} insertion	*dam* (b3387)	3484166–3485431	3513241–3513773	Δ55–242 ER2796_3474 (from internal start at aa 242)
xyl-7		missense	*xylB* (b3564)	3699368 (A)	3726511 (G)	A295V (ER2796_3669)
		missense	*xylA* (b3565)	3700835 (A)	3727978 (G)	H271Y (ER2796_3670)
		silent	*xylA* (b3565)	3700836 (G)	3727979 (A)	N270
		insertion (IS1)	*xylF* (b3566)	3702309–3703085	between 3729451–3729452	Δ100–330; adds 1 aa extension (ER2796_3671)
mtlA2(Fs)	*mtlA2*	–2 frameshift	*mtlA* (b3599)	between 3744696–3744697	3771063–3771064 (GG)	Δ254–637 ER2796_3708 (aa 1–253 + 60 aa), and ER2796_3709 (from internal start at aa 306)

(*Continued*)

Table 3. (*Continued*)

Allele	Old Allele Name (if changed)	Alteration	Genes Affected	ER2796 Sequence	MG1655 Sequence	Amino Acid Changes
rph^{WT} [e]		+1 frameshift	*rph* (b3643)	3787535 (C)	between 3813902–3813903	ER2796_3754
metB1(Fs)	*metB1*	–2 frameshift	*metB* (b3939)	between 4100468–4100469	4126836–4126837 (CG)	Δ48–386; ER2796_4061 (8 aa + aa 48–386)
Δ(fimB-opgB)114::*IS10*(RM^-)[f]	*Δ(mcr-hsd-mrr)114*::*IS10*	deletion of 60,679 bp + insertion of 2 *IS10* elements in direct repeat	*fimBEAICDFGH*, *gntP*, *uxuABR*, *yjiC*, *iraD*, *yjiE*, *iadA*, *yjiGH*, *kptA*, *yjiJKLMN*, *mdtM*, *yjiPRSTV*, *mcrCB*, *symER*, *hsdSMR*, *mrr*, *yjiAXY*, *tsr*, *yjjLMN*, *opgB* (b4312-b4337, b4339-b4342, b4486, b4345-b4347, b4625 [ncRNA], b4348-b4359, respectively)	4511776–4514442 (4511776–4513104 and 4513114–4514442 are *IS10*, 4513105–4513113 is target site duplication [inverted])	4537567–4595455	null, except *opgB* (b4359) Δ671–763; ER2796_4466 (aa 1–670)

[a] The reassignment of *glnV44* (*supE44*) was noted previously [43].

[b] The double mutation (one silent) is in agreement with a previous study [35].

[c] The sequence of *luxS* reported here is identical to a previous study [70], although our alignment differs slightly, moving the frameshift 3 nt and inferring a transition instead of a transversion. The steps that resulted in the shared *luxS11* allele clearly include a base deletion and a base change, but exactly which deletion and which base change depend on the local alignment. Spontaneous unselected transitions are somewhat more frequent than transversions [71], so our alignment may be preferable. The mutation is present in DH1 [72] (see Table 1), an ancestor of the strain used in [70] and may have been present in sibling strains JC1552 (ancestral to ER2796; $RecA^+$) and JC1553 (source of the *recA1* allele of the DH1 and its descendants [73, 74]). The *luxS* and *recA* genes are very close, about 8 kb apart, and introduction of *recA1* was the last step in construction of DH1.

[d] This nonsense mutation, which is common in laboratory *E. coli* strains [75], was most likely ancestral, not introduced by transduction. It may be partially suppressed in this strain. The *rpoS* mutation and the accompanying *supE44* mutation (identified here as *glnX44*) can be traced to strain Y10, very early in the K-12 pedigree [28].

[e] This frameshift mutation presumably restores the wild type state, reverting the frameshift present in early K-12 derivative strains MG1655 and W3110 [76].

[f] The position of the parental *zjj202*::*Tn10* is inferred to be 4597466–4597474 of MG1655 (NC_000913.2), the nine base target sequence that is duplicated upon insertion.

doi:10.1371/journal.pone.0127446.t003

gene, function and pedigree information. The newly found markers have subtle phenotypes but may have been selected by geneticists nonetheless, since stable markers in a healthy background are easier to work with. The wild type state of *rph* promotes growth in minimal media by improving pyrimidine biosynthesis, thus fostering a desirable healthy state. The *rpoS396* allele is at least partially suppressed by the accompanying amber suppressor [36]. At least one attempt to replace the suppressor in this lineage by transduction was foiled (EAR, unpublished observation), consistent with selective pressure to maintain suppression of *rpoS396*. The *luxS11* allele would interfere with "social" interactions mediated by quorum sensing [41], but this lineage has not been used to study this phenomenon.

Of the 15 purposely-introduced markers, the nine found in JC1552 (Fig 2) were generated during early studies of genetic processes, and six more were introduced during more recent studies of genetic processes. These 15 exhibit properties desirable for genetic study: they have strong phenotypes and are stable. 13 of the 15 are indels or nonsense mutations.

Of the full set of 18 markers, three had already been characterized at the sequence level prior to this study. Single-base changes led to the *glnX44* amber suppressor (formerly attributed to the adjacent duplicate tRNA *glnV*) [42, 43], and the *dcm-6* opal and silent mutations [35] (although we did not observe the Q26R mutation described there). As expected, the *mcr-62* allele is an excision of the e*14* prophage. Its sequence is identical to that of DH10B, which carries an independent excision allele in a different lineage [4]. Excision is a relatively frequent event that restores the sequence of *icd* to a presumptively ancestral state by fusion of the N-terminus of *icd* to the original C-terminus, the "pseudogene" *icdC* [44]. We rename this allele "e*14*$^-$ (McrA$^-$)" to match other strains.

The remaining markers from the Bachmann pedigree [28] include an IS2 insertion (*fhuA2*::*IS2*), three frameshifts [*argG6*(Fs), *mtlA2*(Fs) and *metB1*(Fs)], a double missense (*rpsL104*) and a single missense (*trpE31*). The *xyl-7* allele includes four mutations in three of the six genes of the regulon: a conservative missense change in *xylB*, a nonconservative and a silent change in *xylA*, and an IS1 insertion in *xylF*. The *Δ(hisG)1* mutation is unexpectedly found to be a 36 nt deletion, rather than a frameshift mutation. This deletion removes a sequence flanked by four-base direct repeats of ACTG, and one of the repeats. Isolation of a more stable allele during lineage manipulation may account for the conflict with the original report of a revertible allele [36]. Others have reported to CGSC that this allele is non-revertible (John Wertz, personal communication).

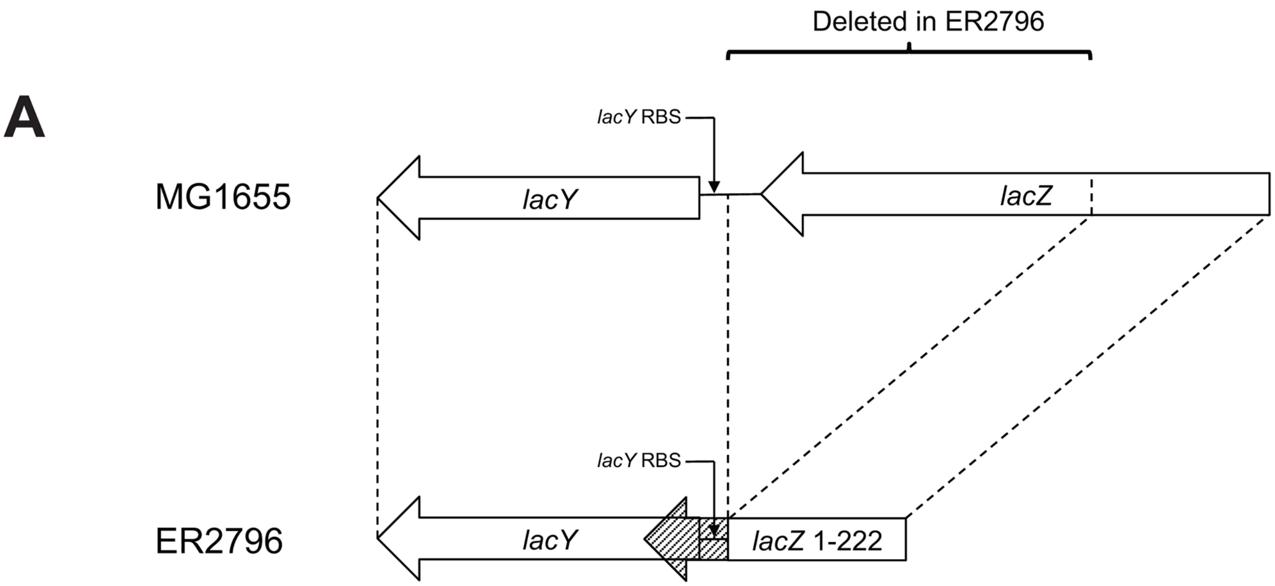
The *ΔlacZ4826* (formerly *Δ(lacZ)r1*) marker is a large *lacZ* deletion that permits LacYA activity. Historically, it was used to characterize the phenomenon of transcriptional polarity [45, 46] in which long untranslated regions in mRNA result in RNA degradation and reduced expression of genes later in in the operon. By bringing the *lacY* translation initiation site close to nonsense mutations early in *lacZ*, it restored much of the activity of *lacYA* that was lost due to early translation termination in the single nonsense mutants. Here we find that the deletion border is actually within the *lacZ-lacY* intergenic region (Fig 4A). The *lacZ* fragment encoding the N-terminal region is extended by 40 codons overlapping the *lacY* start codon (Fig 4B). The LacY activity is weak in the single mutant configuration (determined from growth on melibiose at 42°C) [47], suggesting that translation of the LacZ chimeric protein may compete with LacY initiation when translation is not stopped early.

The remaining three markers were engineered (*dam-16*::*Kan*^{R}) [48], or transposon-mediated [*yedZ501*::*Tn10* [49] and *Δ(fimB-opgB)114*::*IS10*(RM$^-$) [31]]. *Δ(fimB-opgB)114*::*IS10* was mediated by the parental *zjj202*::*Tn10*, which rearranged to remove the unique Tn10 material and adjacent DNA. Deletion of the highly variable immigration control region (ICR) was selected for, yielding an inversion/deletion event that removed adjacent DNA and left a tandem IS10 repeat at the position of the original insertion. The positions of the *yedZ501*::*Tn10* and the distal IS10 from *zjj202* agree with those reported by Nichols [50].

Gene conversion events

We found four instances of clustered SNPs (relative to MG1655) in repeat loci for which a perfect donor copy at a different locus could be identified. These properties strongly suggest gene conversion events. The unidirectional information transfer process can occur by multiple mechanisms [51, 52].

Three of the seven rRNA operons have served as information recipients; in one case a unique donor was identified. The most compelling includes the entire *rrsB* gene and the intergenic region between *rrsB* and *rrlB*. Six SNPs, a deletion of 20 bp and an insertion of 106 bp span coordinates 4,138,303–4,140,042 (corresponding to MG1655 4,166,238–4,166,499) (S1 Table). The sequence aligns perfectly with the corresponding sequence of *rrsE-rrlE*, both in ER2796 and in MG1655 (Fig 5). The contiguous clustering of the six base changes, a 106 bp deletion and a 20 bp insertion strongly suggest unidirectional information transfer from *rrsE-rrlE*



B

```
364924 TAGCGGCAAAAATAATACCCGTATCACTTTTGCTGATATGGTTGATGTCATGTAGCCAAATCGGGAAAAACGGGAAGTAG
         A  A  F  I  I  G  T  D  S  K  S  I  H  N  I  D  H  L  W  I  P  F  F  P  F  Y
                                                *  T  A  L  D  P  F  V  P  L  L
365004 GCTCCCATGATAAAAAAGTAAAAGAAAAAGAATAAACCGAACATCCAAAAGTTTGTGTTTTTTAAATAGTACATAATGGA
       A  G  M  I  F  F  Y  F  F  F  F  L  G  F  M  W  F  N  T  N  K  L  Y  Y  M
                                                                           ← lacY
        S  G  H  Y  F  L  L  L  F  L  I  F  R  V  D  L  L  K  H  K  K  F  L  V  Y  H  I
365084 TTTCCTTACGCGAAATACGGGCAGACATGGCCTGCCCGGTTATTATTATTTTTGACACCAGACCAACTGGTAATGGTAGC
         E  K  S                                     *  K  Q  C  W  V  L  Q  Y  H  Y  R
       GACCGGCGCTCAGCTGGAATTCCGCCGATACTGACGGGCTCCAGGAGTCGTCGCCACCAATCCCCATATGGAAACCGTCG
         G  A  S  L  Q  F  E  A  S  V  S  P  S  W  S  D  D  G  G  I  G  M  H  F  G  D
                           ... [not shown are an additional 2160 bp / 720 aa] ...
       TGCGTTTCACCCTGCCATAAAGAAACTGTTACCCGTAGGTAGTCACGCAACTCGCCGCACATCTGAACTTCAGCCTCCAG
       Q  T  E  G  Q  W  L  S  V  T  V  R  L  Y  D  R  L  E  G  C  M  Q  V  E  A  E  L
       TACAGCGCGGCTGAAATCATCATTAAAGCGAGTGGCAACATGGAAATCGCTGATTTGTGTAGTCGGTTTATGCAGCAACG
        V  A  R  S  F  D  D  N  F  R  T  A  V  H  F  D  S  I  Q  T  T  P  K  H  L  L  S
365120 AGACGTCACGGAAAATGCCGCTCATCCGCCACATATCCTGATCTTCCAGATAACTGCCGTCACTCCAGCGCAGCACCATC
         V  D  R  F  I  G  S  M  R  W  M  D  Q  D  E  L  Y  S  G  D  S  W  R  L  V  M
365200 ACCGCGAGGCGGTTTTCTCCGGCGCGTAAAAATGCGCTCAGGTCAAATTCAGACGGCAAACGACTGTCCTGGCCGTAACC
       V  A  L  R  N  E  G  A  R  L  F  A  S  L  D  F  E  S  P  L  R  S  D  Q  G  Y  G
365280 GACCCAGCGCCCGTTGCACCACAGATGAAACGCCGAGTTAACGCCATCAAAAATAATTCGCGTCTGGCCTTCCTGTAGCC
        V  W  R  G  N  C  W  L  H  F  A  S  N  V  G  D  F  I  I  R  T  Q  G  E  Q  L  W
365360 AGCTTTCATCAACATTAAATGTGAGCGAGTAACAACCCGTCGGATTCTCCGTGGGAACAAACGGCGGATTGACCGTAATG
         S  E  D  V  N  F  T  L  S  Y  C  G  T  P  N  E  T  P  V  F  P  P  N  V  T  I
365440 GGATAGGTCACGTTGGTGTAGATGGGCGCATCGTAACCGTGCATCTGCCAGTTTGAGGGGACGACGACAGTATCGGCCTC
       P  Y  T  V  N  T  Y  I  P  A  D  Y  G  H  M  Q  W  N  S  P  V  V  V  T  D  A  E
365520 AGGAAGATCGCACTCCAGCCAGCTTTCCGGCACCGCTTCTGGTGCCGGAAACCAGGCAAAGCGCCATTCGCCATTCAGGC
        P  L  D  C  E  L  W  S  E  P  V  A  E  P  A  P  F  W  A  F  R  W  E  G  N  L  S
365600 TGCGCAACTGTTGGGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGATGTGCTGCAAGG
         R  L  Q  Q  S  P  R  D  T  R  A  E  E  S  N  R  W  S  A  F  P  P  H  A  A  L
365680 CGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAATCCGTAATCATGGTCAT
       R  N  L  Q  T  V  G  P  N  E  W  D  R  R  Q  L  V  V  A  L  S  D  T  I  M  T  M
                                                                                ← lacZ
```

Fig 4. Comparison of the *lacZY* regions of MG1655 and ER2796. A. Schematic drawing showing the region of MG1655 *lacZ* and *lacZY* intergenic region that is deleted in ER2796. It is oriented forward with respect to the chromosomal sequence, with the operon reversed from the conventional representation. In ER2796, the *lacZ* ORF enodes amino acids 1–222 of MG1655 *lacZ* (white box) fused to 40 amino acids derived from the *lacZY* intergenic region, and overlapping with *lacY* (cross-hatched box). The putative *lacY* ribosome binding site (RBS) is preserved in ER2796. B. DNA and translated protein sequence of the *lacZY* junction, numbered from ER2796. Nucleotides and translated amino acids missing in ER2796 are shown in gray, and those present are shown in black. In ER2796, aa 1–222 of the translated ORF are shown in black, and the 40 aa derived from the intergenic region are shown in red. Start codons of *lacZ* and *lacY* are highlighted, and the putative RBS of *lacY* is underlined. 2160 bp (720 aa) of MG1655 *lacZ* have been removed at the indicated position for brevity.

doi:10.1371/journal.pone.0127446.g004

into *rrsB-rrlB*, i.e. gene conversion, rather than 8 separate mutational events, even if a ready mechanism were available to explain the insertion and deletion.

Two other cases involve clustered changes in *rrl* loci, although more than one donor locus is possible: 4 SNPs in *rrlG* are candidates for a conversion patch from one of four possible donors (*rrlA*, *C*, *E or H*), and 3 SNPs in *rrlD* are candidates for a conversion patch from any of the other six *rrl* loci. This conversion patch is shared with DH10B.

Authentic SNPs do occur in rRNA genes however. No candidate donor was found for two SNPs in *rrlC* in any of 6 K-12 genome sequences at NCBI. All 42 hits (7 operons per genome) had two mismatches at the same positions.

In addition to the ribosomal operons, a fourth gene conversion patch was observed between two prophage sequences. Eight SNPs in the Qin prophage represent conversion to identity with a similar sequence in the Rac prophage. The conversion makes the *ydfK* transcript identical to that of *ynaE*. The *ydfK* upstream untranslated region (UTR) acts as an RNA-mediated thermosensor, and the *ynaE* UTR was inferred to function similarly, from similarity of predicted secondary structure [53].

Indels

There are 42 insertions and deletions in ER2796 compared to MG1655 (S1 Table). Of these, 18 are 1–2 bp in length, resulting in frameshifts in 13 protein-coding genes; 14 are simple gain or loss of mobile elements (discussed below); one is a 16 kb deletion promoted by IS recombination; 4 are selected or engineered deletions (discussed above); and 5 are larger intergenic indels ranging from 8–181 bp in length. Besides the 13 frameshifted protein-coding genes, another has an in-frame deletion, 9 are disrupted by mobile elements or engineered deletions, and 2 RNA genes have 1–2 bp indels, for a total of 25 genes. Finally, large-scale deletions resulted in the loss of 86 genes relative to MG1655, primarily in prophages: 5 in the loss of DLP12, 18 by the loss of e14, 2 by the loss of IS1H at the *flhD* locus, 9 by the loss of CPZ-55, 9 by the IS5R-promoted deletion, and 43 by the loss of the ICR region. Aside from IS sequences, the only genes gained by ER2796 relative to MG1655 are *aph* (ER2796_3475), encoding kanamycin resistance (a consequence of *dam* inactivation), and the 7 unique genes of Tn10, including *tetA* (ER2796_2029), encoding tetracycline resistance.

Polymorphisms

ER2796 contains a total of 249 single-base changes relative to MG1655 (S1 Table), including the 21 that we propose result from 4 gene conversion events (discussed above). These result in 116 missense mutations, 9 nonsense mutations, and 68 silent mutations in protein-coding genes, 11 changes in RNA-coding genes, and 38 base changes in intergenic regions. One of the missense mutations alters a former stop codon to allow readthrough. The nonsense mutations result in known inactivation of the products of *dcm* and *rpoS* (Table 3), and truncation of the products of *rssB*, *ydbH*, *htpX*, *rsmF*, *hycC*, *cptB*, and *galP* (S1 and S2 Tables). The missense mutations occur in a total of 110 different genes, and between missense, nonsense, and RNA-

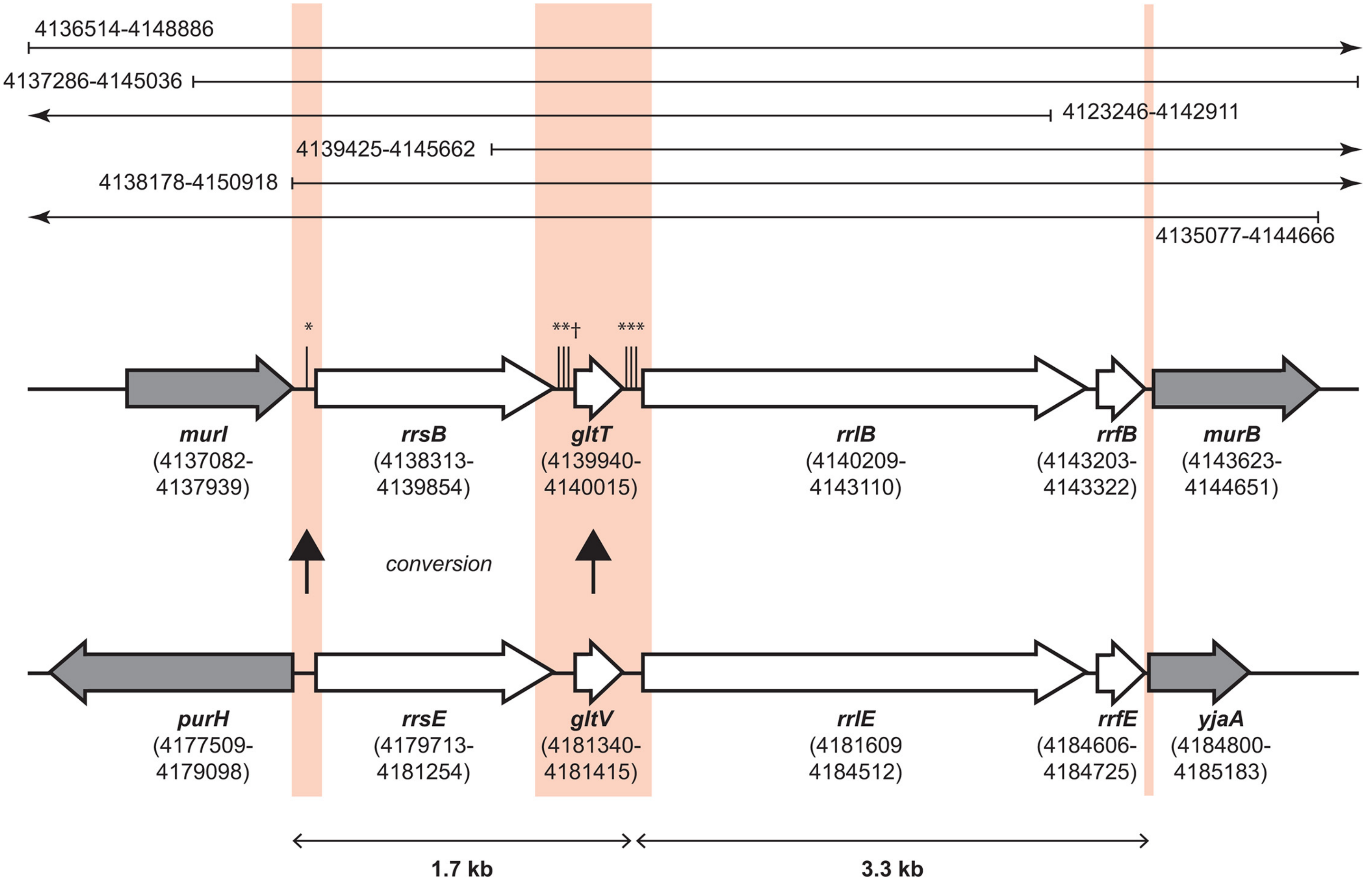


Fig 5. Use of long reads to identify gene conversion events. The schematic alignment shows the paralogous ribosomal gene clusters *rrnB* and *rrnE* from ER2796 (white genes) along with nonhomologous flanking genes (gray). The genes are marked with names and coordinates in ER2796. In ER2796, *rrnB* has been the apparent recipient of a gene conversion event in which *rrnE* served as donor (vertical arrows), and thus both regions are identical. As a result of this event, *rrnB* in ER2796 exhibits minor variations when compared with *rrnB* from its ancestor, MG1655: six SNPs (marked with *) and one indel (marked with †). Red tinted boxes indicate the regions of alteration (left and middle) and delineate the boundaries of the clusters (left and right). Sequencing reads internal to the clusters (i.e., between the outer two red boxes) cannot be mapped uniquely to one locus or the other unless they extend into the nonhomologous flanking regions, and the minor variants within (e.g., the middle red box) cannot be assigned to one cluster or the other without sequencing reads directly connecting them with a flanking region on one side or the other. The long-read library used in this analysis includes numerous reads that connect the unique flanking regions with the internal variants. The mapped coordinates of six example reads from the actual analysis are shown at the top, including some that span both sides of the 5 kb gene cluster. Arrows indicate where a read continues beyond the region shown here.

doi:10.1371/journal.pone.0127446.g005

coding mutations, a total of 124 gene products are altered through polymorphism relative to MG1655 where they occur in both genomes.

By contrast, DH10B contained a total of 132 single-base changes relative to MG1655, resulting in 66 genes with missense and 5 with nonsense mutations [4]. A total of 17 SNPs are shared between ER2796 and DH10B, some of which may have been acquired in the DH10B lineage through a genetic cross with W677, an ancestor of ER2796, or through other crosses. Although it might be suggested that the greater number of SNPs in ER2796 is due to the mutator phenotype resulting from *dam* inactivation, that mutation was introduced only recently in the ER2796 lineage (Fig 1), and ER2796 and DH10B have comparable numbers of SNPs in intergenic regions (38 and 42, respectively).

Mobile elements

Many of the larger indels in ER2796 discussed above result from the gain or loss of mobile elements relative to MG1655. These include four IS1 insertions and one associated deletion, two IS5 insertions and one associated deletion, one IS2 insertion, and two solo IS10 insertions in addition to the four IS10s associated with deliberately introduced changes (discussed above).

The lack of an IS element (IS1H in the case of MG1655) at the regulatory region of the *flhD* operon may lead to a poor motility phenotype in ER2796. The *flhD* operon is the master operon of the flagellar regulon, and the presence of an IS element there has been shown to increase operon expression and is associated with high motility [54]. Of the genes disrupted by IS insertions, *nohD* is a part of the DLP12 prophage not otherwise lost; *xylF* and *fhuA* are part of known alleles (Table 3); *mglA* is part of the beta-methylgalactoside ATP-Binding Cassette transporter, and its inactivation is expected to impair galactoside uptake [55]; *tdcD* participates in threonine degradation; and *rclA* is essential for survival of reactive chlorine stress [56].

Three MG1655 prophages are missing or compromised in ER2796: DLP12 (3.4 kb, partial loss associated with a new IS1A insertion), CPZ-55 (6.8 kb, precise excision), and e*14* (16 kb, precise excision). The largest deletion, 55 kb, was mediated by the parental *zjj202*::*Tn10* insertion, discussed above. The genes lost in these four major events primarily consist of phage-related functions, but the removal of e*14* and the ICR results in the loss of all of the restriction-modification systems from MG1655, namely *mcrA*, *mcrBC*, *mrr*, and *hsdRMS* (EcoKI). In all, 83 genes were completely deleted in these events, and 3 additional genes (*ybcV*, *eutA*, and *opgB*) were disrupted.

Growth properties

Growth curves were determined for MG1655 and ER2796 in Rich medium, and exponential growth rate constants were calculated as 0.0272 for MG1655 and 0.0197 for ER2796. These correspond to doubling times of approximately 25 min for MG1655 and 35 min for ER2796.

Discussion

Gene conversion events

Examination of the complete sequence of ER2796 revealed the occurrence of four gene conversion events between repeated sequences. Gene conversion is a nonreciprocal recombination process in which one copy of a diverged repeat donates its sequence to another, erasing the recipient version. These are evolutionarily important events that have rarely been confirmed in bacterial systems [57, 58]. Intragenomic gene conversion can counter mutational drift of repeated sequences, homogenizing them, or can be programmed to generate variation, as when silent-locus copies are donated to expression loci in antigenic phase variation (e.g., [59]). Here, gene conversion would contribute to the bewildering network of phage interrelationships [60]. The fact that ER2796 has a known pedigree with no horizontal transfer from outside the lineage, together with the long reads enabled by the Pacific Biosystems SMRT sequencing platform, enabled detection of these events. Detection depends on the presence of divergent copies in parent participants, on a complete inventory of parental copies and of the resulting offspring (i.e., complete genomes), and crucially, on correct assembly of long repeats (>5 kb) such as ribosomal operons (Fig 5).

Utility of MTase-deficient strains

The availability of strains of *E. coli* that are completely defective in DNA methylation has obvious advantages for studying the methylation specificity of cloned DNA MTase genes. This has already been realized in a number of studies [25, 26, 61, 62] and will be extremely useful going forward when confirmation of inferences made by whole genome sequencing using SMRT technology need to be confirmed. The absence of the Type I R-M system (EcoKI) means that this strain is more easily transformable since EcoKI is the only known ENase in *E. coli* K-12 recognizing unmethylated DNA. Similarly, the absence of the methylation dependent restriction ENases (McrA, McrBC and Mrr) mean that this strain is also suitable for cloning intact restriction systems that may otherwise be restricted because of the introduced MTase.

It should be noted that the loss of the Dam MTase confers a mutator phenotype on ER2796 and ER3413. Consequently, for SMRT sequencing analysis of the activities of cloned MTases, we perform all plasmid construction and propagation steps in other (Dam$^+$) strains of *E. coli*, utilizing ER2796 or ER3413 only at the final step, namely isolation of total or plasmid DNA for sequencing and methylation analysis. In addition, we routinely check the sequencing reads against the MTase gene as reference to ensure the absence of introduced mutations. In our experience, such mutations are rare. In any case, given the large number of DNA MTases that recognize GATC both in bacterial, archaeal and phage genomes, a strict test of their specificity can only be conducted by cloning the genes into one of these two strains. Already a number of DNA MTases recognizing GATC have been successfully characterized using ER2796 and subtle variations in specificity involving the selectivity for flanking nucleotides have been successfully detected. We anticipate that these two strains will be invaluable as further interest develops in the methylation patterns that provide the epigenetic marks on bacterial and archaeal DNAs.

Supporting Information

S1 Fig. (A) Detection of G^{m6}ATC methylation status by methylation protection assay with MboI and DpnI, and ATGC^{m6}AT by NsiI, in gDNAs from 2796 (lanes 1), 2796 [pUC19:MEcoKIIwt] (lanes 2), 2796 [pUC19:MEcoKIIΔAgeI] (lanes 3) and S17-1 [pRE112:MEcoKIIΔAgeI] (lanes 4). (B) Colony PCR screening for the site-specific recombination event at the *yhdJ* locus in the first round of recombination (clones 9 and 18) and second round of recombination (clones 9–3, 9–7, 9–8, 18–3, 18–4, 18–6, 18–7, 18–8, 18–13, and 18–15). (C) Identification of wild type and mutant alleles of *yhdJ* by AgeI restriction digestion of colony PCR fragments.
(PDF)

S1 Table. List of sequence changes from the reference strain MG1655 to ER2796. (A) Insertions and deletions. (B) SNPs and substitutions.
(PDF)

S2 Table. List of ORFs with nonsynonymous changes from the reference strain MG1655 to ER2796. (A) Disrupted ORFs. (B) ORFs with missense mutations.
(PDF)

S3 Table. List of additional sequence changes from the reference strain MG1655 to ER3413 over and above those in ER2796.
(PDF)

Acknowledgments

The authors are very grateful to John Wertz (CGSC), Martin Marinus (University of Massachusetts Medical School, Worcester) and Ken Rudd (University of Florida) for discussion and advice. Martin Marinus, Nancy Kleckner and the CGSC provided strains during development of this strain.

Author Contributions

Conceived and designed the experiments: BPA EFM AF DRB RJR EAR. Performed the experiments: BPA EFM SA AF DRB. Analyzed the data: BPA EFM SA AF DRB RJR EAR. Contributed reagents/materials/analysis tools: EFM SA AF DRB. Wrote the paper: BPA EFM AF RJR EAR.

References

1. Lederberg J, Tatum EL. Gene recombination in Escherichia coli. Nature. 1946; 158(4016):558. Epub 1946/10/19. PMID: 21001945.
2. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, et al. The complete genome sequence of Escherichia coli K-12. Science. 1997; 277(5331):1453–62. Epub 1997/09/05. PMID: 9278503.
3. Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, et al. Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. Molecular systems biology. 2006; 2:2006 0007. Epub 2006/06/02. doi: 10.1038/msb4100049 PMID: 16738553; PubMed Central PMCID: PMC1681481.
4. Durfee T, Nelson R, Baldwin S, Plunkett G 3rd, Burland V, Mau B, et al. The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. Journal of bacteriology. 2008; 190(7):2597–606. Epub 2008/02/05. doi: 10.1128/JB.01695-07 PMID: 18245285; PubMed Central PMCID: PMC2293198.
5. Ferenci T, Zhou Z, Betteridge T, Ren Y, Liu Y, Feng L, et al. Genomic sequencing reveals regulatory mutations and recombinational events in the widely used MC4100 lineage of Escherichia coli K-12. Journal of bacteriology. 2009; 191(12):4025–9. Epub 2009/04/21. doi: 10.1128/JB.00118-09 PMID: 19376874; PubMed Central PMCID: PMC2698400.
6. Lobner-Olesen A, Skovgaard O, Marinus MG. Dam methylation: coordinating cellular processes. Current opinion in microbiology. 2005; 8(2):154–60. Epub 2005/04/02. doi: 10.1016/j.mib.2005.02.009 PMID: 15802246.
7. Lieb M, Bhagwat AS. Very short patch repair: reducing the cost of cytosine methylation. Molecular microbiology. 1996; 20(3):467–73. Epub 1996/05/01. PMID: 8736526.
8. Kahramanoglou C, Prieto AI, Khedkar S, Haase B, Gupta A, Benes V, et al. Genomics of DNA cytosine methylation in Escherichia coli reveals its role in stationary phase transcription. Nature communications. 2012; 3:886. Epub 2012/06/08. doi: 10.1038/ncomms1878 PMID: 22673913.
9. Militello KT, Simon RD, Qureshi M, Maines R, VanHorne ML, Hennick SM, et al. Conservation of Dcm-mediated cytosine DNA methylation in Escherichia coli. FEMS microbiology letters. 2012; 328(1):78–85. Epub 2011/12/14. doi: 10.1111/j.1574-6968.2011.02482.x PMID: 22150247.
10. Broadbent SE, Balbontin R, Casadesus J, Marinus MG, van der Woude M. YhdJ, a nonessential CcrM-like DNA methyltransferase of Escherichia coli and Salmonella enterica. Journal of bacteriology. 2007; 189(11):4325–7. Epub 2007/04/03. doi: 10.1128/JB.01854-06 PMID: 17400740; PubMed Central PMCID: PMC1913422.
11. Dryden DT, Cooper LP, Murray NE. Purification and characterization of the methyltransferase from the type 1 restriction and modification system of Escherichia coli K12. The Journal of biological chemistry. 1993; 268(18):13228–36. Epub 1993/06/25. PMID: 8514761.
12. Raleigh EA. Organization and function of the mcrBC genes of Escherichia coli K-12. Molecular microbiology. 1992; 6(9):1079–86. Epub 1992/05/01. PMID: 1316984.
13. Sibley MH, Raleigh EA. Cassette-like variation of restriction enzyme genes in Escherichia coli C and relatives. Nucleic acids research. 2004; 32(2):522–34. Epub 2004/01/28. doi: 10.1093/nar/gkh194 PMID: 14744977; PubMed Central PMCID: PMC373321.
14. Palmer BR, Marinus MG. The dam and dcm strains of Escherichia coli—a review. Gene. 1994; 143 (1):1–12. Epub 1994/05/27. PMID: 8200522.
15. Marinus MG, Morris NR. Biological function for 6-methyladenine residues in the DNA of Escherichia coli K12. Journal of molecular biology. 1974; 85(2):309–22. Epub 1974/05/15. PMID: 4600143.
16. Marinus MG, Konrad EB. Hyper-recombination in dam mutants of Escherichia coli K-12. Molecular & general genetics: MGG. 1976; 149(3):273–7. Epub 1976/12/22. PMID: 799245.
17. Fram RJ, Cusick PS, Wilson JM, Marinus MG. Mismatch repair of cis-diamminedichloroplatinum(II)-induced DNA damage. Molecular pharmacology. 1985; 28(1):51–5. Epub 1985/07/01. PMID: 3894930.
18. Karran P, Marinus MG. Mismatch correction at O6-methylguanine residues in E. coli DNA. Nature. 1982; 296(5860):868–9. Epub 1982/04/29. PMID: 7040986.

19. Oshima T, Wada C, Kawagoe Y, Ara T, Maeda M, Masuda Y, et al. Genome-wide analysis of deoxyadenosine methyltransferase-mediated control of gene expression in Escherichia coli. Molecular microbiology. 2002; 45(3):673–95. Epub 2002/07/26. PMID: 12139615.

20. Peterson KR, Wertman KF, Mount DW, Marinus MG. Viability of Escherichia coli K-12 DNA adenine methylase (dam) mutants requires increased expression of specific genes in the SOS regulon. Molecular & general genetics: MGG. 1985; 201(1):14–9. Epub 1985/01/01. PMID: 3932821.

21. Peterson KR, Mount DW. Analysis of the genetic requirements for viability of Escherichia coli K-12 DNA adenine methylase (dam) mutants. Journal of bacteriology. 1993; 175(22):7505–8. Epub 1993/11/01. PMID: 8226701; PubMed Central PMCID: PMC206901.

22. Marinus MG. Recombination is essential for viability of an Escherichia coli dam (DNA adenine methyltransferase) mutant. Journal of bacteriology. 2000; 182(2):463–8. Epub 2000/01/12. PMID: 10629194; PubMed Central PMCID: PMC94297.

23. Loenen WA, Raleigh EA. The other face of restriction: modification-dependent enzymes. Nucleic acids research. 2014; 42(1):56–69. Epub 2013/08/31. doi: 10.1093/nar/gkt747 PMID: 23990325; PubMed Central PMCID: PMC3874153.

24. MacNeil DJ. Characterization of a unique methyl-specific restriction system in Streptomyces avermitilis. Journal of bacteriology. 1988; 170(12):5607–12. Epub 1988/12/01. PMID: 3056907; PubMed Central PMCID: PMC211658.

25. Clark TA, Murray IA, Morgan RD, Kislyuk AO, Spittle KE, Boitano M, et al. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. Nucleic acids research. 2012; 40(4):e29. Epub 2011/12/14. doi: 10.1093/nar/gkr1146 PMID: 22156058; PubMed Central PMCID: PMC3287169.

26. Murray IA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, et al. The methylomes of six bacteria. Nucleic acids research. 2012; 40(22):11450–62. Epub 2012/10/05. doi: 10.1093/nar/gks891 PMID: 23034806; PubMed Central PMCID: PMC3526280.

27. Kong H, Lin LF, Porter N, Stickel S, Byrd D, Posfai J, et al. Functional analysis of putative restriction-modification system genes in the Helicobacter pylori J99 genome. Nucleic acids research. 2000; 28 (17):3216–23. Epub 2000/08/23. PMID: 10954588; PubMed Central PMCID: PMC110709.

28. Bachmann BJ. Derivations and genotypes of some mutant derivatives of Escherichia coli K-12. In: Neidhardt FC, editor. Escherichia coli and Salmonella: Cellular and Molecular Biology, 2nd ed. II. Washington, DC: ASM Press; 1996. p. 2460–88.

29. Brooks JE, Benner JS, Heiter DF, Silber KR, Sznyter LA, Jager-Quinton T, et al. Cloning the BamHI restriction modification system. Nucleic acids research. 1989; 17(3):979–97. Epub 1989/02/11. PMID: 2537955; PubMed Central PMCID: PMC331717.

30. Edwards RA, Keller LH, Schifferli DM. Improved allelic exchange vectors and their use to analyze 987P fimbria gene expression. Gene. 1998; 207(2):149–57. Epub 1998/03/25. PMID: 9511756.

31. Raleigh EA, Trimarchi R, Revel H. Genetic and physical mapping of the mcrA (rglA) and mcrB (rglB) loci of Escherichia coli K-12. Genetics. 1989; 122(2):279–96. Epub 1989/06/01. PMID: 2548920; PubMed Central PMCID: PMC1203701.

32. Kleckner N, Reichardt K, Botstein D. Inversions and deletions of the Salmonella chromosome generated by the translocatable tetracycline resistance element Tn10. Journal of molecular biology. 1979; 127 (1):89–115. Epub 1979/01/05. PMID: 370414.

33. Parker B, Marinus MG. A simple and rapid method to obtain substitution mutations in Escherichia coli: isolation of a dam deletion/insertion mutation. Gene. 1988; 73(2):531–5. Epub 1988/12/20. PMID: 2854098.

34. Marinus MG, Morris NR. Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12. Journal of bacteriology. 1973; 114(3):1143–50. Epub 1973/06/01. PMID: 4576399; PubMed Central PMCID: PMC285375.

35. Dar ME, Bhagwat AS. Mechanism of expression of DNA repair gene vsr, an Escherichia coli gene that overlaps the DNA cytosine methylase gene, dcm. Molecular microbiology. 1993; 9(4):823–33. Epub 1993/08/01. PMID: 7694036.

36. Garrick-Silversmith L, Hartman PE. Histidine-requiring mutants of Escherichia coli K12. Genetics. 1970; 66(2):231–44. Epub 1970/10/01. PMID: 4934198; PubMed Central PMCID: PMC1212491.

37. Zipser D, Newton A. The influence of deletions on polarity. Journal of molecular biology. 1967; 25 (3):567–9. Epub 1967/05/14. PMID: 5340699.

38. Zipser D, Zabell S, Rothman J, Grodzicker T, Wenk M. Fine structure of the gradient of polarity in the z gene of the lac operon of Escherichia coli. Journal of molecular biology. 1970; 49(1):251–4. Epub 1970/04/14. PMID: 4915862.

39. Zhou J, Rudd KE. EcoGene 3.0. Nucleic acids research. 2013; 41(Database issue):D613–24. Epub 2012/12/01. doi: 10.1093/nar/gks1235 PMID: 23197660; PubMed Central PMCID: PMC3531124.

40. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C, et al. EcoCyc: fusing model organism databases with systems biology. Nucleic acids research. 2013; 41(Database issue):D605–12. Epub 2012/11/13. doi: 10.1093/nar/gks1027 PMID: 23143106; PubMed Central PMCID: PMC3531154.

41. Beckwith J. lac: The Genetic System. In: Miller J, Reznikoff W, editors. The Operon. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1978. p. 11–30.

42. Inokuchi H, Kodaira M, Yamao F, Ozeki H. Identification of transfer RNA suppressors in Escherichia coli. II. Duplicate genes for tRNA2Gln. Journal of molecular biology. 1979; 132(4):663–77. Epub 1979/08/25. PMID: 160950.

43. Singaravelan B, Roshini BR, Munavar MH. Evidence that the supE44 mutation of Escherichia coli is an amber suppressor allele of glnX and that it also suppresses ochre and opal nonsense mutations. Journal of bacteriology. 2010; 192(22):6039–44. Epub 2010/09/14. doi: 10.1128/JB.00474-10 PMID: 20833812; PubMed Central PMCID: PMC2976463.

44. Raleigh EA, Kleckner N. Multiple IS10 rearrangements in Escherichia coli. Journal of molecular biology. 1984; 173(4):437–61. Epub 1984/03/15. PMID: 6323719.

45. Zipser D, Newton A. The influence of deletions on polarity. Journal of Molecular Biology. 1967; 25 (3):567. PMID: 15768525299899389946related:-ldDDDcO1doJ.

46. Zipser D, Zabell S, Rothman J, Grodzicker T, Wenk M. Fine structure of the gradient of polarity in the z gene of the lac operon of Escherichia coli. Journal of Molecular Biology. 1970; 49(1):251–4. PMID: 4915862.

47. Beckwith JR. lac: The Genetic System. In: Miller JH, Reznikoff WS, editors. The Operon. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory; 1978. p. 11–30.

48. Parker B, Marinus MG. A simple and rapid method to obtain substitution mutations in *Escherichia coli*: isolation of a *dam* deletion/insertion mutation. Gene. 1988; 73:531–5. PMID: 2854098

49. Singer M, Baker TA, Schnitzler G, Deischel SM, Goel M, Dove W, et al. Collection of strains containing genetically linked alternating antibiotic resistance elements for genetic mapping of *Escherichia coli*. Microbiol Rev. 1989; 53:1–24. PMID: 2540407

50. Nichols BP, Shafiq O, Meiners V. Sequence analysis of Tn10 insertion sites in a collection of Escherichia coli strains used for genetic mapping and strain construction. Journal of Bacteriology. 1998; 180 (23):6408–11. PMID: 9829956; PubMed Central PMCID: PMCPMC107733.

51. Kusano K, Sakagami K, Yokochi T, Naito T, Tokinaga Y, Ueda E, et al. A new type of illegitimate recombination is dependent on restriction and homologous interaction. J Bacteriol. 1997; 179(17):5380–90. PMID: 9286991

52. Dutra BE, Sutera VA Jr, Lovett ST. RecA-independent recombination is efficient but limited by exonucleases. Proc Natl Acad Sci U S A. 2007; 104(1):216–21. PMID: 17182742

53. Raghavan R, Sage A, Ochman H. Genome-wide identification of transcription start sites yields a novel thermosensing RNA and new cyclic AMP receptor protein-regulated genes in Escherichia coli. Journal of Bacteriology. 2011; 193(11):2871–4. doi: 10.1128/JB.00398-11 PMID: 21460078; PubMed Central PMCID: PMCPMC3133129.

54. Barker CS, Pruss BM, Matsumura P. Increased motility of Escherichia coli by insertion sequence element integration into the regulatory region of the flhD operon. Journal of bacteriology. 2004; 186 (22):7529–37. Epub 2004/11/02. doi: 10.1128/JB.186.22.7529–7537.2004 PMID: 15516564; PubMed Central PMCID: PMC524886.

55. Harayama S, Bollinger J, Iino T, Hazelbauer GL. Characterization of the mgl operon of Escherichia coli by transposon mutagenesis and molecular cloning. Journal of bacteriology. 1983; 153(1):408–15. Epub 1983/01/01. PMID: 6294056; PubMed Central PMCID: PMC217387.

56. Parker BW, Schwessinger EA, Jakob U, Gray MJ. The RclR protein is a reactive chlorine-specific transcription factor in Escherichia coli. The Journal of biological chemistry. 2013; 288(45):32574–84. Epub 2013/10/01. doi: 10.1074/jbc.M113.503516 PMID: 24078635; PubMed Central PMCID: PMC3820890.

57. Santoyo G, Martinez-Salazar JM, Rodriguez C, Romero D. Gene Conversion Tracts Associated with Crossovers in Rhizobium etli. Journal of Bacteriology. 2005; 187(12):4116–26. doi: 10.1128/JB.187.12.4116–4126.2005 PMID: 15937174

58. Stewart FJ, Cavanaugh CM. Intragenomic variation and evolution of the internal transcribed spacer of the rRNA operon in bacteria. Journal of Molecular Evolution. 2007; 65(1):44–67. PMID: 17568983

59. Talarico S, Whitefield SE, Fero J, Haas R, Salama NR. Regulation of Helicobacter pylori adherence by gene conversion. Molecular Microbiology. 2012; 84(6):1050–61. doi: 10.1111/j.1365-2958.2012.08073.x PMID: 22519812

60. Casjens S. Prophages and bacterial genomics: what have we learned so far? Molecular microbiology. 2003; 49(2):277–300. Epub 2003/07/31. PMID: 12886937.

61. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, et al. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nature biotechnology. 2012; 30(12):1232–9. Epub 2012/11/10. doi: 10.1038/nbt.2432 PMID: 23138224; PubMed Central PMCID: PMC3879109.

62. Krebes J, Morgan RD, Bunk B, Sproer C, Luong K, Parusel R, et al. The complex methylome of the human gastric pathogen Helicobacter pylori. Nucleic acids research. 2014; 42(4):2415–32. Epub 2013/12/05. doi: 10.1093/nar/gkt1201 PMID: 24302578; PubMed Central PMCID: PMC3936762.

63. Laehnemann D, Pena-Miller R, Rosenstiel P, Beardmore R, Jansen G, Schulenburg H. Genomics of rapid adaptation to antibiotics: convergent evolution and scalable sequence amplification. Genome biology and evolution. 2014; 6(6):1287–301. Epub 2014/05/23. doi: 10.1093/gbe/evu106 PMID: 24850796; PubMed Central PMCID: PMC4079197.

64. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, et al. Escherichia coli K-12: a cooperatively developed annotation snapshot—2005. Nucleic acids research. 2006; 34(1):1–9. Epub 2006/01/07. doi: 10.1093/nar/gkj405 PMID: 16397293; PubMed Central PMCID: PMC1325200.

65. Suzuki S, Ono N, Furusawa C, Ying BW, Yomo T. Comparison of sequence reads obtained from three next-generation sequencing platforms. PloS one. 2011; 6(5):e19534. Epub 2011/05/26. doi: 10.1371/journal.pone.0019534 PMID: 21611185; PubMed Central PMCID: PMC3096631.

66. Fridman O, Goldberg A, Ronin I, Shoresh N, Balaban NQ. Optimization of lag time underlies antibiotic tolerance in evolved bacterial populations. Nature. 2014; 513(7518):418–21. Epub 2014/07/22. doi: 10.1038/nature13469 PMID: 25043002.

67. Jeong H, Barbe V, Lee CH, Vallenet D, Yu DS, Choi SH, et al. Genome sequences of Escherichia coli B strains REL606 and BL21(DE3). Journal of molecular biology. 2009; 394(4):644–52. Epub 2009/09/30. doi: 10.1016/j.jmb.2009.09.052 PMID: 19786035.

68. Archer CT, Kim JF, Jeong H, Park JH, Vickers CE, Lee SY, et al. The genome sequence of E. coli W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of E. coli. BMC genomics. 2011; 12:9. Epub 2011/01/07. doi: 10.1186/1471-2164-12-9 PMID: 21208457; PubMed Central PMCID: PMC3032704.

69. Turner PC, Yomano LP, Jarboe LR, York SW, Baggett CL, Moritz BE, et al. Optical mapping and sequencing of the Escherichia coli KO11 genome reveal extensive chromosomal rearrangements, and multiple tandem copies of the Zymomonas mobilis pdc and adhB genes. Journal of industrial microbiology & biotechnology. 2012; 39(4):629–39. Epub 2011/11/15. doi: 10.1007/s10295-011-1052-2 PMID: 22075923.

70. Surette MG, Miller MB, Bassler BL. Quorum sensing in Escherichia coli, Salmonella typhimurium, and Vibrio harveyi: a new family of genes responsible for autoinducer production. Proceedings of the National Academy of Sciences of the United States of America. 1999; 96(4):1639–44. Epub 1999/02/17. PMID: 9990077; PubMed Central PMCID: PMC15544.

71. Lee H, Popodi E, Tang H, Foster PL. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. Proceedings of the National Academy of Sciences. 2012; 109(41):E2774–83. doi: 10.1073/pnas.1210309109 PMID: 22991466; PubMed Central PMCID: PMCPMC3478608.

72. Hanahan D. Studies on transformation of Escherichia coli with plasmids. Journal of Molecular Biology. 1983; 166(4):557–80. PMID: 6345791.

73. Low B. Formation of merodiploids in matings with a class of Rec- recipient strains of Escherichia coli K12. Proceedings of the National Academy of Sciences of the United States of America. 1968; 60 (1):160–7. PMID: 4873517; PubMed Central PMCID: PMCPMC539096.

74. Bachmann BJ. Derivations and genotypes of some mutant derivatives of *Escherichia coli* K-12. In: Neidhardt FC, editor. *Escherichia coli* and *Salmonella typhimurium*: Cellular and Molecular Biology. 2. Washington, D.C.: American Society for Microbiology; 1987. p. 807–76.

75. Visick JE, Clarke S. RpoS- and OxyR-independent induction of HPI catalase at stationary phase in Escherichia coli and identification of rpoS mutations in common laboratory strains. Journal of bacteriology. 1997; 179(13):4158–63. Epub 1997/07/01. PMID: 9209028; PubMed Central PMCID: PMC179234.

76. Jensen KF. The Escherichia coli K-12 "wild types" W3110 and MG1655 have an rph frameshift mutation that leads to pyrimidine starvation due to low pyrE expression levels. Journal of bacteriology. 1993; 175(11):3401–7. Epub 1993/06/01. PMID: 8501045; PubMed Central PMCID: PMC204738.