J Ind Microbiol Biotechnol (2012) 39:629–639
DOI 10.1007/s10295-011-1052-2

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

Peter C. Turner · Lorraine P. Yomano · Laura R. Jarboe · Sean W. York · Christy L. Baggett · Brélan E. Moritz · Emily B. Zentz · K. T. Shanmugam · Lonnie O. Ingram

Received: 23 August 2011 / Accepted: 19 October 2011 / Published online: 11 November 2011

Abstract *Escherichia coli* KO11 (ATCC 55124) was engineered in 1990 to produce ethanol by chromosomal insertion of the *Zymomonas mobilis pdc* and *adhB* genes into *E. coli* W (ATCC 9637). KO11FL, our current laboratory version of KO11, and its parent *E. coli* W were sequenced, and contigs assembled into genomic sequences using optical *Nco*I restriction maps as templates. *E. coli* W contained plasmids pRK1 (102.5 kb) and pRK2 (5.4 kb), but KO11FL only contained pRK2. KO11FL optical maps made with *Afl*II and with *Bam*HI showed a tandem repeat region, consisting of at least 20 copies of a 10-kb unit. The repeat region was located at the insertion site for the *pdc*, *adhB*, and chloramphenicol-resistance genes. Sequence coverage of these genes was about 25-fold higher than average, consistent with amplification of the foreign genes that were inserted as circularized DNA. Selection for higher levels of chloramphenicol resistance originally produced strains with higher *pdc* and *adhB* expression, and hence improved fermentation performance, by increasing the gene copy number. Sequence data for an earlier version of KO11, ATCC 55124, indicated that multiple copies of *pdc adhB* were present. Comparison of the W and KO11FL genomes showed large inversions and deletions in KO11FL, mostly enabled by IS*10*, which is absent from W but present at 30 sites in KO11FL. The early KO11 strain ATCC 55124 had no rearrangements, contained only one IS*10*, and lacked most accumulated single nucleotide polymorphisms (SNPs) present in KO11FL. Despite rearrangements and SNPs in KO11FL, fermentation performance was equal to that of ATCC 55124.

Keywords Optical mapping · Fermentation · Ethanol · *Escherichia coli* · Genome sequencing

This article is based on a presentation at the 33rd Symposium on Biotechnology for Fuels and Chemicals.

Electronic supplementary material The online version of this article (doi:10.1007/s10295-011-1052-2) contains supplementary material, which is available to authorized users.

P. C. Turner · L. P. Yomano · L. R. Jarboe · S. W. York · C. L. Baggett · B. E. Moritz · K. T. Shanmugam · L. O. Ingram (✉)
Department of Microbiology and Cell Biology,
University of Florida, Box 110700, Gainesville, FL 32611, USA
e-mail: Ingram@ufl.edu

Present Address:
L. R. Jarboe
Chemical and Biological Engineering,
Biorenewables Research Laboratory,
Iowa State University, Ames, IA, USA

E. B. Zentz
OpGen Inc., Gaithersburg, MD, USA

Introduction

Escherichia coli serves as a workhorse biocatalyst for the production of biochemicals and heterologous proteins. The advantages of *E. coli* include the availability of excellent genetic tools, an extensive background knowledge of physiology, and the recognition that commensal strains are safe hosts for biotechnology applications. *E. coli* W (ATCC 9637) is a particularly fast-growing strain that was originally isolated and deposited at the American Type Culture Collection (ATCC) by S. A. Waksman. In addition to faster growth than K-12 strains, *E. coli* W utilizes a wide range of carbon sources including sucrose [22], and can degrade many aromatic compounds [11]. ATCC 11105 [4], a derivative of *E. coli* W, has been used industrially to produce penicillin G acylase [28, 31]. *E. coli* W also served as the

parent for patented biocatalysts that produce ethanol, lactate, alanine, and succinate [13, 18].

E. coli W was used in one of the earliest examples of metabolic engineering. The *Zymomonas mobilis* genes encoding a pyruvate to ethanol pathway (*pdc, adhB*) were assembled into an artificial operon for the production of ethanol, the PET (pyruvate to ethanol) operon [16]. Using an adjacent *cat* gene with its own promoter for selection, this cassette was integrated into the *pflB* region of the *E. coli* W chromosome [25]. A circularized 8.6-kb *Sal*I fragment from pLOI1510 containing the *pdc–adhB–cat* genes flanked by portions of *pflB* was introduced into wild-type *E. coli*, and chloramphenicol-resistant recombinants selected. Expression of *pdc–adhB* was driven by the native *pflB* promoter after integration. Subsequently, a fumarate reductase (*frd*) mutation was added by conjugation from a K-12 strain to produce strain KO11 (ATCC 55124) [25]. The transfer of the *frd* mutation was achieved by selection for the closely linked *zjd*::Tn*10* mutation conferring tetracycline resistance. Strain KO11 was shown to efficiently ferment all hexose and pentose sugars that comprise lignocellulose into ethanol. Following construction of KO11 and the early patent deposit at the ATCC (ATCC 55124), this strain was serially transferred in our laboratory for over 20 years in liquid medium and on plates using either Luria–Bertani (LB) nutrients or corn steep liquor (CSL) media containing various levels of chloramphenicol [37].

Derivatives of KO11 have been developed by a combination of genetic engineering and metabolic evolution (serial transfers with growth-based selection), resulting in improved ethanologens that can grow well in minimal medium [38, 39] and produce a variety of different biochemicals as dominant fermentation products including L-alanine [42], succinate [40, 43], L- and D-lactate [36], and L-malate [41]. In KO11 and its derivatives, growth is obligately linked to the production of a specific fermentation product (oxidation of NADH), providing a powerful basis for strain improvement.

Experiments using adaptive laboratory evolution to select for strains with desired characteristics were pioneered by the Lenski group [23]. Adaptive laboratory evolution together with whole genome sequencing of *E. coli* [2, 15] has been highly successful in identifying a small number of mutations that are crucial for improved growth under specified conditions. As the basis for an analogous investigation of chromosomal changes in KO11-derived biocatalysts, we have sequenced the genomes of the parent *E. coli* W, and our laboratory strain of KO11, denoted KO11FL after 20 years of serial transfers. During the course of our research, the complete genomic sequence of the parent strain *E. coli* W was published [1], and the Department of Energy (DOE) Joint Genome Institute deposited 88 sequence contigs of the *E. coli* W chromosome (GenBank accession NZ_AEDF00000000). DOE researchers also deposited a sequence for the early version of KO11 (ATCC 55124), denoted EKO11(GenBank accession CP002516.1) that differs considerably from our current laboratory strain (KO11FL).

In this study, we report a comparison of KO11FL to the parent *E. coli* W and the earlier version of KO11 from DOE JGI (EKO11; ATCC 55124). Serial cultivation and selection in the laboratory have resulted in extensive IS*10* transposition, several gene deletions and rearrangements, and the development of approximately 25 tandem repeats of the ethanol cassette (*Z. mobilis pdc, adhB,* and *cat*) in KO11FL.

Materials and methods

Genome sequencing and assembly

Genomic DNA was submitted to Integrated Genomics (Chicago, IL) for shotgun 454 sequencing. Reads were assembled with Newbler (Roche). To close gaps between contigs, Sequencher 4.1 (Gene Codes Corporation) was used to assemble Sanger sequencing reads from PCR products spanning neighboring contigs. Next generation sequencing data using Illumina paired-end short-read technology was provided by the Tufts University Core Facility (Boston, MA). These were assembled using Geneious software [12].

Generation and analysis of optical maps

Strains *E. coli* W (ATCC 9637) and KO11FL were grown on LB-agar plates and submitted to OpGen (Gaithersburg, MD) for optical mapping [29]. Briefly, high molecular weight DNA was extracted directly from the cells, and single DNA molecules were immobilized along an optical surface by flowing the DNA through microfluidic channels. Following digestion with a restriction enzyme, the DNA was stained with a fluorescent dye, and the lengths of the fragments measured by fluorescence intensity. Overlapping single molecule maps were assembled using Optical Map Assembler software (OpGen) to create a circular map spanning the entire genome with coverage of approximately 30-fold. MapSolver™ Software was used to place the predicted restriction maps of large contigs (>50 kb) resulting from the Newbler assembly on the optical map scaffold, enabling gaps between contigs to be predicted and filled by PCR.

Annotation

Initial annotation was provided by the xBASE bacterial genome annotation service [7]. The Prokaryotic Genomes

Automatic Annotation Pipeline (PGAAP) service at the National Center for Biotechnology Information (NCBI) was used as a secondary source. Annotation was completed manually by using BLAST at NCBI, EcoCyc [19], and IS Finder [30]. Fully assembled genomic and plasmid sequences for *E. coli* W (WFL) and KO11FL have been deposited at GenBank. The accession numbers for *E. coli* W, W plasmid pRK1, and W plasmid pRK2 are CP002967, CP002968, and CP002969, respectively. Accession numbers for *E. coli* KO11FL and plasmid KO11FLpRK2 are CP002970 and CP002971, respectively. Note plasmid pRK1 was absent in KO11FL.

Comparison of bacterial genomes

The progressive Mauve algorithm [10] was used both as a stand-alone program and as a plugin within Geneious to compare the *E. coli* W and KO11 genomes. Progressive Mauve was suitable for assessing rearrangements, large-scale insertions, and large deletions. For reliably detecting and enumerating small-scale differences [single nucleotide polymorphisms (SNPs) and small insertions or deletions (indels)] between strains, the sequencing reads for one strain were assembled against the template of the other strain with Geneious, and the "Find Variations/SNPs" tool was used to detect polymorphisms, with a minimum variant frequency of 0.9, minimum coverage of 5, and with adjacent variations merged. Reciprocal assemblies were done where possible to confirm differences between strains.

Fermentation conditions

Fermentation performance of KO11 (patent deposit; ATCC 55124) was compared to our current laboratory strain, KO11FL, using complex (Luria broth) and defined mineral salts (NBS; [5]) media. Frozen stock cultures (−80°C) were grown overnight on plates containing 5% xylose using both media. Seed cultures were grown in each respective medium (250-ml flask containing 100 ml broth; 4–6 h incubation, 37°C, 150 rpm). Small fermenters (500 ml containing 350 ml broth, 14% xylose) were inoculated at an initial density of 16 mg dcw l^{-1}. Growth and ethanol were monitored [24] during incubation (37°C, 150 rpm). 2 N KOH was added automatically to maintain pH 6.5.

Results and discussion

Sequencing of *E. coli* W (ATCC 9637) and comparison with other *E. coli* strains

The sequence of the *E. coli* W genome was assembled from 454 data using an optical map as a guide for contig placement [21]. Subsequently W was resequenced using the Illumina platform. Figure 1 shows the workflow for sequencing, optical mapping, and annotation. The W genome spans 4,897,452 bases and encodes 4,626 proteins, of which 801 are not present in *E. coli* K-12 strain MG1655 [3]. *E. coli* W was originally reported to carry three plasmids [31], but only two were found: pRK1 (102.5 kb) and pRK2 (5.4 kb). pRK1 belongs to the incompatibility group I1 (IncI1), and is similar to the 100-kb plasmid pSE11-1 from *E. coli* SE11 [26]. pRK2 is 99% identical to pSE11-5 (5.4 kb) from the same strain. Our sequence for *E. coli* W (denoted WFL) is in excellent agreement with the sequence of Archer et al. [1] with one exception: the G segment of bacteriophage Mu was in the opposite orientation in their sequence relative to ours. In spite of the many differences between *E. coli* W and *E. coli* K-12, the W genome is colinear with the K-12 genome, as are most *E. coli* genomes that have been sequenced.

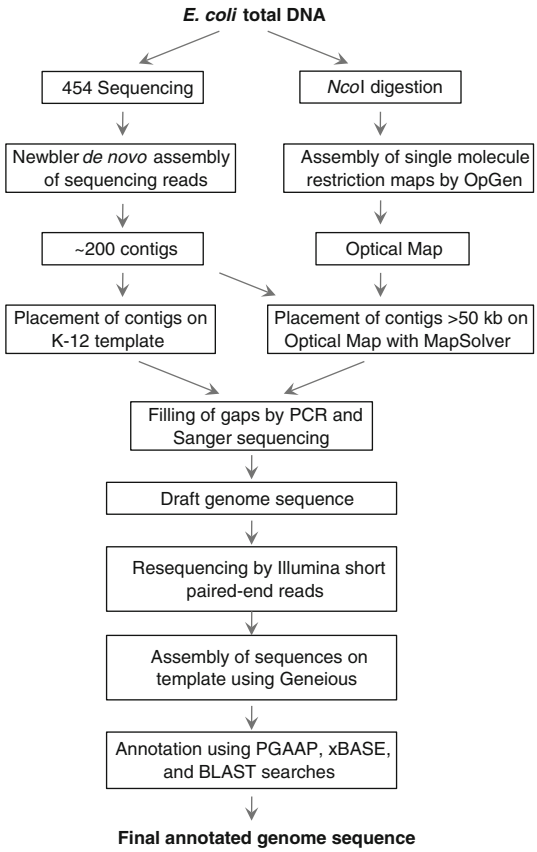

Fig. 1 Workflow for sequencing, optical mapping, sequence assembly, and annotation

Restriction maps of entire chromosomes provide a broad and minimally biased basis for phylogenetic comparisons of *E. coli* strains. The relationship of *E. coli* W was investigated by comparing the *Nco*I optical map with the predicted *Nco*I maps of all available complete *E. coli* genomes using MapSolver™ software (OpGen) and the UPGMA clustering

method. From this comparison, *E. coli* W (Fig. 2) is most closely related to the commensal strains IAI1 [34] and SE11 [26] in the B1 phylogenetic group. K-12 strains cluster with the A group, a sister group of B1, and pathogenic strains are found in groups D, E, and B2 [14]. Very similar trees (not shown) were obtained when the *Afl*II and *Bam*HI maps for *E. coli* W predicted in silico were compared with in silico maps for the other *E. coli* strains. The overall grouping of strains and the closeness of *E. coli* W and SE11 (Fig. 2) agree with the tree created by more traditional methods using the sequences of selected housekeeping genes [1]. Using genomic restriction maps to construct phylogenies therefore appears to be a reliable method. The similarities between the optical maps of the W and SE11 chromosomes and between the sequences of plasmids pRK1/pSE11-1 and pRK2/pSE11-5 suggest that SE11 is the closest known relative of *E. coli* W.

Sequencing and optical mapping of KO11

The strategy used to sequence *E. coli* W (Fig. 1) was applied to strain KO11FL. Assembling contigs for KO11FL following Roche 454 sequencing proved a major challenge. The initial assembly of the 454 reads with Newbler produced a total of 192 contigs ranging in size from 103 to 393,989 bp. Many gaps were filled manually by PCR amplification and sequencing, including gaps that were bordered by repetitive DNA such as rRNA operons, *rhs* genes, and mobile elements including IS*621* (6 copies), IS*609* (4 copies), and IS*3* (3 copies). Some gaps resulted from the presence of multiple copies of the IS*10* element in KO11FL, although IS*10* was not present in the parent strain W. In some cases, adjoining contig ends were predicted by aligning KO11FL contigs using either W or K-12 as template, but the gaps could not be readily bridged by PCR, suggesting that the gaps were too large, or that the contig ends were adjacent in W and K-12 but not in KO11. To solve this problem, an optical map for KO11FL was constructed after digestion with *Nco*I. Aligning the contigs with the optical map allowed ends of neighboring contigs to be identified, and also provided an estimate of length for missing sequences [21]. This strategy was used to fully assemble the genomic sequence of KO11FL. The complete sequence obtained from Roche 454 pyrosequencing was confirmed using Illumina technology. The final genomic sequence for KO11FL was 5,021,182 bases in length. Plasmid pRK2 is present (5.4 kb), differing from the pRK2 in W by a single base. The larger plasmid, pRK1, was absent in KO11FL.

A comparison of the optical maps of W and KO11FL revealed large-scale rearrangements, including inversions and deletions. Although W and K-12 were colinear, this was not the case for KO11. In addition, an unusual 160-kb

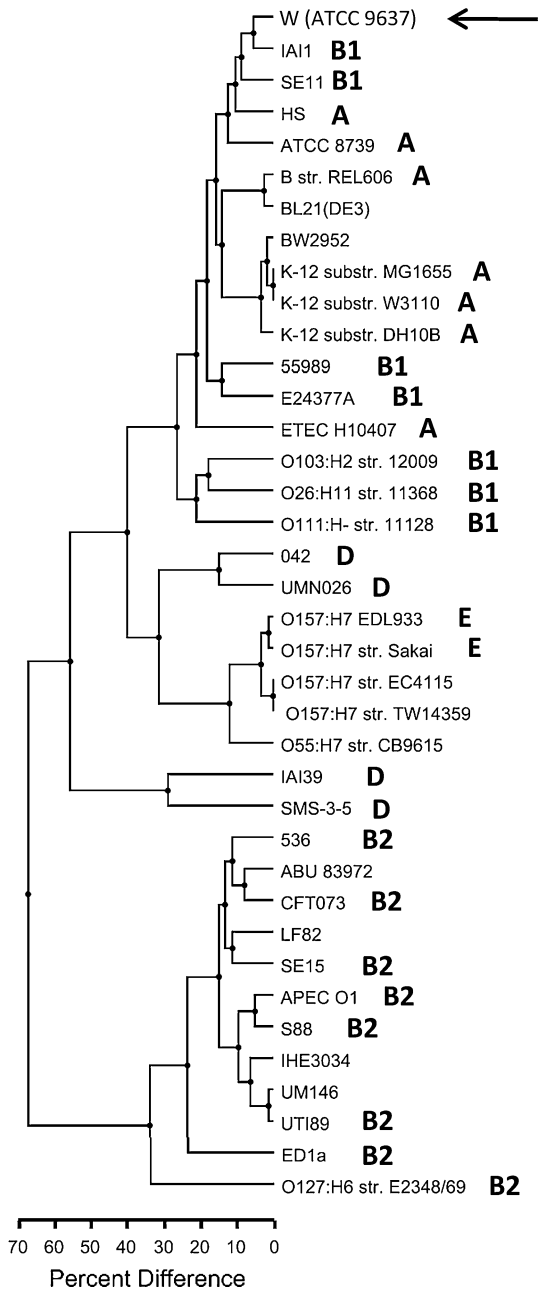

Fig. 2 Phylogeny of *E. coli* genomes based on MapSolver™ alignment of optical *Nco*I restriction maps. The *Nco*I restriction map for W was clustered with in silico restriction maps for available sequenced *E. coli* strains. The phylogenetic group that each strain belongs to is shown where known

region was found in the KO11 *Nco*I map that was not present in W (Fig. 3) or in any known *E. coli* strain in the databases. Additional optical mapping of the KO11 genome with *Afl*II and *Bam*HI confirmed the existence of this region, and estimated that it consisted of 34 and 24 repeats, respectively, of a 10-kb sequence (Fig. 3). For the *Afl*II map, each 10-kb region was made up of a 7.6-kb and a 2.5-kb fragment. For the *Bam*HI map, fragments were approximately 5.5 and 4.8 kb. The variation in the apparent

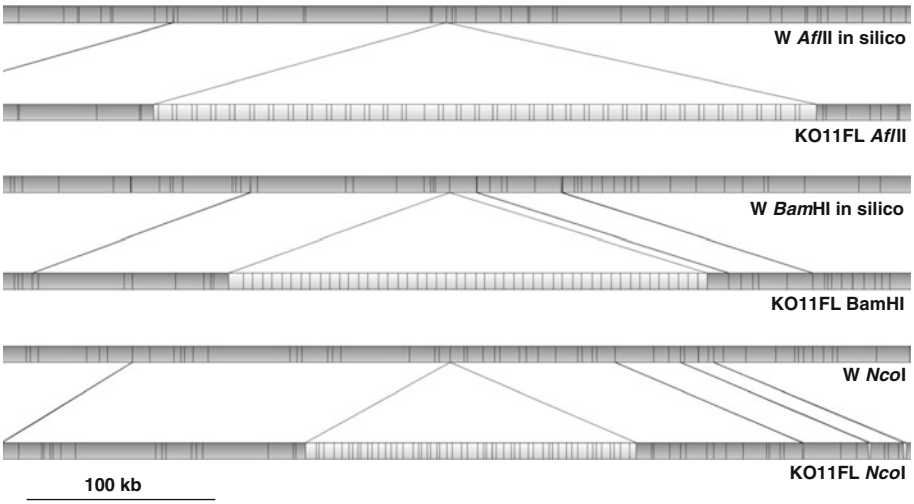

Fig. 3 Optical maps for W (ATCC 9637) and KO11FL. The whole genome optical maps were aligned using MapSolver™ (OpGen), and a 360-kb segment spanning the unusual region in KO11FL is shown. The *top two lines* are *Afl*II maps, the *middle two lines* are *Bam*HI, and the *lower two lines* are *Nco*I. The fragments are *shaded* where they correspond between the two strains, and *white* when they are present in only one strain. Note the regular repeat structure of the inserted region evident from the *Afl*II and *Bam*HI maps

extent of the repeat region with different restriction enzymes may reflect difficulties in aligning tandem repeats with the OpGen assembly software, in much the same way that DNA sequence assembly programs sometimes fail to correctly assemble repetitive sequences. Misassembly of the DNA fragment data could result in under- or overestimation of the number of repeats. By looking at the *Afl*II data for individual DNAs spanning the tandem repeat region, multiple molecules were seen with approximately 20 copies of the 10-kb repeat. Although the number of repeats need not be constant in a cell population, the average number of repeats in the chromosome was at least 20. Contigs on either side of the "repeat region" flanked the *pflB* region that was used as the insertion site of the PET operon containing the *Z. mobilis pdc adhB* and *cat* genes. However, the 10-kb length of each repeat exceeded the 8.6-kb size of the original PET construct (circularized *Sal*I fragment of pLOI510 [25]) by approximately 1.4 kb. This difference was accounted for by the presence in the repeat of the insertion sequence IS*10* (1.3 kb).

Sequence coverage using Illumina data (Fig. 4) provided an independent estimate of copy number for the repeat sequence. The *pdc*, *adhB*, and *cat* genes had approximately 25-fold higher coverage than the regions to either side, and regions elsewhere in the genome. A similar higher coverage (approximately 21-fold) for the *pdc*, *adhB*, and *cat* genes was seen with the original 454 sequence data (not shown). There is precedent for detecting large genomic duplications in *E. coli* by higher sequence coverage [9]. No other regions in the KO11 genome were found to be at higher than expected coverage. The optical mapping and sequence coverage data are consistent with tandem duplication of the *pdc*, *adhB*, and *cat* genes, which were introduced as a circular DNA [25]. Recombination between monomer circles could have created dimers that were integrated into the chromosome by single crossover events, creating tandem repeats. We propose that selection for increased levels of chloramphenicol resistance [25] led to amplification of the inserted genes to 20–25 copies, thereby increasing gene expression and improving fermentation. Routine propagation of KO11 in our laboratory involved growth on LB plates with chloramphenicol alternating between levels of 50 and 600 μg/ml. Plates were incubated under argon in sealed containers, in the presence of trace amounts of oxygen. Each time, selections were made for large colonies to maintain vigor and high expression of ethanol genes.

The DOE JGI sequence for an early version of KO11 (ATCC 55124) was deposited as GenBank accession CP002516.1, and the genes prefixed with EKO11. The sequence includes versions of plasmids pRK1 (pEKO1101, GenBank accession CP002517) and pRK2 (pEKO1102, GenBank accession CP002518). Our current version of KO11 (KO11FL, GenBank accession CP002970) is quite different (Table 1). Also, the JGI genome sequence for KO11 was reported to contain only one copy of the inserted *pdc*, *adhB*, and *cat* genes. However, we were able to estimate the sequence coverage for each of the 127 KO11 (ATCC 55124) contigs deposited in GenBank before the genome assembly was completed. Coverage was calculated from the contig length and the number of sequencing reads for each contig. Average coverage of contigs #123, 268, 276, and 277 that together span the *pdc–adhB–cat* region was respectively 16.4-, 15.4-, 16.5-, and 19.2-fold higher than for the surrounding regions and the genome as a whole. Thus amplification of the *pdc–adhB–cat* genes occurred early, prior to the patent deposit of KO11 (ATCC 55124).

Rearrangements and IS*10* elements in KO11

The genomes of W and KO11FL were aligned using progressive Mauve [10], an algorithm that is capable of

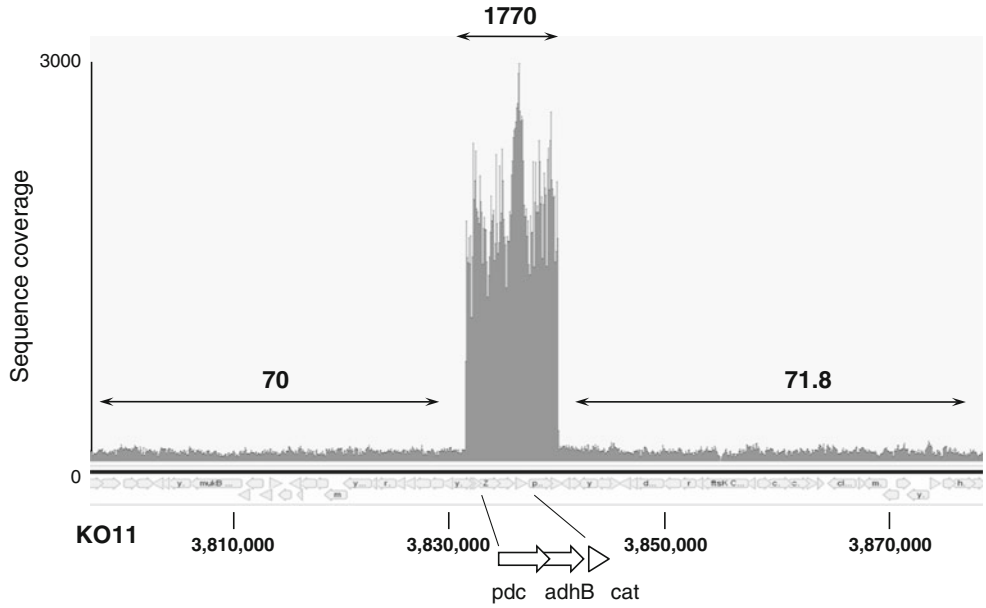

Fig. 4 Sequence coverage over the KO11FL region including the *Z. mobilis pdc* and *adhB* genes. A template containing only one copy of the foreign inserted genes was used for this assembly of Illumina short-read paired-end data. IS*10* was absent from the template. *Numbers* at the *bottom* are coordinates on the template. The *numbers over* the *arrows* show the average sequence coverage for the indicated regions. Note the high sequence coverage (1,770) for the *pdc-adhB-cat* region, approximately 25-fold higher than coverage to the *left* and *right*

Table 1 Comparison of KO11 (ATCC 55124) and KO11FL

	KO11 (ATCC 55124)	KO11FL
Plasmid pRK1 (102.5 kb)	Present	Absent
Plasmid pRK2 (5.4 kb)	Present	Present
Estimated copy number of *pdc–adhB–cat* genes	~19	~21-25
IS*10* insertion sites	1	30
IS*621*	5	6
IS*609*	4	4
IS*Ec31*	2[a]	2
IS*3*	3	3
IS*150*	1[a]	1
Major rearrangements relative to W	0	4
Total DNA deleted relative to W (kb)	0	122
Total SNPs and indels relative to W	1,382	1,473
SNPs in K-12-derived region relative to W	1,369	1,370
SNPs outside K-12 region relative to W	13	103

Numbers listed for transposons are chromosomal insertion sites for IS*10*, and chromosomal copy number for other insertion sequences

[a] pEKO1101, the pRK1 equivalent in KO11(ATCC 55124), contains one additional copy of each of IS*150* and IS*Ec31*

handling the large-scale rearrangements found in many bacterial chromosomes. In order for alignment to occur, the genomes of W and KO11FL were split by progressive Mauve into eight locally colinear blocks (LCBs).

One major inversion between LCB1 and LCB8 at the ends of the W genome (Fig. 5) was apparently caused by a duplication of prophage Mu sequences, as these were present in inverted orientation at the breakpoints in KO11FL. Note that W contains only one copy of prophage Mu. All of the other observed inversions and deletions were caused by recombination between the insertion sequence IS*10*. The genome of KO11FL contained a large number of IS*10* insertions (Table 1), in contrast with W, which lacks IS*10*. Outside the repeat region in KO11FL, there are 29 intact IS*10* s (#1–26, 28–30, each 1,337 bp), and one partial IS*10* (#22A) which is immediately adjacent to IS*10* #22. In addition, a single copy of IS*10* (#27) is present in each of the estimated 25 copies of the tandem repeat. The most likely step for the introduction of IS*10* into KO11 involved P1 transduction of an *frd* deletion from *E. coli* K-12 strain DW12 [*zjd*::Tn*10* Δ(*frdABCD*)] into the Hfr strain KL282 by selection for tetracycline resistance conferred by the linked Tn*10* insertion [25]. From there *zjd*::Tn*10* and Δ(*frdABCD*) were transferred to KO4, an intermediate in the construction of KO11, by conjugation. Tn*10* contains two copies of IS*10* flanking the tet^R gene. Although loss of tetracycline resistance was selected during strain construction, IS*10* transposition occurred at some point.

The sites of the IS*10* insertions and the functions of genes inactivated by IS*10* insertion or adjacent to an insertion site are listed in Table 2. Transposition of IS*10* occurs randomly and is accompanied by duplication of a 9-bp sequence within the target gene [20]. Each IS*10* should lie between two

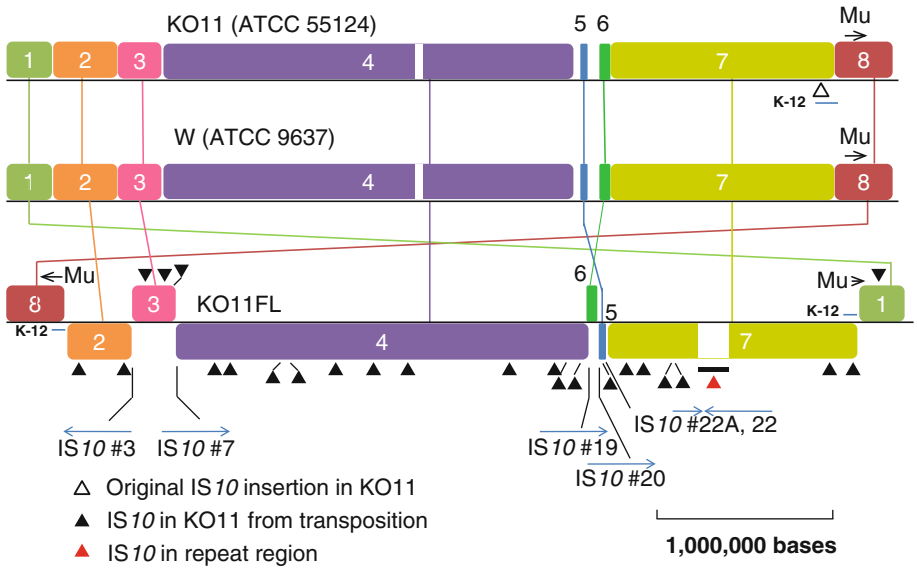

Fig. 5 Alignment of W (ATCC 9637) and KO11FL genomes by Mauve. The genome of *E. coli* W (*center*) is the reference genome, with the early KO11 strain ATCC 55124 (GenBank accession CP002516.1) shown *above* and KO11FL (this work) *below*. Note that the genomes of W (ATCC 9637) and KO11-JGI are in reverse orientation. Locally colinear blocks (LCBs) of the same *color* are conserved in the two strains, and connected by *lines*. Blocks in KO11FL are in the same orientation as in W if they appear *above* the *black line*, and in the opposite orientation if they appear *below*. *White areas* indicate regions that are present in one strain but not the other. The tandem repeat region in KO11FL is represented by the *black bar below* the *white area* in the KO11FL genome. IS*10* insertions that are not associated with chromosomal rearrangements are shown by *black triangles*. The original IS*10* insertion in KO11 (ATCC 55124) is shown by the *open black triangle* within the region introduced from K-12 (*short blue bar*). IS*10* s within the tandem repeat in KO11FL are represented by a *single red triangle*. IS*10* insertions located at block junctions, i.e., associated with rearrangements, are numbered and their orientation shown. The phage Mu segments in KO11FL are indicated

copies of a 9-bp repeat, which will be different for each insertion site. The bulk (25/30) of the IS*10*s in KO11FL conform to this pattern, and therefore these IS*10*s have not contributed to rearrangements following insertion. In these cases the sequences flanking each IS*10* insertion in KO11FL are adjacent in the W genome. However, if homologous recombination has occurred between IS*10* elements at different locations, the intervening region can be either inverted or deleted, if the IS*10*s are in opposite or the same orientation, respectively. Another consequence of such rearrangements is that the remaining IS*10* elements are flanked by non-identical 9-bp sequences (Fig. 6). Recombination in KO11FL between IS*10* #3 and #7, which have opposite orientations, has caused inversion of LCB3 (Fig. 5) containing genes KO11FL_03245 to KO11FL_04375 [including K-12 orthologs b3659 (*setC*)–b3874 (*yihN*)] relative to W, and has shuffled the 9-bp repeats to either side of the original IS*10* insertion sites (Fig. 6).

Rearrangements have also been promoted by IS*10* #19, 20, and 22A/22 (adjacent partial and intact IS*10*). The final consequences are inversion and transposition of *feaR* (b1384)–*mcbR* (b1450), shown as LCB6, transposition of *marB* (b1532)–*ydeN* (b1498), shown as LCB5, and two separate deletions. The first deletion of 18 kb removes genes *bcgA*–*ydfK* matching WFL_08140–WFL_08230, and the second of 64 kb removes genes *yncD*–*ydeM* matching WFL_07730–WFL_07955 and corresponding to K-12 genes b1451–b1497. Some deletions in KO11FL appear within LCB alignments in Fig. 4, and include a 40-kb lambda prophage from *intS* (WFL_12425)–*torI* (WFL_12720) in W, which appears as the white region in LCB4 in the W genome. Supplementary Table S1 lists all genes deleted in KO11FL.

The genes present in W but absent from KO11FL were grouped by function (Suppl. Fig. S1), and the proportions in the different groups compared with all genes from *E. coli* W. From this analysis the prophage group contains the largest proportion (44%) of the deleted genes. The remaining genes missing from KO11FL did not include any in the following groups: amino acid metabolism, cell structure, cofactors, DNA processes, fatty acids and phospholipids, and nucleotide metabolism. Presumably the 122 kb of DNA absent from KO11 did not contain any genes that were essential for growth or fermentation.

In contrast to our current laboratory KO11FL strain, the sequence of the original KO11 from DOE JGI (ATCC 55124, GenBank accession CP002516.1) reveals an IS*10* insertion at only a single site, between *yjeJ* and *yjeK*. One complete and two partial copies of IS*10* are present at this location. Our KO11FL contains a remnant of IS*10* at the

Table 2 IS*10* insertions in KO11FL

Insertion	Inserted into	Function of target or adjacent gene(s)
IS*10* #1	*yjcS*	Predicted alkyl sulfatase
IS*10* #2	*ompL*	Predicted outer membrane porin L
IS*10* #3	*ompL//KO11FL_03245**	Predicted outer membrane porin L//hypothetical protein
IS*10* #4	*mdtL*	Drug/chloramphenicol transport protein (MFS family)
IS*10* #5	*KO11FL_04085*	Hypothetical protein
IS*10* #6	*yihF*	Hypothetical protein
IS*10* #7	*ompL//KO11FL_04385**	Predicted outer membrane porin L//hypothetical protein
IS*10* #8	*slp–KO11FL_05250*[†]	Starvation lipoprotein—hypothetical protein
IS*10* #9	*KO11FL_05645*	Hypothetical protein
IS*10* #10	*yqjG*	Predicted S-transferase
IS*10* #11	*ebgC*	Evolved β-D-galactosidase, β subunit; cryptic gene
IS*10* #12	*KO11FL_08470*	Fragment of type III secretion system protein
IS*10* #13	*hydN*	Electron transport protein
IS*10* #14	*yfgF*	Cyclic di-GMP phosphodiesterase
IS*10* #15	*yeeN*	Hypothetical protein
IS*10* #16	arpB_1	Fragment of putative ankyrin repeat protein
IS*10* #17	*intQ–rspB*[†]	Predicted defective phage integrase—predicted oxidoreductase
IS*10* #18	*cspB–essQ*[†]	Phage cold shock protein—phage prediced S lysis protein
IS*10* #19	*pinQ//feaR**	Predicted phage site-specific recombinase//regulatory protein for 2-phenylethylamine catabolism
IS*10* #20	*mcbR//bcgA**	DNA-binding transcriptional dual regulator//6-phospho-β-glucosidase
IS*10* #21	ydeK–lsrK[†]	Predicted lipoprotein—autoinducer-2 kinase
IS*10* #22A, #22[a]	*ydeN//feaR**	Putative sulfatase//transcriptional activator for 2-phenylethylamine catabolism
IS*10* #23	*ynaI*	Conserved inner membrane protein, MscS family
IS*10* #24	*chaA*	Sodium ion:proton antiporter
IS*10* #25	*csgF*	Curli assembly component
IS*10* #26	*ycdT*	Diguanylate cyclase
IS*10* #27	*focA–ycaO*[†]	Formate transporter—protein involved in β-methylthiolation of ribosomal protein S12
IS*10* #28	*phoA*	Alkaline phosphatase
IS*10* #29	*KO11FL_22295–KO11FL_22305*[†]	Lateral flagellin—lateral flagellar transcriptional regulator
IS*10* #30	*idnK*	D-Gluconate kinase, thermosensitive

* Two genes separated by // indicates that the regions to either side of the IS*10* match genes that are not adjacent in *E. coli* W

[†] Two genes separated by a dash (–) indicates that the IS*10* is located in an intergenic region between coding sequences

[a] IS*10* #22A (partial IS*10*) and #22 are immediately adjacent

same location, resulting from nearly precise excision of IS*10* [27]. The genomes of W and the early KO11 sequenced by DOE JGI are colinear (Fig. 5), indicating that none of the IS*10*-promoted rearrangements had occurred at the time of deposit (ATCC 55124), and that the prophage Mu-promoted inversion took place after construction of the original KO11. None of the IS*10*-promoted 18-kb and 64-kb deletions or the 40-kb prophage deletion described above for KO11FL had occurred in ATCC 55124 sequenced by DOE JGI. Although amplification of the *pdc–adhB–cat* region (15–19 copies) was inferred from sequence coverage data for KO11 (ATCC 55124), the copy number of contig 214 matching part of IS*10* was 3.8, well below the coverage expected if every copy of the repeat contained IS*10* in this early KO11 strain. The occurrence of tandem copies of *pdc–adhB–cat* therefore appears to predate insertion of IS*10* into this region.

Small-scale sequence differences between KO11FL and W

We identified sites where the sequence of KO11FL was different from the sequence of the parent strain W. The total number of point mutations, including SNPs and small indels, was 1,473 in the regions of W and KO11FL that were common to both strains. Of these, 1,370 mapped to a region that had been introduced from K-12 into KO11

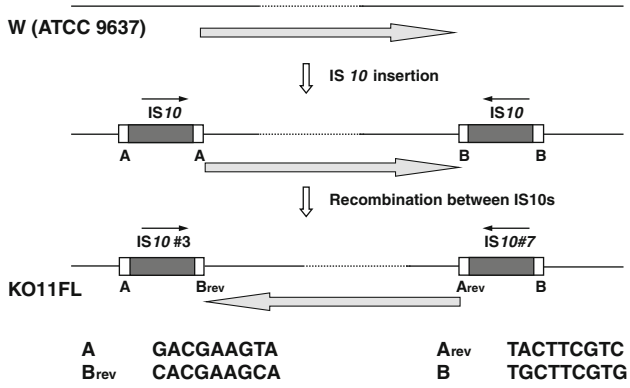

Fig. 6 Inversion of genes in between two IS*10* elements. The *top line* shows the arrangement of genes in the parental strain W (ATCC 9637), with the region that becomes inverted indicated by the *grey arrow*. The *next line* shows insertion of IS*10* at two sites, #3 and #7, generating 9-bp repeats A A and B B flanking each IS*10*. Homologous recombination between the two IS*10*s generates the arrangement seen in KO11FL (*lower line*), with inversion of the region between the IS*10*s, including the 9-bp repeats A and B, which become A_{rev} and B_{rev}. The experimentally determined sequences of A and B_{rev} to either side of IS*10* #3 and A_{rev} and B adjacent to IS*10* #7 are shown. The IS*10* elements #4, 5, and 6 within the inverted region are not represented here

during strain construction [25]. The endpoints of this 125-kb region were within the *uvrA* and *mutL* genes, and were precisely defined by the tight clustering of sequence differences. As expected, the remnant IS*10* sequence was inside this region. Of the 1,370 SNPs within the K-12 derived *uvrA–mutL* region, 1,165 were within coding sequences. Of these, 973 were synonymous, 192 caused single amino acid substitutions, and none caused protein truncation.

The locations and protein consequences of the 103 SNPs/indels that were outside of K-12-derived genes are listed in Supplementary Table S2. For the 103 differences, one was within a tRNA gene (*aspU*), 20 were intergenic, and the remaining 82 were within protein coding sequences (CDSs). Of the 82 within CDSs, 21 were synonymous, 53 caused single amino acid substitutions, 5 created premature stop codons, and 3 caused frameshifts by insertion of one base (two occurrences) or 2 bases (one occurrence). The bulk of the differences between KO11FL and W (101/103) were not present in early KO11 sequenced by DOE JGI, suggesting that they had occurred during propagation of KO11 in our laboratory over the past 20 years. The two mutations in common in KO11FL and DOE JGI sequence, relative to *E. coli* W, caused an Asn to Lys substitution in PtsG, the glucose-specific PTS permease, and a premature stop codon in the *malP* gene encoding maltodextrin phosphorylase, which functions in glycogen degradation.

Comparison of ethanol production (14% xylose)

The performance of the patent deposit strain of KO11 (ATCC 55124) and strain KO11FL were very similar in

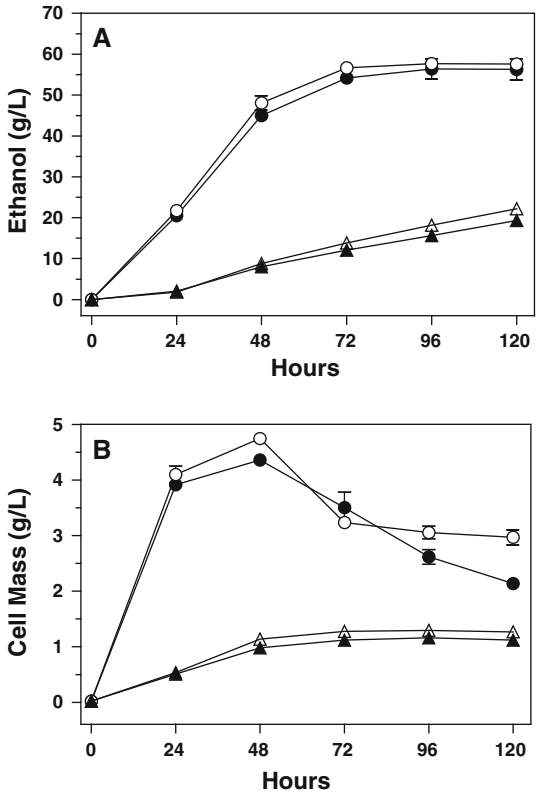

Fig. 7 Comparison of fermentation performance of KO11 (ATCC 55124) and KO11FL. Ethanol production and growth were measured in complex (Luria broth, LB) and defined (NBS) media containing 14% xylose. **a** Ethanol production. **b** Cell growth. *Open circles*, KO11 (patent deposit ATCC 55124) in LB; *filled circles*, KO11FL (current laboratory strain) in LB; *open triangles*, KO11 (ATCC 55124) in NBS; *filled triangles*, KO11FL in NBS

both complex and mineral salts media despite the large number of additional mutations and chromosomal rearrangements in KO11FL. Both grew rapidly in Luria both and produced 58 g l^{-1} ethanol after 96 h (Fig. 7). In defined mineral salts NBS medium, however, neither strain effectively fermented 14% xylose. Subsequent ethanologenic strains [17, 24, 36] have been metabolically evolved to grow and ferment well in defined mineral salts media.

Conclusions

The finding that KO11FL contains tandem repeats of the inserted *pdc–adhB–cat* region provides a mechanism for high-level expression, and is consistent with selection for increased chloramphenicol resistance. The presence of multiple copies of IS*10* and IS*10*-promoted rearrangements in KO11FL was unexpected. In KO11FL the mutations caused by IS*10* did not improve performance. However, in other bacterial strains, IS*10* insertion has proven beneficial [6, 8, 32, 33], and in the ethanologen EMFR9 insertion of IS*10* into *yqhC* increased resistance to furfural [35].

Sequence analysis of further strains will yield a more complete picture of the kinds of mutations that have occurred, and a better understanding of how adaptive laboratory evolution operates.

Acknowledgments The authors thank Savita Shanker at the DNA Sequencing core at the University of Florida for Sanger sequencing of plasmids and PCR products, and Dibyendu Kumar at the UF Bacterial Genome Finishing Program for assistance with bridging gaps between contigs. We acknowledge research support by grants from the US Department of Energy (DE-FG36-08GO88142), US Department of Agriculture, National Institute of Food and Agriculture (2011-10006-30358), and Myriant Technologies. L.O. Ingram is a consultant for Myriant Technologies and a minor stock holder (less than 4%).

References

1. Archer CT, Kim JF, Jeong H, Park JH, Vickers CE, Lee SY, Nielsen LK (2011) The genome sequence of *E. coli* W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of *E. coli*. BMC Genomics 12:9
2. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461:1243–1247
3. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462
4. Burkholder PR (1951) Determination of vitamin B12 with a mutant strain of *Escherichia coli*. Science 114:459–460
5. Causey TB, Zhou S, Shanmugam KT, Ingram LO (2003) Engineering the metabolism of *Escherichia coli* W3110 for the conversion of sugar to redox-neutral and oxidized products: homoacetate production. Proc Natl Acad Sci U S A 100:825–832
6. Chao L, McBroom SM (1985) Evolution of transposable elements: an IS*10* insertion increases fitness in *Escherichia coli*. Mol Biol Evol 2:359–369
7. Chaudhuri RR, Loman NJ, Snyder LA, Bailey CM, Stekel DJ, Pallen MJ (2008) xBASE2: a comprehensive resource for comparative bacterial genomics. Nucleic Acids Res 36:D543–D546
8. Chou HH, Berthet J, Marx CJ (2009) Fast growth increases the selective advantage of a mutation arising recurrently during evolution under metal limitation. PLoS Genet 5:e1000652
9. Conrad TM, Joyce AR, Applebee MK, Barrett CL, Xie B, Gao Y, Palsson BO (2009) Whole-genome resequencing of *Escherichia coli* K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. Genome Biol 10:R118
10. Darling AE, Mau B, Perna NT (2010) Progressive mauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147
11. Diaz E, Ferrandez A, Prieto MA, Garcia JL (2001) Biodegradation of aromatic compounds by *Escherichia coli*. Microbiol Mol Biol Rev 65:523–569
12. Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A (2010) Geneious v5.3. Available from http://www.geneious.com
13. Geddes CC, Nieves IU, Ingram LO (2011) Advances in ethanol production. Curr Opin Biotechnol 22:312–319
14. Gordon DM, Clermont O, Tolley H, Denamur E (2008) Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. Environ Microbiol 10:2484–2496
15. Herring CD, Raghunathan A, Honisch C, Patel T, Applebee MK, Joyce AR, Albert TJ, Blattner FR, van den Boom D, Cantor CR, Palsson BO (2006) Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. Nat Genet 38:1406–1412
16. Ingram LO, Conway T, Clark DP, Sewell GW, Preston JF (1987) Genetic engineering of ethanol production in *Escherichia coli*. Appl Environ Microbiol 53:2420–2425
17. Jarboe LR, Grabar TB, Yomano LP, Shanmugan KT, Ingram LO (2007) Development of ethanologenic bacteria. Adv Biochem Eng Biotechnol 108:237–261
18. Jarboe LR, Zhang X, Wang X, Moore JC, Shanmugam KT, Ingram LO (2010) Metabolic engineering for production of biorenewable fuels and chemicals: contributions of synthetic biology. J Biomed Biotechnol 2010:761042
19. Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muniz-Rascado L, Bonavides-Martinez C, Paley S, Krummenacker M, Altman T, Kaipa P, Spaulding A, Pacheco J, Latendresse M, Fulcher C, Sarker M, Shearer AG, Mackie A, Paulsen I, Gunsalus RP, Karp PD (2011) EcoCyc: a comprehensive database of *Escherichia coli* biology. Nucleic Acids Res 39:D583–D590
20. Kleckner N (1981) Transposable elements in prokaryotes. Annu Rev Genet 15:341–404
21. Latreille P, Norton S, Goldman BS, Henkhaus J, Miller N, Barbazuk B, Bode HB, Darby C, Du Z, Forst S, Gaudriault S, Goodner B, Goodrich-Blair H, Slater S (2007) Optical mapping as a routine tool for bacterial genome sequence finishing. BMC Genomics 8:321
22. Lee SY, Chang HN (1993) High cell density cultivation of *Escherichia coli* W using sucrose as a carbon source. Biotechnol Lett 15:971–974
23. Lenski RE, Mongold JA, Sniegowski PD, Travisano M, Vasi F, Gerrish PJ, Schmidt TM (1998) Evolution of competitive fitness in experimental populations of *E. coli*: what makes one genotype a better competitor than another? Antonie Van Leeuwenhoek 73:35–47
24. Miller EN, Jarboe LR, Yomano LP, York SW, Shanmugam KT, Ingram LO (2009) Silencing of NADPH-dependent oxidoreductase genes (*yqhD* and *dkgA*) in furfural-resistant ethanologenic *Escherichia coli*. Appl Environ Microbiol 75:4315–4323
25. Ohta K, Beall DS, Mejia JP, Shanmugam KT, Ingram LO (1991) Genetic improvement of *Escherichia coli* for ethanol production: chromosomal integration of *Zymomonas mobilis* genes encoding pyruvate decarboxylase and alcohol dehydrogenase II. Appl Environ Microbiol 57:893–900
26. Oshima K, Toh H, Ogura Y, Sasamoto H, Morita H, Park SH, Ooka T, Iyoda S, Taylor TD, Hayashi T, Itoh K, Hattori M (2008) Complete genome sequence and comparative analysis of the wild-type commensal *Escherichia coli* strain SE11 isolated from a healthy adult. DNA Res 15:375–386
27. Ross DG, Swan J, Kleckner N (1979) Nearly precise excision: a new type of DNA alteration associated with the translocatable element Tn*10*. Cell 16:733–738
28. Schumacher G, Sizmann D, Haug H, Buckel P, Bock A (1986) Penicillin acylase from *E. coli*: unique gene-protein relation. Nucleic Acids Res 14:5713–5727
29. Schwan WR, Briska A, Stahl B, Wagner TK, Zentz E, Henkhaus J, Lovrich SD, Agger WA, Callister SM, DuChateau B, Dykes CW (2010) Use of optical mapping to sort uropathogenic *Escherichia coli* strains into distinct subgroups. Microbiology 156:2124–2135
30. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M (2006) ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 34:D32–D36

31. Sobotkova L, Grafkova J, Stepanek V, Vacik T, Maresova H, Kyslik P (1999) Indigenous plasmids in a production line of strains for penicillin G acylase derived from *Escherichia coli* W. Folia Microbiol (Praha) 44:263–266
32. Stoebel DM, Dorman CJ (2010) The effect of mobile element IS*10* on experimental regulatory evolution in *Escherichia coli*. Mol Biol Evol 27:2105–2112
33. Stoebel DM, Hokamp K, Last MS, Dorman CJ (2009) Compensatory evolution of gene regulation in response to stress by *Escherichia coli* lacking RpoS. PLoS Genet 5:e1000671
34. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguenec C, Lescat M, Mangenot S, Martinez-Jehanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Medigue C, Rocha EP, Denamur E (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet 5:e1000344
35. Turner PC, Miller EN, Jarboe LR, Baggett CL, Shanmugam KT, Ingram LO (2010) YqhC regulates transcription of the adjacent *Escherichia coli* genes *yqhD* and *dkgA* that are involved in furfural tolerance. J Ind Microbiol Biotechnol 38:431–439
36. Wang X, Miller EN, Yomano LP, Zhang X, Shanmugam KT, Ingram LO (2011) Overexpression of NADH-dependent oxidoreductase *fucO* increases furfural tolerance in *Escherichia coli* strains engineered for the production of ethanol and lactate. Appl Environ Microbiol 77:5132–5140
37. Yomano LP, York SW, Ingram LO (1998) Isolation and characterization of ethanol-tolerant mutants of *Escherichia coli* KO11 for fuel ethanol production. J Ind Microbiol Biotechnol 20:132–138
38. Yomano LP, York SW, Shanmugam KT, Ingram LO (2009) Deletion of methylglyoxal synthase gene (*mgsA*) increased sugar co-metabolism in ethanol-producing *Escherichia coli*. Biotechnol Lett 31:1389–1398
39. Yomano LP, York SW, Zhou S, Shanmugam KT, Ingram LO (2008) Re-engineering *Escherichia coli* for ethanol production. Biotechnol Lett 30:2097–2103
40. Zhang X, Jantama K, Shanmugam KT, Ingram LO (2009) Reengineering *Escherichia coli* for succinate production in mineral salts medium. Appl Environ Microbiol 75:7807–7813
41. Zhang X, Wang X, Shanmugam KT, Ingram LO (2011) L-Malate production by metabolically engineered *Escherichia coli*. Appl Environ Microbiol 77:427–434
42. Zhang X, Jantama K, Moore JC, Shanmugam KT, Ingram LO (2007) Production of L-alanine by metabolically engineered *Escherichia coli*. Appl Microbiol Biotechnol 77:355–366
43. Zhang X, Jantama K, Moore JC, Jarboe LR, Shanmugam KT, Ingram LO (2009) Metabolic evolution of energy-conserving pathways for succinate production in *Escherichia coli*. Proc Natl Acad Sci U S A 106:20180–20185