JB
Journals.ASM.org

Newly Identified Genetic Variations in Common *Escherichia coli* MG1655 Stock Cultures

Peter L. Freddolino,* Sasan Amini,* and Saeed Tavazoie*

Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, USA

We have recently identified seven mutations in commonly used stocks of the sequenced *Escherichia coli* strain MG1655 which do not appear in the reference sequence. The mutations are likely to cause loss of function of the *glpR* and *crl* genes, which may have serious implications for physiological experiments using the affected strains.

The *Escherichia coli* K-12 strain MG1655 is, by virtue of its early sequencing (3) and the subsequent body of literature characterizing it, one of the most celebrated and best-studied organisms in biology. Several minor corrections of the MG1655 genome have been published (10); in addition, variations between different purported MG1655 stocks have been noted (1, 23). Given that variations between stocks of the same strain may nevertheless cause substantial physiological effects (as in both cases noted above), it is crucial for researchers to be aware of any idiosyncrasies of strains available to them for use in their experiments, particularly in cases such as that of MG1655, for which there may be an expectation of uniformity due to the availability of a carefully vetted sequence. We have recently encountered a set of variations in several MG1655 stocks, including strains available from the American Type Culture Collection (ATCC) and Yale Coli Genetic Stock Center (CGSC). Two of the variations appear likely to have physiological implications under a variety of common laboratory conditions and may confound experiments using the affected strains. In addition, some of the variations identified here appear to have been erroneously reported as novel mutations arising from laboratory evolution experiments, possibly due to direct comparisons of evolved strains to a reference sequence without analysis of the founder strain.

Strains were obtained either directly from ATCC (ATCC 700926 and ATCC 47076) or via the chain of custody described below (HSG001). PCR primers were ordered from IDT (Coralville, IA). PCR template DNA was obtained for each strain by pelleting overnight-grown culture, resuspending the cells in 0.625% Triton X-100, and incubating the suspension for 40 s at 99°C. PCR was performed using ExTaq Hot Start (TaKaRa Bio, Otsu, Shiga, Japan) following the manufacturer's instructions. PCR products were purified using a QIAquick PCR purification kit (Qiagen, Hilden, Germany) and run on a 1% agarose gel containing 1 μg/ml ethidium bromide at 120 V for 60 min. PCR products from selected loci were sequenced via Sanger sequencing by Genewiz, Inc. (South Plainfield, NJ), to confirm all variations.

The mutations investigated here were identified based on being repeatedly observed in several high-throughput sequencing runs by members of the Tavazoie lab using slightly different protocols. In a representative case, genomic DNA from HSG001 cells (see below for lineage) was isolated using a DNEasy blood and tissue kit (Qiagen, Hilden, Germany) and prepared for Illumina sequencing by following the published Illumina genomic DNA sequencing protocol (revision B). Reagents for sample preparation were obtained as a NEBNext kit from New England Biolabs (Ipswich, MA). Mutations relative to the reference sequence were identified both by using breseq (2) and by aligning high-quality reads to the genome using bwa (17) followed by single-nucleotide polymorphism (SNP) and small indel calling using samtools (18).

During whole-genome sequencing of the MG1655-derived strains obtained from a recent set of directed evolution experiments, we identified several apparent mutations and genomic rearrangements, most notably including an insertion in the middle of the *crl* coding region and a frameshift mutation in the *glpR* open reading frame (ORF). Sequencing of the same loci of the parental strain (referred to as HSG001) revealed that it contained many of the noted features. HSG001 is an MG1655 clone that was obtained from the laboratory of F. Blattner (University of Wisconsin) in August 2003. Assuming that our parental strain had, simply by chance, acquired a set of mutations, we obtained a strain from ATCC (ATCC 700926) purportedly matching the reference sequence (GenBank accession U00096.2). The mutations that we had identified were also found in two separately ordered samples of ATCC 700926, but only a subset was present in an older MG1655 isolate available from ATCC, ATCC 47076 (deposited by G. Weinstock). We found identical results for strains from the Coli Genetic Stock Center (CGSC) at Yale University: at the loci considered here, the strain deposited after the MG1655 sequencing project (CGSC7740) contained all of the noted mutations, whereas the older CGSC isolate deposited by M. Guyer (CGSC6300; used as the original strain for the MG1655 sequencing project) contained only the subset present in ATCC 47076. The deviations from the MG1655 reference sequence found in each strain that we considered are summarized in Table 1 (note that while CGSC6300 and ATCC 47076 are equivalent at the loci considered in this study, the strains may not be identical at all other positions). The mutations to *glpR* and *crl* genes are particu-

Received 29 August 2011 **Accepted** 28 October 2011

Published ahead of print 11 November 2011

Address correspondence to Saeed Tavazoie, st2744@columbia.edu.

* Present address: P. L. Freddolino and S. Tavazoie, Department of Biochemistry and Molecular Biophysics & The Initiative in Systems Biology, Columbia University, New York, New York, USA; S. Amini, Illumina, Inc., San Diego, California, USA.

Supplemental material for this article may be found at http://jb.asm.org/.

doi:10.1128/JB.06087-11

TABLE 1 Locations and types of newly identified deviations between ATCC700926, ATCC47076, and the MG1655 reference sequence[a]

Coordinate	Genomic region name	Type	MG1655 reference sequence[b]	W3110 reference sequence[c]	ATCC 47076/CGSC6300	ATCC 700926/CGSC7740	CGSC4474 (W3110)
257899	*crl*	Insertion	G	G	G	IS1 insertion	G
547694	*ylbE*	SNP[d]	A	A	G[e]	G[e,f,g]	G
547832	*ylbE*	Insertion	—	—	G[e]	G[e,f,g]	G
1298719	*oppA-ychE*	Insertion	T	T	T	IS*5* insertion	IS*5* insertion[h]
2171386	*gatC*	Insertion	—	—	—	CC[f]	—
3558478	*glpR*	Deletion	G	G	G	—[f,i]	G
3957957	*ppiC-yifO*	SNP	C	T	T[e]	T[e,f]	T

[a] All variations were identified from whole-genome sequencing data and confirmed by Sanger sequencing of the specific loci (Genewiz, Inc., South Plainfield, NJ). The mutations are shown in genomic context in Note S1 in the supplemental material. We do not include in this table any differences between MG1655 and W3110 strains in the considered regions except for the variations specifically identified between MG1655 stocks and the corresponding reference sequence. —, gap.
[b] GenBank accession no. U00096.2.
[c] GenBank accession no. AP009048.1.
[d] SNP, single-nucleotide polymorphism.
[e] Also noted in reference 22.
[f] Also noted in reference 5.
[g] Also noted in reference 15.
[h] Differs in orientation from that in MG1655 strains.
[i] Also noted in reference 8.

larly likely to have effects relevant to experiments using the affected strains, as they represent, respectively, a frameshift mutation and a large insertion in the middle of the ORFs for a pair of regulatory proteins. A frameshift in the *gatC* gene is also likely to abolish the ability of affected strains to grow on galactitol as a sole carbon source.

The GlpR protein acts as a repressor of genes involved in glycerol 3-phosphate metabolism, and loss of *glpR* gene function has been shown to yield constitutive expression of genes involved in glycerol catabolism in typical rich medium, although cyclic AMP receptor protein (CRP)-mediated repression still occurs in the presence of glucose (12). The nonsense allele that we observed in newer MG1655 stocks was also recently noted in strains obtained from the *E. coli* genome project by Gubellini and coworkers (8) and presumably arose because the sequenced strain was, at various points during the sequencing process, grown on glycerol as a carbon source (F. Blattner, personal communication). Different defective *glpR* alleles have also been observed previously in other K-12 lineages (12, 14) and may likewise reflect the common use of glycerol as a carbon source in laboratory cultures.

Crl modulates the balance of different sigma factors in RNA polymerase holoenzyme formation, particularly the balance of σ^{38} and σ^{70} factors (24), but also affects σ^{32} usage (7). The effects of Crl are known to be particularly important in mediating the onset of the stationary phase (24). While we are unaware of previous reports of the specific *crl* allele noted in Table 1, deletions, including that of the *crl* gene, have been observed to occur with high frequency in strains stored for long periods in stab cultures (6), apparently because a reduction in σ^{38} activity may provide a selective advantage under starvation conditions in laboratory culture (6).

Galactitol import and utilization in *E. coli* occurs through the products of a single gene cluster, *gatYZABCD*, which are expressed constitutively due to an insertion in the repressor *gatR* gene (19). In ATCC 700926 and CGSC7740, we observe a frameshift in the *gatC* gene (the phosphotransferase gene that acts as a galactitol transporter [19]), leading to truncation after 311 residues (out of 451). This mutation almost certainly yields a nonfunctional GatC and may also exert polar effects on the *gatD* gene, the galactose-1-phosphate dehydrogenase gene required for galactitol usage (16).

In addition to the *glpR*, *crl*, and *gatC* mutations described above, the insertion between *ychE* and *oppA* genes is also likely to have regulatory consequences given that it falls between an identified Lrp binding site and the *oppABCDF* operon that it represses (4). Thus, this insertion likely decouples *oppABCDF* operon expression from regulation by Lrp and causes increased expression of the regulon under at least some conditions. OppABCDF is an ABC transporter (20) that has been shown to act in the uptake of small oligopeptides (9) and in the recycling of cell wall peptides (11). The *ychE-oppA* insertion that we identified for some MG1655 strains occurs just 30 bp downstream of an IS*2* insertion found in the *E. coli* K-12 W3110 reference sequence (GenBank accession no. AP009048.1) (10). Surprisingly, our own copy of W3110 (CGSC4474) does not contain the expected IS*2* element but instead contains an IS*5* element inserted at a location identical to that found in the MG1655 strains, except that the IS*5* element is inserted in opposite orientations between the two strains. While IS*5* insertion exhibits a clear sequence preference (6), the fixation of an insertion at this particular intergenic site independently in several separate K-12 lineages strongly suggests that these insertions confer some fitness advantage under laboratory conditions. Based on the function of OppABCDF, it is possible to speculate that the overexpression of this transporter may be beneficial when cells are frequently grown in medium containing an abundance of small peptides.

As seen in Table 1, the remaining three noted mutations (two point mutations in the *ylbE* pseudogene and an insertion in the *ppiC-yifO* intergenic region) were also present in the ancestral MG1655 strain (ATCC 47076/CGSC6300); all three were also recently noted to be present in MG1655 by Skovgaard et al. (22), and those in the *ylbE* gene were noted by Lee and Palsson (15), although in the latter case they were stated to have occurred midway through a directed evolution experiment. Given that the *ylbE* gene is known to be a pseudogene (13), it appears unlikely that the mutations in it would have any measureable effect; the insertion present at position 547832 causes a frameshift that extends the

TABLE 2 Primer pairs used for amplification and sequencing of specific genomic regions[a]

Genomic region name	Forward sequence	Reverse sequence	Left end	Right end
crl	CAGGAAATCACCGACTGGAT	CGACGTCGGTGCTACGTATT	257733	258407
oppA-ychE	GGCATTGGGGATTGAATTTAT	CAGAACGCCAGCTGCTACTA	1298405	1299250
glpR	GAACTTTCTGCCAGCGTCAC	CCTGCTCTGGTGGTGGTATC	3557529	3559014
ylbE	AAAACGTCGCCGTGATTAAC	CAACCTGTGGTCGTTCATTG	547429	549078
ppiC-yifO	CAGTGCTGCTGCTGTTTTTG	TTGATCAGCAGATTCGTTGG	3957813	3958210
gatC	CGCACCTCATCCTGATGTTT	CACTGCCAAAGTGGATCGTA	2170812	2172421

[a] Shown are primer pairs used for amplification and sequencing of specific genomic regions, including the genomic locations of the leftmost and rightmost ends of the primers (using one-indexed genomic coordinates). The locations of all primers are shown in genomic context, along with the mutations that they are designed to test, in Note S1 in the supplemental material.

ylbE_1 ORF past the first stop codon present in the reference sequence, but it does not extend the full length of the predicted ancestral *ylbE* ORF (13). The single-nucleotide polymorphism in the *ppiC-yifO* intergenic region occurs immediately downstream of the *yifO* ORF, approximately 90 bp before the transcription start site for the *ppiC* gene (13). In principle, the regulation of the *ppiC* gene (a proline isomerase gene [21]) could be affected if a regulatory site were disrupted by this change, but to our knowledge, the regulation of the *ppiC* gene remains unstudied. For both mutations identified for the *ylbE* gene, W3110 (CGSC4474) contained the same variations as MG1655, at odds with the W3110 reference sequence (GenBank accession no. AP009048.1). In the case of the SNP identified in the *ppiC-yifO* intergenic region, both the W3110 reference sequence and the strain that we considered contained the variant allele (T instead of C at position 3957957), suggesting that the presence of a T at this position is in fact correct for both MG1655 and W3110.

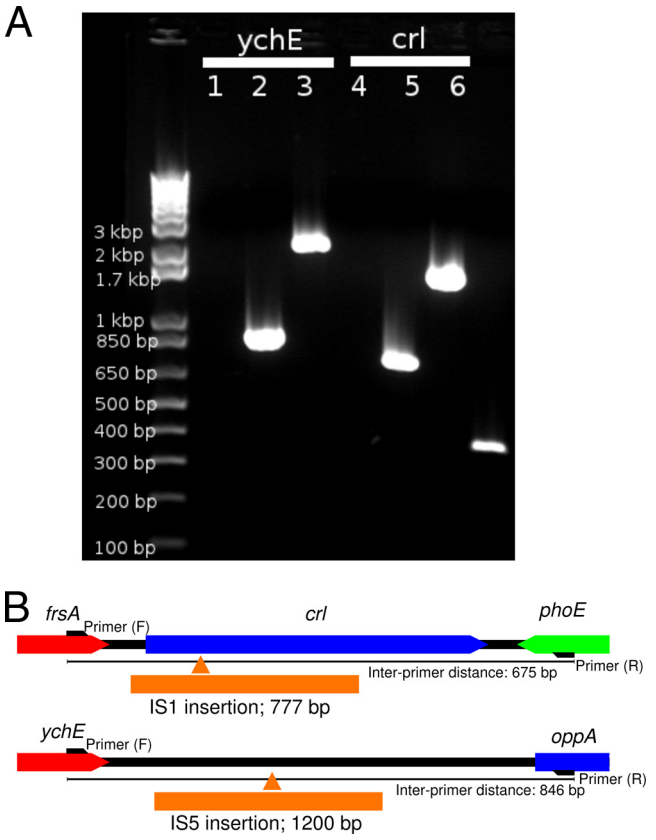

FIG 1 Comparison of PCR product sizes from ATCC 47076 and ATCC 700926 cells by amplifying the *ychE* and *crl* genomic regions. Primer sets are given in Table 2. (A) Gel showing amplified fragment sizes in the absence and presence of insertions. The leftmost lane contains Invitrogen 1 kb Plus DNA ladder. Relevant lanes: 1, *ychE* primers, no template DNA; 2, *ychE* primers, ATCC 47076 DNA; 3, *ychE* primers, ATCC 700926 DNA; 4, *crl* primers, no template DNA; 5, *crl* primers, ATCC 47076 DNA; 6, *crl* primers, ATCC 700926 DNA. Gel lanes from an unrelated experiment have been cropped from the image. Expected product sizes based on the reference sequence are ~750 bp (*crl* gene) and ~850 bp (*ychE* gene). (B) Schematic showing the location of each insertion relative to nearby genes and the primers. Interprimer distances are in the absence of the insertion.

Several variants identified here have, in at least one recent case (5), been presented as novel mutations arising during laboratory evolution experiments, on the basis of direct comparisons between sequencing results for the evolved strain with the MG1655 reference sequence. Given the rate at which bacterial populations evolve, it is crucial in the analysis of laboratory evolution experiments to ensure that any mutations arising in evolved strains were not actually present in the laboratory stock used to begin the experiment.

Given the potential physiological implications of the domestication-related mutations identified here (particularly those to the *crl* and *glpR* genes), the use of an MG1655 isolate with a minimal number of deviations from the reference sequence (e.g., ATCC 47076) is prudent for physiological experiments, particularly for studies involving temperature shifts or growth on glycerol. Strains may be tested for the variations that we identified using the primer pairs shown in Table 2; the *crl* and *oppA-ychE* insertions may then be detected by gel electrophoresis (results for ATCC 700926 and ATCC 47076 are shown in Fig. 1), and the remainder can be identified by DNA sequencing.

Because we found the mutations described here in MG1655 stocks obtained from several common sources (ATCC, CGSC, and another university laboratory), it is prudent for researchers with any doubt regarding the status of their own strains to test them. More generally, variations such as these underscore the need in any laboratory evolution experiment to sequence the parental strain used to start the experiment as well as any offspring of interest, even if the parental strain is ostensibly sequenced and was obtained from a stock collection. Due to the rapidity with which genetic variation arises in bacterial populations, researchers should be aware that any reference sequence represents a single snapshot of the target genome and may not be perfectly matched even by close descendants of the sequenced isolate.

ACKNOWLEDGMENTS

We are grateful to Yirchung Liu for assistance in identification of the *crl* insertion and to Frederick Blattner and Guy Plunkett III for useful discussions regarding the role of reference sequences.

This work was supported by an NIAID award to S.T. (5R01AI077562).

REFERENCES

1. **Barker CS, Prüß BM, Matsumura P.** 2004. Increased motility of *Escherichia coli* by insertion sequence element integration into the regulatory region of the *flhD* operon. J. Bacteriol. **186**:7529–7537.
2. **Barrick J, et al.** 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli.* Nature **461**:1243–1247.
3. **Blattner FR, et al.** 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277**:1453–1462.
4. **Cho B-K, Barrett CL, Knight EM, Park YS, Palsson BØ.** 2008. Genome-scale reconstruction of the Lrp regulatory network in *Escherichia coli.* Proc. Natl. Acad. Sci. U. S. A. **105**:19462–19467.
5. **Fabich AJ, et al.** 2011. Genotype and phenotypes of an intestine-adapted *Escherichia coli* K-12 mutant selected by animal passage for superior colonization. Infect. Immun. **79**:2430–2439.
6. **Faure D, et al.** 2004. Genomic changes arising in long-term stab cultures of Escherichia coli. J. Bacteriol. **186**:6437–6442.
7. **Gaal T, Mandel MJ, Silhavy TJ, Gourse RL.** 2006. Crl facilitates RNA polymerase holoenzyme formation. J. Bacteriol. **188**:7966–7970.
8. **Gubellini F, et al.** 2011. Physiological response to membrane protein overexpression in *E. coli*. Mol. Cell Proteomics **10**:M111.007930.
9. **Guyer C, Morgan DG, Staros JV.** 1986. Binding specificity of the periplasmic oligopeptide binding protein from Escherichia coli. J. Bacteriol. **168**:775–779.
10. **Hayashi, K., et al.** 2006. Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. Mol. Syst. Biol. **2**:2006.0007.
11. **Hiles ID, Gallagher MP, Jamieson DJ, Higgins CF.** 1987. Molecular characterization of the oligopeptide permease of Salmonella typhimurium. J. Mol. Biol. **195**:125–142.
12. **Holtman CK, Thurlkill R, Pettigrew DW.** 2001. Unexpected presence of defective glpR alleles in various strains of *Escherichia coli.* J. Bacteriol. **183**:1459–1461.
13. **Keseler IM, et al.** 2009. EcoCyc: a comprehensive view of *Escherichia coli* biology. Nucleic Acids Res. **37**:D464–D470.
14. **Kinnersley MA, Holben WE, Rosenzweig F.** 2009. E unibus plurum: genomic analysis of an experimentally evolved polymorphism in Escherichia coli. PLoS Genet. **5**:e1000713.
15. **Lee D-H, Palsson BØ.** 2010. Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a nonnative carbon source, L-1,2-propanediol. Appl. Environ. Microbiol. **76**:4158–4168.
16. **Lengeler J.** 1977. Analysis of mutations affecting the dissmilation of galactitol (dulcitol) in Escherichia coli K 12. Mol. Gen. Genet. **152**:83–91.
17. **Li H, Durbin R.** 2009. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics **25**:1754–1760.
18. **Li H, et al.** 2009. The sequence alignment/map format and samtools. Bioinformatics **25**:2078–2079.
19. **Nobelmann B, Lengeler J.** 1996. Molecular analysis of the gat genes from Escherichia coli and of their roles in galactitol transport and metabolism. J. Bacteriol. **178**:6790–6795.
20. **Pearce SR, et al.** 1992. Membrane topology of the integral membrane components, OppB and OppC, of the oligopeptide permease of *Salmonella typhimurium.* Mol. Microbiol. **6**:47–57.
21. **Rahfeld J-U, Schierhorn A, Mann K, Fischer G.** 1994. A novel peptidyl-prolyl cis/trans isomerase from Escherichia coli. FEBS Lett. **343**:65–69.
22. **Skovgaard O, Bak M, Løbner-Olesen A, Tommerup N.** 2011. Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. Genome Res. **21**:1388–1393.
23. **Soupene E, et al.** 2003. Physiological studies of *Escherichia coli* strain MG1655: growth defects and apparent cross-regulation of gene expression. J. Bacteriol. **185**:5611–5626.
24. **Typas F, Barembruch C, Possling A, Hengge R.** 2007. Stationary phase reorganisation of the Escherichia coli transcription machinery by Crl protein, a fine-tuner of σ^s activity and levels. EMBO J. **26**:1569–1578.