Chapter 10
Probing the Deep Genetic Basis of a Novel Trait in *Escherichia coli**

Tanush Jagdish, J. Jeffrey Morris, Brian D. Wade and Zachary D. Blount

Abstract Evolution innovates by repurposing existing genetic elements to produce new functions. However, the range of new functions and traits this evolutionary tinkering can produce is limited to those that are supported and enabled by the rest of the genome. The full complement of genes in a genome required for a novel trait to manifest constitutes the trait's "deep" genetic basis. The deep genetic basis of novel traits can be very difficult to determine under most circumstances, leaving it understudied despite its critical importance. Novel traits that arise during highly tractable microbial evolution experiments present opportunities to correct this deficit. One such novel trait is aerobic growth on citrate (Cit^+), which evolved in one of twelve populations in the Long-Term Evolution Experiment with *Escherichia coli* (LTEE). We sought to uncover the deep genetic basis of this trait by transforming 3,985 single gene knockout mutants from the Keio collection with a plasmid that can confer

Tanush Jagdish
Department of Microbiology and Molecular Genetics, Michigan State University; BEACON Center for the Study of Evolution in Action, East Lansing, MI 48824, USA; and Department of Physics and of Biology, Kalamazoo College, Kalamazoo, MI 49006, USA. Currently at: Program for Systems Biology, Harvard University, Cambridge, MA 02138, USA e-mail: tanush@g.harvard.edu

J. Jeffrey Morris
Department of Microbiology and Molecular Genetics, Michigan State University; BEACON Center for the Study of Evolution in Action, East Lansing, MI 48824, USA; and Department of Biology, University of Alabama at Birmingham, Birmingham, AL 35924, USA

Brian D. Wade
Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI 48824, USA

Zachary D. Blount
Department of Microbiology and Molecular Genetics, Michigan State University; BEACON Center for the Study of Evolution in Action, East Lansing, MI 48823, USA; and Department of Biology, Gambier, OH 43022, USA e-mail: zachary.david.blount@gmail.com

* This paper was externally peer-reviewed.

W. Banzhaf et al. (eds.), *Evolution in Action: Past, Present and Future*,
Genetic and Evolutionary Computation, https://doi.org/10.1007/978-3-030-39831-6_10

aerobic growth on citrate. In our preliminary screen, we identified 111 genes putatively necessary for expression of the Cit$^+$ trait. Of these, $\sim$ 32% are involved in core metabolic pathways, including the TCA and glycolysis pathways. Another $\sim$ 22% encode a variety of transporter proteins. The remaining genes are either of unknown function or uncertain involvement with citrate metabolism. Our work demonstrates how novel traits that are built upon pre-existing functions can depend on the activity of a large number of genes, hinting at an unappreciated level of complexity in the evolution of relatively simple new functions.

Key words: Experimental Evolution, Microbial Evolution, Evolutionary Innovation, Evolution Experiments, Cit$^+$

10.1 Introduction

The living world is astonishingly diverse. Ecologically significant, qualitatively new traits have played an important role in the origin of this diversity [1]. Evolutionary innovations allow lineages to escape competitive pressures in their ancestral niches by invading new niches. Adaptation to the new niches can then drive divergence, speciation, and increased diversity [50]. Diversification driven by novel trait evolution has likely been particularly consequential in microbes, where the origin of novel traits and speciation are thought to be synonymous [12] and thus responsible for the estimated billions to trillions of extant microbial clades [37, 49].

Microbial lineages can evolve novel traits in two distinct ways: acquisition of genes from other lineages through horizontal gene transfer (HGT) or the origination of new genes through modification of existing sequences. HGT can rapidly disseminate new traits among a diverse and distantly related community of microbes and has been implicated in the spread of antibiotic resistance [32, 55], virulence factors [19, 21], and even entire metabolically related gene clusters [2]. However, truly novel traits ultimately arise by the modification of genetic information that did not originally encode them [23].

Such genetic modification, or 'evolutionary tinkering', occurs through four distinct mechanisms [1, 16]. First, new genes can arise *de novo* from mutations that induce expression of previously non-coding DNA that fortuitously yield functional polypeptides [6, 20]. Second, duplications can lead to neofunctionalization, in which mutations of redundant gene copies confer novel functions [13, 41, 43]. Third, pre-existing gene components recombine to yield new functions in a process called 'domain shuffling' [8, 23, 26]. And finally, via a mechanism directly relevant to our work reported here, mutation or recombination can place existing genes under new regulatory control, co-opting functionality by changing the physiological or developmental context within which the genes are expressed [8].

The mutational event that immediately causes the manifestation — or "actualization" — of a novel trait has been the principal concern of most research into the mechanistic bases of evolutionary innovations [1, 16], but it is only one part of a

larger evolutionary process [7, 8]. A new trait can only be actualized if prior evolution has "potentiated" its emergence, either by mechanistically increasing the rate of actualizing mutations, or by epistatically enabling those mutations to produce the trait upon their occurrence [8, 9]. Moreover, new traits almost always first appear in a weak form. The effectiveness of a new trait therefore requires a potentially open-ended, selection-mediated accumulation of mutations that "refines" its functionality and improves its fitness contribution [8, 44, 45]. In this process, potentiation makes the innovation possible, actualization brings it into being, and refinement makes it functionally effective [8].

Prior evolution crucially determines the potential for evolving a novel trait [10]. A new trait can only be actualized if it is mutationally reachable from an organism's existing genetic state [7, 8]. This is to say, it must be possible for the genetic information needed for the trait to arise from sequences that already exist in the genome. One facet of this principle is the necessary interrelationships between genes within a genome. All genes exist in a genome that includes both the set of regulatory elements that govern their expression and that of the other genes with which they interact. Genes also function in a broader organismal context in which their expressed products interact with those encoded by other genes. Typically, gene products only produce a given function when working in tandem with many other gene products. One consequence of this interdependence is that a new trait can only evolve in a lineage in which it will be supported by the existing genomic and organismal context. This broader context that is a necessary part of the potentiation of a novel trait's evolution may be called its "deep" genetic basis.

Consider, for example, an organism that grows on substrate A. However, the organism's environment also contains substrate B, which the organism cannot metabolize. Suppose a mutation in a gene produces an enzyme for converting substrate B to substrate A once it is in the cell, thus allowing the organism to survive on substrate B. The capacity to grow on substrate B will only manifest if the organism has the means to transport substrate B into the cell. Evolution of growth on substrate B is hence contingent on and potentiated by not only the presence of the gene that can be mutated to produce the new enzyme, but also the broader genetic context that includes the requisite transporter gene.

In real organisms, the broader context that determines evolutionary potential is exponentially more complicated [51, 56]. Organisms possess a wide array of integrated traits ranging from survival and stress tolerance to mating and reproduction [52, 54], and the genes and pathways coding for any given trait often also affect other functions at different levels of causality [25, 33, 34]. These interactions between traits and their underlying genetic bases are highly complex and difficult to understand [24, 29, 30]. Adding to this challenge is the fact that novel traits emerge over evolutionary time, which can range from hundreds to thousands of generations [10, 39, 40]. This system-wide, multi-level complexity coupled with the infeasible timescales of novel trait evolution makes it inherently challenging to identify the full set of interacting genetic elements underlying a new trait.

Novel traits sometimes arise during long-term evolution experiments with highly tractable microbial model organisms, providing opportunities to examine their deep

genetic bases [17]. One such instance arose during the Long-Term Evolution Experiment with *Escherichia coli* (LTEE). The LTEE was begun in 1988 with the founding of twelve populations of *E. coli* from a single clone. These populations have since been evolved for more than 70,000 generations of daily 1:100 serial transfer in Davis and Mingioli minimal medium supplemented with glucose (DM25) [15, 34, 35]. Throughout the experiment, population samples have been frozen every 500 generations, providing a complete, viable fossil record of the evolution in each population [35]. DM25 also contains a high concentration of citrate (500 mg/L), which is added as an iron chelating agent and constitutes a potential second carbon and energy source. However, despite possessing a complete TCA cycle, *E. coli* is partly defined as a species by its inability to grow aerobically on citrate (Cit^-) [22, 47]. This Cit^- phenotype is due to an inability to transport citrate into the cell when oxygen is present, but most *E. coli* strains can grow fermentatively on citrate using a transporter expressed only under anoxia [42].

Despite having the cellular machinery to potentially evolve aerobic growth on citrate (Cit^+), spontaneous Cit^+ mutants of *E. coli* are extremely rare under most conditions [22, 53]. Nonetheless, a weak Cit^+ variant appeared in one LTEE population after 31,000 generations [9]. Later genomic analysis showed that the trait was actualized by a duplication that placed the previously silent citrate transporter gene, *citT*, under the control of a promoter that directs expression under aerobic metabolism [8]. The Cit^+ subpopulation remained a minority until shortly after 33,000 generations, when refined variants better able to exploit the citrate resource evolved and rose to high frequency in the population, concurrently leading to a several-fold increase in the size of the population. This process of refinement is ongoing and has involved further adaptation to the idiosyncratic physiology of the Cit^+ trait [5].

The long-delayed and singular evolution of the Cit^+ trait in the LTEE was contingent upon the particular history of the population in which it arose. The ongoing research into this history has revealed the complexity of the interactions underlying evolutionary potential. A series of "replay" experiments with clones isolated from the population's fossil record showed that later clones had a significantly higher rate of mutation to Cit^+, and thus a greater potential to evolve the trait than did earlier ones [9]. This potentiation arose in part from adaptation to acetate-based ecological interactions that evolved in the population. Quandt et al. [45] identified a series of mutations that occurred in the citrate synthase gene, *gltA*, of the lineage in which Cit^+ eventually arose. These mutations altered carbon flow into the TCA cycle, improving growth on acetate. As citrate is also metabolized via the TCA cycle, they also pre-adapted the lineage to growth on citrate, and rendered the Cit^+ actualizing mutation beneficial when it eventually arose. These mutations hence allowed the weak initial Cit^+ variant lineage to remain in the population long enough to accumulate refining mutations that improved growth on citrate [45]. Ironically, the potentiating mutations in *gltA* compensated for the anti-potentiating effects of earlier mutations that were beneficial to growth on glucose [36]. Moreover, even after potentiation, the Cit^+ actualizing mutation's fitness benefit prior to 30,000 generations was too low to permit it to outcompete other beneficial mutations available to the population [36].

Much has therefore been gleaned about the evolutionary potentiation of the Cit$^+$ trait during the LTEE. However, the trait's evolution also depended on the full complement of genes necessary for citrate metabolism that existed in the ancestral genome prior to the start of the LTEE. These genes that interact and support aerobic growth on citrate, the trait's deep genetic basis, remain a mystery. Here we describe an initial exploration of the deep genetic basis of the novel Cit$^+$ trait, in which we sought to identify the non-essential genes required for aerobic growth on citrate in the presence of the actualizing mutation.

10.2 Methods

10.2.1 Long-Term Evolution Experiment with E. coli

The LTEE has been described in greater detail elsewhere [35]. Briefly, twelve populations of *E. coli* B were founded in 1988 and have since been evolved for more than 70,000 generations of serial batch culture in Davis and Mingioli minimal medium supplemented with 25 mg/L glucose ("DM25"; [15, 34]. The populations are diluted 100-fold every 24 hours into fresh medium to a final volume of 10 mL and maintained at 37 °C with 120 rpm orbital aeration.

10.2.2 Screening for Genes Necessary for Aerobic Growth on Citrate

The pZB*rnk-citT* plasmid is a pUC19-based recombinant plasmid that contains a copy of the novel *rnk-citT* module that actualized the Cit$^+$ trait in Ara-3 [8]. This high copy number plasmid confers a Cit$^+$ phenotype in both *E. coli B* and *E. coli K12*. The Keio collection is made up of 3985 strains of *E. coli* K12 in which a single, non-essential gene has been deleted and replaced with a kanamycin resistance cassette [4]. Genes required for aerobic growth on citrate were screened by chemically transforming each Keio strain with pZB*rnk-citT* and testing each transformant for the capacity to grow aerobically on citrate using the procedure described below.

Each knock-out strain to be transformed was grown on LB plates supplemented with kanamycin at a final working concentration of 50 μg/mL. Three to five colonies of each strain were then resuspended in 300 μL of 50 mM $CaCl_2$. The suspensions were incubated on ice for 15 minutes, after which $\sim$100 ng of plasmid DNA was added to the suspensions. Following 45 minutes of further incubation on ice, the cells were heat-shocked at 42 °C for 1 minute before being returned to ice for 5 minutes. A 500 μL volume of LB was added to the cells, shaken to mix, and the full volume directly spread on Christensen's citrate indicator plates supplemented with 50 μg/mL kanamycin and 100 μg/mL ampicillin [3, 11]. The plates were then in-

cubated for 10 days at 37 °C. Plates were assessed after 5 and 10 days of incubation for development of a hot pink coloration of the medium indicative of growth on citrate. Those Keio strain transformants able to grow on citrate as indicated by a color change were presumed to have deletions in genes that were not required for aerobic growth on citrate. Those that did not produce a color change were considered to have deletions of candidate genes required for aerobic growth on citrate. Thirty six mutants from the Keio collection were unused as viable transformants for those mutants could not be obtained. This screen was carried out twice to rule out false positives.

10.2.3 Curation and Analysis of Gene Functions

Functions of candidate genes were first manually curated using the EcoCyc database [28]. All 111 functions were then grouped into five major categories: Metabolism, Membrane-Related Proteins, Stress Response, Transcription, and Motility. Genes that could not be classified into groups of three or more were labelled as "Other", and those with undefined functions were labelled "Unknown".

10.2.4 Analysis of the Mutational History in Candidate Genes

Good et al. [18] conducted whole genome, whole population sequencing of all frozen populations between 0 and 60,000 generations across all 12 LTEE lines. We downloaded the annotated sequence data made publicly available by Good et al. [18] and looked for mutations in any of the 111 candidate genes in all 12 LTEE lines over the course of 60,000 generations. In order to ensure we only looked at mutations that were nearing or had already reached fixation, we only considered mutations that had reached a frequency of 0.95. All mutational analysis was carried out using R (version 3.5.1). Analysis scripts and raw data are deposited at https://github.com/tjagdish/DeepGeneticBasis

10.3 Results

We screened the entire Keio *E. coli* K-12 gene knockout collection for non-essential genes required for aerobic growth on citrate by transforming each constituent knockout strain with a plasmid, pZB*rnk-citT*, which can confer a Cit^+ phenotype in a wild type genetic background. We identified 111 candidate genes that are putatively necessary for aerobic growth on citrate (Appendix). We manually categorized each candidate gene by the primary function identified for it by the EcoCyc *E. coli* database ([27], Fig. 10.1). Thirty-six percent of the candidates are involved in core

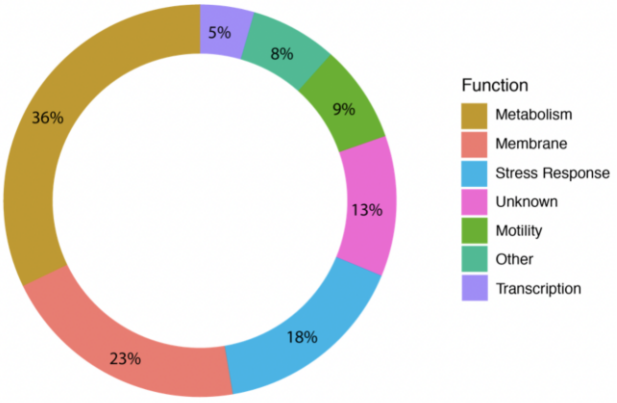

Fig. 10.1: Functional Categories of Genes Putatively Necessary for Aerobic Growth on Citrate. Genes grouped by functions assigned by EcoCyc [27]. Genes assigned to the 'other' category include those associated with fimbriae (2%), metalloproteases (2%), lipoproteins (2%), cell division (1%), cell shape (1%), and curli formation (1%)

metabolic pathways, including *sucA, sucB* and *sdhB*, which encode TCA cycle enzymes, and others that encode glycolysis pathway enzymes (Fig. 10.1). Another ∼23% encode for membrane-associated proteins, such as *tatB, macB,* and *garP*, which are involved in protein translocation, antibiotic export, and galactarate transport, respectively (Fig. 10.1).

The remaining candidates are either of unknown function or uncertain involvement with citrate metabolism. These genes include a substantial number that are likely involved in bacterial stress response, such as multi-drug efflux pumps (*emrD* and *ybhF*), and regulators of acid resistance and biofilm formation (*ymgB*). Other candidates include transcription regulators (5), metalloproteases (3), lipoproteins (3), and genes involved in regulating fimbriae (3), cell division (1), cell shape (1), and curli formation (2).

Genes necessary for growth on citrate would be logical targets of selection in the Cit$^+$ population. To determine if this has been the case for the 111 candidate mutations we identified, we searched the whole-metagenome sequence data for the citrate-using population from Good et al. [18]. We found that mutations had fixed or nearly fixed in 8 of the candidate genes by 60,000 generations (Table 10.2). However, all mutations occurred in Ara-3 after ∼35,000 generations, at which point Ara-3 had evolved a mutator phenotype [8].

All 12 LTEE populations have been evolving for over 70,000 generations. Ara-3 remains the only population in which aerobic growth on citrate has evolved. An ongoing question is that of whether it might evolve in any of the other 11 populations. Mutations that impair or eliminate the function of genes required for the Cit$^+$ trait would presumably reduce the likelihood of evolving it. We therefore examined

Ara-1	Ara-2	Ara-3	Ara-4	Ara-5	Ara+2	Ara+3		Ara+4	Ara+6	
bisC	*ilvG*	*aroG*	*emrD*	*iscR*	*sucB*	*phoR*	*aroG*	*aroG*	*aroG*	*ybaY*
rarD	*nadB*	*atpI*	*flgI*			*rarD*	*atpI*	*phoR*	*csgG*	*ycbS*
sucB	*phoR*	*csgG*	*gatY*			*rimL*	*dinJ*		*dedA*	*ydiJ*
yahK	*treA*	*emrD*	*mltD*			*sdhB*	*flgE*		*erfK*	*yegV*
ycbS	*ybhF*	*flgF*	*sucA*			*sfmA*	*flgF*		*exbD*	*yejB*
yegV	*ydhP*	*ybhF*	*yahB*			*sucA*	*flgI*		*flgE*	*yfiB*
	ydjZ	*ycbS*	*ydiS*			*sucB*	*garL*		*hcaE*	*ygbI*
	yhjK	*ydiJ*	*ydjZ*			*tauC*	*garP*		*ilvG*	*yadC*
	yneJ		*yegV*			*thiD*	*hcaE*		*pfkA*	*wcaK*
			yihF			*yadC*	*ilvD*		*phoR*	*treA*
			yjgZ			*yahB*	*mltD*		*rimL*	*sucB*
						ybaY	*ygbI*			
						ydhP	*yggN*			
						yejB	*yibQ*			
						ypdE	*yjiB*			

Fig. 10.2: Genes putatively identified as necessary for aerobic growth on citrate in which mutations have fixed in LTEE populations by 60,000 generations. Ara-1, Ara-2, Ara-3, Ara-4, Ara+3 and Ara+6 have all evolved heightened mutation rates over the course of evolution. No fixed mutations in candidate genes were identified in populations Ara-6, Ara+1, and Ara+5

the metagenome sequences for the other LTEE populations for mutations in the 111 genes our screen identified.

We found an array of mutations in genes putatively essential for aerobic growth on citrate across the other LTEE populations. The vast majority of these mutations occurred in the five populations, Ara-1, Ara-2, Ara-3, Ara-4, Ara+3, and Ara+6, in which elevated mutation rates have also evolved over the course of the experiment (Fig. 10.3; Fig. 10.2). However, we did identify mutations in three non-mutator populations: a single mutation in a noncoding region of the transcription regulator gene, *iscR*, fixed in Ara-5 by 45,000 generations; a missense mutation in a key oxoglutarate dehydrogenase component gene, *sucB*, fixed in Ara+2 by 47,000 generations; and in Ara+4 a missense mutation in the glycolysis regulating phosphofructokinase gene, *phoR*, fixed by 18,000 generations, as well as the insertion of a mobile genetic element in the aromatic amino acid biosynthesis gene *aroG*. We identified no mutations in any candidate genes in Ara-6, Ara+1, or Ara+5.

10.4 Discussion

The Cit$^+$ trait that arose in the LTEE would seem to be a simple innovation, given that the trait can be conferred in the ancestral genetic background by activating expression of the *citT* gene that encodes a citrate-C4-dicarboxylate antiporter [8]. Despite this apparent simplicity, the trait's manifestation required the activity of other gene products, and thus the presence of other preexisting genes. These genes

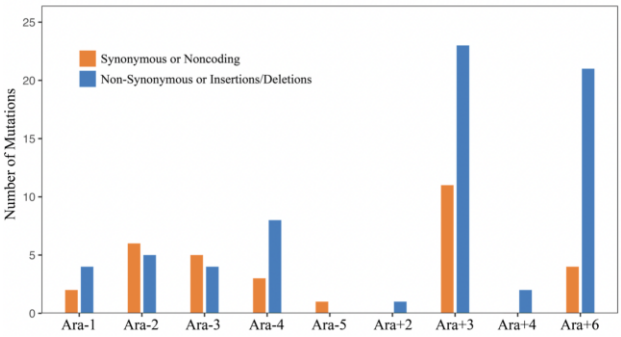

Fig. 10.3: Fixed mutations in candidate genes across 12 LTEE populations by 60,000 generations.

constitute the deep genetic basis of the Cit$^+$ trait, and we sought to identify them. In total, our preliminary screen showed that 3%, or 111 out of 4288 non-essential genes in *E. coli* were required for aerobic growth on citrate. Moreover, considering the possible role of epistatic interactions and the conservative nature of our screen, this number is likely to be an underestimate [14, 46]. Our findings therefore suggest that the deep genetic basis of the novel Cit$^+$ trait is quite broad, highlighting the integrated nature of the organism.

The largest group of genes that we identified as putatively necessary for aerobic growth on citrate belong to core metabolism. *E. coli* metabolizes exogenous citrate via the TCA cycle, making the genes that encode the steps in the cycle necessary [31, 38, 47]. Aerobic growth on citrate as a sole carbon source also creates problems for biosynthesis as glycolysis is bypassed. Growth on citrate as a sole carbon source thus requires the capacity to feed carbon into the gluconeogenesis pathways, as well as to produce the TCA intermediates and key amino acid precursors 2-oxyglutarate and succinyl-CoA. A further complication is caused by the physiology of the citT transporter. Every molecule of citrate imported requires the simultaneous export of a C4-dicarboxylate, and specifically the TCA intermediates of succinate, fumarate, or malate [42]. Consequently, as few as 2 carbons per citrate molecule are available for both catabolism and anabolism (Fig. 10.4). Given these considerations, several genes might be predicted as necessary for aerobic growth on citrate.

Predicted genes include isocitrate dehydrogenase (*icd*), which is necessary to yield 2-oxoglutarate because the 2-oxoglutarate decarboxylase (SucAB) reaction is irreversible; isocitrate lyase (*aceA*), which is necessary to bypass the CO2-producing SucAB and SucCD reactions via the glyoxylate shunt; and citrate synthase (*gltA*), which is necessary to pass carbons harvested via glyoxylate back into the TCA cycle to reach 2-oxoglutarate and succinyl-CoA. Interestingly, our screen did not yield any of these genes. This odd result might be an artifact of the screen's design. These genes would be essential in minimal medium with citrate as a sole carbon source. However, the Christensen's citrate agar on which we conducted our phenotypic screen likely contains sufficient amino acids and other metabolites to compensate for anabolic deficiencies caused by loss of any one of these key en-

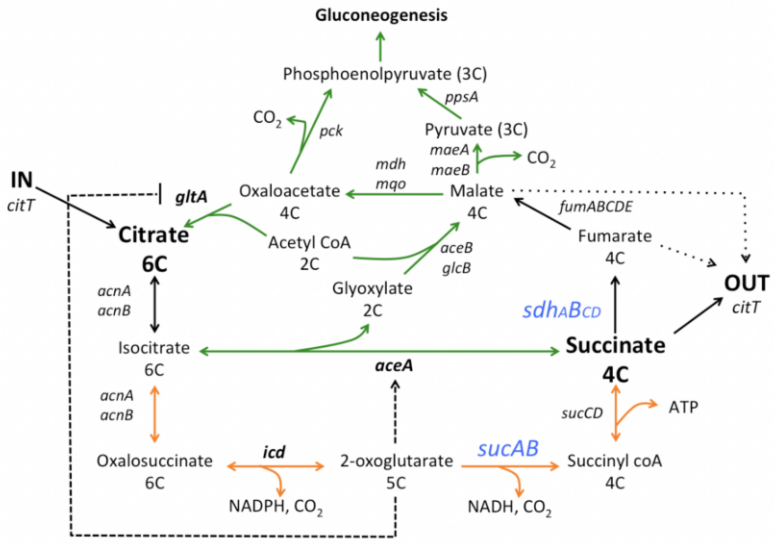

Fig. 10.4: Citrate metabolic pathways in Ara-3 Cit$^+$. Extracellular citrate is exchanged for the intracellular C4-dicarboxylate TCA intermediates, succinate, fumarate, and malate by the citT antiporter. This physiology means perhaps only 2 carbon atoms are available per citrate molecule for both catabolism and anabolism. The citrate can be metabolized in two ways. In the catabolic pathway (orange arrows), both carbons are lost as CO_2 (leaving none for biosynthesis), but substantial energy is conserved in the form of NAD(P)H. In the anabolic pathway (green arrows), no energy is gained, but the carbons are harvested as glyoxylate, and can either be passed back into the TCA cycle or into gluconeogenesis for biosynthesis of amino acids and other necessary metabolites. Where two gene names are given (e.g. *acnA, acnB*) for a reaction, the genes are redundant; where genes are listed as a complex (e.g. *sucAB*), all gene products are needed to catalyze the reaction. Because of redundant genes and/or pathways, only 3 genes (*gltA, icd*, and *aceA*, shown in bold) may be predicted as essential for aerobic growth on citrate as a sole carbon source. None of these genes were identified in our screen, but others, shown in blue, were. Regulatory effects of 2-oxoglutarate are indicated with dashed lines.

zymes. We plan to follow up on these preliminary findings with further screens using Christensen's citrate broth from which we will exclude yeast extract, which should allow us to evaluate this hypothesis.

Some of the TCA genes we identified in our screen are not strictly essential, but their absence could lead to the over-accumulation of intermediates, negatively affecting regulation, and causing unbalanced growth. For instance, deletion of either of the subunits of 2-oxoglutarate decarboxylase (*sucA* or *sucB*) eliminated the capacity to grow aerobically on citrate. This effect could be due to an accumulation of 2-oxoglutarate levels that would inhibit citrate synthase, preventing the use of citrate-derived carbon for amino acid biosynthesis. Similarly, loss of one of the four subunits of succinate dehydrogenase, SdhB, eliminated the Cit$^+$ trait. SdhB is the

cytoplasmic subunit responsible for passing electrons from SdhA, which binds both succinate and FAD, to the electron transport chain [48]. It is possible that loss of SdhB sequesters succinate in a form that cannot be accessed by citT, stopping the flow of citrate into the cell.

The second largest subset of essential gene candidates were membrane proteins and cell wall synthesis enzymes. The 22 membrane proteins include ABC transporter families such as *macB* and *yhhJ*, and a diversity of importers, exporters, and symporters. Several genes involved in peptidoglycan synthesis and remodeling (*ddlA, erfK, dacA*) were also identified. This subset is more puzzling, as there is no obvious direct relationship between these genes and citrate metabolism. While Cit$^+$ requires an aerobically functioning citT membrane transporter, an *rnk-citT* module (where rnk is an aerobic promoter) was available to the tested cells via a high-copy-number plasmid [8]. Thus, the removal of the citT gene from the genome, or any other transporter gene, should in principle not affect Cit$^+$.

The role of most of the genes we have identified in aerobic growth on citrate is unclear. This lack of obvious connection is perhaps most clearly seen in the cell appendage biosynthesis genes our screen showed as putatively necessary to the Cit$^+$ trait. These genes included those involved in flagellar biosynthesis (*flgE, flgF, flgI*, and *flgJ*), fimbrial assembly (*ycbS*), and curli synthesis (*csgABFG*). Similarly, 18 candidate genes are involved in stress response functions in *E. coli*, including multidrug efflux pumps (*emrD* and *ybhF*), reactive oxygen defenses (*sodB*), acid resistance regulators (*ymgB*), and response regulators for phosphate starvation (*phoR*). At least some of these genes may be experimental artifacts of a screening procedure that exposed the cells to multiple stressors, including exposure to two antibiotics, treatment in high concentrations of calcium chloride, and cold.

Might the candidate genes we identified be targets of refining mutations that improved aerobic growth on citrate after the evolution of the Cit$^+$ trait? To answer this question, we examined the metagenome sequence data generated by Good et al. [18] for the Ara-3 population through 60,000 generations. We found mutations in 8 candidate genes, though none reached high frequency until well after the Cit$^+$ clade evolved an elevated mutation rate around 35,000 generations [8]. The earliest mutation to rise to high frequency occurred at around 37,000 generations in the gene *csgG*, which encodes an outer membrane lipoprotein involved in curli biosynthesis. Due to the elevated mutation rate, however, it is unclear if these mutations are beneficial or reached high frequency by hitchhiking with some other beneficial allele.

Only 3 of the 8 mutated candidate genes in Ara-3, *aroG, atpI*, and *ydiJ*, are related to central metabolism. These genes have a total of six mutations, all of which are either synonymous or occur in noncoding regions. While it is possible that these mutations might have beneficial fitness effects, it seems more likely that they are hitchhikers. The five other candidate genes, *csgG, flgF, ycbS, emrD*, and *ybhF*, which are involved in biofilm, flagellar, and fibrial biosynthesis, and encode multidrug efflux pumps, respectively, have a total of 12 mutations. Five of these mutations are synonymous or noncoding, but the remaining seven are missense or indel mutations. The latter seven mutations would presumably affect gene function, and potentially

conflict with our findings that they are necessary for the Cit$^+$ trait. Later work will examine these mutations, their effects, and will determine if their identification in the screen was perhaps an experimental artifact of some sort. Broadly, however, the rarity of mutations in candidate genes in Ara-3 after the evolution of Cit$^+$ is consistent with their being necessary for the trait and suggests that mutations in them are generally detrimental.

If the 111 genes we identified are actually necessary for aerobic growth on citrate, then mutations that impair or eliminate their function would seemingly reduce the likelihood of evolving the Cit$^+$ trait. We identified nonsynonymous mutations and indels in multiple candidate genes in 7 of the other 11 LTEE populations. We do not yet know enough about the effects of these mutations on the function of the genes in question. However, any that impair or eliminate function would likely reduce or foreclose the possibility of the evolution of Cit$^+$ in the respective populations. It will be interesting to examine how the evolvability of Cit$^+$ varies between the populations in which mutations have occurred in candidate genes, and Ara-6, Ara+1, and Ara+5, in which they have not.

Novel traits are not the result solely of the genetic changes that immediately underlie them. Those genetic changes that actualize a trait must always occur in a genetic background containing an integrated set of genes and gene products that allow them to produce that new trait. Though the importance of this deep genetic basis of novel traits is in a sense obvious, it has rarely been examined. The preliminary work we have described gives a glimpse into the deep genetic basis of the Cit$^+$ trait that arose in the LTEE. The trait was actualized by a mutation that activated the expression of a citrate transporter gene, *citT*, when oxygen was present via the cooption of an alternate promoter, that of the aerobically expressed gene *rnk*. Despite this apparent simplicity, our results show that the manifestation of the Cit$^+$ trait depends on the activity of more than 100 other genes. This finding shows how the additive nature of novel traits in evolution means that even relatively minor, seemingly simple novel traits nonetheless depend on a foundation of many pre-existing elements that interact in complex and highly integrated ways.

Indeed, our findings suggest that what genes might be involved in the manifestation of a trait may be anything but obvious due to this complexity and integration. Prior work has shown that the evolution of the Cit$^+$ trait was historically contingent upon several potentiating mutations that arose over the course of the Ara-3 population's history during the experiment, and the twisted paths of evolution during that history. Our findings here show that the trait was contingent upon not simply this history during the experiment. Indeed, it was contingent upon the much longer history that preceded the experiment, over which evolution constructed an organism with the full complement of genes that interacted in such a way as to support the manifestation of the trait once the actualizing mutation took place. Such contingent histories are necessary to provide the proper deep genetic basis that underlies the evolution of all novel traits. Given the role of novel traits in evolution, our work argues that it is time to take this deeper historical contingency seriously.

Acknowledgements We thank Neerja Hajela for her outstanding technical expertise, logistical aid, and general moral support. This project would not have been possible without her. We also thank the undergraduate technicians who provided logistical support for our work, including Camorrie Bradley, Michelle Mize, Maia Rowles, Kiyana Weatherspoon, Rafael Martinez, Jamie Johnson, Devin Lake, and Jessica Baxter. We are grateful to Richard Lenski for helpful discussions, advice, funding, lab space, and his continued mentorship. We are especially grateful to Erik Goodman for his extensive and path-defining leadership of the BEACON Center for the Study of Evolution in Action, and in whose honor this work is submitted. Mark Kauth, Kyle Card, and Nkrumah Grant made critical contributions to the design of this project. This research was supported in part by grants from the National Science Foundation (currently DEB-1451740) and the BEACON Center for the Study of Evolution in Action (DBI-0939454). T.J. acknowledges support from the Internship Stipend Program at the Center for Career and Professional Development at Kalamazoo College.

References

1. Andersson, D.I., Jerlstroem-Hultqvist, J., Naesvall, J.: *Evolution of new functions de novo and from preexisting genes*. Cold Spring Harb Perspect Biol **7**(6), a017996 (2015)
2. Araviad, L., Tatusov, R.L., Wolf, Y.I., Roland Walker, D., Koonin, E.: *Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles*. Trends in Genetics **14**, 442–444 (1999)
3. Atlas, R.: *Handbook of Microbiological Media for the Examination of Food*. CRC, Boca Raton, FL (2006)
4. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., Mori, H.: *Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection*. Molecular Systems Biology **2**, 2006.0008 (2006)
5. Bajic, D., Vila, J.C.C., Blount, Z.D., Sanchez, A.: *On the deformability of an empirical fitness landscape by microbial evolution*. Proceedings of the National Academy of Sciences of the United States of America **115**, 11,286–11,291 (2018)
6. Begun, D.J., Lindfors, H.A., Kern, A.D., Jones, C.D.: *Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade*. Genetics **176**(2), 1131–1137 (2007)
7. Blount, Z.D.: *A case study in evolutionary contingency*. Stud Hist Philos Sci Part C: Stud Hist Philos Biol Biomed Sci **58**, 82 – 92 (2016)
8. Blount, Z.D., Barrick, J.E., Davidson, C.J., Lenski, R.E.: *Genomic analysis of a key innovation in an experimental Escherichia coli population*. Nature **489**, 513–518 (2012)
9. Blount, Z.D., Borland, C.Z., Lenski, R.E.: *Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America **105**(23), 7899–7906 (2008)
10. Blount, Z.D., Lenski, R.E., Losos, J.B.: *Contingency and determinism in evolution: Replaying life's tape*. Science **362**, eaam5979 (2018)
11. Christensen, W.: *Hydrogen sulfide production and citrate utilization in the differentiating of enteric pathogens and coliform bacteria*. Res Bull Weld County Health Dept Greeley Colo **1**, 3–16 (1949)
12. Cohan, F.M.: *Bacterial species and speciation*. System Biology **50**(4), 513–524 (2001)
13. Conant, G.C., Wolfe, K.H.: *Turning a hobby into a job: How duplicated genes find new functions*. Nature Reviews Genetics **9**, 938–950 (2008)
14. Dasmeh, P., Girard, E., Serohijos, A.W.R.: *Highly expressed genes evolve under strong epistasis from a proteome-wide scan in E. coli*. Scientific Reports **7**, 15844 (2017)
15. Davis, B.D., Mingioli, E.S.: *Mutants of Escherichia coli requiring methionine or vitamin B12*. Journal of Bacteriology **60**, 17–28 (1950)

16. Ding, Y., Zhou, Q., Wang, W.: *Origins of new genes and evolution of their novel functions*. Annual Review of Ecology, Evolution, and Systematics **43**, 345–363 (2012)
17. Elena, S.F., Lenski, R.E.: *Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation*. Nature Reviews Genetics **4**, 457–469 (2003)
18. Good, B.H., McDonald, M.J., Barrick, J.E., Lenski, R.E., Desai, M.M.: *The dynamics of molecular evolution over 60,000 generations*. Nature **551**, 45–50 (2017)
19. Groisman, E.A., Ochman, H.: *Pathogenicity islands: Bacterial evolution in quantum leaps*. Cell **87**, 791–794 (1996)
20. Guerzoni, D., McLysaght, A.: *De novo origins of human genes*. PLoS Genetics **7**(11), e1002381 (2011)
21. Hacker, J., Blum-Oehler, G., Muehldorfer, I., Tschaepe, H.: *Pathogenicity islands of virulent bacteria: Structure, function and impact on microbial evolution*. Molecular Microbiology **23**, 1089–1097 (1997)
22. Hall, B.G.: *Chromosomal mutation for citrate utilization by Escherichia coli k-12*. Journal of Bacteriology **151**, 269–273 (1982)
23. Jacob, F.: *Evolution and tinkering*. Science **196**(4295), 1161–1166 (1977)
24. Jerison, E.R., Desai, M.M.: *Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments*. Current Opinion in Genetics & Development **35**, 33–39 (2015)
25. Jerison, E.R., Kryazhimskiy, S., Mitchell, J.K., Bloom, J.S., Kruglyak, L., Desai, M.M.: *Genetic variation in adaptability and pleiotropy in budding yeast*. eLife **6**, e27167 (2017)
26. Kawashima, T., Kawashima, S., Tanaka, C., Murai, M., Yoneda, M., Putnam, N.H., Rokhsar, D.S., Kanehisa, M., Satoh, N., Wada, H.: *Domain shuffling and the evolution of vertebrates*. Genome Research **19**(8), 1393–1403 (2009)
27. Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M., Karp, P.D.: *EcoCyc: A comprehensive database resource for Escherichia coli*. Nucleic Acids Research **33**(suppl-1), D334–D337 (2005)
28. Keseler, I.M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martnez, C., Caspi, R., et al.: *The EcoCyc database: Reflecting new knowledge about Escherichia coli K-12*. Nucleic Acids Research **45**, D543–D550 (2016)
29. Khan, A.I., Dinh, D.M., Schneider, D., Lenski, R.E., Cooper, T.F.: *Negative epistasis between beneficial mutations in an evolving bacterial population*. Science **332**, 1193–1196 (2011)
30. Kryazhimskiy, S., Rice, D.P., Jerison, E.R., Desai, M.M.: *Global epistasis makes adaptation predictable despite sequence-level stochasticity*. Science **344**, 1519–1522 (2014)
31. Lara, F.J.S., Stokes, J.L.: *Oxidation of citrate by Escherichia coli*. Journal of Bacteriology **63**(3), 415–420 (1952)
32. Lawrence, J.G., Ochman, H.: *Reconciling the many faces of lateral gene transfer.* Trends in Microbiology **10**, 1–4 (2002)
33. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M., Wray, N.R.: *Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood*. Bioinformatics **28**, 2540–2542 (2012)
34. Lenski, R.E.: *Experimental studies of pleiotropy and epistasis in Escherichia coli. i. variation in competitive fitness among mutants resistant to virus T4*. Evolution **42**, 425–432 (1988)
35. Lenski, R.E., Rose, M.R., Simpson, S.C., Tadler, S.C.: *Long-term experimental evolution in Escherichia coli. I. adaptation and divergence during 2,000 generations*. American Naturalist **138**(6), 1315–1341 (1991)
36. Leon, D., D'Alton, S., Quandt, E.M., Barrick, J.E.: *Innovation in an E. coli evolution experiment is contingent on maintaining adaptive potential until competition subsides*. PLoS Genetics **14**, 1–22 (2018)
37. Locey, K.J., Lennon, J.T.: *Scaling laws predict global microbial diversity*. Proceedings of the National Academy of Sciences of the United States of America **113**(21), 5970–5975 (2016)
38. Luetgens, M., Gottschalk, G.: *Why a co-substrate is required for anaerobic growth of Escherichia coli on citrate*. Journal of General Microbiology **119**, 63–70 (1980)
39. Meyer, J.R., Dobias, D.T., Medina, S.J., Servilio, L., Gupta, A., Lenski, R.E.: *Ecological speciation of bacteriophage lambda in allopatry and sympatry*. Science **354**, 1301–1304 (2016)

40. Meyer, J.R., Dobias, D.T., Weitz, J.S., Barrick, J.E., Quick, R.T., Lenski, R.E.: *Repeatability and contingency in the evolution of a key innovation in phage lambda*. Science **335**, 428–432 (2012)
41. Ohno, S.: *Evolution by gene duplication*. Springer-Verlag (1970)
42. Pos, K.M., Dimroth, P., Bott, M.: *The Escherichia coli citrate carrier CitT: A member of a novel eubacterial transporter family related to the 2-oxoglutarate/malate translocator from spinach chloroplasts*. Journal of Bacteriology **180**, 4160–4165 (1998)
43. Qian, W., Zhang, J.: *Genomic evidence for adaptation by gene duplication*. Genome Research **24**, 1356–1362 (2014)
44. Quandt, E.M., Deatherage, D.E., Ellington, A.D., Georgiou, G., Barrick, J.E.: *Recursive genomewide recombination and sequencing reveals a key refinement step in the evolution of a metabolic innovation in Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America **111**, 2217–2222 (2014)
45. Quandt, E.M., Gollihar, J., Blount, Z.D., Ellington, A.D., Georgiou, G., Barrick, J.E.: *Fine-tuning citrate synthase flux potentiates and refines metabolic innovation in the Lenski evolution experiment*. eLife **4**, e09696 (2015)
46. Remold, S.K., Lenski, R.E.: *Pervasive joint influence of epistasis and plasticity on mutational effects in Escherichia coli*. Nature Genetics **36**, 423–426 (2004)
47. Reynolds, C.H., Silver, S.: *Citrate utilization by Escherichia coli: Plasmid- and chromosome-encoded systems*. Journal of Bacteriology **156**, 1019–1024 (1983)
48. Ruprecht, J., Yankovskaya, V., Maklashina, E., Iwata, S., Cecchini, G.: *Structure of Escherichia coli succinate: Quinone oxidoreductase with an occupied and empty quinone-binding site*. Journal of Biological Chemistry **284**, 29836–29846 (2009)
49. Schloss, P.D., Handelsman, J.: *Status of the microbial census*. Microbiol Mol Biol Rev **68**(4), 686–691 (2004)
50. Schluter, D.: *Evidence for ecological speciation and its alternative*. Science **323**(5915), 737–741 (2009)
51. Segrè, D., DeLuna, A., Church, G.M., Kishony, R.: *Modular epistasis in yeast metabolism*. Nature Genetics **37**, 77–83 (2005)
52. Suding, K.N., Goldberg, D.E., Hartman, K.M.: *Relationships among species traits: Separating levels of response and identifying linkages to abundance*. Ecology **84**, 1–16 (2003)
53. Van Hofwegen, D.J., Hovde, C.J., Minnich, S.A.: *Rapid evolution of citrate utilization by Escherichia coli by direct selection requires citT and dctA*. Journal of Bacteriology **198**, 1022–1034 (2016)
54. Vasi, F., Travisano, M., Lenski, R.E.: *Long-term experimental evolution in Escherichia coli. ii. changes in life-history traits during adaptation to a seasonal environment*. The American Naturalist **144**, 432–456 (1994)
55. von Wintersdorff, C.J.H., Penders, J., van Niekerk, J.M., Mills, N.D., Majumder, S., van Alphen, L.B., Savelkoul, P.H.M., Wolffs, P.F.G.: *Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer*. Front Microbiol **7**, 173 (2016)
56. Wang, Z., Liao, B.Y., Zhang, J.: *Genomic patterns of pleiotropy and the evolution of complexity*. Proceedings of the National Academy of Sciences of the United States of America **107**, 18034–18039 (2010)

Appendix

Genes identified as putatively necessary for aerobic growth on citrate and their annotations as reported by EcoCyc.

Genes	Group	Function
ada	stress response	Adaptive response transcriptional regulator
aroG	metabolism	Amino acid biosynthesis
-	-	-
atpI	metabolism	ATP synthase accessory factor
bioC	stress response	Biotin scavenging and protection from oxygen damage
chaC	metabolism	γ-glutamylcyclotransferase with relatively low catalytic efficiency
chpB	stress response	ChpB is the toxin component of the ChpB-ChpS toxin-antitoxin system
csgA	motility	Curlin subunit
csgB	motility	Curlin subunit
csgF	membrane	Curlin subunit
csgG	membrane	Membrane lipoprotein
dacA	membrane	Penicillin-binding protein, antibiotic resistance
ddlA	metabolism	Peptidoglycan biosynthesis
dedA	membrane	Inner membrane protein with trans-membrane domains
dinJ	stress response	Antitoxin to YafQ toxin
dksA	transcription	Initates transcription by binding to RNA polymerase
ecnA	stress response	Antidote/toxin system which may function as a starvation adaptation
emrD	stress response	Multidrug efflux pump
eptB	metabolism	Encodes the Ca2+-induced phosphoethanolamine transferase
-	-	-
erfK	cell shape	Covalent attachment of peptidoglycan to the outer membrane
exbD	metabolism	Active transport of iron-siderophore complexes and vitamin B12 across the outer membrane
FabR	metabolism	Fatty acid biosynthesis regulator
flgE	motility	Flagellar hook protein
flgF	motility	Flagellar basal-body rod protein
flgI	motility	Flagellar P-ring protein
fliJ	motility	Flagellar biosynthesis protein
fur	transcription	DNA-binding transcriptional dual regulator
galP	membrane	Galactose symporter
garL	metabolism	α-dehydro-β-deoxy-D-glucarate aldolase
garP	membrane	Galactarate/glucarate/glycerate transporter
garY	stress response	Catalytic subunit of the GarYZ tagatose-1,6-bisphosphate aldolase complex. Responds to low pH, acetate stress, and glucose limitation
gmlX	metabolism	Catabolism of DNA as carbon source
hcaE	unknown	Unknown
hslO	stress response	Stress regulated chaperone
htrL	unknown	Unknown
hyfC	metabolism	Hydrogenase-4 component C
ilvD	metabolism	Isoleucine and valine biosynthesis
ilvG	metabolism	Isoleucine biosynthesis
iscR	transcription	Iron-sulfur cluster regulator
ligB	mismatch repair	NAD+-dependent DNA ligase
lsrC	membrane	ABC transporter membrane subunit
macB	membrane	ABC exporter
mltD	membrane	Membrane-bound transglycosylase
murP	membrane	N-acetylmuramic acid PTS transport system
nadB	metabolism	NAD biosynthesis I pathway
nohA	transcription	Transcription regulator of nohA
ompW	membrane	Receptor for Colicin S4
opgC	membrane	ABC transporter
pgaA	metabolism	Phosphoacetylglucosamine-4-kinase
pfkA	metabolism	6-phosphofructokinase-1 (PFK-1)
phoR	stress response	Responds to variations in the level of extracellular inorganic phosphate
ppiB	metabolism	Conformational isomerization of prolyl residues in peptides
rarD	membrane	Putative transporter
rimL	metabolism	Acetyltransferase
sdhB	metabolism	Succinate/quinone oxidoreductase
slmA	motility	Putative fimbrial protein
yangB	unknown	Unknown

Genes	Group	Function
slmA	motility	Putative fimbrial protein
sodB	stress response	Superoxide dismutases, implicated in the response to a large number of environmental changes
-	-	2-oxoglutarate decarboxylase, thiamine-requiring
sucA	metabolism	Conversion of 2-oxoglutarate (2-ketoglutarate) to succinyl-CoA and carbon dioxide
suiC	metabolism	Assembly of iron-sulfur clusters
tatB	membrane	Inner membrane component of the twin arginine translocation (Tat) complex
tauC	membrane	ABC transporter
thiD	metabolism	Cfosynthesis of thiamine
treA	metabolism	Hydrolysis of trehalose into two molecules of D-glucose
trkD	membrane	Potassium ion uptake under hyper-osmotic stress at a low pH
ulaE	metabolism	Isomerization of L-ribulose 5-phosphate to D-xylulose 5-phosphate
ulaF	metabolism	Isomerization of L-ribulose 5-phosphate to D-xylulose 5-phosphate
wcaK	metabolism	Colanic acid biosynthesis
yadC	motility	Fimbrial operon
yahB	unkown	Unknown
yahK	metabolism	NADPH-dependent aldehyde reductase activity
yhiQ	stress response	Cysteine detoxification
yhaY	stress response	yhaY has supercoiling-dependent transcription, which is associated with the osmotic stress response
yhhA	metabolism	Pyridoxal phosphate phosphatase
yhhF	stress response	Multidrug ABC exporter
-	-	-
ycbS	motility	Fimbrial operon
yceO	stress response	Acid stress response
ygeX	unknown	Unknown
ydaS	stress response	Toxic protein; expression of YdaS leads to loss of cell viability
ydhP	membrane	Putative transporter
ydiJ	metabolism	FAD-linked oxidoreductase
ydiS	metabolism	Putative oxidoreductase with FAD/NAD(P)-binding domain
ydjJ	metabolism	Putative zinc-binding dehydrogenase
ydtZ	membrane	Putative transmembrane protein
yeaM	stress response	DNA-binding transcriptional repressor
-	-	-
yeaW	metalloprotease	2Fe-2S cluster-containing protein
yegV	metabolism	Putative sugar kinase
yciB	membrane	Inner membrane subunit of a putative ATP-dependent oligopeptide uptake system
yfiB	biofilm	Biofilm formation
yghI	unkown	Unknown
ygjN	biofilm	Biofilm formation
yhaC	unknown	Unknown
yhhJ	membrane	ABC transporter
yhjK	membrane	Inner membrane protein
yiaY	metabolism	NAD-dependent L-threonine dehydrogenase activity
yibQ	metabolism	Nucleoside diphosphatase
yidQ	unknown	Putative lipoprotein
yihF	unknown	Unknown
yjaZ	unknown	Unknown
yjbE	membrane	Putative membrane protein
yjjB	metabolism	Putative succinate exporter
ymfD	metabolism	e14 prophage; Putative SAM-dependent methyltransferase
ymfJ	unknown	e14 prophage; protein YmfJ
ynsJ	transcription	Putative LysR-type DNA-binding transcriptional regulator
yoaD	membrane	Inner membrane protein
yrfF	metalloprotease	Metalloendopeptidase with broad specificity
yggB	stress response	Acid stress response
zapA	cell division	Cell division protein
tatE	unknown	Unknown
yfcG	unknown	Transcriptional regulator
ymfC	unknown	Unknown