Links between Transcription, Environmental Adaptation and Gene Variability in *Escherichia coli*: Correlations between Gene Expression and Gene Variability Reflect Growth Efficiencies

Jean-Paul Feugeas*,[1,2] Jerome Tourret,[1,2,3] Adrien Launay,[1,2] Odile Bouvet,[1,2] Claire Hoede,[4] Erick Denamur,[1,2,5] and Olivier Tenaillon*,[1,2]

[1]INSERM, UMR 1137, Infection, Antimicrobiens, Modélisation, Evolution (IAME), Paris, France

[2]Faculté de Médecine, Universités Paris Diderot et Paris Nord—Sorbonne Paris Cité, Paris, France

[3]AP-HP, Unité de Transplantation, GH Pitié-Salpêtrière Charles Foix et Université Pierre et Marie Curie, Paris, France

[4]INRA, MIAT, Plateforme Bio-Informatique GenoToul, Castanet-Tolosan Cedex, France

[5]AP-HP, Laboratoire de Génétique Moléculaire, GH Paris Nord Val de Seine, Paris, France

***Corresponding author:** E-mail: jean-paul.feugeas@inserm.fr; olivier.tenaillon@inserm.fr.

Associate editor: Deepa Agashe

Abstract

Gene expression is known to be the principle factor explaining how fast genes evolve. Highly transcribed genes evolve slowly because any negative impact caused by a particular mutation is magnified by protein abundance. However, gene expression is a phenotype that depends both on the environment and on the strains or species. We studied this phenotypic plasticity by analyzing the transcriptome profiles of four *Escherichia coli* strains grown in three different culture media, and explored how expression variability was linked to gene allelic diversity. Genes whose expression changed according to the media and not to the strains were less polymorphic than other genes. Genes for which transcription depended predominantly on the strain were more polymorphic than other genes and were involved in sensing and responding to environmental changes, with an overrepresentation of two-component system genes. Surprisingly, we found that the correlation between transcription and gene diversity was highly variable among growth conditions and could be used to quantify growth efficiency of a strain in a medium. Genetic variability was found to increase with gene expression in poor growth conditions. As such conditions are also characterized by down-regulation of all DNA repair systems, including transcription-coupled repair, we suggest that gene expression under stressful conditions may be mutagenic and thus leads to a variability in mutation rate among genes in the genome which contributes to the pattern of protein evolution.

Key words: **transcription, adaptation, mutation, evolution.**

Introduction

Gene expression is a phenotype that is determined by a combination of environmental and genetic effects. Changes in environmental conditions trigger coordinated changes in the transcription of genes, allowing a physiological adaptive response. This phenotypic plasticity and its connections to metabolic pathways have been extensively studied in the *E. coli* laboratory strain K-12. These studies revealed the key contribution of transcriptional regulatory networks and of a diverse set of proteins such as RpoS and H-NS (Sheldon et al. 2012; Imam et al. 2015). The other aspect of gene expression diversity, the diversity based on genetic differences between strains, is more poorly characterized, but is expected to result in differences in transcript expression (Le Gall et al. 2005; Okuda et al. 2007) which are critical for the life style variability seen within bacterial species. Genetic differences in promoter regions and position effects correlate with functional features (Bryant et al. 2014; Meysman et al. 2014), but small variations in coding regions can also directly or indirectly modulate transcription through regulatory networks.

Gene expression is also a major determinant of gene evolution (Drummond and Wilke 2009). From eukaryotes to prokaryotes, the most highly expressed genes are the ones showing the lowest rate of evolution. Selection appears to drive this pattern. Any mutation that generates a less effective or less stable protein will lead to a fitness loss proportional to protein abundance (Drummond and Wilke 2008). Consequently, as gene expression increases, a larger fraction of mutations are counter-selected. In microbes, gene expression is also a major factor in codon bias (Plotkin and Kudla 2011). The use of codons recognized by abundant tRNAs is favored in highly expressed genes as it increases the fidelity and speed of translation and generates more error-free proteins while mobilizing less ribosomes to do so. The Codon Adaptation Index (CAI) itself is sometimes used as a proxy for expression, based on the correlation between high expression and codon bias (Plotkin and Kudla 2011). Additionally, gene expression level is also a major determinant of horizontal gene transferability, high expression hampering horizontal gene transfer (HGT) (Park and Zhang 2012). Therefore, it is widely accepted that high expression is associated with an increased gene stability.

Mutation and drift are the other forces that shape gene evolution. High gene expression is connected to a reduced diversity through selection but might enhance mutation rates. In all organisms, including *Escherichia coli*, transcription has been indeed found to be mutagenic (Chen and Zhang 2013). During transcription, the coding strand is exposed as a single strand. Therefore, it is more prone to chemical alterations that may lead to mutations after translesional replication (Wright et al. 2013). The noncoding strand might be partially protected by the action of transcription-coupled repair (TCR). TCR is a repair mechanism centered on Mfd protein in *E. coli*, which mediates the recruitment of nucleotide excision-repair proteins at DNA lesions encountered by the RNA polymerase (Selby and Sancar 1993). Nevertheless, mutants may also arise from damages to the transcribed DNA strand. In this case, miscoding of the damaged DNA during transcription may sometimes lead to the production of a mutant protein that could help the cell to undergo DNA replication and to fix the mutation in daughter cells (Bridges 1994). This could contribute to adaptive mutation, a process in which new mutants arise from a nondividing cell population during selection (Morreall et al. 2015). Whatever the mechanism is, transcription-driven mutagenesis (TDM) reveals an increasing tension between mutation and selection when expression is high. A hypothesis suggests that this tension may be partially resolved by the creation of DNA secondary structures: Bases that are paired in local secondary DNA structures are protected from chemical alterations during transcription and are thus less sensitive to TDM. Consistently, an excess of these structures has been found in highly expressed genes, and may contribute to the lower mutation rate observed among these genes (Hoede et al. 2006; Wright et al. 2013). However, whether or not selection may be strong enough to favor local modulation of mutation rates is still debated. Some authors have suggested that secondary structures may have been selected for their role in RNA stability (Chursov et al. 2013).

Most of the studies linking expression to gene evolution have been performed with the assumption that gene expression is variable throughout the genome only, with any variation attributable to environmental factors being neglected. The data used to infer expression patterns are usually based on transcriptomic studies conducted on a reference strain in a reference rich medium (Drummond and Wilke 2009). Hence, the contribution of gene expression variability to gene evolution remains to be explored and there is a need for analyses combining the genetical and environmental aspects. Recently, an analysis taking into account both environmental and genetic diversity in expression patterns was performed on *E. coli*, and suggested that that highly differentially expressed genes showed increased coding-sequence dissimilarity (Vital et al. 2015). To delve further into the analysis of the impact of transcription diversity and gene evolution, we designed an experimental strategy (fig. 1) in which we first analyze the transcriptomes of four *E. coli* strains in three different culture media (step 1, influence of genetic and environmental factors on transcription). We then compute several indices of genetic diversity in these four strains and in an additional set of 114 phylogenetically diverse *E. coli* strains (step 2). Finally, we explore how gene expression may be linked to genetic diversity (step 3).

Results

As described in figure 1, our aim was to first identify determinants of gene expression variability across media and strains, then determine markers of genetic diversity, and finally search for connections between transcription and nucleotide diversity or gene loss. The four *E. coli* strains that we chose were isolated from different habitats and had various pathogenicity potentials (Touchon et al. 2009; Sabarly et al. 2011) (table 1): IAI1 (phylogroup B1) and K-12 MG1655 (phylogroup A) are human commensals, O157:H7 Sakai (phylogroup E) is responsible for hemorrhagic diarrhea in humans, and CFT073 (phylogroup B2) was isolated from the blood of a patient suffering from septicemia of urinary origin. Strains were cultured under micro-aerobic conditions in lysogeny broth (LB), urine (U) and a minimal medium supplemented with gluconate (G). We selected these culture media for their different nutrient content and the stress they induce. LB is a complex and rich medium; urine is also a complex medium but lacks iron and induces osmotic and oxidative stresses (Aubron et al. 2012; Roux et al. 2012). The gluconate minimal medium is also relatively poor and stressful and was chosen because gluconate is utilized in the intestine by *E. coli* (Chang et al. 2004). Transcript levels of the 3,281 genes present in all four of the *E. coli* strains were quantified. Obtaining this core transcriptome was not a trivial task because genetic variability may have introduced bias if microarray probes had been designed specifically for one strain. This is the reasoning that led us to design an oligonucleotide array containing probes for all of the genes of the four genomes ("pan-coli" chip, see Materials and Methods). We observed that expression levels depend on both the strain under study and culture conditions, as also found by Vital et al. (2015). In our experimental conditions, the proportion of variance in expression levels between strains or between media was comparable (53% and 47% of the total variance, respectively) and we decided to focus on analyzing these two effects separately. For the sake of clarity we called "medium-dependent" genes, the genes whose expression variability depended primarily on the medium and "strain-dependent" genes the genes whose expression variability depended primarily on the strain, although this distinction might vary according to strains and media used.

Genes Whose Expression Variability Depends Primarily on the Medium ("Medium-Dependent" Genes) Are Involved in Metabolic Pathways

We first identified genes whose pattern of expression under the 12 conditions was mostly determined by the medium in

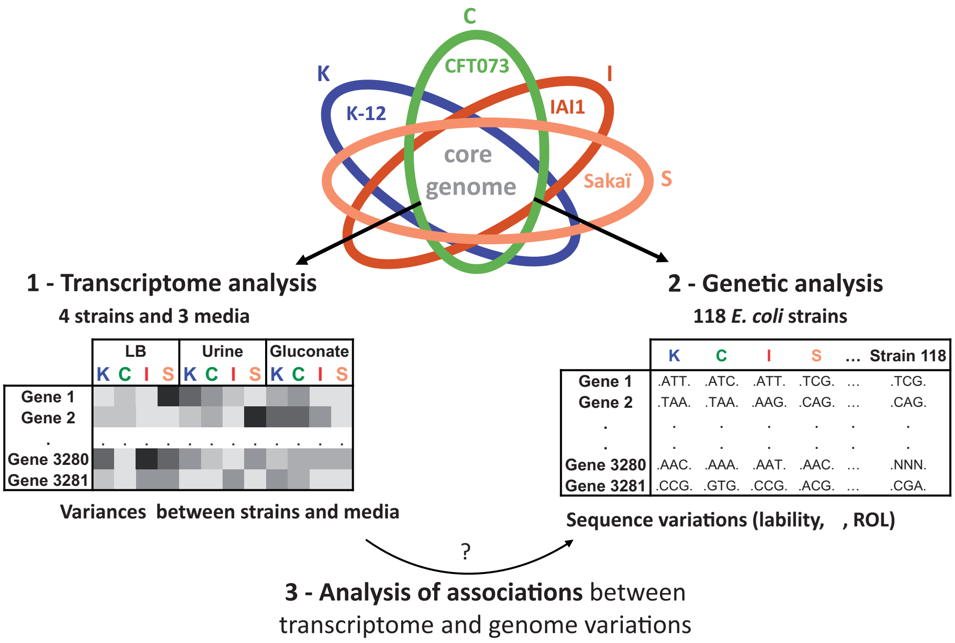
C
CFT073
K
K-12
I
IAI1
S
Sakaï
core genome
1 - Transcriptome analysis
4 strains and 3 media
LB
Urine
Gluconate
K C I S K C I S K C I S
Gene 1
Gene 2
Gene 3280
Gene 3281
Variances between strains and media
2 - Genetic analysis
118 *E. coli* strains
K C I S … Strain 118
Gene 1 .ATT. .ATC. .ATT. .TCG. … .TCG.
Gene 2 .TAA. .TAA. .AAG. .CAG. … .CAG.
Gene 3280 .AAC. .AAA. .AAT. .AAC. … .NNN.
Gene 3281 .CCG. .GTG. .CCG. .ACG. … .CGA.
Sequence variations (lability, , ROL)
?
3 - Analysis of associations between transcriptome and genome variations

Fig. 1. General study design. The transcriptomes of four *Escherichia coli* strains cultured in three media were simultaneously analyzed (step 1). Genetic markers of diversity among the four studied strains and an additional set of 114 *E. coli* strains were determined (step 2). Links between transcription and genetic diversity were then explored (step 3). θ is the per site theta Watterson estimator. ROL is the rate of loss index defined by Silander and Ackermann. C=CFT073, K=K-12, S=Sakaï, I=IAI1.

Table 1. Strains and Culture Media.

E. coli Strains	Phylogroups	Pathogenicity	Media	Designation	Growth (OD at H18)
K-12 MG1655	A	Commensal	LB	KL	0.6
			Urine	KU	0.2
			Gluconate	KG	0.1
IAI1	B1	Commensal	LB	IL	0.9
			Urine	IU	0.2
			Gluconate	IG	0.3
CFT073	B2	Uropathogenic	LB	CL	0.4
			Urine	CU	0.1
			Gluconate	CG	0.1
Sakaï O157:H7	E	Enterohaemorrhagic	LB	SL	0.7
			Urine	SU	0.2
			Gluconate	SG	0.4

Note.—C, CFT073; K, K-12; S, Sakaï; I, IAI1; G, Gluconate; U, Urine; L, LB. Media and strains provided 12 transcriptome analyses (CG, CU, CL, KG, KU, KL, SG, SU, SL, IG, IU, IL).

which strains were grown. We found >400 such genes, and most of them were involved in bacterial metabolism (fig. 2 and supplementary tables S1 and S2, Supplementary Material online). The first branch of the supervised hierarchical clustering of these genes distinguished LB from the other two culture media (fig. 2). The second branch in the nonLB group was then split into two sub-groups, corresponding to the less energy-rich media, urine and gluconate. The expression pattern of these genes was in agreement with proteomic data obtained for the same strains grown under similar conditions. Hence, metabolic pathway usage is determined by the nutrients present in the culture medium (Sabarly et al. 2016). Details of the genes that were recruited in the different media are given in the captions of supplementary tables S1 and S2, Supplementary Material online.

Genes Whose Expression Variability Depends Primarily on the Strain ("Strain-Dependent" Genes) Are Involved in Sensing and Responding to Environmental Changes

We identified genes whose expression pattern in the 12 transcriptomes was mostly driven by genetic diversity between strains rather than by environmental conditions. We identified 887 such strain-dependent genes (supplementary tables S3 and S4, Supplementary Material online). As expected, they discriminated the four strains in a supervised hierarchical clustering (fig. 3B). Gene sequence variability may be a causative factor for strain-dependent gene expression variation. We estimated the phylogenetic distance between the strains by calculating the average nucleotide identity (ANI). There was a correlation between ANI and transcriptomic distances (mean $r = 0.87$, $P < 0.1$), resulting in similar clustering

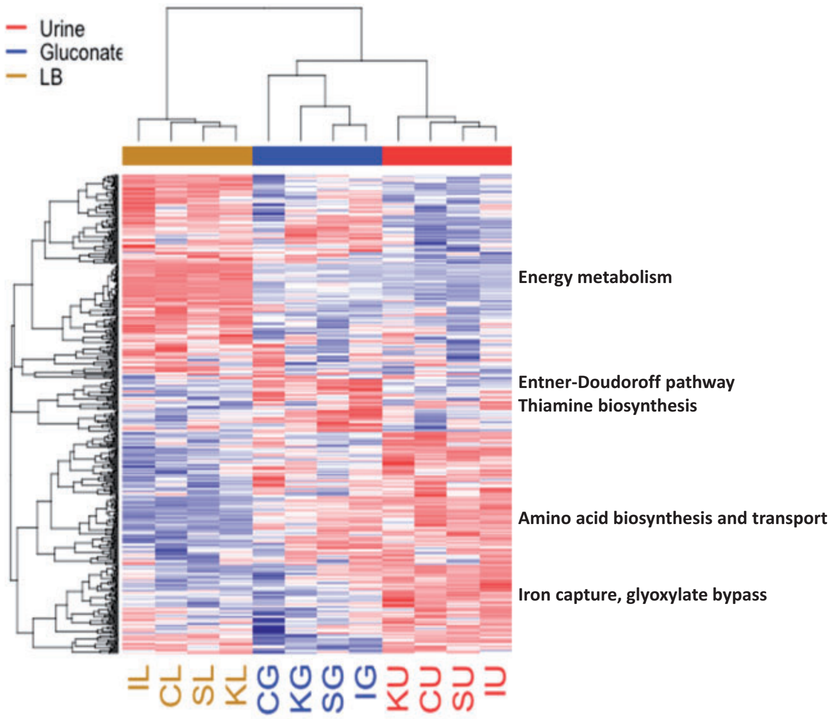

FIG. 2. Hierarchical clustering of the four *Escherichia coli* strains cultured in three media with "medium-dependent" genes. The 404 most "medium-dependent" genes (rows) and the 12 experimental conditions (columns) were clustered using the rank correlation between transcript expressions as similarity measure (see also supplementary table S1, Supplementary Material online). C = CFT073, K = K-12, S = Sakaï, I = IAI1, G = Gluconate, U = Urine, L = LB. High transcript levels are marked in red and low levels are marked in blue.

dendrograms when only the expression of "strain" genes were considered (fig. 3A and B). Phylogenetic distances and gene expression variations between strains were not correlated when the expression of all core genes were considered (data not shown). This was also described by Vital et al. (2015) who found that global gene expression distances among strains mainly depended on their physiological state. Interestingly, many of the strain-dependent genes were linked to processes implicated in sensing and responding to the environment, such as tyrosine kinase signaling, transmembrane transport and two-component signal transduction systems (TCS) ($P < 4 \times 10^{-3}$). Examples of genes that are differentially expressed in the four strains are shown in figure 3B and described in the supplementary data (supplementary tables S3 and S4, Supplementary Material online).

Analysis of Genetic Variability

To explore gene evolution determinants, we first computed two markers of gene variability of the core genome of the studied strains across 114 other *E. coli* strains (supplementary table S5, Supplementary Material online). These markers were gene lability and the Watterson estimator of gene diversity, θ (Watterson 1975). Gene lability (L) can be computed as $L = 1-$gene presence frequency, and measures the probability of a gene being lost. L had a mean value of 0.03 within a range of 0–0.96. The Watterson estimator of gene diversity, θ, is an estimate of genetic diversity based on the number of polymorphic sites observed within a gene. The mean θ value of the 118 strains was 0.009 per bp, and ranged from 0 to 0.1.

Because most of the genes we studied were conserved among *E. coli* strains, we also used the Rate of Loss index (ROL) based on gene loss among other bacterial species. The "ROL all" was established by Silander and Ackermann (2009) for 3,845 orthologs in "all bacteria" (448 taxa), and it reflects the rate of loss of the ortholog along the bacterial phylogeny (mean value: 51.7, range: 0.3–124). Its mean value in the set of core genes we studied was 46.7 (range: 0.3–124).

Medium-Dependent Transcription Variability Correlates Negatively with Genetic Variability, While Strain-Dependent Variability Correlates Positively

In the third step of our study we looked for associations between transcription and genetic variability. We found that medium-dependent genes had a lower lability and diversity than other genes (0.01 vs. 0.03, P value $= 6 \times 10^{-4}$, and 0.017 vs. 0.019, P value $= 3 \times 10^{-4}$, respectively). Conversely, strain-dependent genes had a higher lability and diversity than other genes (0.04 vs. 0.02, P value $= 2 \times 10^{-4}$, 0.018 vs. 0.023, P value $= 2 \times 10^{-16}$). Similarly, strain-dependent genes had higher values of ROL than medium-dependent genes (51.3 vs. 41.4, P value $< 10^{-4}$). These findings are consistent with those of Vital *et al.*, who linked transcriptional variations to genetic variability and observed that highly differentially expressed genes showed increased coding-sequence dissimilarity (Vital et al. 2015). "Differential expression", in this case, was measured within each environmental condition and so is close to what we call strain variability.

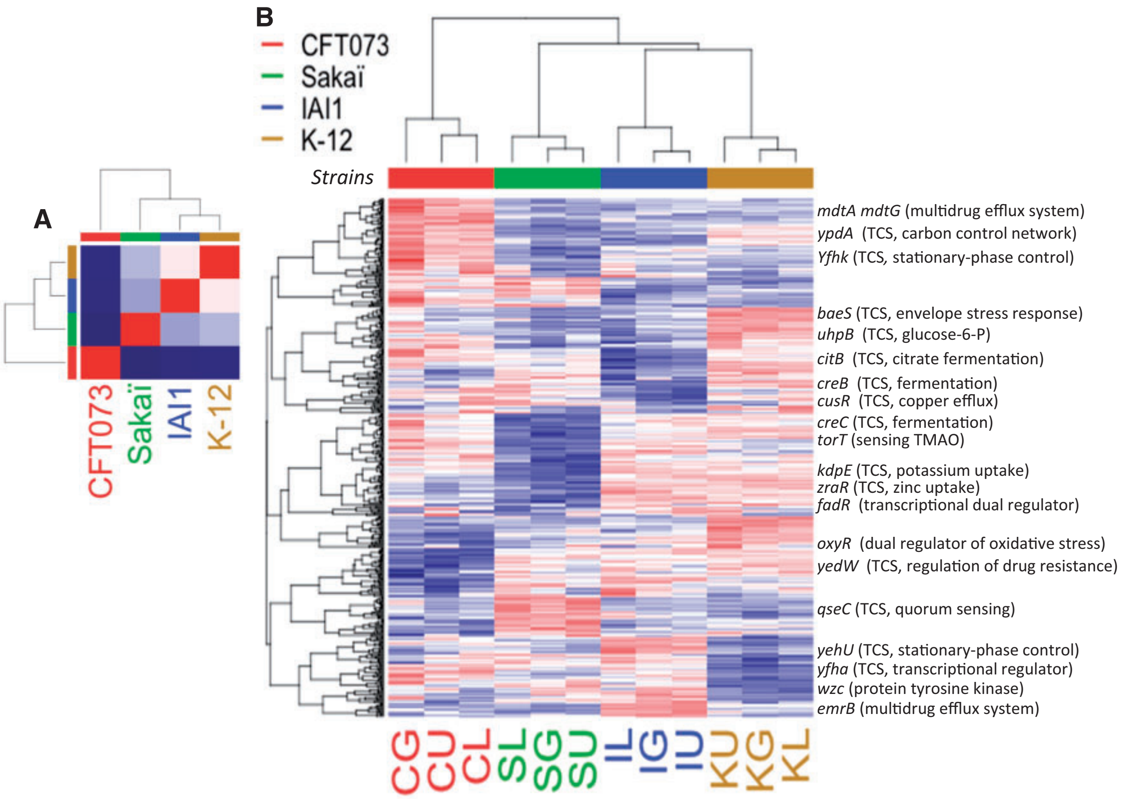

FIG. 3. Hierarchical clustering of the four *Escherichia coli* strains with genetic distances and with "strain dependent" gene expression. (*A*) The four strains were clustered using ANI as the similarity measure (see also supplementary table S8, Supplementary Material online). The dendrogram showed that the distance between CFT073 and the other strains was more significant than between the other strains (Sakaï, IAI1 and K-12), as previously observed. (*B*) The 500 most "strain-dependent" genes (rows) and the 12 experimental conditions (columns) were clustered using the rank correlation between transcript expressions as similarity measure (see also supplementary table S3, Supplementary Material online). C=CFT073, K=K-12, S=Sakaï, I=IAI1, G=Gluconate, U=Urine, L=LB. High transcript levels are marked in red and low levels are marked in blue.

We found that medium-dependent genes were more likely to be coded on the leading strand than were strain-dependent genes, although the difference is not quite statistically significant (57% vs. 54%, $P < 0.2$). This is consistent with the possibility that strain-dependent genes experience weaker negative selection than do medium-dependent genes (Srinivasan et al. 2015).

These effects were further studied with a quantitative approach rather than by comparison of groups of genes. For each gene, we quantified the variance among strains and among culture media (fig. 4). Variance of transcripts among media within strains was called "medium variance" and variance of transcripts among strains within media was named "strain variance". We found that medium variance was negatively correlated with gene lability ($\tau = -0.11$, $P < 2 \times 10^{-16}$), gene diversity ($\tau = -0.16$, $P < 2 \times 10^{-16}$) and ROL ($\tau = -0.09$, $P = 6 \times 10^{-14}$), whereas strain variance was positively correlated with gene lability ($\tau = 0.09$, $P = 1.5 \times 10^{-13}$), gene diversity ($\tau = 0.21$, $P < 2 \times 10^{-16}$) and ROL ($\tau = 0.1$, $P < 2 \times 10^{-16}$). Hence, we found that gene expression variability shaped gene evolution in two ways: Genes showing a great variability of expression among media had a low variability, and, on the contrary, genes whose expression was variable among strains had a high variability.

Associations between Transcriptomic Variance and Gene Variability Are Coupled with Selection

By definition, medium-dependent genes were important for all strains whereas the importance of strain-dependent genes was limited to one or two particular strains. Nichols et al. (2011) performed large-scale growth measurements of gene knock-out collections to determine gene essentiality across a large set of conditions. We observed that transcriptomic variations caused by culture media were higher in the essential genes they defined than in other genes (2.0 vs. 1.5, $P < 2.5 \times 10^{-7}$). The log of medium to strain variance was also higher in these essential genes: 0.16 versus -0.22, $P < 7 \times 10^{-11}$. Therefore, we hypothesized that the association between genetic variability and medium or strain variance was coupled to the intensity of selection acting on those genes. We computed nonsynonymous (Ka) and synonymous (Ks) substitution rates for each gene using divergence between K-12 and a closely related species *Escherichia* clade V, strain E1118. The Ka/Ks ratio can be used to infer the intensity of selective pressure exerted on a gene. A low ratio reflects an efficient purge by natural selection of the mutations affecting the proteins. As expected, the Ka/Ks values of the core genes shared by the four *E. coli* strains and the clade V bacteria were much lower than one (mean value: 0.05), illustrating that the core genome was under strong selective pressure.

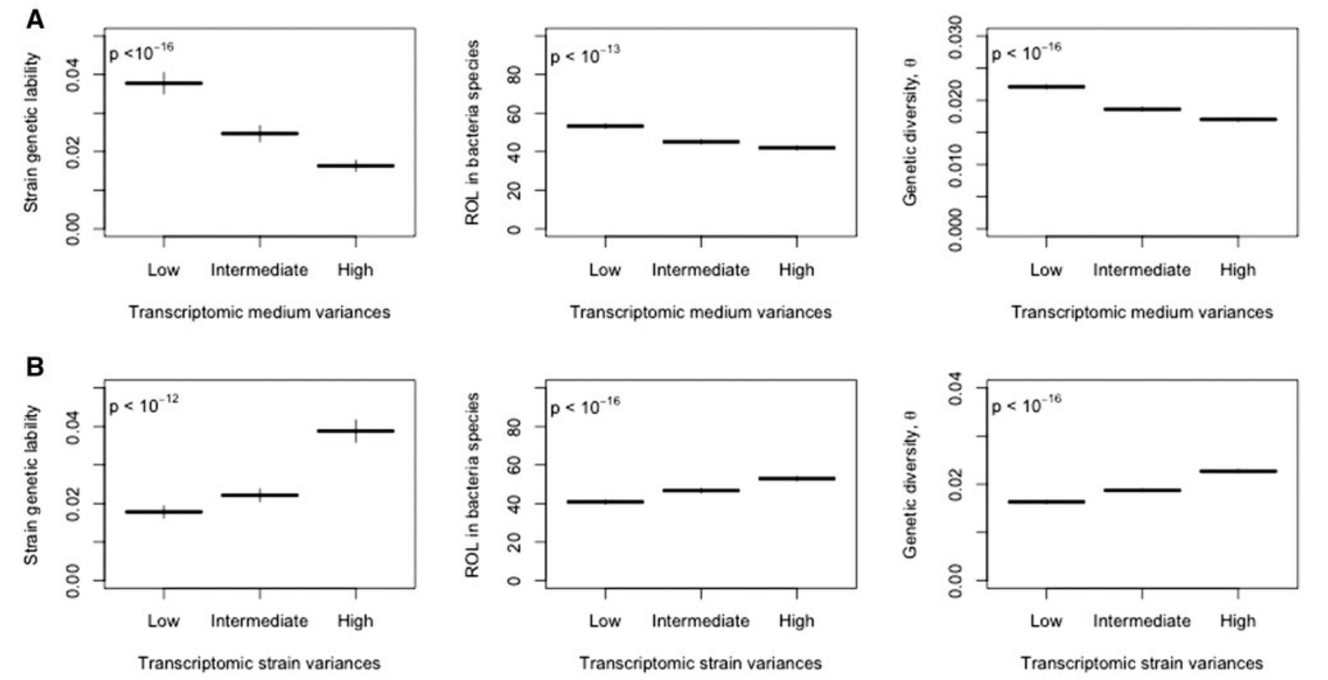

Fig. 4. Associations between transcriptomic “medium” and “strain” variances and genetic variability. Genes were classified into three groups (low, intermediate or high) according to their transcriptomic “medium” or “strain” variances (fig. 4*A* and *B*, respectively). Vertical bars indicate SEs. Associations between transcriptomic variances and genetic variability (strain lability, theta estimator and ROL index) were analyzed with Kendall’s correlation test (*P* indicates *P* values obtained with this test). θ is the per site theta Watterson estimator. ROL is the rate of loss index defined by Silander and Ackermann.

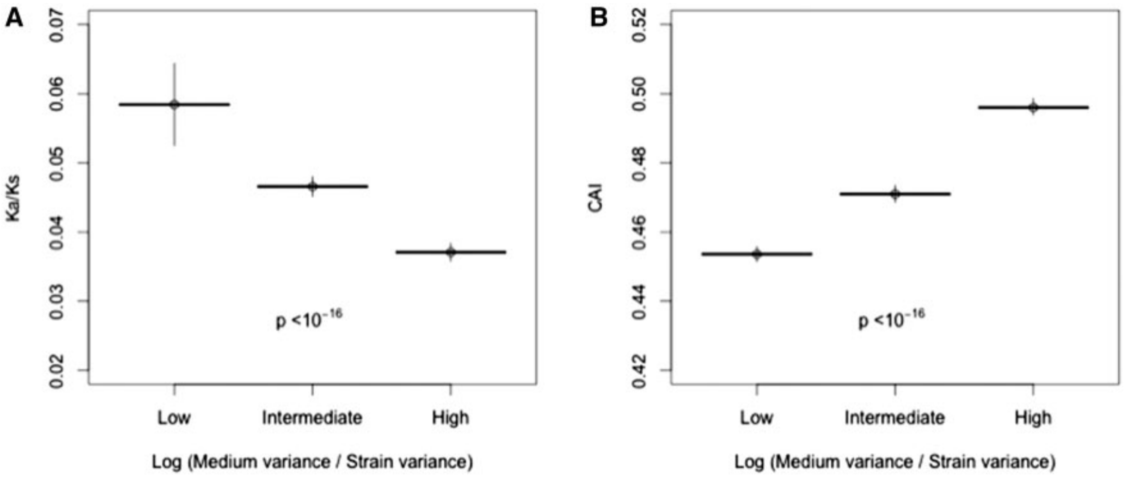

Fig. 5. Associations between medium/strain variance ratio and Ka/Ks and CAI. Genes were classified into three groups (low, intermediate and high) according to the log of their medium/strain variances ratio. Vertical bars indicate SEs and *P*, the *P* value obtained with Kendall’s correlation test. (*A*) Association between medium/strain variances log ratio and Ka/Ks (nonsynonymous and synonymous substitution rate computed between K-12 and *Escherichia* clade V, strain E1118). (*B*) Association between medium/strain variances log ratio and CAI.

Interestingly, the Ka/Ks ratio tended to decrease further with high medium variance ($\tau = -0.10$, $P = 7\times10^{-16}$), confirming that a greater selection was exerted on medium-dependent genes than on other genes. In contrast, the ratio increased with strain variance ($\tau = 0.11$, $P < 2\times10^{-16}$), revealing a somehow lower efficiency of selection being exerted on strain-dependent genes. This result can be summarized as a negative correlation between the medium versus strain variances ratio and the Ka/Ks ratio ($\tau = -0.11$, $P < 2\times10^{-16}$, fig. 5*A*).

The medium/strain variances ratio was also positively correlated with the CAI, a measure of codon bias ($\tau = 0.13$, $P < 9\times10^{-26}$, fig. 5*B*). A high codon bias reveals an efficient selection for optimal synonymous codons. It reflects the efficiency of selection exerted on a gene. Consequently, its association with the medium/strain variance ratio confirmed that

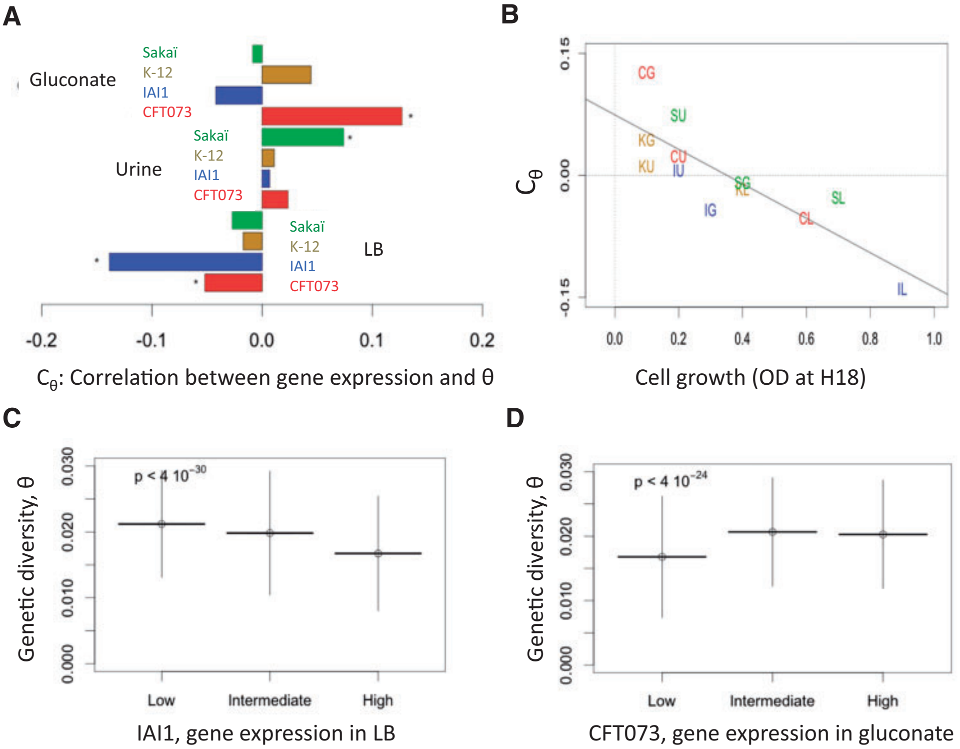

FIG. 6. Variations of the correlation between gene variability and transcript levels. (*A*) The correlation between gene variability (θ Watterson estimator) and transcription was named C_θ and calculated in the 12 culture conditions. Significant correlations are indicated with an asterisk. (*B*) Relation between C_θ (*y* axis) and cell growth (OD at 610 nm after 18 h of culture). (*C*) Example of negative correlation between mean transcript levels and genetic variability for IAI1 in LB medium. (*D*) Example of positive correlation between mean transcript levels and genetic variability for CFTO73 in gluconate medium. C=CFT073, K=K-12, S=Sakaï, I=IAI1, G=Gluconate, U=Urine, L=LB.

the links between genetic and transcriptomic variability were partly due to differences in selective pressure.

Correlation between Transcript Levels and Genetic Variability Depends on Growth Rate and on Strain Adaptation to Its Culture Medium

To explore the correlation between expression and genetic diversity we computed the values of the 12 correlation coefficients between transcript levels and Watterson's estimator of diversity, θ. We performed the analysis with data controlled for overall expression (column-scaled data, supplementary table S9, Supplementary Material online) and controlled for gene expression mean level (row-scaled data, supplementary table S11, Supplementary Material online). In both cases the correlations appeared to be highly variable. In the data controlled for overall cell expression, they ranged from −0.36 to −0.06. These negative values indicate that a high level of transcription (relatively to the average gene transcription level in the cells) is associated with gene stability, which is expected. When we used data controlled for gene expression mean level, we found both negative and positive values of the correlation that we named after C_θ (fig. 6 and supplementary figs. S1–S4, Supplementary Material online). The values of C_θ ranged from −0.14 to 0.13 and were the most negative in optimal LB medium, but close to null or positive in the less energetic media, urine or gluconate, suggesting that adaptation of the strain to the media could modulate C_θ (fig. 6A). To test this correlation, we quantified bacterial growth in each medium by the measure of the OD_{610} after 18 h of culture (table 1). There was a strong negative correlation between C_θ and bacterial growth efficiency ($r = -0.83$, $P = 8\times10^{-4}$ and $\tau = -0.7$, $P = 2\times10^{-3}$), suggesting that C_θ was negatively associated with strain adaptation to the media (fig. 6B). Using strain lability or the ROL index instead of θ, provided similar results (data not shown).

Our observations suggest that C_θ can be used as a proxy for the growth efficiency of a strain in a given media. If the strain grew efficiently in a media we observed a strong negative C_θ, if it grew poorly we observed a positive one. To confirm this result, we looked at the expression pattern in the condition giving the most positive value of C_θ: CFT073 in gluconate (fig. 6). We found that genes that were important for gluconate utilization and genes coding for the ATP synthase complexes were not induced (supplementary table S6, Supplementary Material online). This showed that the greatest C_θ value came from a strain-medium combination in which physiological adaptation was particularly poor.

To further explore the link between strain physiology in a given medium and C_θ, we identified genes whose expression showed a correlation with C_θ. About 852 gene transcripts were significantly correlated with C_θ (supplementary table S7, Supplementary Material online), some positively, others negatively (ranging from −0.93 to 0.94). Genes belonging to stress pathways were over-represented in the list of genes positively associated with C_θ (C_θ positive genes) (fig. 7). Examples of such genes were *yegS* (coding for a lipid kinase

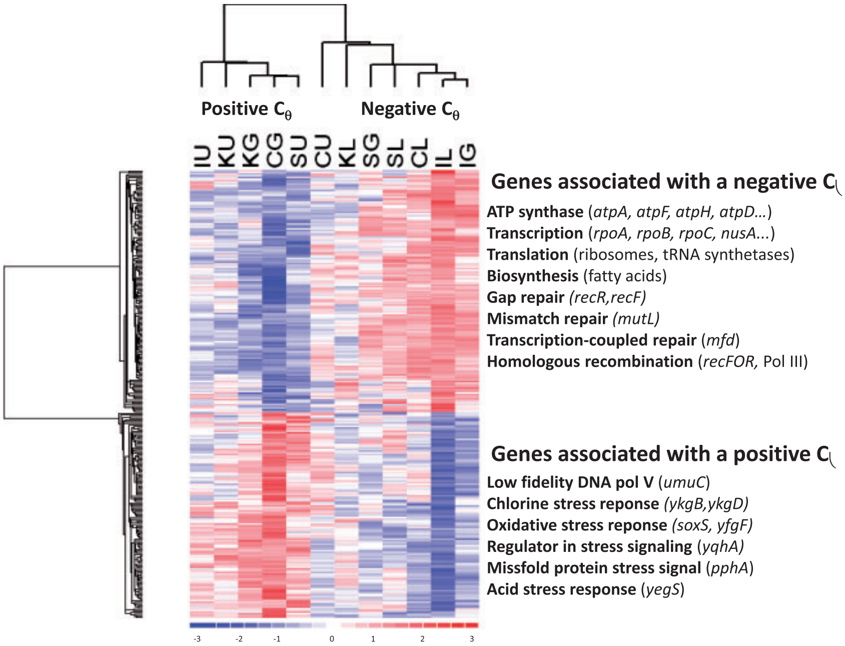

Fig. 7. Hierarchical clustering of the four *Escherichia coli* strains cultured in three media with core genes correlated to C_θ. Genes whose transcription correlated to C_θ (supplementary table S7, Supplementary Material online) and the 12 experimental conditions were clustered using the rank correlation between transcript levels as the similarity measure. C=CFT073, K=K-12, S=Sakaï, I=IAI1, G=Gluconate, U=Urine, L=LB. High transcript levels are marked in red and low levels are marked in blue.

involved in the response to acid stress; Maharjan et al. 2013), *ykgB* and *ykgD* (coding for enzymes involved in the chlorine stress response), *yqhA* (coding for an inner membrane protein, regulator in stress signaling), *pphA* (coding for a misfolded protein stress signal), *rcsA* (coding for an enzyme involved in sensing perturbations of the outer membrane and peptidoglycan; Salscheider et al. 2014), *soxS* and *yfgF* (involved in the oxidative stress response), *cspH* (coding for a cold-shock stress protein), and the global regulator *rpoS* (which controls the general stress response in *E. coli*; Maharjan and Ferenci 2015). Low-fidelity DNA polymerase V genes (*umuC* and *umuD*) were also positively associated with C_θ (see below).

On the contrary, pathways indicating efficient growth of the strain were over-represented in genes negatively associated with C_θ (fig. 7 and supplementary table S7, Supplementary Material online). These biological processes included ATP production (genes coding for ATP synthase and NADH-quinone oxidoreductase), DNA replication (genes coding for DNA polymerase III and nucleotide metabolism) transcription (RNA polymerase genes, except *rpoS*), translation (ribosomal genes and tRNA synthetase genes) and lipid biosynthesis (genes involved in terpenoid and fatty acid biosynthesis). High-fidelity DNA polymerase and DNA repair genes were also negatively associated with C_θ (see below).

Accordingly, we found that medium-dependent genes tended to have a negative value of correlation with C_θ (-0.06 vs. 0.03 for other genes, $P < 10^{-6}$), whereas strain-dependent genes tended to have a positive one (0.05 vs. -0.04 for other genes, $P < 10^{-7}$).

Contribution of Growth Efficiency and Gene Expression to Explain Gene Diversity

While the previous observations suggest an interesting association between C_θ and growth efficiency they do not provide a clear understanding on the underlying mechanisms. We first decided to disentangle the contribution of gene mean expression across conditions to that of expression according to growth efficiency, without considering C_θ. For each gene, we computed its mean gene expression across the 12 conditions used and computed the correlation between gene expression and the OD at 18 h. Using a multivariate regression, we could show that both mean expression and correlation with growth efficiency could be used to predict gene variability individually or in combination. The *R*-squared, which is a statistical measure of how close the data are to the fitted regression, had the following values: $r^2 = 8.1\%$ for gene mean expression, $r^2 = 5.1\%$ for pattern of expression, $r^2 = 11.7\%$ for both, all *P* values being $< 2 \times 10^{-16}$. In both cases the association with gene variability was negative, which suggests that: (1) as expected highly expressed genes evolved more slowly than others and (2) preferential expression in conditions of poor growth efficiency increased gene variability.

We then explored if genes having similar expression in nonstressful conditions had an increased genetic variability when their expression was high in conditions of poor growth. We binned genes according to their expression level in the conditions showing the lowest C_θ (IAI1 strain in LB); 1,000 bins were used. Associations between gene expression and gene variability were computed inside each bin in the condition showing the highest C_θ, (CFT073 strain in gluconate medium) and provided a weak but significantly positive correlation ($\tau = 0.05$, $P = 0.003$). This suggested that genes with equal expression level in good growth conditions had a higher diversity if they were more expressed in conditions of poor growth.

Positive C_θ in Low-Energy Media May Be Partly Explained by Differential Expression of DNA Repair Genes

Interestingly, the expression pattern of repair pathway genes was found to be linked with C_θ. Conditions in which C_θ was negative showed a relatively high transcription of *mutS* and *mutL* (mismatch repair), *mfd* (transcription-coupled repair), *recA*, *recFOR* and DNA polymerase III (gap repair, homologous recombination and high fidelity DNA replication, respectively) (fig. 7). This suggested that, in a nonstressful environment, DNA lesions are prevented by efficient DNA repair in spite of active transcription and replication. On the other hand, in conditions characterized by a positive C_θ, we observed a down-regulation of the aforementioned repair pathways and an active transcription of the low-fidelity DNA polymerase IV (*dinB*) and V (*umuCD*) genes known to introduce mutations by translesion synthesis and to play a role in adaptive mutations (Corzett et al. 2013). This differential expression of repair mechanism genes could contribute to the positive correlation found between transcription in a stressful environment and genetic diversity. The balance between the mutagenic effect of transcription and the intensity of purifying selection in highly expressed genes could be modulated by the availability of DNA repair enzymes. Selection may predominantly be exerted on genes highly expressed under unstressful conditions whereas TDM may prevail in genes highly expressed under stressful conditions.

Finally, we did not find an association between a gene's correlation with C_θ and whether it is transcribed from the lagging or leading strand that could have explained the above observation by strand-dependent variation in mutation rate.

Validation with External Data and Other Bacterial Species

To validate our results with another bacterial species subjected to different growth conditions, we carried out similar analyses with a public data set of gene expression measurements for *Salmonella enterica* grown under various stresses, taken at mid exponential growth phase (Kröger et al. 2013). Each gene expression (rows, see Materials and Methods) was scaled to the expression of the same gene found in a pooled RNA sample. We calculated Watterson's estimator, θ, in five *Samonella enterica* strains (see Materials and Methods) and used the same ROL index as for *E. coli*, as it was determined over "all" bacteria (including *E. coli* and *S. enterica*) (Silander and Ackermann 2009). Correlation coefficients between relative transcript levels and θ (C_θ) or ROL (C_{ROL}) were calculated for each condition. We confirmed that C_θ (ranging from -0.139 to 0.127) and C_{ROL} (ranging from -0.33 to 0.25) could be positive or negative and were both positively correlated with the level of stress. For instance, they were higher in pH 3 than in pH 4, higher at 15 °C than at 25 °C, higher in SPI2 medium (phosphate carbon nitrogen minimal medium inducing *Salmonella* pathogenicity island 2) than in LB, and higher in SPI2 medium with low magnesium than in SPI2 medium with a normal level of magnesium (supplementary fig. S5, Supplementary Material online). In the same data set, Kröger *et al.* also provide transcriptomes from bacteria cultured in optimal LB medium but at various growth phases. We found that C_θ and C_{ROL} were negative during exponential phases (minimum in the mid-exponential phase) and positive in stationary phases (maximum in late stationary phase), in which growth rate is low but also in which bacteria develop a multiple-stress-resistant state (fig. 8). As observed in *E. coli*, genes associated with positive C_θ were often genes implicated in stress responses (such as *yegS* and *pphA*), and genes associated with negative C_θ were often DNA repair genes (such as *mfd* or *endoV*). These results thus confirm that both stress and low growth rate are associated with positive C_θ and C_{ROL} values. As they are observed within species (C_θ) and between species (C_{ROL}) it may be said that both polymorphism (within species) and divergence (between species) are linked with gene expression.

Horizontally Transferred Genes Correlate Positively with C_θ

Horizontally transferred genes, defined by the use of Ecogene database annotations, are known to evolve faster than other genes (Lerat et al. 2005; Davids and Zhang 2008). The main explanation for this observation is the low selective pressure exerted on these genes compared with the core genome. As correlation with C_θ could be computed with any genes and not just on core genes, we also studied *E. coli* genes outside the core genomes (strain by strain analysis), and we confirmed that HGT genes had a stronger correlation with C_θ than other genes (0.13 vs. -0.05, $P < 2\times10^{-16}$). HGT genes did not by themselves explain our results (which remained very similar whether or not HGT genes were included, data not shown) but HGT evolution might be affected by this association between transcription and stressful conditions. Prophage sequences (which are usually not part of the core genome) in particular were more highly transcribed under stressful conditions (urine and gluconate in our experimental conditions) than in unstressful medium (LB) (data not shown). Consequently, these sequences had high correlations with C_θ (0.40 for prophage sequences vs. -0.05 for nonHGT sequences in K-12, $P < 2\times10^{-16}$). It is therefore possible that transcription under conditions lacking repair mechanisms could be an additional factor in the variability of phage sequences, in addition to mutations linked to phage proliferation and modular exchanges between prophages (Brüssow et al. 2004).

Discussion

One of the dogmas in molecular evolution is that highly expressed genes evolve slowly. However, this key observation in genomics relies on the postulate that expression of genes is

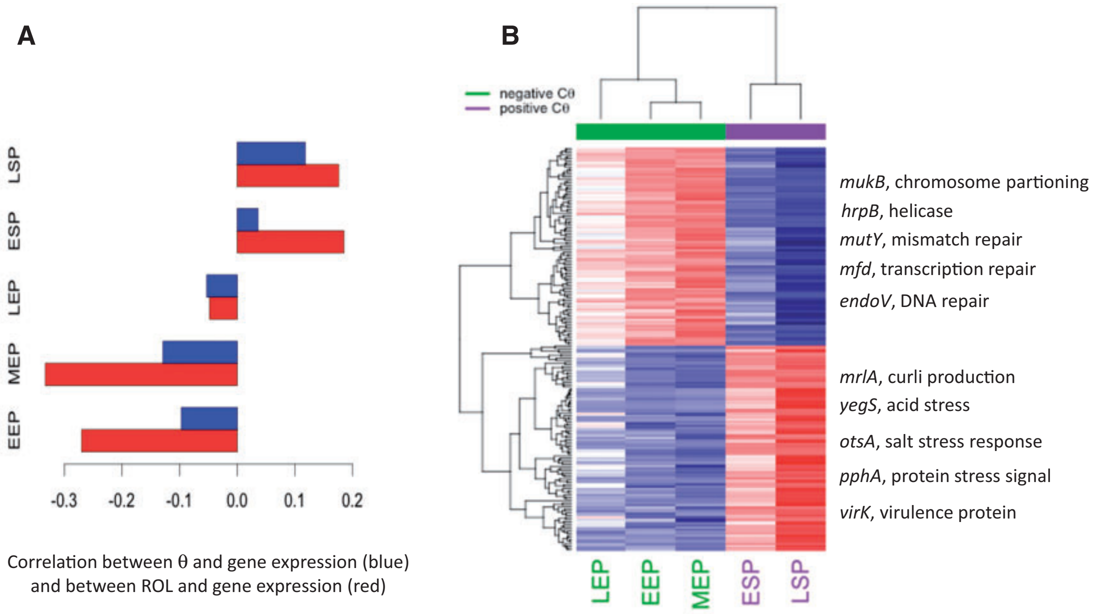

Fig. 8. C_θ and C_{ROL} in *Salmonella enterica* serovar Typhimurium. (*A*) The correlation coefficients between gene variability and expression level were named C_θ (correlation of gene expression with θ Watterson estimator computed in four *Salmonella* strains) and C_{ROL} (correlation with Rate Of Loss across "all bacteria"). Transcriptomes are from Kröger et al. (2013). Bacteria were grown in LB and harvested at six different growth phases: early exponential growth phase (EEP), mid-exponential growth phase (MEP), late exponential growth phase (LEP), early stationary phase (ESP) and late stationary phase (LSP). (*B*) Genes whose transcription correlated to C_θ in the six different growth phases were clustered using the rank correlation between transcript levels as the similarity measure. High transcript levels are marked in red and low levels are marked in blue.

constant, while in fact transcription may depend on the environment and on the strain. We decided to investigate how expression variability among environments and strains could affect gene evolution and diversity within *E. coli*. Genes whose expression pattern depended mainly on the medium, named medium-dependent genes, typically showed a strong pattern of induction in one medium in all strains. These genes were involved in basic cell functions whose response to the culture medium is central for physiological adaptation. Strain-dependent genes were classified as genes whose expression pattern varied mostly according to strains, possessing, e.g., an expression pattern that was specific to one or two strains in all culture media. Their sequence variability may be a causative factor for their expression variation, as they had similar phylogenetic and transcriptomic trees (fig. 3). Interestingly, they were associated with the response to environmental changes, as some of these genes were involved in multidrug efflux pumping or membrane transport. For instance, *nhaR* (Na^+/H^+ antiporter Regulator) was selectively up-regulated in the Sakaï strain and was recently found to be implicated in extra-intestinal virulence (Lescat et al. 2014). Moreover, many TCSs belonged to this class of strain-dependent genes. TCSs are used by bacteria to respond to their environment (Tobe 2008). Therefore, TCS expression variability suggests variability in the way strains "feel" their environment.

We then looked for links between transcript variability and markers of genetic diversity. Medium-dependent genes had a low genetic variability (fig. 4*A*) whereas strain-dependent genes had higher diversity (fig. 4*B*). The intensity of the stabilizing selection acting on these different categories of genes may explain such a pattern. Genes whose medium expression pattern is conserved among strains presumably include key genes for bacterial physiological adaptation. Genes whose expression varies according to the strain are genes that are less constrained, and could therefore tolerate higher nucleotide diversity. Such genes appear to be expressed at a low level in some strains, their overall expression being lower than of medium dependent genes (10.4 vs. 11.5 with the column-scaled data, supplementary table S9, Supplementary Material online, $P < 10^{-4}$), favoring even further mutation accumulation. Moreover, lower Ka/Ks ratios were associated with a higher "medium" signature, whereas higher Ka/Ks ratios were associated with a higher "strain" signature (fig. 5). The CAI, which reflects the intensity and efficiency of selection acting on synonymous codons, increased for medium-dependent genes and was low for strain-dependent genes (fig. 5). Overall, these correlations supported the idea that the balance between stabilizing selection and mutation pressure tilts more in favor of stabilizing selection for medium-dependent genes than for strain-dependent genes.

However, increased selection may not be the only driving force for the patterns that we observed. All genomic profiles are the result of three evolutionary forces: Mutation, selection and drift. In finite populations, slightly beneficial or deleterious mutations behave as though they were neutral, and so their accumulation in genomes is more strongly affected by mutation rate than by selection. Consequently, changes in mutation rate may also contribute to our observations. While a negative correlation between expression level and diversity is typically reported, we found that this correlation was modulated by the specific conditions of the experiment. Using data corrected for gene mean expression, the correlation was negative under conditions of efficient growth, but was positive when growth was poor. The correlation coefficient between expression and gene diversity (C_{θ}) was strongly associated with growth efficiency measured experimentally by optical density after 18 h of culture. That suggested that C_{θ} could be used to quantify growth efficiency of a strain in a medium. In order to test if results were specific to *E. coli*, to our data set or to our experimental methodologies, we used a recently published RNAseq data set (Kröger et al. 2013) obtained from a single strain of *Salmonella enterica* in many conditions. We recovered both negative and positive values of C_{θ}. The positive values being found in more stressful conditions (supplementary fig. S5, Supplementary Material online and fig. 8). Our observations thus hold true in another bacterial species: Increased expression under stressful conditions leads to increased variability. Moreover similar results were found for both polymorphism and divergence, suggesting that the underlying molecular mechanisms are robust.

Selection may explain the negative correlation between expression and diversity: The more a gene is expressed, the more costly the effect of deleterious mutations is, and so the more likely it is that mutations be purged by selection. Yet, it cannot directly explain why increased expression in stressful conditions is linked to increased gene variability. Evolution in stressful conditions is likely to be as frequent as proper growth conditions, as *E. coli* cycles between the gut and nutrient poor secondary environments (Tolla et al. 2015) and even within the gut, feast and famine cycles are found (Breton et al. 2016). The positive association between increased values of genetic variability and an increased transcription in stress conditions suggests that expression is not the only determinant of genetic variability and that growth conditions are also important. Variability of mutation rate arising from growth conditions and stress-induced mutagenesis, in addition to variability of selective pressure, might contribute to this variations of C_{θ}. We indeed found that repair genes, including the TCR gene, *mfd*, were highly expressed under conditions of negative C_{θ} and repressed under conditions of positive C_{θ} (figs. 7 and 8*B*). This means that, under conditions of poor growth, high expression may be linked to high mutagenesis because repair mechanisms are down-regulated. This is reminiscent of earlier results suggesting that stress-induced mutations may contribute to adaptive evolution (Bjedov et al. 2003). Many authors suggested that increased mutagenesis under stressful conditions may have been selected to promote adaptation to stress (Tenaillon et al. 2004). Some even suggested that this mutability targeted genes directly involved in the stress response (Galhardo et al. 2007). These hypotheses were highly controversial from an evolutionary perspective, but lead to the discovery of complex modes of diversification, from the recruitment of error-prone polymerases (Galhardo et al. 2009) to the transient amplification of genes (Sano et al. 2014). Wright and Kim (Kim and Jinks-Robertson 2012; Wright et al. 2013) connected these hypotheses to expression, as transcription exposes ssDNA and increases the rate of point mutation rate as well as diverse types of rearrangements such as deletions, duplications, inversions and translocations. As low growth rate, stress and defective TCR repair may occur simultaneously, miscoding of the damaged DNA during transcription in nondividing cells could also contribute to adaptive mutation (Morreall et al. 2015). While our data cannot be used to test whether any increase in mutagenesis has been selected for, they suggest that expression under stressful conditions seems to favor increased diversity and may preferentially target genes involved in stress response or environmental sensing.

The variability of mutation rate along the chromosome is also debated (Touchon et al. 2009). For instance, noncore genes and HGT genes have been shown to evolve faster than core genes (Davids and Zhang 2008). Yet, decoupling the effect of selection from the effect of mutation is not a trivial task. Using different genomic approaches, Martincorena et al. (2012) suggested that variability in mutation rate along the chromosome existed, and that highly expressed genes had a lower mutation rate. However, they did not explore the underlying mechanism. In a previous study, we suggested that the existence of DNA secondary structure might limit mutagenesis due to transcription. We found an excess of such structures in the genes showing high expression under optimal growth conditions (more precisely, in genes with high codon bias) (Hoede et al. 2006). Yet, it is difficult to determine whether the observed reduced mutation rate results from a direct selective pressure on mutation rate (which requires extremely efficient selection) or if it is a by-product of other selective pressures acting on RNA, such as RNA stability, for example (Park et al. 2013). The data presented here suggest a more global mechanism to explain fluctuation in mutation rate along the chromosome. Down-regulation of repair mechanisms coupled to high expression under poor growth conditions may lead to increased mutagenesis. This scenario is also consistent with the higher evolutionary rate found in horizontally acquired genes (Davids and Zhang 2008). We found that HGT genes have a higher expression under poor growth conditions. We suggest that the observed high rate of evolution of HGT genes is not only due to their low level of expression in optimal growth conditions, but also to their enhanced transcription under stressful conditions. This model might reconcile the conflicting results of Martincorena et al. (2012), who suggest that nonrandom mutation rates across genomes are due to the protection of highly expressed genes and of Maddamsetti et al. (2012), who argue that this heterogeneous pattern is due to the variability of HGT genes. In the present study, HGT genes do not account for our results, but our results suggest a scenario consistent with their high evolutionary rate.

In conclusion, this work suggests that variability in gene expression is linked to genetic polymorphism through a balance between selection and TDM that fluctuates according to the growth conditions. The mutagenic effect of transcription may be seen preferentially in genes expressed under poor growth conditions. As a result, genes highly expressed only under poor growth conditions may evolve faster, present lower codon bias and higher Ka/Ks ratios as the balance between mutation and selection tilts in favor of mutation for these genes.

Materials and Methods

Bacterial Strains and Culture Conditions

Three different media were used and named as followed: LB, gluconate (minimal medium with 10 mM gluconate) and urine (pool of filtered urine collected from seven healthy male volunteers according to the principles expressed in the declaration of Helsinki, with informed consent of the participants and in agreement with INSERM ethics regulation; table 1). The volunteers did not take any medication at least 1 month before urine sampling. Four *E. coli* strains were cultured: K-12 (K-12 MG1655; Blattner et al. 1997), IAI1 (Picard et al. 1999), CFT073 (Mobley et al. 1990) and Sakaï (O157:H7 Sakaï; Watanabe et al. 1996). Twelve core transcriptomes were obtained in triplicates and named KL, IL, CL, SL, KG, IG, CG, SG, KU, IU, CU, SU (K = K-12, I = IAI1, C = CFT073, S = Sakaï, L = LB, G = gluconate, U = urine). About 10^6 cells were inoculated into 40 ml of culture medium at 37 °C under micro-aerobic condition with a low flask to medium ratio under constant agitation (Partridge et al. 2007; Somerville and Proctor 2013). Growth rates were determined by OD_{610} measurement after 18 h of growth, as previously described (Sabarly et al. 2016). For RNA extraction, cells were harvested at the end of exponential phase, defined as the time when the second order derivative of growth as a function of time becomes negative. This time point was reached at an OD_{610} of 0.6, 0.2 and 0.3, and after a culture duration of 3 h, 3 h 30 min and 10 h, for cultures in LB, urine and gluconate, respectively. About 5 ml of LB cultures and 20 ml of urine and gluconate cultures (each containing $1–4\times10^8$ colony forming units) were removed from flasks and immediately mixed with 7 ml (LB cultures) or 25 ml (urine and gluconate cultures) of RNA protect (Qiagen) in order to block RNAses. The culture-RNA protected mixes were than centrifuged and pellets were kept frozen at − 80 °C until RNA extraction (Khodursky et al. 2003).

RNA Extraction and cDNA Synthesis

Culture pellets were allowed to thaw on ice and total RNA was extracted with NucleoSpin RNA II kit (Macherey-Nagel GmbH & Co KG, Germany) according to the manufacturer's instruction. RNA concentration and integrity were analyzed using an Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Santa Clara, CA).

Ten micrograms of each total prokaryotic RNA extraction were used to generate cDNA with Invitrogen's SuperScript double-stranded cDNA Synthesis Kit (Thermo Fisher Scientific Inc., Waltham, MA). cDNA was then sent to Roche Nimblegen Inc. (500 South Rosa Road, Madison, WI) which performed the Cy3-labeling and hybridization of samples to the "pan-coli" chip (see below).

Design of a "Pan-Coli" Oligonucleotide Microarray and Data Processing

A whole-genome open reading frame (ORF) array for *E. coli* strains was designed by Roche Nimblegen Inc. (500 South Rosa Road, Madison, WI) based on published complete genome sequences (Touchon et al. 2009). Whenever the genes were long enough, six oligomeric probe pairs per ORF were generated (128,595 probes). Each pair consisted of a full-match oligomer and a close oligomer which differed by only one mismatch. Subtraction of the single-mismatch reading to the full-match reading for the six probe pairs allowed a highly specific quantification of each gene expression. Raw data were extracted from scanned images using NimbleScan software (version 2.6.0.0). Log2 normalized expression values were generated using the Robust Multichip Average (RMA) algorithm (Irizarry et al. 2003). The numbers of studied genes were 4,398, 5,446, 4,244 and 5,347 for IAI1, CFT073, K-12 and Sakaï, respectively. The four strains shared a core genome of 3,281 genes. Each gene of each strain had its own specific set of six probe pairs (full match and single mismatch oligomers) in order to avoid potential artifacts due to minor sequence variations. For the study of the core transcriptome, microarray data were merged into one single data matrix with the 3,281 core gene transcript expression levels (rows), and the four bacteria in the three media (12 columns). A quantile normalization of columns provided a first level of processed data in which transcript levels had the same distribution in all bacteria (supplementary table S9, Supplementary Material online). We then standardized both rows and columns after successive column–row standardizations and performed further analyses with those data (supplementary table S11, Supplementary Material online). This approach is recommended in microarrays studies to prevent incidental gene differences from obscuring the actual effects of interest (Efron 2009).

Transcriptome Variability

Most of the statistical analyses were carried out with the R 3.1.2 statistical software (R Foundation for Statistical Computing, Vienna, Austria). To find genes significantly associated with strains or with media we applied the linear model (lm R command) of analysis of variance. Variances of transcript levels of one strain in the different media or of the different strains in one medium were estimated by mean squares (between group variations divided by the number of degrees of freedom). In order to improve the reliability of our results, we used a resampling procedure with replicates: Medium and strain *P* values were calculated in 47 different resampling groups of 12 samples (4 strains × 3 media) (supplementary table S10, Supplementary Material online). Concordance between *P* values could be estimated for each gene as the percentage of the same significant *P* values found in the 47 groups. The mean concordance was 98% ± 0.07 and 97% ± 0.09 for medium and strain genes, respectively (supplementary table S10, Supplementary Material online). In other words, the 47 different groups of 12 samples provided the same lists of medium and strain genes (supplementary tables S1 and S3, Supplementary Material online). Specific interactions between media and strains were determined by comparing linear models with or without interaction (with ANOVA R command). False discovery rates (FDR) associated with *P* values were calculated using the p.adjust R command (Benjamini and Hochberg method). Associations with FDR of < 0.05 were considered significant. Gene annotations such as bioprocess classification or CAI were downloaded from

MAGE web site (Microbial Genome Annotation & Analysis Platform, www.genoscope.cns.fr, last accessed May 31, 2016). Genes acquired by HGT were annotated with Ecogene database (www.ecogene.org, last accessed May 31, 2016). Over- or under-represented annotations in the selected genes were estimated with hypergeometric tests. Hierarchical clustering was performed with the heatmap3 R package, using rank correlation as the similarity measure and average linkage as the clustering method.

Genetic Variability

Average nucleotide identity (ANI) between K-12, IAI1, CFT073, Sakaï *E. coli* strains, was calculated according to Goris et al. (2007) using the Konstantinidis' laboratory website (http://enve-omics.ce.gatech.edu/ani, last accessed May 31, 2016). Gene lability and diversity were determined in those strains and 114 phylogenetically diverse other *E coli* strains whose genome sequences were publicly available (Lescat et al. 2014) (supplementary table S5, Supplementary Material online). The sequences were aligned and studied for diversity with a homemade Perl script (Touchon et al. 2009), MUSCLE v3.8.31 program (Edgar 2004) and the PopGenome R package. Gene lability was defined as the percentage of strains in which the gene did not exist, using K-12 gene names as the reference (supplementary table S1, Supplementary Material online). Gene lability was measured for the 3,281 core genes of K-12, IAI1, CFT073, and Sakaï among the other 114 *E coli* strains. Gene diversity among strains was evaluated with the per site θ Watterson estimator (Watterson 1975). For *Salmonella enterica*, five strains were analyzed: 4/74 [the strain used by Krögler et al. (2013)], SL1344, 14028S, LT2 and CT18 (whose sequences were downloaded from the Mage website, www.genoscope.cns.fr/agc/microscope, last accessed May 31, 2016). Their core genome was identified using the cd-hit-v4.6.4 program (Li and Godzik 2006). Gene instability over all bacterial species was estimated with the ROL "all" index (rate of loss in "all bacteria") defined by Silander and Ackermann (2009), which is similar to lability but uses different species rather than different strains within a species. We calculated the ratio of the number of nonsynonymous substitutions per nonsynonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks) with the seqinr package (Charif et al. 2005).

Associations between Transcriptome and Genetic Variability

To provide easy-to-read graphs, mean transcript levels and variances were discretized in three class intervals (low, intermediate, high) with two quantile breaks (classInt R package). Associations between transcript levels and genetic variability (strain lability, θ estimator and ROL index) were evaluated by correlation tests and linear regression, genetic variability being the dependent variable. Correlation between transcript levels and the θ estimator was named C_θ and calculated for the 12 points (KL, IL, CL, SL, KG, IG, CG, SG, KU, IU, CU, SU). Correlations between transcript levels and variations of C_θ among these 12 conditions were then calculated for each gene. We carried out the same calculation with transcritpomic data from *Salmonella enterica* provided by Krögler et al. (2013). We similarly named C_{ROL} the correlation between transcript levels and ROL.

Correlations were determined with the cor.test R command using the nonparametric Kendall's τ statistic and the Pearson's *r* coefficient. Multivariate regressions were carried out with the glm R command.

Supplementary Material

Supplementary figure S5 and tables S1–S11 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was partially supported by the "Fondation pour la Recherche Médicale" (J.T. and E.D.) and an ERC grant (ERC-2012-StG_20111109, O.T.). We thank Harry Kemble for carefully editing the article.

References

Aubron C, Glodt J, Matar C, Huet O, Borderie D, Dobrindt U, Duranteau J, Denamur E, Conti M, Bouvet O. 2012. Variation in endogenous oxidative stress in *Escherichia coli* natural isolates during growth in urine. *BMC Microbiol*. 12:120.

Bjedov I, Tenaillon O, Gérard B, Souza V, Denamur E, Radman M, Taddei F, Matic I. 2003. Stress-induced mutagenesis in bacteria. *Science* 300:1404–1409.

Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453–1462.

Breton J, Tennoune N, Lucas N, Francois M, Legrand R, Jacquemot J, Goichon A, Guérin C, Peltier J, Pestel-Caron M, et al. 2016. Gut commensal *E. coli* proteins activate host satiety pathways following nutrient-induced bacterial growth. *Cell Metab*. 23:324–334.

Bridges BA. 1994. Starvation-associated mutation in *Escherichia coli*: a spontaneous lesion hypothesis for "directed" mutation. *Mutat Res*. 307:149–156.

Brüssow H, Canchaya C, Hardt W-D. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev*. 68:560–602, table of contents.

Bryant JA, Sellars LE, Busby SJW, Lee DJ. 2014. Chromosome position effects on gene expression in *Escherichia coli* K-12. *Nucleic Acids Res*. 42:11383–11392.

Chang D-E, Smalley DJ, Tucker DL, Leatham MP, Norris WE, Stevenson SJ, Anderson AB, Grissom JE, Laux DC, Cohen PS, et al. 2004. Carbon nutrition of *Escherichia coli* in the mouse intestine. *Proc Natl Acad Sci U S A*. 101:7427–7432.

Charif D, Thioulouse J, Lobry JR, Perrière G. 2005. Online synonymous codon usage analyses with the ade4 and seqinR packages. *Bioinforma Oxf Engl*. 21:545–547.

Chen X, Zhang J. 2013. No gene-specific optimization of mutation rate in *Escherichia coli*. *Mol Biol Evol*. 30:1559–1562.

Chursov A, Frishman D, Shneider A. 2013. Conservation of mRNA secondary structures may filter out mutations in *Escherichia coli* evolution. *Nucleic Acids Res*. 41:7854–7860.

Corzett CH, Goodman MF, Finkel SE. 2013. Competitive fitness during feast and famine: how SOS DNA polymerases influence physiology and evolution in *Escherichia coli*. *Genetics* 194:409–420.

Davids W, Zhang Z. 2008. The impact of horizontal gene transfer in shaping operons and protein interaction networks–direct evidence of preferential attachment. *BMC Evol Biol*. 8:23.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.

Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet*. 10:715–724.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.

Efron B. 2009. Are a set of microarrays independent of each other? *Ann Appl Stat*. 3:922–942.

Galhardo RS, Do R, Yamada M, Friedberg EC, Hastings PJ, Nohmi T, Rosenberg SM. 2009. DinB upregulation is the sole role of the SOS response in stress-induced mutagenesis in *Escherichia coli*. *Genetics* 182:55–68.

Galhardo RS, Hastings PJ, Rosenberg SM. 2007. Mutation as a stress response and the regulation of evolvability. *Crit Rev Biochem Mol Biol*. 42:399–435.

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol*. 57:81–91.

Hoede C, Denamur E, Tenaillon O. 2006. Selection acts on DNA secondary structures to decrease transcriptional mutagenesis. *PLoS Genet*. 2:e176.

Imam S, Noguera DR, Donohue TJ. 2015. An integrated approach to reconstructing genome-scale transcriptional regulatory networks. *PLoS Comput Biol*. 11:e1004103.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostat Oxf Engl*. 4:249–264.

Khodursky AB, Bernstein JA, Peter BJ, Rhodius V, Wendisch VF, Zimmer DP. 2003. *Escherichia coli* spotted double-strand DNA microarrays: RNA extraction, labeling, hybridization, quality control, and data management. *Methods Mol Biol*. 224:61–78.

Kim N, Jinks-Robertson S. 2012. Transcription as a source of genome instability. *Nat Rev Genet*. 13:204–214.

Kröger C, Colgan A, Srikumar S, Händler K, Sivasankaran SK, Hammarlöf DL, Canals R, Grissom JE, Conway T, Hokamp K, et al. 2013. An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium. *Cell Host Microbe*. 14:683–695.

Le Gall T, Darlu P, Escobar-Páramo P, Picard B, Denamur E. 2005. Selection-driven transcriptome polymorphism in *Escherichia coli*/Shigella species. *Genome Res*. 15:260–268.

Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol*. 3:e130.

Lescat M, Reibel F, Pintard C, Dion S, Glodt J, Gateau C, Launay A, Ledda A, Cruvellier S, Tourret J, et al. 2014. The conserved nhaAR operon is drastically divergent between B2 and non-B2 *Escherichia coli* and is involved in extra-intestinal virulence. *PloS One* 9:e108738.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.

Maddamsetti R, Hatcher PJ, Cruveiller S, Médigue C, Barrick JE, Lenski RE. 2012. Horizontal gene transfer may explain variation in θs. ArXiv12100050 Q-Bio [Internet]. [cited 2016 May 31]. Available from: http://arxiv.org/abs/1210.0050.

Maharjan R, Ferenci T. 2015. Mutational signatures indicative of environmental stress in bacteria. *Mol Biol Evol*. 32:380–391.

Maharjan RP, Gaffé J, Plucain J, Schliep M, Wang L, Feng L, Tenaillon O, Ferenci T, Schneider D. 2013. A case of adaptation through a mutation in a tandem duplication during experimental evolution in *Escherichia coli*. *BMC Genomics* 14:441.

Martincorena I, Seshasayee ASN, Luscombe NM. 2012. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* 485:95–98.

Meysman P, Collado-Vides J, Morett E, Viola R, Engelen K, Laukens K. 2014. Structural properties of prokaryotic promoter regions correlate with functional features. *PloS One* 9:e88717.

Mobley HL, Green DM, Trifillis AL, Johnson DE, Chippendale GR, Lockatell CV, Jones BD, Warren JW. 1990. Pyelonephritogenic *Escherichia coli* and killing of cultured human renal proximal tubular epithelial cells: role of hemolysin in some strains. *Infect Immun*. 58:1281–1289.

Morreall J, Kim A, Liu Y, Degtyareva N, Weiss B, Doetsch PW. 2015. Evidence for retromutagenesis as a mechanism for adaptive mutation in *Escherichia coli*. *PLoS Genet*. 11:e1005477.

Nichols RJ, Sen S, Choo YJ, Beltrao P, Zietek M, Chaba R, Lee S, Kazmierczak KM, Lee KJ, Wong A, et al. 2011. Phenotypic landscape of a bacterial cell. *Cell* 144:143–156.

Okuda S, Kawashima S, Kobayashi K, Ogasawara N, Kanehisa M, Goto S. 2007. Characterization of relationships between transcriptional units and operon structures in *Bacillus subtilis* and *Escherichia coli*. *BMC Genomics* 8:48.

Park C, Chen X, Yang J-R, Zhang J. 2013. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. 110:E678–E686.

Park C, Zhang J. 2012. High expression hampers horizontal gene transfer. *Genome Biol Evol*. 4:523–532.

Partridge JD, Sanguinetti G, Dibden DP, Roberts RE, Poole RK, Green J. 2007. Transition of *Escherichia coli* from aerobic to micro-aerobic conditions involves fast and slow reacting regulatory components. *J Biol Chem*. 282:11230–11237.

Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N, Bingen E, Elion J, Denamur E. 1999. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infect Immun*. 67:546–553.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*. 12:32–42.

Roux A, Xu Y, Heilier J-F, Olivier M-F, Ezan E, Tabet J-C, Junot C. 2012. Annotation of the human adult urinary metabolome and metabolite identification using ultra high performance liquid chromatography coupled to a linear quadrupole ion trap-Orbitrap mass spectrometer. *Anal Chem*. 84:6429–6437.

Sabarly V, Aubron C, Glodt J, Balliau T, Langella O, Chevret D, Rigal O, Bourgais A, Picard B, de Vienne D, et al. 2016. Interactions between genotype and environment drive the metabolic phenotype within *Escherichia coli* isolates. *Environ Microbiol*. 18:100–117.

Sabarly V, Bouvet O, Glodt J, Clermont O, Skurnik D, Diancourt L, de Vienne D, Denamur E, Dillmann C. 2011. The decoupling between genetic structure and metabolic phenotypes in *Escherichia coli* leads to continuous phenotypic diversity. *J Evol Biol*. 24:1559–1571.

Salscheider SL, Jahn A, Schnetz K. 2014. Transcriptional regulation by BglJ-RcsB, a pleiotropic heteromeric activator in *Escherichia coli*. *Nucleic Acids Res*. 42:2999–3008.

Sano E, Maisnier-Patin S, Aboubechara JP, Quiñones-Soto S, Roth JR. 2014. Plasmid copy number underlies adaptive mutability in bacteria. *Genetics* 198:919–933.

Selby CP, Sancar A. 1993. Molecular mechanism of transcription-repair coupling. *Science* 260:53–58.

Sheldon JR, Yim M-S, Saliba JH, Chung W-H, Wong K-Y, Leung KT. 2012. Role of rpoS in *Escherichia coli* O157:H7 strain H32 biofilm development and survival. *Appl Environ Microbiol*. 78:8331–8339.

Silander OK, Ackermann M. 2009. The constancy of gene conservation across divergent bacterial orders. *BMC Res Notes*. 2:2.

Somerville GA, Proctor RA. 2013. Cultivation conditions and the diffusion of oxygen into culture media: the rationale for the flask-to-medium ratio in microbiology. *BMC Microbiol*. 13:9.

Srinivasan R, Scolari VF, Lagomarsino MC, Seshasayee ASN. 2015. The genome-scale interplay amongst xenogene silencing, stress response and chromosome architecture in *Escherichia coli*. *Nucleic Acids Res*. 43:295–308.

Tenaillon O, Denamur E, Matic I. 2004. Evolutionary significance of stress-induced mutagenesis in bacteria. *Trends Microbiol*. 12:264–270.

Tobe T. 2008. The roles of two-component systems in virulence of pathogenic *Escherichia coli* and Shigella spp. *Adv Exp Med Biol*. 631:189–199.

Tolla DA, Kiley PJ, Lomnitz JG, Savageau MA. 2015. Design principles of a conditional futile cycle exploited for regulation. *Mol Biosyst*. 11:1841–1849.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet*. 5:e1000344.

Vital M, Chai B, Østman B, Cole J, Konstantinidis KT, Tiedje JM. 2015. Gene expression analysis of *E. coli* strains provides insights into the role of gene regulation in diversification. *ISME J*. 9:1130–1140.

Watanabe H, Wada A, Inagaki Y, Itoh K, Tamura K. 1996. Outbreaks of enterohaemorrhagic *Escherichia coli* O157:H7 infection by two different genotype strains in Japan, 1996. *Lancet* 348:831–832.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*. 7:256–276.

Wright BE, Schmidt KH, Minnick MF. 2013. Kinetic models reveal the in vivo mechanisms of mutagenesis in microbes and man. *Mutat Res*. 752:129–137.