SCIENCE CHINA
Life Sciences

SPECIAL TOPIC:
• RESEARCH PAPER •

doi: 10.1007/s11427-014-4734-y

Understanding the commonalities and differences in genomic organizations across closely related bacteria from an energy perspective

MA Qin[1,2†], CHEN Xin[1,3†], LIU Chao[1,4], MAO XiZeng[1,2], ZHANG HanYuan[1,3], JI Fei[1,2], WU ChunGuo[1,3] & XU Ying[1,2,3*]

[1]*Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA;*
[2]*BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA,*
[3]*College of Computer Science and Technology, Jilin University, Changchun 130012, China;*
[4]*Provincial Hospital Affiliated to Shandong University, Jinan 250021, China*

Received May 25, 2014; accepted July 11, 2014

The availability of a large number of sequenced bacterial genomes facilitates in-depth studies about why genes (operons) in a bacterial genome are globally organized the way they are. We have previously discovered that (the relative) transcription-activation frequencies among different biological pathways encoded in a genome have a dominating role in the global arrangement of operons. One complicating factor in such a study is that some operons may be involved in multiple pathways with different activation frequencies. A quantitative model has been developed that captures this information, which tends to be minimized by the current global arrangement of operons in a bacterial (and archaeal) genome compared to possible alternative arrangements. A study is carried out here using this model on a collection of 52 closely related *E. coli* genomes, which revealed interesting new insights about how bacterial genomes evolve to optimally adapt to their environments through adjusting the (relative) genomic locations of the encoding operons of biological pathways once their utilization and hence transcription activation frequencies change, to maintain the above energy-efficiency property. More specifically we observed that it is the frequencies of the transcription activation of pathways relative to those of the other encoded pathways in an organism as well as the variation in the activation frequencies of a specific pathway across the related genomes that play a key role in the observed commonalities and differences in the genomic organizations of genes (and operons) encoding specific pathways across different genomes.

genomic organization, transcription activation frequency, pathway modeling, comparative genomics analysis

Citation: Ma Q, Chen X, Liu C, Mao XZ, Zhang HY, Ji F, Wu CG, Xu Y. Understanding the commonalities and differences in genomic organizations across closely related bacteria from an energy perspective. Sci China Life Sci, doi: 10.1007/s11427-014-4734-y

E. coli is by far the best-studied group of organisms on earth. Substantial amount of information has been derived about this class of model bacteria in the past century [1], which has served as the foundation for biological studies of many other bacterial organisms. For example, fundamental concepts in bacteriology such as operons [2], regulons [3] and two-component systems [4], all discovered in *E. coli*, have been mapped to other bacteria. With the rapid advancement of the (next-generation) sequencing techniques, bacterial genomes are being sequenced and functionally annotated at an accelerated rate, which has enabled studies aimed to ad-

†Contributed equally to this work
*Corresponding author (email: xyn@bmb.uga.edu)

dress deeper and broader biological questions, including how various functional systems such as metabolic and regulatory pathways are encoded in an *E. coli* genome [5–7]. As of now, 52 complete genomes of the *E. coli* species have been sequenced and made publicly available. Comparative analyses of these closely related genomes have led to a number of interesting observations. For example, different strains of *E. coli* share a high proportion of the protein-encoding genes [8].

These sequenced genomes within the *E. coli* species have also offered a unique opportunity for in-depth studies of the organizational principles of functional elements encoded in a genome, for example, the rules that may govern how operons are globally arranged in a genome. We have previously discovered one general organizational principle of bacterial genomes: that is, more frequently activated biological pathways tend to have their component operons localized into a smaller number of DNA-folding domains, termed *supercoil domains* or *supercoils* [9,10]. More specifically, the total number of supercoil unfoldings in the folded chromosome of a bacterial cell, needed to transcriptionally activate all the obligatory pathways during the life cycle of a cell, tends to be minimized [11]. Our interpretation of this observation is that a bacterial genome tends to evolve so that its operons are globally arranged in such a way that the total energy for unfolding (and then refolding) the relevant DNA segments needed to make the relevant genes transcriptionally accessible in response to various stimuli during the life cycle of a bacterial cell is minimized. We have defined a simple model for estimating this total energy along with a prediction of the set of supercoils, i.e., a partition of the whole genome into 10–100 kb genomic regions, that achieves the minimal energy [11], which is termed as the energy score of a genome throughout this paper. This model explains very well the current global arrangement of operons in a variety of bacterial and archaeal genomes, which fall into a wide range of taxonomical groups, as we have previously demonstrated [11,12]. In this paper, we present a study on elucidation of the possible reasons for the observed diversity in the genomic organizations of operons across the 52 *E. coli* genomes, using this model.

We have applied the model to the 52 *E. coli* genomes. It is worth noting that our model requires the information of the transcriptional activation frequency of each pathway, which can be estimated based on a large number of genome-scale transcriptomic data collected on multiple conditions as we have done in our previous studies [11]. However, such data are not available for the organisms under consideration except for one, *E. coli* K_12_MG1655. To alleviate this issue, we have systematically tested a simplified version of the model, specifically by assuming all pathways have the same transcriptional activation frequencies, and found that this version achieves comparable results in estimating the set of supercoils that achieve the minimal energy score [12], hence substantially expanding the scope of genomes that we can study using this model.

Using this simplified model, we have predicted the supercoils for each of the 52 genomes and then compared the supercoils across different *E. coli* genomes, to address what may have caused the observed diversity in the genomic organizations of operons across the 52 genomes. The main finding indicates that the diversity in terms of the numbers of supercoils encoding the same pathways across different gnomes is largely due to the variation in the relative (transcription) activation frequencies of the pathways.

Throughout this paper, we use genes and operons (including single-gene operons) interchangeably.

1 Materials and methods

1.1 Data

The sequenced genomes of 52 *E. coli* strains are downloaded from the NCBI download site as of 11/01/2013 and the details about these genomes can be found in Table 1. Five hundred and fifty-five microarray gene-expression datasets of four *E. coli* strains are retrieved from the GEO database [46] and the M3D database [47] for estimating the frequencies of transcription activation of individual pathways. Specifically, 40 datasets are found and retrieved for K_12_W3110 [48], 26 sets for O157_H7_Sakai [49], 23 sets for E24377A [50] and 466 sets for K_12_MG1655 [47]. Biological pathway information used in this study was collected from the KEGG database [7]. The operon data for the 52 genomes are obtained from the DOOR2 database [5], which are needed for the supercoil prediction and orthologous gene mapping across the genomes. In addition, the growth rate data of 10 *E. coli* strains, measured under the same growth condition with citric acid and acetate among other nutrients, are retrieved from [51].

1.2 Analyses of transcriptomic data

The microarray gene-expression data used in this study are normalized using the AFFY package in Bioconductor [52]. In addition, the 466 sets of microarray data of K_12_MG1655 are grouped into seven collections corresponding to seven generally studied physiological conditions, namely anaerobic, heat shock, exponential growth, nitrogen limitation, oxidative, stationary growth, and SOS response. A set of marker genes for each condition have been collected from the literature, which collectively define the condition. A microarray dataset among the 466 datasets is assigned to a condition-specific subset if the marker genes of the dataset show the same expression pattern as the defining expression pattern of the condition. The detailed information can be found in [11].

Table 1 Fifty-two *E. coli* strains with their IDs, energy scores, numbers of supercoils and relative reference to their sequences

E. coli strain	Chromosome	Energy scores	#Supercoils	Reference
K_12_substr__MG1655_uid57779	NC_000913	4115.19	79	[13]
O157_H7_Sakai_uid57781	NC_002695	4233.48	85	[14]
CFT073_uid57915	NC_004431	4214.08	83	[15]
K_12_substr__W3110_uid161931	NC_007779	4135.24	77	[16]
coli_UTI89_uid58541	NC_007946	4197.22	82	[17]
536_uid58531	NC_008253	4238.2	72	n/a
APEC_O1_uid58623	NC_008563	4115.99	79	[18]
HS_uid58393	NC_009800	4091.26	78	n/a
E24377A_uid58395	NC_009801	4179.27	83	[19]
K_12_substr__DH10B_uid58979	NC_010473	4035.45	77	[20]
SMS_3_5_uid58919	NC_010498	4273.11	82	[21]
O157_H7_EC4115_uid59091	NC_011353	4208.03	89	[22]
SE11_uid59425	NC_011415	4153.56	77	[23]
O127_H6_E2348_69_uid59343	NC_011601	4146.9	74	[24]
IAI1_uid59377	NC_011741	4145.65	71	[25]
S88_uid62979	NC_011742	4153.41	77	[25]
ED1a_uid59379	NC_011745	4158.46	88	[25]
55989_uid59383	NC_011748	4127.08	79	[25]
IAI39_uid59381	NC_011750	4225.11	81	[25]
UMN026_uid62981	NC_011751	4236.82	85	[25]
LF82_uid161965	NC_011993	4124.2	74	n/a
BW2952_uid59391	NC_012759	4073.42	77	[26]
BL21_Gold_DE3_pLysS_AG__uid59245	NC_012947	4064.92	74	n/a
B_REL606_uid58803	NC_012967	4108.71	75	[27]
BL21_DE3__uid161947	NC_012971	4059.22	72	[27]
O157_H7_TW14359_uid59235	NC_013008	4210.32	86	[28]
O103_H2_12009_uid41013	NC_013353	4204.99	88	[29]
O26_H11_11368_uid41021	NC_013361	4277.7	92	[29]
O111_H__11128_uid41023	NC_013364	4247.44	85	[29]
SE15_uid161939	NC_013654	4163.36	67	[30]
O55_H7_CB9615_uid46655	NC_013941	4215.95	86	[31]
KO11FL_uid52593	NC_016902	4118.33	77	[32]
DH1_uid161951	NC_017625	4065.91	76	n/a
IHE3034_uid162007	NC_017628	4148.23	84	[33]
ABU_83972_uid161975	NC_017631	4182.54	76	[34]
UM146_uid162043	NC_017632	4125.2	78	[35]
ETEC_H10407_uid161993	NC_017633	4136.91	82	[36]
O83_H1_NRG_857C_uid161987	NC_017634	4104.77	72	[37]
W_uid162011	NC_017635	4067.85	76	n/a
DH1_uid162051	NC_017638	4061.41	77	[38]
UMNK88_uid161991	NC_017641	4192.17	88	[39]
NA114_uid162139	NC_017644	4111.79	82	[40]
O7_K1_CE10_uid162115	NC_017646	4235.53	87	[41]
clone_D_i2__uid162047	NC_017651	4161.67	76	[42]
clone_D_i14__uid162049	NC_017652	4161.67	76	[42]
O55_H7_RM12579_uid162153	NC_017656	4126.13	83	[43]
KO11FL_uid162099	NC_017660	4130.9	81	[32]
P12b_uid162061	NC_017663	4048.21	79	[44]
W_uid162101	NC_017664	4101.29	79	n/a
Xuzhou21_uid163995	NC_017906	4114.72	82	[44]
O104_H4_2011C_3493_uid176127	NC_018658	4130.08	77	[45]
O104_H4_2009EL_2071_uid176128	NC_018661	4149.56	79	[45]

1.3 Measuring the conservedness of supercoils across 52 genomes

Orthologous gene groups are identified across the 52 genomes using our in-house tool GOST [53], followed by a Markovian clustering algorithm MCL (with the parameter *I* set at 2) [54] to determine if a gene is present in more than one genome. It is worth noting that GOST is a combinatorial optimization tool with integration of sequence-similarity and contextual (working partners in operons) information, which can overcome intrinsic problems in orthology mapping across bacterial genomes, faced by sequence similarity-based methods, when orthology mapping involves gene fusions and horizontal gene transfers. At the end, 14203 orthologous gene groups were identified, which are used to calculate of the similarity score for a pair of supercoils in two genomes. Specifically, the similarity between two supercoils is defined as the percentage of the number of the shared orthologous gene groups out of the total number of unique genes in the two supercoils.

We have constructed a network consisting of all pairs of supercoils with similarity score above 0.6 across all 52 genomes. In this supercoil similarity network, an orthologous supercoil group (OSG) is defined as a maximal set of similar supercoils. The level of conservedness of an OSG is defined as the size of the OSG divided by 52.

2 Results and discussion

We have previously discovered that the global arrangement of operons in a bacterial (and also archaeal) genome follows some simple rules [55]. Specifically we have observed that any bacterial genome can be partitioned into a set of supercoils [9,10], each ranging between 10 and 100 kb, so that the total number of supercoils encoding genes in each biological pathway weighted by its activation frequency across all pathways is minimized, compared to the alternative arrangements of the operons [11]. The following C^+ function captures the essence of this observation (with slight variation). That is, operons are arranged into a collection of supercoils so that the following function is minimized:

$$C^+ = \sum_{i=1}^{N} f_i \sum_{j=1}^{R_i} (1 + \alpha(Q_{ij} - w_{ij})),$$

where N is the number of pathways encoded in the current genome, R_i represents the number of supercoils containing operons encoding the ith pathway, f_i is the activation frequency of the ith pathway, Q_{ij} and w_{ij} represent the number of all operons encoding the ith pathway and that in the jth supercoil, respectively, and α is a scaling factor currently set at 1. We have previously shown that the partition of a genome into supercoils that minimizes the above function can be efficiently calculated [12].

We have calculated the energy score C^+ for each of the 52 genomes along with the associated partition of the genome to a collection of supercoils using this model, with f_i being set to 1 for all pathways. In this calculation, 104 KEGG biological pathways (Supplementary File 1 for the names of the pathways), representing all largely complete *E. coli* pathways in the database, are considered in our energy score calculation and supercoil-domain partition. The number of operons encoding these pathways across the 52 genomes ranges from 793 to 868, with 822 as the mean. Table 1 shows the energy score, along with the number of the supercoils in each genome. The boundaries of these supercoils in each genome can be accessed in Supplementary File 2.

It is worth emphasizing that the C^+ function is designed to reflect the total energy required for unfolding (and refolding) all the relevant supercoils for the transcriptional activation of all the pathways encoded in each genome. We have compared the calculated energy scores with the growth rates of 10 *E. coli* strains, which have published growth-rate data collected under the same condition in the literature [51]. Interestingly, the growth rates of these 10 strains have a strong negative correlation with their respective energy scores with Pearson Correlation Coefficient (PCC)=–0.8 as shown in Figure 1, as we expected. While this example does not necessarily imply that our model can be used to predict the relative growth rates, it does suggest that the model captured something informative relevant to the energy efficiency of a bacterial cell.

2.1 Pathway C^+ values versus activation frequencies

We have further evaluated our model against the transcriptional activation frequencies of the four *E. coli* strains with some transcriptomic data available: K_12_W3110, O157_H7_Sakai, E24377A, and K_12_MG1655. For each organism, the transcription-activation frequency of a pathway is estimated based on the number of datasets having the pathway over-expressed divided by the total number of datasets under consideration for the organism, as done in [55] (see METHODS). 104 KEGG pathways are included in this analysis. Figure 2 shows the relationships between the C^+ values of individual pathways and their estimated activation frequencies across the whole genomes, where the 104 pathways are ranked in the increasing order of their C^+ values and evenly divided into seven groups, denoted as 1–7 (with 15 pathways in each of the first six groups and 14 pathways in group 7).

In addition, we have also compared the C^+ values of K_12_MG1655 across seven different sets of conditions, which should trigger the activation of very different sets of pathways (i.e., the activation frequencies of the pathways should be largely different across these seven sets of conditions). We grouped the 466 sets of transcriptomic data into seven subsets, each representing one set of conditions re-

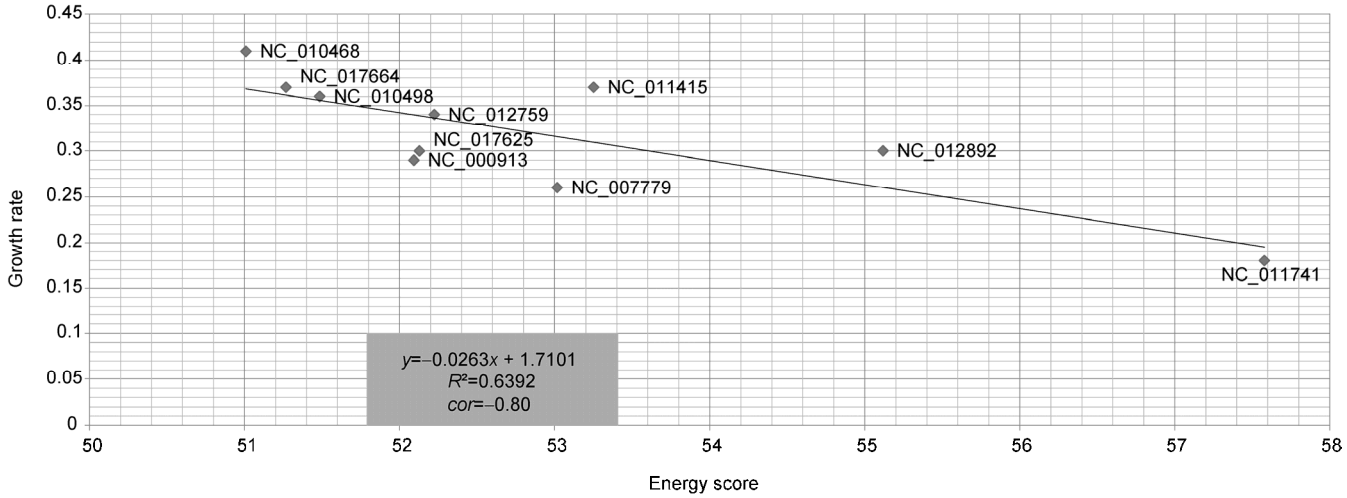

Figure 1 The correlation between the growth rates and the corresponding energy scores of 10 *E. coli* strains. The *x*-axis and *y*-axis represent energy scores and growth rates, respectively. Note: the energy scores in the *x*-axis is normalized by dividing the number of supercoils of each genome.

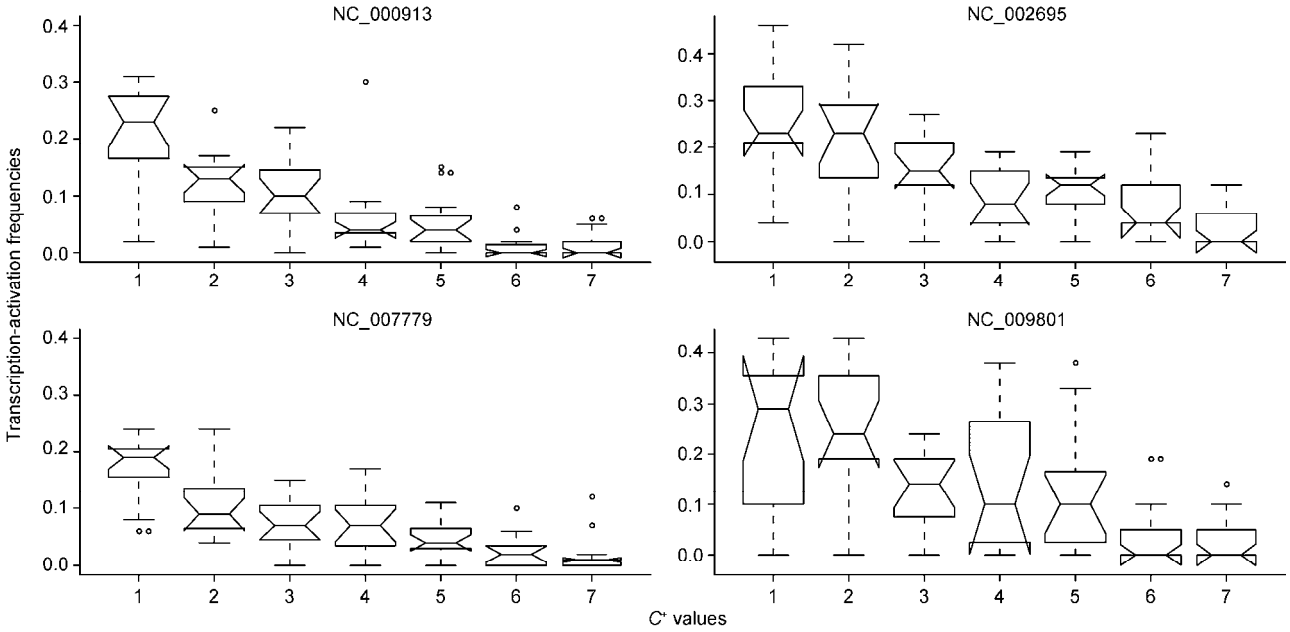

Figure 2 Relationship between the C^+ values (*x*-axis) and the transcription-activation frequencies (*y*-axis) of 104 pathways for four *E. coli* genomes. The organisms are represented by their chromosomal IDs.

lated to one of the following general growth conditions: anaerobic, heat shock, exponential growth, nitrogen limitation, oxidative, stationary growth, and SOS response. The detailed grouping scheme can be found in Supplementary File 3, which was published previously [11]. It is worth noting that the activation frequencies calculated here are over each subset rather than the whole set as done above. Figure 3 shows a strong negative correlation between the C^+ values and the estimated activation frequencies for all conditions except for the SOS response, where the negative correlation is not as strong as those under the other six growth conditions. We speculate that a possible reason for this is that the organization of the pathways responding to SOS is not fully optimized possibly due to that the SOS condition covers a wide range of different stresses, each of which may trigger different sets of responding pathways, and such conditions may not have occurred as often as the other conditions so the genomes have not been well trained under such conditions and hence the genomic organization of the relevant pathways is not optimized.

2.2 Organizational differences of pathways in genomes versus their activation frequencies

We noted that 2113 genes are shared (through orthology) by all the 52 genomes (Supplementary File 4 for details),

which encode 94 of the 104 pathways under consideration. Interestingly some of these 94 pathways have their encoding genes clustered into a similar set of supercoils across different genomes while others have their genes scattered across a wide range of different collections of supercoils. Figure 4 shows two examples with very distinct distributions of the number of supercoils encoding a pathway.

The question that we are interested in addressing is what may have caused the diversity in the genomic arrangements of individual pathways, or specifically in terms of the number of supercoils that encode each pathway across different genomes. To answer this question, we have calculated the standard deviation (SD) of the number of supercoils that encode each of the 94 pathways shared across all the 52 genomes. Figure 5A shows the distribution of the SD values across the 94 pathway, revealing the diversity in the genomic organization of pathways in different organisms.

We hypothesize that the diversity level in the number of supercoils encoding the same pathway across different genomes may be related to the variation in the activation frequencies of the pathway in different genomes. Since we do not have large numbers of transcriptomic data for the vast majority of the 52 organisms to estimate the activation frequency of each pathway, we have used the C^+ values and their relative ranking among all pathways' C^+ values to approximate the relative activation frequencies. While this

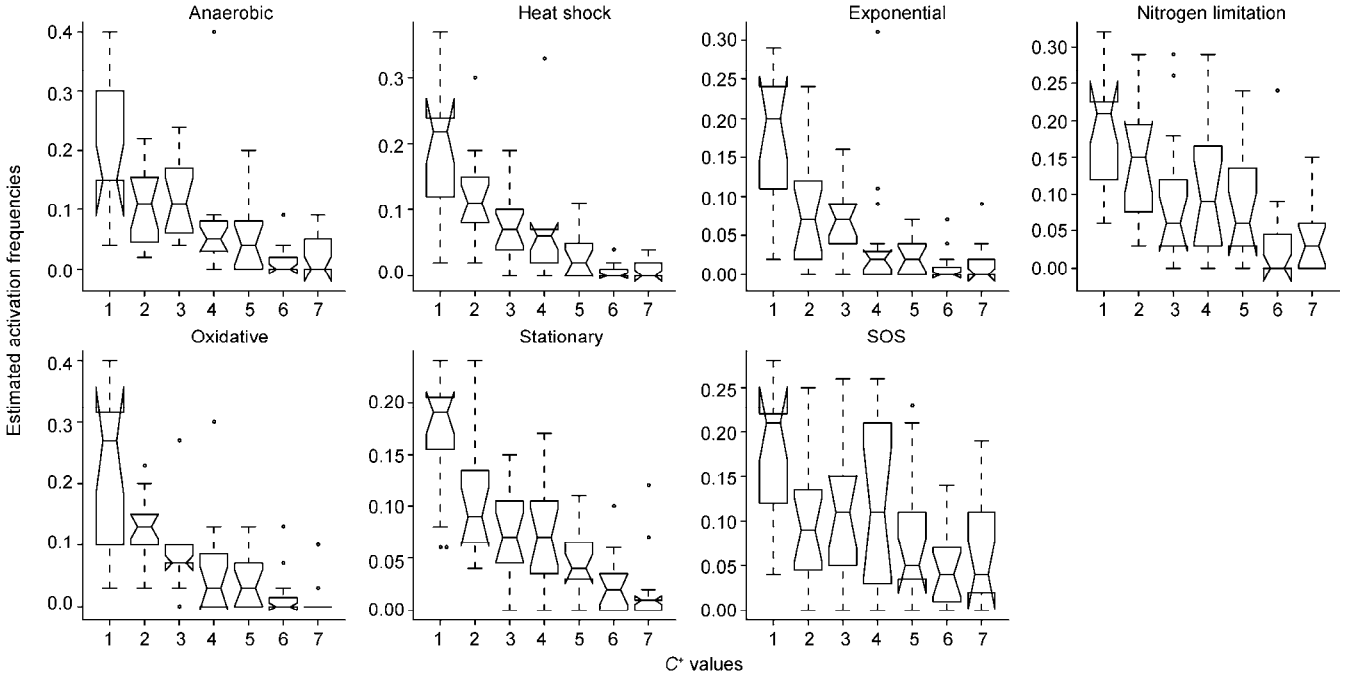

Figure 3 Correlations between the C^+ values (*x*-axis) and the estimated activation frequencies (*y*-axis) of *E. coli* K_12_MG1655 (NC_000913) under each of the seven sets of conditions. The seven panels are for seven conditions with the names given on top of each panel.

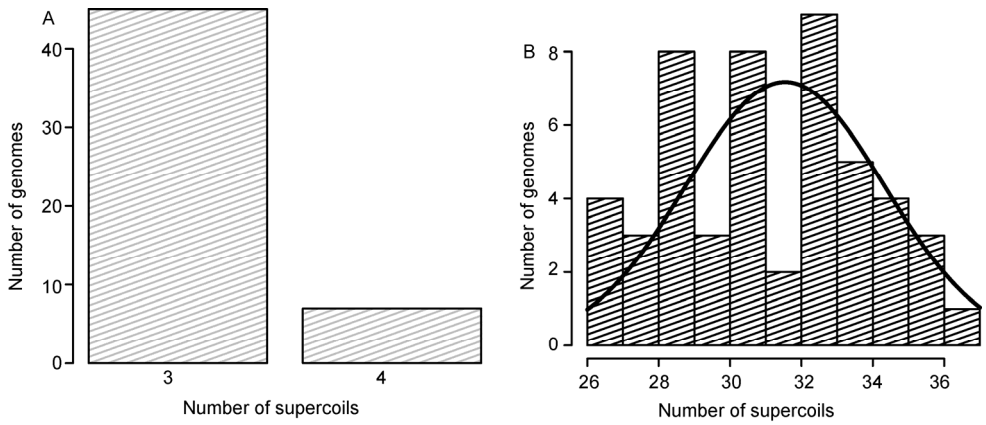

Figure 4 The distribution of the number of supercoils encoding a pathway across 52 genomes. For each panel, the *x*-axis represents the number of supercoils and the *y*-axis is the number of genomes using a specific number of supercoils to encode a pathway. A, The arachidonic acid metabolism pathway. B, The two-component system pathway.

may be a crude estimation, C^+ values of individual pathways in general do reflect the activation frequencies as shown in Figures 2 and 3, hence offering a justification for doing so. Specifically, the C^+ values of all the 94 pathways are sorted in the increasing order for each organism; and then the SD of the C^+ value ranking distribution for each pathway is calculated across the 52 genomes. We noted that there is a strong correlation between the SD of the numbers of encoding supercoils for each pathway (shown in Figure 5A) and the SD of the ranking distribution of the pathway's C^+ values calculated here (with Spearman correlation score 0.66). Figure 5B shows a box plot of the two values across all 94 pathways, hence providing an affirmative answer to our hypothesis.

2.3 Supercoil conservation versus functional diversity of genes encoded in the supercoil

We have also examined the genomic arrangement issue of the encoding genes of pathways from a supercoil-centric perspective. Specifically, we noted that some supercoils are used by all 52 genomes, i.e., supercoils containing the same sets of genes (through orthology) across all 52 genomes, referred to as conserved supercoils, while other supercoils are used only by a few genomes. We have developed a method for calculating the orthologous supercoil group (OSG), which is defined as the maximal group of highly similar supercoils across the 52 genomes. The size of each OSG is defined as the number of similar supercoils in the group, referred to as the level of conservedness of the supercoil. A total of 314 OSGs are determined using a simple procedure given in METHODS. Figure 6A shows the distribution of the conservedness of the 314 OSGs.

An analysis was carried out to examine the relationship between the conservedness of each OSG and the variation of the activation frequencies of pathways with encoding genes in the OSG across the 314 OSGs. Figure 6B shows a box plot of the two values across 314 OSGs, revealing a

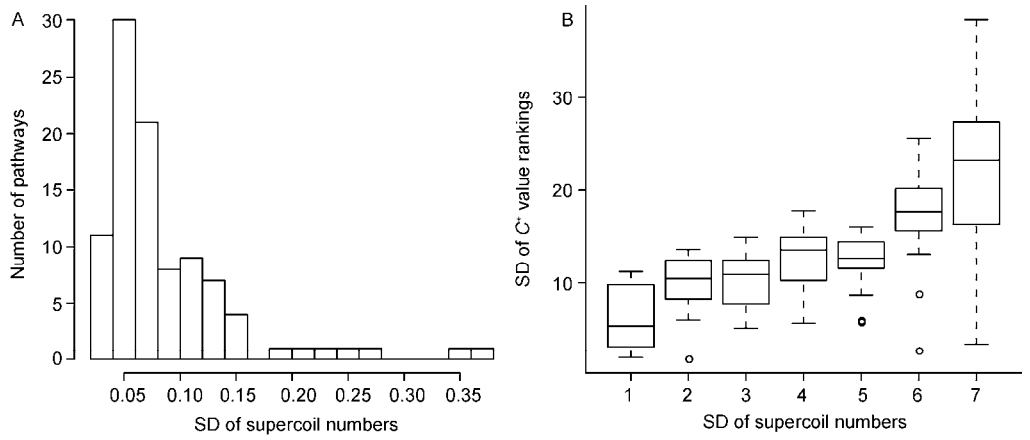

Figure 5 A relevant factor for the diversity in the genomic arrangements of individual pathways. A, The distribution of the standard deviation of the number of supercoils encoding each of the 94 pathways across the 52 genomes. B, Correlation between the SD of each pathway calculated in (A) and the SD of the rankings of the C^+ values of the pathway across the 52 genomes, where the 94 pathways are divided into seven groups in a similar way to that in Figure 2.

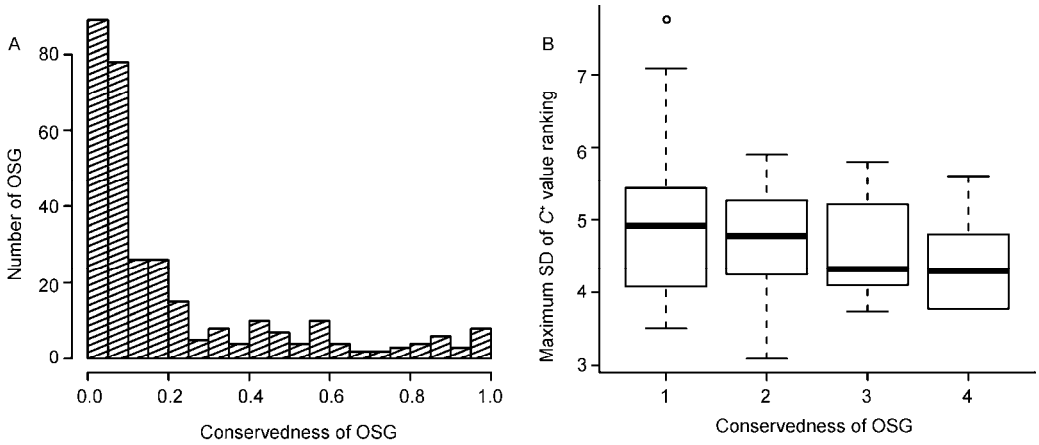

Figure 6 A relevant factor for the diversity in the conservedness of OSG. A, Distribution of the conservedness of 314 OSGs across the 52 genomes. B, Correlation between the conservedness in (A) and the maximum SD of corresponding C^+ values rankings which is calculated similarly to that in Figure 5B. The 314 OSGs are divided into four groups in a similar way to that in Figure 2.

new relationship concerning the conserved level of each supercoil.

3 Conclusion

The availability of the 52 closely related *E. coli* genomes enabled us to study the problem of the global arrangements of operons across different and related genomes from the perspective of bacterial genome evolution driven by energy efficiency. The data presented here provide further evidence to an observation that we have previously made: the global arrangement of operons in a bacterial genome is largely determined by the transcription-activation frequencies of individual pathways encoded by these operons in a way that the total DNA-unfolding (and refolding) energy needed to make the relevant pathways transcriptionally accessible tends to be minimized [11]. The analysis here on 52 closely related genomes allowed us to examine this observed organizational principle at a finer resolution by studying how genomes evolve to optimally adapt to their environments by adjusting their genomic organization to ensure that more frequently used pathways will involve a fewer number of relevant supercoils. Detailed information is revealed regarding how genes encoding a pathway change their genomic locations, specifically clustering into certain supercoils, to maintain the above energy-efficiency property, once the activation frequency of a pathway changes in response to the changing environment.

This work was supported in part by National Science Foundation (#NSF DEB-0830024 and NSF MCB-0958172) and by the US Department of Energy's BioEnergy Science Center grant through the Office of Biological and Environmental Research. The BioEnergy Science Center is a US Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Funding for open access charge: US Department of Energy's BioEnergy Science Center.

1 Breed RS, Conn HJ. The Status of the Generic Term Bacterium Ehrenberg 1828. J Bacteriol, 1936, 31: 517–518
2 Jacob F, Perrin D, Sanchez C, Monod J. Operon: a group of genes with the expression coordinated by an operator. C R Hebd Seances Acad Sci, 1960, 250: 1727–1729
3 Manson McGuire A, Church GM. Predicting regulons and their cis-regulatory motifs by comparative genomics. Nucleic Acids Res, 2000, 28: 4523–4530
4 Stock AM, Robinson VL, Goudreau PN. Two-component signal transduction. Ann Rev Biochem, 2000, 69: 183–215
5 Mao X, Ma Q, Zhou C, Chen X, Zhang H, Yang J, Mao F, Lai W, Xu Y. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. Nucleic Acids Res, 2014, 42: D654–659
6 Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernandez S, Alquicira-Hernandez K, Lopez-Fuentes A, Porron-Sotelo L, Huerta AM, Bonavides-Martinez C, Balderas-Martinez YI, Pannier L, Olvera M, Labastida A, Jimenez-Jacinto V, Vega-Alvarado L, Del Moral-Chavez V, Hernandez-Alvarez A, Morett E, Collado-Vides J. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res, 2013, 41: D203–213
7 Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res, 2014, 42: D199–205
8 Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev, 2005, 15: 589–594
9 Dillon SC, Dorman CJ. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nat Rev Microbiol, 2010, 8: 185–195
10 Benza VG, Bassetti B, Dorfman KD, Scolari VF, Bromek K, Cicuta P, Lagomarsino MC. Physical descriptions of the bacterial nucleoid at large scales, and their biological implications. Rep Prog Phys Phys Soc, 2012, 75: 076602
11 Ma Q, Yin Y, Schell MA, Zhang H, Li G, Xu Y. Computational analyses of transcriptomic data reveal the dynamic organization of the Escherichia coli chromosome under different conditions. Nucleic Acids Res, 2013, 41: 5594–5603
12 Ma Q, Xu Y. Global Genomic arrangement of bacterial genes is closely tied with the total transcriptional efficiency. Genom Proteom Bioinform, 2013, 11: 66–71
13 Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. The complete genome sequence of *Escherichia coli* K-12. Science, 1997, 277: 1453-1462
14 Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res, 2001, 8: 11–22
15 Welch RA, Burland V, Plunkett G, 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci USA, 2002, 99: 17020–17024
16 Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, Kosuge T, Mori H, Perna NT, Plunkett G, 3rd, Rudd KE, Serres MH, Thomas GH, Thomson NR, Wishart D, Wanner BL. *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. Nucleic Acids Res, 2006, 34: 1–9
17 Chen SL, Hung CS, Xu J, Reigstad CS, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer RR, Ozersky P, Armstrong JR, Fulton RS, Latreille JP, Spieth J, Hooton TM, Mardis ER, Hultgren SJ, Gordon JI. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. Proc Natl Acad Sci USA, 2006, 103: 5977–5982
18 Johnson TJ, Wannemeuhler YM, Scaccianoce JA, Johnson SJ, Nolan LK. Complete DNA sequence, comparative genomics, and prevalence of an IncHI2 plasmid occurring among extraintestinal pathogenic *Escherichia coli* isolates. Antimicrob Agents Chemother, 2006, 50: 3929–3933
19 Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J Bacteriol, 2008, 190: 6881–6893
20 Durfee T, Nelson R, Baldwin S, Plunkett G, 3rd, Burland V, Mau B, Petrosino JF, Qin X, Muzny DM, Ayele M, Gibbs RA, Csorgo B, Posfai G, Weinstock GM, Blattner FR. The complete genome sequence of *Escherichia coli* DH10B: insights into the biology of a laboratory workhorse. J Bacteriol, 2008, 190: 2597–2606
21 Fricke WF, Wright MS, Lindell AH, Harkins DM, Baker-Austin C, Ravel J, Stepanauskas R. Insights into the environmental resistance

gene pool from the genome sequence of the multidrug-resistant environmental isolate *Escherichia coli* SMS-3-5. J Bacteriol, 2008, 190: 6779–6794

22 Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. Genomic anatomy of *Escherichia coli* O157:H7 outbreaks. Proc Natl Acad Sci USA, 2011, 108: 20142–20147

23 Oshima K, Toh H, Ogura Y, Sasamoto H, Morita H, Park SH, Ooka T, Iyoda S, Taylor TD, Hayashi T, Itoh K, Hattori M. Complete genome sequence and comparative analysis of the wild-type commensal *Escherichia coli* strain SE11 isolated from a healthy adult. DNA Res, 2008, 15: 375–386

24 Iguchi A, Thomson NR, Ogura Y, Saunders D, Ooka T, Henderson IR, Harris D, Asadulghani M, Kurokawa K, Dean P, Kenny B, Quail MA, Thurston S, Dougan G, Hayashi T, Parkhill J, Frankel G. Complete genome sequence and comparative genome analysis of enteropathogenic *Escherichia coli* O127:H6 strain E2348/69. J Bacteriol, 2009, 191: 347–354

25 Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguenec C, Lescat M, Mangenot S, Martinez-Jehanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Medigue C, Rocha EP, Denamur E. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet, 2009, 5: e1000344

26 Ferenci T, Zhou Z, Betteridge T, Ren Y, Liu Y, Feng L, Reeves PR, Wang L. Genomic sequencing reveals regulatory mutations and recombinational events in the widely used MC4100 lineage of *Escherichia coli* K-12. J Bacteriol, 2009, 191: 4025–4029

27 Jeong H, Barbe V, Lee CH, Vallenet D, Yu DS, Choi SH, Couloux A, Lee SW, Yoon SH, Cattolico L, Hur CG, Park HS, Segurens B, Kim SC, Oh TK, Lenski RE, Studier FW, Daegelen P, Kim JF. Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). J Mol Biol, 2009, 394: 644–652

28 Kulasekara BR, Jacobs M, Zhou Y, Wu Z, Sims E, Saenphimmachak C, Rohmer L, Ritchie JM, Radey M, McKevitt M, Freeman TL, Hayden H, Haugen E, Gillett W, Fong C, Chang J, Beskhlebnaya V, Waldor MK, Samadpour M, Whittam TS, Kaul R, Brittnacher M, Miller SI. Analysis of the genome of the Escherichia coli O157:H7 2006 spinach-associated outbreak isolate indicates candidate genes that may enhance virulence. Infect Immun, 2009, 77: 3713–3721

29 Ogura Y, Ooka T, Iguchi A, Toh H, Asadulghani M, Oshima K, Kodama T, Abe H, Nakayama K, Kurokawa K, Tobe T, Hattori M, Hayashi T. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. Proc Natl Acad Sci USA, 2009, 106: 17939–17944

30 Toh H, Oshima K, Toyoda A, Ogura Y, Ooka T, Sasamoto H, Park SH, Iyoda S, Kurokawa K, Morita H, Itoh K, Taylor TD, Hayashi T, Hattori M. Complete genome sequence of the wild-type commensal *Escherichia coli* strain SE15, belonging to phylogenetic group B2. J Bacteriol, 2010, 192: 1165–1166

31 Zhou Z, Li X, Liu B, Beutin L, Xu J, Ren Y, Feng L, Lan R, Reeves PR, Wang L. Derivation of *Escherichia coli* O157:H7 from its O55:H7 precursor. PLoS One, 2010, 5: e8700

32 Turner PC, Yomano LP, Jarboe LR, York SW, Baggett CL, Moritz BE, Zentz EB, Shanmugam KT, Ingram LO. Optical mapping and sequencing of the Escherichia coli KO11 genome reveal extensive chromosomal rearrangements, and multiple tandem copies of the *Zymomonas mobilis* pdc and adhB genes. J Ind Microbiol Biotechnol, 2012, 39: 629–639

33 Moriel DG, Bertoldi I, Spagnuolo A, Marchi S, Rosini R, Nesta B, Pastorello I, Corea VA, Torricelli G, Cartocci E, Savino S, Scarselli M, Dobrindt U, Hacker J, Tettelin H, Tallon LJ, Sullivan S, Wieler LH, Ewers C, Pickard D, Dougan G, Fontana MR, Rappuoli R, Pizza M, Serino L. Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic *Escherichia coli*. Proc Natl Acad Sci USA, 2010, 107: 9072–9077

34 Zdziarski J, Brzuszkiewicz E, Wullt B, Liesegang H, Biran D, Voigt B, Gronberg-Hernandez J, Ragnarsdottir B, Hecker M, Ron EZ, Daniel R, Gottschalk G, Hacker J, Svanborg C, Dobrindt U. Host imprints on bacterial genomes—rapid, divergent evolution in individual patients. PLoS Pathog, 2010, 6: e1001078

35 Krause DO, Little AC, Dowd SE, Bernstein CN. Complete genome sequence of adherent invasive *Escherichia coli* UM146 isolated from Ileal Crohn's disease biopsy tissue. J Bacteriol, 2011, 193: 583

36 Crossman LC, Chaudhuri RR, Beatson SA, Wells TJ, Desvaux M, Cunningham AF, Petty NK, Mahon V, Brinkley C, Hobman JL, Savarino SJ, Turner SM, Pallen MJ, Penn CW, Parkhill J, Turner AK, Johnson TJ, Thomson NR, Smith SG, Henderson IR. A commensal gone bad: complete genome sequence of the prototypical enterotoxigenic *Escherichia coli* strain H10407. J Bacteriol, 2010, 192: 5822–5831

37 Nash JH, Villegas A, Kropinski AM, Aguilar-Valenzuela R, Konczy P, Mascarenhas M, Ziebell K, Torres AG, Karmali MA, Coombes BK. Genome sequence of adherent-invasive *Escherichia coli* and comparative genomic analysis with other *E. coli* pathotypes. BMC Genomics, 2010, 11: 667

38 Suzuki S, Ono N, Furusawa C, Ying BW, Yomo T. Comparison of sequence reads obtained from three next-generation sequencing platforms. PLoS One, 2011, 6: e19534

39 Fernandez-Alarcon C, Singer RS, Johnson TJ. Comparative genomics of multidrug resistance-encoding IncA/C plasmids from commensal and pathogenic *Escherichia coli* from multiple animal sources. PLoS One, 2011, 6: e23415

40 Avasthi TS, Kumar N, Baddam R, Hussain A, Nandanwar N, Jadhav S, Ahmed N. Genome of multidrug-resistant uropathogenic Escherichia coli strain NA114 from India. J Bacteriol, 2011, 193: 4272–4273

41 Lu S, Zhang X, Zhu Y, Kim KS, Yang J, Jin Q. Complete genome sequence of the neonatal-meningitis-associated *Escherichia coli* strain CE10. J Bacteriol, 2011, 193: 7005

42 Reeves PR, Liu B, Zhou Z, Li D, Guo D, Ren Y, Clabots C, Lan R, Johnson JR, Wang L. Rates of mutation and host transmission for an *Escherichia coli* clone over 3 years. PLoS One, 2011, 6: e26907

43 Kyle JL, Cummings CA, Parker CT, Quinones B, Vatta P, Newton E, Huynh S, Swimley M, Degoricija L, Barker M, Fontanoz S, Nguyen K, Patel R, Fang R, Tebbs R, Petrauskene O, Furtado M, Mandrell RE. *Escherichia coli* serotype O55:H7 diversity supports parallel acquisition of bacteriophage at Shiga toxin phage insertion sites during evolution of the O157:H7 lineage. J Bacteriol, 2012, 194: 1885-1896

44 Liu B, Hu B, Zhou Z, Guo D, Guo X, Ding P, Feng L, Wang L. A novel non-homologous recombination-mediated mechanism for *Escherichia coli* unilateral flagellar phase variation. Nucleic Acids Res, 2012, 40: 4530–4538

45 Ahmed SA, Awosika J, Baldwin C, Bishop-Lilly KA, Biswas B, Broomall S, Chain PS, Chertkov O, Chokoshvili O, Coyne S, Davenport K, Detter JC, Dorman W, Erkkila TH, Folster JP, Frey KG, George M, Gleasner C, Henry M, Hill KK, Hubbard K, Insalaco J, Johnson S, Kitzmiller A, Krepps M, Lo CC, Luu T, McNew LA, Minogue T, Munk CA, Osborne B, Patel M, Reitenga KG, Rosenzweig CN, Shea A, Shen X, Strockbine N, Tarr C, Teshima H, van Gieson E, Verratti K, Wolcott M, Xie G, Sozhamannan S, Gibbons HS, Threat Characterization C. Genomic comparison of *Escherichia coli* O104:H4 isolates from 2009 and 2011 reveals plasmid, and prophage heterogeneity, including shiga toxin encoding phage stx2. PLoS One, 2012, 7: e48228

46 Barrett T, Edgar R. Reannotation of array probes at NCBI's GEO database. Nat Methods, 2008, 5: 117

47 Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS. Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. Nucleic Acids Res, 2008, 36: D866–870

48 Lemuth K, Hardiman T, Winter S, Pfeiffer D, Keller MA, Lange S, Reuss M, Schmid RD, Siemann-Herzberg M. Global transcription and metabolic flux analysis of *Escherichia coli* in glucose-limited fed-batch cultivations. Appl Environ Microbiol, 2008, 74: 7002–7015

49 Abe H, Miyahara A, Oshima T, Tashiro K, Ogura Y, Kuhara S, Ogasawara N, Hayashi T, Tobe T. Global regulation by horizontally transferred regulators establishes the pathogenicity of *Escherichia coli*. DNA Res, 2008, 15: 25–38
50 Sahl JW, Rasko DA. Analysis of global transcriptional profiles of enterotoxigenic *Escherichia coli* isolate E24377A. Infect Immun, 2012, 80: 1232–1242
51 Eashwar Rajaraman MAE. Transcriptional analysis and adaptive evolution of *Escherichia coli* growing on acetate. Dissertation for Doctoral Degree. Athens: University of Georgia, 2012
52 Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol, 2004, 5: R80
53 Li G, Ma Q, Mao X, Yin Y, Zhu X, Xu Y. Integration of sequence-similarity and functional association information can overcome intrinsic problems in orthology mapping across bacterial genomes. Nucleic Acids Res, 2011, 39: e150
54 Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res, 2002, 30: 1575–1584
55 Yin Y, Zhang H, Olman V, Xu Y. Genomic arrangement of bacterial operons is constrained by biological pathways encoded in the genome. Proc Natl Acad Sci USA, 2010, 107: 6310–6315

Supporting Information

Supplementary Files 1–4

The supporting information is available online at life.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.