Gene 663 (2018) 178–188

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Essentiality, conservation, evolutionary pressure and codon bias in bacterial genomes

Maddalena Dilucca[a,*], Giulio Cimini[b,c], Andrea Giansanti[a,d]

[a] Dipartimento di Fisica, "Sapienza" University of Rome, Rome 00185, Italy
[b] IMT School for Advanced Studies, Lucca 55100, Italy
[c] Istituto dei Sistemi Complessi (ISC)-CNR, Rome 00185, Italy
[d] INFN Roma1 Unit, Rome 00185, Italy

ARTICLE INFO

Keywords:
Bacteria
Essentiality
Conservation
Selective pressure
Codon bias

ABSTRACT

Essential genes constitute the core of genes which cannot be mutated too much nor lost along the evolutionary history of a species. Natural selection is expected to be stricter on essential genes and on conserved (highly shared) genes, than on genes that are either nonessential or peculiar to a single or a few species. In order to further assess this expectation, we study here how essentiality of a gene is connected with its degree of conservation among several unrelated bacterial species, each one characterised by its own codon usage bias. Confirming previous results on *E. coli*, we show the existence of a universal exponential relation between gene essentiality and conservation in bacteria. Moreover, we show that, within each bacterial genome, there are at least two groups of functionally distinct genes, characterised by different levels of conservation and codon bias: i) a core of essential genes, mainly related to cellular information processing; ii) a set of less conserved non-essential genes with prevalent functions related to metabolism. In particular, the genes in the first group are more retained among species, are subject to a stronger purifying conservative selection and display a more limited repertoire of synonymous codons. The core of essential genes is close to the minimal bacterial genome, which is in the focus of recent studies in synthetic biology, though we confirm that orthologs of genes that are essential in one species are not necessarily essential in other species. We also list a set of highly shared genes which, reasonably, could constitute a reservoir of targets for new anti-microbial drugs.

1. Introduction

From an evolutionary point of view, all living species are in a process of adaptation to the environments they happen to live in. This process rests on the incorporation of genetic mutations into the genomes at the level of populations of species, which evolves on time-scales far longer than the time-scale of a generation. Signals from this process can be searched for in the sequences of single genes, of several genes within one single species, and among several species. In a previous work we have shown that, in *E. coli*, essentiality and degree of conservation of genes are subtly correlated with the codon bias displayed by their sequences (Dilucca et al., 2015). In this work we extend those observations to a set of unrelated bacterial species, by elaborating on the connection between gene essentiality and conservation, and their relation with codon bias.

Individual genes in the genome of a given species contribute differentially to the survival and propagation of the organisms of that species. According to their known functional profiles and based on experimental evidences, genes can be divided into two categories: essential and nonessential ones (Fang et al., 2005; Gerdes et al., 2003). Essential genes are not dispensable for the survival of an organism in the environment it lives in and the functions they encode are, therefore, considered as fundamental for life, irrespective of environmental changes (Fang et al., 2005; Peng and Gao, 2014). On the other hand, nonessential genes are those which are dispensable (Lin et al., 2010), being related to functions that can be silenced without lethal effects for the phenotype. Naturally, each species has adapted to one or more evolving environments and, plausibly, genes that are essential for one species may be not essential for another one. However, the set of genes that are essential in several bacterial species should encode for functions that are fundamental for life. As suggested by a quite broad literature, essential genes are more conserved than nonessential ones

Abbreviations: COG, clusters of orthologous genes; DAMBE, data analysis in molecular biology and evolution; DEG, database of essential genes; E. coli, escherichia coli; ERI, evolutionary retention index; Nc, number of effective codons; RSCU, relative synonymous codon usage

* Corresponding author.
E-mail address: maddalena.dilucca@roma1.infn.it (M. Dilucca).

https://doi.org/10.1016/j.gene.2018.04.017
Received 31 October 2017; Received in revised form 25 March 2018; Accepted 9 April 2018
Available online 18 April 2018
0378-1119/

Table 1
Functional specialization of essential and nonessential genes according to COG clusters. Figures indicate the percentages of essential and nonessential genes within a given COG (sums of these figures for table subsections are reported in boldface). COGs are sorted by percent essentiality.

COG ID	Functional classification	% E	% NE
	Information storage and processing		
J	Translation, ribosomal structure and biogenesis	0.25	0.05
K	Transcription	0.06	0.10
L	Replication, recombination and repair	0.08	0.07
		0.39	**0.22**
	Cellular processes and signaling		
D	Cell cycle control, cell division, chromosome partitioning	0.03	0.01
T	Signal transduction mechanisms	0.02	0.07
M	Cell wall/membrane/envelope biogenesis	0.10	0.08
N	Cell motility	0.01	0.03
O	Posttranslational modification, protein turnover, chaperones	0.04	0.05
		0.20	**0.24**
	Metabolism		
C	Energy production and conversion	0.07	0.08
G	Carbohydrate transport and metabolism	0.06	0.10
E	Amino acid transport and metabolism	0.06	0.12
F	Nucleotide transport and metabolism	0.05	0.03
H	Coenzyme transport and metabolism	0.08	0.06
I	Lipid transport and metabolism	0.07	0.05
P	Inorganic ion transport and metabolism	0.03	0.08
		0.41	**0.52**

(Alvarez-Ponce et al., 2016; Hurst and Smith, 1999; Ish-Am et al., 2015; Jordan et al., 2002; Luo et al., 2015). It is worth noting that he term " conservation" has at least a twofold meaning. On the one hand, a gene is conserved if orthologous copies are found in the genomes of many species, as measured by the Evolutionary Retention Index (ERI) (Bergmiller et al., 2012; Gerdes et al., 2003). On the other hand, a gene is (evolutionarily) conserved when it is subject to a purifying evolutionary pressure which disfavors mutations (Hurst, 2002; Hurst and Smith, 1999). In this second meaning a conserved gene is, generically, a slowly evolving gene. The ratio K_a/K_s of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site is a widely used measure, though not exempt from criticism, to assess whether a gene is under Darwinian selection or a purifying evolutionary pressure.

Beyond essentiality and conservation, in our analysis we also consider the degeneracy of the genetic code, due to the fact that the same amino acid is encoded by different codon triplets (synonymous codons). Usage frequencies of synonymous codons vary significantly between different organisms, and also between proteins within the same organism (Kanaya et al., 2001). This phenomenon, known as *codon usage bias*, can be measured by various indices (see Roth et al., 2012 for an overview); we use here statistical indicators such as the effective number of codons and the relative synonymous codon usage.

Our analysis reveals that those genes which are more conserved among bacterial species are also prone to be essential. Moreover, the codon usage in these conserved genes is, in general, more optimized than in less conserved genes. We have also shown that essential, conserved genes tend to be subject to a relatively more purifying evolutionary pressure. We argue that the set of genes with the highest degree of conservation (ERI = 1, see Table 4) could include putative novel targets for novel anti bacterial strategies, as suggested with rather similar arguments by Dötsch et al. (2010).

2. Materials and methods

2.1. Bacterial genomes

In this work we consider a set of 45 bacterial genomes from unrelated species, whose details are provided in Table 3. Nucleotide sequences from complete bacterial genomes were downloaded from the FTP server of the National Center for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_genbank/Bacteria/) (Benson et al., 2013). Note that 31 of the 45 species we collected are also present in the dataset selected by Gerdes et al. (2003) in their seminal paper on *E. coli*'s essential genes.

2.2. Conservation and essentiality

We use the Evolutionary Retention Index (ERI) of Gerdes et al. (2003) as a proxy for gene conservation. We compute the ERI of a gene as the fraction of genomes in Table 3 that have at least an ortholog of the given gene. A low ERI value means that a gene is specific, common to a small number of genomes, whereas, high ERI is a characteristic of highly shared, conserved, possibly universal genes.

In order to investigate gene essentiality we use the Database of Essential Genes (DEG), available at http://www.essentialgene.org (Luo et al., 2015). DEG classifies a gene as either essential or nonessential on the basis of a combination of experimental evidence (null mutations or transposons) and general functional considerations. DEG collects genomes from Bacteria, Archea and Eukarya, with different degrees of coverage (Luo et al., 2014; Zhang and Lin, 2009). Of the 45 bacterial genomes we have collected, only 24 are covered—in toto or partially—by DEG, as indicated in Table 3.

2.3. Clusters of orthologous genes

We use the database of orthologous groups of proteins (COGs), available at http://ncbi.nlm.nih.gov/COG/, for the functional annotation of gene sequences (Galperin et al., 2015). We consider 15 functional classes given by COGs, excluding the generic categories R and S for which functional annotation is too general or missing. Given a group of genes in a genome, we evaluate the conditional probability that these genes belong to a specific COG as:

$$P(\text{COG}|\text{group}) = P(\text{group}|\text{COG})P(\text{COG})/P(\text{group}), \quad (1)$$

where P(group) is the size of the group with respect to the genome, P(COG) is the fraction of the genome belonging to the COG, and P(group|COG) is the fraction of genes in a given COG that belong to the group.

2.4. K_a/K_s

K_a/K_s is the ratio of nonsynonymous substitutions per nonsynonymous site (K_a) to the number of synonymous substitutions per synonymous site (K_s) (Hurst, 2002). This parameter is widely accepted as

Table 2
Ranking of COGs according to Z-scores of K_a/K_s. We count ± 1 for each time the Z-score is smaller than −1 or bigger than +1, respectively.

COG ID	J	F	K	O	E	I	D	C	T	H	G	P	L	N	M
Z-score	−20	−19	−14	−13	−13	−13	−12	−12	−9	−9	−6	−5	−5	4	5

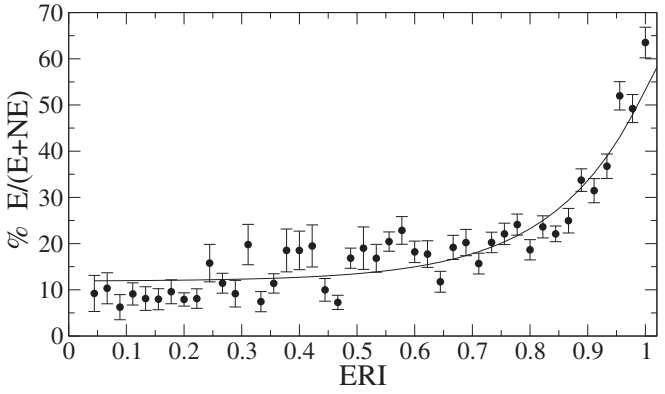

Fig. 1. Degree of essentiality versus ERI values. Genes from the 24 DEG-annotated genomes in Table 3 are aggregated into bins of ERI, and the fraction of essential genes in each bin is computed. Error bars are root mean square deviations, expressing the variability in percent essentiality from genome to genome. The solid line shows the exponential fit of the data $y = y_0 + A\exp(Bx)$, which returns $y_0 = 11.(8)$, $A = 0.06(3)$, $B = 6.4(8)$ and a coefficient of determination $R^2 = 0.90$. Note that since in small genomes essential genes outnumber nonessential ones (as we show below), we exclude from fitted data the two small genomes of *myge* and *mypu*. We also exclude *stpn* because it is poorly covered by DEG.

a straightforward and effective way of separating genes subject to purifying selection ($K_a/K_s < 1$) from genes subject to positive Darwinian selection ($K_a/K_s > 1$). There are different methods to evaluate this ratio, though the alternative approaches are quite consistent among themselves. For the sake of comparison, we have used here the K_a/K_s estimates by Luo et al. (2015) which are based on method by Nei and Gojobori (1986). Note that each genome has a specific average level of K_a/K_s.

To study the patterns of K_a/K_s in the various COGs, we use Z-score values:

$$Z_g[(K_a/K_s)|COG] = \frac{\langle K_a/K_s\rangle_{COG,g} - \langle K_a/K_s\rangle_g}{\sigma_g/\sqrt{N_g}}, \tag{2}$$

where $\langle K_a/K_s\rangle_{COG,g}$ is the average of the ratio within a given COG in a genome g, $\langle K_a/K_s\rangle_g$ and σ_g are the average value of K_a/K_s and its standard deviation over the whole genome g, and N_g is the number of genes in the genome (we use the standard deviation of the mean as we are comparing average values).

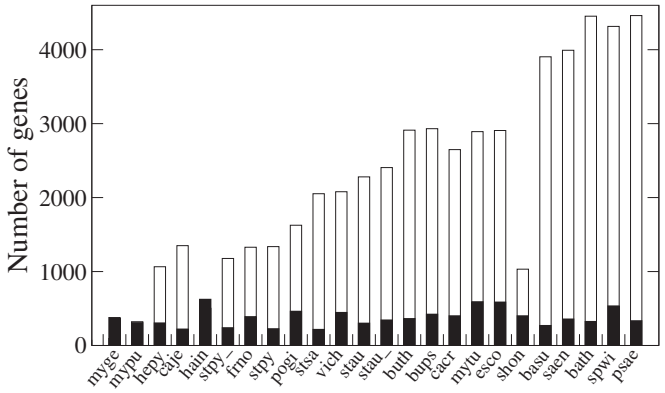

Fig. 2. Essential and nonessential genes in each bacteria. On the horizontal axis, DEG-annotated species from Table 3 are sorted according to the size of their genome. Black and white bars represent the number of essential and nonessential genes. The number of essential genes is basically constant in all species (average value = 378 ± 115), while the number of nonessential genes increases with the size of the genomes.

To study the patterns of K_a/K_s in a given group of genes (e.g. percentiles of ERI as in Fig. 5), we also use Z-score values:

$$Z_g[(K_a/K_s)|\text{group}] = \frac{\langle K_a/K_s\rangle_{\text{group},g} - \langle K_a/K_s\rangle_g}{\sigma_g/\sqrt{N_g}}, \tag{3}$$

where $\langle K_a/K_s\rangle_{\text{group},g}$ is the average of the ratio within a given group in a genome g, $\langle K_a/K_s\rangle_g$ and σ_g are the average value of K_a/K_s and its standard deviation over the whole genome g, and N_g is the number of genes in the genome (we use the standard deviation of the mean as we are comparing average values). In the following we will analyze K_a/K_s patterns with respect to group of genes with similar values of ERI, as indicated by $Z_g[(K_a/K_s)|ERI]$ values (eventually discerning essential and nonessential genes), as well as with respect to COGs, as indicated by $Z_g[(K_a/K_s)|COG]$ values.

Note that K_a/K_s values can differ much by magnitude not only among genes in different genomes but even for genes in the same genome, and their genome-specific distribution is rather broad with a high peak at zero (see below). In such a situation, larger values can in principle bias arithmetic averages. On the other hand, using alternative methods like the geometric mean, the arithmetic mean of the logarithms or the harmonic mean cannot be used because of the frequent zero values. Using the median and the median absolute deviation in a Mann-Whitney-Wilcoxon U test instead of arithmetic means and standard deviations in Z-score tests to compare distributions leads to results—reported in the Supporting Information—which are highly consistent with what we obtained with Z-scores, though more noisy due to the vanishing of the median in many distributions.

2.5. Codon bias

There are several methods and indices to estimate the degree of codon usage bias in a gene. For an overview of current methods, their classification and rationale see Roth et al. (2012). We use here two basic statistical indicators: the Number of Effective Codons (N_c) and the Relative Synonymous Codon Usage (*RSCU*).

N_c measures of the effective diversity of the codons used to code a given protein (Wright, 1990). In principle, N_c ranges from 20 (when just one single codon is used to code each one of the amino acids) to 61 (when the entire degeneracy of the genetic code is fully deployed, and each amino acid is coded by all its synonymous codons on an equal footing). Given a sequence of interest, the computation of N_c starts from F_α, a quantity defined for each family α of synonymous codons (one for each amino acid):

$$F_\alpha = \sum_{k=1}^{m_\alpha} \left(\frac{n_{k\alpha}}{n_\alpha}\right)^2, \tag{4}$$

where m_α is the number of different codons in α (each one appearing $n_{1_\alpha}, n_{2_\alpha}, ..., n_{m_\alpha}$ times in the sequence) and $n_\alpha = \sum_{k=1}^{m_\alpha} n_{k_\alpha}$. N_c then weights these quantities on a sequence:

$$N_c = N_s + \frac{K_2\sum_{\alpha=1}^{K_2} n_\alpha}{\sum_{\alpha=1}^{K_2}(n_\alpha F_\alpha)} + \frac{K_3\sum_{\alpha=1}^{K_3} n_\alpha}{\sum_{\alpha=1}^{K_3}(n_\alpha F_\alpha)} + \frac{K_4\sum_{\alpha=1}^{K_4} n_\alpha}{\sum_{\alpha=1}^{K_4}(n_\alpha F_\alpha)} \tag{5}$$

where N_s is the number of families with one codon only and K_m is the number of families with degeneracy m (the set of 6 synonymous codons for leucine can be split into one family with degeneracy 2, similar to that of phenylalanine, and one family with degeneracy 4, similar to that, of proline). In this paper we evaluate N_c by using the implementation provided in DAMBE 5.0 (Xia, 2013).

The relative synonymous codon usage ($RSCU_i$) of each codon i is estimated as:

$$RSCU_i = \frac{X_i}{\frac{1}{N_i}\sum_{j=1}^{n_i} X_j} \tag{6}$$

where X_i is the number of occurrences, either in a gene or in the whole

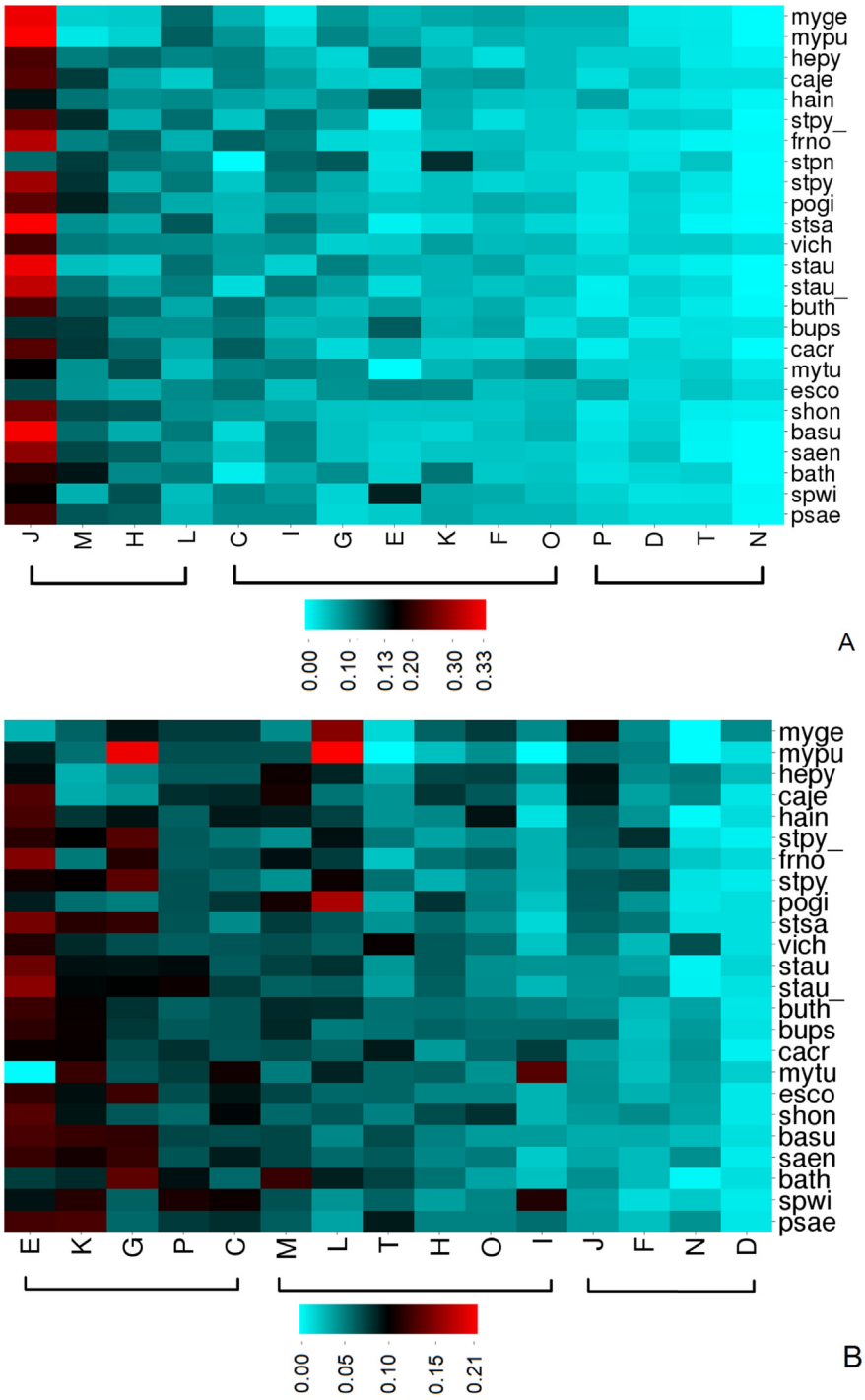

Fig. 3. For each genome we estimated the conditional probabilities $P(COG|E)$ (panel A) and $P(COG|NE)$ (panel B) for an essential and a nonessential gene to belong to a given COG. Genomes are ranked from top to bottom by the size of their genomes. COGs are ranked, separately in both panels and from left to right, according to their overall incidence. In panel A, 51% of the essential genes belong, in different proportions, to COGs J, M, H and L; 40% to C, I, G, E, K, F and O and the remaining 10% to P, D, T and N. In panel B, 49% of the nonessential genes belong to E, K, G, P, and C; 38% to M, L,T, H, O and I; the remaining 13% to J, F, N, and D.

genome, of codon *i*. The sum in the denominator runs over n_i, the degeneracy of the family of synonymous codons *i* belongs to. For each codon *i*, its $RSCU_i$ is comprised between zero (no usage) and 1 (when only that codon is used among its synonymous alternatives). We evaluate these values by using DAMBE 5.0 (Xia, 2013).

The *RSCU* values of the various codons can be grouped together as the 64 components (including the start codon ATG and the stop codons TAA, TAG and TGA—which are differently used by different species) of vectors which measure codon usage bias in a given bacterial species.

To detect different patterns of codon usage between species we use heat maps drawn with CIMMiner (http://discover.nci.nih.gov/cimminer), and we cluster *RSCU* vectors using Euclidean distances and the Average Linkage cluster algorithm.

3. Results and discussion

3.1. Essentiality and conservation in bacterial genes

Fig. 1 shows the percentage of essential genes within genes with a given value of ERI (which we recall operationally encodes the degree of conservation of a gene). The observed exponential dependence generalises to several unrelated species a basic result on *E. coli*, by Gerdes et al. (2003) (see Fig. 3 therein), and the fit parameters we find are strictly consistent with those reported in that paper. This points to the existence of a universal exponential correlation between gene essentiality and conservation in bacteria. Indeed, the fact that essential genes should be more evolutionarily conserved than nonessential ones has

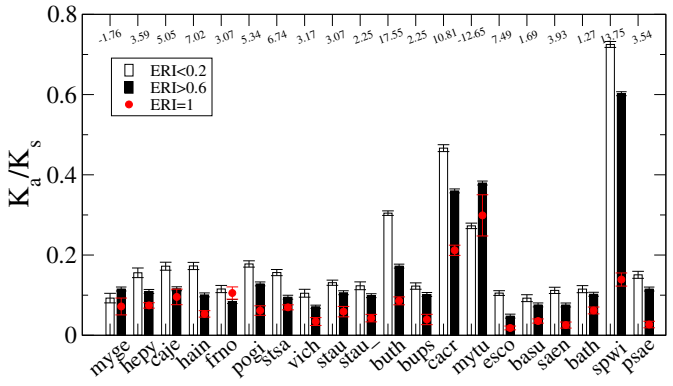

Fig. 4. Average values and standard deviations of K_a/K_s for specific and conserved genes in each genome. Specific genes have ERI < 0.2 and conserved genes have ERI > 0.6. According to a two-sample z statistic, whose values are shown at the top of the panel, with the exception of *myge* and *mytu*, the average value of K_a/K_s is significantly higher for specific genes. Red points denote averages and errors of K_a/K_s for the most conserved genes (those with ERI = 1). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

been previously shown, following different approaches (Gong et al., 2008; Jordan et al., 2002). Our result confirms those earlier observations and leads to conclude that the more a gene is shared, the more it is likely to be essential. This point will be further investigated in the next section.

Fig. 2 further shows that the number of essential genes is rather constant among bacterial genomes. In small genomes with < 1000 genes, most genes are essential. Then, as the size of the genome increases, the number of nonessential genes increases proportionally. Note that *shon* does not follow the trend: it is a species with a peculiar metabolism and, at present, is poorly covered by DEG. Independently from the genome size, each bacterial species has a core of about 500 essential genes. This observation can be related to recent experiments in synthetic biology, devoted to the in vitro assembly of artificial bacteria with minimal genomes, limited to those genes which are necessary to sustain basic life processes (Gil et al., 2004; Glass et al., 2006; Hutchison et al., 2016). In particular, the synthetic bacterium designed and synthesized by Hutchison et al. (2016) has a genome constituted by 473 genes from *Mycoplasma mycoides*, a species whose genome contains 475 genes and which is evolutionarily close to the *Mycoplasma genitalium* considered here (*myge*). Of the genes in myge, 80% are annotated as essential and the remaining 20% have no annotation yet; clearly there are still unknown functions that could be, nevertheless, essential for life.

It is tempting to suppose that the core of essential genes in the bacterial species of Fig. 2 constitutes a kind of minimal, universal and conserved genome, made by genes that have an orthologous in all species. But this is not the case. We have checked that only 83 genes are strictly retained (ERI = 1) among all the DEG-annotated bacterial species we consider (and are reported in Table 4). Among them, no one is essential in all species, but only in a fraction $f(E)$ of the bacteria. Thus, essentiality does not strictly imply orthology: genes that are essential for one species may be not essential for another one. Indeed, the experimentally determined essentiality might disagree between species for a variety of reasons, for instance because experimental conditions that are near-optimal for one species, maybe demanding for another. Concerning strictly retained genes in general, as shown in Table 1 they have a quite restricted repertoire of functions, limited to COGs J (*translation, ribosomal structure and biogenesis*: 49 cases), K (*transcription*: 7 cases), L (*replication, recombination and repair*: 7 cases) and O (*Post translational modification, protein turnover, chaperones*: 8 cases). Hence, more than half of these genes correspond to ribosomal proteins with different degree of shared essentiality, as evaluated by $f(E)$. We have checked in DrugBank that, several of these genes, as expected, are targets of antimicrobial drugs in *E. coli*, as shown with bold COG ids in Table 4, and we note that all these targets have a shared essentiality of at least 0.56. It is then tempting to suppose that the set of strictly retained genes is a reservoir of highly druggable genes, characterised both by highly shared orthology and essentiality, to be further exploited in the design of next generation antimicrobial drugs (Chessher, 2012). This result somehow specializes what Luo et al. (2015) found on the same set of bacterial species: "essential genes in the functional COG categories G, H, I, J, K and L tend to be more evolutionarily conserved than the corresponding nonessential genes in bacteria". This kind of general statement deserves more investigation. First of all, in the next section we consider how essential and nonessential genes are partitioned into different COGs.

3.2. Functional specialization of essential and nonessential genes

The heat maps of Fig. 3 represent conditional probabilities $P(COG|E)$ and $P(COG|NE)$ that essential and nonessential genes belong to the different COGs, for the various bacterial species we consider. Essential and nonessential genes have different functional spectra. In both panels, a banded vertical structure emerges which roughly separates COGs into three groups. In particular, 51% of essential genes fall into J, M, H and L, whereas 49% of nonessential genes belong to E, K, G, P and C. Table 1 synthetically shows that essential genes dominate functions related to *information storage and processing*, whereas nonessential genes prevail among the set of functions related to *metabolism*. Functions related to *cellular processes and signaling* appear to be equally shared between essential and nonessential genes. In the next section, using the criteria of the K_a/K_s ratio, we challenge the sensible statement that essential genes are subject to a stricter purifying selection than nonessential genes. If that were true, then each COG would exhibit a signature of either purifying or positive selection, on the basis of the fraction of essential genes that belong to it.

3.3. Selective pressure, conservation and essentiality

In this section we firstly consider how evolutionary pressure, as represented by the ratio K_a/K_s, correlates with the degree of retention (conservation) of bacterial genes. Note that each bacterial genome has its own level of evolutionary pressure (see Figs. 4 and 10). We thus compare, within each genome, the evolutionary pressure that is exerted over more or less conserved genes. Using the thresholds of ERI used in Dilucca et al. (2015), Fig. 4 shows that more conserved genes (with ERI > 0.6) significantly display lower values of K_a/K_s than less conserved genes (with ERI < 0.2). Interestingly, genes belonging to the core of 83 strictly conserved genes of ERI = 1, mentioned above, have levels of K_a/K_s that are systematically below the average value of the more conserved genes. This last observation stresses once more that the most conserved genes, those involved in more basic and universal functions, tend to be subject to a relatively purifying, conservative selection. Since highly conserved genes are also prone to be essential, as shown in Fig. 1, our observation confirms the previous conclusion by Luo et al. (2015) that "essential genes are more evolutionarily conserved (they are characterised by a significantly lower K_a/K_s) than nonessential ones in most of the bacteria".

Looking for a general relationship between the evolutive pressure exerted on a gene and its degree of conservation as measured by the ERI, in Fig. 5 we show that when the degree of retention increases, the Z-score of K_a/K_s systematically decreases, becoming more and more negative. This observation stresses again that those genes which are common to several species are subject to a purifying, more constrained evolution. Note that, comparatively, essential genes have systematically a Z-score which is more negative than for nonessential genes, indicating that they are, for each degree of retention, subject to a more purifying evolution.

Table 3
List of bacterial genomes. For each genome we report organism name, abbreviation, class, ncBI RefSeq, size (number of genes) and percentage of COG. Classes are: *Alphaproteobacteria* (1), *Betaproteobacteria* (2), *Gammaproteobacteria* (3), *Epsilonproteobacteria* (4), *Actinobacteria* (5), *Bacilli* (6), *Bacteroidetes* (7), *Clostridia* (8), *Deinococci* (9), *Mollicutes* (10), *Spirochaetales* (11), *Aquificae* (12), *Cyanobacteria* (13), *Chlamydiae* (14), *Fusobacteria* (15), *Thermotoga* (16). Asterisks denote genomes considered by Gerdes et al. (2003). For those genomes annotated in the Database of Essential Genes (DEG) we report the number of essential (E) and nonessential (NE) genes, as well as the coverage of essentiality.

Organism	Abbr.	Class	ncBI RefSeq	Size	% COG	E	NE	Cov. (%)
Agrobacterium tumefaciens (fabrum)	agtu*	1	nc_003062	2765	83.34			
Aquifex aeolicus VF5	aqae*	12	nc_000918	1497	86.65			
Bacillus subtilis 168	basu*	6	nc_000964	4175	76.84	271	3904	100
Bacteroides thetaiotaomicron VPI-5482	bath	7	nc_004663	4778	68.22	325	4453	100
Brucella melitensis bv. 1 str. 16 M	brme*	1	nc_003317.1	2059	93.50			
Buchnera aphidicola Sg uid57913	busg*	3	nc_004061	546	100			
Burkholderia pseudomallei K96243	bups	2	nc_006350	3398	88.80	423	2932	98.8
Burkholderia thailandensis E264	buth	2	nc_007651	3276	81.76	364	2912	100
Campylobacter jejuni	caje*	4	nc_002163	1572	82.49	222	1350	100
Caulobacter crescentus	cacr*	1	nc_011916	3885	65.55	402	2649	78.5
Chlamydia trachomatis D/UW-3/CX	chtr*	14	nc_000117.1	894	71.75			
Clostridium acetobutylicum ATCC 824	clac*	8	nc_003030.1	3602	77.80			
Corynebacterium glutamicum ATCC 13032	cogl*	5	nc_003450.3	2959	74.54			
Deinococcus radiodurans R1	dera*	9	nc_001263.1	2629	72.86			
Escherichia coli K-12 MG1655	esco*	3	nc_000913.3	4004	86.98	587	2907	87.3
Francisella novicida U112	frno	3	nc_008601	1719	82.71	390	1329	100
Fusobacterium nucleatum ATCC 25586	funu*	15	nc_003454.1	1983	79.65			
Haemophilus influenzae Rd KW20	hain*	3	nc_000907.1	1610	93.28	625	503	70
Helicobacter pylori 26695	hepy*	4	Nc_000915.2	1469	76.90	305	1065	93.3
isteria monocytogenes EGD-e	limo*	6	nc_003210.1	2867	84.33			
Mesorhizobium loti MAFF303099	melo*	1	nc_002678.2	6743	80.33			
Mycoplasma genitalium G37	myge	10	nc_000908	475	80.84	378	94	99.37
Mycoplasma pneumoniae M129	mypn*	10	nc_000912.1	648	68.62			
Mycoplasma pulmonis UAB CTIP	mypu	10	nc_002771	782	71.57	309	321	80.56
Mycobacterium tuberculosis H37Rv	mytu*	5	nc_000962.3	3936	74	592	2892	88.5
Neisseria gonorrhoeae FA 1090 uid57611	nego*	2	nc_002946	1894	76.07			
Porphyromonas gingivalis ATCC 33277	pogi	7	nc_010729	2089	65.46	463	1626	100
Pseudomonas aeruginosa UCBPP-PA14	psae*	3	nc_008463	5892	82.97	335	4461	81.4
Ralstonia solanacearum GMI1000	raso*	2	nc_003295.1	3436	81.22			
Rickettsia prowazekii str. Madrid E	ripr*	1	nc_000963.1	8433	87.76			
Salmonella enterica serovar Typhi	saen	3	nc_004631	4352	78.28	358	3992	99.96
Shewanella oneidensis MR-1	shon	3	nc_004347	4065	69.68	402	1032	32.28
Sinorhizobium meliloti 1021	sime*	1	nc_003047.1	3359	90.26			
Sphingomonas wittichii RW1	spwi	1	nc_009511	4850	83.89	535	4315	100
Staphylococcus aureus N315	stau*	6	nc_002745.2	2582	81	302	2280	100
Staphylococcus aureus ncTC 8325	stau_	6	nc_007795	2767	71.25	345	2406	100
Streptococcus pneumoniae TIGR4	stpn*	9	nc_003028.3	1814	85			
Streptococcus pyogenes MGAS5448	stpy	6	nc_007297	1865	77.52	227	1337	83.86
Streptococcus pyogenes NZ131	stpy_	6	nc_011375	1700	80.45	241	1177	83.41
Streptococcus sanguinis	stsa	6	nc_009009	2270	79.94	218	2052	100
Synechocystis sp. PCC 6803	sysp*	13	nc_000911.1	3179	76.96			
Thermotoga maritima MSB8	thma*	16	nc_000853.1	1858	86.64			
Treponema pallidum Nichols	trpa*	11	nc_000919.1	1036	71.50			
Vibrio cholerae N16961	vich*	3	nc_002505	2534	85	447	2079	99.68
Xylella fastidiosa 9a5c	xyfa*	3	nc_002488	2766	62.96			

Getting to the functional annotation provided by COGs, Luo et al. (2015) also show that essential genes in each of the COGs G, H, I, J, K and L tend to be significantly more evolutionarily conserved than nonessential genes belonging to the same COGs. It would be then natural to conclude that the nature of the evolutionary pressure that is exerted on the genes belonging to a COG depends on its content of essential genes. From Table 1 it is possible to rank bacterial COGs and their functions by their content of essential genes. In particular, note that COGs J, M, H and L contain > 51% of the annotated essential genes, by the way, a recent experiments have re-confirmed in *basu* that precisely these COGs are enriched in essential genes (see Fig. 2 in Koo et al., 2017). One would then conclude that the genes in these COGs should be under a more conservative evolutionary pressure than those belonging to the rest of the COGs. To elucidate this point, we evaluated the Z-scores of K_a/K_s over the genes of each COG with respect to the average value of this ratio over the genomes they come from (results in Table 2 and in Fig. 6). According to this analysis, one would conclude that the rank order of the evolutionary pressure on the COGs would be J, F, K, O, E, I, D, C, T, H, G, P, L, N, M, going from relatively purifying to more diversifying selection. In the first four ranks we find J, F, K and O, at variance with the expectation based on the content of essential genes. From one point of view one could think that the observed discrepancy between the ranking based on the Z-score and the one based on the essentiality content should depend on the limited coverage of the available dataset. At present not all genes in the genomes we have investigated are annotated for essentiality and, even-worse, not all the genes have been attributed to a COG class (see Table 3 to check for the coverage of the essentiality and COG annotation). From another point of view, our Z-score statistics in Fig. 6 is based on a set of 39.804 genes (annotated for K_a/K_s and COG) over a total of 127.012. We believe our Z-score statistics is sufficiently representative of the overall evolutionary pressure exerted over the COG classes. On the basis of the data in Fig. 6, we propose a tentative distinction between a set of relatively more evolutionarily conserved COGs (J, F, K, and O) and, on the other side, a set of more adaptive ones (P, L, N, and M). This distinction should be further tested, along with the progressive annotation of bacterial genomes. Notably COG J, the set of genes which is more exhaustively annotated, has the highest percentage of essential genes and

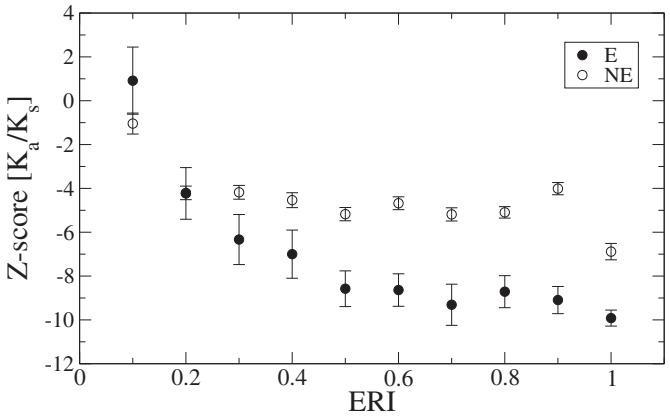

Fig. 5. Z-scores of K_a/K_s, relative to the average value in the species, for essential and nonessential genes within binned ERI values, for the DEG-annotated genomes of Table 3. Error bars are root mean square deviations for the genes falling in each bin. Z-scores decrease with the degree of conservation: the more a gene is retained among different species, the more is subject to a purifying selection. The two trends are well separated (with the exception of less conserved genes): average Z-scores of essential genes are systematically lower than those of nonessential genes, confirming that essential genes are subject to a more purifying, conservative evolutionary pressure.

with little uncertainty is, in all genomes, under a conservative evolutionary pressure. To conclude this section we focus on the set of 83 genes which are highly retained in all species (those with ERI = 1) and which display a restricted set of functional specializations. In Fig. 7 we show the histogram of the K_a/K_s values for this core of genes in all DEG-annotated species of Table 3. The plot indicates that even the genes with orthologous copies in all the genomes here considered rarely have values of K_a/K_s bigger than 1. This shows that they generally are under an overall purifying selection. It is worth to note that this distribution of K_a/K_s is consistent with the distributions of the same parameter over the different COGs. This last observation points out that a sufficiently large set of bacterial genes display a similar distribution of K_a/K_s, which in turn indicates that in general these genes are subject to an overall purifying selection ($K_a/K_s < 1$). Nevertheless, through the relative comparison of the individual Z-scores of genes in different genomes and in different COGs, we can sensibly assess that the different COGs are under diverse evolutive pressure and constraints.

3.4. Codon bias patterns in bacterial genes

Previous observations (see Plotkin and Kudla, 2011 and data therein) point to the fact that each bacterial species has a specific pattern and level of codon bias, which is strongly shared by all its genes; codon bias in specialized categories of genes appears to be just a modulation of the distinctive codon bias of the species. To check this statement, we compute *RSCU* values of each codon for our set of bacterial genomes, and plot results in Fig. 8—where both codon bias patterns and genomes are clustered according to similarity in the codon usage. The emerging striped structure indicates that these bacteria cluster into at least four groups, characterised by different patterns of codon usage (as measured by *RSCU*). Note that this grouping is not driven solely by the GC content of the genomes, as only the first group turns out to be well separated by the others in terms of the group-specific GC content distribution (See Fig. 3 and Tables 1 and 2 in Supporting Information). Overall, this is just a preliminary exploration suggesting that there should be a strong correlation between codon bias patterns of each species and his evolutionary history. Further work is needed, in our opinion, to search for hidden ecological determinants behind this rough classification based on basic codon bias.

Following this line of reasoning, we check whether the essentiality of a gene has a signature in its codon usage bias. For each of our DEG-annotated genome, we thus compute *RSCU* values separately for essential and nonessential genes. With the exception of *buth* and *stsa*, the two *RSCU* vectors are very similar for each genome. This indicates that the change in codon bias induced by essentiality, if any, is weak with respect to the prevailing codon bias signature of the species.

3.5. Conservation and codon bias of bacterial genes

In order to investigate whether the codon bias of a bacterial gene is correlated with its degree of conservation, we plot in Fig. 9 values of N_c

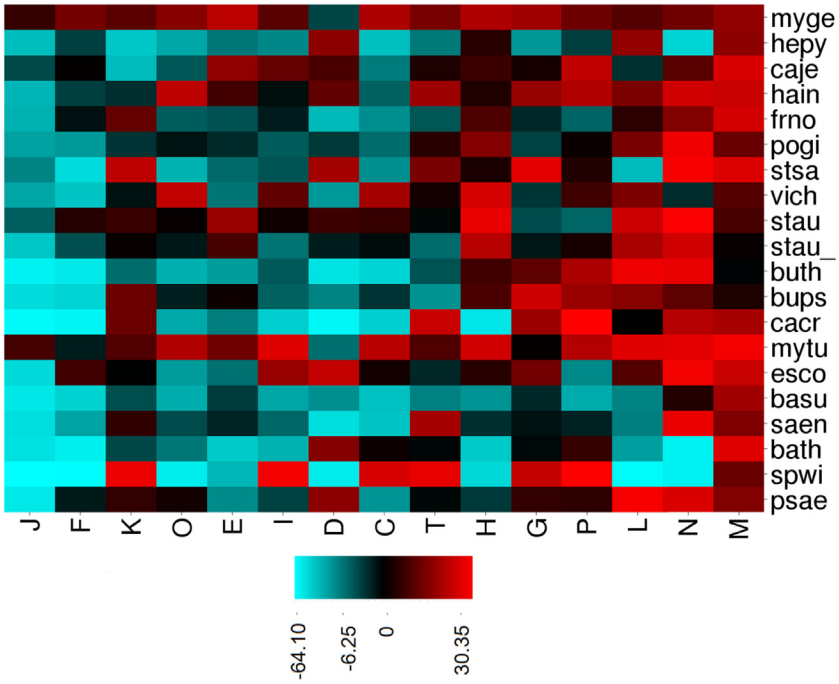

Fig. 6. Z-score of the K_a/K_s ratios for each COG, with respect to the average value in each bacteria. In the color scale, red means significant positive Z-score (selective pressure), whereas, green indicates significant negative Z-score (purifying pressure). Bacteria are ordered according to their genome size (from top to bottom), while COGs are ordered from left to right according to the ranking of Table 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
List of strictly retained genes (ERI = 1). These 83 genes have orthologous copies in all the 45 bacteria of Table 3, and are sorted in decreasing order of *f(E)*, where *f(E)* and *f(NE)* are the fraction of genomes in which each gene is annotated as essential and nonessential, respectively. *f(E)* + *f(NE)* = 1 only when the gene is fully annotated. Note the prevalence of COG J. In bold we report the genes which are antibiotic drug targets in *E. coli* (See: https://www.drugbank.ca/).

COG ID	Gene	Protein name	*f(E)*	*f(NE)*
COG0442J	proS	Prolyl-tRNA synthetase	0.92	0.08
COG0092J	rpsC	30S ribosomal protein S3	0.88	0.04
COG0049J	rpS7	30S ribosomal protein S7	0.88	0.08
COG0097J	rplF	50S ribosomal protein L6	0.88	0.08
COG0233J	frr	Ribosome recycling factor	0.88	0.12
COG0215J	cysS	Cysteinyl-tRNA synthetase	0.88	0.08
COG0201U	secY	Preprotein translocase subunit SecY	0.84	0.08
COG0088J	rplD	50S ribosomal protein L4	0.84	0.04
COG0087J	rplC	50S ribosomal protein L3	0.84	0.08
COG0528F	pyrH	Uridylate kinase	0.84	0.16
COG0525J	valS	Valyl-tRNA synthetase	0.84	0.08
COG0172J	serS	Seryl-tRNA synthetase	0.84	0.12
COG0197J	rplP	50S ribosomal protein L16	0.8	0.16
COG0090J	rplB	50S ribosomal protein L2	0.8	0.12
COG0202K	rpoA	DNA-directed RNA polymerase subunit alpha	0.8	0.12
COG0305L	dnaB	Replicative DNA helicase	0.8	0.16
COG0541U	ffh	Signal recognition particle protein	0.8	0.12
COG0264J	tsf	Elongation factor Ts	0.8	0.12
COG0522J	rpsD	30S ribosomal protein S4	0.8	0.12
COG0195 K	nusA	Transcription elongation factor NusA	0.8	0.12
COG0102J	rplM	50S ribosomal protein L13	0.8	0.04
COG0552U	ftsY	Signal recognition particle-docking protein FtsY	0.8	0.12
COG0080J	rplK	50S ribosomal protein L11	0.76	0.12
COG0100J	rpS11	30S ribosomal protein S11	0.76	0.12
COG0086K	rpoC	DNA-directed RNA polymerase subunit beta	0.76	0.2
COG0072J	pheT	Phenylalanyl-tRNA synthetase subunit beta	0.76	0.2
COG0093J	rplN	50S ribosomal protein L14	0.72	0.16
COG0211J	rpmA	50S ribosomal protein L27	0.72	0.12
COG0255J	rpmC	50S ribosomal protein L29	0.72	0.2
COG0186J	rpsQ	30S ribosomal protein S17	0.72	0.16
COG0089J	rplW	50S ribosomal protein L23	0.72	0.2
COG0091J	rplV	50S ribosomal protein L22	0.72	0.2
COG0592L	dnaN	DNA polymerase III subunit beta	0.72	0.16
COG0200J	rplO	50S ribosomal protein L15	0.72	0.16
COG0244J	rplJ	50S ribosomal protein L10	0.72	0.16
COG0180J	trpS	Tryptophanyl-tRNA synthetase	0.72	0.2
COG0018J	argS	Arginyl-tRNA synthetase	0.72	0.2
COG0587L	polC-2	DNA polymerase III subunit alpha	0.72	0.24
COG0185J	rpsS	30S ribosomal protein S19	0.68	0.2
COG0222J	rplL	50S ribosomal protein L7/L12	0.68	0.2
COG0013J	alaS	Alanyl-tRNA synthetase	0.68	0.28
COG0536R	obgE	GTPase ObgE	0.68	0.24
COG0576O	grpE	Co-chaperone GrpE	0.68	0.24
COG0223J	fmt	Methionyl-tRNA formyltransferase	0.68	0.24
COG0199J	rpsN	30S ribosomal protein S14	0.64	0.24
COG0081J	rplA	50S ribosomal protein L1	0.64	0.24
COG0203J	rplQ	50S ribosomal protein L17	0.64	0.24
COG0563F	adk	Adenylate kinase	0.64	0.32
COG0188L	gyrA	DNA gyrase subunit A	0.64	0.24
COG0188L	gyrA	DNA gyrase subunit A	0.64	0.24
COG0228J	rpsP	30S ribosomal protein S16	0.64	0.28
COG0568 K	rpoD	RNA polymerase sigma factor	0.64	0.36
COG0209F	nrdE	Ribonucleotide-diphosphate reductase subunit alpha	0.64	0.32
COG0143J	metS	Methionyl-tRNA synthetase	0.64	0.32
COG0238J	rpsR	30S ribosomal protein S18	0.6	0.28
COG0336J	trmD	tRNA (guanine-N(1)-)-methyltransferase	0.6	0.28
COG0443O	dnaK	Molecular chaperone DnaK	0.6	0.36
COG1214O	–	Glycoprotease	0.6	0.32
COG0103J	rpsI	30S ribosomal protein S9	0.56	0.32
COG0261J	rplU	50S ribosomal protein L21	0.56	0.28
COG0480J	fus	Elongation factor G	0.56	0.4
COG0335J	rplS	50S ribosomal protein L19	0.56	0.32
COG0250 K	–	Transcription antitermination protein NusG	0.52	0.32
COG0184J	rpsO	30S ribosomal protein S15	0.52	0.36
COG0231J	efp	Elongation factor P	0.48	0.48
COG0050J	tuf	Elongation factor Tu	0.48	0.36
COG0781 K	–	Transcription termination/antitermination protein NusB	0.48	0.4
COG0236IQ	–	Acyl carrier protein	0.48	0.32
COG0629L	ssb	Single-strand binding protein family	0.48	0.44
COG0858J	–	Ribosome-binding factor A	0.4	0.36
COG0484O	–	DnaJ domain-containing protein	0.4	0.56
COG0484O	–	DnaJ domain-containing protein	0.4	0.56

(*continued on next page*)

Table 4 (*continued*)

COG ID	Gene	Protein name	*f*(*E*)	*f*(*NE*)
COG0484O	–	DnaJ domain-containing protein	0.4	0.56
COG0691O	smpB	SsrA-binding protein	0.28	0.52
COG0359J	rplI	50S ribosomal protein L9	0.24	0.68
COG0571 K	rnc	Ribonuclease III	0.24	0.6
COG1136 V	–	ABC transporter ATP-binding protein	0.2	0.76
COG1136 V	–	ABC transporter ATP-binding protein	0.2	0.76
COG1136 V	–	ABC transporter ATP-binding protein	0.2	0.76
COG0084L	–	TatD family deoxyribonuclease	0.16	0.8
COG0012J	–	GTP-dependent nucleic acid-binding protein EngD	0.08	0.84
COG0313R	–	Tetrapyrrole (corrin/porphyrin) methylase protein	0.08	0.76
COG0544O	tig	Trigger factor	0	0.88

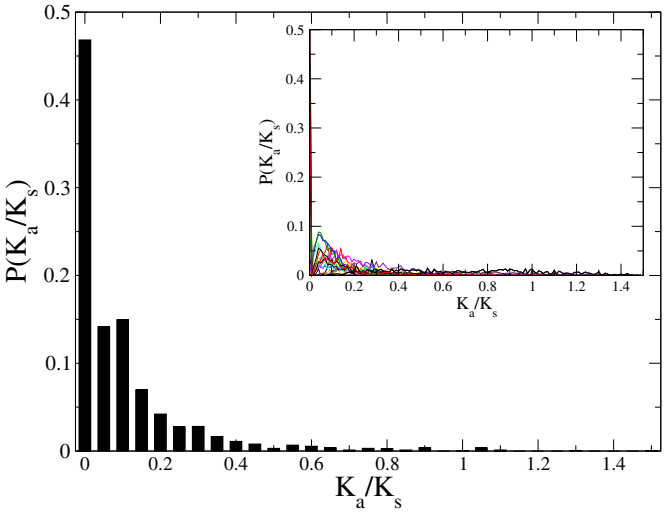

Fig. 7. Frequency distribution of K_a/K_s for the 83 genes with ERI = 1, those having orthologous in all the genomes of the DEG-annotated species of Table 3. These genes tend to display values of K_a/K_s concentrated between 0 and 0.2. Indeed, the average value is 0.085, the standard deviation is 0.004 and the median is 0.070, and only in very few cases values bigger than 1 are observed. Inset: same distributions for all genes within individual bacterial genomes.

(normalised within the species) for genes with given values of ERI. As we did in Dilucca et al. (2015) for *E. coli*, for each bacterial genome it is possible to separate groups of genes with different patterns of codon bias on the basis of their degree of conservation: those with ERI < 0.2, those with $0.2 <$ ERI < 0.6 and those with ERI > 0.6. The evolutionary codon adaptation indeed tends to be higher for genes that are more conserved (genes with ERI > 0.6 have lower values of N_c). Recall also from Fig. 1 that groups of genes with ERI < 0.6 have a probability of being essential that is < 0.2. From these observation, we can conclude that the more a gene is conserved, the more it displays a selected choice of synonymous codons.

3.6. Codon bias and evolutionary pressure

As a conclusive observation, we correlate average N_c values with corresponding average value of K_a/K_s in different bacterial genomes (Fig. 10). We show the same plot, but calculated with median values of K_a/K_s in Fig. 3 in Supplementary material. Bacterial species appear to be separated in at least three clusters, corresponding to different ranges of average values of N_c, and average values of K_a/K_s are consistent with the frequency distribution of Fig. 7. The few outliers, namely 11 (*buth*), 13 (*cacr*) and 19 (*spwi*), are the species with the highest K_a/K_s ratios and the lowest N_c values: an optimized choice of codons seems to be required to be under a more selective evolutive pressure, remember that lower values of N_c indicate more selective choice of synonymous codons. It would be interesting to have data on other bacterial genomes to complete the phase diagram correlating codon bias with evolutionary pressure, of which our Fig. 10 is just a preliminary sketch, in order to deeply investigate the possibly subtle connection between codon bias at the genetic level with the propensity to mutate at the protein sequence level.

4. Conclusions

Inspired by the results by Luo et al. (2015), with this work we further contribute to the elucidation of the intricate connections among gene essentiality, conservation, codon usage bias and evolutionary pressure. In particular, we extended the investigation we performed on *E. coli* (Dilucca et al., 2015) to several bacterial species. That essentiality, conservation, evolutionary pressure and codon bias in bacterial genes, and also in general should be strongly connected is one of those views that are widely shared among life-scientist. A view that rests on a broad literature. Our work does not convey radically new messages and the results we have shown largely confirm shared views, but the perhaps modest merit of our observations resides in having shown, carefully, quantitative correlations that have been extended to several genomes. A unified view or theoretical model of the complex interplay among the quantities we have considered here is not yet available. We made a quite unsatisfactory attempt at a ternary representation of codon usage, retention index and selective pressure in the same plot, for essential and non essential genes. The heat maps we get (see Fig. 4 in Supplementary material) do not convey any clear emerging pattern and we believe that going beyond binary correlations would require in the future a focussed effort by several researchers.

Going back to our findings: as a first result, we have shown that there is a universal exponential correlation between gene essentiality and degree of conservation: genes with high values of the evolutionary retention index (ERI) are more likely to be essential (Fig. 1). We have then observed that the number of essential genes is rather conserved among bacterial species. Small bacterial genomes are composed mainly by essential genes but, as the size of the genome increases, the number of nonessential genes increases proportionally (Fig. 2). The set of around 500 essential genes in a given bacterial genome is however not composed by genes having orthologs in all the species: essentiality does not imply orthology. This is true also for the core of 83 genes which are strictly retained (ERI = 1) in all the species here considered. These genes have a peculiar functional repertoire (mainly COGs J, but also K, L and O, see Table 4), and while they are not always essential they have, however, a probability of being essential not < 0.56. This set could thus represent an optimal reservoir of potential targets for new antimicrobial components (Fields et al., 2017).

Regarding functional classification, we have considered how the different clusters of orthologous genes (COGs) accommodate essential and nonessential genes (Fig. 3). These two groups turn out to have a complementary spectrum of functions (Table 1): essential genes mainly fall into COGs J, M, H, and L, and prevail in functions related to

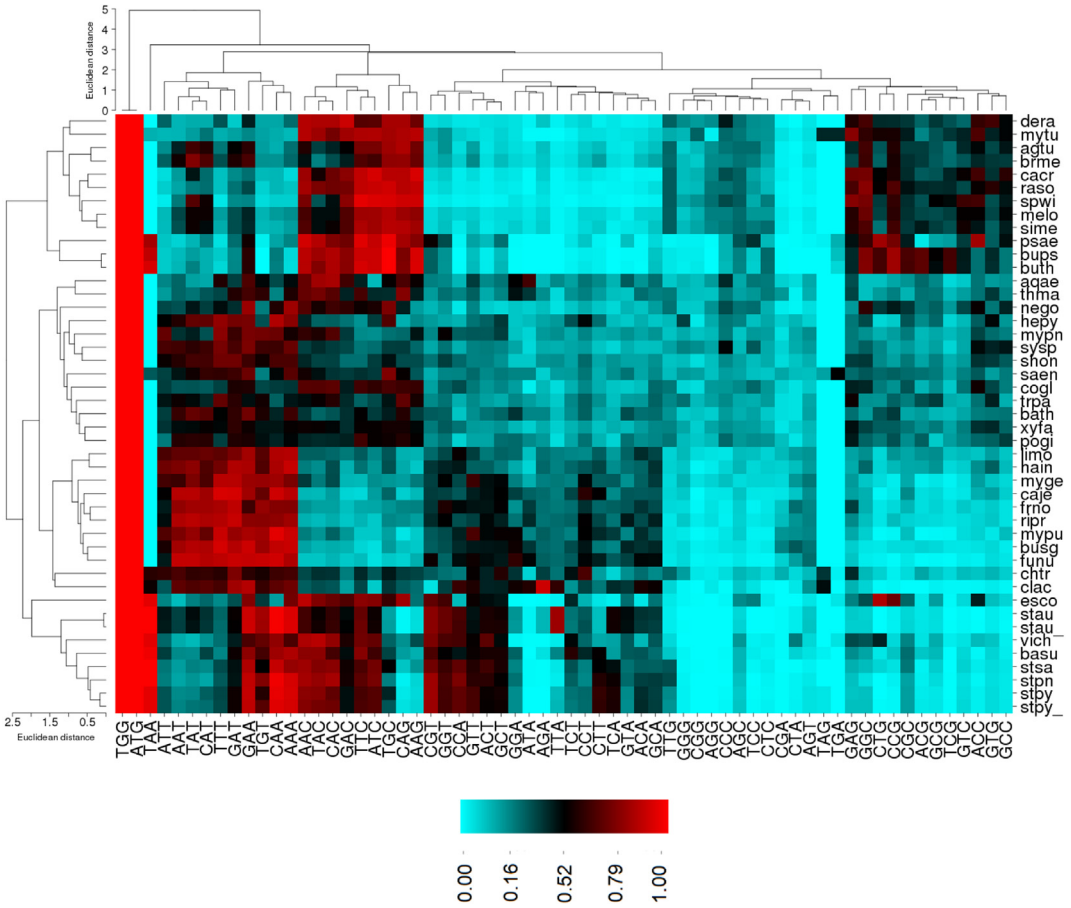

Fig. 8. RSCU values of individual codons in different species. Both genomes and groups of codons are clustered by similarity of codon usage. Note that bacterial strains of *mypu*, *shon*, *stpy* and *stpy_* are missing in the dataset of Luo et al. (2015) as well as in the figure.

information processing (translation/transcription, replication, recombination and repair), whereas, nonessential genes mainly belong to COGs E, K, G, C and P, with prevalence in metabolic functions (production and transport of energy and basic cellular constituents). Since essentiality implies a certain degree of evolutionary conservation, genes and functions of the first group of COGs should be under a relatively purifying selection, whereas, the second group of functions should be more prone to diversifying selection. Indeed, we have shown in Figs. 4 and 5 and Figs. 4 that more conserved (shared) genes feature significantly lower values of K_a/K_s than less conserved genes.

The distribution of K_a/K_s values of Fig. 7 shows that, overall, bacterial genes are under purifying evolutionary pressure, as K_a/K_s is hardly > 1. Nevertheless, through the relative comparison of the individual Z-scores of K_a/K_s for genes in different genomes and in different COGs of Fig. 6, we could sensibly assess that the different COGs are under different evolutionary pressure and constraints. We have thus

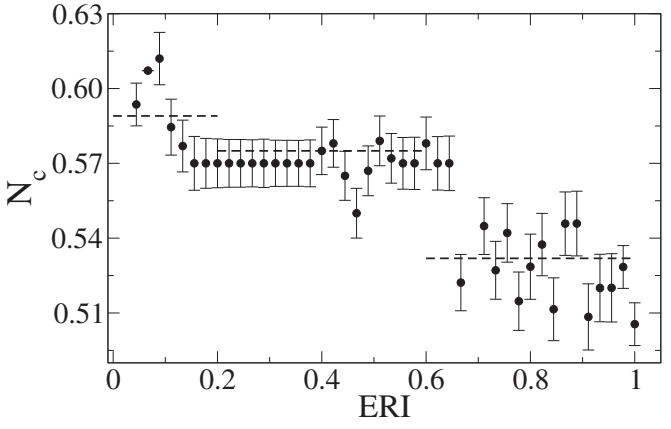

Fig. 9. Codon bias measured by N_c for bacterial genes having similar values of ERI. Note that N_c values have significantly different averages among bacterial species (Ran and Higgs, 2012) and thus, for the sake of comparison, they have been normalised within each species between 0 and 1. The dashed lines represent average codon bias levels of genes in the groups of ERI < 0.2 (specific genes), of 0.2 < ERI < 0.6, and of ERI > 0.6 (conserved genes).

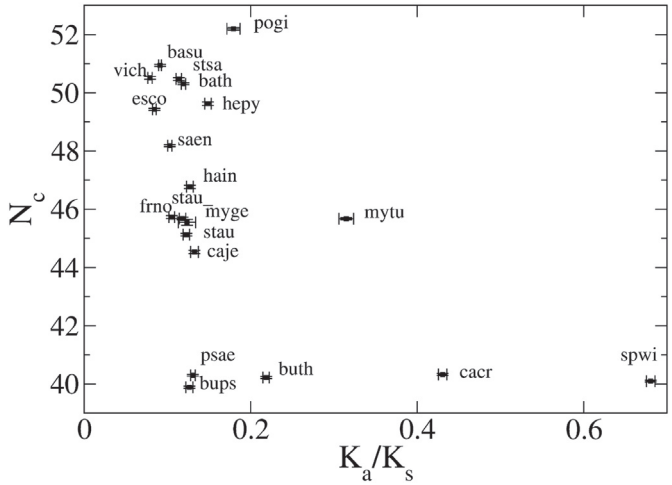

Fig. 10. Average values of (unnormalised) N_c and K_a/K_s for each bacterial species, with error bars denoting root mean square errors.

proposed a new tentative distinction between a set of relatively more evolutionarily conserved COGs (J, F, K, and O) and a set of more adaptive ones (P, L, N, and M). Such a distinction is clearly at variance with the one coming from the analysis of essential genes we discussed above. Detailing the terms of this contradiction requires further investigation, particularly for understanding the relevance of the coverage of the databases for the consistency between the test based on the COG enrichment in essential genes with that based on the Z-scores of K_a/K_s.

Using *RSCU* vectors and the effective number of codons N_c, we have finally shown that it is possible to finely classify bacteria following their codon usage patterns (Fig. 8). This classification still requires a consistent interpretation, possibly based on the analysis of ecological relationships among species. We have also shown in Fig. 9 that specific and conserved (shared) genes make slightly different use of synonymous codons: more conserved genes have a reduced number of effective codons, a clear indication that conservation of a gene rests on some kind of evolutionary optimization in the use of synonymous codons. Distinguishing essential from nonessential genes does not change the overall classification, indicating that each bacterial species has its own strong signature in codon bias. This specificity of the bias suggest where to proceed in the next future, with further investigations on the relevance of codon bias in phylogeny reconstructions and in the prediction of protein-protein interaction networks.

Acknowledgments

We warmly thank Dr. Hao Luo from the Department of Physics of Tianjin University for sharing with us his evaluations of K_a/K_s for the genomes in Table 3, that are also used by Luo et al. (2015).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2018.04.017.

References

Alvarez-Ponce, D., Sabater-Muñoz, B., Toft, C., Ruiz-González, M.X., Fares, M.A., 2016. Essentiality is a strong determinant of protein rates of evolution during mutation accumulation experiments in *Escherichia coli*. Genome Biol. Evol. 8 (9), 2914–2927. http://dx.doi.org/10.1093/gbe/evw205.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2013. GenBank. Nucleic Acids Res. 41 (D1), D36–D42. http://dx.doi.org/10.1093/nar/gks1195.

Bergmiller, T., Ackermann, M., Silander, O.K., 2012. Patterns of evolutionary conservation of essential genes correlate with their compensability. PLoS Genet. 8 (6), e1002803. http://dx.doi.org/10.1371/journal.pgen.1002803.

Chessher, A., 2012. Evaluating the suitability of essential genes as targets for antibiotic screening assays using proteomics. Protein Cell 3 (1), 5–7. http://dx.doi.org/10.1007/s13238-011-1135-x.

Dilucca, M., Cimini, G., Semmoloni, A., Deiana, A., Giansanti, A., 2015. Codon bias patterns of *E. coli*'s interacting proteins. PLoS ONE 10 (11). http://dx.doi.org/10.1371/journal.pone.0142127. e0142,127.

Dötsch, A., Klawonn, F., Jarek, M., Scharfe, M., Blöcker, H., Häussler, S., 2010. Evolutionary conservation of essential and highly expressed genes in *Pseudomonas aeruginosa*. BMC Genomics 11 (1), 234. http://dx.doi.org/10.1186/1471-2164-11-234.

Fang, G., Rocha, E., Danchin, A., 2005. How essential are nonessential genes? Mol. Biol. Evol. 22 (11), 2147–2156. http://dx.doi.org/10.1093/molbev/msi211.

Fields, F.R., Lee, S.W., McConnell, M.J., 2017. Using bacterial genomes and essential genes for the development of new antibiotics. Biochem. Pharmacol. 134, 74–86. http://dx.doi.org/10.1016/j.bcp.2016.12.002.

Galperin, M.Y., Makarova, K.S., Wolf, Y.I., Koonin, E.V., 2015. Expanded microbial genome coverage and improved protein family annotation in the cog database. Nucleic Acids Res. 43 (D1), D261. http://dx.doi.org/10.1093/nar/gku1223.

Gerdes, S., Scholle, M., Campbell, J., Balazsi, G., Ravasz, E., Daugherty, M., Somera, A., Kyrpides, N., Anderson, I., Gelfand, M., et al., 2003. Experimental determination and system level analysis of essential genes in *Escherichia coli* mg1655. J. Bacteriol. 185 (19), 5673–5684. http://dx.doi.org/10.1128/JB.185.19.5673-5684.2003.

Gil, R., Silva, F.J., Peretó, J., Moya, A., 2004. Determination of the core of a minimal bacterial gene set. Microbiol. Mol. Biol. Rev. 68 (3), 518–537. http://dx.doi.org/10.1128/MMBR.68.3.518-537.2004.

Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchison, C.A., Smith, H.O., Venter, J.C., 2006. Essential genes of a minimal bacterium. Proc. Nat. Acad. Sci. 103 (2), 425–430. http://dx.doi.org/10.1073/pnas.0510013103.

Gong, X., Fan, S., Bilderbeck, A., Li, M., Pang, H., Tao, S., 2008. Comparative analysis of essential genes and nonessential genes in *Escherichia coli* k12. Mol. Genet. Genomics 279 (1), 87–94. http://dx.doi.org/10.1007/s00438-007-0298-x.

Hurst, L.D., 2002. The K_a/K_s ratio: diagnosing the form of sequence evolution. Trends Genet. 18, 486–487. http://dx.doi.org/10.1016/S0168-9525(02)02722-1.

Hurst, L.D., Smith, N.G., 1999. Do essential genes evolve slowly? Curr. Biol. 9 (14), 747–750. http://dx.doi.org/10.1016/S0960-9822(99)80334-0.

Hutchison, C.A., Chuang, R.-Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., Pelletier, J.F., Qi, Z.-Q., Richter, R.A., Strychalski, E.A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K.S., Smith, H.O., Glass, J.I., Merryman, C., Gibson, D.G., Venter, J.C., 2016. Design and synthesis of a minimal bacterial genome. Science 351 (6280). http://dx.doi.org/10.1126/science.aad6253. aad6253.

Ish-Am, O., Kristensen, D.M., Ruppin, E., 2015. Evolutionary conservation of bacterial essential metabolic genes across all bacterial culture media. PLoS ONE 10 (4). http://dx.doi.org/10.1371/journal.pone.0123785. e0123,785.

Jordan, I.K., Rogozin, I.B., Wolf, Y.I., Koonin, E.V., 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12 (6), 962–968. http://dx.doi.org/10.1101/gr.87702.

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., Ikemura, T., 2001. Codon usage and trna genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with cg-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol. 53 (4), 290–298. http://dx.doi.org/10.1007/s002390010219.

Koo, B.-M., Kritikos, G., Farelli, J.D., Todor, H., Tong, K., Kimsey, H., Wapinski, I., Galardini, M., Cabal, A., Peters, J.M., Hachmann, A.-B., Rudner, D.Z., Allen, K.N., Typas, A., Gross, C.A., 2017. Construction and analysis of two genome-scale deletion libraries for Bacillus subtilis. Cell Syst. 4 (3). http://dx.doi.org/10.1016/j.cels.2016.12.013. 291–305.e7.

Lin, Y., Gao, F., Zhang, C.-T., 2010. Functionality of essential genes drives gene strand-bias in bacterial genomes. Biochem. Biophys. Res. Commun. 396 (2), 472–476. http://dx.doi.org/10.1016/j.bbrc.2010.04.119.

Luo, H., Gao, F., Lin, Y., 2015. Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. Sci. Rep. 5, 13210. http://dx.doi.org/10.1038/srep13210.

Luo, H., Lin, Y., Gao, F., Zhang, C.-T., Zhang, R., 2014. Deg 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic Acids Res. 42 (D1), D574–D580. http://dx.doi.org/10.1093/nar/gkt1131.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3 (5), 418. http://dx.doi.org/10.1093/oxfordjournals.molbev.a040410.

Peng, C., Gao, F., 2014. Protein localization analysis of essential genes in prokaryotes. Sci. Rep. 4, 6001. http://dx.doi.org/10.1038/srep06001.

Plotkin, J.B., Kudla, G., 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. 12, 32–42. http://dx.doi.org/10.1038/nrg2899.

Ran, W., Higgs, P.G., 2012. Contributions of speed and accuracy to translational selection in bacteria. PLoS ONE 7 (12), e51,652. http://dx.doi.org/10.1371/journal.pone.0051652.

Roth, A., Anisimova, M., Cannarozzi, G.M., 2012. Measuring Codon Usage Bias. Oxford University Press Inc., New York, pp. 189–217.

Wright, F., 1990. The "effective number of codons" used in a gene. Gene 87 (1), 23–29. http://dx.doi.org/10.1016/0378-1119(90)90491-9.

Xia, X., 2013. Dambe5: a comprehensive software package for data analysis in molecular biology and evolution. Mol. Biol. Evol. 30 (7), 1720–1728. http://dx.doi.org/10.1093/molbev/mst064.

Zhang, R., Lin, Y., 2009. Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res. 37 (suppl 1), D455–D458.