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

Maddalena Dilucca,1 Giulio Cimini,2,3 and Andrea Giansanti1,4

1 Dipartimento di Fisica, Sapienza University of Rome, Rome – Italy
2 IMT School for Advanced Studies, Lucca – Italy
3 Istituto dei Sistemi Complessi (ISC)-CNR, Rome – Italy
4 INFN Roma1 unit, Rome – Italy

Essential genes constitute the core of genes which cannot be mutated too much nor lost along the adaptive 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 correlation 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 genes with prevalent functions related to metabolism. The genes in the first group are more retained among species, are subject to a relatively purifying conservative selection and display a more selected choice 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 othologues 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 could constitute a reservoir of targets for new anti-microbial drugs.

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 of the populations of the 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 [1]. 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 [2,3]. 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 [3,4]. Essential genes constitute a kind of minimal set, required by a living cell to effectively respond to environmental changes. On the other hand, nonessential genes are those which are dispensable [5], 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 plausibly encode for functions that are fundamental for life.

As suggested by a quite wide literature, essential genes are more conserved than nonessential ones [6–10]. The term “conservation” has however a twofold meaning. On 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) [2,11]. On the other hand, a gene is (evolutionarily) conserved when it is subject to a purifying evolutionary pressure which disfavors mutations [6,12], as measured by the ratio $K_a/K_s$ of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site. In this second meaning a conserved gene is, generically, a slowly evolving gene.

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 [13]. This phenomenon, known as codon usage bias, can be measured by various indices (see [14] for an overview).

In a nutshell, our analysis reveals that those genes which are more conserved among 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 I[1]) could include putative novel targets for novel anti bacterial strategies, as suggested with rather similar arguments by Dötsch et al. [12].

MATERIALS AND METHODS

Bacterial genomes

In this work we consider a set of 45 bacterial genomes from unrelated species, whose details are provided in Table I[1]. 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/) [10]. Note that 31 of the 45 species we collected are also present in the dataset selected by Gerdes et al. [2] in their seminal paper on E. Coli’s essential genes.

Conservation and essentiality

We use the Evolutionary Retention Index (ERI) [2] as a proxy for gene conservation. We compute the ERI of a gene as the fraction of genomes in Table I[1] 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 www.essentialgene.org [8]. 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 [17, 18]. Of the 45 bacterial genomes we have collected, only 24 are covered—in toto or partially—by DEG, as indicated in Table I[1].

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 [19]. 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:

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

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

$K_a/K_s$

$K_a/K_s$ is the ratio of nonsynonymous substitutions per nonsynonymous site ($K_a$) to the number of synonymous substitutions per synonymous site ($K_s$) [12]. This parameter is widely accepted as a straightforward and effective way of separating genes subject to purifying evolutionary selection ($K_a/K_s < 1$) from genes subject to positive selective Darwinian evolution ($K_a/K_s > 1$). 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. [8] which are based on the Nej and Gojobori method [20]. 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:

$$Z_g(\langle K_a/K_s \rangle_{\text{COG}}, g) = \frac{\langle K_a/K_s \rangle_{\text{COG}, g} - \langle K_a/K_s \rangle_{g}}{\sigma_g / \sqrt{N_g}},$$

where $\langle K_a/K_s \rangle_{\text{COG}, g}$ is the average of the ratio within a given COG in a genome $g$, $\langle K_a/K_s \rangle_{g}$ and $\sigma_g$ 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).

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 [14]. We use here two basic statistical indicators: the Number of Effective Codons ($N_e$) and the Relative Synonymous Codon Usage (RSCU).

$N_e$ measures of the effective diversity of the codons used to code a given protein [21]. In principle, $N_e$ 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 $N_e$ starts from $F_\alpha$, a quantity defined for each family $\alpha$ of synonymous codons (one for each amino acid):

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

where $m_\alpha$ is the number of different codons in $\alpha$ (each one appearing $n_1, n_2, \ldots, n_m$ times in the sequence)
TABLE I. 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), Bacteroides (6), Bacteroidetes (7), Clostridia (8), Deinococci (9), Mollicutes (10), Spirochaetales (11), Aquificae (12), Cyanobacteria (13), Chlamydiae (14), Fusobacteria (15), Thermoanaerobacterales (16). Asterisks denote genomes considered in [2]. For those genomes annotated in the Database of Essential Genomes (DEG), highlighted in gray, we report the number of essential (E) and nonessential (NE) genes, as well as the coverage of essentiality assessment.

| 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   | 1407 | 86.65 |     |    |         |
| Bacillus subtilis    168        | basu  | 6     | nc_000964   | 4175 | 76.84 | 271 | 3904 | 100     |
| Bacteroides thetaotaomicron VPI-5482 | bath | 7     | nc_004663   | 4778 | 68.22 | 323 | 4453 | 100     |
| Brucella melitensis bv. 1 str. 16M | brme* | 1     | nc_003317.1 | 2059 | 93.50 |     |    |         |
| Buchnera aphidicola Sg mid57913 | busg* | 3     | nc_004061   | 514  | 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         | caer* | 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 | 2969 | 74.54 |     |    |         |
| Deinococcus radiodurans K1     | dera* | 9     | nc_001263.1 | 2629 | 72.86 |     |    |         |
| Escherichia Coli K-12 MG1655   | eso*  | 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 | fnuu* | 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     | Ne_000915.2 | 1469 | 76.90 | 305 | 1065 | 93.3    |
| *isteria monocytogenes EGD-e    | limo* | 6     | nc_003210.1 | 2867 | 81.33 |     |    |         |
| Mesorhizobium loti MAFF303099   | melo* | 1     | nc_002678.2 | 6743 | 80.33 |     |    |         |
| Mycoplasma genitalium G37       | myge  | 10    | nc_000008   | 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_000662.3 | 3936 | 74.59 | 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 | pseae | 3     | nc_008463   | 5892 | 82.97 | 335 | 4611 | 81.4    |
| Ralstonia solanacearum GMI1000  | raso* | 2     | nc_003295.1 | 3436 | 81.22 |     |    |         |
| Rickettsia prowazekii str. Madrid E | rir* | 1     | nc_000663.1 | 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 I021      | 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 | 81.02 | 320 | 2280 | 100     |
| Staphylococcus aureus nTc 8325  | stau* | 6     | nc_007795   | 2776 | 71.25 | 345 | 2406 | 100     |
| Streptococcus pneumoniae TIGR4 | stpn* | 9     | nc_003028.3 | 1814 | 85    |     |    |         |
| Streptococcus pyogenes MGA55448 | 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      | syys* | 13    | nc_000911.1 | 3179 | 76.96 |     |    |         |
| Thermotoga maritima MSB8        | thma* | 16    | nc_008533.1 | 1858 | 86.64 |     |    |         |
| Treponema pallidum Nichols      | trpa* | 11    | nc_000919.1 | 1036 | 71.50 |     |    |         |
| Vibrio cholerae N16961          | vich* | 3     | nc_002505   | 2534 | 85.44 | 447 | 2079 | 99.68   |
| Xylella fastidiosa 9a5c         | xyfa* | 3     | nc_002188   | 2766 | 62.96 |     |    |         |
and \( n_\alpha = \sum_{k=1}^{m_\alpha} n_{k_\alpha} \). \( N_c \) then weights these quantities on a sequence:

\[
N_c = N_s + \frac{K_2}{\sum_{\alpha=1}^{m_\alpha} n_\alpha} + \frac{K_3}{\sum_{\alpha=1}^{m_\alpha} n_\alpha} + \frac{K_4}{\sum_{\alpha=1}^{m_\alpha} n_\alpha},
\]

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, e.g., of proline). In this paper we evaluate \( N_c \) by using the implementation provided in DAMBE 5.0.

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

\[
RSCU_i = X_i \sqrt{\frac{1}{n} \sum_{j=1}^{n} X_j},
\]

where \( X_i \) is the number of occurrences, either in a gene or in the whole 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. 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.

\section*{RESULTS AND DISCUSSION}

\subsection*{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 \( E. \) \( C. \) by Gerdes et al. (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 been previously shown, following different approaches. 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 less than 1000 genes, most genes are essential (as established by Gerdes et al. 2001), while the number of nonessential genes increases proportionally to the size of the genomes (as shown in the Inset, where the regression line is \( y = 0.84(2)x + 20(6) \) with \( R^2 = 0.81 \)).
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 [24–26]. In particular, the synthetic bacterium designed and synthesized in [26] 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 DAG-annotated bacterial species we consider (and are reported in Table I). Among them, no one is essential in all species, but only in a fraction f(E) of the bacteria. Thus, essentiality does not imply orthology: genes that are essential for one species may be not essential for another one. However, as also shown in Table I strictly conserved genes 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). Interestingly, several of these genes are targets of antimicrobial drugs in E. Coli, as shown with bold COG ids in Table I and all these targets have a shared essentiality of at least 0.56. It is again 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 [27]. This result somehow specialises what Luo et al. [2] 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保守 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.

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 I synthesizes these observations, confirming 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 evolution than nonessential genes. If that were true, then each COG could be considered as subject either to a prevalent purifying selection or to an evolutionary pressure more prone towards Darwinian selection, on the basis of the fraction of essential genes that belong to it.

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 3 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 I, 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, i.e., 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 3, our observation confirms the previous conclusion by Luo et al. [2] that “essential genes are more evolutionarily conserved (i.e., 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 evolutionary pressure exerted on a gene and its degree of conservation as measured by the ERI, in Fig 6 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.

Getting to the functional annotation provided by COGs, Luo et al. [8] 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
TABLE III. 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 |
|--------|--------------------------|------|------|
| 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**                  |      |      |
| 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**                  |      |      |
| 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**                  |      |      |

FIG. 3. Functional specialization of essential and nonessential genes. 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.

pressure that is exerted on the genes belonging to a COG depends on its content of essential genes. From Table III 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 more than 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 figure 2 in [28]). 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 IV 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 pu-
FIG. 4. Histogram of $K_a/K_s$ for specific and conserved genes in each bacteria. Specific genes have ERI < 0.2 and conserved genes have ERI > 0.6. With the exception of myge, the average value of $K_a/K_s$ is significantly higher for specific genes. p-values of the Mann-Whitney-Wilcoxon U test are shown at the top of the panel. Red points denote averages of $K_a/K_s$ for the most conserved genes (those with ERI = 1).

FIG. 5. Evolutionary pressure and degree of conservation (ERI) for essential and nonessential genes. For each gene of the DEG-annotated genome of Table I, we compute its Z-score of $K_a/K_s$ relative to the average value in the species, and plot these values for essential and nonessential genes within binned ERI values. 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.

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 specialisations. 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 I. The plot indicates that even the genes gated are annotated for essentiality and, even-worse, not all the genes have been attributed to a COG class (see Table IV to check for the coverage of the essentiality and COG annotation). From another point of view, our Z-score statistics in figure 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 with little uncertainty is, in all genomes, under a conservative evolutionary pressure.

TABLE IV. Ranking of COGs according to Z-scores of $K_a/K_s$.

| COG ID | J | F | K | O | E | I | D | C | T | H | G | P | L | N | M |
|--------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Z-score| -20| -19| -14| -13| -13| -12| -12| -9| -9| -8| -6| -5| -5| 4 | 5 |

FIG. 6. Evolutionary pressure on different COGs. The Z-score of the $K_a/K_s$ ratios is evaluated in each COG with respect to the average value in each bacterial species. 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 IV.
Codon bias patterns in bacterial genes

Previous observations (see, e.g., [29] 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). 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.

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 Nc (normalised within the species) for genes with given values of ERI. As we did in [1] 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 Nc). Recall also from Fig 1 that groups of genes with ERI < 0.6 have a probability of being essential that is less than 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.

Codon bias and evolutionary pressure

As a conclusive observation, we correlate average Nc values with corresponding average values of K_a/K_s in different bacterial genomes (Fig 10). Bacterial species appear to be separated in at least three clusters, corresponding to different ranges of average values of Nc, and average values of K_a/K_s are consistent with the frequency distribution of Fig 7. The few outliers, namely 11 (bath), 13 (cacr) and 19 (spwi), are the species with the highest K_a/K_s ratios and the lowest Nc values: an optimized choice of codons seems to be required to be under a more selective evolutive pressure, remember that lower values of Nc 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
FIG. 8. **RSCU vectors of codon bias values for each 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. [8] as well as in the figure.

FIG. 9. **Correlation between codon bias and conservation in bacterial genomes.** The codon bias index $N_c$ is averaged over sets of bacterial genes having the same values of ERI. Note that $N_c$ values have significantly different averages among bacterial species [30] 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 $\text{ERI} < 0.2$ (specific genes), of $0.2 < \text{ERI} < 0.6$, and of $\text{ERI} > 0.6$ (conserved genes).

to deeply investigate the possibly subtle connection between codon bias at the genetic level with the propensity to mutate at the protein sequence level.

FIG. 10. **Relation between codon bias and evolutionary pressure.** $N_c$ (unnormalised) and $K_{a}/K_{s}$ are averaged over each bacterial species. Error bars are root mean square errors.

**CONCLUSIONS**

Inspired by the results by Luo et al. [8], in this work we elaborated on the intricate entanglement between gene essentiality, conservation, codon usage bias and evolutionary pressure. To task, we extended the investigation
we performed on E. Coli to several bacterial species.

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 bacteria is however not composed by genes having orthologous 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 polarized functional repertoire (mainly COGs J, but also K, L and O, see Table I), and while they are not always essential they do have a minimal probability of being so of 0.56. This set could thus represent an optimal reservoir of potential targets for new antimicrobial components.

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 III): essential genes mainly fall into COGs J, M, H, and L, while nonessential genes mainly belong to COGs E, K, G, C and P, with prevalence in metabolic and cellular 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 evolution, whereas, the second group of functions should be more prone to selective Darwinian evolution. Indeed, we have shown in Figs 4 and 5 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 greater than 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 8, we could sensibly assess that the different COGs are under different evolutionary pressure and constraints. We have thus proposed a new tentative distinction between a set of relatively more evolvable and conserved COGs (J, F, K, and O) and a set of more adaptive ones (P, L, N, and M). Such a distinction is clearly a matter of variance with the one coming from the analysis of essential genes we discussed above. Detailing the term 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 $N_e$, we have finally shown that it is possible to finely classify bacteria following their codon usage patterns (Fig 9). 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 I that are also considered in [8].

* maddalena.dilucca@roma1.infn.it

[1] M. Dilucca, G. Cimini, A. Semmoloni, A. Deiana, and A. Giansanti, PLoS ONE 10, e0142127 (2015).
[2] S. Gerdes, M. Scholle, J. Campbell, G. Balazsi, E. Ravasz, M. Daugherty, A. Somera, N. Kyrpides, I. Anderson, M. Gelfand, et al. Journal of Bacteriology 185, 5673 (2003).
[3] G. Fang, E. Rocha, and A. Danchin, Molecular Biology and Evolution 22, 2147 (2005).
[4] C. Peng and F. Gao, Scientific Reports 4, 6001 (2014).
[5] Y. Lin, F. Gao, and C.-T. Zhang, Biochemical and Biophysical Research Communications 396, 472 (2010).
[6] L. D. Hurst and N. G. Smith, Current Biology 9, 747 (1999).
[7] I. K. Jordan, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin, Genome research 12, 962 (2002).
[8] H. Luo, F. Gao, and Y. Lin, Scientific Reports 5, 13210 (2015).
[9] O. Ish-Am, D. M. Kristensen, and E. Ruppin, PLoS ONE 10, e0123785 (2015).
[10] D. Alvarez-Ponce, B. Sabater-Muñoz, C. Toft, M. X. Ruiz-González, and M. A. Fares, Genome Biology and Evolution 8, 2914 (2016).
[11] T. Bergmiller, M. Ackermann, and O. K. Silander, PLoS Genetics 8, e1002803 (2012).
[12] L. D. Hurst, Trends in Genetics 18, 486 (2002).
[13] S. Kanaya, Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura, Journal of Molecular Evolution 53, 290 (2001).
[14] A. Roth, M. Anisimova, and G. M. Cannarozzi, “Measuring codon usage bias,” in Codon evolution: mechanisms and models (Oxford University Press Inc. (New York), 2012) pp. 189–217.
[15] A. Dötsch, F. Klawonn, M. Jarek, M. Scharfe, H. Blöcker, and S. Häussler, BMC genomics 11, 234 (2010).
[16] D. A. Benson, M. Cavanaugh, K. Clark, T. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, Nucleic Acids Research 41, D36 (2013).
[17] R. Zhang and Y. Lin, Nucleic Acids Research 37, D455 (2009).
[18] H. Luo, Y. Lin, F. Gao, C.-T. Zhang, and R. Zhang, Nucleic Acids Research 42, D574 (2014).
[19] M. Y. Galperin, K. S. Makarova, Y. T. Wolf, and E. V. Koonin, Nucleic Acids Research 43, D261 (2015).
[20] M. Nei and T. Gojobori, Molecular Biology and Evolution 3, 418 (1986).
[21] F. Wright, Gene 87, 23 (1990).
[22] X. Xia, Molecular Biology and Evolution 30, 1720 (2013).
[23] X. Gong, S. Fan, A. Bilderbeck, M. Li, H. Fang, and S. Tao, Molecular Genetics and Genomics 279, 87 (2008).
[24] R. Gil, F. J. Silva, J. Peretó, and A. Moya, Microbiology and Molecular Biology Reviews 68, 518 (2004).
[25] J. I. Glass, N. Assad-Garcia, N. Alperovich, S. Yoospe, M. R. Lewis, M. Maruf, C. A. Hutchison, H. O. Smith, and J. C. Venter, Proceedings of the National Academy of Sciences 103, 425 (2006).
[26] C. A. Hutchison, R.-Y. Chuang, V. N. Noskov, N. Assad-Garcia, T. J. Deerinck, M. H. Ellisman, J. Gill, K. Kannan, B. J. Karas, L. Ma, J. F. Pelletier, Z.-Q. Qi, R. A. Richter, E. A. Strychalski, L. Sun, Y. Suzuki, B. Tsvelanove, K. S. Wise, H. O. Smith, J. I. Glass, C. Merriman, D. G. Gibson, and J. C. Venter, Science 351, aad6253 (2016).
[27] A. Chessher, Protein & Cell 3, 5 (2012).
[28] B.-M. Koo, G. Kritikos, J. D. Farelli, H. Todor, K. Tong, H. Kimsey, I. Wapinski, M. Galardini, A. Cabal, J. M. Peters, A.-B. Hachmann, D. Z. Rudner, K. N. Allen, A. Typas, and C. A. Gross, Cell Systems 4, 291 (2017).
[29] J. B. Plotkin and G. Kudla, Nature Reviews Genetics 12, 32 (2011).
[30] W. Ran and P. G. Higgs, PLoS ONE 7, e51652 (2012).
[31] F. R. Fields, S. W. Lee, and M. J. McConnell, Biochemical Pharmacology 134, 74 (2017).