Comparative analysis of essential genes and nonessential genes in Escherichia coli K12

Xiaodong Gong
Shaohua Fan
Amy Bilderbeck
Mingkun Li
Hongxia Pang
Shiheng Tao

Springer-Verlag
Mol Genet Genomics

演讲者：吴帆

Abstract

Genes can be classified as essential or nonessential based on their indispensability for a living organism. Previous researches have suggested that essential genes evolve more slowly than nonessential genes and the impact of gene dispensability on a gene's evolutionary rate is not as strong as expected. However, findings have not been consistent and evidence is controversial regarding the relationship between the gene indispensability and the rate of gene evolution.

Abstract

Understanding how different classes of genes evolve is essential for a full understanding of evolutionary biology, and may have medical relevance in the design of new antibacterial agents. We therefore performed an investigation into the properties of essential and nonessential genes. Analysis of **evolutionary conservation**, **protein length distribution** and **amino acid usage** between essential and nonessential genes in ***Escherichia coli K12*** demonstrated that **essential genes are relatively preserved throughout the bacterial kingdom when compared to nonessential genes.**

Background

Individual genes within a given species genome contribute differentially to the survival and propagation of the organism. According to their **known functional** profiles, genes can be divided into two categories: essential and nonessential genes. Essential genes are indispensable to cellular life and constitute a minimal gene set required to support a living cell. Nonessential genes are those which have been shown to be dispensable.

Here, we study essential genes by comparing them with nonessential genes in the preferred model organism, *Escherichia coli K12*.

Background

Recent investigations have led to the identification of an increasing number of essential genes in a number of different organisms.

Wilson et al. (1977) proposed that **essential genes evolving more slowly than nonessential genes**. This has more recently been termed the ‘‘knockout-rate’’ prediction (Hust and Smith 1999), a model of the relationship between **gene importance and rate of gene evolution** which has been tested with varying results.

Hust and Smith (1999) concluded that **there was no difference in evolutionary rate between essential and nonessential genes.**

Hirsh and Fraser (2001) found that **genes with smaller fitness effects evolve faster** and suggested that the evolutionary rate is negatively correlated with the fitness effect when the fitness effect is weak (<0.5).

Background

Jordan et al. (2002) found that **essential** bacterial genes are **more evolutionarily conserved than nonessential genes**.

Yang et al.(2003) found **a negative relationship between gene importance and gene evolutionn rate** in duplicate yeast genes.

Zhang and He (2005) showed that **protein evolutionary rate was significantly affected by dispensability**, even when the gene expression level was controlled for and duplicate genes excluded. They also found that: (a) the effect of gene dispensability on the evolutionary rate declines in strength with evolutionary time, and (b) protein dispensability measured in a single species predicts the short-term rate of protein evolution in other species.

Background

Despite the evidence suggesting that **protein evolutionary rate** is related to factors such as the **fitness effect** and **gene essentiality**, it is notable that the fitness effect of genes is often weakly correlated with the evolutionary rate, and furthermore that a gene's rate of evolution fails to predict whether it can be classified as essential or nonessential. This may perhaps reflect the fact that **gene importance may change during the process of evolution.**

Thus, **a gene may be essential in one species but not in others, leading to between-species variation in the gene's rate of evolution.**

Background

In order to further improve our understanding of the impact of a gene's importance on gene evolution and to therefore gain insights into the fundamental biological processes underlying evolution, we investigated the properties of essential and nonessential genes in *E. coli*. This included comparative analysis of **amino acid usage, protein length distribution, and evolutionary conservation** between essential genes and nonessential genes in *E. coli*.

Materials and methods

The lists of essential and nonessential genes were obtained from the **Profiling of E. coli Chromosome database** (http://www.shigen.nig.ac.jp/ecoli/pec/index.jsp), which classified genes as essential, nonessential or unknown on the basis of a combination of **experimental** evidence and **functional** annotations.

In addition, most **ribosomal structural genes** and **coding unique tRNAs genes** are classified as essential. Conversely, genes involved in **flagellation motility** and **chemotaxis** were classified as nonessential. Finally a total of 250 essential genes and 3,260 nonessential genes with unique Blattner numbers (Bnums) (Blattner et al. 1997) were collected.

We then retrieved DNA and protein sequences from E. coli K12 genome **U00096.2** (Hayashi et al. 2006; ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K12/) based on the Bnums, and removed 93 nonessential genes, which were discontinued from NCBI general were considered. In this way the sequences of 234 essential protein coding genes and 16 essential RNA genes, and 3,071 nonessential protein coding genes and 96 nonessential RNA genes were obtained.

Materials and methods

The *E. coli K12* homologs from the genomes of 236 diverse bacterial species were obtained from the **PEC database** (http://www.shigen.nig.ac.jp/ecoli/pecv3/index.jsp), which identified putative orthologs by the use of blast program, setting the E value at each value of $10^{-1}, 10^{-20}, 10^{-50}$.

The **evolutionary retention indexes** (ERI) for each gene was computed by dividing the number of genomes carrying orthology (No) by the total number of genomes searched (Nt): **ERI = No/Nt**.

Conservation data of every amino acid site in E. coli protein was downloaded from the **CoSMoS database** (Liuet al. 2006; http://www.biology.las.umich.edu/cosmos). From this database PSI-BLAST was used to identify homologs of *E. coli K12* proteins in the **RefSeq database**.

Materials and methods

The BLAST output was parsed and used to generate a fasta file for each individual *E. coli* protein containing the *E. coli* sequence itself and all homolog sequences, and the fasta files were then aligned using MUSCLE, a novel method **for performing sequence alignments** (Edgar 2004), **to extract amino acid conservation information.**

Protein length information was retrieved from *E. coli K12* genome which was downloaded from **Genbank**. A Mann–Whitney test was used to test whether there was significant difference between the average protein length of essential and nonessential genes.

Materials and methods

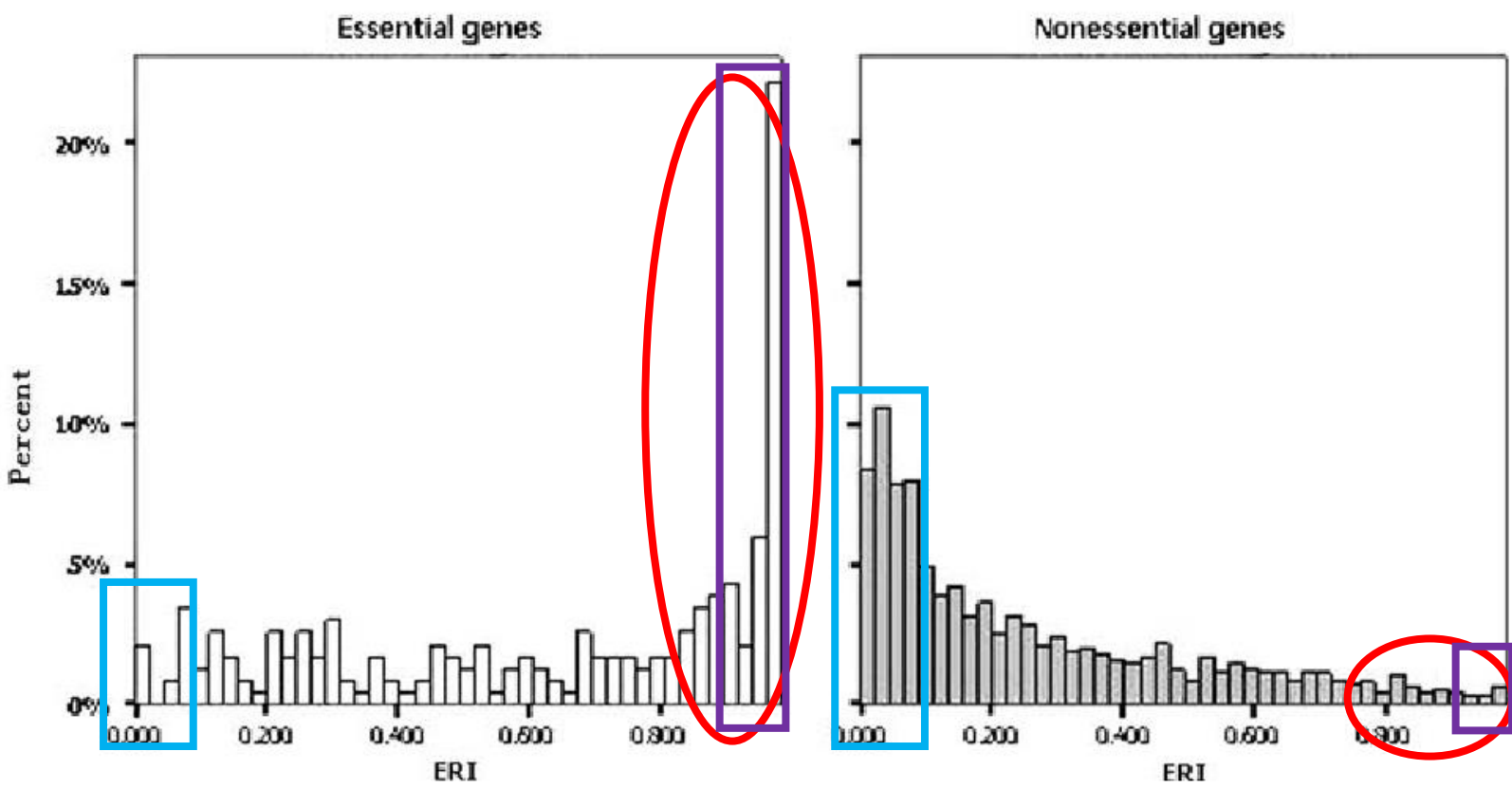
$$c = \frac{(\hat{p}_1 - \hat{p}_2)}{\sigma(\hat{p}_1 - \hat{p}_2)}$$

where $\hat{p}_1 = x_1/n_1$ (x_1 = number of single amino acids in essential genes; n_1 = overall numbers of amino acids in essential genes); $\hat{p}_2 = x_2/n_2$ (x_2 = number of single amino acids in nonessential genes; n_2 = overall number of amino acids in nonessential genes); and

$$\sigma(\hat{p}_1 - \hat{p}_2) = \sqrt{(x_1 + x_2/n_1 + n_2) \quad (1 - (x_1 + x_2/n_1 + n_2)) \quad (1/n_1 + 1/n_2)}$$

Resu
and no

/ The EI
have or
how pe
conserv
essentia
betwee



Higher ERI values are associated **with a sharp increase** in the proportion of essential genes, whereas the proportion of nonessential genes yielding high ERI values is relatively **low**.

A larger proportion of nonessential genes (~39%) are poorly conserved (ERI<=0.1) in comparison to essential genes (6.84%); conversely, a smaller proportion of nonessential genes (1.65%) are highly conserved (ERI>0.9) compared to essential genes (33.33%).

Result

These findings are **consistent** with those of Jordan (2002) which indicated that essential genes are more evolutionarily conserved than nonessential genes in bacteria, and together support the hypothesis that essential genes are more preserved throughout the bacterial kingdom.

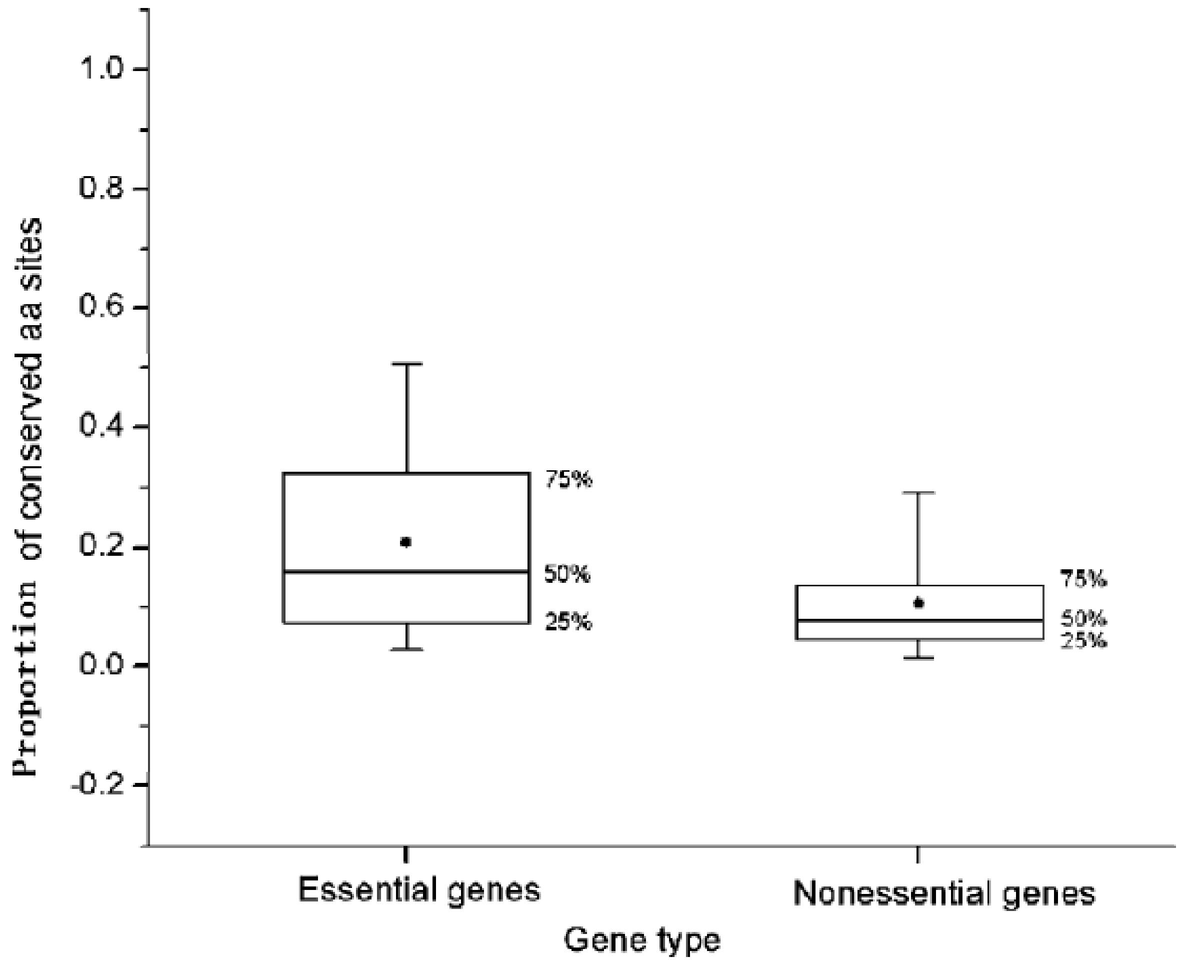
This suggests that across the course of evolutionary history, **essential genes are subject to a more intense selective procedure.**

Result

The conservation of gene function is partly determined by the conservation of **special amino acid sites**. Therefore conservation patterns at the level of amino acid sites might be expected to differ between essential and nonessential genes.

/ The extent of amino acid conservation of each protein was calculated by dividing the number of amino acids that match in that column by the total number of proteins in the multiple alignments. Amino acid sites with a conservation value at or above a cut-off value of **0.6** were considered to be ‘‘highly conserved’’. The proportion of highly conserved sites in every protein was then calculated.

Result



Statistical tests result indicates that **essential genes contain a significantly higher fraction of highly conserved sites than nonessential genes.** Results were comparable when analysis was performed using a lower cut-off value of 0.3 (data not shown).

Distribution for essential and nonessential genes can be seen in Fig. 2. What that the proportion of highly conserved sites is higher in essential genes than in nonessential genes, and furthermore that the range of ratio is wider in essential genes than in nonessential genes.

Result (Comparison on protein length distribution)

It has been reported that conserved proteins are, on average, longer than poorly conserved genes in the bacterium *E. coli*, the archaeon *Archaeoglobus fulgidus*, and the eukaryotes *S. cerevisiae, Dosophila melanogaster*, and *Homo sapiens* (Lipman et al. 2002). **We therefore performed analyses to investigate the relationship between protein length and gene importance in *E. coli K12.***

Table 1 Summary of protein lengths of essential and nonessential genes

Gene type	Number of proteins	Minimum length	Maximum length	Average length	Standard deviation
EG	234	31	1,407	343.2	249.9
NEG	3,071	14	2,367	329	209.9

EG essential genes, *NEG* nonessential genes

In Table 1, the average lengths of protein products were 343.2 and 329.9 for essential and nonessential genes. **It is not to be statistically significant.**

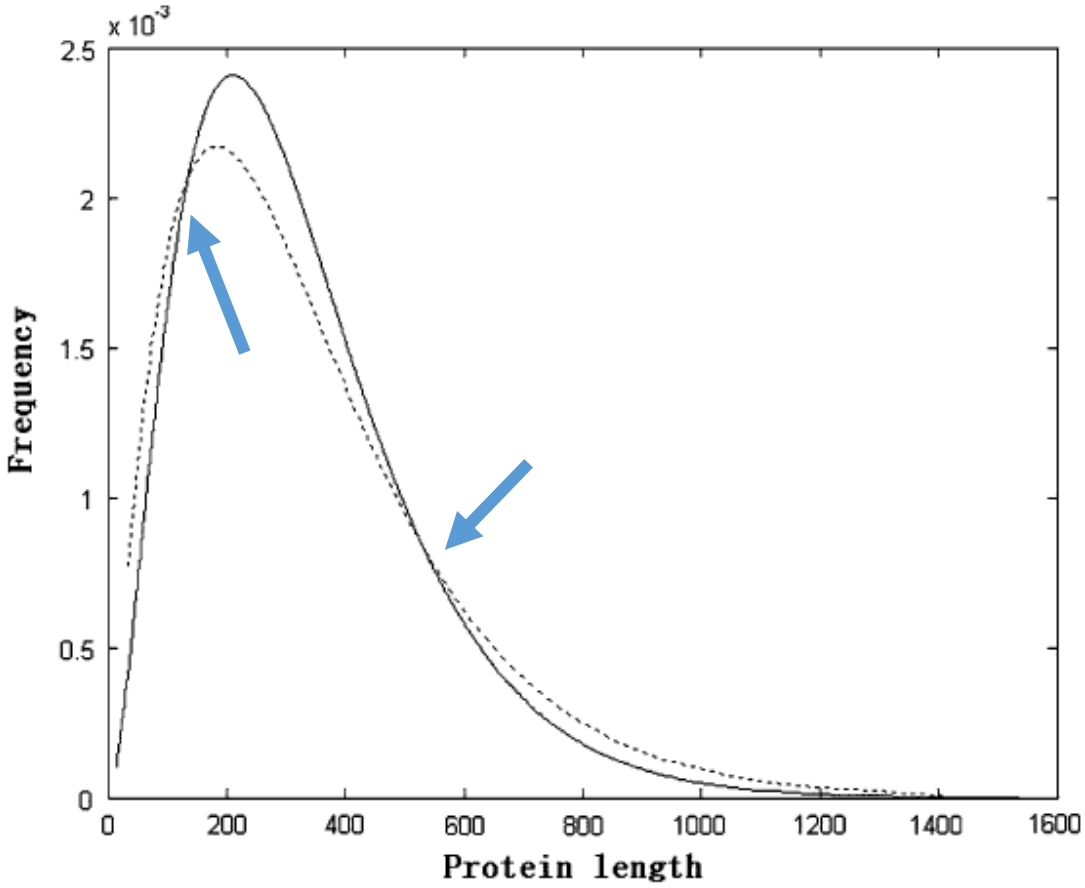
Result

To further analyze protein lengths in the two gene categories, we computed the protein-length distributions for essential genes and nonessential genes. The observed protein-length distributions fitted **a gamma distribution** (Fig. 3) with goodness-of-fit P = 0.99 and P = 0.22 for essential and nonessential genes, respectively.

In the case of gamma distribution, the protein length variation can be measured by a shape parameter, α. **The lower the a value, the great the variation.** The estimation of α for the protein products of **essential genes was 2.15**, and for the protein products of **nonessential genes was 2.75**, indicating that the protein-length variation of essential genes is relatively greater than nonessential genes.

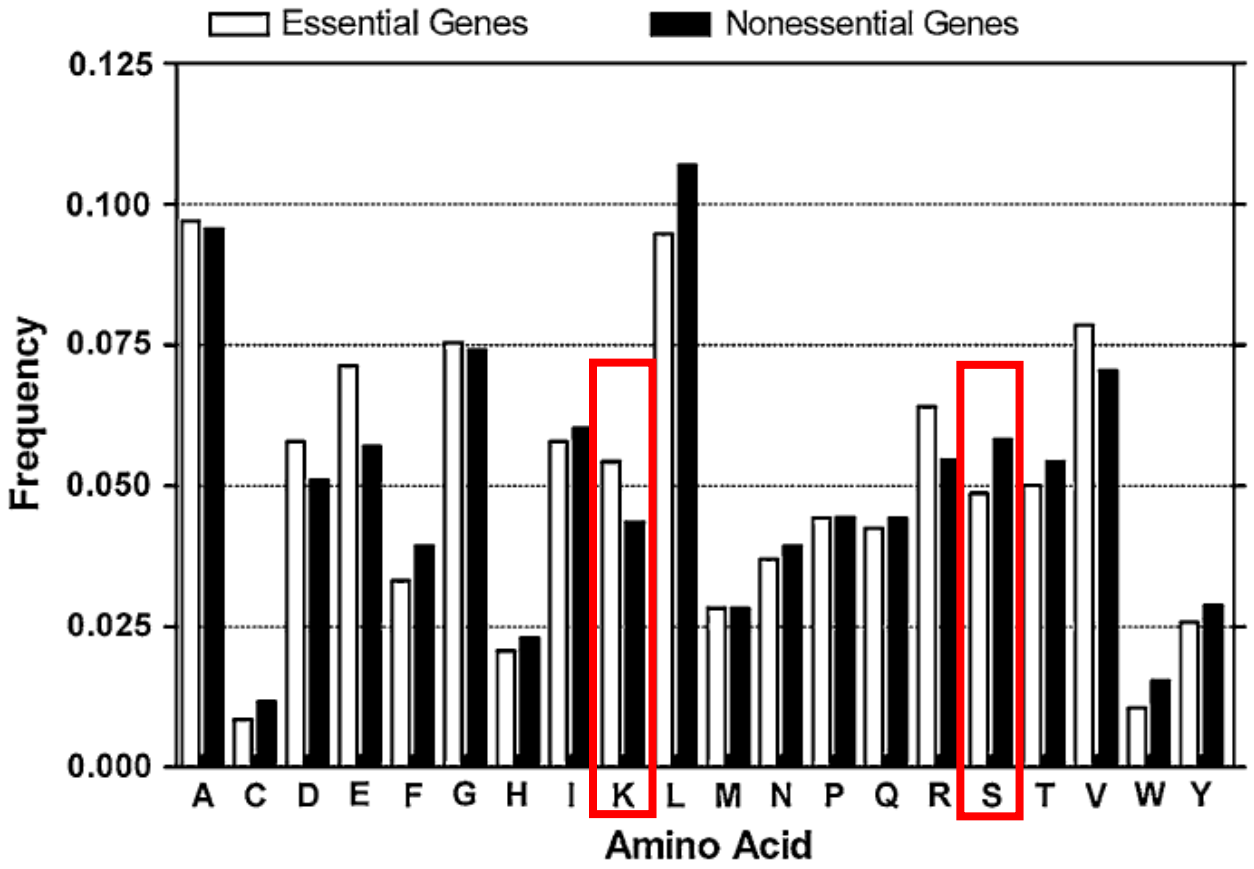
Result

Fig.3 The ***broken line*** is the protein-length distribution for **essential genes**, and the ***solid line*** is the protein-length distribution for **nonessential genes**.



As shown in Fig. 3, we also found that the proportion of **small proteins** (<139 amino acids) and the proportion of **big proteins** (>534 amino acids) are both significantly greater in essential genes than in nonessential genes, and the proportion of medium protein (>=139 amino acids, <=534 amino acids) is significantly greater in nonessential genes than in essential genes (X2 test, P = 0.0082).

Result (Compa



Comparative analysis of the protein compositions of essential and nonessential genes (Fig. 4) **very similar tendencies** in overall amino acid usage

Ala (A), Asp (D), Glu (E), Gly (G), Ile (I), Leu (L), Arg (R), Thr (T) and Val (V) are preferred amino acids in both essential and nonessential genes; only compositions of Lys (K), Ser (S) are **differentiated** between the two categories of gene, with S and K being preferred in nonessential and essential genes, respectively.

Amino acid	EG-frequency	NEG-frequency	*P* value	Physicochemical signatures	Energetic cost
A	0.096982	0.095533	0.0947	HYD	11.7
C	0.008454	0.011677	0.0001**	POL	24.7
D	0.057798	0.051075	*0.0001***	CHAR	12.7
E	0.07127	0.056937	*0.0001***	CHAR	15.3
F	0.033145	0.039311	0.0001**	HYD	52.0
G	0.075416	0.074113	0.0891	POL	11.7
H	0.020656	0.022964	0.0002**	CHAR	38.3
I	0.057798	0.060195	0.0025**	HYD	32.3
K	0.054324	0.043519	*0.0001***	CHAR	30.3
L	0.094691	0.106905	0.0010*	HYD	27.3
M	0.028301	0.028218	0.4474	HYD	34.3
N	0.03693	0.039393	0.0004**	POL	14.7
P	0.044289	0.044373	0.4606	HYD	20.3
Q	0.042458	0.044214	0.0094**	POL	16.3
R	0.064049	0.054600	*0.0001***	CHAR	27.3
S	0.048634	0.058194	0.0001**	POL	11.7
T	0.050029	0.054230	0.0001**	POL	18.7
V	0.078517	0.070410	*0.0001***	HYD	23.3
W	0.010521	0.015359	0.0001**	HYD	74.3
Y	0.025736	0.028776	0.0001**	POL	50.0

Discuss

In this work we used two distinct measures, the **ERI value** and the **proportion of highly conserved amino acid sites**, to estimate the evolutionary conservation of a gene. Both methods of analysis showed that **essential genes are more conserved than nonessential genes**, providing support for the ‘‘Knockout-rate’’ hypothesis. But results do not suggest that essential and nonessential genes can be categorized purely by their conservation properties./

In our analysis of gene distribution within proteins of different lengths, we observed a greater proportion of essential genes than nonessential genes in **small proteins** and in **large proteins**. We suggest that this may be due to the essential genes products like ribosomal protein is small and the hub protein is large.

Discuss

Our results also indicated that the usage of most amino acids **shows similar trends** in both proteins. This may mean that amino acids in the two types of genes have undergone **similar selective processes** during evolution, but **to different extents.**

Physicochemical properties are **the key factors** contributing to variation in amino acid usage. We found that almost all charged amino acids (except His) were preferred amino acids and were more biased among essential proteins, whereas other polar or hydrophobic amino acids were either biased among nonessential proteins or showed no significant difference between the two types of proteins.

This may partly be because **charged amino acids** (Lys, Arg and Glu) are found at higher frequencies in proteins involved in **genetic information processing**, whilst **hydrophobic amino acids** are of less frequency in this type of protein and are more common in proteins related to **environmental information processing**.

Discuss

In summary, we have described some of the key properties of essential and nonessential genes, and suggested how these findings inform our understanding of genetic evolution. Results suggest that ***E. coli*** **essential genes suffered stronger selective pressure over evolutionary history than nonessential genes**, and **charged amino acids are more biased in essential gene products**. In addition, **there were significant differences between the protein-length distributions for essential and nonessential genes**. Together, these findings clarify our understanding of the evolution of essential genes and contribute to our understanding of the processes underpinning the genesis of life. These results may be useful in making predictions about essential genes in nonexperimental work, and may aid the search for potential antibacterial drug targets in poorly understood pathogens.

Thank you