Molecular & Biochemical Parasitology 225 (2018) 54–66

Contents lists available at ScienceDirect

Molecular & Biochemical Parasitology

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

Whole genome analysis of codon usage in *Echinococcus*

Lucas L. Maldonado[a,*], Georgina Stegmayer[b], Diego H. Milone[b], Guilherme Oliveira[c], Mara Rosenzvit[a], Laura Kamenetzky[a,*]

[a] IMPaM, CONICET, Facultad de Medicina, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina
[b] sinc(i)-FICH-UNL-CONICET, Ciudad Universitaria, Santa Fe, Argentina
[c] Instituto Tecnológico Vale, Belém, Brazil

ARTICLE INFO

Keywords:
Codon usage bias
Selection
Cestode
Genome-wide analysis

ABSTRACT

The species of the genus *Echinococcus* are parasitic platyhelminths that cause echinococcosis and exert a global burden on public and animal health. Here we performed codon usage bias and comparative genomic analyses using whole genome and expression data of three *Echinococcus* species. The study of 4,710,883 codons, two orders of magnitude more than in previous research works, showed that the codon usage in *Echinococcus* genes is biased towards the pyrimidines T and C ending codons, with an average effective number of codons equal to 57 revealing a low codon usage bias. The gene annotations and the expression profile of 7613 genes allowed to accurately determine 27 optimal codons for the *Echinococcus* species, most of them ending in G/C. Approximately the 30% of *Echinococcus* genes analysed exhibits higher codon usage bias as well as a higher expression profile. Neutrality-plots demonstrated that the selection pressure is the main evolutionary force shaping the codon usage with a contribution of 80%. Comparative genome analyses among several tapeworm species revealed that codon usage patterns are a conserved trait in cestodes parasites. Since cestodes parasites take advantage of the host protein synthesis pathways, this study could provide valuable information associated with the parasite-host relationship that would be useful to determine which host's factors are relevant for shaping the codon usage.

1. Introduction

In the universal genetic code each codon's family contains between 1 and 6 codons depending on the amino acid for which they code. When the amino acids are encoded by two or more alternative codons, they are called synonymous codons. Although the choice between synonymous codons does not alter the primary structure of a protein, it is well known that the use of alternative synonymous codons is a non-random process [1–3]. Eighteen out of twenty amino acids are encoded by multiple synonymous codons (exceptions being methionine and tryptophan) and the probability of occurrence is not equally distributed. The phenomenon where synonymous codons are not used with equal frequencies during translation of genes is called codon usage bias (CUB). CUB is a common phenomenon in a wide variety of organisms, including prokaryotes and eukaryotes [2,4,5] and the pattern of CUB can vary among different species of organisms [2]. Even within a particular species many genes show differential patterns of codon usage which results in enhancing the efficiency and accuracy of the translation of particular genes into proteins [1,2,6,7].

The evolutionary forces that model the codon usage patterns consist of the balance between mutational bias and natural selection for the accuracy and efficiency of the translation, as well as mutation - selection - drift [8–11]. This phenomenon is widespread in all species contributing significantly to the evolution of the genome. Several studies performed on different organisms have determined which factors are responsible for shaping the CUB. These include the gene expression level, GC content, gene length, hydrophobicity and aromaticity of proteins, the content of iso-acceptor tRNAs in genomes [9,12,13]; biased gene conversion [14,15]; sex-biased gene expression [16]; mRNA structure [17,18], recombination rates, RNA stability [17,19–22], protein structure [23]; evolutionary age of genes [24]; environmental stress [25] and nitrogen availability [26]. Moreover, It is further believed that population size can influence CUB within and among species [9,27–30].

With the advent of whole genome sequencing techniques, studies of the patterns of codon usage in different prokaryotes and eukaryotes organisms have been carried out [31] and research is now aimed at analysing the codon usage pattern holistically, rather than only in

* Corresponding authors.
E-mail addresses: lucas.l.maldonado@gmail.com (L.L. Maldonado), gstegmayer@sinc.unl.edu.ar (G. Stegmayer), dmilone@sinc.unl.edu.ar (D.H. Milone), oliveirag@gmail.com (G. Oliveira), mrosenzvit@fmed.uba.ar (M. Rosenzvit), lkamenetzky@fmed.uba.ar (L. Kamenetzky).

https://doi.org/10.1016/j.molbiopara.2018.08.001
Received 16 June 2018; Received in revised form 20 July 2018; Accepted 1 August 2018
Available online 03 August 2018
0166-6851/

particular genes [3]. The study of codon usage patterns has a relevant importance in understanding the basic features of the molecular organization of a genome. Genomic research on the patterns of CUB, its causes and consequences, as well as the identification of the evolutionary forces that intervene in evolution are relevant in genomic studies for the understanding of the biology of any organism and for the accurate application of methodologies such as heterologous gene expression [32,33], the design of degenerate primers [34], as well as the prediction of gene functions [35] and the design of attenuated vaccines [36–38]. The study of CUB patterns is also useful to predict genes with high expression levels. This relies on the facts that CUB of highly-expressed genes need abundant ribosomes and matching tRNAs properly for an efficient translation conducing to the optimization of particular codons for the translation of particular genes [12,39–43].

Species of the genus *Echinococcus* are Platyhelminths parasites causatives of echinococcosis considered neglected tropical diseases (NTDs) and neglected zoonotic diseases (NZDs) that are prioritized by the World Health Organization (WHO) (http://www.who.int/neglected_diseases/diseases/en/). The cystic and alveolar echinococcoses are animals and human diseases caused by the larval stage of the tapeworms. Cystic echinococcosis (CE) is caused by *E. granulosus sensu lato*, being *E. granulosus* G1 and *E. canadensis* G6/G7 the main responsible for the global burden on public health [44] as well as *E. multilocularis* that is the causative of alveolar echinococcosis [45].

For Platyhelminths species few codon usage researches have been carried out and most of them were performed previous to the genomic era using a low representative number of sequences [46–51]. Even though, important approaches have been achieved despite the scarce genomic information, the low number of genes analysed and the lack of expression data. For the genus *Echinococcus* a high codon usage variation among the genes and GC ending codon bias have been reported. Furthermore it has been suggested that codon usage is the result of an equilibrium between mutational bias and natural selection and that natural selection could be acting in presumably highly expressed genes [51]. In further studies carried on in *Schistosoma mansoni,* it has been shown that there is a tendency to use A and T ending codons [52] and that mutational bias is the main force shaping the codon usage leaded by a genomic structure of isochoro-like [50,53]. Recently CUB researches have been performed on *Taenia* spp [54–56]. A general tendency of low CUB and GC ending codons in highly expressed genes has been described in all the *Taenia* species analysed so far. Furthermore, it has been observed that both mutational bias and natural selection shape the codon usage in these organisms.

Within the last 10 years, advancements have taken place in *Echinococcus* biology and genetics and full sequencing of *E. canadensis* G7 [57], *E. granulosus* G1 and *E. multilocularis* [58,59] have provided new tools to better understand the parasite biology and host-parasite interactions, however a whole genome description of codon usage patterns has not been performed yet using whole genome data. In this study, we analysed the nucleotide composition, the expression level and the codon usage pattern of nuclear genes of *E. canadensis* G7, *E. granulosus* G1 and *E. multilocularis* supported by transcriptomic datasets and in relation with the genomic context. The analysis includes 4,710,883 codons and more than 7613 genes that were strictly filtered and analysed giving confidence to our results.

2. Methods

2.1. Data collection and sequence pre-processing

A total of 11,435, 10,552 and 10,274 CDS of *E. canadensis* G7, *E. multilocularis* and *E. granulosus* G1 respectively were downloaded from the WormBase Parasite (https://parasite.wormbase.org/). The sequences were aligned with clean RNA-seq reads from protoscoleces of *Echinococcus* and showed > 99.0% of mapped reads. This suggested that the quality of the genes data was reliable. The CDS lacking the start and/or stop codons or containing premature stop codons were strictly filtered out. Among the remaining CDS, the orthologs were selected according to Maldonado et al. [57]. The remaining data set contained 2754, 2417 and 2442 CDSs of *E. canadensis* G7, *E. multilocularis* and *E. granulosus* G1 respectively. The CUB analyses were performed with CodonW 1.4.4 (J Peden, http://codonw.sourceforge.net/). Highly and lowly expressed genes were identified according to Tsai et al. and Zheng et al. RFPKM analysis [58,59] and used as controls to create the CAI, CBI and Fop indexes for CUB analyses. The CAI and Fop indexes were used as an indication of gene expression level under the assumption that translational selection can optimize gene sequences according to their expression levels.

2.2. Nucleotide composition analyses

The total GC content of the CDS as well as the GC content of the first (P1), second (P2), and third (P3) codon positions were calculated using custom PERL scripts. To correct the inequality composition at the third codon position [60], the three stop codons (UAA, UAG, and UGA) were excluded in the calculation of P3, and the two single codons for methionine (AUG) and tryptophan (UGG) were excluded from P1, P2, and P3. Furthermore, sliding windows analyses were performed at chromosomic levels. The *Echinococcus* chromosomes were divided into non-overlapping stretches of 5 Kb windows size and the nucleotide composition along the chromosomes was analysed. The GC content of the CDS was compared with the surrounded genomic context such as introns and 5′p and 3′p flanking regions of the genes.

2.3. Codon usage indices

The following codon indices were determined: relative synonymous codon usage (RSCU) [61], effective number of codons (ENc) [62], codon adaptation index (CAI) [61,63], codon bias index (CBI) [43], optimal frequency of codons (Fop) [10], General Average Hydropathicity (GRAVY) [61], aromaticity (Aromo) [64] and GC-content at the first, second and third codon positions (GC1, GC2 and GC3), frequency of either a G or C at the third codon position of synonymous codons (GC3s), the average of GC1 and GC2 (GC12), synonymous codon usage order (SCUO), the GC content of the 5′p and 3′p flanking regions (5p_GC, and 3p_GC), number of introns (Intron_number), length of introns (Intron_L) and the GC content of introns (Intron_GC), Mutational responsive index (MRI) and Translational selection (TrS2). Correlation analysis was performed to evaluate the relationships among the indexes.

RSCU is the ratio of the observed frequency of codons relative to the expected frequency of codons under a uniform synonymous codon usage [61]. In the absence of any CUB, the RSCU values would be 1. A codon that is used less frequently than expected will have an RSCU value of less than 1 and if a codon is used more frequently than expected the value is higher than 1 [61].

ENc indicates the degree of codon bias for individual genes. Over a range of values from 20 to 61, lower values indicate higher codon bias, while ENc equal to 61 means that all codons are used with equal probability [62,65].

CAI values measure the extent of bias toward preferred codons in highly expressed genes. CAI values range between 0 and 1.0, with higher CAI values indicating higher expression and higher CUB [61,63]. The set of sequences used to calculate CAI values in this study were the genes with high RFPKMs obtained by Tsai et al. [58] and Zheng et al. [59] so that it can provide an indication of genes expression level under the assumption that translational selection would optimize gene sequences according to their expression levels.

CBI is another measure of directional codon bias, based on the degree of preferred codons used in a gene, like to the frequency of optimal codons. It measures the extent to which a gene uses a subset of optimal codons. In genes with extreme codon bias, CBI will be equal to 1,

whereas in genes with random codon usage the CBI values will be equal to 0 [43].

Fop is a species-specific measure of bias towards particular codons that appear to be translationally optimal in particular species. It can be calculated as the ratio between the frequency of optimal codons and the total number of synonymous codons. Its values range from 0 if a gene contains no optimal codons to 1 if a gene is entirely composed of optimal codons [10]. The determination of optimal codons was carried out based on the axis 1 ordination, the top and bottom 5% of genes were regarded as the high and low bias datasets, respectively. Codon usage in the two data sets was compared using chi-square tests, with the sequential Bonferroni correction to assess significance according to Peden [66]. Optimal codons were defined as those used at significantly higher frequencies (p-value < 0.01) in highly expressed genes compared with the frequencies in genes expressed at low levels.

GRAVY values were calculated as a sum of the hydropathy values of all the amino acids encoded by the codons in the gene product divided by the total number of residues in the sequence of the protein. The more negative the GRAVY value, the more hydrophilic the protein is, whereas while the more positive the GRAVY value, the more hydrophobic the protein [61].

Aromo values denote the frequency of aromatic amino acids (Phe, Tyr, Trp) encoded by the codons in the gene product [64].

MRI is a factor of codon usage bias which is based on nucleotide composition of the gene and is a measure of the mutational drift in codons. A positive value of MRI indicates directional mutation pressure while a negative value of MRI indicates that translational selection operates on the gene. It was calculated according to Gatherer and McEwan [67,68] and is the difference between scaled chi-square (SCS) values and corrected scaled chi-square values (CSCS). Both SCS and CSCS are calculated based on the standard chi-square formula.

TrS2 estimates the codon-anticodon interaction efficiency revealing bias in favour of optimal codon-anticodon energy and represents the translational efficiency of a gene. TrS2 value > 0.5 shows bias in favour of translational selection according to Gouy and Gautier [69–71].

2.4. ENc-plots

The ENc-plot was used to analyse the influence of base the composition on the codon usage in a genome [18]. The ENc values were plotted against GC3s values and a standard curve was generated to show the functional relationship between ENc and GC3s values under mutational bias rather than selection pressure. In genes where codon choice is constrained only by a G + C mutational bias, the predicted ENc values will lie on or close to the GC3s standard curve. However, the presence of other factors, such as selection effects, causes the values to deviate considerably from the expected GC3s curve. The values of ENc range from 20 (when only one codon is used per amino acid) to 61 (when all codons are used with equal probability). The predicted values of ENc were calculated according to Hartl et al. [18].

2.5. Parity rule 2 plot

The parity rule 2 (PR2) analysis was performed in order to evaluate the impact of mutation and selection on CUB. This analysis evaluates the relative roles of mutation pressure and selective codon-usage bias on the variability of the DNA G + C content. If mutation bias is the cause of codon bias, GC or AT should be used proportionally among the degenerate codon groups in a gene. In contrast, natural selection for codon choice would cause unproportioned use of G and C (A and T) [72,73]. PR2 states that at equilibrium and under no-strand-bias conditions, the equimolar frequencies are expected between A and T (A = T) and between G and C (G = C) without regard to G + C content of DNA. PR2 is an intra strand rule that is statistically expected at equilibrium within each strand when both strands are equally susceptible to mutation and selection. [72]. Deviations from PR2 are analysed in terms of an excess of the number of guanines relative to cytosines or adenines relative to thymines and the bias is measured by GC and AT skews, (G)/(G + C) and (A)/(A + T), respectively. Now if we assume that there are no mutational pressure or selection pressure to influence the composition of the two DNA strands, the Chargaff's rules should be in force for each of the two strands and not only for double-stranded DNA.

2.6. Neutrality plots

The neutrality plot was used to measure the degree of relative neutrality when selection pressure plays a prime role in evolution [60,73]. In the neutrality plot, P12 was used as the ordinate and P3 as abscissa. Since P3 is the least restrictive position and therefore the most dependent on mutational bias, if a P12 value is as neutral as P3 to selection, all the points should be distributed along the diagonal slope line equal to 1, indicating that the bias in codon usage is only determined by a mutational bias. On the other hand, if the curve of the neutrality plot tends to be parallel to the horizontal axis (slope equal to 0) this non-correlation between P12 and P3 indicates complete selective constraints and the absence of mutation pressure [60,61,73,74].

2.7. Correspondence analysis (CA)

Correspondence analysis (CA) was used to explore the codon usage variation among genes as the multivariate statistical method that resolves the high-dimensional codon-frequency data by reducing them to a limited number of variables called principal axes. The axes represent and allow to identify the most prominent factors contributing to the variation among the genes. Since there are a total of 59 synonymous codons (including 61 sense codons, minus the unique Met and Trp codons), the degrees of freedom were reduced to 40 at removing variations caused by the unequal usage of amino acids during the correspondence analysis of RSCU [75]. The data was normalized according to Sharp and Li [61] in order to define the relative adaptiveness of each codon [66,76].

2.8. Software

The software and indexes used here have been widely used in thousands of codon usage researches in many organisms, and in related organism such as those of the class Cestoda [51,54–5677,78]. Hereby and for comparative reasons the study of codon usage in *Echinococcus* species was assessed using the following programs: CodonW (V.1.4.2) (http://sourceforge.net/projects/ codonw/), CodonO [79,80], and CUSP (http:// emboss.sourceforge.net/apps/release/6.6/emboss/apps/cusp.html). These programs were used to calculate CUB indices, such as GC, GC3s (G + C content at the third position of codons), and silent base compositions (A3s, T3s, C3s, and G3s, which indicate the frequency of codons with A, U, C, or G, respectively, at the synonymous third position). GRAVY, Aromo, RSCU and ENc values were also calculated and included in the COA analysis. All the graphs in this study were plotted with R (www.r-project.org) and the statistical analyses were carried out with the Hmisc package (https://cran.r-project.org/web/packages/Hmisc/). The tRNA genes in the *Echinococcus* genomes were searched using the tRNAscan-SE program with the eukaryote-specific parameters [81] and the gene copy number was used as an estimation of the cellular tRNA abundance.

2.9. Clustering analysis

A total of 35 codon usage parameters of the three *Echinococcus* species were integrated into an input matrix to feed the integration and clustering algorithm *omeSOM [82] (see the full parameters used in supplementary table 2.1). The optimum map size was set according to Kohonen's rule [83], that is k = 5√D, where D is total number of genes

in the matrix. The features were normalized with standard z-score. To validate the stability of the clusters found, k-means algorithm was run as well, and the clustering partitions obtained with both algorithms were compared with the Rand Index (RI) [84], obtaining a 99.0% matching. To simplify the analysis, the resulting clusters were scored and ordered according to internal compactness:

$$\overline{C_j} = \frac{1}{|\Omega_j|} \sum_{\forall X_i \in \Omega_j} ||X_i - W_j||_2$$

where $|\ |$ stands for set cardinality, x_i are data samples (gene features); w_j are the cluster (or neurons) centroids; and W_j are the set of samples in each neuron [85].

3. Results

3.1. Nucleotide composition of Echinococcus genes

The complete set of genes of *Echinococcus* species were pre-processed as described in Materials and Methods, and the remaining 7613 CDS were subjected to codon usage analysis which involved a total of 1,849,107, 1,425,315 and 1,436,461 codons of *E. canadensis* G7, *E. multilocularis* and *E. granulosus* G1, respectively. The nucleotide content analysis showed a higher GC content in coding regions than in the whole genomes and it was similar among the three *Echinococcus* species. The average percentages of GC were 49.8% for *E. canadensis* G7 (ranged from 33.3% to 60.4%), 49.6% for *E. multilocularis* (ranged from 35.7% to 59.8%) and 49.8% for *E. granulosus* G1 (ranged from 34.2% to 63.4%) (Fig. 1 A, B and C). Furthermore, the nucleotide composition of each triplet position was different from each other, and the same pattern was observed for the three species. In the first position the preferred nucleotide is a G; in the second position, the most restrictive position, the preferred nucleotides are A/T, and in the third codon position, the least restrictive, the preferred nucleotides are T/C (pyrimidines), (Fig. 1 D, E, F and G). In the three *Echinococcus* species the highest percentage of GC is observed in the first position of the codon (GC1) with 53.7%, followed by the third position (GC3) with 48.6% and finally the second position of the codon (GC2) with approximately 42.5% which is in accordance with the whole genome GC content (Supplementary Table S1.1). The non-overlapping 5 Kb windows size of *Echinococcus* chromosomes showed high GC content heterogeneity along and among the chromosomes revealing stretches with high and low GC content. However the GC content of the genes is remarkably higher than the genomic regions in which they are embedded. There are only few genes whose GC content seems to follow the patterns of the genomic GC content. Furthermore, compositional bias analysis at introns and flanking regions of the genes within the genome showed a lower GC content than in the CDS and in the third codon position. All the differences in means were significant (p-value < 0.001, confidence level of 99.9%) (Supplementary File 1).

3.2. Codon usage in Echinococcus species

The codon usage in the genes of *Echinococcus* species was assessed and the overall RSCU values for 59 sense codons were calculated. In the three *Echinococcus* species, 13 codons are the most frequently used. The RSCU values are shown in Supplementary Table S1.2. The over-represented codons (RSCU values > 1.15) were: GCU (Ala), CGU, CGC, CGA (Arg), GGU (Gly), AUU (Ile), CUU, CUC, CUG (Leu), UCC (Ser), ACC (Thr), UAC (Tyr), GUG (Val). Nevertheless most of the codons showed RSCU values close to 1 indicating that *Echinococcus* genes do not have a strong CUB. RSCU values showed no differences among the three *Echinococcus* species. Indeed, the correlation analysis of RSCU among the three *Echinococcus* species exhibited a high positive Spearman correlation (r = 0.99, p-value < 0.01) (Fig. 2).

3.3. Optimal codons and tRNA abundance

A total of 134, 849 and 242 tRNA genes were predicted in the genomes of *E. canadensis* G7, *E. multilocularis* and *E. granulosus* G1 respectively. The average RSCU values of high and low expressed genes were analysed to identify the optimal codons. In this regard, 28, 27 and 29 codons were determined to be optimal codons in *E. canadensis* G7, *E. multilocularis* and *E. granulosus* G1 respectively (Supplementary Table S1.3). These codons were found to be used significantly with more frequency in highly expressed genes (p-value < 0.01) according to the chi-square test. All of the optimal codons, except GGU and CGU contained G or C in the third position. Conversely, low expressed genes exhibited codons whose third position contained more frequently A or U. In addition, since the optimal codons tend to correspond to high expression levels of tRNAs genes or with high tRNA gene copy numbers [6], we used the tRNA gene copy numbers as an approach to measure the tRNA abundance in the cell. These analyses showed that 4 out of 28, 13 out of 27 and 8 out of 29 optimal codons correspond to the most abundant tRNAs genes (Supplementary Table S1.3) for *E. canadensis* G7, *E. multilocularis* and *E. granulosus* G1 respectively. However it should be considered that the tRNA gene copy number may be under-estimated because of the presence of gaps in the genomes. Besides, since the tRNA genes are small and are ordered in clusters [57], the presence of a gap can conduce to the loss of several genes, situation that is worsened due to the difficulty in assembling short repetitive regions. Furthermore, a significant positive correlation between CAI and Fop was also observed supporting that highly expressed genes have optimized codons.

3.4. Relationship between ENc and GC3s

The ENc-GC3s plot is widely used to determine whether the codon usage of a gene is influenced by mutation and selection. Firstly, a standard curve was generated under the assumption that the codon usage is influenced only by mutational bias rather than selection pressure [18]. In genes whose codon choice is limited only by a G + C mutation bias, the ENc values are above or very close to the standard curve. However, the presence of other phenomena such as selection causes values to deviate considerably below the expected curve. This analysis showed that most of the genes distributed below the expected curve, and only a small number of genes distributed along or above it, suggesting that the codon usage of most of the *Echinococcus* genes is influenced by other factors in addition to the mutational bias (Fig. 3 A, B and C). To gain a better understanding of the CUB the observed and expected ENc values were evaluated by means of a frequency distribution plot of $(ENc_{exp}\text{-}ENc_{obs})/ENc_{exp}$,$(ENc_{exp}\text{-}ENc_{obs})/ENc_{exp}$ (Fig. 3 D, E and F). This analysis showed a peak of maximum frequency at 0.04 and a distribution that ranged from -0.1 to 0.25 indicating that most of the genes have ENc values that are slightly different from the expected. Most of the genes distributed toward positive values which indicates a skewed codon usage, however there are also genes that distributed toward negative values indicating a random codon usage in those genes.

3.5. PR2 bias plot analyses

According to PR2, if the codon usage in genes of any organism is determined only by a mutational bias, the G + C and A + T should be used proportionally in all synonymous codons families. Conversely, if a selection force shapes the codon usage, the nucleotide proportions of the third triplet position are not necessarily proportional to each other [72]. This analysis showed that most of the genes distributed in the lower left quadrant (Supplementary File 2), which means that C and T (pyrimidines) are used more frequently than G and A (purines) in the codons of the genes of the *Echinococcus* species. This result generated a vector pointing to the lower left quadrant (-0.468; -0.438), providing additional evidence that other factors in addition to mutational bias

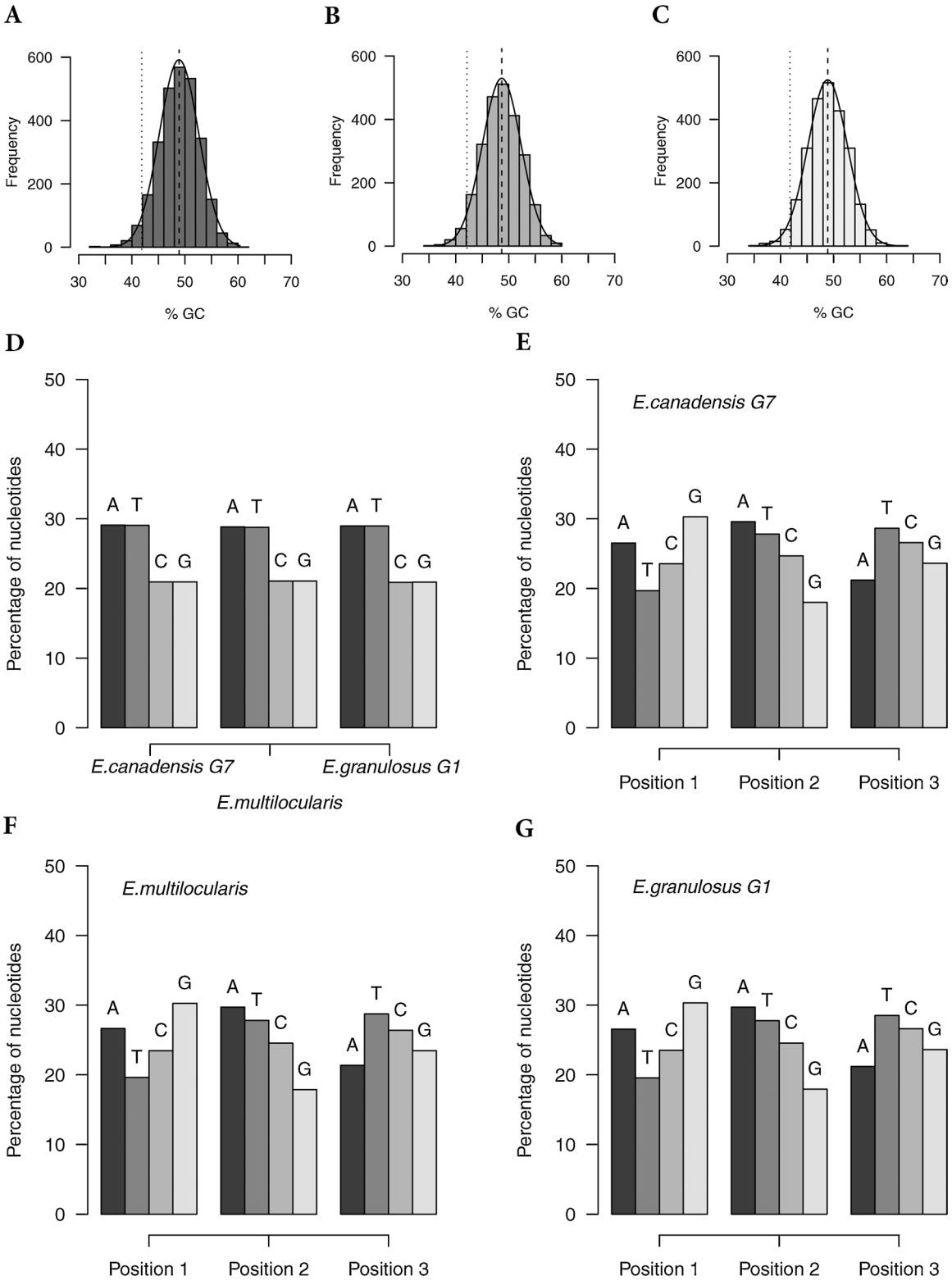

Fig. 1. *Echinococcus* genes distribution according to GC content. Gene frequency is represented by a histogram plot. Genome GC percentage (fine dotted line). Gene GC percentage (thick dotted line). A). *E. canadensis* G7, B). *E. multilocularis* and C) *E. granulosus* G1. D). Percentage of total nucleotides in *Echinococcus* genomes E). Percentage of nucleotides per position in the codon in *E. canadensis* G7. F). Percentage of nucleotides per position in the codon in *E. multilocularis.* G). Percentage of nucleotides per position in the codon in *E. granulosus* G1.

contribute to shaping the codon usage.

3.6. Neutrality plot analysis

In order to identify the main evolutionary forces that shape CUB in *Echinococcus* species, neutrality plots (P12 vs P3) analyses were

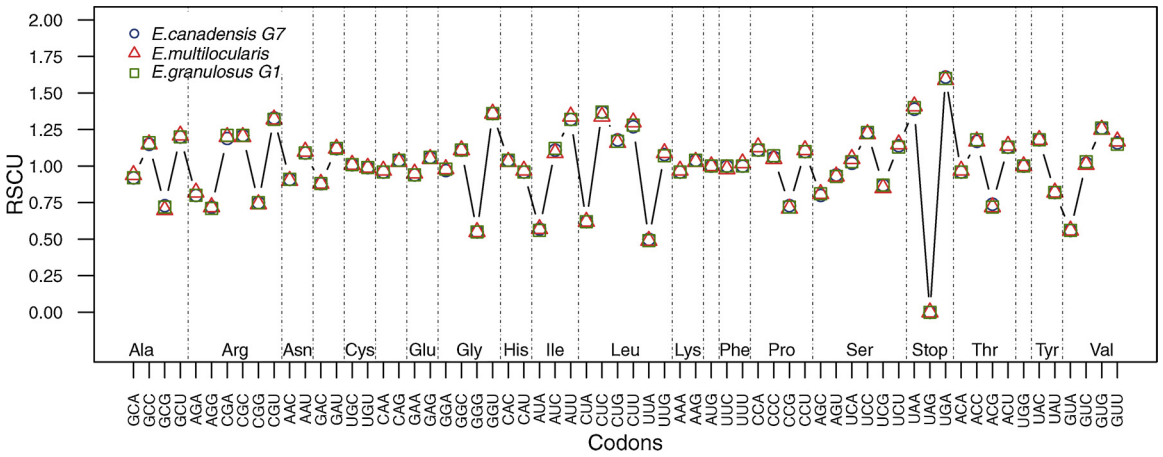

Fig. 2. *Echinococcus* RSCU values per codon. The amino acids encoded by the codons are indicated below (Spearman correlation r = 0.998; p-value < 0.01).

performed. The relationship between the GC content of P12 and the GC content of P3 was plotted and the correlation parameters were calculated (Fig. 4). These results showed significant and high positive correlations between P12 and P3 for the three *Echinococcus* species (p-values ~ 0). The regression line exhibited a slope < 1and showed that the degree of relative neutrality involved in shaping the codon usage was 21.3%, 19.5% and 21.8% for *E. canadensis* G7, *E. multilocularis* and *E. granulosus* G1 respectively. The remaining 78.7%, 81.5% and 78.2% owes to the contribution of the selection pressure. This result demonstrated that selection pressure is the main evolutionary force that

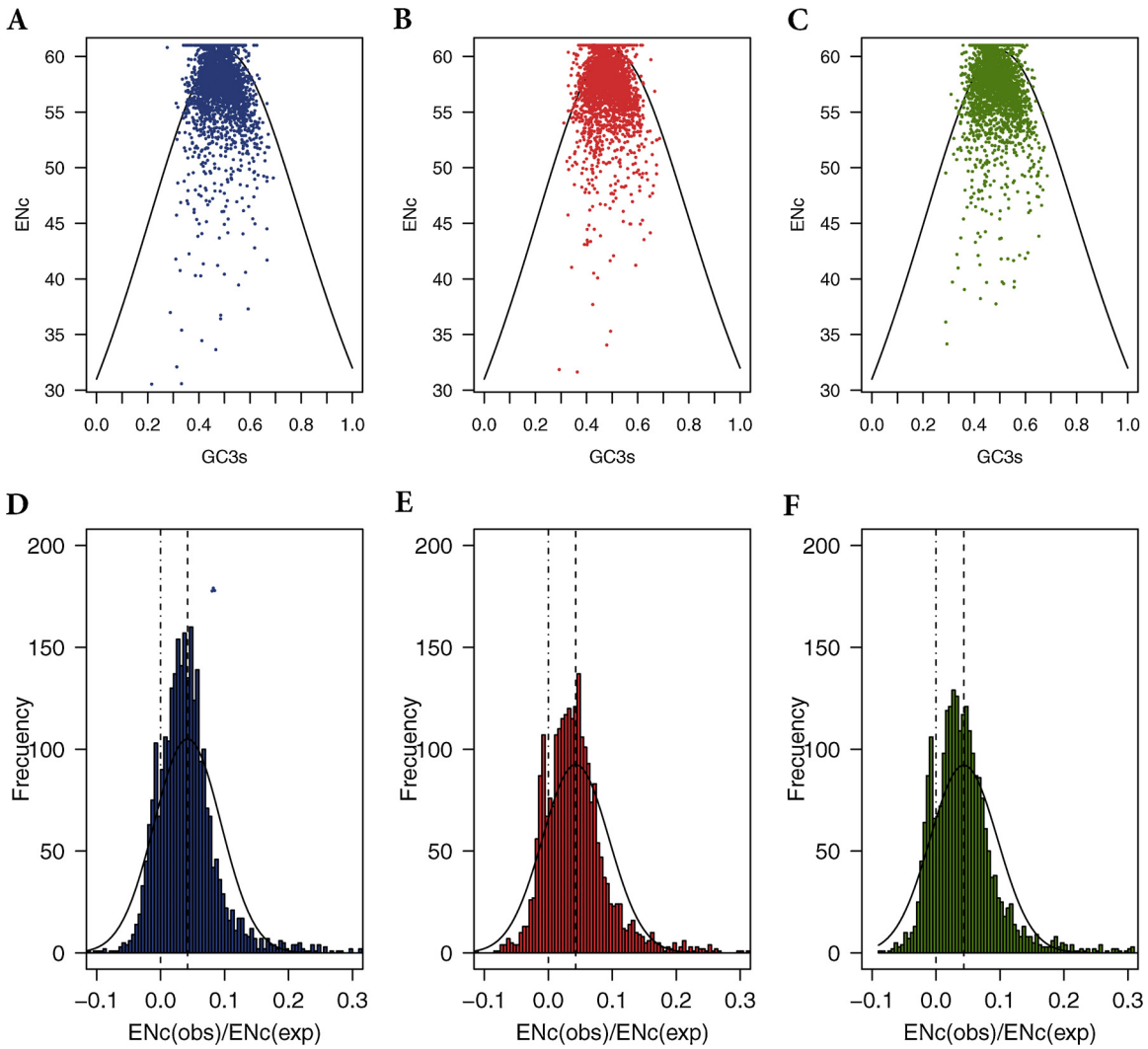

Fig. 3. Distribution of the effective number of codons (ENc) in relation to the GC3s content of *Echinococcus* genes A). *E. canadensis* G7, B). *E. multilocularis* and C) *E. granulosus* G1. The standard curve of ENc values is indicated (black solid line). Genes distribution according to the relationship between the ENc observed and the ENc expected of D). *E. canadensis* G7, E). *E. multilocularis,* and F). *E. granulosus* G1. The relationship (ENcexp - ENcobs) / Encexp is represented in the horizontal axis.

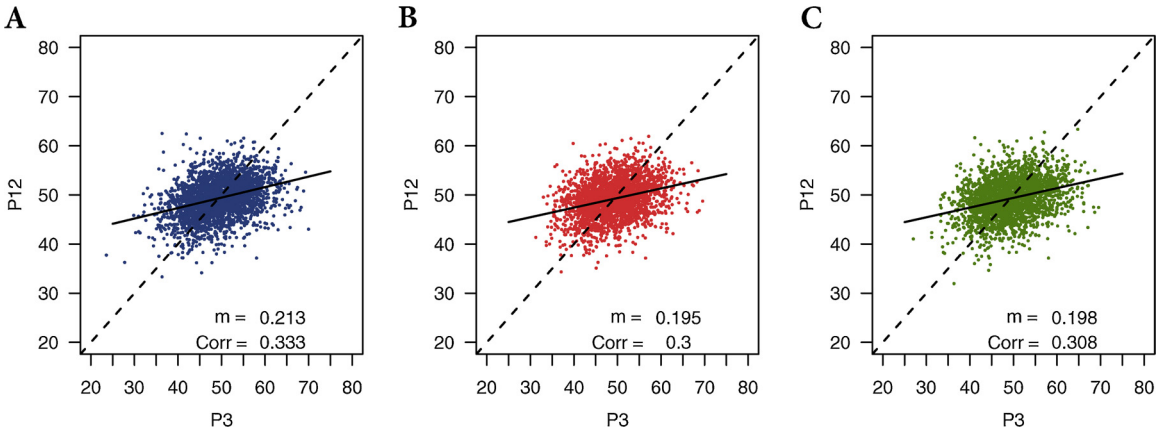

Fig. 4. Neutrality plots according to the relationship between GC composition in the first and second position of the synonymous codons (P12) with the GC composition in the third position of the synonymous codons (P3) in A). *E. canadensis* G7, B). *E. multilocularis* and C). *E. granulosus* G1. P12 represents the GC mean in the first and second position of the codons (GC1s and GC2s), while P3 represents the GC content in the third position of the codons (GC3). The black line shows the linear regression of P12 versus P3. The intersection of the regression curve and the diagonal represents the point at which P12 equals P3.

shapes the CUB in *Echinococcus*.

3.7. Mutational responsive index (MRI) and translational selection (TrS2)

The average MRI values for *E. canadensis* G7, *E. multilocularis* and *E. granulosus* G1 were 0.016, 0.017 and 0.017 respectively. Furthermore, the average TrS2 values were 0.70, 0.70 and 0.71 for *E. canadensis* G7, *E. multilocularis* and *E. granulosus* G1 respectively. The values of MRI ~ 0 indicate that mutational pressure does not exert a relevant influence on CUB. In contrast, the high values of PrS2 (> 0.5) suggest that translational selection plays a major role in shaping the codon usage of *Echinococcus* genes.

3.8. Correspondence analysis of Echinococcus genes

In order to explore the variation of RSCU in *Echinococcus* genes and to further determine the main factors that are involved in shaping the CUB, correspondence analyses and Spearman correlations were performed (Supplementary Table S2.1). COA was used to study the variation trends among the CDS by reducing the codon data into a limited number of axes. This analysis showed that the first axis accounted for the 9.64% of the total RSCU variation while the other three axes accounted for the 5.22%, 3.68% and 3.39% of the data in *E. canadensis* G7. Regarding *E. multilocularis,* the axis1 explained the 8.9% of the total variation in RSCU while axis1, axis2 and axis3 explained the 5.34%, 3.64% and 3.43% of the total variation respectively. On the other hand, for *E. granulosus* G1, the axis1, axis2, axis3 and axis4 accounted for the 9.27%, 5.23%, 3.66% and 3.47% of the total variation of RSCU, respectively (Supplementary File 3). Indeed, the first 20 PCs accounted for the 64% of the total data variation.

The COA applied on the synonymous codon usage allowed to distinguish the variability of codons composition along the PCs axes. The positions of the codons in the main axes are closer to those that share similar features. This analysis showed that the codons distribution in the space of the axis1 and axis2 were grouped mainly according to their nucleotide content in the third codon position (Fig. 5). COA applied on the CDS of the *Echinococcus* species allowed to distinguish the variation of pattern of codon usage and the responsible factors for such variation. The CUB of *Echinococcus* genes were better explained by the first two axes, which showed high significant correlation with ENc (p-values < 0.01). (Supplementary Table S2.1). The factors that exert a higher RSCU variability were identified based on the significance of the Spearman's correlation coefficients between ENc and all the parameters measured in these CUB analyses of *Echinococcus* genes. The positions of the genes along the axes identified by the correspondence analysis and whose codon usage parameters correlated significantly with ENc (p-values < 0.05) were selected to be displayed as a scatter plot along the principal axes as shown in the following sections.

3.9. Nucleotides composition effect on codon bias

The genes distribution was plotted along the axis1 and axis2 and were discriminated based on the GC content of the third codon positions (Fig. 6 A, B, E, F, I and J) and based on the total nucleotides composition (Supplementary File 4). This analysis showed that the Axis1 was the component that better dispersed the genes according to the GC content. Indeed, the *Echinococcus* genes with low GC content distributed toward the left and those with high GC content distributed toward the right side of the axis1. No differences in the genes distribution related to GC content were observed among the three *Echinococcus* species, however, *E. canadensis* G7 and *E. multilocularis* seem to present a distribution with a lower dispersion. The genes distribution was also plotted along the axis1 and according to the ENc values. The ENc vs Axis1 plot showed a slight tendency to group the genes with a high percentage of GC toward low ENc values. However, it can be observed that some genes with lower ENc values showed a GC content lower than 45%. This group of genes were used for further analyses (supplementary Table S2.2). The same pattern was observed for both, the total GC content and the GC content in the third codon position.

3.10. Effect of gene expression level on CUB

In order to accurately determine the relationship between the CUB and the gene expression levels, we firstly verified the relationship between the CAI and the gene expression levels based on the RFPKM values of *Echinococcus* species obtained by Tsai et al. [58]. This analysis showed a significant positive correlation index of 0.32 (p-value < 0.05), demonstrating that the CAI is suitable to be used as an approach for the gene expression level. Multivariate correlation analyses showed a high positive correlation of CAI with the first two Axes for the three *Echinococcus* species (Supplementary Table S2.1) and with ENc, %GC, %GC3 and gene size. COA showed that the Axis1 was the component that better dispersed the genes based on the CAI values. As shown in Fig. 6 (panels C, D G, H, K and L), the lower the ENc values, the higher the CAI values are which is in accordance with the negative correlation between CAI and ENc. Regarding their GC content, the genes with higher GC content distributed towards higher CAI values, which is in accordance with the positive Spearman's correlation coefficients (r ~ 0.5, p-value < 0.01) (Fig. 6). The same pattern was observed for the three *Echinococcus* species.

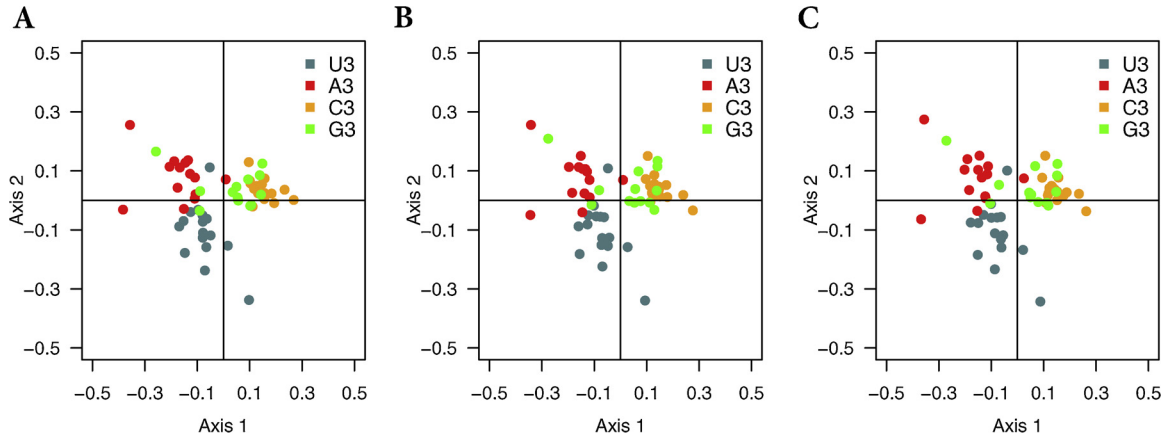

Fig. 5. Codon distribution along the main components axes (PC1 and PC2) in A). *E. canadensis* G7 B). *E. multilocularis* and C). *E. granulosus* G1. The codons are represented by dots in different colours according to the base in the third position of the codon.

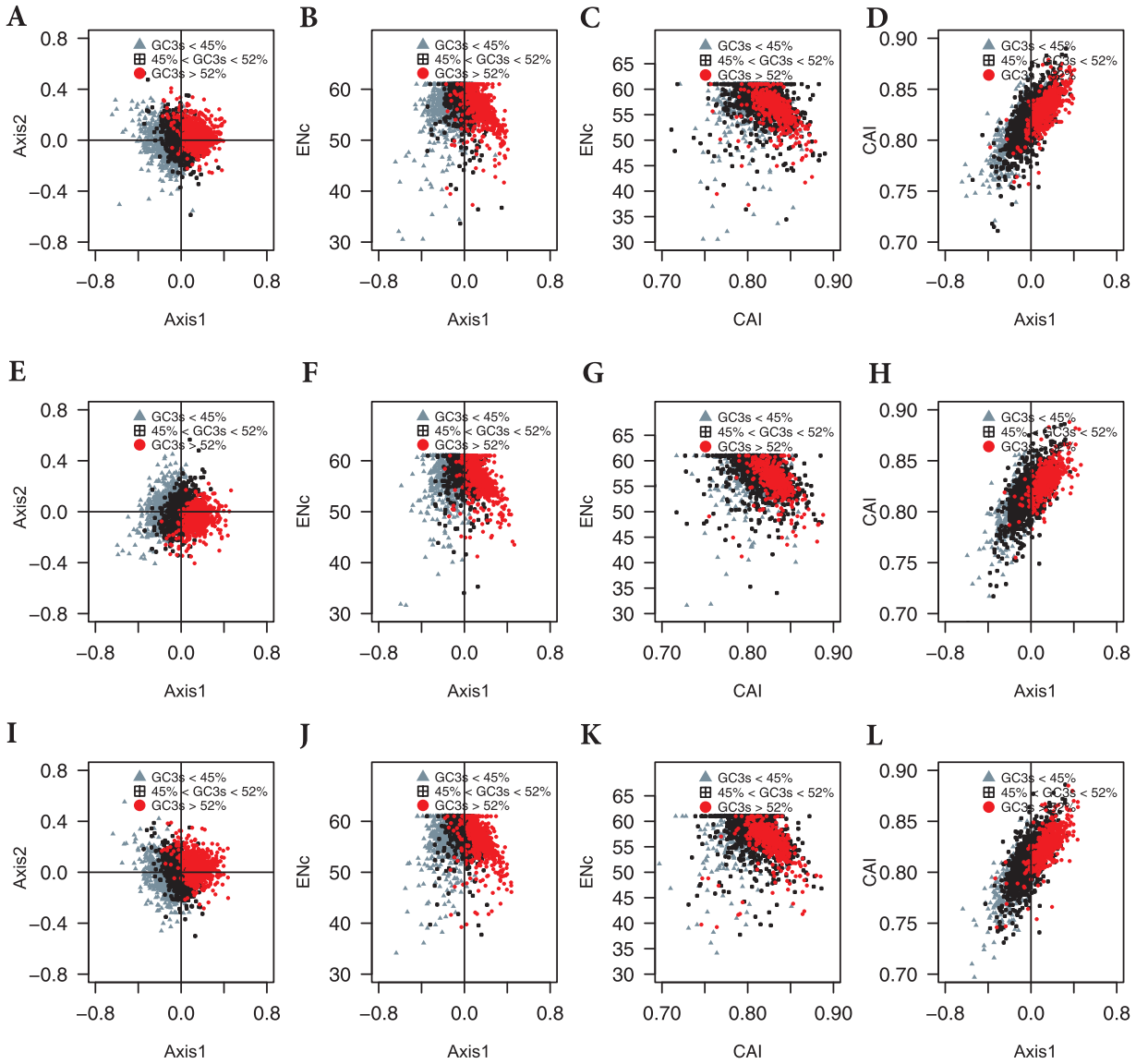

Fig. 6. Effect of gene GC content in the third codons position. Genes distribution in the coordinates of the main components of the correspondence analysis in relation to GC content in the third position of the synonymous codons. PC1 vs PC2 and PC1 vs ENc values are shown for *E. canadensis* G7 (A and B), *E. multilocularis* (E and F) and *E. granulosus* G1 (I and J). The points represent the distribution of the genes of each species are coloured according to the GC content. Effect of the gene expression level on CUB. Correspondence analysis of ENc, CAI and GC content. Gene distribution of CAI vs ENc and CP1 vs CAI values for *E. canadensis* G7 (C and D), *E. multilocularis* (G and H) and *E. granulosus* G1 (K and L). The points represent the distribution of the genes of each species coloured according to the total GC content.

3.11. Effect of hydrophobicity, aromaticity and gene length of encoded proteins on CUB

To assess whether the CUB is influenced by Hf, Arom and the genes length, correlation analyses and COA were performed. This analysis showed low but significant correlations between Hf and the ENc. Conversely, the correlation between Arom and ENc was not statistically significant for any of the *Echinococcus* species (Supplementary Table S2.1). These results suggested that the codon usage variation is influenced by the degree of the proteins hydrophobicity, but not by the aromaticity. Nevertheless, ENc vs Hf and Hf vs Axis1 plots showed neither a clear influence of Hf on the CUB nor on the gene expression level (Supplementary File 5).

With regard to the effect that the protein length exert on the CUB, a significant low positive correlation between L_aa and ENc was observed for the genes of the three *Echinococcus* species ($r = 0.04$, p-value < 0.05) (Supplementary Table S2.1). The smallest proteins ($100 \leq$ L_aa ≤ 350), (~38.5 KDa) showed a significant higher positive correlation between L_aa and ENc ($r = 0.15$, p-value < 0.05) indicating that in this L_aa range the higher the size, the lower the CUB is. COA and a scatter plot of *Echinococcus* genes were performed to analyse the relationship between ENc and L_aa and were also discriminated based on the CAI values. As shown in Supplementary File 6, the largest genes exhibited high ENc values, whereas smaller genes showed a wide range of ENc and CAI values.

3.12. Comprehensive and integrated analysis of the effect of different factors on the CUB in Echinococcus species

In order to obtain a global understanding of the influence of certain factors on CUB, the *Echinococcus* genes were classified as follows: GC < 42% and GC > 55%; GC3 < 38% and GC3 > 60%; arom < 0.035 and arom > 0.15; Hf < -0.9 and Hf > 0.46; ENc < 50 and ENc = 61 and CAI < 0.75 and CAI > 0.85. Under these criteria a total of 850, 755 and 801 genes of *E. canadensis* G7, *E. multilocularis* and *E. granulosus* G1 were identified respectively (Supplementary Table S2.2). The total genes distribution of the above-mentioned groups was scattered along the principal axes (Fig. 7). This analysis showed that the genes distributed along the Axis1 mostly according to the degree of GC, GC3s content and the gene expression levels. Indeed, the genes with higher percentage of GC and higher expression level distributed towards positive values of the axis1. In relation to the aromaticity and hydrophobicity, the genes with high and low Arom and Hf values scattered close to the intersection of the Axis1 and Axis2. The genes with higher hydrophobicity distributed towards higher values along the Axis1 in the three *Echinococcus* species. On the other hand, the genes with higher aromaticity scattered toward lower values along the Axis1. Finally, genes with lower ENc values scattered toward higher values along the Axis1.

3.13. Genes involved in the maximum variability of codon usage identified by COA and clustering analysis

In order to understand the relationship between the biological function and the CUB of the *Echinococcus* genes, those genes identified in the previous section were annotated and classified according to the GO terms of the Molecular Function category (MF). MF GO terms were assigned to 519 out of 850 for *E. canadensis* G7 genes, 458 out of 755 for *E. multilocularis* genes and 472 out of 801 for *E. granulosus* G1 genes. This analysis showed that the categories "Binding" (~53%) and "Catalytic activity" (~23,6%) were the most abundant categories, which is in accordance with the annotations described by Maldonado et al. [57]. Nevertheless, the group of genes that showed extreme and correlated CAI and ENc values, (i.e., the genes with optimized codons) have fundamental functions involved in the cellular cycle such as kinases, proteins involved in the messenger RNA biogenesis, nucleotide biogenesis, DNA repair and recombination proteins, transporters, cytoskeleton proteins, exosome and ribosomal proteins (Supplementary File S2.3). In addition, further analyses using clustering algorithm based on Self-Organizing Maps [82] grouped the genes of the three *Echinococcus* species into 439 clusters based on their codon usage parameters (Supplementary Table S2.4). This clustering analysis allowed to identify the 80.2% of the previously defined orthology groups [57] based only on codon usage parameters without the requirement of sequence alignment, which reinforces the fact that the codon usage pattern is a conserved trait in *Echinococcus* species.

4. Discussion

CUB is a generalized feature of the genomes of many organisms that is deeply influenced by evolutionary phenomena, and results basically from the balance between mutational bias and natural selection. The extent of CUB use to be specific for individual genes groups within a genome of a particular organism and is determined by multiple factors. In this study based on a general assumption that mutation, selection, and random drift represent the three main forces involved that give rise to CUB [6,11,22,86–88], the degree of the contribution that the mutational bias and natural selection pressure exert on the genes of three

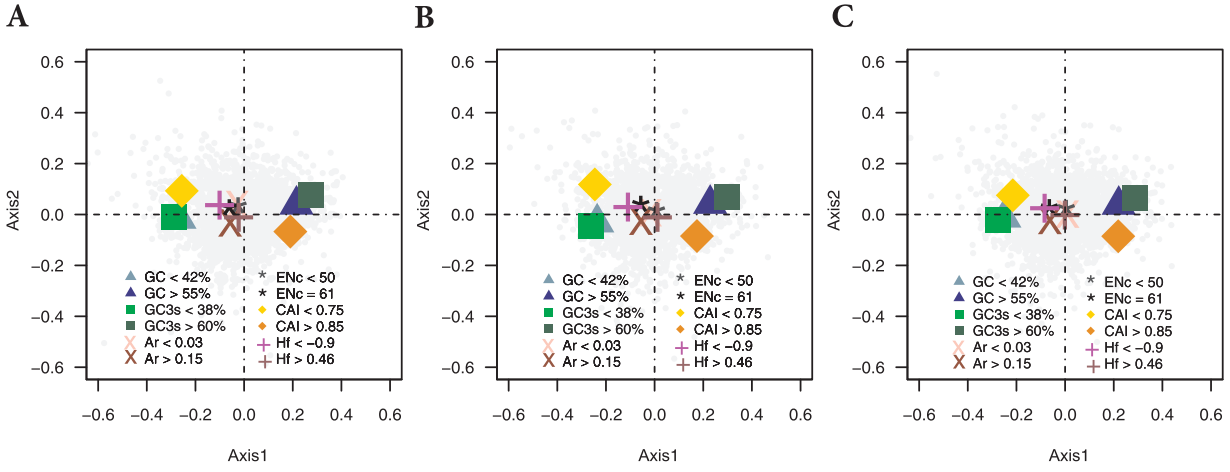

Fig. 7. Integrated analysis of the influence of different factors on the CUB in *Echinococcus*. Genes distribution in principal components of the *Echinococcus*. A). *E. canadensis* G7, B). *E. multilocularis* and C). *E. granulosus* G1. The plot symbols referenced in the graphs represent the mean of the distribution of the genes grouped according to supplementary Table S4. The light grey points represent the total distribution of the *Echinococcus* genes.

Echinococcus species were determined and compared to each other. Furthermore, nucleotide compositional analysis and the influence of several factor on CUB was characterised.

The nucleotide composition of the genes and the genome itself is considered one of the most important factors that shape the codon usage patterns of the genes, being the GC content the reflect of the general trend of codon mutation in relation to the genomic context [55]. In contrast to previous reports [46,48,51], here we demonstrated by means of whole genome analysis of three *Echinococcus* species that the codon usage is biased towards the pyrimidines T and C ending codons. Since in our studies, we analysed two orders of magnitude more than in the previous research works (~4.700.000 codons against ~20.000), these differences could be possibly due to sampling limitations in the pre-genomic era. Furthermore, we identified a heterogeneous compositional bias in different regions of the genome and we observed that the GC content of the CDS is remarkably higher than the neighbouring non-coding regions such as introns, and the 5′p and 3′p flanking regions. Indeed, only a few genes match the GC content patterns of the non-coding genome regions.

An incremented level of GC3s in certain genes has been reported by Fernández et al. in presumably highly expressed genes [51]. The optimal codons are usually defined as codons that are present significantly more often in highly expressed genes relative to their frequency in lowly expressed genes [10]. In our study we identified 28, 27 and 29 optimal codons for *E. canadensis* G7, *E. multiloculais* and *E. granulosus* G1 respectively; most of them ending in G or C except the codons GGU and CGU. This pattern has also been observed in *Taenia* spp [54–5677] suggesting a conserved trait in cestodes. However, these studies showed that not more than the 50% of the optimized codons correlated with the copy number of the tRNA genes. This low correlation may be explained by different patterns of mutation in *Echinococcus* genomes in comparison with those species in which codon usage and tRNA coevolution play a major role [6,10,39,89]. As a general rule, certain regions of the genome, those enriched in G + C, are also enriched in highly expressed genes. This compositional pattern could explain why CUB in highly expressed genes are enriched in G + C content, especially in the third synonymous codon position that is the least restrictive to vary. This would also explain the non-significant association with isoaceptor tRNAs and why most of the putative optimal codons are all G or C ending. However the low correlation between the GC content of CDS and the GC content of introns, 5′p and 3′p regions suggests that mutational bias plays a minor role and that natural selection is acting at the synonymous third codon position increasing the frequency of optimal codons in highly expressed genes as previously suggested by Fernández et al. [51].

Some studies performed on *Drosophila* sp. and *Caenorhabditis* sp. have demonstrated that highly expressed genes have higher CUB, explaining that this phenomenon is due to the preferential use of synonymous optimal codons in fundamental genes in order to ensure the effectiveness and accuracy of the translation of these particular genes [12,29]. In this regard, the same phenomenon could be intervening in a low number of *Echinococcus* genes, for which the correlation between tRNAs and optimal codon exists. All these hypotheses are supported, since a significant negative correlation was observed between ENc and CAI and between ENc and Fop and a positive correlation between CAI and Fop which also correlates with a higher GC content in the third codons position.

The average ENc values of *Echinococcus* genes (~57) indicate a random codon usage with no strong CUB. Correlation analysis between ENc and GC3s showed that mutational bias plays a minor role in choosing the codon usage. ENc analyses provide a method for quantifying CUB; however, this analysis alone is not enough to determine the exact contributions of natural selection and mutational pressure [90,91]. Therefore, neutrality analyses and PR2 were performed to measure the evolutionary forces that shape the codon usage in *Echinococcus*. According to these analyses, the selection pressure is the main evolutionary force shaping the codon usage in the three *Echinococcus* species with an 80% of contribution, whereas the remaining 20% is attributable to mutational bias. Hereby, the translational selection pressure or the contribution of an external selection pressure or a combination of both may play a crucial role in driving the CUB in the *Echinococcus* genes. This statement is also supported by the results of MRI and TrS2 analyses, which indicated that mutational pressure exerts a minor role, whereas translational selection pressure plays a relevant role in CUB.

In contrast to the CUB observed in *Echinococcus* genes, studies in mammals, yeasts and flies [22,66] have shown a wider range of ENc and GC3s values with a genes distribution whose CUB clearly depends on the nucleotide compositional bias [66,73,92]. Indeed, there is scarce evidence supporting that the variation in CUB has been shaped by selection in favour of translation efficiency [93]. Our studies performed on *E. canadensis* G7, *E. multilocularis* and *E. granulosus* G1 in addition to the analyses carried out on some species of Platyhelminths suggest that the low CUB is a general trait among cestodes. As here demonstrated for *Echinococcus* species, the codon usage analyses of the tapeworms *T. saginata* [54], *T. multiceps* [55] and *T. solium* [56] have also shown GC ending codons in highly expressed genes, similar mean values of ENc (~57) and similar genes distribution on the ENc-GC3s plots using almost randomly all the synonymous codons. However, a different contribution of selection pressure has been reported in some of these organisms. The percentage of selection pressure is lower in *T. solium* (~50%) [56] than in the *Echinococcus* species (~80%), in *T. saginata* [54] and in *T. multiceps* (~90%) [55]. Although this codon usage pattern may suppose that the codons are poorly optimized for the most of the genes of these parasites, this phenomenon could give them advantages to use almost any isoacceptor tRNA in any particular situation depending on the availability and tRNAs abundance. Indeed, the codon variability and low CUB could provide these parasitic organisms with an adaptive advantage to develop in the host environment under adverse conditions.

The COA allowed to disperse the genes along their axes according to their most outstanding characteristics; which were mainly the GC content and the expression level. In contrast, other factors such as the gene size, the aromaticity and the hydrophobicity seem to exert a minor influence on the codon usage. In relation to the gene length influence, those genes that encodes for proteins shorter than 350 amino acids (~38.5 KDa) exhibited a significant positive correlation between ENc, gene length and CAI values suggesting that shorter genes tend to exhibit higher bias and higher expression levels. In this regard, a selection phenomenon could intervene in favour of decreasing the length of highly expressed genes, explaining the negative correlation with high bias [21].

It is quite evident that the CUB is a complex phenomenon and may involve many factors. Many of them are difficult to address in *Echinococcus* species due to the complexity of its life cycle. However, the analyses performed here described for the first time the codon usage patterns in three *Echinococcus* species using whole genome and expression data. These analyses provide the basis for further research contributing to the identification of new genes, as well as to the development of molecular genetic engineering and evolutionary studies of the species of this genus. In addition, the identification of the *Echinococcus* optimal codons provide valuable information that must be considered when designing genetic engineering assays to express foreign genes. CUB and gene expression levels are generally attributed to selection in order to prioritize the efficiency of translation of genes that are critical to an organism's survival [9,12,40,94,95]. In this regard, *Echinococcus* genes with high CUB and high expression levels were associated with fundamental pathways in which the most abundance proteins were: protein kinases that may be associated with cellular cycle, proteins involved in the messenger RNA biogenesis, membrane trafficking proteins, DNA repair and recombination proteins, transporters, cytoskeleton proteins, exosome and ribosomal proteins. This group

of genes must be further explored to identify the essential genes, by applying in silico approach that could be helpful in searching potential therapeutic drug targets against the echinococcosis. Although in many species the natural selection associated with the accuracy and efficiency of the translation exerts a considerable influence on shaping the codon usage, the way the codons choice is carried out is not always so clear and therefore it is not possible to explain the overall codon usage variation. For this reason, it is common to observe that some genes exhibit a codon usage that is mainly determined by mutation and drift, and others exhibit a codon usage that results from the balance between mutational bias and selection pressure [94,96,97]. Some evolutionary studies have been carried out in order to explore the degeneration and the redundancy of the translation machinery of the genetic code in relation to the infection with parasites that make use of the host protein synthesis pathways. These studies have suggested that the features of the translation machinery could have emerged as adaptations of the parasite organisms in favour of reducing the energetic costs of the infection [98]. Therefore, the study of the *Echinococcus* codon usage could provide valuable information associated with the parasite-host relationship that would be helpful to determine which host's factors are relevant for shaping the parasite codon usage and favour the development, survival and parasite's infections.

5. Conclusions

In this study we analysed the codon usage of three *Echinococcus* species and determined the degree of contribution that mutational bias and natural selection pressure exert on the codon usage. Here we demonstrated that CUB is strongly influenced by natural selection and that the codon usage patterns are highly conserved in *Echinococcus* species but are quite different from the corresponding mammal host counterpart. The ENc values are negatively correlated with CAI indicating that highly expressed genes have a significant higher codon usage bias. The availability of high quality gene annotation and expression data allowed us to accurately determine 27 optimal codons, most of them ending in G/C and to update the codon usage of the *Echinococcus* species. The codon usage profile determined in this work will be useful for estimating the gene expression level of parasite and will provide new insights to perform data driven wet assays.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the Ministerio de Ciencia, Tecnología e Innovación Productiva BR/RED 1413 (L.K.), PIP 1122015 (M. R.), PIP 0100361 (L. K.); Ministerio de Educación Argentina y Brasil CAPG-BA 070/13 (L.K. and G.O.) and Agencia Nacional de Promoción Científica y Tecnológica SNCAD-C-AC-15 (L.K.).

Authors' contributions

L.M. performed the bioinformatics analysis and wrote the manuscript. L. M. and L. K. designed the study. L. M, L. K., M. R. and G. O. wrote and revised the manuscript. G. S. designed and implemented the SOM- analysis. All authors read and approved the manuscript.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in http://parasite.wormbase.org/Echinococcus_canadensis_prjeb8992/Info/Index and https://www.ebi.ac.uk/ena/data/view/PRJEB8992. All data generated or analysed during this study are included in this published article in supplementary material.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.molbiopara.2018.08.001.

References

[1] R. Hershberg, D.A. Petrov, Selection on codon Bias, Annu. Rev. Genet. 42 (2008) 287–299, https://doi.org/10.1146/annurev.genet.42.110807.091442.
[2] P.M. Sharp, L.R. Emery, K. Zeng, Forces that influence the evolution of codon bias, Philos. Trans. R. Soc. B Biol. Sci. 365 (2010) 1203–1212, https://doi.org/10.1098/rstb.2009.0305.
[3] J.B. Plotkin, G. Kudla, Synonymous but not the same: the causes and consequences of codon bias, Nat. Rev. Genet. 12 (2011) 32–42, https://doi.org/10.1038/nrg2899.
[4] R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pavé, Codon catalog usage and the genome hypothesis, Nucleic Acids Res. 8 (1980) 197, https://doi.org/10.1093/nar/8.1.197-c.
[5] C. Bermudez-Santana, C. Attolini, T. Kirsten, J. Engelhardt, S.J. Prohaska, S. Steigele, P.F. Stadler, Genomic organization of eukaryotic tRNAs, BMC Genom. 11 (270) (2010), https://doi.org/10.1186/1471-2164-11-270.
[6] E.P.C. Rocha, Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization, Genome Res. 14 (2004) 2279–2286, https://doi.org/10.1101/gr.2896904.
[7] R. Hershberg, D.A. Petrov, General rules for optimal codon choice, PLoS Genet. 5 (2009) e1000556, , https://doi.org/10.1371/journal.pgen.1000556.
[8] P.M. Sharp, G. Matassi, Codon usage and genome evolution, Curr. Opin. Genet. Dev. 4 (1994) 851–860 (Accessed September 11, 2017), http://www.ncbi.nlm.nih.gov/pubmed/7888755.
[9] S.K. Behura, D.W. Severson, Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes, Biol. Rev. 88 (2013) 49–61, https://doi.org/10.1111/j.1469-185X.2012.00242.x.
[10] T. Ikemura, Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system, J. Mol. Biol. 151 (1981) 389–409, https://doi.org/10.1016/0022-2836(81)90003-6.
[11] M. Bulmer, The selection-mutation-drift theory of synonymous codon usage, Genetics 129 (1991) 897–907, https://doi.org/10.1002/yea.320070702.
[12] L. Duret, D. Mouchiroud, Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 4482–4487, https://doi.org/10.1073/pnas.96.8.4482.
[13] C. Chen, D.A. Ridzon, A.J. Broomer, Z. Zhou, D.H. Lee, J.T. Nguyen, M. Barbisin, N.L. Xu, V.R. Mahuvakar, M.R. Andersen, K.Q. Lao, K.J. Livak, K.J. Guegler, Real-time quantification of microRNAs by stem-loop RT-PCR, Nucleic Acids Res. 33 (2005) e179, https://doi.org/10.1093/nar/gni178.
[14] G. Marais, D. Mouchiroud, L. Duret, Neutral effect of recombination on base composition in Drosophila, Genet. Res. 81 (2003) 79–87, https://doi.org/10.1017/S0016672302006079.
[15] N. Galtier, Gene conversion drives GC content evolution in mammalian histones, Trends Genet. 19 (2003) 65–68, https://doi.org/10.1016/S0168-9525(02)00002-1.
[16] T.M. Hambuch, J. Parsch, Patterns of synonymous codon usage in *Drosophila melanogaster* genes with sex-biased expression, Genetics 170 (2005) 1691–1700, https://doi.org/10.1534/genetics.104.038109.
[17] D.B. Carlini, Y. Chen, W. Stephan, The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr, Genetics 159 (2001) 623–633 (Accessed September 11, 2017), http://www.ncbi.nlm.nih.gov/pubmed/11606539.
[18] D.L. Hartl, E.N. Moriyama, S.A. Sawyer, Selection intensity for codon bias, Genetics 138 (1994) 227–234, https://doi.org/10.3168/jds.S0022-0302(75)84789-8.
[19] H. Akashi, Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift, Gene 205 (1997) 269–278, https://doi.org/10.1016/S0378-1119(97)00400-9.
[20] Y. Chen, D.B. Carlini, J.F. Baines, J. Parsch, J.M. Braverman, S. Tanda, W. Stephan, RNA secondary structure and compensatory evolution, Genes Genet. Syst. 74 (1999) 271–286, https://doi.org/10.1266/ggs.74.271.
[21] E.N. Moriyama, J.R. Powell, Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*, Nucleic Acids Res. 26 (1998) 3188–3193, https://doi.org/10.1093/nar/26.13.3188.
[22] J.R. Powell, E.N. Moriyama, Evolution of codon usage bias in Drosophila, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 7784–7790, https://doi.org/10.1073/pnas.94.15.7784.
[23] M. Oresic, M.H.H. Dehn, D. Korenblum, D. Shalloway, Tracing specific synonymous codon-secondary structure correlations through evolution, J. Mol. Evol. 56 (2003) 473–484, https://doi.org/10.1007/s00239-002-2418-x.
[24] Y. Prat, M. Fromer, N. Linial, M. Linial, Codon usage is associated with the evolutionary age of genes in metazoan genomes, BMC Evol. Biol. 9 (285) (2009), https://doi.org/10.1186/1471-2148-9-285.
[25] H. Goodarzi, N. Torabi, H.S. Najafabadi, M. Archetti, Amino acid and codon usage profiles: adaptive changes in the frequency of amino acids and codons, Gene 407 (2008) 30–41, https://doi.org/10.1016/j.gene.2007.09.020.
[26] E.A. Seward, S. Kelly, Dietary nitrogen alters codon bias and genome composition in

parasitic microorganisms, Genome Biol. 17 (226) (2016), https://doi.org/10.1186/s13059-016-1087-9.
[27] O.G. Berg, Selection intensity for codon bias and the effective population size of Escherichia coli, Genetics 142 (1996) 1379–1382 (Accessed September 11, 2017), http://www.ncbi.nlm.nih.gov/pubmed/8846914.
[28] P.C. Dedon, T.J. Begley, A system of RNA modifications and biased codon use controls cellular stress response at the level of translation, Chem. Res. Toxicol. 27 (2014) 330–337, https://doi.org/10.1021/tx400438d.
[29] H. Akashi, Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA, Genetics 139 (1995) 1067–1076 (accessed September 24, 2017), http://www.ncbi.nlm.nih.gov/pubmed/7713409.
[30] G.A.T. McVean, J. Vieira, Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila, Genetics 157 (2001) 245–257, https://doi.org/10.1371/journal.pgen.1005069.
[31] L. Duret, Evolution of synonymous codon usage in metazoans, Curr. Opin. Genet. Dev. 12 (2002) 640–649, https://doi.org/10.1016/S0959-437X(02)00353-2.
[32] J.F. Kane, Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli, Curr. Opin. Biotechnol. 6 (1995) 494–500, https://doi.org/10.1016/0958-1669(95)80082-4.
[33] C. Gustafsson, S. Govindarajan, J. Minshull, Codon bias and heterologous protein expression, Trends Biotechnol. 22 (2004) 346–353, https://doi.org/10.1016/j.tibtech.2004.04.006.
[34] Y. Zheng, W.-M. Zhao, H. Wang, Y.-B. Zhou, Y. Luan, M. Qi, Y.-Z. Cheng, W. Tang, J. Liu, H. Yu, X.-P. Yu, Y.-J. Fan, X. Yang, Codon usage bias in Chlamydia trachomatis and the effect of codon modification in the MOMP gene on immune responses to vaccination, Biochem. Cell Biol. 85 (2007) 218–226, https://doi.org/10.1139/o06-211.
[35] K. Lin, Y. Kuang, J.S. Joseph, P.R. Kolatkar, Conserved codon composition of ribosomal protein coding genes in *Escherichia coli*, *Mycobacterium tuberculosis* and *Saccharomyces cerevisiae*: lessons from supervised machine learning in functional genomics, Nucleic Acids Res. 30 (2002) 2599–2607 (Accessed January 30, 2018), http://www.ncbi.nlm.nih.gov/pubmed/12034849.
[36] S. Mueller, D. Papamichail, J.R. Coleman, S. Skiena, E. Wimmer, Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity, J. Virol. 80 (2006) 9687–9696, https://doi.org/10.1128/JVI.00738-06.
[37] J.R. Coleman, D. Papamichail, S. Skiena, B. Futcher, E. Wimmer, S. Mueller, Virus attenuation by genome-scale changes in codon pair Bias, Science (80-.) 320 (2008) 1784–1787, https://doi.org/10.1126/science.1155761.
[38] R.L.Y. Fan, S.A. Valkenburg, C.K.S. Wong, O.T.W. Li, J.M. Nicholls, R. Rabadan, J.S.M. Peiris, L.L.M. Poon, Generation of live attenuated influenza virus by using codon usage Bias, J. Virol. 89 (2015) 10762–10773, https://doi.org/10.1128/JVI.01443-15.
[39] S.K. Behura, M. Stanke, C.A. Desjardins, J.H. Werren, D.W. Severson, Comparative analysis of nuclear tRNA genes of *Nasonia vitripennis* and other arthropods, and relationships to codon usage bias, Insect Mol. Biol. 19 (2010) 49–58, https://doi.org/10.1111/j.1365-2583.2009.00933.x.
[40] H. Akashi, Translational selection and yeast proteome evolution, Genetics 164 (2003) 1291–1303 (Accessed September 11, 2017), http://www.ncbi.nlm.nih.gov/pubmed/12930740.
[41] A. Coghlan, K.H. Wolfe, Relationship of codon bias to mRNA and concentration protein length in *Saccharomyces cerevisiae*, Yeast 16 (2000) 1131–1145, https://doi.org/10.1002/1097-0061(20000915)16:12<1131::AID-YEA609>3.0.CO;2-F.
[42] H. Grosjean, W. Fiers, Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes, Gene 18 (1982) 199–209, https://doi.org/10.1016/0378-1119(82)90157-3.
[43] J.L. Bennetzen, B.D. Hall, Codon selection in yeast, J. Biol. Chem. 257 (1982) 3026–3031 (Accessed September 11, 2017), http://www.jbc.org/content/257/6/3026.full.pdf.
[44] M.A. Cucher, N. Macchiaroli, G. Baldi, F. Camicia, L. Prada, L. Maldonado, H.G. Avila, A. Fox, A. Gutiérrez, P. Negro, R. López, O. Jensen, M. Rosenzvit, L. Kamenetzky, Cystic echinococcosis in South America: systematic review of species and genotypes of *Echinococcus granulosus* sensu lato in humans and natural domestic hosts, Medline 21 (2016) 166–175, https://doi.org/10.1111/tmi.12647.
[45] C.M. Budke, A. Casulli, P. Kern, D.A. Vuitton, Cystic and alveolar echinococcosis: successes and continuing challenges, PLoS Negl. Trop. Dis. 11 (2017), https://doi.org/10.1371/journal.pntd.0005477.
[46] J. Ellis, D.A. Morrison, B. Kalinna, Comparison of the patterns of codon usage and bias between *Brugia*, *Echinococcus*, *Onchocerca* and *Schistosoma* species, Parasitol. Res. 81 (1995) 388–393, https://doi.org/10.1007/BF00931499.
[47] J.L. Milho, J.W. Tracy, Updated codon usage in schistosoma, Exp. Parasitol. 80 (1995) 353–356, https://doi.org/10.1006/expr.1995.1046.
[48] F. Alvarez, B. Garat, H. Musto, M. Picon, R. Ehrlich, Tendencies in *Echinococcus* sp. codon usage, Mem. Inst. Oswaldo Cruz 88 (1993) 345–346, https://doi.org/10.1590/S0074-02761993000200029.
[49] B.H. Kalinna, D.P. McManus, Codon usage in Echinococcus, Exp. Parasitol. 79 (1994) 72–76, https://doi.org/10.1006/expr.1994.1063.
[50] H. Musto, H. Romero, H. Rodríguez-Maseda, Heterogeneity in codon usage in the flatworm Schistosoma mansoni, J. Mol. Evol. 46 (1998) 159–167, https://doi.org/10.1007/PL00006291.
[51] V. Fernández, A. Zavala, H. Musto, Evidence for translational selection in codon usage in *Echinococcus* spp, Parasitology 123 (2001) 203–209, https://doi.org/10.1017/S0031182001008150.
[52] J.T. Ellis, D.A. Morrison, *Schistosoma mansoni*: patterns of codon usage and bias, Parasitology 110 (Pt 1) (1995) 53–60, https://doi.org/10.1017/S003118200008104X.
[53] G. Lamolle, A.V. Protasio, A. Iriarte, E. Jara, D. Simón, H. Musto, An isochore-like structure in the genome of the flatworm *Schistosoma mansoni*, Genome Biol. Evol. 8 (2016) 2312–2318, https://doi.org/10.1093/gbe/evw170.
[54] X. Yang, X. Luo, X. Cai, Analysis of codon usage pattern in *Taenia saginata* based on a transcriptome dataset, Parasit. Vectors 7 (2014) 527, https://doi.org/10.1186/s13071-014-0527-1.
[55] X. Huang, J. Xu, L. Chen, Y. Wang, X. Gu, X. Peng, G. Yang, Analysis of transcriptome data reveals multifactor constraint on codon usage in Taenia multiceps, BMC Genom. 18 (308) (2017), https://doi.org/10.1186/s12864-017-3704-8.
[56] X. Yang, X. Ma, X. Luo, H. Ling, X. Zhang, X. Cai, Codon usage bias and determining forces in *Taenia solium* genome, Korean J. Parasitol. 53 (2015) 689–697, https://doi.org/10.3347/kjp.2015.53.6.689.
[57] L.L. Maldonado, J. Assis, F.M.G. Araújo, A.C.M. Salim, N. Macchiaroli, M. Cucher, F. Camicia, A. Fox, M. Rosenzvit, G. Oliveira, L. Kamenetzky, The *Echinococcus canadensis*(G7) genome: a key knowledge of parasitic platyhelminth human diseases, BMC Genom. 18 (2017) 204, https://doi.org/10.1186/s12864-017-3574-0.
[58] I.J. Tsai, M. Zarowiecki, N. Holroyd, A. Garciarrubio, A. Sanchez-Flores, K.L. Brooks, A. Tracey, R.J. Bobes, G. Fragoso, E. Sciutto, M. Aslett, H. Beasley, H.M. Bennett, J. Cai, F. Camicia, R. Clark, M. Cucher, N. De Silva, T.A. Day, P. Deplazes, K. Estrada, C. Fernández, P.W.H. Holland, J. Hou, S. Hu, T. Huckvale, S.S. Hung, L. Kamenetzky, J.A. Keane, F. Kiss, U. Koziol, O. Lambert, K. Liu, X. Luo, Y. Luo, N. Macchiaroli, S. Nichol, J. Paps, J. Parkinson, N. Pouchkina-Stantcheva, N. Riddiford, M. Rosenzvit, G. Salinas, J.D. Wasmuth, M. Zamanian, Y. Zheng, X. Cai, X. Soberón, P.D. Olson, J.P. Laclette, K. Brehm, M. Berriman, The genomes of four tapeworm species reveal adaptations to parasitism, Nature 496 (2013) 57–63, https://doi.org/10.1038/nature12031.
[59] H. Zheng, W. Zhang, L. Zhang, Z. Zhang, J. Li, G. Lu, Y. Zhu, Y. Wang, Y. Huang, J. Liu, H. Kang, J. Chen, L. Wang, A. Chen, S. Yu, Z. Gao, L. Jin, W. Gu, Z. Wang, L. Zhao, B. Shi, H. Wen, R. Lin, M.K. Jones, B. Brejova, T. Vinar, G. Zhao, D.P. McManus, Z. Chen, Y. Zhou, S. Wang, The genome of the hydatid tapeworm *Echinococcus granulosus*, Nat. Genet. 45 (2013) 1168–1175, https://doi.org/10.1038/ng.2757.
[60] N. Sueoka, Directional mutation pressure and neutral molecular evolution, Proc. Natl. Acad. Sci. U. S. A. 85 (1988) 2653–2657, https://doi.org/10.1073/pnas.85.8.2653.
[61] P.M. Sharp, W.H. Li, The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Res. 15 (1987) 1281–1295, https://doi.org/10.1093/nar/15.3.1281.
[62] F. Wright, The "effective number of codons" used in a gene, Gene 87 (1990) 23–29, https://doi.org/10.1016/0378-1119(90)90491-9.
[63] S. Lee, S. Weon, S. Lee, C. Kang, Relative codon adaptation index, a sensitive measure of codon usage bias, Evol. Bioinform. 2010 (2010) 47–55, https://doi.org/10.4137/EBO.S4608.
[64] J.R. Lobry, C. Gautier, Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes, Nucleic Acids Res. 22 (1994) 3174–3180, https://doi.org/10.1093/nar/22.15.3174.
[65] J.A. Novembre, Accounting for background nucleotide composition when measuring codon usage Bias, Mol. Biol. Evol. 19 (2002) 1390–1394, https://doi.org/10.1093/oxfordjournals.molbev.a004201.
[66] J.F. Peden, Analysis of Codon Usage, (1999), https://doi.org/10.1016/j.biosystems.2011.06.005.
[67] N.R. McEwan, D. Gatherer, The mutational-response index and coton bias in genes from a Frankia nif operon, Theor. Appl. Genet. 96 (1998) 716–718, https://doi.org/10.1007/s001220050793.
[68] D. Gatherer, N.R. McEwan, Small regions of preferential codon usage and their effect on overall codon bias - the case of the plp gene, Biochem. Mol. Biol. 43 (1991) 107–114.
[69] M. Gouy, C. Gautier, Codon usage in bacteria: correlation with gene expressivity, Nucleic Acids Res. 10 (1982) 7055–7074, https://doi.org/10.1093/nar/10.22.7055.
[70] A. Uddin, S. Chakraborty, Codon usage pattern of genes involved in central nervous system, Mol. Neurobiol. 55 (1) (2018) 1–12, https://doi.org/10.1007/s12035-018-1173-y.
[71] A. Uddin, M.N. Choudhury, S. Chakraborty, Factors influencing codon usage of mitochondrial ND1 gene in pisces, aves and mammals, Mitochondrion 37 (2017) 17–26, https://doi.org/10.1016/j.mito.2017.06.004.
[72] N. Sueoka, Intrastrand parity rules of DNA base composition and usage biases of synonymous codons, J. Mol. Evol. 40 (1995) 318–325, https://doi.org/10.1007/BF00163236.
[73] N. Sueoka, Y. Kawanishi, DNA G + C content of the third codon position and codon usage biases of human genes, Gene 261 (2000) 53–62, https://doi.org/10.1016/S0378-1119(00)00480-7.
[74] N. Sueoka, Directional mutation pressure, selective constraints, and genetic equilibria, J. Mol. Evol. 34 (1992) 95–114, https://doi.org/10.1007/BF00182387.
[75] M.J. Greenacre, Theory and Applications of Correspondence Analysis, Academic Press, 1984 (Accessed May 27, 2018), https://books.google.com.ar/books/about/Theory_and_Applications_of_Correspondenc.html?id=LsPaAAAAMAAJ&redir_esc=y.
[76] H. Suzuki, R. Saito, M. Tomita, A problem in multivariate analysis of codon usage data and a possible solution, FEBS Lett. 579 (2005) 6499–6504, https://doi.org/10.1016/j.febslet.2005.10.032.
[77] L. Chen, T. Liu, D. Yang, X. Nong, Y. Xie, Y. Fu, X. Wu, X. Huang, X. Gu, S. Wang, X. Peng, G. Yang, Analysis of codon usage patterns in Taenia pisiformis through annotated transcriptome data, Biochem. Biophys. Res. Commun. 430 (2013)

1344–1348, https://doi.org/10.1016/j.bbrc.2012.12.078.
[78] G.A. Mazumder, A. Uddin, S. Chakraborty, Comparative analysis of codon usage pattern and its influencing factors in *Schistosoma japonicum* and *Ascaris suum*, Acta Parasitol. 62 (2017) 748–761, https://doi.org/10.1515/ap-2017-0090.
[79] M.C. Angellotti, S.B. Bhuiyan, G. Chen, X.F. Wan, CodonO: codon usage bias analysis within and across genomes, Nucleic Acids Res. 35 (2007) 132–136, https://doi.org/10.1093/nar/gkm392.
[80] X.F. Wan, J. Zhou, D. Xu, CodonO: a new informatics method for measuring synonymous codon usage bias within and across genomes, Int. J. Gen. Syst. 35 (2006) 109–125, https://doi.org/10.1080/03081070500502967.
[81] T.M. Lowe, S.R. Eddy, TRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, Nucleic Acids Res. 25 (1996) 955–964, https://doi.org/10.1093/nar/25.5.0955.
[82] D.H. Milone, G.S. Stegmayer, L. Kamenetzky, M. López, J.M. Lee, J.J. Giovannoni, F. Carrari, *omeSOM: a software for clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants, BMC Bioinform. 11 (438) (2010), https://doi.org/10.1186/1471-2105-11-438.
[83] F. Bação, V. Lobo, M. Painho, Self-organizing Maps As Substitutes for K-Means Clustering, Springer, Berlin, Heidelberg, 2005, pp. 476–483, https://doi.org/10.1007/11428862_65.
[84] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66 (1971) 846–850, https://doi.org/10.1080/01621459.1971.10482356.
[85] G. Stegmayer, D.H. Milone, L. Kamenetzky, M.G. Lopez, F. Carrari, A biologically inspired validity measure for comparison of clustering methods over metabolic data sets, IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (2012) 706–716, https://doi.org/10.1109/TCBB.2012.10.
[86] H. Akashi, R.M. Kliman, A. Eyre-Walker, Mutation pressure, natural selection, and the evolution of base composition in Drosophila, Genetica 102–103 (1998) 49–60, https://doi.org/10.1023/A:1017078607465.
[87] R.M. Kliman, J. Hey, The effects of mutation and natural-selection on codon bias in the genes of Drosophila, Genetics 137 (1994) 1049–1056 (Accessed September 11, 2017), http://www.genetics.org/content/137/4/1049.
[88] P.M. Sharp, G. Matassi, Codon usage and genome evolution, Curr. Opin. Genet. Dev. 4 (1994) 851–860, https://doi.org/10.1016/0959-437X(94)90070-1.
[89] T. Ikemura, Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer R, J. Mol. Biol. 158 (1982) 573–597, https://doi.org/10.1016/0022-2836(82)90250-9.
[90] A. Kawabe, N.T. Miyashita, Patterns of codon usage bias in three dicot and four monocot plant species, Genes Genet. Syst. 78 (2003) 343–352, https://doi.org/10.1266/ggs.78.343.
[91] X. Jia, S. Liu, H. Zheng, B. Li, Q. Qi, L. Wei, T. Zhao, J. He, J. Sun, Non-uniqueness of factors constraint on the codon usage in Bombyx mori, BMC Genom. 16 (2015) 356, https://doi.org/10.1186/s12864-015-1596-z.
[92] G. D'Onofrio, D. Mouchiroud, B. Aïssani, C. Gautier, G. Bernardi, Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins, J. Mol. Evol. 32 (1991) 504–510, https://doi.org/10.1007/BF02102652.
[93] P.M. Sharp, M. Stenico, J.F. Peden, A.T. Lloyd, Q.M. Centre, Codon usage: mutational bias, translational selection, or both? Biochem. Soc. Trans. 21 (1993) 835–841, https://doi.org/10.1042/bst0210835.
[94] P.M. Sharp, W.H. Li, An evolutionary perspective on synonymous codon usage in unicellular organisms, J. Mol. Evol. 24 (1986) 28–38, https://doi.org/10.1007/BF02099948.
[95] D.B. Carlini, W. Stephan, *In vivo* introduction of unpreferred synonymous codons into the *Drosophila* Adh gene results in reduced levels of ADH protein, Genetics 163 (2003) 239–243 (Accessed September 11, 2017), http://www.ncbi.nlm.nih.gov/pubmed/12586711.
[96] O.G. Berg, M. Martelius, Synonymous substitution-rate constants in *Escherichia coli* and *Salmonella typhimurium* and their relationship to gene expression and selection pressure, J. Mol. Evol. 41 (1995) 449–456, https://doi.org/10.1007/BF00160316.
[97] M. Bulmer, Are codon usage patterns in unicellular organisms determined by selection - mutation balance? J. Evol. Biol. 1 (1988) 15–26, https://doi.org/10.1046/j.1420-9101.1988.1010015.x.
[98] N.M. Krakauer, D.C. Jansen, A. Valleriani, Biological Evolution and Statistical Physics, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, https://doi.org/10.1007/3-540-45692-9.