Analysis of synonymous codon usage patterns in sixty-four different bivalve species

Marco Gerdol, Gianluca De Moro, Paola Venier, Alberto Pallavicini

Synonymous codon usage bias (CUB) is defined as the non-random usage of codons encoding the same amino acid across different genomes. This phenomenon is common to all organisms and the real weight of the many factors involved in its shaping still remains to be fully determined. So far relatively little attention has been put in the analysis of CUB in bivalve mollusks due to the limited genomic data available. Taking advantage of the massive sequence data generated from next generation sequencing projects, we explored codon preferences in 64 different species pertaining to the six major evolutionary lineages in Bivalvia. We detected remarkable differences across species, which are only partially dependent on phylogeny. While the intensity of CUB is mild in most organisms, a heterogeneous group of species (including Arcida and Mytilida, among the others) display higher bias and a strong preference for AT-ending codons. We show that the relative strength and direction of mutational bias, selection for translational efficiency and for translational accuracy contribute to the establishment of synonymous codon usage in bivalves. Although many aspects underlying bivalve CUB still remain obscure, we provide for the first time an overview of this phenomenon in this large, commercially and environmentally important, class of marine invertebrates.
Marco Gerdol¹, Gianluca De Moro¹, Paola Venier², Alberto Pallavicini¹*

¹Department of Life Sciences, University of Trieste, Via Giorgieri 5, 34127 Trieste (TS), Italy
²Department of Biology, University of Padova, Via U.Bassi 58/ B, 35121 Padova (PD), Italy

*Corresponding author: Alberto Pallavicini, pallavic@units.it, Department of Life Sciences, University of Trieste, Via Giorgieri 5, 34127 Trieste (TS), Italy

Authors e-mail addresses

Marco Gerdol: mgerdol@units.it
Gianluca De Moro: gdemoro@units.it
Paola Venier: paola.venier@unipd.it
Alberto Pallavicini: pallavic@units.it
Abbreviations

CAI: Codon Adaptation Index
CDS: coding sequence
CUB: Codon Usage Bias
ENC: Effective Number of Codons
GC$_3$: GC content at the third codon base
NGS: Next Generation Sequencing
ORF: Open Reading Frame
RNA-seq: RNA-sequencing
RSCU: Relative Synonymous Codon Usage
1. Introduction

Codon usage bias (CUB), intended as the non-random usage of synonymous codons in the protein translation process, can be observed in virtually all organisms. This phenomenon widely varies across different species and it is expected to significantly influence molecular genome evolution (Hershberg & Petrov, 2008; Sharp, Emery & Zeng, 2010; Plotkin & Kudla, 2011).

The mechanisms behind CUB are complex and not completely understood, since a large number of different intertwined biological factors are correlated with the choice of optimal codons. These include the GC content, both at gene and at whole genome level (Sueoka & Kawanishi, 2000; Zeeberg, 2002; Wan et al., 2004; Palidwor, Perkins & Xia, 2010), gene length, structure, expression levels and transcriptional efficiency (Gouy & Gautier, 1982; Sharp, Tuohy & Mosurski, 1986; Bains, 1987; Eyre-Walker, 1996; Duret & Mouchiroud, 1999), protein structure and amino acid composition (D’Onofrio et al., 1991; Xie et al., 1998), tRNA abundance (Ikemura, 1985), selection, mutational bias and random drift (Bulmer, 1991; Sharp et al., 1993; Kliman & Hey, 1994).

While CUB has been extensively studied in many viruses, prokaryotes, as well as in a number of eukaryote model species (Stenico, Lloyd & Sharp, 1994; Powell & Moriyama, 1997; Ermolaeva, 2001; Jenkins & Holmes, 2003; Gu et al., 2004; Mitreva et al., 2006; Vicario, Moriyama & Powell, 2007; Behura & Severson, 2012), so far little attention has been focused on non-model invertebrates. In particular, the large phylum of Mollusca has been almost completely neglected, even though it comprises more than 9,000 species including some of a great relevance as sea food and as sentinel organisms for coastal water biomonitoring (Gosling, 2003).

For a long time, genomic studies in Bivalvia have been limited by the lack of sequence data. However, the recent advances in the field of high throughput sequencing permitted to unravel the
genomes of *Crassostrea gigas* and *Pinctada fucata* (Takeuchi et al., 2012; Zhang et al., 2012) and to obtain a massive amount of transcriptomic data, useful for large-scale comparative studies (Suárez-Ulloa et al., 2013).

The only study performed so far on codon usage in Bivalvia was based on ESTs generated by Sanger sequencing and targeted a single species, the Pacific oyster *C. gigas* (Sauvage et al., 2007). Here, we provide the first comprehensive study of CUB in bivalves: based on the analysis of 2,846 evolutionarily conserved protein-coding genes in 64 different species, we calculated codon frequencies and Relative Synonymous Codon Usage (RSCU) values for each species, thus identifying both preferred and avoided codons, and calculating the overall CUB at the species transcriptome level.

Our data highlight significant differences among the analyzed species and clearly identify a bivalve subgroup with an increased codon bias, comprising Mytilida, Arcida and several different species of the Imparidentia lineage. We discuss the evolution of CUB in Bivalvia in relation with the possible underlying factors such as species phylogeny, mutational bias and natural selection. Overall, the results of these analysis bring new insights on the evolution of bivalve genomes and on the major forces driving the evolution of codon usage in bivalves and will provide a reference for improving the annotation of protein-coding genes in future bivalve genome sequencing efforts.

2. Materials and methods

2.1. Data sources
We considered two bivalve mollusk species with a fully sequenced genome (C. gigas and P. fucata) (Takeuchi et al., 2012; Zhang et al., 2012) and 62 other species whose transcriptome has been sequenced using next generation sequencing technologies and deposited in public sequence databases. When both 454 Life Sciences and Illumina-generated sequencing reads were available for a same species, the latter were chosen due to higher throughput and lower rate of sequencing errors. Namely, Illumina reads were used for species Anadara trapezia, Arctica islandica, Argopecten irradians, Astarte sulcata, Atrina rigida, Azumapecten farreri, Bathymodiolus platifrons, Cardites antiquata, Cerastoderma edule, Corbicula fluminea, Crassostrea angulata, Crassostrea cortesiensis, Crassostrea hongkongensis, Crassostrea virginica, Cycladicama cumingii, Cyrenoida floridana, Diplodonta sp. VG-2014, Donacilla cornea, Elliotio complanata, Ennucula tenuis, Eucrassatella cumingii, Galeomma turtoni, Glossus humanus, Hiatella arctica, Lampsilis cardium, Lamychaena hians, Mactra chinensis, Margaritifera margaritifera, Mercenaria campechiensis, Meretrix meretrix, Mizuhopecten yessoensis, Mya arenaria, Myochama anomoiodes, Mytilus californianus, Mytilus edulis, Mytilus galloprovincialis, Mytilus trossulus, Neotrigonia margaritacea, Ostrea chilensis, Ostrea edulis, Ostrea lurida, Ostreola stentina, Pecten maximus, Perna viridis, Pinctada martensii, Placopecten magellanicus, Polymesoda caroliniana, Pyganodon grandis, Ruditapes decussatus, Ruditapes philippinarum, Sinonovacula constricta, Solemya velum, Sphaerium nucleus, Uniomerus tetralasmus and Villosa lienosa (Qin et al., 2012; Ghiselli et al., 2012; Chen et al., 2013; Meng et al., 2013; Gerdol et al., 2014; Fu et al., 2014; Zhang et al., 2014; Pauletto et al., 2014; de Sousa et al., 2014; Zhao et al., 2014; Cornman et al., 2014; Zapata et al., 2014; Prentis & Pavrasovic, 2014; González et al., 2015). 454 Life Sciences sequences were used for Bathymodiolus azoricus, Geukensia demissa, Laternula elliptica, Mimachlamys nobilis, Pinctada maxima, Saccostrea glomerata, and
From Tegillarca granosa (Clark et al., 2010; Philipp et al., 2012; Egas et al., 2012; Jones et al., 2013; Fields et al., 2014). Details about the data used for the different species are provided in Table S1.

Sequence data were processed as follows: predicted CDS from the fully sequenced genomes of C. gigas (release 9) and P. fucata were retrieved from http://oysterdb.cn and http://marinegenomics.oist.jp/pinctada_fucata, respectively. De novo transcriptome assemblies were performed for all the other 62 bivalve species with the CLC Genomics Workbench (v.7.5, CLC Bio, Aarhus, Denmark) using the de novo assembly tool with “automatic word size” and “automatic bubble size” parameters selected, and setting the minimum allowed contig length to 300 bp.

In all transcriptomes, ORFs (Open Reading Frames) longer than 100 codons were predicted with TransDecoder (http://transdecoder.sourceforge.net). We selected the predicted CDS of C. gigas, and of one representative species for the Imparidentia (R. decussatus), Protobranchia (S. velum) and Palaeoheterodonta (P. grandis) lineage to identify a subset of evolutionarily conserved protein-coding genes with a 1:1 orthology ratio across Bivalvia. This was achieved by performing reciprocal tBLASTx searches (the e-value threshold was set a 1x10^{-10} and only hits displaying sequence identity > 50% were considered). This procedure resulted in a selection of 2,846 conserved protein-coding genes, whose orthologous sequences were retrieved in the remaining 60 species. Due to the heterogeneous tissue and developmental stage origin, the different sequencing platforms and depth applied, several of these evolutionarily sequences could not be identified or were fragmented in some transcriptomes. In order to ensure a minimum quality criteria, all the selected species had to display at least 25% of the sequences included in the dataset of evolutionarily conserved genes, with an average length > 500 nucleotides. A
number of additional transcriptomes derived from publicly available data did not meet such
criteria and were therefore not included in our analyses (**Table S2**).

### 2.2. Codon frequencies and codon usage statistics

The sets of evolutionarily conserved genes retrieved for each species were individually processed
with the *cusp* tool of the EMBOSS package (Rice, Longden & Bleasby, 2000) obtaining codon
frequencies and GC composition for each codon position. RSCU values for each individual
codon were calculated for each species as described by Sharp and colleagues (Sharp, Tuohy &
Mosurski, 1986). The effective number of codons (ENC) for each species was calculated
according to Wright (1990) using the EMBOSS *chips* tool, summing codons over all sequences
(Rice, Longden & Bleasby, 2000). The sENC-X values were determined for every amino acid
for each species and scaled to a range of values between 0 and 1 according to Moriyama and
Powell (1997). EMBOSS *chips* was also used to calculate ENC for individual genes whenever
necessary. We identified a reference set of 50 highly expressed genes for the calculation of
Codon Adaptation Index (CAI) based on the average expression in *C. gigas* digestive gland
(SRA: SRX093412), gills (SRA: SRX093414) and hemocytes (SRA: SRX093417) RNA-seq
libraries and their inclusion in the above mentioned set of 2,846 genes conserved across bivalves.

Gene expression was calculated as TPM (Transcripts Per Million) (Wagner, Kin & Lynch,
2012), with the RNA-seq mapping tool included in the CLC Genomics Workbench 8.5 (Aarhus,
Denmark), setting length and similarity fraction parameters to 0.75 and 0.98 and
insertion/deletion/mismatch penalties to 3. Orthologous genes were used for CAI calculation in
other species. CAI values were computed with CAI calculator 2 (Wu, Culley & Zhang, 2005).

The gene expression levels of *M. galloprovincialis* transcripts were calculated using the digestive
gland (SRA: SRX126945-8), gills (SRA: SRX389466) and hemocytes (SRA: SRX389338) RNA-seq libraries (Gerdol et al., 2014; Moreira et al., 2015).

Scatter plots were generated between ENC and the average GC content calculated at the third codon position (GC₃) for each species, between ENC and sENCx and between ENC and CAI; Pearson correlation coefficients and linear regression analyses were computed with R 3.1.0 (http://www.r-project.org).

2.3. Hierarchical clustering analysis

The Relative Synonymous Codon Usage (RSCU) values calculated for the 59 informative codons in each species were used to build a tabular file. STOP, ATG (encoding Met) and TGG (encoding Trp) codons were excluded from this analysis. This file was used as an input for Cluster 3 (de Hoon et al., 2004), thus generating a species distance matrix. Hierarchical clustering was performed by using the Euclidean distance as a similarity metric and complete linkage as a clustering method.
3. Results and discussion

3.1. CUB varies across bivalve species

The codon frequencies and RSCU values calculated for the 64 bivalve species analyzed are displayed in Tables S3 and S4. As shown in the comparative overview of Fig. 1, RSCU values are usually very similar in closely related species, such as in the case of Mytilida, Ostreoidea and Unionida (panels A, B and C), but marked differences can be observed in an higher-order comparison (panel D).

Different species clearly show a different tendency to the preferential usage of specific codons, as exemplified by the average Effective Number of Codons (ENC) (Wright, 1990) values in Table 1. Overall, the observed ENC values range between 40.65 (in the Chinese surf clam M. chinensis) and 56.80 (G. turtoni) across the analyzed species, while the theoretical value is comprised between 21 (if only a single codon is used for each amino acid) and 61 (if all codons are used with equal frequency). Most bivalve species display a weak CUB, using on average over 50 out of the 61 available codons, and only a limited number of bivalve species display an ENC value comparable to that of other invertebrates (the ENC range is 45-48 in Drosophila and nematodes) (Powell & Moriyama, 1997; Mitreva et al., 2006; Vicario, Moriyama & Powell, 2007).

3.2. Species clustering based on CUB does not reflect the evolutionary history of Bivalvia

We computed RSCU values for the 59 informative codons of each species to perform a hierarchical clustering of species with Cluster 3, thereby investigating the role of CUB in the
evolution of bivalve genomes. The resulting dendrogram is shown in Fig. 2. Although phylogeny and CUB-based clustering are in agreement, in several cases, up to the order level, the six major lineages of Anomalodesmata, Archiheterodonta, Imparidentia, Palaeoheterodonta, Protobranchia and Pteriomorphia expected from molecular phylogeny (Bieler et al., 2014) are hardly distinguishable. This observation is consistent with data previously reported for nematodes by Mitreva and colleagues (Mitreva et al., 2006), who reported a connection between codon distribution and phylogeny only for closely related species, up to the genus level. In bivalves such a relationship seems to extend a bit further, in some cases up to one of the six major evolutionary lineages, as for instance in Palaeoheterodonta, which include the freshwater mussels of the order Unionida and the saltwater clams of the order Trigoniida, or Archiheterodonta, which only comprise four relatively small extant families. While in some cases (e.g. Mytilida and Ostreoidea) all the species maintain a similar usage of synonymous codons (Fig. 1), in others (e.g. Venerida) remarkable differences among species are clearly visible.

In essence, the clustering based on codon usage divides the bivalve species into two largely divergent groups:

(I) The first group is very heterogeneous, comprising 43 species with ENC > 52 (with the exception of L. hians). Two subgroups are detectable: group I-a comprises all the Pectinida and Ostreoidea species (Pteriomorphia), two Anomalodesmata (M. anomides and L. elliptica), two Protobranchia (E. tenuis and S. velum) and the Imparidentia G. turtoni, M. arenaria and S. constricta. The species pertaining to this subgroup show a weak CUB, with ENC 54-58 and GC₃ very close to 50% (averaging ~49%).
The subgroup I-b comprises Unionida and Trigoniida (Palaeoheterodonta), *Pinctada* spp. (Pteriomorphia, Pterioidea), eight unrelated Imparidentia and the three Archiheterodonta species. Compared to subgroup I-a, the observed CUB is slightly higher, including species with ENC ranging from ~52 to ~54, but GC$_3$ is remarkably lower than group Ia, averaging ~43%.

(II) The second major group comprises the remaining 20 species, including the Pteriomorphia groups Mytilida, Arcida and Pinnoidea (represented by the lone species *A. rigida*), together with various Imparidentia species classified as Cardioidea, Cyamiidae, Mactroidea, Sphaeriidae and Venerida, implying that this well-defined group of bivalves with high CUB (ENC ~40-52) and low GC$_3$ (30-40%, with the outlier *M. chinensis* reaching ~23%) comprises phylogenetically distant species.

3.3. **Codon usage is biased towards A/T ending codons in Bivalvia**

Here we report the most commonly used codon(s) for each amino acid, designing as “preferred codons” all the codons used more frequently than expected (RSCU > 1), and we investigate the correlation between their frequency and overall codon usage bias (negative ENC) across the bivalve species analyzed (Vicario, Moriyama & Powell, 2007).

**Figure 3, panel A** summarizes the number of bivalve species where a given codon is preferred, evidencing that, despite the difference in overall ENC values, several preferred and avoided codons are shared by most, if not by all, bivalves: this is the case, for example, of ACA (Thr), AGA and AGT (Arg), TCA (Ser), TGT (Cys), CCA (Pro), GCT (Ala) and GGA (Gly). On the other hand, the RSCU values of ACG (Thr), AGC (Ser), TCG and CTA (Leu), TGC (Cys), CCG
(Pro), CGC and CGG (Arg), GCG (Val) and GGG (Gly) are always lower than 1, indicating that these codons are avoided in all species. In general, A/T-ending codons appear to be preferred over those ending in G/T, but some notable exceptions exist, including the two C-starting codons encoding the six-fold degenerate amino acids Ser and Arg. While G-ending codons are not uncommon, C-ending codons are almost invariably avoided.

However, when the correlation between codon frequencies and overall CUB of bivalve species is taken into account, the important weight of A/T ending codons on bivalve codon bias becomes evident, (Fig. 3, panel B). Indeed, the high CUB of all the species pertaining to clustering group II (Fig. 2) appears to be mostly resulting from an increased use of A/T-ending codons over those ending in G/C, also explaining the significant negative correlation between ENC and GC across species (see Section 3.4 below). This correlation is significant for most codons with, once again, the exception of the C-starting codons encoding the six-fold degenerate amino acids Leu and Arg. On the contrary, the frequency of G/C ending codons is significantly and negatively correlated with CUB in all cases, with the exception of TTG (Leu).

We explored in detail the contribution of different amino acids to CUB in bivalves by calculating s-ENCx values for each amino acid and correlating this parameter to the overall CUB of each species. This measure is a variation of ENC (Wright, 1990) which is scaled in a range from 0-1 for each amino acid independently from the level of redundancy, and which can be used to estimate the relative intensity of CUB across the 18 degenerate amino acids (Moriyama & Powell, 1997). As reported in Table S5 and Fig. 4, the sENC-x values of all amino acids negatively correlate with ENC with significant p-values, including those with relatively low average s-ENCx values. The only exception is represented by of Gln, which is likely related to the fact that it is the only two-fold degenerate amino acid to display a strong preference for a
G/C-ending (CAG) over an A/T-ending codon (CAA) (see Fig. 3, panel A). Overall, Arg is
certainly the amino acid which accounts for the greatest CUB in bivalves, as highlighted by the
high average s-ENCx value (0.32), followed by Pro, Thr, Cys, Ala, Gly, Ser and Leu, all
characterized by values > 0.1.

3.4. Effects of mutational bias and selection on CUB in Bivalvia

Although the knowledge of the factors underlying the preferential use of certain codons has
remarkably increased over the past few decades, the real contribution of the several potentially
contrasting forces involved in shaping CUB still remains a matter of debate in the scientific
community.

Mutational bias, intended as the non-randomness of mutational patterns, certainly has a major
role in determining codon usage in prokariotes, as well as in many eukaryotes (Muto & Osawa,
1987; Sémon, Lobry & Duret, 2006). A critical parameter tightly linked to mutational bias is
genomic GC content, which in turn strongly influences the coding CG content (Knight, Freeland
& Landweber, 2001) and GC3 (Mitreva et al., 2006). Actually, all amino acids, with the
exception of the non-degenerate methionine and tryptophan, tolerate A/T to C/G changes in the
third codon position and thus, in organisms where mutational bias is dominant over other factors,
the occurrence of G/C-ending codons should follow the total genomic GC content. However,
even in such cases some deviations are observed, in particular for the 6-fold degenerated Arg and
Leu, which also tolerate synonymous mutations at the first codon position (Palidwor, Perkins &
Xia, 2010).

In bivalves, average GC3 appears to be tightly related to ENC (Fig. 5), as these two variables
could be strongly correlated by linear regression with $R^2 = 0.92$ (p-value = 2.86x10⁻³⁶). Overall,
while GC$_3$ can assume values as low as ~23 in *M. chinensis*, GC$_3$ and AT$_3$ contents become even in species displaying weak CUB (and high ENC). Based on these observations and the well-documented correlation between genomic GC content and coding GC$_3$, one could hypothesize that the bivalves pertaining to the clustering group II possess a similar genomic GC content, lower than all the bivalve species included in the heterogeneous clustering group I, thus explaining their common placement in a well-distinct cluster regardless of the reported phylogenetic distance (Fig. 2). On the contrary, the bivalve transcriptomes of the clustering group Ia show a GC$_3$ not far from 50%, accompanied by ENC values higher than 54 (Fig. 5), which would suggest that the absence of a strong CUB is linked to a lack of mutational bias. However, the genomic data currently available for bivalves do not support this view, since the differences in genomic GC content among species are minimal. The fully sequenced genomes of *P. fucata* and *C. gigas* possess a relatively low G/C composition (33.69% and 32.33%, respectively) (Takeuchi et al., 2012; Zhang et al., 2014) and partial data from *A. farreri* and *M. galloprovincialis* indicate a similar GC content for these two species (35.75% and 31.65%, respectively) (Zhao et al., 2012; Nguyen, Hayes & Ingram, 2014). Therefore, while the A/T-rich nature of most bivalve genomes is consistent with a role of mutational bias towards the preferential choice of A/T-ending codons in bivalves (Section 3.3), other factors might be taken into account to explain the differences in CUB among these marine organisms. Among these, selection for translational speed and accuracy are certainly among the most relevant, as they can potentially overcome and mask the effects of mutational bias on CUB. This is the case, for example, of *Drosophila* spp. and of *Strongylocentrotus purpuratus*, which preferentially use G/C ending codons despite having A/T rich genomes (Vicario, Moriyama & Powell, 2007; Kober & Pogson, 2013). The selection for translational speed is evident in many
unicellular and multicellular organisms, where the preferential use of optimal codons by highly
expressed genes, with the aim to maximize the rate of elongation during protein synthesis
(Marais & Duret, 2001), has been clearly demonstrated (Gouy & Gautier, 1982; Powell &
Moriyama, 1997; Duret, 2000). Nevertheless, the correlation between gene expression and CUB
is not universally applicable, as the selection for translational speed appears to be weak in some
species (Hiraoka et al., 2009).

Natural selection also potentially shapes codon usage to improve translational accuracy, reducing
the risk of missense errors during the translational process (Akashi, 1994; Eyre-Walker, 1996).
Even though this model of selection predicts a higher CUB in genes encoding longer proteins, a
positive correlation between CUB and protein length has only been observed in bacteria, while
on the opposite an unexpected negative correlation has been described in a number of
eukaryotes, including S. cerevisiae, D. melanogaster and C. elegans (Moriyama & Powell, 1998;
Duret & Mouchiroud, 1999). Even though differences in selective constraints between
prokaryotes and eukaryotes have been evoked to explain this contradiction, the relationship
between CUB and protein length still remains obscure (Marais & Duret, 2001).

The investigation of these two selective forces in bivalves is complicated by the
incomplete/fragmented nature of de novo assembled transcriptomes and by the limited gene
expression resources available. We correlated CUB with gene expression (in three tissues:
digestive gland, gills and hemocytes) and with ORF (protein) length in C. gigas, the most
appropriate bivalve species for this purpose due to the completeness of its genome annotation
and availability of gene expression data. We observed a significant positive correlation between
CUB (negative ENC) and gene expression in the three tissues analyzed, as well as between CUB
and ORF (protein) length (Table 2), which would suggest that both selection for translational
speed and accuracy are actively shaping codon usage in oyster. However, we also observed a
highly significant, negative correlation between GC3 and gene expression and between GC3 and
protein length, which seem to contradict the mutational bias given by the A/T-rich nature of the
oyster genome. This observation matches the results obtained in a previous work conducted with
limited expression data based on Sanger EST sequencing, which suggested that translational
selection acts as a contrasting force to mutational bias in oyster, effectively counteracting its
action in highly expressed genes (Sauvage et al., 2007). Our data further indicate that, besides
the selection for translational speed, also the selection for translational accuracy provides a
contribution to the selection of G/C-ending codons in oyster.

The contrasting action of mutational bias and selection becomes particularly evident while taking
into consideration the correlation between CAI, a directional measure of CUB which is based on
a reference set of highly expressed genes (Sharp & Li, 1987), and ENC, which on the other hand
is a non-directional measure which does not permit to appreciate the contribution of opposite
forces (in this case mutational bias towards A/T-ending codons and selection towards G/C-
ending codons). Indeed, in oyster and in all the other species clusterized in group I (see Fig. 2),
the scatter in the correlation plot between CAI and ENC appears to be quite relevant (Fig. 6;
Paerson correlation coefficients are -0.44 for C. gigas and -0.38 for P. magellanicus). This
indicates that the mutational bias towards A/T is counterbalanced by natural selection in favor of
G/C-ending in a relevant number of genes. However, the significant correlation between ENC
and GC3 at the whole protein-coding transcriptome level (Paerson correlation = 0.16, p-value =
5.42x10^{-148}) indicates that the weight of A/T mutational bias is still dominating over that of G/C
selection in most oyster genes.
Overall, it is likely that in all species pertaining to clustering group I, whose coding GC$_3$ sensibly deviates from the frequency expected by genomic GC content, the mutational bias towards A/T-ending codons is countered by the opposite forces of selection for translational speed and accuracy, leading to moderate/low CUB.

On the other hand, the correlation between CAI and ENC in the bivalve species pertaining to clustering group II is highly significant (Fig. 6, Paerson correlation coefficients are -0.87 for *M. chinensis* and -0.71 for *M. californianus*). To better interpret this result, we extended the analysis performed in oyster to a representative species of this group, *M. galloprovincialis*, limiting our calculations to full-length protein-coding transcripts (Table 2). Overall, like in oyster, the significant positive correlation between gene expression and GC$_3$ over three different tissues indicates that selection for translational speed acts towards G/C-ending codons. However, unlike oyster, this appears to have no effect on CUB on the whole transcriptome scale, as the correlation between gene expression and ENC was not significant. In addition, we observed that, in contrast with oyster, the selection for translational accuracy is biased towards A/T-ending codons (i.e. GC$_3$ and protein length are negatively correlated). Overall, the *M. galloprovincialis* average GC$_3$ is 30.80%, a value which is very close to the GC genomic content, further indicating that mutational bias and selection for translational accuracy, which both act in favor of A/T-ending, are largely dominating over the opposite selection for translational speed in mussel.

4. Conclusion

In this paper we presented a comprehensive and well supported overview of CUB in 64 different bivalve species, including both marine and freshwater species, based on the analysis of nearly 3,000 evolutionarily conserved genes. To the best of our knowledge, this is the first time that
CUB is systematically investigated in such a large and important class of invertebrates, the largest one after Insecta. We show that bivalves can be divided into two distinct groups, based on the intensity of the bias towards the use of optimal codons. While in many species CUB appears to be relatively weak, Mytilida, Arcida and several Imparidentia species show a remarkable preference for A/T-ending codons.

Given the poor correlation between CUB and bivalve taxonomy, other factors are expected to drive the evolution of CUB towards the same direction in distantly related species. We investigated this issue in *C. gigas* and *M. galloprovincialis*, two species displaying low and high CUB, respectively. Our analyses pointed out that:

(i) Bivalves are subject to mutational bias towards A/T-ending codons, due to the low GC content of their genomes. Although the nucleotide genome composition should be reflected by a low GC$_3$, some species display an almost even nucleotide content at the third codon position, which could be explained by the contrasting action exerted by other factors.

(ii) Gene expression is positively correlated with GC$_3$, which indicates a selection towards G/C-ending codons, acting to maximize the rate of elongation of protein synthesis for highly expressed transcripts. The different intensity of this selective force, opposed to mutational bias, might partly explain differences in CUB among species with similar genomic GC content.

(iii) Although the selection for translational accuracy, evidenced by the correlation between protein length and CUB, is also active in bivalves, the direction of this selective force varies across species, either reinforcing (in *M. galloprovincialis*) or counteracting (in *C. gigas*) mutational bias.
In conclusion, multiple factors contribute to CUB in bivalves, and the specific weight of each of them is still difficult to be determined with certainty considering the limited genomic resources available for most species. However, based on the data collected so far, the different intensity of the opposing forces represented by mutational bias towards A/T-ending codons and by selection for translational speed towards G/C-ending codons appear to be two major players in this process. Further study will be certainly needed to ascertain whether this model can be extended to all bivalve species once additional genomic resources will become available. Such analyses could also benefit from the integration of additional data such as the abundance of isoaccepting tRNAs and data concerning effective population size of each species.

This large scale analysis supports the progressive understanding of molecular genome evolution in bivalves and it is potentially useful for many different applications. For example, as we have previously explained, codon usage is known to widely vary across genes based on their expression level. The calculation of CAI could be used to predict the expression level and the transcriptional efficiency of unknown genes and to assess the adaptation of viral genes to their bivalve hosts (Sharp & Li, 1987). Codon bias also has profound implications on codon-based phylogenetic reconstructions, which could be optimized to avoid the over-estimation of divergence between distantly related species and the under-estimation of divergence between closely related ones (Christianson, 2005).

But, above all, one of the key applications of codon usage in the post-genomic era is certainly the annotation of newly sequenced genomes. As the raw genomic sequences of new organisms become available, new tools are required to efficiently identify genes and to predict their structure and boundaries, in particular in non-model invertebrate species, which show a very high proportion of genes without similarity to sequences deposited in public databases. The
oyster genome has already evidenced that homology-based annotation methods perform rather poorly for certain gene families, highlighting the need for integrating additional parameters in gene prediction algorithms to optimize annotation (Gerdol, Venier & Pallavicini, 2014). A number of algorithms which take into account species-specific CUB information have been developed, trying to identify protein coding sequences based on their congruence with a reference codon usage table (Gribskov, Devereux & Burgess, 1984). This task is particularly important and complicated in Bivalvia, since genomic analyses in this taxa are quickly expanding and only two species present a completely sequenced and annotated genome (Suárez-Ulloa et al., 2013).
References

Akashi H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* 136:927–935.

Bains W. 1987. Codon distribution in vertebrate genes may be used to predict gene length. *Journal of Molecular Biology* 197:379–388.

Behura SK., Severson DW. 2012. Comparative analysis of Codon usage bias and Codon context patterns between dipteran and hymenopteran sequenced genomes. *PLoS ONE* 7.

Bieler R., Mikkelsen PM., Collins TM., Glover EA., González VL., Graf DL., Harper EM., Healy J., Kawauchi GY., Sharma PP., Staubach S., Strong EE., Taylor JD., Tēmkin I., Zardus JD., Clark S., Guzmán A., McIntyre E., Sharp P., Giribet G. 2014. Investigating the Bivalve Tree of Life – an exemplar-based approach combining molecular and novel morphological characters. *Invertebrate Systematics* 28:32–115.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897–907.

Chen H., Zha J., Liang X., Bu J., Wang M., Wang Z. 2013. Sequencing and de novo assembly of the Asian clam (Corbicula fluminea) transcriptome using the Illumina GAIIx method. *PloS One* 8:e79516.

Christianson ML. 2005. Codon usage patterns distort phylogenies from or of DNA sequences. *American Journal of Botany* 92:1221–1233.

Clark MS., Thorne MA., Vieira FA., Cardoso JC., Power DM., Peck LS. 2010. Insights into shell deposition in the Antarctic bivalve Laternula elliptica: gene discovery in the mantle transcriptome using 454 pyrosequencing. *BMC Genomics* 11:362.

Cornman RS., Robertson LS., Galbraith H., Blakeslee C. 2014. Transcriptomic Analysis of the Mussel Elliptio complanata Identifies Candidate Stress-Response Genes and an Abundance of Novel or Noncoding Transcripts. *PLoS ONE* 9:e112420.
D’Onofrio G., Mouchiroud D., Aïssani B., Gautier C., Bernardi G. 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. Journal of Molecular Evolution 32:504–510.

Duret L. 2000. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends in genetics: TIG 16:287–289.

Duret L., Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proceedings of the National Academy of Sciences 96:4482–4487.

Egas C., Pinheiro M., Gomes P., Barroso C., Bettencourt R. 2012. The Transcriptome of Bathymodiolus azoricus Gill Reveals Expression of Genes from Endosymbionts and Free-Living Deep-Sea Bacteria. Marine Drugs 10:1765–1783.

Ermolaeva MD. 2001. Synonymous codon usage in bacteria. Current Issues in Molecular Biology 3:91–97.

Eyre-Walker A. 1996. Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy? Molecular Biology and Evolution 13:864–872.

Fields PA., Eurich C., Gao WL., Cela B. 2014. Changes in protein expression in the salt marsh mussel Geukensia demissa: evidence for a shift from anaerobic to aerobic metabolism during prolonged aerial exposure. The Journal of Experimental Biology 217:1601–1612.

Fu X., Sun Y., Wang J., Xing Q., Zou J., Li R., Wang Z., Wang S., Hu X., Zhang L., Bao Z. 2014. Sequencing-based gene network analysis provides a core set of gene resource for understanding thermal adaptation in Zhikong scallop Chlamys farreri. Molecular Ecology Resources 14:184–198.

Gerdol M., De Moro G., Manfrin C., Milandri A., Riccardi E., Beran A., Venier P., Pallavicini A. 2014. RNA sequencing and de novo assembly of the digestive gland transcriptome in Mytilus galloprovincialis fed with toxigenic and non-toxic strains of Alexandrium minutum. BMC research notes 7:722.
Gerdol M., Venier P., Pallavicini A. 2014. The genome of the Pacific oyster Crassostrea gigas brings new insights on the massive expansion of the C1q gene family in Bivalvia. Developmental and comparative immunology accepted: in press.

Ghiselli F., Milani L., Chang PL., Hedgecock D., Davis JP., Nuzhdin SV., Passamonti M. 2012. De Novo assembly of the Manila clam Ruditapes philippinarum transcriptome provides new insights into expression bias, mitochondrial doubly uniparental inheritance and sex determination. Molecular Biology and Evolution 29:771–786.

González VL., Andrade SCS., Bieler R., Collins TM., Dunn CW., Mikkelsen PM., Taylor JD., Giribet G. 2015. A phylogenetic backbone for Bivalvia: an RNA-seq approach. Proceedings of the Royal Society B: Biological Sciences 282.

Gosling EM. 2003. Bivalve Molluscs: Biology, Ecology and Culture. Oxford, UK.

Gouy M., Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Research 10:7055–7074.

Gribskov M., Devereux J., Burgess RR. 1984. The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. Nucleic Acids Research 12:539–549.

Gu W., Zhou T., Ma J., Sun X., Lu Z. 2004. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Research 101:155–161.

Hershberg R., Petrov DA. 2008. Selection on codon bias. Annual Review of Genetics 42:287–299.

Hiraoka Y., Kawamata K., Haraguchi T., Chikashige Y. 2009. Codon usage bias is correlated with gene expression levels in the fission yeast Schizosaccharomyces pombe. Genes to Cells: Devoted to Molecular & Cellular Mechanisms 14:499–509.

de Hoon MJL., Imoto S., Nolan J., Miyano S. 2004. Open source clustering software. Bioinformatics (Oxford, England) 20:1453–1454.
Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* 2:13–34.

Jenkins GM., Holmes EC. 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Research* 92:1–7.

Jones DB., Jerry DR., Forêt S., Konovalov DA., Zenger KR. 2013. Genome-wide SNP validation and mantle tissue transcriptome analysis in the silver-lipped pearl oyster, Pinctada maxima. *Marine Biotechnology (New York, N.Y.)* 15:647–658.

Kliman RM., Hey J. 1994. The effects of mutation and natural selection on codon bias in the genes of Drosophila. *Genetics* 137:1049–1056.

Knight RD., Freeland SJ., Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology* 2:research0010.

Kober KM., Pogson GH. 2013. Genome-wide patterns of codon bias are shaped by natural selection in the purple sea urchin, Strongylocentrotus purpuratus. *G3 (Bethesda, Md.)* 3:1069–1083.

Marais G., Duret L. 2001. Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. *Journal of Molecular Evolution* 52:275–280.

Meng X., Liu M., Jiang K., Wang B., Tian X., Sun S., Luo Z., Qiu C., Wang L. 2013. De Novo Characterization of Japanese Scallop Mizuhopecten yessoensis Transcriptome and Analysis of Its Gene Expression following Cadmium Exposure. *PLoS ONE* 8.

Mitreva M., Wendl MC., Martin J., Wylie T., Yin Y., Larson A., Parkinson J., Waterston RH., McCarter JP. 2006. Codon usage patterns in Nematoda: analysis based on over 25 million codons in thirty-two species. *Genome Biology* 7:R75.
Moreira R., Pereiro P., Canchaya C., Posada D., Figueras A., Novoa B. 2015. RNA-Seq in Mytilus galloprovincialis: comparative transcriptomics and expression profiles among different tissues. *BMC Genomics* 16:728.

Moriyama EN., Powell JR. 1997. Codon usage bias and tRNA abundance in Drosophila. *Journal of Molecular Evolution* 45:514–523.

Moriyama EN., Powell JR. 1998. Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. *Nucleic Acids Research* 26:3188–3193.

Muto A., Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proceedings of the National Academy of Sciences* 84:166–169.

Nguyen TTT., Hayes BJ., Ingram BA. 2014. Genetic parameters and response to selection in blue mussel (Mytilus galloprovincialis) using a SNP-based pedigree. *Aquaculture* 420–421:295–301.

Palidwor GA., Perkins TJ., Xia X. 2010. A general model of codon bias due to GC mutational bias. *PloS One* 5:e13431.

Pauletto M., Milan M., Moreira R., Novoa B., Figueras A., Babucci M., Patarnello T., Bargelloni L. 2014. Deep transcriptome sequencing of Pecten maximus hemocytes: a genomic resource for bivalve immunology. *Fish & Shellfish Immunology* 37:154–165.

Philipp EER., Wessels W., Gruber H., Strahl J., Wagner AE., Ernst IMA., Rimbach G., Kraemer L., Schreiber S., Abele D., Rosenstiel P. 2012. Gene expression and physiological changes of different populations of the long-lived bivalve Arctica islandica under low oxygen conditions. *PloS One* 7:e44621.

Plotkin JB., Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics* 12:32–42.

Powell JR., Moriyama EN. 1997. Evolution of codon usage bias in Drosophila. *Proceedings of the National Academy of Sciences* 94:7784–7790.
Prentis PJ., Pavasovic A. 2014. The Anadara trapezia transcriptome: A resource for molluscan physiological genomics. *Marine Genomics* 18, Part B:113–115.

Qin J., Huang Z., Chen J., Zou Q., You W., Ke C. 2012. Sequencing and de novo analysis of Crassostrea angulata (Fujian oyster) from 8 different developing phases using 454 GSFlx. *PloS One* 7:e43653.

Rice P., Longden I., Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics: TIG* 16:276–277.

Sauvage C., Bierne N., Lapègue S., Boudry P. 2007. Single Nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster Crassostrea gigas. *Gene* 406:13–22.

Sémon M., Lobry JR., Duret L. 2006. No Evidence for Tissue-Specific Adaptation of Synonymous Codon Usage in Humans. *Molecular Biology and Evolution* 23:523–529.

Sharp PM., Stenico M., Peden JF., Lloyd AT. 1993. Codon usage: mutational bias, translational selection, or both? *Biochemical Society Transactions* 21:835–841.

Sharp PM., Emery LR., Zeng K. 2010. Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365:1203–1212.

Sharp PM., Li WH. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* 15:1281–1295.

Sharp PM., Tuohy TM., Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research* 14:5125–5143.

de Sousa JT., Milan M., Bargelloni L., Pauletto M., Matias D., Joaquim S., Matias AM., Quillien V., Leitão A., Huvet A. 2014. A Microarray-Based Analysis of Gametogenesis in Two Portuguese Populations of the European Clam Ruditapes decussatus. *PloS ONE* 9.

Stenico M., Lloyd AT., Sharp PM. 1994. Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. *Nucleic Acids Research* 22:2437–2446.
Suárez-Ulloa V., Fernández-Tajes J., Manfrin C., Gerdol M., Venier P., Eirín-López JM. 2013. Bivalve
Omnics: State of the Art and Potential Applications for the Biomonitoring of Harmful Marine
Compounds. *Marine Drugs* 11:4370–4389.

Sueoka N., Kawanishi Y. 2000. DNA G+C content of the third codon position and codon usage biases of
human genes. *Gene* 261:53–62.

Takeuchi T., Kawashima T., Koyanagi R., Gyoja F., Tanaka M., Ikuta T., Shoguchi E., Fujiwara M., Shinzato
C., Hisata K., Fujie M., Usami T., Nagai K., Maeyama K., Okamoto K., Aoki H., Ishikawa T.,
Masaoka T., Fujiwara A., Endo K., Endo H., Nagasawa H., Kinoshita S., Asakawa S., Watabe S.,
Satoh N. 2012. Draft genome of the pearl oyster Pinctada fucata: a platform for understanding
bivalve biology. *DNA research: an international journal for rapid publication of reports on genes
and genomes* 19:117–130.

Vicario S., Moriyama EN., Powell JR. 2007. Codon usage in twelve species of Drosophila. *BMC
Evolutionary Biology* 7:226.

Wagner GP., Kin K., Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM
measure is inconsistent among samples. *Theory in Biosciences = Theorie in Den Biowissenschaften*
131:281–285.

Wan X-F., Xu D., Kleinhofs A., Zhou J. 2004. Quantitative relationship between synonymous codon usage
bias and GC composition across unicellular genomes. *BMC Evolutionary Biology* 4:19.

Wright F. 1990. The “effective number of codons” used in a gene. *Gene* 87:23–29.

Wu G., Culley DE., Zhang W. 2005. Predicted highly expressed genes in the genomes of Streptomyces
coelicolor and Streptomyces avermitilis and the implications for their metabolism. *Microbiology
(Reading, England)* 151:2175–2187.

Xie T., Ding D., Tao X., Dafu D. 1998. The relationship between synonymous codon usage and protein
structure. *FEBS Letters* 434:93–96.
Zapata F., Wilson NG., Howison M., Andrade SCS., Jörger KM., Schrödl M., Goetz FE., Giribet G., Dunn CW. 2014. Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. *Proceedings of the Royal Society of London B: Biological Sciences* 281:20141739.

Zeeberg B. 2002. Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes. *Genome Research* 12:944–955.

Zhang G., Fang X., Guo X., Li L., Luo R., Xu F., Yang P., Zhang L., Wang X., Qi H., Xiong Z., Que H., Xie Y., Holland PWH., Paps J., Zhu Y., Wu F., Chen Y., Wang J., Peng C., Meng J., Yang L., Liu J., Wen B., Zhang N., Huang Z., Zhu Q., Feng Y., Mount A., Hedgecock D., Xu Z., Liu Y., Domazet-Lošo T., Du Y., Sun X., Zhang S., Liu B., Cheng P., Jiang X., Li J., Fan D., Wang W., Fu W., Wang T., Wang B., Zhang J., Peng Z., Li Y., Li N., Wang J., Chen M., He Y., Tan F., Song X., Zheng Q., Huang R., Yang H., Du X., Chen L., Yang M., Gaffney PM., Wang S., Luo L., She Z., Ming Y., Huang W., Zhang S., Huang B., Zhang Y., Qu T., Ni P., Miao G., Wang J., Wang Q., Steinberg CEW., Wang H., Li N., Qian L., Zhang G., Li Y., Yang H., Liu X., Wang J., Yin Y., Wang J. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* 490:49–54.

Zhang L., Li L., Zhu Y., Zhang G., Guo X. 2014. Transcriptome analysis reveals a rich gene set related to innate immunity in the Eastern oyster (*Crassostrea virginica*). *Marine Biotechnology (New York, N.Y.)* 16:17–33.

Zhao C., Zhang T., Zhang X., Hu S., Xiang J. 2012. Sequencing and analysis of four BAC clones containing innate immune genes from the Zhikong scallop (*Chlamys farreri*). *Gene* 502:9–15.

Zhao X., Yu H., Kong L., Liu S., Li Q. 2014. Comparative Transcriptome Analysis of Two Oysters, *Crassostrea gigas* and *Crassostrea hongkongensis* Provides Insights into Adaptation to Hypo-Osmotic Conditions. *PLoS ONE* 9:e111915.
Figures
Figure 1: RSCU values (Y axis) in four Mytilida (panel A), Ostreoidea (panel B) and Unionida (panel C) species. Panel D shows a comparison between representative species from the three above mentioned orders and the Pectinida *Pecten maximus*. Codons are ordered by decreasing RSCU value on the X axis, based on *Mytilus galloprovincialis* (panels A and D), *Crassostrea gigas* (panel B) and *Elliptio complanata* (panel C).
Figure 2. Clustering of bivalve species according to the variation of codon usage. The
dendrogram was inferred with Cluster 3 by hierarchical clustering, using Euclidean distance as a
similarity metric and an average linkage clustering method. Effective number of codons (ENC)
and GC₃ metric for each species are also displayed. A three letter code near the species name
indicates the taxonomical classification according to Bieler et al. (2014). In detail, capital letters
identify one of the six major evolutionary lineages and lowercase letters identify the order. ARC:
Archiheterodonta; ANO: Anomalodesmata; IMP: Imparidentia; PAL: Palaeoheterodonta; PRO:
Protobranchia; PTE: Pteriomorphia; ada: Adapedonta; arc: Arcida; car: Carditoidea; cle:
Cleidothaeridae; cra: Crassatelloidea; cya: Cyamioidea; gal: Galeommatoidea; gas:
Gastrochaenidae; mac: Mactroidea; myi: Myida; myt: Mytilida; ncl: Nuculoidea; pec: Pectinida;
pin: Pinnoidea; pte: Pterioidea; ost: Ostreoidea; sol: Solemyoidea; sph: Sphaerioidea; tri:
Trigoniida; uni: Unionida; ven: Veneroidea.
**Figure 3**: Panel A: number of bivalve species (out of the 64 selected for this study) where a given codon was preferred (RSCU > 1). Panel B: Pearson correlation coefficient between the frequency of each codon and overall species CUB (negative ENC); NS = non-significant correlation, based on F-test of linear regression.
Figure 4. Simple linear regression analysis exemplifying the different contribution of four amino acids (Asn, Arg, Ser and His) to synonymous codon usage bias. ENC values for each species are plotted on the X axis and represent a measure of synonymous codon usage (lower ENC values indicate a stronger CUB). sENC-x values are plotted on the Y axis and represent the relative intensity of CUB for each amino acid in each species. R squared correlation values are shown for each regression line. Detailed data for all amino acids are reported in Table S5.
Figure 5. Regression line defining the correlation between ENC and GC$_3$ in bivalve species. Species pertaining to the bivalve clustering group Ia, Ib and II (see Figure 2) are marked as black, red and white circles, respectively. The P-value of the F-test of linear regression is 2.86x10$^{-36}$. 
Figure 6. Scatter plot of CAI (X axis) vs ENC (Y axis) for four representative bivalve species: *Mactra chinensis*, *Mytilus californianus*, *Crassostrea gigas* and *Placopecten magellanicus*. The reference set of highly expressed genes for each species is based on the orthologous genes of *C. gigas* (see Materials and methods).
| Species                | Taxonomic classification * | ENC  |
|------------------------|----------------------------|------|
| Galeomma turtoni       | Imparidentia               | Galeommatoidea 56.80 |
| Mya arenaria           | Imparidentia               | Myida 56.76 |
| Myochama anomoides     | Anomalodesmata             | Cleidotheridea 56.52 |
| Placopesten magellanicus | Pteriomorphia              | Pectinida 56.44 |
| Mizuhopecten yessoensis| Pteriomorphia              | Pectinida 56.25 |
| Azumaopecten farrei    | Pteriomorphia              | Pectinida 56.10 |
| Ennucula tenuis        | Protobranchia              | Nuculoidea 55.98 |
| Solemya velum          | Protobranchia              | Solemyidea 55.73 |
| Laternula elliptica    | Anomalodesmata             | Cleidotheridea 55.72 |
| Argopesten iradians    | Pteriomorphia              | Pectinida 55.65 |
| Ostrea lurida          | Pteriomorphia              | Ostreoida 55.61 |
| Pecten maximus         | Pteriomorphia              | Pectinida 55.61 |
| Sinonovacula constricta| Pteriomorphia              | Adapedonta 55.51 |
| Mimachlamys nobilis    | Pteriomorphia              | Pectinida 55.47 |
| Crassostrea virginica  | Pteriomorphia              | Ostreoida 55.33 |
| Ostreola stentina      | Pteriomorphia              | Ostreoida 55.25 |
| Crassostrea gigas      | Pteriomorphia              | Ostreoida 55.24 |
| Crassostrea angulata   | Pteriomorphia              | Ostreoida 55.13 |
| Glossus humanus        | Imparidentia               | Venerida 55.04 |
| Crassostrea corteziensis| Pteriomorphia             | Ostreoida 54.95 |
| Pinctada fucata        | Pteriomorphia              | Pterioidea 54.90 |
| Crassostrea hongkongensis| Pteriomorphia            | Ostreoida 54.88 |
| Polymesoda caroliniana | Imparidentia               | Venerida 54.71 |
| Ostrea edulis          | Pteriomorphia              | Ostreoida 54.70 |
| Ostrea chilensis       | Pteriomorphia              | Ostreoida 54.46 |
| Saccostrea glomerata   | Pteriomorphia              | Ostreoida 54.46 |
| Astarte sukata         | Archiheterodonta           | Crassatelloidea 53.95 |
| Pinctada martensi      | Pteriomorphia              | Pterioidea 53.89 |
| Cornicula fluminea     | Imparidentia               | Venerida 53.88 |
| Pinctada maxima        | Pteriomorphia              | Pterioidea 53.47 |
| Hiastella arctica      | Imparidentia               | Adapedonta 53.42 |
| Arctica islandica      | Imparidentia               | Venerida 53.42 |
| Cardites antiquata     | Archiheterodonta           | Carditoidea 52.92 |
| Neotrigonia margaritacea| Palaeoheterodonta         | Trigonida 52.87 |
| Villosa lienosa        | Palaeoheterodonta          | Unionida 52.76 |
| Margarifera margatifta | Palaeoheterodonta          | Unionida 52.68 |
| Eucrassatella cumingii | Archiheterodonta           | Crassatelloidea 52.66 |
| Ellipto complanata     | Palaeoheterodonta          | Unionida 52.64 |
| Uniomerus tetalasmus    | Palaeoheterodonta          | Unionida 52.62 |
Table 1. Effective number of codons (ENC) values in bivalves, ordered from the least to the most biased species. * based on the revised classification of bivalves by Bieler et al. (2014).
|                                | Crassostrea gigas | Mytilus galloprovincialis |
|--------------------------------|-------------------|---------------------------|
| Coding GC₃                     | 49.27%            | 30.80%                    |
| Global ENC                    | 55.24             | 45.58                     |
| Genomic GC content            | 33.69%            | 31.65%                    |
| Mutational bias               | towards A/T-ending codons | towards A/T-ending codons |
| Correlation between CUB and protein length | 0.03 (p-value 9.14X10⁻⁹) | 0.09 (p-value 3.75X10⁻¹⁸) |
| Correlation between GC₃ and protein length | 0.05 (p-value 9.69X10⁻¹⁹) | -0.10 (p-value 5.43X10⁻¹⁵) |
| Selection for translational accuracy | towards G/C-ending codons | towards A/T-ending codons |
| Correlation between CUB and gene expression (hemocytes) | 0.04 (p-value 8.99X10⁻¹²) | 0 (NS)                     |
| Correlation between GC₃ and gene expression (hemocytes) | 0.03 (p-value 1.44X10⁻⁸) | 0.07 (p-value 3.99X10⁻¹¹) |
| Correlation between CUB and gene expression (digestive gland) | 0.05 (p-value 1.16X10⁻¹⁷) | 0 (NS)                     |
| Correlation between GC₃ and gene expression (gills) | 0.07 (p-value 4.20X10⁻²⁰) | 0 (NS)                     |
| Correlation between CUB and gene expression (gills) | 0.06 (p-value 6.14X10⁻²⁰) | 0.12 (p-value 1.04X10⁻²⁷) |
| Selection for translational speed | towards G/C-ending codons | towards G/C-ending codons |
| Correlation between CUB and GC3 | -0.16 (p-value 5.42x10⁻¹⁴) | -0.53 (p-value 0)         |

**Table 2.** Influence of mutational bias and selection on codon usage bias in *Crassostrea gigas* and *Mytilus galloprovincialis*. Pearson correlation coefficients and p-values of F-test for linear regression analysis are shown.
1

Relative synonymous codon usage across bivalves

RSCU values (Y axis) in four Mytilida (panel A), Ostreoidea (panel B) and Unionida (panel C) species. Panel D shows a comparison between representative species from the three above mentioned orders and the Pectinida Pecten maximus. Codons are ordered by decreasing RSCU value on the X axis, based on Mytilus galloprovincialis (panels A and D), Crassostrea gigas (panel B) and Elliptio complanata (panel C).
Clustering of bivalve species according to the variation of codon usage

The dendrogram was inferred with Cluster 3 by hierarchical clustering, using Euclidean distance as a similarity metric and an average linkage clustering method. Effective number of codons (ENC) and GC\textsubscript{3} metric for each species are also displayed. A three letter code near the species name indicates the taxonomical classification according to Bieler et al. (2014). In detail, capital letters identify one of the six major evolutionary lineages and lowercase letters identify the order. ARC: Archiheterodonta; ANO: Anomalodesmata; IMP: Imparidentia; PAL: Palaeoheterodonta; PRO: Protobranchia; PTE: Pteriomorphia; ada: Adapedonta; arc: Arcida; car: Carditoidea; cle: Cleidothaeridae; cra: Crassatelloidea; cya: Cyamioidea; gal: Galeommatoidea; gas: Gastrochaenidae; mac: Mactroidea; myi: Myida; myt: Mytilida; ncl: Nuculoidea; pec: Pectinida; pin: Pinnoidea; pte: Pterioidea; ost: Ostreoidea; sol: Solemyoidea; sph: Sphaerioidea; tri: Trigoniida; uni: Unionida; ven: Veneroidea.
3

Codon usage bias in bivalves in mainly due to A/T-ending codons

Panel A: number of bivalve species (out of the 64 selected for this study) where a given codon was preferred (RSCU > 1). Panel B: Paerson correlation coefficient between the frequency of each codon and overall species CUB (negative ENC); NS = non-significant correlation, based on F-test of linear regression.

| preferred | A  | T  | C  | G  |
|-----------|----|----|----|----|
| AA        | 54 | 48 | 16 | 10 |
| AT        | 28 | 58 | 19 |    |
| AC        | 64 | 25 | 9  |    |
| AG        | 64 | 64 | 0  | 44 |
| TA        |    | 39 | 25 |    |
| TT        | 28 | 59 | 5  | 52 |
| TC        | 64 | 60 | 17 |    |
| TG        |    | 64 | 0  |    |
| CA        | 4  | 54 | 10 | 60 |
| CT        | 0  | 19 | 2  | 49 |
| CC        | 64 | 57 | 7  |    |
| CG        | 3  | 4  | 0  |    |
| GA        | 49 | 60 | 4  | 14 |
| GT        | 21 | 42 | 1  | 52 |
| GC        | 63 | 64 | 35 |    |
| GG        | 64 | 52 | 2  |    |

| ENC       | A  | T  | C  | G  |
|-----------|----|----|----|----|
| AA        | 0.86 | 0.9 | -0.9 | -0.86 |
| AT        | 0.82 | 0.34 | -0.92 | / |
| AC        | 0.84 | 0.71 | -0.87 | -0.77 |
| AG        | 0.95 | 0.9 | -0.89 | -0.59 |
| TA        |    | 0.94 | -0.94 | / |
| TT        | 0.93 | 0.91 | -0.91 | 0.24 NS |
| TC        | 0.91 | 0.84 | -0.88 | -0.76 |
| TG        |    | 0.75 | -0.75 | / |
| CA        | 0.85 | 0.91 | -0.91 | -0.85 |
| CT        | 0.1 NS | 0.2 NS | -0.92 | -0.91 |
| CC        | 0.79 | 0.68 | -0.87 | -0.73 |
| CG        | -0.79 | -0.11 NS | -0.88 | -0.87 |
| GA        | 0.93 | 0.92 | -0.92 | -0.93 |
| GT        | 0.85 | 0.9 | -0.86 | -0.93 |
| GC        | 0.58 | 0.86 | -0.89 | -0.78 |
| GG        | 0.54 | 0.75 | -0.86 | -0.83 |

**times preferred**

**correlation**

PeerJ reviewing PDF | (2015:03:4547:1:1:NEW 20 Nov 2015)
Simple linear regression analysis exemplifying the different contribution of four amino acids (Asn, Arg, Ser and His) to synonymous codon usage bias.

ENC values for each species are plotted on the X axis and represent a measure of synonymous codon usage (lower ENC values indicate a stronger CUB). sENC-x values are plotted on the Y axis and represent the relative intensity of CUB for each amino acid in each species. R squared correlation values are shown for each regression line. Detailed data for all amino acids are reported in Supplementary Table 5.
Regression line defining the correlation between ENC and GC₃ in bivalve species

Species pertaining to the bivalve clustering group Ia, Ib and II (see Figure 2) are marked as black, red and white circles, respectively. The P-value of the F-test of linear regression is 2.86x10⁻³⁶.
CAI vs ENC plot

Scatter plot of CAI (X axis) vs ENC (Y axis) for four representative bivalve species: *Mactra chinensis, Mytilus californianus, Crassostrea gigas* and *Placopecten magellanicus*. The reference set of highly expressed genes for each species is based on the orthologous genes of *C. gigas* (see Materials and methods).
Table 1 (on next page)

Effective Number of Codons (ENC) in bivalves

Effective number of codons (ENC) values in bivalves, ordered from the least to the most biased species. * based on the revised classification of bivalves by Bieler et al. (2014).
| Species                     | Taxonomic classification * | ENC  |
|----------------------------|----------------------------|------|
| Galeomma turtoni           | Imparidentia                | 56.80|
| Mya arenaria               | Imparidentia                | 56.76|
| Myochama anomoides         | Anomalodesmata              | 56.52|
| Placopecten magellanicus   | Pteriomorphia               | 56.44|
| Mizuhopecten yessoensis    | Pteriomorphia               | 56.25|
| Azumapecten farreri        | Pteriomorphia               | 56.10|
| Ennucula tenuis            | Protobranchia               | 55.98|
| Solemya velum              | Protobranchia               | 55.73|
| Laternula elliptica        | Anomalodesmata              | 55.72|
| Argopecten iradians        | Pteriomorphia               | 55.65|
| Ostrea lurida              | Pteriomorphia               | 55.61|
| Pecten maximus             | Pteriomorphia               | 55.61|
| Sinonovacula constricta    | Imparidentia                | 55.51|
| Mimachlamys nobilis        | Pteriomorphia               | 55.47|
| Crassostrea virginica      | Pteriomorphia               | 55.33|
| Ostreola stentina          | Pteriomorphia               | 55.25|
| Crassostrea gigas          | Pteriomorphia               | 55.24|
| Crassostrea angulata       | Pteriomorphia               | 55.13|
| Glossus humanus            | Imparidentia                | 55.04|
| Crassostrea cortziensis    | Pteriomorphia               | 54.95|
| Pinctada fucata            | Pteriomorphia               | 54.90|
| Crassostrea hongkongensis  | Pteriomorphia               | 54.88|
| Polymesoda caroliniana     | Imparidentia                | 54.71|
| Ostrea edulis              | Pteriomorphia               | 54.70|
| Ostrea chilensis           | Pteriomorphia               | 54.46|
| Saccostrea glomerata       | Pteriomorphia               | 54.46|
| Astarte sulcata            | Archiheterodonta            | 53.95|
| Pinctada martensi          | Pteriomorphia               | 53.89|
| Corbicula fluminea         | Imparidentia                | 53.88|
| Pinctada maxima            | Pteriomorphia               | 53.47|
| Hiatella arctica           | Imparidentia                | 53.42|
| Arctica islandica          | Imparidentia                | 53.42|
| Cardites antiquata         | Archiheterodonta            | 52.92|
| Neotrigonia margaritacea   | Palaeoheterodonta           | 52.87|
| Villosa lienosa            | Palaeoheterodonta           | 52.76|
| Margaritifera margatifera  | Palaeoheterodonta           | 52.68|
| Eucrassatella cunningii    | Archiheterodonta            | 52.66|
| Ellipto complanata         | Palaeoheterodonta           | 52.64|
| Unio merus tetratusm           | Palaeoheterodonta   | 52.62|
| Cyrenoida floridana        | Imparidentia                | 52.35|
| Pyganodon grandis          | Palaeoheterodonta           | 52.32|
| Lampsilis cardium          | Palaeoheterodonta           | 52.29|
| Species                                | Class       | Family       | Suborder | Percentage |
|---------------------------------------|-------------|--------------|----------|------------|
| Donacilla cornea                      | Imparidentia| Mactroidea   |          | 52.18      |
| Meretrix meretrix                     | Imparidentia| Venerida     |          | 51.50      |
| Lamychaena hians                      | Imparidentia| Gastrochaenida|        | 51.07      |
| Sphaerium nucleus                     | Imparidentia| Sphaeridae   |          | 51.01      |
| Mercenaria campechiensis              | Imparidentia| Venerida     |          | 50.98      |
| Cerastoderma edule                    | Imparidentia| Venerida     |          | 50.92      |
| Ruditapes decussatus                  | Imparidentia| Pinnoidea    |          | 50.21      |
| Ruditapes philipinarum                | Imparidentia| Venerida     |          | 50.13      |
| Atrina rigida                         | Imparidentia| Cyamiidae    |          | 49.62      |
| Diplodonta sp.                        | Imparidentia| Mytilida     |          | 49.40      |
| Geukensia demissa                     | Imparidentia| Cyamiidae    |          | 47.42      |
| Cycladicama cumingii                  | Imparidentia| Mytilida     |          | 47.25      |
| Perna viridis                         | Imparidentia| Mytilida     |          | 47.21      |
| Bathymodiolus azoricus                | Imparidentia| Mytilida     |          | 46.65      |
| Anadara trapezia                      | Imparidentia| Arcida       |          | 45.72      |
| Tegillarca granosa                    | Imparidentia| Arcida       |          | 45.69      |
| Mytilus galloprovincialis             | Imparidentia| Mytilida     |          | 45.58      |
| Bathymodiolus platifrons              | Imparidentia| Mytilida     |          | 45.46      |
| Mytilus edulis                        | Imparidentia| Mytilida     |          | 45.34      |
| Mytilus trossulus                     | Imparidentia| Mytilida     |          | 45.26      |
| Mytilus californianus                 | Imparidentia| Mytilida     |          | 45.02      |
| Mactra chinensis                      | Imparidentia| Mactroidea   |          | 40.65      |
Table 2 (on next page)

Influence of mutational bias and selection on codon usage bias in *Crassostrea gigas* and *Mytilus galloprovincialis*

Paerson correlation coefficients and p-values of F-test for linear regression analysis are shown.
|                                | **Crassostrea gigas** | **Mytilus galloprovincialis** |
|--------------------------------|-----------------------|-------------------------------|
| Coding GC$_3$                  | 49.27%                | 30.80%                        |
| Global ENC                     | 55.24                 | 45.58                         |
| Genomic GC content             | 33.69%                | 31.65%                        |
| Mutational bias                |                       |                               |
| towards A/T-ending codons      |                       | towards A/T-ending codons     |
| Correlation between CUB and protein length | 0.03 (p-value 9.14X10^-4) | 0.09 (p-value 3.75X10^-8)     |
| Correlation between GC$_3$ and protein length | 0.05 (p-value 9.69X10^-19) | -0.10 (p-value 5.43X10^-13)   |
| Selection for translational accuracy |                       |                               |
| towards G/C-ending codons      |                       | towards A/T-ending codons     |
| Correlation between CUB and gene expression (hemocytes) | 0.04 (p-value 8.99X10^-12) | 0 (NS)                       |
| Correlation between GC$_3$ and gene expression (hemocytes) | 0.03 (p-value 1.44X10^-4) | 0.07 (p-value 3.99X10^-13)    |
| Correlation between CUB and gene expression (digestive gland) | 0.05 (p-value 1.16X10^-17) | 0 (NS)                       |
| Correlation between GC$_3$ and gene expression (digestive gland) | 0.06 (p-value 1.04X10^-20) | 0.11 (p-value 2.67X10^-29)    |
| Correlation between CUB and gene expression (gills) | 0.07 (p-value 4.20X10^-23) | 0 (NS)                       |
| Correlation between GC$_3$ and gene expression (gills) | 0.06 (p-value 6.14X10^-23) | 0.12 (p-value 1.04X10^-29)    |
| Selection for translational speed |                       |                               |
| towards G/C-ending codons      |                       | towards G/C-ending codons     |
| Correlation between CUB and GC3 | -0.16 (p-value 5.42x10^-148) | -0.53 (p-value 0)            |
| Prevailing factor at the whole protein-coding transcriptome scale | mutational bias         | Mutational bias and selection for translational accuracy |