GC Content of Early Metazoan Genes and Its Impact on Gene Expression Levels in Mammalian Cell Lines

Ismail Sahin Gul1,2, Jens Staal1,2, Paco Hulpiau1,2, Evi De Keuckelaere1,2, Kai Kamm2, Tom Deroo1,2, Ellen Sanders1,2, Katrien Staes1,2, Yasmine Driege1,2, Yvan Saeys1,4, Rudi Beyaert1,2, Ulrich Technau5, Bernd Schierwater3, and Frans van Roy1,2,*

1Center for Inflammation Research, Flanders Institute for Biotechnology (VIB), Ghent, Belgium
2Department of Biomedical Molecular Biology, Ghent University, Belgium
3Institut für Tierökologie und Zellbiologie (ITZ), Division of Ecology and Evolution, Stiftung Tierärztliche Hochschule Hannover, Hannover, Germany
4Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
5Department of Molecular Evolution and Development, Faculty of Life Sciences, University of Vienna, Austria

*Corresponding author: E-mail: frans.vanroy@ugent.be.

Accepted: February 20, 2018

Abstract

With the genomes available for many animal clades, including the early-branching metazoans, one can readily study the functional conservation of genes across a diversity of animal lineages. Ectopic expression of an animal protein in, for instance, a mammalian cell line is a generally used strategy in structure–function analysis. However, this might turn out to be problematic in case of distantly related species. Here we analyzed the GC content of the coding sequences of basal animals and show its impact on gene expression levels in human cell lines, and, importantly, how this expression efficiency can be improved. Optimization of the GC3 content in the coding sequences of cadherin, alpha-catenin, and paracaspase of Trichoplax adhaerens dramatically increased the expression of these basal animal genes in human cell lines.

Key words: Trichoplax adhaerens, Placozoa, genome evolution, GC content, ectopic expression.

Introduction

Genomic sequences for several early diverging animals have become available during the last decade. These include the placozoan Trichoplax adhaerens (Placozoa), the demosponge Amphimedon queenslandica (Porifera) and the sea anemone Nematostella vectensis (Cnidaria) (Putnam et al. 2007; Srivastava et al. 2008; Srivastava et al. 2010). They help us to gain insight into the evolution of animal multicellularity, in development and in the mechanisms for division, adhesion, death, and specialization of cells. The transition from simple unicellular eukaryotes to multicellular organisms necessitates proteins that are able to mediate connections between individual cells. Members of the cadherin and armadillo superfamilies are essential key players in this important process. For instance, our previous research on the evolution of the cadherin and armadillo superfamilies revealed that only a few representative members of these superfamilies can be found in those early branching metazoa, including T. adhaerens and N. vectensis (Hulpiau and van Roy 2011; Gul et al. 2017).

Placozoans are small (1–3 mm) disk-shaped marine invertebrates that were initially discovered in 1883 on the walls of a seawater aquarium (Schulze 1883; for review see Schierwater 2005). Since then the enigmatic placozoans have drawn much attention from biologists because of their morphological simplicity and their basal position in metazoan evolution (Schierwater et al. 2016). In contrast to other metazoa, placozoans have half a dozen somatic cell types and lack any extracellular matrix (ECM) and any obvious basal lamina (Schierwater et al. 2009; Smith et al. 2014). This simple body plan also includes the lack of any organs, a nervous system, muscle cells and of any kind of symmetry. Although the genome of the placozoan T. adhaerens was sequenced as early as 2008 (Srivastava et al. 2008), there is unfortunately no method described to specifically introduce DNA constructs in
Placozoa. Alternatively, a most commonly preferred strategy to study the function of ancestral genes is to use human or mouse cell lines as an ectopic expression host (Luthringer et al. 2011; Selvan et al. 2015; Siau et al. 2016). However, it is not straightforward to perform (co) immunoprecipitation experiments, subcellular localization or any other type of functional conservation study in a heterologous animal host expression system whose protein synthesis machinery is separated by hundreds of millions of years of evolution. In ongoing work in our laboratories, where we study T. adhaerens and N. vectensis cadherins and catenins, we encountered problems in expressing the recombinant T. adhaerens proteins in human cell lines. Hence, we analyzed the GC content of coding sequences from basal metazoan animals and investigated its impact on gene expression levels in human cell lines, and how these can be improved.

In this study, we have expressed variants of a human, N. vectensis and T. adhaerens cadherin (CDH1) and catenin (CTNNA2) in HEK293T and HeLa cell lines to show that altering the GC3 to the level of the host expression system dramatically improves the expression of the GC poor genes.

Materials and Methods

Expression Plasmids

Plasmids expressing the full-length (FL) human zN-catenin (encoded by CTNNA2 in man) and FL human E-cadherin (encoded by CDH1 in man) were described previously (Goossens et al. 2007; Pieters et al. 2016). cDNA fragments encoding an N-terminal fragment of human zN-catenin (amino acids [AA] 24–206) or the cytoplasmic tail of human E-cadherin (AA 729–882 = C-terminal end) were amplified for Gateway cloning (Invitrogen) with gene-specific primers containing the AttB sites (supplementary table S1, Supplementary Material online). The amplified fragments were precipitated with polyethylene glycol and inserted in vectors pdcDNAMyc and pdcDNA-FLAGMyc produced by Gateway LR cloning. LR reactions between these Gateway Entry clones and the destination vector pdcDNA-FLAGMyc produced the expression plasmids: pdcDNA-FLAGMyc-Nv-Ctnna2-Nterm (encoding AA 33–267 of N. vectensis Ctnna2), pdcDNA-FLAGMyc-Nv-Cdh1-Cyto (AA 4175–4357 of N. vectensis Cdh1), pdcDNA-FLAGMyc-Ta-Ctnna2-Nterm (AA 9–262 of T. adhaerens Ctnna2), pdcDNA-FLAGMyc-Ta-Cdh1-Cyto (AA 4082–4270 of T. adhaerens Cdh1). The protein and DNA sequences encoded by the cloned N. vectensis and T. adhaerens cDNAs are provided in the supplementary table S2, Supplementary Material online. A plasmid expressing the (FL) T. adhaerens paracaspase (Pcasp) was described previously (Hulpiau et al. 2016). AttB-site-containing Ta-Cdh1-Cyto-GC45 (guanine and cytosine [GC] content: 45), Ta-Cdh1-Cyto-GC50, Ta-Cdh1-Cyto-GC60, Ta-Ctnna2-Nterm-GC60, Ta-Pcasp-GC40, Ta-Pcasp-GC45, Ta-Pcasp-GC50 and Ta-Pcasp-GC55 oligonucleotides with increased GC contents were ordered from Integrated DNA Technologies as gBlocks. Synthetic DNAs were amplified using the attB elongation primers (supplementary table S1, Supplementary Material online) and inserted into pDONR207 (cadherin and catenin constructs) (Invitrogen) or pDONR221 (paracaspase constructs) (Invitrogen) by the BP recombination. LR reactions between these Gateway Entry clones and the destination vectors pdcDNA-FLAGMyc and pdcDNA-FLAGMyc produced the expression plasmids: pdcDNA-FLAGMyc-Ta-Cdh1-Cyto-GC45 (AA 4082–4270), pdcDNA-FLAGMyc-Ta-Cdh1-Cyto-GC50 (AA 4082–4270), pdcDNA-FLAGMyc-Ta-Cdh1-Cyto-GC60 (AA 4082–4270), pdcDNA-FLAGMyc-Ta-Ctnna2-Nterm-GC60 (AA 9–262), pdcDNA-FLAGMyc-Ta-Pcasp-GC40 (AA 1–480), pdcDNA-FLAGMyc-Ta-Pcasp-GC45 (AA 1–480), pdcDNA-FLAGMyc-Ta-Pcasp-GC50 (AA 1–480), and pdcDNA-FLAGMyc-Ta-Pcasp-GC55 (AA 1–480).

Cell Lines and Transfections

HeLa and HEK293tsA1609neo (in short HEK293T) cells were cultured in Dulbecco’s modified Eagle’s medium (DMEM) supplemented with 10% fetal calf serum, 2 mM glutamine, and 0.4 mM sodium pyruvate. Cells were cultured at 37°C in a humified atmosphere containing 5% CO2. For transient transfection of HeLa and HEK293T cells, 5 x 106 cells were

seeded the day before in a 75 cm² cell culture flask. Cells were allowed to reach about 70% confluency and then transfected with the calcium phosphate method using 5 μg of Myc-tagged expression vector (see above). At 6–8 h post transfection, the medium was replaced with 15 ml of fresh medium.

SDS-PAGE and Western Blotting

Forty-eight hours posttransfection cells were washed once with ice-cold phosphate buffered saline (PBS) and lysed with ice-cold Laemmli lysis buffer. For subsequent SDS-PAGE analysis, we used an 8% polyacrylamide gel and loaded a total of 60 μg protein per lane. After blotting to PVDF membranes (Millipore), a subsequent blocking step was performed in the presence of 0.1% (vol/vol) Tween-20 (Sigma) and 5% (w/vol) nonfat dry milk. Membranes were then immunostained overnight with primary antivinculin (monoclonal mouse anti-Flag-HRP, Sigma, dilution 1/1,000) or anti-Myc (monoclonal mouse anti-Myc-HRP, Sigma, dilution 1/10,000) or anti-Flag (monoclonal mouse anti-Flag-HRP, Sigma, dilution 1/1,000) antibodies in the same buffer at 4°C. After several washing steps, the membranes were incubated with secondary horse-radish peroxidase (HRP)-conjugated antibody and detection was by enhanced chemiluminescence (ECL) (Amersham GE Healthcare).

mRNA Quantification

Cells were washed with PBS after which Trizol reagent (Life Technologies) was added. Following a short incubation time, cells were scraped from the tissue culture flasks and harvested. Total RNA was isolated using the Bio-Rad RNA extraction kit. cDNA synthesis was done with the iScript cDNA synthesis kit according to the manufacturer’s instructions (Bio-Rad). RT-qPCR reactions were performed using the SensiFAST SYBR No-ROX Kit (GC Biotech) and detection was performed using a LightCycler 480 Instrument (Roche Diagnostics). For normalization of gene expression levels the geometric mean of the most stable reference genes was used (Vandesompele et al. 2002). Two primer pairs were designed for the Myc tag (forward #1Myc 5'-TCTTTTGTGCAGATCCTCA-3', reverse #1Myc 5'–GCTTTTGCTCCATTCATT-3'), reverse #2Myc 5'–GCTCATTTCATTCAAGTCCTC-3', and one primer pair for the Flag tag (forward #1Flag 5'–AGAACCCACTGGTACCTTC-3' and reverse #1Flag 5’-CATCCGTACCTTCATTGCTCA-3'). The primers for the reference genes were designed as follows: RPL13A primers (forward 5’-CCTGGAGGAAGAGGAGAAAAGA-3’, reverse 5’-TTGAGGACATCTGTGATTGTC-3’), UBC primers (forward 5’-ATTGGGTCGCGTTCTTG-3’, reverse 5’-TTGCTTGACATCTGCATGG-3’) and YWHAZ primers (forward 5’-ACTTTTGTGACATTTGCGTCA-3’, reverse 5’-CCGCCAGAACCAGAT-3’).

Data Acquisition and Sequence Analyses

Trichoplax adhaerens proteomics data covering the expression abundance of 6,500 T. adhaerens proteins were taken from Ringrose et al. (2013). Uniprot accession IDs were converted to the NCBI RefSeq Nucleotide IDs using the Retrieve/ID mapping tool of the UniProt database (http://www.uniprot.org/; last accessed March 15, 2018). The nucleotide sequences were imported into the CLC Main workbench (http://www.clcbio.com/products/clc-main-workbench/; last accessed March 15, 2018) and coding sequences were extracted. To calculate the counts of nucleotides, frequency of codons and nucleotide frequency in codon positions, sequence statistics for each entry were created using the CLC Main workbench. This data was then merged and compared with the Log 10 abundance factors of the proteomics data. Codon usage data of Homo sapiens were obtained from http://gtmdb.ucsc.edu/; last accessed March 15, 2018. To calculate the frequency of codons and nucleotide frequency in codon positions of representative bilaterians, coding sequences for protein-coding genes of H. sapiens, Mus musculus, Gallus gallus, Danio rerio, Ciona intestinalis, Lingula anatina, Crasostrea gigas, Octopus bimaculoides, Daphnia pulex, Tribolium castaneum, Drosophila melanogaster, and Caenorhabditis elegans were downloaded from Ensembl, (http://www.ensembl.org/info/data/ftp/index.html; last accessed March 15, 2018) or (http://metazoa.ensembl.org/info/website/ftp/index.html; last accessed March 15, 2018). The coding sequence data for N. vectensis and Monosiga brevicollis were obtained from http://genome.jgi.doe.gov/; last accessed March 15, 2018, for Mnemiopsis leidyi and Pleurobrachia bachei from https://neurobase.rc.ufl.edu/; last accessed March 15, 2018, for Oscarella carmela and Haliclona amboinensis from http://www.compagen.org/; last accessed March 15, 2018, for Capsaspora owczarzaki and Dictyostelium discoideum from http://protists.ensembl.org/info/website/ftp/index.html; last accessed March 15, 2018, and for Amphimedon queenslandica from http://amphimedon.qcloud.qcif.edu.au/downloads.html; last accessed March 15, 2018. Detailed information on these data sets and calculations are provided in supplementary table S3, Supplementary Material online. To calculate the nucleotide frequency in codon positions the Biopython SeqUtils module was used. The GC and GC3 plots were created in R using the ggplot2 package.

The nucleotide sequences of the expression plasmids (pdcDNAMyc backbone plus insert) were subjected to a compositional sequence analysis using the CpGPlot and Isochore tools from EMBOSS (European Molecular Biology Open Software Suite, (http://www.ebi.ac.uk/Tools/sqseqstats/; last accessed March 15, 2018). All profiles were obtained using an overlapping window of 500 base pairs.
Results

Low GC(3) Content of T. adhaerens Genes in Comparison with Other Organisms

At the onset of this project we investigated the GC (or Guanine-Cytosine content) and GC3 (proportion of G and C in the third position of the codons) content of the protein coding sequences (CDS) of 22 metazoan and nonmetazoan species using the available data from the genome databases (supplementary table S3, Supplementary Material online). Figure 1 shows that the mean GC3 content of the four vertebrate CDS is above 55%. In contrast, none of the GC3 contents of the investigated early diverging metazoans (or nonbilaterian animals) exceeds 50%. In figure 2, we have compared the distribution of the GC and GC3 contents of the CDS of the vertebrate species and nonbilaterian animals. The mean of the GC content of the nonbilaterian T. adhaerens CDS is 38%, and thus the lowest seen in any metazoan genome so far investigated (fig. 1). Figure 2D shows the distribution of the GC content of the 11,520 predicted T. adhaerens CDS. Remarkably, 80% of the T. adhaerens CDS have GC3 contents between 25% and 35% (fig. 2D), with a mean of 31%. In sharp contrast, the GC3 content of the mammalian CDS ranges from 10% to almost 100% (Alvarez-Valin et al. 2002), with a mean of 59% (figs. 1 and 2A).

Expression of T. adhaerens cDNAs in Human Cell Lines

In a first set of experiments, we compared the protein and mRNA expression levels in HeLa cells for constructs encoding Myc-tagged cytoplasmic tails of cadherins from human (Hs-Cdh1-Cyto), the starlet sea anemone N. vectensis (Nv-Cdh1-Cyto), and the placozoan T. adhaerens (Ta-Cdh1-Cyto). These coding sequences were cloned downstream of the CMV promoter in the pdcDNAMyc vector. Except for these insert regions, the plasmid constructs used were the same for all three expression vectors (supplementary fig. S1, Supplementary Material online) and equal amounts of pdcDNAMyc-(Hs/Nv/Ta)-Cdh1-Cyto vectors were transfected. After 2 days of incubation, the cells were lysed and the amount of the pdcDNAMyc-(Hs/Nv/Ta)-Cdh1-Cyto proteins expressed was quantified by Western-blotting using an anti-Myc antibody.

The wild-type (WT) Hs-Cdh1-Cyto and Nv-Cdh1-Cyto protein fragments, encoded by GC3-rich sequences (57% and 63%, respectively), were readily expressed in transfected HeLa cells (fig. 3A). However, expression of the GC3-poor Ta-Cdh1-Cyto (GC3: 21%; WT sequence) was undetectable. Next, we quantified the mRNA expression levels of (Hs/Nv/Ta)-Cdh1-Cyto using real-time RT-PCR. The amount of Hs-Cdh1-Cyto mRNA was over 3-fold higher than the amount of Nv-Cdh1-Cyto and Ta-Cdh1-CytoWT mRNAs. Despite the high Hs-Cdh1-Cyto levels, the expression of the Hs-Cdh1-Cyto protein was not higher than the expression of the Nv-Cdh1-Cyto protein (fig. 3A and B). Similarly, while the Nv-Cdh1-Cyto and Ta-Cdh1-CytoWT mRNA levels were comparable, the amount of expressed Nv-Cdh1-Cyto and Ta-Cdh1-CytoWT proteins were remarkably different (fig. 3A and B).

We have extended this observation by an expression experiment in HeLa cells for the N-terminal region of xN-catenin (Ctnna2) of, respectively, H. sapiens, N. vectensis, and T. adhaerens. Quite similar to the previous experiment, the levels of Hs-Ctnna2-Nterm and Nv-Ctnna2-Nterm proteins were readily detectable; however no Ta-Ctnna2-NtermWT was detectable (fig. 3C), despite high expression levels of the encoding mRNA (fig. 3D). Similar results were obtained when the expression efficiencies of (Hs/Nv/TaWT)-Cdh1-Cyto or of (Hs/Nv/TaWT)-Ctnna2-Nterm constructs were compared in HEK293T cells (supplementary fig. S2, Supplementary Material online).

To test whether the expression of the T. adhaerens cDNAs could be modulated by changing their GC content, we designed two synthetic variants of Ta-Cdh1-Cyto. The GC3 content of the first variant has been enriched from 21% (WT, with GC: 35%) to 51% (with GC: 45%), and the second enhanced variant had a GC3 content as high as 99% (with GC: 60%). All of these codon modifications were synonymous changes and therefore did not change the resulting AA sequence of the Ta-Cdh1-Cyto protein (supplementary fig. S3, Supplementary Material online). Remarkably, Ta-Cdh1-Cyto protein synthesis increased significantly with increasing GC3/GC content of the encoding cDNA. Although the WT Ta-Cdh1-Cyto protein was not detectable at all, the Ta-Cdh1-Cyto-GC45 encoded protein was expressed at weak but detectable levels (fig. 3A). When the GC3 content of the Ta-Cdh1-Cyto mRNA was further increased to 99% (GC: 60%), the encoded protein was expressed 11-fold stronger than in case of the protein encoded by Ta-Cdh1-Cyto-GC45. qPCR experiments demonstrated a similar tendency between GC content and mRNA levels in HeLa cells. Ta-Cdh1-Cyto-GC45 and Ta-Cdh1-Cyto-GC60 mRNAs were expressed, respectively, 2- and 3-fold stronger than the WT Ta-Cdh1-Cyto mRNA (fig. 3B). We have repeated the ectopic expression of Ta-Cdh1-Cyto constructs in HEK293T cells, and in addition to the two previously used GC enriched versions of Ta-Cdh1-cyto (GC45 and GC60), we expressed a version of Ta-Cdh1-cyto with a GC content of 50%. Supplementary figure S4, Supplementary Material online shows the gradual increase in mRNA and protein levels when the GC3 content of Ta-Cdh1-Cyto was increased. We obtained similar results when we tested the GC enhanced version of Ta-Ctnna2-Nterm. When the GC3 content of Ta-Ctnna2-Nterm was increased from 36% to 94%, there was an 18- and 2-fold increase at protein and mRNA levels, respectively (fig. 3C and D). These results were confirmed upon transfection of HEK293T cells.
The expression experiments discussed above were performed on partial gene fragments of cadherins and catenins. To demonstrate the wider applicability of our findings, we have performed an additional expression experiment on a full-length *T. adhaerens* paracaspase (Ta-Pcasp) construct (Hulpiau et al. 2016). In addition to the WT Ta-Pcasp (with GC: 37% and GC3: 32%), we have designed and expressed four different GC enriched Ta-Pcasp constructs in HEK293T cells: Ta-Pcasp-GC40 (with GC3: 40), Ta-Pcasp-GC45 (with GC3: 55%), Ta-Pcasp-GC50 (with GC3: 70%) and Ta-Pcasp-GC55 (with GC3: 86%). The resulting Western blot and qPCR data confirmed the gradual increase in mRNA and protein levels when the GC content of this additional *T. adhaerens* cDNA was enriched (fig. 3E and F). Altogether, these experiments indicate that the GC content in CDS strongly affects the expression efficiency of the *T. adhaerens* genes in human cells. In view of the data presented here (fig. 3 and supplementary figs. S2 and S4, Supplementary Material online), we cannot exclude that both transcription and translation efficiencies are enhanced in function of increasing GC3 content. Hence, the majority of the *T. adhaerens* genes may require a major nucleotide optimization to increase their ectopic expression in mammalian cell lines.

**Correlation between *T. adhaerens* Gene Expression and GC Content**

It has been shown that the human genome is highly compositionally compartmentalized (Bernardi 1995), and abundantly.
expressed genes are localized in the GC-richest isochores (Bernardi 2007). As we showed above, the T. adhaerens CDs have a low GC and GC3 content compared with those of vertebrates. Figure 4 shows the relationship between the GC content and the respective abundance factor of T. adhaerens proteins using the published expression data (Ringrose et al. 2013). Remarkably, these results suggest that, similar to the human genes, abundantly expressed T. adhaerens genes have a higher GC content (fig. 4).

Next, based on the available expression data we have generated a codon usage table for T. adhaerens and compared this with the codon usage table for the human genome. Supplementary figure S5, Supplementary Material online, shows that, for both data sets, 14 (out of 61) sense codons have a usage above 2%. However, while for T. adhaerens six out of seven pure A/T codon combinations are present in this window, we found only one for H. sapiens (AAA = K codon, 2.44% usage). Moreover, similar to the Drosophila melanogaster genome (Vicario et al. 2007), the codon usage in the human genome is biased toward G/C-ending codons (or GC3), as 11 out of 14 GC3 codons have ≥2% usage (supplementary fig. S5B, Supplementary Material online). In contrast, in the T. adhaerens genome, the codon usage is biased toward A/T-ending codons (12 out of 14 AT3 codons have ≥2% usage) (supplementary fig. S5A, Supplementary Material online).

**Discussion**

Over the last two decades, GC(3) content dynamics in the context of vertebrate evolution have been extensively studied (Costantini et al. 2009; Romiguier et al. 2010). Most of the mammalian genomes are composed of GC-poor (L1 and L2 families, GC% ranging between 36% and 39%) and GC-rich isochores (H1-H3 families, GC% ranging between 43% and 55%) (Costantini et al. 2006). Previous studies indicate that GC3 can be used to detect isochore existence and structure (Bernardi [2001]; Romiguier et al. [2010]; but see Elhaik et al. [2009] for opposing arguments), and GC3-rich human genes are generally located...
in H3 isochores (Alvarez-Valin et al. 2002). Similar to other vertebrate genomes, the mean GC3 content of the human coding sequences is above 55% (fig. 1). In nonvertebrate metazoans the mean GC3 content of coding sequences is often significantly lower than 50% (supplementary table S3, Supplementary Material online), and can drop to as low as 31% in *T. adhaerens*. In contrast, the fruit fly (*D. melanogaster*) has a much higher GC3% of 65% (figs. 1 and 2).

In this study, we analyzed the GC(3) content of the genes in three nonbilaterian species *N. vectensis, T. adhaerens*, and *A. queenslandica* and analyzed the ectopic expression in human cell lines of two proteins from both *N. vectensis* and *T. adhaerens*. The data presented in figure 3 and supplementary figures S2 and S4, Supplementary Material online, clearly show that “optimizing” the GC3 content of the *T. adhaerens* genes causes an increase in both protein and mRNA expression levels in mammalian cell lines. It should be noted that in order to obtain such increased expression for *T. adhaerens* genes, we did not specifically optimize either the codon usage in view of the host (human), like in the previously reported “preferred human codon-optimized method” (Kudla et al. 2006; Inouye et al. 2015), or the Kozak sequences. Instead, we changed the GC content in the third position of each codon without change on the encoded AA (supplementary fig. S3 and table S2, Supplementary Material online). This suggests that synonymous substitutions in CDS towards G/C codon combinations have a positive effect on the expression levels in mammalian cell lines. At least in part, this was due to increased steady-state mRNA levels, as observed before for human proteins and green fluorescent protein (GFP) (Kudla et al. 2006; Inouye et al. 2015), but we do not exclude an additional and significant enhancing effect at the translational level.

The positive effect of GC3 content on gene expression also raises important questions regarding the neutral theory of evolution (Kimura 1968, 1983), which assumes that most mutations are either “silent” or have a negligible effect on fitness. Especially synonymous substitutions, which do not alter the AA sequence, are considered neutral and thought to be propagated by chance alone. This assumption is also the

---

![Fig. 3.](https://example.com/fig3.png)

**Fig. 3.** — Expression of human (Hs), *N. vectensis* (Nv), and *T. adhaerens* (Ta) cDNAs in HEK293T cells. Equal amounts of pdcDNA(Myct-Hs/Nv/Ta)-Cdh1-Cyto, pdcDNA(Myct-Hs/Nv/Ta)-Ctnna2-Nterm and pdcDNAFlag-Ta-Pcasp plasmids were used to transfect HEK293T cells. (A, C, E) Protein levels were analyzed by Western blotting using an anti-Myc antibody or an anti-Flag antibody as indicated; an anti-vinculin antibody was used as a control. (B, D, F) Expression of (Hs/Nv/Ta)-Cdh1-Cyto, (Hs/Nv/Ta)-Ctnna2-Nterm and Ta-Pcasp mRNA. After 24 h, total cellular RNA was isolated and analyzed by qRT-PCR. As negative controls, untransfected HEK or HEK293T cells were used as well as transfection with empty pdcDNA(Myct) plasmid (mock). Values have been normalized such that the values of the different WT constructs of *T. adhaerens* are equal to 1. The results are representative of three experiments.
basis for $dN/dS$ calculations (Hurst 2002), which are commonly used to detect positively selected genes. The ratio of synonymous (“neutral”) substitutions per synonymous site is considered the background mutation rate and compared with the ratio of nonsynonymous (protein altering) substitutions per nonsynonymous site. However, more recently it has been suggested that mutations at synonymous sites are not necessarily neutral (for reviews, see Chamary et al. 2006; Shabalina et al. 2013). Just as one example, in human it has been found that the GC3 content and synonymous substitution rates differ between constitutive and alternatively expressed exons of the same gene (Iida and Akashi 2000), suggesting varying constraint between different synonymous sites. There is also strong evidence that a great portion of 4-fold degenerate sites (all changes are synonymous) in mammalian and avian genes are under selective constraint (Eory et al. 2010; Kunstner et al. 2011). Similarly, the effect of GC3 content on gene expression reported here suggests that synonymous substitutions are not always neutral and may have an effect on an organism’s fitness. It thus further corroborates the view that $dN/dS$ ratios need to be cautiously interpreted for estimating natural selection.

Optimizing heterologous gene expression is a key step to produce and study proteins of interest. It has been shown that high GC content leads to increased mRNA levels in mammalian cells (Kudla et al. 2006). This phenomenon might cause a major problem for scientists who perform experimental studies to gain insight into the functional conservation of certain genes and gene families. More than 5,000 T. adhaerens CDSSs, including the orthologs of E-cadherin, $\alpha$- and $\beta$-catenin, p53, and DNA Topoisomerase II, have low GC values (<39%). These GC-poor coding sequences may behave as GC-poor families in the mammalian context and probably cannot be expressed in mammalian cell lines as we clearly demonstrated for T. adhaerens E-cadherin and $\alpha$-catenin in HeLa (fig. 3) and HEK293T (supplementary figs. S2 and S4, Supplementary Material online) cells. Our work, together with previous studies of expression of human proteins, GFP and luminescence proteins in mammalian cell lines (Kudla et al. 2006; Inouye et al. 2015), suggest that GC optimization is essential for the efficient production of proteins of interest in mammalian cell lines. Our work demonstrates that this is especially true for proteins of phylogenetically distant animal groups, such as those of lower metazoans like T. adhaerens, which are important for studying protein evolution in particular and animal evolution in general.

**Supplementary Material**

Supplementary data are available at Genome Biology and Evolution online.

**Acknowledgments**

We thank our colleagues and Prof. Dr Klaas Vandepoele (Ghent University) for helpful discussions. This work was supported by the Research Foundation—Flanders (FWO-Vlaanderen, Award G.0320.11 N), the Belgian Science Policy (Interuniversity Attraction Poles—Award IAP7/07), and the Special Research Fund of Ghent University (Award BOF 01J14211). B.S. acknowledges support from the German Science Foundation (DFG Schi-277/27-1).

**Literature Cited**


Associate editor: Maria Costantini