Theory Biosci.
DOI 10.1007/s12064-012-0167-y

ORIGINAL PAPER

The correlation of genome size and DNA methylation rate in metazoans

Marcus Lechner · Manja Marz · Christian Ihling · Andrea Sinz · Peter F. Stadler · Veiko Krauss

Received: 17 May 2012 / Accepted: 3 October 2012

Abstract Total DNA methylation rates are well known to vary widely between different metazoans. The phylogenetic distribution of this variation, however, has not been investigated systematically. We combine here publicly available data on methylcytosine content with the analysis of nucleotide compositions of genomes and transcriptomes of 78 metazoan species to trace the evolution of abundance and distribution of DNA methylation. The depletion of CpG and the associated enrichment of TpG and CpA dinucleotides are used to infer the intensity and localization of germline CpG methylation and to estimate its evolutionary dynamics. We observe a positive correlation of the relative methylation of CpG motifs with genome size. We tested this trend successfully by measuring total DNA methylation with LC/MS in orthopteran insects with very different genome sizes: house crickets, migratory locusts and meadow grasshoppers. We hypothesize that the observed correlation between methylation rate and genome

Electronic supplementary material The online version of this article (doi:10.1007/s12064-012-0167-y) contains supplementary material, which is available to authorized users.

M. Lechner
Institut für Pharmazeutische Chemie, Philipps-Universität Marburg, Marbacher Weg 6, 35037 Marburg, Germany
e-mail: lechner@staff.uni-marburg.de

M. Marz
RNA Bioinformatics and High Throughput Analysis, Friedrich Schiller University, Leutragraben 1, 07743 Jena, Germany
e-mail: manja@uni-jena.de

C. Ihling · A. Sinz
Department of Pharmaceutical Chemistry and Bioanalytics, Institute of Pharmacy, Martin-Luther-University Halle-Wittenberg, Wolfgang-Langenbeck-Straße 4, 06120 Halle, Germany
e-mail: christian.ihling@pharmazie.uni-halle.de

A. Sinz
e-mail: andrea.sinz@pharmazie.uni-halle.de

P. F. Stadler
Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany

P. F. Stadler
Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany

P. F. Stadler
RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie, Deutscher Platz 5e, 04103 Leipzig, Germany

P. F. Stadler
Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Wien, Austria

P. F. Stadler
Center for Non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, 1870 Frederiksberg, Denmark

P. F. Stadler (✉)
Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA
e-mail: studla@bioinf.uni-leipzig.de

V. Krauss (✉)
Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany
e-mail: veiko@bioinf.uni-leipzig.de

V. Krauss
Group of Developmental Genetics, Institute for Biology, Martin Luther University Halle-Wittenberg, Weinbergweg 10, 06120 Halle, Germany

Published online: 07 November 2012

size is due to a dependence of both variables from long-term effective population size and is driven by the accumulation of repetitive sequences that are typically methylated during periods of small population sizes. This process may result in generally methylated, large genomes such as those of jawed vertebrates. In this case, the emergence of a novel demethylation pathway and of novel reader proteins for methylcytosine may have enabled the usage of cytosine methylation for promoter-based gene regulation. On the other hand, persistently large populations may lead to a compression of the genome and to the loss of the DNA methylation machinery, as observed, e.g., in nematodes.

Keywords Evolution of genome size · Suppression of transposon activity · Germline methylation rate · Cytosine methylation · CpG bias · CpNpG bias

Introduction

In animals, up to 12 % of cytosines (C) are methylated (Regev et al. 1998; Grunau et al. 2001; Varriale and Bernardi, 2006; Varriale and Bernardi, 2006; Zemach et al. 2010; Feng et al. 2010). DNA methylation in animals as well as in plants is associated with the suppression of promoters (Goll and Bestor 2005; Laurent et al. 2010). Within active genes, DNA methylation is restricted to the gene body, i.e., to the region of transcriptional elongation (Simmen et al. 1999; Zemach et al. 2010). In this context, DNA methylation apparently suppresses a possibly interfering initiation of transcription (Maunakea et al. 2010). This effect may be dependent on the positions of nucleosomes and/or histone modifications as histone H4K36me3 (Hodges et al. 2009; Choi et al. 2009; Nanty et al. 2011).

Another evolutionary old function of DNA methylation is the suppression of activity of transposable elements (TEs), which was demonstrated for model organisms of plants, animals, and fungi, as well as for some other unicellular eukaryotes including *Entamoeba* and *Dictyostelium* (Yoder et al. 1997; Walsh and Bestor 1999; Krauss and Reuter 2011). Transposon suppression may be accomplished by compaction of nucleosomes (Choy et al. 2010), by stabilization of DNA duplexes (Thalhammer et al. 2011), by inhibition of transcription factor binding or by mutational decay through deamination of methylcytosine (mC) to thymine. Transposable elements constitute up to 45 % of metazoan genomes (The Human Genome International Sequencing Consortium 2001). Both the fraction of the genome occupied by TEs and the diversity of TEs grow with genome size (Lynch 2007; Feschotte et al. 2009). The hypothesis that TE control is a main function of cytosine methylation (Yoder et al. 1997) therefore implies that a greater fraction of cytosines should be methylated in larger genomes, i.e., one would expect a positive correlation of genome size and the partition of cytosines that are methylated. However, earlier analyses of total methylation (Regev et al. 1998) did not observe a significant correlation.

More recently, it was suggested that DNA methylation has a fundamentally different distribution in invertebrates compared to vertebrates (Suzuki and Bird 2008; Zemach et al. 2010). Within the invertebrate genomes studied so far, cytosine methylation is concentrated at broadly expressed genes (Nanty et al. 2011), whereas the global methylation in vertebrates is interrupted only by small regions that mostly contain promoters (Molaro et al. 2011). However, these differences may simply result from a different activity and not from a different targeting of the methylation apparatus. Consistent with this point of view, methylation of TEs was reported also from diverse invertebrates such as *Ciona*, *Echinus*, *Drosophila*, *Medauroidea*, and *Locusta* (Bird et al. 1979; Krauss et al. 2009; Feng et al. 2010; Krauss and Reuter 2011; Robinson et al. 2011). Global methylation patterns of vertebrates, furthermore, seem to had evolved gradually during the radiation of deuterostomes (Okamura et al. 2010). Thus, changes of methylation rates may explain the observed changes of methylation patterns.

Alternative to a measurement of the cytosine methylation rate, expressed as the percentage of methylcytosines among all genomic cytosines, an analysis of CpG depletion may be used to estimate the extent of germline DNA methylation (Bird 1980). In mammals, the amount of CpG→TpG/CpA substitutions caused by cytosine methylation was determined to be about ten times higher than the rate of a transversion (Arndt et al. 2003). Thus, dinucleotide abundances could be profoundly influenced by a significant methylation of CpG positions (Duret and Galtier 2000). CpG depletion rates in primate genomes depend causally on the rate of DNA methylation. They are higher in most types of transposons than in surrounding sequences (Elango et al. 2008). Moreover, Simmen (2008) shows, using vertebrate and invertebrate genomes, that CpG depletion and TpG/CpA overrepresentation rates are strongly correlated with each other and with the occurrence of DNA methylation. In contrast to CpG, methylation of CpH (H = A,T, or C) is much weaker and much less prevalent in animals (Laurent et al. 2010). It exhibits a similar genome-wide distribution as CpG methylation in human cells. In contrast to the situation in plants, specific aberrant allocations of CpH methylation are also unknown in other metazoa (Regev et al. 1998; Krauss et al. 2009; Walsh et al. 2010; Zemach et al. 2010; Feng et al. 2010). Thus, CpG methylation data are the appropriate model to study the evolution of cytosine methylation in animals.

Genome sizes and CpG methylation rates are simultaneously available for only 29 animal species (Grunau et al. 2001; Gregory 2010). A scatter plot of these data suggests a significant correlation between the genome size and the ratio of methylcytosines and CpG dinucleotides (mC/CpG, see Supplementary Material 1). A closer inspection of these data, however, shows that vertebrates (with large genomes) have mC/CpG >40 %, while invertebrates (with smaller genomes) have mC/CpG <40 %. The available data also have a very uneven phylogenetic distribution: 12 of the 29 species are mammals, 6 are insects and only 11 represent other taxa. Instead of a systematic trend with genome size, the available methylation data could in principle also be explained by group-specific properties.

Since this sampling bias cannot be corrected easily, we have to resort to evaluating the traces that are left by DNA methylation in the genome sequence itself. To this end, we evaluate di- and trinucleotide composition data for 78 public available genome and transcriptome data sets. The depletion of CpG and the associated enrichment of TpG and CpA dinucleotides were used to infer intensity and localization of germline CpG methylation. Our data support the view that the level of methylation is positively correlated with genome size. In addition, we generated evidence that at least some TEs are methylated both in vertebrates and invertebrates. Fully methylated genomes were found to be exactly restricted to the taxon of jawed vertebrates (gnathostomes). Here, evolution of a novel demethylation pathway and of novel reader molecules appeared to enable the usage of cytosine methylation to temporarily silence endogenous genes. To further support the relationship between genome size and methylation rate within a certain taxon that contain species with widely diverging genome sizes, we measured the total cytosine methylation of crickets, locusts and grasshoppers (all Orthoptera) using combined liquid chromatography and mass spectrometry (LC/MS) These original data are consistent with the supposed positive correlation between genome size and cytosine methylation. Finally, we discuss possible consequences of our findings for the interpretation of epigenetic modifications and their roles during evolution.

Materials and methods

Genome size and sequence data

Genome size data were retrieved from the Animal Genome Size Database (Gregory 2010). We used medians of the available measurements. For nine species without known genome sizes, other species of the same genera were utilized. The data sets underlying the analyses and figures of this contribution can be downloaded from http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/10-026.

Dnmt, MBD, AID, Kaiso, and Zbtb38 genes of metazoans were sampled from genome project and EST databases using `tblastn` (Gertz et al. 2006). In particular, we used the human orthologs to retrieve sequences from finished and unfinished genome projects deposited at the NCBI database. In addition, single trace sequences were screened using `blastn`. The orthology of these candidate sequences was verified by `blastx` analysis using the protein refseq database of NCBI. We counted only such sequences as originated from orthologous genes which return a more than 10^5-fold lower E value for the orthologous human protein than for other human proteins. In case of MBD proteins, we searched for orthologs of the human MBD2 protein that show a preferential methyl-binding domain structure (Ho et al. 2008). MBD orthologs that did not meet this condition were not counted.

Transposable elements data were extracted from `RepBase` (Jurka et al. 2005). We used all autonomous, non-autonomous and simple elements of the eukaryotic taxon group. Species with < 15 entries were excluded. Genomic regions similar to `RepBase` entries were located by `blast` with E value $\leq 10^{-10}$ and their sequence identity to the most similar `RepBase` sequence was determined. The level of sequence similarity approximates the relative age. Nucleotide frequencies for CpG and CpNpG calculations were determined using `galculator`.[1]

CpG and CpNpG Bias

CpG dinucleotides are depleted because cytosines in a CpG context are preferential methylated and mC then frequently mutates to thymine (Simmen 2008). Hence the frequency of CG dinucleotides decreases at the expense of the mutation product TpG and its complement CpA (Bird 1980; Hodges et al. 2009). We measured the CpG depletion and the accompanying TpG and CpA enrichment as parallel consequences of CpG-oriented methylation in the germline by the ratio of the resulting relative dinucleotide bias

$$\eta = \frac{p^*_{\mathrm{TG}} + p^*_{\mathrm{CA}}}{2p^*_{\mathrm{CG}}} \tag{1}$$

where $p^*_{XY} = p_{XY}/(p_X p_Y) | \, X, Y \in \{\mathrm{A, T, C, G}\}$. p_X is the absolute mononucleotide frequency of X and p_{XY} is the absolute dinucleotide frequency of XY. Thus, p^*_{XY} can be considered as quotient of the observed and the expected dinucleotide content.

[1] Available at http://www.bioinf.uni-leipzig.de/Software/galculator/.

We use η instead of the CpG observed/expected measure p^*_{CG} because both CpG depletion and TpG and CpA enrichment are connected to the CpG methylation in the germline (Simmen 2008). Therefore, η should be a better measure than p^*_{CG} for the influence of cytosine methylation on dinucleotide content. In contrast to p^*_{CG} (Duret and Galtier 2000), furthermore, we find that the parameter η is not significantly correlated with GC content (see Supplementary Material 2). Analogously, we determined η_{CNG} using the frequency p_{CNG} of CNG motifs instead of p_{CG}.

Independent contrasts

The analysis of independent contrasts required a phylogenetic tree with corresponding branch lengths. The tree topology is a composite of the phylogenies described by Hejnol et al. (2009); Consortium (2007); Holterman et al. (2006); Kriegs et al. (2006); Chen et al. (2004), see Fig. 2. To approximate the branch lengths, ribosomal small subunit sequences were taken from the ribosomal RNA database `silva` and aligned using `SINA Web Aligner` (eukaria ssu r106) (Pruesse et al. 2007). For *Takifugu rubripes*, *Helobdella robusta* and *Pongo pygmaeus* we have chosen sequences of closely related species (*Takifugu pardalis*, *Helobdella stagnalis* and *Pongo abelii*). *Euprymna scolopes* and *Drosophila grimshawi* had to be removed from the analysis because the most related species were already included in the data set which then comprised 76 out of 78 species. The alignment was trimmed to columns with gaps in <50 % of all species. Branch lengths were then estimated for the tree topology of Fig. 1 by `PhyML` (Guindon and Gascuel 2003) using the HKY85 model (see Supplementary Data).

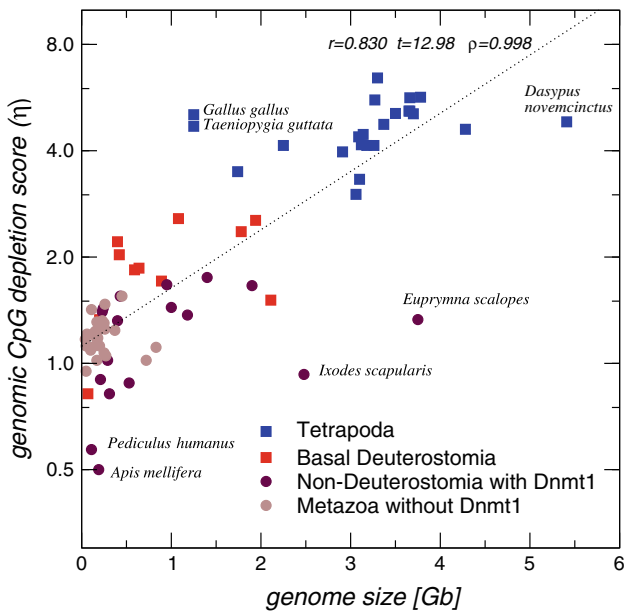

Fig. 1 The parameter η, which measures CpG depletion and TpG + CpA enrichment, correlates with metazoan genomes sizes. The data covers all metazoan species for which genome sequences are currently available (30 vertebrates and 48 invertebrates). Seven outliers are indicated by their species names. ρ Spearman rank correlation, r coefficient of correlation, t t test statistic

The analysis of independent contrasts was then performed using the `PDAP`-plugin (Midford et al. 2010) for `mesquite` (Maddison and Maddison 2001). Utilizing the `PDAP` diagnostic chart, we calculated the contrasts of η versus the positivized contrasts of genome size in gigabases and derived the Pearson product-moment correlation coefficient as well as a two tailed p value.

Total DNA methylation analysis in Orthoptera

Adult *Chorthippus parallelus* (meadow grasshopper) females were trapped in the vicinity of Leipzig (Saxony, Germany). Juvenile specimens of *Locusta migratoria* (migratory locust) and of *Acheta domestica* (house cricket) were used from commercial stocks. All animals were fed last time at least 24 h before use. Only heads and legs were selected for DNA preparation to avoid contamination with plant or bacterial DNA. RNA-free DNA was isolated using the DNeasy Blood & Tissue Kit (QIAGEN). Combined liquid chromatography and mass spectrometry analysis (LC/MS) of the samples were conducted on an Agilent 1200 analytical HPLC system coupled to an LTQ-Orbitrap XL mass spectrometer (Thermo Fisher Scientific). Two micrograms from at least five separate DNA extractions of each species were degraded to single nucleosides using 2 U of DNA degradase Plus (ZYMO research) according to the protocol of the provider. Standards contained equal amounts of all four nucleotides, supplemented with 0.25, 0.5, 1.0, 1.5, 5.0 or 10.0 % of 5-methylcytosine (Fermentas) and degraded to nucleosides by the same method as the probes. A positive control of herring sperm DNA (6.76 % mC/C) and four negative controls made from adult *Drosophila melanogaster* specimens (0.002 ± 0.005 % mC/C) were also run.

The samples were injected by an autosampler and separated on a reversed column (Jupiter C18, 2 mm × 250 mm, Phenomenex) using a solvent system of A (0.1 % formic acid in water) and B (0.1 % formic acid in 50 % methanol). The column was equilibrated with A and maintained at 100 % A for 5 min following injection, then a linear gradient from 0–50 % B during 9 min, followed by another linear gradient from 50–100 % B during 3 min was applied. The eluent was kept at 100 % B for 20 min. During separation the column was maintained at 25 °C and the flow rate was 200 µl/min. By applying a post-column flow split roughly 10 % of the eluate was introduced to the mass spectrometer and ionized by electrospray ionization

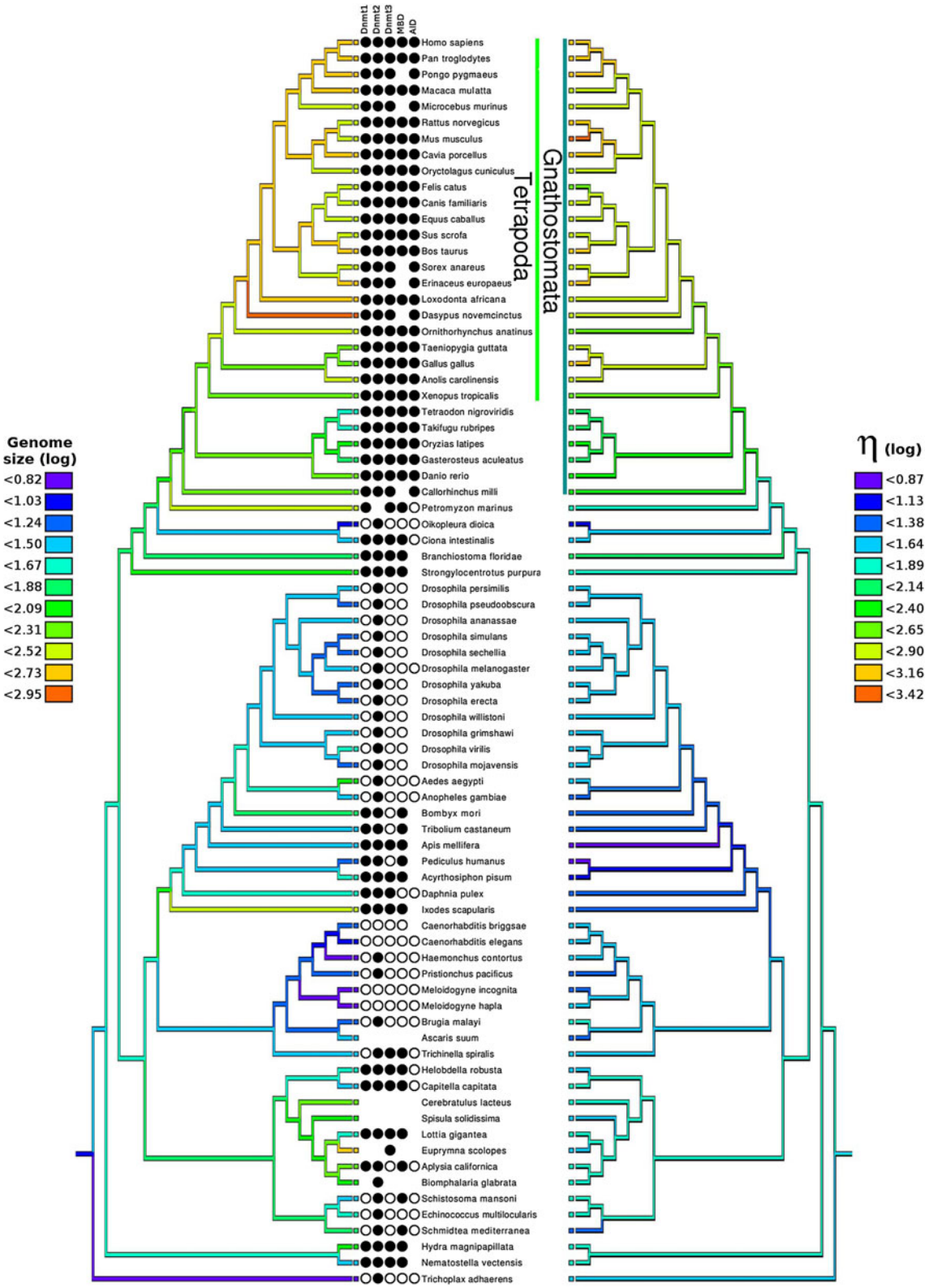

Fig. 2 Co-evolution of genome size and η. Data were log transformed and shifted into a similar numerical range to facilitate the comparison. The genome size is thus expressed as $2 + \log(m[pg])$, η as $\log \eta$. The phylogenetic distribution of components of the DNA methylation apparatus (Dnmt1-3, MBD, and AID) is indicated. *Filled circles* denote presence of a gene, *empty circles* designate its absence

(spray voltage 5 kV, sheath gas flow rate 25 a.u., capillary temperature 200 °C). The ion chromatograms of the nucleosides were acquired in SRM mode (selected reaction monitoring) in the linear ion trap, by monitoring the specific transitions of the nucleosides during their respective retention times (dC: *m/z* 228.1–112.1, 7–8.5 min, mdC: *m/z* 242–126.1, 11–13 min, dA: *m/z* 252.1–136.1, 14–15.5 min, dG: *m/z* 268.1–152.1, 15.5–17 min, dT: *m/z* 243.1–127.1, 18-19.5 min). The chromatographic peak areas were integrated using the `Genesis algorithm` (`Xcalibur` version 2.0.7, Thermo Fisher Scientific) and the methylation rate of cytosine was calculated based on the peak areas determined for cytosine and 5-methylcytosine.

Results

CpG depletion and TpG/CpA enrichment correlate with metazoan genome sizes

CpG-biased methylation causes a specific mutation pressure on CpG dinucleotides: CpG dinucleotides are depleted because cytosines in a CpG context are preferential methylated and mC frequently mutates to thymine (Simmen 2008). While CpG becomes depleted, TpG and its complement CpA are observed at an elevated level (Bird 1980; Hodges et al. 2009). This process causes an increase of, e.g., the human substitution rate of about one order of magnitude, and about a third of all SNPs between human cell lines (Kondrashov 2003; Shoemaker et al. 2010). Moreover, DNA methylation correlates also in rather weakly methylated insects with CpG depletion (Elango et al. 2009; Krauss et al. 2009; Walsh et al. 2010). The resulting signal can be used to investigate methylation in metazoan species for which mC data are not available. If methylation is related to genome size, we expected to observe a correlation of the CpG depletion and TpG + CpA enrichment ratio η with genome size.

We computed η for 78 metazoans (Fig. 1). To correct for the non-independence of our data from the phylogenomic position of the used species, we implemented an independent contrasts analysis. This method uses topology and branch lengths of a tree corresponding to the data ("Materials and methods") under the assumption of a Brownian motion model of character evolution. An independent contrasts analysis which comprises 76 of those 78 species ("Materials and methods") attests a significant Pearson product-moment correlation coefficient of 0.4 (p value 0.0004) between genome size and η (Table 1, Supplementary Data). An analysis without the seven outliers labeled in Fig. 1 noticeably improves the correlation. We comment on these exceptional genomes in "Discussion". In addition, all three subgroups of species containing taxa with genes coding for Dnmt1 methyltransferases show a positive correlation within the groups, albeit not as significant. In contrast, Metazoa without Dnmt1, revealing no species with cytosine methylation rates above 0.17 % (Supplementary Material 1), exhibit a significant negative correlation (r −0.57, p value 0.002). This data suggests that the positive correlation between η and genome size depends, indeed, on the methylation rate.

Table 1 Results of the independent contrast analysis concerning the correlation between η and genome size, Fig. 1

Group	n	n^*	mC/C in %	r	p
All	76	46	0–12.04	0.40	0.0004
Without outliers	69	40	0–12.04	0.84	0
Tetrapoda	23	23	2.25–10.90	0.40	0.057
Basal Deuterostomia	11	10	3.28–12.04	0.26	0.42
Non-Deuterostomia with Dnmt1	17	17	0.12–8.40	0.18	0.5
Metazoa without Dnmt1	27	0	0–0.17	−0.57	0.002

Given are the number of species in each group (n), the number of Dnmt1-expressing species (n^*), the range of measured total methylation rates of the group (mC/C in %), the pearson product-moment correlation coefficient (r) and the corresponding p value

All 30 species of jawed vertebrates (gnathostomes), the lancelet *Branchiostoma*, the sea urchin *Strongylocentrotus*, the nemertean *Cerebratulus* and the snails *Biomphalaria* and *Aplysia* exhibit a highly significant CpG depletion and a concurrent enrichment of TpG and CpA [$\eta > 1.58$ (Karlin and Cardon 1994)]. Given that species such as *Ciona intestinalis*, *Hydra magnipapillata* and *Nematostella vectensis* revealed mC/CpG rates between 10 and 30 % (Grunau et al. 2001; Zemach et al. 2010) and η values between 1.32 and 1.44, we can expect that within all genomes with $\eta \geq 1.32$ CG dinucleotides become methylated at least in part. This includes the majority of Chordata (32 of 33 species), Mollusca (5 of 5), Cnidaria (2 of 2), Annelida (2 of 2), Platyhelminthes (2 of 3), Echinodermata (1 of 1) and Nemertea (1 of 1), but excludes Nematoda (1 of 9), Arthropoda (0 of 21) and Placozoa (0 of 1). Hence, 46 species in 8 of the 10 metazoan phyla under consideration methylate a significant fraction of cytosines in CpG context at least in the germline. Cytosine methylation of large parts of the genome thus appears to be an ancient property of metazoans.

Figure 2 shows Gnathostomes (i.e., vertebrates excluding jawless ones such as *Petromyzon*) had evolved at even higher methylation rates. In contrast, several times during evolution, the extent of CpG methylation has been reduced: in arthropods (e.g., *Pediculus*), nematodes, urochordates (*Oikopleura*) and planarians (*Schmidtea*). With the exception of *Schmidtea*, genome sizes were reduced together with CpG methylation levels.

Evidence for strong germline methylation in transposons

The positive correlation of genome size with CpG methylation argues for a relatively high methylation rate of repetitive sequences. Thus, we wondered if cytosine methylation within invertebrates is strongly concentrated at coding sequences as recently suggested (Zemach et al. 2010). To evaluate the relative methylation rate of transposable elements, we compared η between transposons and whole genomes for 35 species using TE sequences extracted from RepBase (Jurka et al. 2005), Fig. 3. In invertebrates and fishes, similar values of η for TE and whole genome sequences support an equivalent DNA methylation level within TEs and in the remaining genome. In contrast, all tetrapods show a lower η in transposon-derived sequences than in the complete genome. This observation is only consistent with a preferential methylation of LTR and LINE elements in human cells (Meissner et al. 2008) as well as with the uniformly high methylation around repetitive DNA in mouse (Feng et al. 2010) if the value of η is not yet saturated by DNA methylation within the younger part of TE insertions.

Therefore, we investigated η as a function of the relative age of the TE insertions, determined by the amount of sequence divergence (see "Materials and methods"). Figure 4a shows, that η increases systematically with the age of TE insertions up to values above the genomic average value in mammals and birds (tetrapods). This is consistent with the report by Arndt and Hwa (2005) that the CpG mutation rate within repetitive elements increased substantially during the evolution of mammals. A much weaker increase was observed for some invertebrates, including *Strongylocentrotus purpuratus*, *Ciona savignyi*, *Nasonia vitripennis*, *Acyrthosiphon pisum* (all with Dnmt1) and *Schistosoma mansoni* (without Dnmt1). In contrast, no increase was found for the species *Bombyx mori*, *N. vectensis* (both with Dnmt1), *D. melanogaster* and *Caenorhabditis elegans* (both without Dnmt1).

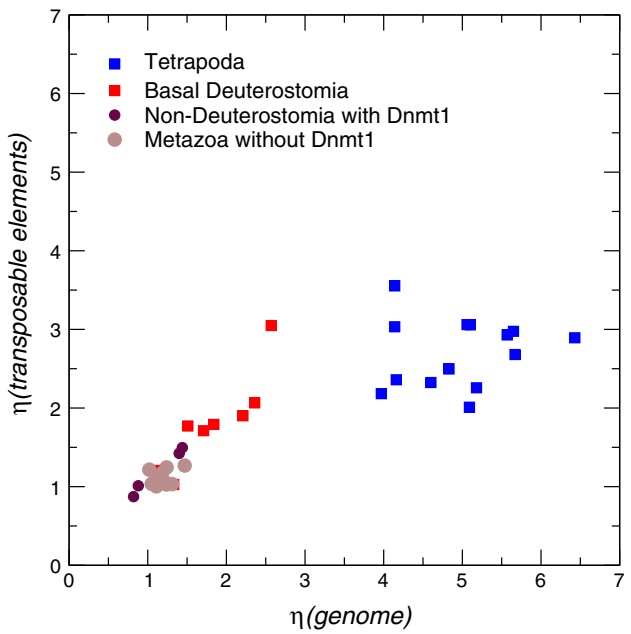

Fig. 3 η of transposable elements compared to genomic background. Metazoan transposable elements were extracted from RepBase (Jurka et al. 2005) for all species with more than 14 entries. Transposable elements of tetrapods exhibit a systematically reduced value of η

In green plants, transposons are specifically methylated at CpNpG sites (Zemach et al. 2010). CpNpG methylation was found also in mammal genomes, its distribution, however, remains unknown (Clark et al. 1995; Lister et al. 2009). To evaluate whether CpNpG methylation is localized similarly in animals and plants, we determined η_{CNG} values as function of TE insertion age, Fig. 4b. We found a weak, but significant rise of η_{CNG} with the age of TE insertions, well above the genomic η average, in basal chordates (Kendall Tau Rank Correlation Coefficient $\tau = 0.467$) and in invertebrates with Dnmt1 ($\tau = 0.46$). Intriguingly, tetrapods ($\tau = 0.087$) did not deviate strongly from other animals in that respect. In contrast, four species show large variations of η_{CNG} values in aged TE insertions: *N. vitripennis* (with Dnmt1), *Drosophila pseudoobscura*, *S. mansoni* and *C. elegans* (all without Dnmt1). Thus, a weak, but in most cases significant values of CpNpG depletion and TpNpG + CpNpA enrichment (η_{CNG}) were specifically found in animal transposon copies located in genomes with at least one active Dnmt. While we cannot exclude that other causes are responsible for these observations, we suspect that TE copies in most animals are CpG and CpNpG methylated in the germline.

Evidence that promoter-related CpG islands are restricted to Gnathostomata

CpG islands, initially described for mammals, were also found in birds, frogs and fishes, but not in invertebrate chordates as *Ciona* (Elango and Yi 2008). They form regions of low methylation intensity and are characterized by a much more moderate loss of CpG dinucleotides than all other parts of the genome. CpG islands are found most often at promoters because a strong methylation is incompatible with initiation of transcription in animals. Thus, local selection against methylation maintains CpG islands in globally strongly methylated genomes. Most effectively, this could be accomplished (1) by general methylation and (2) by local demethylation, as shown in fishes and mammals (Rai et al. 2008; Bhutani et al. 2010). Figure 1 reveals that higher mutation biases ($\eta > 2$) are

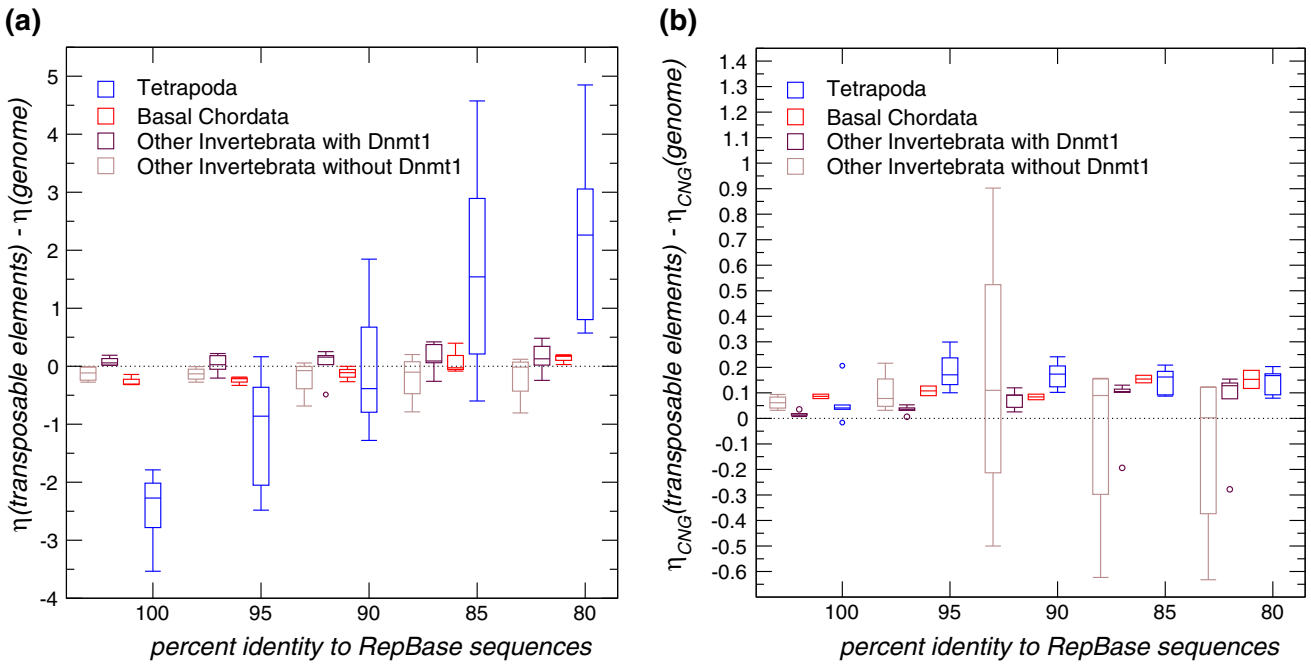

Fig. 4 Relative values of η and η_{CNG} of transposable elements compared to the genomic background shown as a function of the insertion age of TEs cyan in a *boxplot*. The relative ages of TE insertions are determined by the amount of divergence from the `RepBase` TE appropriate consensus sequences. The genomic background levels are marked by a *dashed line*

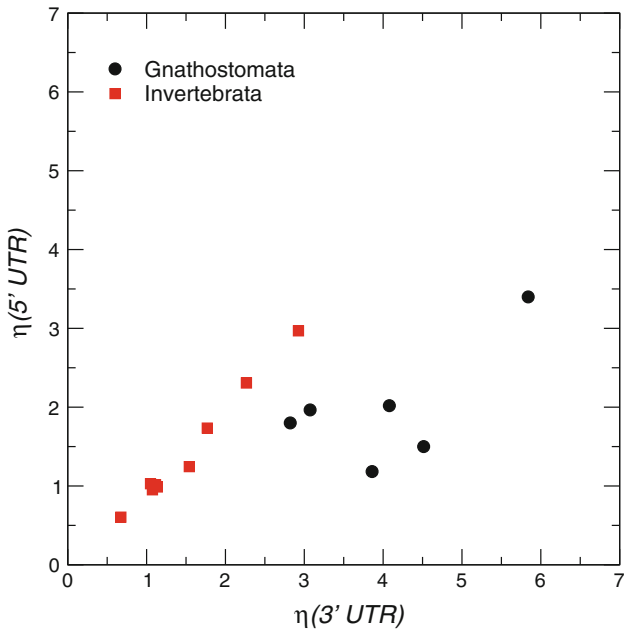

Fig. 5 Emergence of demethylated 5′UTRs in jawed vertebrates. η was calculated for 5′UTR and 3′UTR of coding genes in 13 species. In gnathostomes, a relatively low η of the 5′UTR (part of the promoter) contrasted with high η of the 3′UTR. Invertebrates show only small differences between both UTRs

restricted to jawed vertebrates (gnathostomes) including cartilaginous fishes (*Callorhinchus milii*). To evaluate the methylation status of promoters, we compared η of 5′UTRs, which represent a part of the promoter (FitzGerald et al. 2006), and in 3′UTRs (control) of 15 species, Fig. 5. Contrary to earlier expectations, η values were much higher in 3′UTRs than in 5′UTRs only in gnathostomes. In 5′UTRs, η was even lower in these species than in the analyzed invertebrates.

Unfortunately, no data to evaluate η in the UTRs of the lamprey *Petromyzon marinus* (a jawless vertebrate) and of the cartilaginous fish *C. milii* were available. To overcome this limitation, we investigated chordate genomes for the presence of orthologs of the activation-induced cytidine deaminase (AID), a necessary component of one of the cytosine demethylation pathways of vertebrates (Rai et al. 2008; Popp et al. 2010), which is only found in jawed vertebrates (including *Callorhinchus*), Fig. 2. Although two paralogs of the AID/APOBEC protein family were identified in lampreys, these are not orthologs of AID or APOBEC2 (Conticello et al. 2005; Rogozin et al. 2007; Conticello 2008). Only AID and APOBEC2, but not other paralogs from the AID/APOBEC protein family are known to be involved in demethylation (Rai et al. 2008). In addition, the methyl-specific transcription factors Kaiso and Zbtb38 (Sasai and Defossez 2010; Bogdanovic and Veenstra 2009) were also found only in gnathostomes (data not shown), supporting that differential promoter methylation as a pathway of gene regulation originated in the last common ancestor of the gnathostomes.

Total methylation appears to be related with genome sizes also in Orthopteran species

The genomes of the 78 sequenced metazoa have a size between 0.04 and 5.41 Gb, whereas the 4,972 metazoan genomes, whose sizes have been determined, vary between 0.019 and 129.9 Gb (Gregory 2010). Thus, there exists a bias to include preferentially animals with smaller genomes into the analysis. To confirm with independent data that (1) cytosine methylation is positively correlated to genome sizes and that (2) this correlation is not a taxonomic effect, we performed comparative measures of total methylation within a selected metazoan taxon. 19 of the bioinformatically analyzed genome sequences are from insects, a taxon that is well known for the small, sparsely methylated genomes of its best-studied species. Insect genomes studied so far contain typically <1 % mC (Regev et al. 1998; Marhold et al. 2004; Walsh et al. 2010; Zemach et al. 2010) and have $\eta < 1.32$. Therefore, we decided to measure total DNA methylation of selected species of the insect order Orthoptera. This highly diverse order contains many species with large genomes of highly variable sizes (1.516–16.56 Gb) (Gregory 2010).

We evaluated species whose genome sizes (1) were determined independently three or more times and (2) are highly diverged from each other (Gregory 2010). The mC content was determined by LC/MS and confirmed the correlation between the cytosine methylation rate and the genome size, Fig. 6: First, the house cricket *A. domestica* has the smallest genome size average of 2.03 ± 0.17 Gb and revealed a mC/C ratio of 0.28 ± 0.15 %. Second, the genome of the migratory locust *L. migratoria* is about three times larger (5.76 ± 0.48 Gb). Its methylation rate of 1.25 ± 0.79 % is significantly higher as well. Our measurements are consistent with the mC/C ratio of 0.96 % cited in (Regev et al. 1998), although we have not been able to backtrack the source of this number in the literature. Third, the genome of the meadow grasshopper *C. parallelus* belongs to the largest genomes of the group (13.26 ± 0.98 Gb). Its mC/C ratio of 4.06 ± 0.68 % is, in fact, comparable to that of mammals (4.32 ± 1.24 %). Fourth, we added recently determined LC/MS cytosine methylation from three different tissues and genome size data of the desert locust *Schistocerca gregaria*, Fig. 6 (Boerjan et al. 2011). Although these experiments comprise only four genomes, this observations indicates that the high mC/C values in vertebrates are not exceptional for metazoans and might be dependent on genome size.

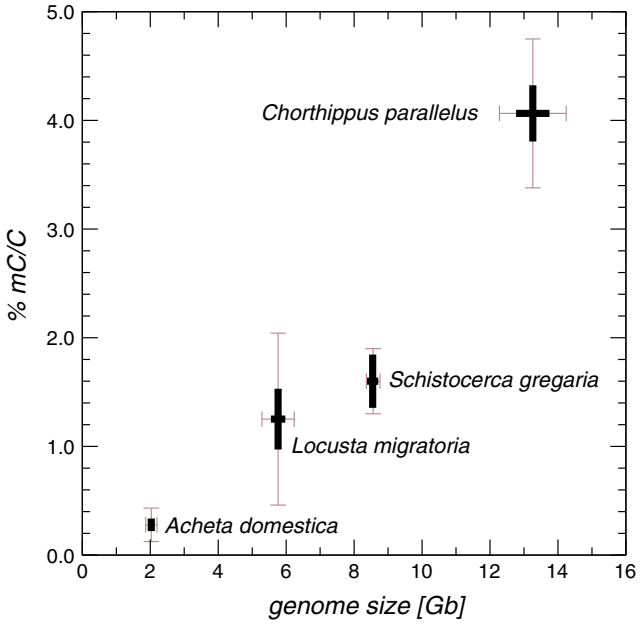

Fig. 6 Relationship between genome size and cytosine methylation rate in four species of orthopterans. *Error bars* indicate standard deviations of the sample distribution (*thin*) and standard deviations of the means (*thick*) of three to five measurements of genome size (Gregory 2010) and of seven to eight (this study) or three (Boerjan et al. 2011) cytosine methylation measurements using LC/MS

Discussion

Total methylation

Analysis of the genomic CpG depletion and TpG+CpA enrichment measure η in animals shows that the level of DNA methylation correlates positively with genome size. Dinucleotide data are available preferentially for smaller, already sequenced genomes. To overcome this limitation, we determined methylation rates for three species of a selected taxon with larger genome sizes and larger genome size differences (Orthoptera, Insecta). The results of this analysis are consistent with the predicted positive correlation of DNA methylation with genome size.

Promoters generally show the weakest methylation of all functionally defined elements of the genomes [see, e.g., (Zemach et al. 2010; Feng et al. 2010)]. An active demethylation of promoters as found at least in mammals mark the evolutionary emergence of regulative, differential methylation of functional promoters [for review see (Wu and Zhang 2010)]. The restriction of demethylated 5′UTRs, of one pathway of active demethylation and of some CpG methylation readers to jawed vertebrates suggest that this evolutionary novelty originated in the last common ancestors of the gnathostomes.

In some species, labeled as outliers in Fig. 1, genome sizes seem not to co-evolve with the level of DNA methylation. The genomes of *Ixodes*, *Euprymna*, and *Dasypus* appear to have expanded substantially in the recent past, see Fig. 2. It can be speculated that their DNA methylation machineries have not had sufficient time to adapt to the current genome size. In contrast, the genomes of birds shrank relative to other tetrapods, possibly due to a higher

deletion bias, as proposed for pufferfishes (Neafsey and Palumbi 2003). Corresponding genome-wide adaptations of the methylation rate might be limited by a local need for TE silencing and therefore should proceed more slowly than genome amplification or reduction. A second argument for the evolutionary stability of cytosine methylation in the birds *Gallus* and *Taeniopygia* is the importance of DNA methylation for developmental gene silencing in vertebrates (Elango and Yi 2008).

As depicted in Fig. 1 the genomes of the insects *Apis mellifera*, *A. pisum* and *Pediculus humanus* remarkably show a trend opposite to that obtained for vertebrates: CpG dinucleotides are enriched, while TpG and CpA are both depleted, i.e., $\eta \ll 1$ (Wang et al. 2006; Elango et al. 2009). In EST data of both species, however, all three types of dinucleotides show nearly the expected values, i.e., $\eta \approx 1$ (data not shown). For *Apis*, it was demonstrated repeatedly that exons are methylated preferentially, see, e.g., (Wang et al. 2006; Elango et al. 2009; Feng et al. 2010; Zemach et al. 2010). It may be speculate that the biased dinucleotide content is result of a context-dependent DNA repair mechanism that compensates for the mutational effect caused by CpG methylation. This could be achieved by repairing a T-G mismatch systematically to C-G and not to T-A provided the position 3′ of T is occupied by G. This would suppress mainly CpG→TpG substitutions in coding regions which are frequently methylated, while non-coding, typically unmethylated regions would accumulate CpG at the expense of TpG and CpA, explaining the unexpectedly small genome-wide value of η. Such a hypothetical mechanism would naturally explain the increase of η in exons.

To evaluate this possibility, we searched the proteomes of those insects to find candidate member proteins of such a pathway. Using reciprocal `blastp`, we identified an ortholog of thymine DNA glycosylase (TDG) in all examined insect species. This enzyme excises thymine from G-T mispairs with a preference for lesions in a CpG-T context (Morgan et al. 2007). Interestingly, *Drosophila* TDG is five to tenfold less effective than the human ortholog in this respect (Hardeland et al. 2003). The relative activities of TDG enzymes of other insects are unknown to date but might be responsible for the observed nucleotide biases. This bias is even more visible in AT-rich genomes, e.g., in *Apis* and *Pediculus* (AT content 67.3 and 68.5 %, respectively). Accordingly, twofold more TpG and CpA than CpG increases (1) the absolute TpG/CpA to CpG mutation rate and (2) their effect on η.

Methylation of TE sequences

We found evidence for methylation of repetitive elements in most analyzed genomes of Metazoans that express the DNA methyltransferase Dnmt1 (Fig. 4). Zemach et al. (2010) reported for *C. intestinalis*, *N. vectensis*, *A. mellifera*, and *B. mori* that methylation is weaker within repeats than in neighboring sequences. The authors conclude, based on data collected from six metazoan species only, that selective methylation of transposons is not conserved between plants and animals. Our results do not contradict this suggestion. However, we prefer not to exclude that preferential methylation of transposons is conserved between plants, fungi and vertebrates, as beyond our own results also substantial other evidence is consistent with a preferential methylation of TEs in at least some invertebrates. First, *Ciona* appears to contain at least a subset of repeats that are preferentially methylated (Zemach et al. 2010; Feng et al. 2010). Second, the general incompleteness of genome assemblies may be source of a substantial bias, given that the conclusions of Zemach et al. (2010) and Feng et al. (2010) are based only on mappable repetitive sequences. For *A. mellifera*, a mCpG rate of 3.09 % was measured for unassembled and unoriented contigs (comprising more than 18 % of CpGs in the genome), while chromosomal linkage groups exhibit an mCpG rate of only 0.93 % (Feng et al. 2010). Thus unmapped, typically repetitive, sequences may be more strongly methylated than mapped sequences. Third, all invertebrate species which became genome wide evaluated for DNA methylation patterns (*Acyrthosiphon*, *Apis*, *Bombyx*, *Ciona*, *Drosophila*, *Nematostella*, and *Tribolium*) have relatively small genome sizes (180−525 MB), compared to the analyzed vertebrates *Danio*, *Homo*, *Mus*, and *Tetraodon* (420−3,500 MB) (Gregory 2010).

Nevertheless, we found relatively low values of η for transposons, generally not exceeding the values obtained for the whole genomes of the corresponding species. These low η or η_{CNG} values within TE *consensus* sequences might be explained by the selection for activity, whereas the increase of the bias in older copies is consistent with a mutagenic effect of CpG and CpNpG methylation, respectively. Perhaps most interestingly, TE sequences of invertebrates with Dnmt1 (a main DNA methyltransferase) show higher η values than their genomic average, while TE sequences of invertebrates without Dnmt1 show exactly the opposite behavior (Fig. 4a) or large group-internal differences, depending on the organism measured (Fig. 4b). This is true, in particular, for later stages of mutational decay and supports germline cytosine methylation of transposons in all metazoans with DNA methylation apparatus. This does not necessarily imply, however, that transposons are preferentially methylated in these genomes because selection should be stronger in coding sequences and could have erased, at least partially, methylation-caused nucleotide sequence bias. In addition, Feuerbach et al. (2011) have recently found that the CpG mutation rate within AluSx SINE elements of primates depends mainly not on the age

of the copies but on the methylation status of the surrounding sequences. Thus, small TEs may escape DNA methylation also in strongly methylated genomes. While amplification of TEs might enlarge genomes and the methylation rate, not all TE copies need to be methylated to the same or a higher extent than the remaining genome. High-quality TE sequence data analyzed using available tools (Arndt et al. 2003; Feuerbach et al. 2011) may reveal a diversity of TE methylation patterns in the near future.

A hypothesis

We suggest that the correlation between genome size and the level of cytosine methylation might be based on their common dependence on long-term effective population size. Lynch (2006) pointed out that reductions of long-term effective population size are associated with dramatic expansions in genome size, most of which reflect changes in non-coding regions. This is why selection against weakly deleterious insertions is abolished in smaller populations. Estimations of the effective population size reviewed by Charlesworth (2009) show that population sizes of large mammals are between 10^4 and 4×10^4, whereas the population sizes of small invertebrates as *D. melanogaster* (10^6) and *Caenorhabditis remanei* (2×10^6) are much larger. We could not add more numbers for comparison as the estimates are methodically difficult, and thus rare.

However, the within-species nucleotide divergence for silent sites (Π_s), determined for at least 36 metazoan species using sampled alleles of protein-coding genes (Lynch 2006), is proportional to the product of long-term effective population size and mutation rate. It is difficult to relate Π_s to known methylation rates, however, because both numbers are known simultaneously only for four invertebrate species. Therefore, we plotted averages of ten groups containing at least two species with Π_s data (birds, fishes, mammals, nematodes, crustaceans, orthopterans, lepidopterans, dipterans, molluscs and invertebrate deuterostomes), see Fig. 7. The result is consistent with the proposed relationship. Additional, phylogenetically more balanced data will be necessary, however, to evaluate our hypothesis further.

Two model species deviate strongly from the suggested dependencies. First, wild mice have a population size near the reported invertebrate estimates (Eyre-Walker et al. 2002) but nevertheless maintained methylation rates similar to larger mammals. Here, the about 100-fold higher TE transposition rate (mouse versus humans) may play an important role (Maksakova et al. 2006). In addition, the Π_s value of the rat is congruent with those of other mammals. Thus, the divergence time might not have been sufficient to decrease the methylation rate of the mouse. Second, the high self-fertilizing rates of some nematodes, especially

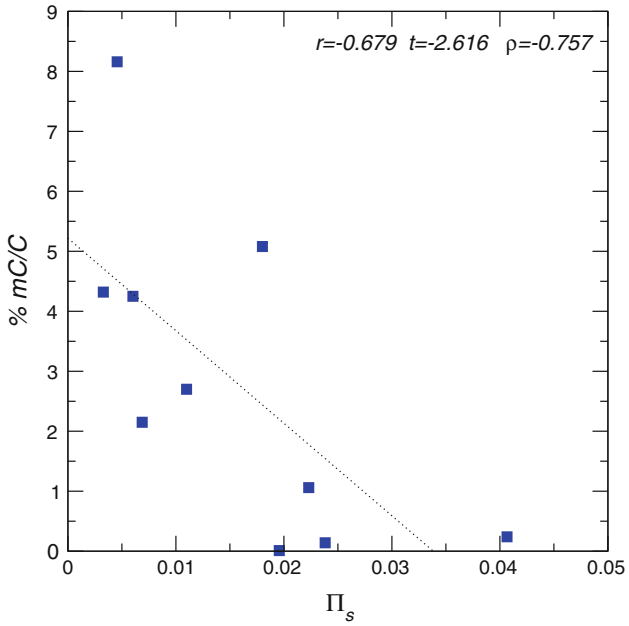

Fig. 7 The relationship between cytosine methylation rate and within-species nucleotide divergence for silent sites (Π_s) is correlated negatively. Compared are the averages of ten metazoan taxa. ρ Spearman rank correlation, r coefficient of correlation, t t test statistic

C. elegans, decrease both TE activity and effective population size (Charlesworth 2009; Hu et al. 2011), two factors that are expected to have opposite influences on genome size and methylation rate. Thus, differences in long-term effective population size, here represented by Π_s, are certainly not the sole cause for differences in genome size or methylation rate, but might have played an important role in establishing those. Genome size, in addition, is only a crude estimate for deleterious consequences of TE activity. Data on absolute TE abundances would represent a better measure. At present, however, such estimates are rare and unreliable (Feschotte et al. 2009).

DNA methylation within metazoan genomes is strongest in exons and repetitive elements. The function of DNA methylation is to inactivate transposons, stabilize repetitive sequences and suppress non-functional sites of transcriptional initiation (Yoder et al. 1997; Maunakea et al. 2010). In other words, DNA methylation acts mainly against deleterious DNA sequence variation that would not persist if purifying selection against such mutations would be successful. Because the efficiency of natural selection is directly dependent on long-term effective population size, the importance of DNA methylation increases with shrinking population size and with increasing genome size, which in turn is associated with increased retention rates of TEs and accelerated emergence of promoter-like sequences by local mutations. We predict that this hypothesis of "junk-masking" may explain only the broad patterns, as we

have found ample evidence of a stochastic distribution around the general expectations.

Intriguingly, if the proposed evolutionary mechanism is valid, the primary function of DNA methylation is likely a defense against the increasing deleterious effects caused by growing genomes, not the invention of an additional level of regulation. Such erroneous effects or functions are, e.g., disruptions of endogenous genes by TE activity, modulations of functional and non-functional transcription by TE-based regulative sequences, non-homologous recombinations at TE insertion sites and activations of cryptic promoters (promoter-like sequences) by mutations within gene structures overstretched by large introns. Our hypothesis is consistent with the suggestion that cytosine methylation constrains the effective size of the genome through the selective exposure of regulatory sequences remaining hypomethylated (Rollins et al. 2006). If genome and methylation rate had grown under the influence of a persistently low long-term effective population size, the entire genome may became transfigured by DNA methylation as in jawed vertebrates. Now a novel form of mitotically stable silencing of endogenous genes could arise in response to the emergence of a novel active demethylation pathway.

Acknowledgments This work was supported in part by *DFG GRK-1384 and MA5082/1-1*.

References

Arndt PF, Burge CB, Hwa T (2003a) DNA sequence evolution with neighbor-dependent mutation. J Comput Biol 10:313–322

Arndt PF, Hwa T (2005) Identification and measurement of neighbor-dependent nucleotide substitution processes. Bioinformatics 21:2322–2328

Arndt PF, Petrov DA, Hwa T (2003b) Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. Mol Biol Evol 20:1887–1896

Bhutani N, Brady JJ, Damian M, Sacco A, Corbel SY, Blau HM (2010) Reprogramming towards pluripotency requires AID-dependent DNA demethylation. Nature 463:1042–1047

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499–1504

Bird AP, Taggart MH, Smith BA (1979) Methylated and unmethylated DNA compartments in the sea urchin genome. Cell 17: 889–901

Boerjan B, Sas F, Ernst UR, Tobback J, Lemière F, Vandegehuchte MB, Janssen CR, Badisco L, Marchal E, Verlinden H, Schoofs L, Loof AD (2011) Locust phase polyphenism: Does epigenetic precede endocrine regulation? Gen Comp Endocrinol 173:120–128

Bogdanovic O, Veenstra GJC (2009) DNA methylation and methyl-CpG binding proteins: developmental requirements and function. Chromosoma 118:549–565

Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195–205

Chen WJ, Ortí G, Meyer A (2004) Novel evolutionary relationship among four fish model systems. Trends Genet 20:424–431

Choi JK, Bae JB, Lyu J, Kim TY, Kim YJ (2009) Nucleosome deposition and DNA methylation at coding region boundaries. Genome Biol 10:R89

Choy JS, Wei S, Lee JY, Tan S, Chu S, Lee TH (2010) DNA methylation increases nucleosome compaction and rigidity. J Am Chem Soc 132:1782–1783

Clark SJ, Harrison J, Frommer M (1995) CpNpG methylation in mammalian cells. Nat Genetics 10:20–27

Consortium DG (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218

Conticello S (2008) The AID/APOBEC family of nucleic acid mutators. Genome Biol 9:229

Conticello SG, Thomas CJ, Petersen-Mahrt SK, Neuberger MS (2005) Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. Mol Biol Evol 22:367–377

Duret L, Galtier N (2000) The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. Mol Biol Evol 17:1620–1625

Elango N, Hunt BG, Goodisman MA, Yi SV (2009) DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. Proc Natl Acad Sci USA 106:11206–11211

Elango N, Kim SH, Vigoda E, Yi SV (2008) Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. PLoS Comput Biol 4:e1000015

Elango N, Yi SV (2008) DNA methylation and structural and functional bimodality of vertebrate promoters. Mol Biol Evol 25:1602–1608

Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D (2002) Quantifying the slightly deleterious mutation model of molecular evolution. Mol Biol Evol 19:2142–2149

Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, Ukomadu C, Sadler KC, Pradhan S, Pellegrini M, Jacobsen SE (2010) Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci USA 107:8689–8694

Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D (2009) Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol Evol 1:205–220

Feuerbach L, Lyngsø RB, Lengauer T, Hein J (2011) Reconstructing the ancestral germline methylation state of young repeats. Mol Biol Evol 470:1777–1784

FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C (2006) Comparative genomics of *Drosophila* and human core promoters. Genome Biol 7:R53

Gertz EM, Yu YK, Agarwala R, Schäffer AA, Altschul SF (2006) Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol 4:41

Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. Annu Rev Biochem 74:481–514

Gregory TR (2010) Animal genome size database. http://www.genomesize.com/

Grunau C, Renault E, Rosenthal A, Roizes G (2001) MethDB–a public database for DNA methylation data. Nucleic Acids Res 29:270–274. http://www.methdb.de

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Hardeland U, Bentele M, Jiricny J, Schär P (2003) The versatile thymine DNA-glycosylase: a comparative characterization of the human, *Drosophila* and fission yeast orthologs. Nucleic Acids Res 31:2261–2271

Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguñà J, Bailly X, Jondelius U, Wiens M, Müller WEG, Seaver E, Wheeler WC, Martindale MQ, Giribet G, Dunn

CW (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci 276:4261–4270
Ho KL, McNae IW, Schmiedeberg L, Klose RJ, Bird AP, Walkinshaw MD (2008) MeCP2 binding to DNA depends upon hydration at methyl-CpG. Mol Cell 29:525–531
Hodges C, Bintu L, Lubkowska L, Kashlev M, Bustamante C (2009) Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. Science 325:626–628
Holterman M, van der Wurff A, van den Elsen S, van Megen H, Bongers T, Holovachov O, Bakker J, Helder J (2006) Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades. Mol Biol Evol 23:1792–1800
Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottilar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KFX, Peer YVD, Grigoriev IV, Nordborg M, Weigel D, Guo YL (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet 43:476–481
Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467
Karlin S, Cardon LR (1994) Computational DNA sequence analysis. Annu Rev Microbiol 48:619–654
Kondrashov AS (2003) Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mutat 21:12–27
Krauss V, Eisenhardt C, Unger T (2009) The genome of the stick insect *Medauroidea extradentata* is strongly methylated within genes and repetitive DNA. PLoS ONE 4:e7223
Krauss V, Reuter G (2011) DNA methylation in *Drosophila*—a critical evaluation. Prog Mol Biol Transl Sci 101:177–191
Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, Schmitz J (2006) Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol 4:e91
Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J, Wei CL (2010) Dynamic changes in the human methylome during differentiation. Genome Res 20:320–331
Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315–322
Lynch M (2006) The origins of eukaryotic gene structure. Mol Biol Evol 23:450–468
Lynch M (ed) (2007) The origins of genome architecture. Sinauer Associates, Sunderland
Maddison WP, Maddison DR (2001) Mesquite: a modular system for evolutionary analysis. http://mesquiteproject.org
Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, Mager DL (2006) Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet 2:e2
Marhold J, Rothe N, Pauli A, Mund C, K K, Brueckner B, Lyko F (2004) Conservation of DNA methylation in dipteran insects. Insect Mol Biol 13:117–123
Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen CB, Zhao Y, Turecki G, Delaney A, Varhol R, Thiessen N, Shchors K, Heine VM, Rowitch DH, Xing X, Fiore C, Schillebeeckx M, Jones SJM, Haussler D, Marra MA, Hirst M, Wang T, Costello JF (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466:253–257
Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454:766–770
Midford PE, Garland T, Maddison WP (2010) PDAP:PDTREE: a translation of the PDTREE application of Garland et al.'s phenotypic diversity analysis programs, version1.14
Molaro A, Hodges E, Fang F, Song Q, McCombie WR, Hannon GJ, Smith AD (2011) Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell 146:1029–1041
Morgan MT, Bennett MT, Drohat AC (2007) Excision of 5-halogenated uracils by human thymine DNA glycosylase. robust activity for DNA contexts other than cpg. J Biol Chem 282: 27578–27586
Nanty L, Carbajosa G, Heap GA, Ratnieks F, van Heel DA, Down TA, Rakyan VK (2011) Comparative methylomics reveals gene-body H3K36me3 in *Drosophila* predicts DNA methylation and CpG landscapes in other invertebrates. Genome Res 21:1841–1850
Neafsey DE, Palumbi SR (2003) Genome size evolution in pufferfish: a comparative analysis of diodontid and tetraodontid pufferfish genomes. Genome Res 13:821–830
Okamura K, Matsumoto KA, Nakai K (2010) Gradual transition from mosaic to global DNA methylation patterns during deuterostome evolution. BMC Bioinforma 11(Suppl 7):S2
Popp C, Dean W, Feng S, Cokus SJ, Andrews S, Pellegrini M, Jacobsen SE, Reik W (2010) Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature 463:1101–1115
Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35:7188–7196
Rai K, Huggins IJ, James SR, Karpf AR, Jones DA, Cairns BR (2008) DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45. Cell 135:1201–1212
Regev A, Lamb M, Jablonka E (1998) The role of DNA methylation in invertebrates: Developmental regulation or genome defense? Mol Biol Evol 15:880–891
Robinson KL, Tohidi-Esfahani D, Lo N, Simpson SJ, Sword GA (2011) Evidence for widespread genomic methylation in the migratory locust, *Locusta migratoria* (Orthoptera: Acrididae). PLos One 6:e28167
Rogozin IB, Iyer LM, Liang L, Glazko GV, Liston VG, Pavlov YI, Aravind L, Pancer Z (2007) Evolution and diversification of lamprey antigen receptors: evidence for involvement of an AID-APOBEC family cytosine deaminase. Nat Immunol 8:647–656
Rollins RA, Haghighi F, Edwards JR, Das R, Zhang MQ, Ju J, Bestor TH (2006) Large-scale structure of genomic methylation patterns. Genome Res 16:157–163
Sasai N, Defossez PA (2010) Many paths to one goal? The proteins that recognize methylated DNA in eukaryotes. Int J Dev Biol 53:323–334
Shoemaker R, Deng J, Wang W, Zhang K (2010) Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res 20:883–889
Simmen MW (2008) Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. Genomics 92:33–40
Simmen MW, Leitgeb S, Charlton J, Jones SJ, Harris BR, Clark VH, Bird A (1999) Nonmethylated transposable elements and methylated genes in a chordate genome. Science 283:1164–1167

Suzuki MM, Bird AP (2008) Dna methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9:465–476

Thalhammer A, Hansen AS, El-Sagheer AH, Brown T, Schofield CJ (2011) Hydroxylation of methylated CpG dinucleotides reverses stabilisation of DNA duplexes by cytosine 5-methylation. Chem Commun Camb 47:5325–5327

The Human Genome International Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Varriale A, Bernardi G (2006a) DNA methylation and body temperature in fishes. Gene 385:111–121

Varriale A, Bernardi G (2006b) DNA methylation in reptiles. Gene 385:122–127

Walsh CP, Bestor TH (1999) Cytosine methylation and mammalian development. Genes Dev 13:26–34

Walsh TK, Brisson JA, Robertson HM, Gordon K, Jaubert-Possamai S, Tagu D, Edwards OR (2010) A functional dna methylation system in the pea aphid, *Acyrthosiphon pisum*. Insect Mol Biol 19 Suppl 2:215–228

Wang Y, Jorda M, Jones PL, Maleszka R, Ling X, Robertson HM, Mizzen CA, Peinado MA, Robinson GE (2006) Functional CpG methylation system in a social insect. Science 314:645–647

Wu SC, Zhang Y (2010) Active DNA demethylation: many roads lead to rome. Nat Rev Mol Cell Biol 11:607–620

Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13:335–340

Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328:916–919