J Mol Evol (2014) 78:217–233
DOI 10.1007/s00239-014-9612-5

ORIGINAL ARTICLE

Molecular Evolution and Functional Divergence of the Metallothionein Gene Family in Vertebrates

Nina Serén · Scott Glaberman · Miguel A. Carretero · Ylenia Chiari

Received: 19 November 2013 / Accepted: 1 February 2014 / Published online: 21 February 2014

Abstract The metallothionein (MT) gene superfamily consists of metal-binding proteins involved in various metal detoxification and storage mechanisms. The evolution of this gene family in vertebrates has mostly been studied in mammals using sparse taxon or gene sampling. Genomic databases and available data on MT protein function and expression allow a better understanding of the evolution and functional divergence of the different MT types. We recovered 77 MT coding sequences from 20 representative vertebrates with annotated complete genomes. We found multiple MT genes, also in reptiles, which were thought to have only one MT type. Phylogenetic and synteny analyses indicate the existence of a eutherian MT1 and MT2, a tetrapod MT3, an amniote MT4, and fish MT. The optimal gene-tree/species-tree reconciliation analyses identified the best root in the fish clade. Functional analyses reveal variation in hydropathic index among protein domains, likely correlated with their distinct flexibility and metal affinity. Analyses of functional divergence identified amino acid sites correlated with functional divergence among MT types. Uncovering the number of genes and sites possibly correlated with functional divergence will help to design cost-effective MT functional and gene expression studies. This will permit further understanding of the distinct roles and specificity of these proteins and to properly target specific MT for different types of functional studies. Therefore, this work presents a critical background on the molecular evolution and functional divergence of vertebrate MTs to carry out further detailed studies on the relationship between heavy metal metabolism and tolerances among vertebrates.

Keywords CDS · Functional analysis · Gene duplication · Gene tree · Genomic database · Reconciliation

Scott Glaberman—See Disclaimer.

Electronic supplementary material The online version of this article (doi:10.1007/s00239-014-9612-5) contains supplementary material, which is available to authorized users.

N. Serén · M. A. Carretero · Y. Chiari
CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos Campus Agrário de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

S. Glaberman
U.S. Environmental Protection Agency, Office of Pesticide Programs (MC-7507P), Washington, DC 20460, USA

Y. Chiari (✉)
Department of Biology, University of South Alabama, LSCB 123, 5871 USA Dr. N, Mobile, AL 36688, USA
e-mail: yle@yleniachiari.it

Introduction

Gene families are a set of genes sharing sequence, and often functional, homology. The evolution of gene families is considered an important driver of species evolution (Ohno 1970; Demuth et al. 2006 and references therein). Gene families mainly evolve as a result of duplication and loss events, often associated with gain of adaptive function through neofunctionalization or subfunctionalization (e.g., Chang and Duda 2012; Kondrashov 2012; Zhang 2003). In neofunctionalization, duplicated genes may undergo an accelerated rate of mutation in one of the recently duplicated copies when freed from selective constraints, potentially leading to new function (e.g., Ohno 1970; Zhang et al. 1998). In subfunctionalization, paralogs gradually take on multiple functions once maintained by the original single copy gene, leading to specialized genes with no overlapping functions (e.g., Force et al. 1999;

Prince and Pickett 2002; see also Kondrashov et al. 2002). Independently of the underlying mechanisms driving the process, functional divergence of genes following duplication may promote increased gene diversity and novel gene functions (Kondrashov et al. 2002; Prince and Pickett 2002; Zhang 2003), possibly facilitating organismal adaptation to environmental conditions (e.g., Brown et al. 1998; Lenormand et al. 1998; Kondrashov 2012).

The metallothionein (MT) gene superfamily is known for its high turnover through gene duplication and loss (Capdevila and Atrian 2011). In some mammals, such as mice and humans, the presence of multiple MT genes has been associated with varying levels of gene expression, as well as gains and losses of function (e.g., Garrett et al. 1998; Moleirinho et al. 2011; Tío et al. 2004; reviewed in Blindauer and Leszczyszyn 2010). MTs are ubiquitous low-molecular weight proteins and polypeptides of extremely high metal and sulfur content (Nordberg and Nordberg 2009). These inducible proteins exhibit essential metal-binding properties and have several roles in metabolism, homeostasis, and kinetics of metals such as transport, storage and detoxification of metal ions in cells (e.g., Carpenè et al. 2007; Nordberg and Nordberg 2009; Palmiter 1998). Metal affinity varies among the different MT types (Nordberg 1989). MTs have also recently gained attention in biomedical studies, due to their proposed involvement in cancer or neurological diseases (reviewed in Hidalgo et al. 2009). Despite this increasing understanding of the physiological properties and function of MTs, very little is known about the intra- and inter-specific variation of these genes in vertebrates. It has been shown in invertebrates that both duplication events and molecular changes in the regulatory and coding regions of MT genes produce species-specific differences in terms of expression, tertiary structure, metal-binding affinity, and metal resistance (Dallinger et al. 1997; Shaw et al. 2007; reviewed in Dallinger and Höckner 2013).

Although MT genes are widely represented across all three domains of life, they have not been equally studied in different taxonomic groups. In vertebrates, a large number of biochemical, molecular, and chemical studies have been carried out on these multi-functional genes in mammals (mostly in humans and mice), which contrasts with the very limited data available for other groups (reviewed in Blindauer and Leszczyszyn 2010; Hidalgo et al. 2009). Different MT classification systems have been developed for all organisms based on protein structure (e.g., Fowler et al. 1987; Nordberg and Kojima 1979; Palacios et al. 2011; Valls et al. 2001; Vašák and Armitage 1986; reviewed in Nordberg and Nordberg 2009) or using both protein structure and phylogenetic relationships (Binz and Kägi 1999; Moleirinho et al. 2011). An earlier classification, based on protein structure, divides MTs into three classes, including proteinaceous MTs closely related to those in mammals (called class I), proteinaceous MTs that lack this close resemblance (class II), and non-proteinaceous MT-like polypeptides (class III) (Fowler et al. 1987; reviewed in Blindauer and Leszczyszyn 2010; Miles et al. 2000). However, the currently adopted classification scheme in available genomic and protein databases of MT genes are primarily based on the phylogenetic relationships among mammalian MT sequences, which are known for their great functional diversity (e.g., Capdevila and Atrian 2011; Vašák and Meloni 2011). According to this classification system, MTs fall into at least four subgroups: MT1, MT2, MT3, and MT4, which are generally differentially expressed and induced, and show diverse metal-binding affinities (e.g., Tío et al. 2004; reviewed in Davis and Cousins 2000; Miles et al. 2000; Vašák and Meloni 2011). There is a partial correspondence shown between the early classification and the more recent mammalian MT subgroups (e.g., MT1 and 2 to class I) (reviewed in Palacios et al. 2011). Recent studies suggest that the later classification system based on mammalian MTs is unsuited for molecular evolution studies at a taxonomically large scale due to differences in underlying physiologies between mammals and other organisms (reviewed in Capdevila and Atrian 2011). For example, studies on other taxonomic groups, including plants and bacteria, underline a departure from the classical mammalian (humans and mice) amino acid composition, biochemical metal-binding characteristics, and protein folding (reviewed in Vašák and Meloni 2011; see also Villarreal et al. 2006).

The increasing availability of annotated genomic databases provides an incredible resource to study functional diversification and evolution of genes and gene families (e.g., Koonin 2009; Yanai et al. 2000). In vertebrates, available gene and protein databases have recently been used to infer MT gene family origin and evolution through comparative analyses (e.g., Moleirinho et al. 2011; Guirola et al. 2012; Trinchella et al. 2008, 2012). However, these studies have been based either on sparse taxon or gene sampling when species with fully sequenced genomes have been considered (e.g., Moleirinho et al. 2011) or they have been based on cDNA or protein data (Guirola et al. 2012; Trinchella et al. 2008, 2012). In the latter case, proper distinction between different genes versus different isoforms cannot be assessed without genomic sequencing of the target gene or comparison of untranslated regions (UTRs). In addition, for cDNA data, all existing genes cannot be easily detected if they are not expressed, for example, as a result of a lack of response to metal treatment or differential expression in time and space. This could produce misleading estimates of duplication and loss events and of the evolutionary history and functional divergence of the MT gene family.

In this study, we use annotated complete genomes from 20 representative species spanning the major vertebrate

taxonomic groups (mammals, birds, fish, reptiles, and amphibians) to complement the current knowledge of the molecular evolution and functional divergence of the MT gene family in vertebrates. The specific objectives of this study are to: (1) identify actual MT genes in many different vertebrate species, (2) estimate the number of MT genes in vertebrates, (3) infer the root of the MT gene tree in vertebrates, (4) infer possible different selective pressures among the main MT types, and (5) study functional divergence among MT types and identify amino acid sites potentially correlated with this difference. These analyses, combined with the high quality of the genomic dataset used in this study, provide a critical molecular evolutionary basis for examining the functional significance of MT genes in important physiological processes.

Materials and Methods

Dataset Assembly and Characteristics

The dataset was initially constructed based on MT genes of representative vertebrate taxonomic groups and species retrieved from the Ensembl 68 database (Flicek et al. 2011) (data collected on September 25, 2012). Representative vertebrates were selected to have data for each of the major higher taxonomic groups (mammals, birds, fish, reptiles, and amphibians) and when possible for multiple species with annotated genomes in each of these groups (Online Resource 1). MT genes for all selected species were first retrieved using Ensembl Comparative Genomics search tool for orthologs and paralogs of all human identified functional MT genes (MT1A, MT1B, MT1E, MT1F, MT1G, MT1H, MT1M, MT1X, MT2A, MT3, and MT4) except for MT5, which is testis specific and was not included in our analysis. The obtained genes were double checked using the BLAT tool on UCSC Genome Bioinformatics (Kent 2002), the BioMart, BLAST/BLAT tools of Ensembl, the NCBI genomic database (data collected on January 29, 2013), and the Ensembl 70 release (data checked on January 28, 2013). Sequences were further checked to account for annotation discrepancies between databases. For sequences showing partial poor annotation in both the Ensembl and NCBI databases, information from the two databases was combined (Online Resources 1 and 2).

To properly identify the CDS (coding sequence) products of distinct genes rather than distinct transcripts of the same gene, exon and intron sequences and lengths were compared. Only CDS corresponding to different genes (distinct exons and intron sequences) were retained for our analyses. Furthermore, for genes with multiple CDS, only the ones coding for a product above 50 and below 70 amino acids were retained, which is in accordance with the characteristic amino acid length of MT proteins in vertebrates. When this parameter was matched by more than one transcript, we maintained only the transcript tagged as CCDS, based on the Consensus CDS project (Pruitt et al. 2009). A complete list of species, CDS (with relative Ensembl and NCBI accession numbers), chromosome location when available, and intron/exon gene characteristics of the CDS used for this work are indicated in Online Resource 1. The final dataset contained ten mammals (*Bos taurus, Canis lupus familiaris, Equus caballus, Homo sapiens, Monodelphis domestica, Mus musculus, Ornithorhynchus anatinus, Pan troglodytes, Rattus norvegicus,* and *Sus scrofa*), three birds (*Gallus gallus*, *Meleagris gallopavo*, and *Taeniopygia guttata*), two reptiles (*Anolis carolinensis* and *Pelodiscus sinensis*), one amphibian (*Xenopus tropicalis*) and four fish (*Danio rerio, Oryzias latipes, Takifugu rubripes*, and *Tetraodon nigroviridis*).

Sequence alignment was carried out on CDS nucleotide sequences in MEGA 5 (Tamura et al. 2011) using the Clustal W option. The alignment was further checked by eye. Terminal stop codons were removed from all CDS prior to analyses. Nucleotide alignment of the dataset used in this work can be accessed on www.researchgate.net/publication/260137321_MT_alignment?ev=prf_pub. The absolute number of variable and conserved nucleotide and amino acid sites was calculated in MEGA.

Phylogenetic Analyses

Prior to phylogenetic analyses, the degree of saturation was estimated for all the codon positions together and for the third codon position alone in DAMBE 5.3 (Xia 2013), since the presence of substitution saturation in the data, if not taken into account, may produce misleading phylogenetic results (e.g., Chiari et al. 2012). The estimate of the degree of saturation present in the dataset was based on the comparison between the index of substitution saturation (ISS), calculated from the data and a critical value (ISS.c) at which the sequence signal fails to recover the true tree. The calculation was performed under different topologies (symmetrical and asymmetrical); if ISS was not recovered to be smaller than ISS.c, this was interpreted as an indication of substitution saturation in the dataset (see Xia et al. 2003 for further details).

Phylogenetic analyses were run on the nucleotide and amino acid datasets. Maximum likelihood (ML) analysis was performed in PhyML 3.0 (Guindon et al. 2010). ML analysis on CDS was carried out with a K80+G substitution model (tr/tv = 1.7917; gamma shape = 0.7250; proportion of invariable sites = 0) as estimated by the AICc (Akaike information criterion corrected for finite sample

size) in jModeltest2 (Darriba et al. 2012; Guindon and Gascuel 2003), to account for the small size of the dataset used. ML analysis on amino acid data was carried out with FLU substitution model as estimated by the AICc in Prottest3 (Darriba et al. 2011) using PhyML to estimate the gamma factor and the proportion of invariable sites. ML analysis was run with 1,000 bootstrap replicates for both nucleotide and amino acid data. ML analyses were also repeated using the fast approximation to the likelihood ratio test using the aLRT method (Anisimova and Gascuel 2006) and the SH-branch support, as implemented in PhyML.

Bayesian analyses were carried out on the nucleotide and amino acid datasets in MrBayes 3.2. (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). Bayesian analyses on the nucleotide dataset were run by applying (1) the same model of evolution to all codon positions or (2) a two-partitioned mixed model (1st+2nd codon and 3rd codon positions). The partitioned mixed model was applied to our dataset as an alternative model of substitution to take into account the higher substitution rate of the third codon position compared to the first and second. For these analyses, we used number of substitutions (n_{st}), proportion of invariable sites (p_{invar}), and rates according to jModeltest2. In the analysis with the partitioned mixed model we used $n_{st} = 6$, gamma = equal, and $p_{invar} = 0$ for the 1st+2nd codon position, and $n_{st} = 6$, rates = gamma, and $p_{invar} = 0$ for the 3rd codon position. The Bayesian amino acid analysis was run with rates = gamma and $p_{invar} = 0$. The other parameters were left to be estimated by MrBayes. Phylogenetic analyses were run on nucleotide and amino acid datasets and using a partitioned mixed model (Bayesian analysis) to further take into account potential saturation occurring in the data not detected by the saturation test.

All Bayesian analyses were performed with two runs, each with four Markov chain Monte Carlo (MCMC) chains (one cold and three hot). The analyses were run for 50 million generations to allow the standard deviation of split frequencies to reach a value below 0.01. Trees and associated model parameters were sampled every 1,000 generations. The first 25 % of the obtained trees were discarded as burnin and the 50 % majority-rule Bayesian consensus retained. Admixture and convergence of chains and runs were checked with Tracer v1.5 (Rambaut and Drummond 2009). To compare the best model of evolution used for the Bayesian analysis of the nucleotide data, we calculated the Bayes factor for the two distinct models (single or partitioned mixed model). The Bayes factor value (K) was calculated by the ratio between harmonic means (average for all runs) of likelihoods for the models/tree topology comparison (e.g., A and B, see Nylander et al. 2004 for further information). An A/B ratio of $K > 1$ indicates that the A model is more strongly supported than B, while a value of $K < 1$ indicates the opposite. A value of $K = 1$ suggests that the difference between the two models is not important. Bayes factor was also used to compare distinct tree topologies obtained by the nucleotide and amino acid Bayesian analyses. In this case, the Bayes factor was given by the ratio of the harmonic means obtained by the Bayesian analyses on the nucleotide data run without any input tree (Bayesian analysis ran as described above) or by giving a prior on the tree topology to constrain the monophyly of tetrapod MT3 clade and tetrapod MT3—amniote MT4 (as recovered in the Bayesian analysis run on the amino acid data).

Synteny Analysis

Synteny analysis was performed using Genomicus version 72.01 (Muffato et al. 2010; Louis et al. 2012) by searching neighboring genes of representative MTs for mammals and fish: the MT3 (ENSMUSG00000031760) from *M. musculus* and the MT2 (ENSDARG00000041623) from *D. rerio*, respectively (genes with Ensembl MT nomenclature). Choosing different MTs as input did not change the output results. Genomicus returned a comparative alignment of the MT genes on a chromosome and their neighboring genes for tetrapods and fish separately, according to the two searches. Ortholog and paralog genes are identified in Genomicus according to the Ensembl annotation. Genomicus alignment criteria are based on pairwise comparison between species to identify pairwise synteny blocks assuming that order of genes on the chromosome reflects accurately the order and orientation of genes in their last common ancestor (Louis et al. 2012). Because Genomicus relies on Ensembl gene annotation, for MT and neighboring genes that were not retrieved in some of the species (see Online Resources 1 and 2), we manually searched the genome of these species on NCBI (*O. anatinus, P. sinensis, S. scrofa, X. tropicalis*).

Reconciliation Analysis

To infer the best root for the vertebrate MT tree, we carried out a reconciliation analysis of the gene and species trees (see Doyon et al. 2011 for a review). The reconciliation was performed by a parsimony-based approach as implemented in NOTUNG 2.6. (Chen et al. 2000; Durand et al. 2006; Vernot et al. 2008) using an unrooted gene tree with multifurcations (uncertainties) and a constructed binary species tree. A species tree including all the species in our dataset was built based on Chiari et al. (2012), Li et al. (2007), and the Tree of Life Web Project (2012). The gene tree used corresponded to the one obtained from the Bayesian analysis run on the nucleotide dataset with one

model of evolution. To test for the influence of alternative gene tree reconstruction on the reconciliation results, the analysis was also repeated using the phylogenetic tree obtained from the Bayesian analysis run on the amino acid dataset. For the reconciliation analyses, we chose the default parameters and the posterior probability values as obtained from the Bayesian analyses for the trees used (edge weight threshold/posterior probability values (pp) = 0.9, duplication = 1.5, loss = 1.0). The edge weight threshold identifies nodes that are not supported with a posterior probability equal or greater than 0.9 (chosen threshold for this study) and that can be rearranged during the reconciliation. This allows obtaining the optimal reconciliation considering different tree topologies from the ones used as the input at nodes with support below the threshold value. The cost/weight of gene loss was considered lower than the one for duplication so that losses may occur more frequently than duplications in the inferred reconciliation. This allows accounting for possible non-sequenced or non-retrieved data in our dataset. Because there may be many possible reconciliations of a gene tree within a species tree, the optimal reconciliation corresponds to the one with the lower cost of duplication and loss (see Doyon et al. 2011 for further information).

Analysis of Variation in Selective Pressure Among MT Types

To estimate the possible variation in selective pressure among main MT clades (defined according to the results of the phylogenetic analyses, eutherian MT1 and MT2, eutherian or tetrapod MT3, and amniote MT4, Fig. 1; fish clade excluded from this analysis), we applied a model of coding sequence evolution allowing variation of the selective pressure among branches. Selective pressure is calculated by comparing synonymous (d_S) versus non-synonymous (d_N) substitution rates. Synonymous substitutions are silent substitutions as they do not involve an amino acid change, differently from non-synonymous substitutions. The analysis of variation in selective pressure among main MT clades was performed with the codeml program in the PAML 4.7 package (Yang 2007). The analyses were performed both on the unrooted Bayesian nucleotide (one model of evolution) and amino acid trees to take into account the influence of alternative tree topologies on the selection pressure results. This ML-based analysis can estimate different ω (d_N/d_S) within the tree by letting the user apply different weights of selective pressure among evolutionary lineages. The parameter ω is, therefore, first estimated by running the model with one single ω across all lineages (model = 0 option; hypothesis H = 0), and then by allowing the program to estimate from the data distinct ω parameters for chosen clades (model = 2 option; alternative hypotheses). This permits testing the different selection rates among branches by assigning different ω estimates to these branches. Branch lengths and transition/transversion are also estimated separately for each run. Both ambiguity characters and alignment gaps were treated as undetermined nucleotides (option Cleandata = 0). Runs were carried out with one single ω across sites (option $N_{sites} = 0$). To test for convergence of the runs, several simulations were run with multiple initial starting values of ω (0.2 and 2) in H0 (hypothesis with one single ω across the tree, see below). For the analysis run using the Bayesian nucleotide tree, we also tested the influence of different kappa on H0 (k = transition/transversion rate; $k = 2.0241$ obtained with $\omega = 0.2$ and $k = 3.5834$ as previously calculated in jModeltest2). Alternative hypotheses (H1–H3, Table 1a) were subsequently formulated to test for different selective pressure, ω, between branches. Tests were aimed at evaluating whether there was a difference in the mutation rates among the main MT clades (Table 1), to assess whether they evolved under similar selective pressures. To statistically compare the different evolutionary hypotheses, we applied the LRT (Likelihood Ratio Test), which is a statistical test based on the likelihood ratio between the null and alternative hypotheses (LR) following the χ^2 distribution of this statistic with degrees of freedom (df) equal to the difference between the number of parameters (n_p) of the alternative hypothesis and the H0. The LRT rejects the H0, when the LR is considered too small by the given significance level (p value < 0.05).

Functional Analyses

To further analyze the existence of functional protein divergence among main MT clades, we calculated the grand average of hydropathicity (GRAVY) index using the referenced hydropathic index for amino acids as in Kyte and Doolittle (1982). The hydropathic index attributes a fixed value to an amino acid according to the hydrophobic or hydrophilic properties of its side chain (Kyte and Doolittle 1982). The GRAVY index of a sequence corresponds to the sum of the hydropathic values of each amino acid in the sequence divided by the number of residues in the sequence. This calculation was performed using the GRAVY Calculator (2013) web application. A higher positive score indicates greater hydrophobicity, meaning higher water repellency to non-polar molecules. Since the structure and folding of a protein define its function, differences in overall GRAVY index and in the hydropathic plots can provide information about functional divergence among protein types and have been used as an indication of the flexibility of the protein (e.g., Capasso et al. 2003, 2005). We calculated the average, maximum, minimum, and standard

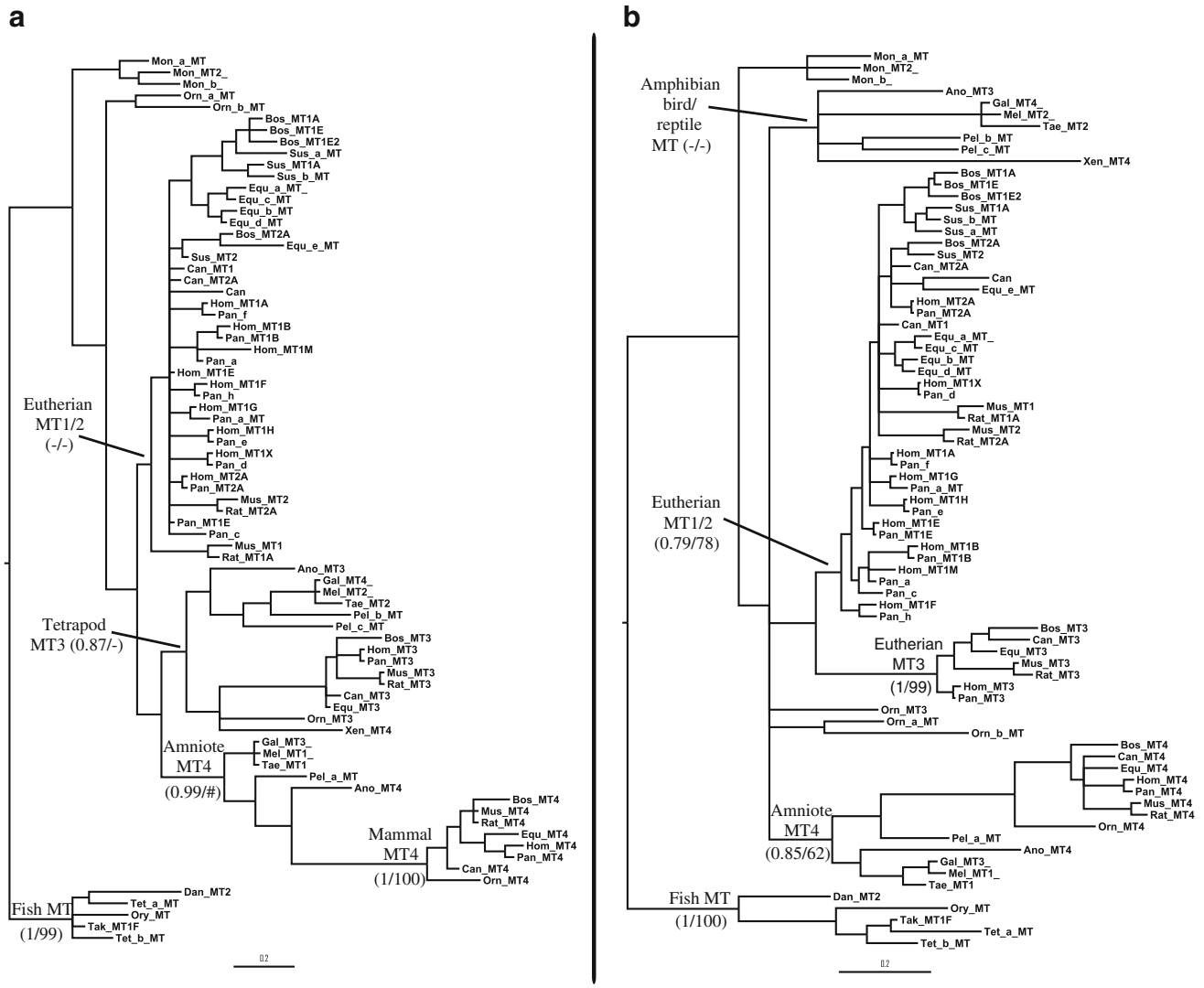

Fig. 1 Amino acid (**a**) and nucleotide (one single model of evolution) (**b**) Bayesian consensus trees (50 % majority-rule). Values are given for the following order: posterior probability/bootstrap (%). Values replaced by "−" when below 0.70 pp or 60 % bootstrap. When a clade is not recovered by the analysis, it is indicated with "#". Bootstrap values and posterior probabilities for all the nodes are reported in the legend of Online Resource 3

deviation (SD) GRAVY index for the main MT clades. These values were obtained on the basis of the GRAVY index of all the sequences contained within each main MT clade (Fig. 1). Furthermore, for each of the main MT clades, a "clade" amino acid consensus sequence was manually built by eye. This step was not done for the tetrapod MT3 clade due to the high divergence of its sequences and the greater number of mammalian sequences that would bias the consensus toward a mammalian MT3 consensus sequence. For the consensus amino acid clade sequences, the hydrophobicity plots were obtained following the Kyte–Doolittle method using the Protein Hydrophobicity Plots Generator (2013).

Functional divergence among main MT clades was further investigated using DIVERGE v.2 (Gu and Vander Velden 2002) by assessing the Type I and Type II functional divergences among main MT clades (defined in the program as clusters). Gu (2001) recognizes two main types of functional divergence for duplicated genes: Type I (Gu 1999) is characterized by amino acids that are highly conserved in one cluster, but variable in the other, taking into account the phylogeny and sequence variation across the tree. Type I divergence is correlated to different functional constraints between duplicate genes with consequent site-specific rate differences (Gu 1999). Type II (Gu 2006) can be considered as a sub-category of Type I divergence, and it is characterized by a "cluster-specific functional divergence" (Lichtarge et al. 1996) due to site-specific changes of amino acid physiochemical properties (e.g., charge, hydrophobicity). For this analysis, the sequences Equ_b_MT, Mel_MT1_,

Table 1 Test of selective pressure

a Models chosen to test variability of selective pressure (H0–H3) for the following tree topologies: unrooted Bayesian nucleotide/unrooted Bayesian amino acid tree (see "Materials and Methods" section for additional information). "ω" indicates dN/dS. "MT1/2, MT3, amniote MT4 and tetrapod MT3" next to the "ω" indicate all internal branches tested for of eutherian MT1 and MT2, eutherian MT3, amniote MT4, and tetrapod MT3, respectively (see "Materials and Methods" section for further explanations)

H0	$\omega_{MT1/2} = \omega_{MT3\ (or\ \omega tetrapodMT3)} = \omega_{amnioteMT4} = \omega_{others}$
H1	$\omega_{MT1/2} \neq \omega_{MT3} \neq \omega_{others\ or\ \omega MT1/2} \neq \omega_{tetrapodT3} \neq \omega_{others}$
H2	$\omega_{MT1/2} \neq \omega_{amnioteMT4} \neq \omega_{others}$
H3	$\omega_{MT3} \neq \omega_{amnioteMT4} \neq \omega_{others\ or\ \omega tetrapodMT3} \neq \omega_{amnioteMT4} \neq \omega_{others}$

b Results of variable selective pressure among main MT clades according to the different hypotheses (H1–H3). Values are given as results using: unrooted nucleotide Bayesian tree topology/unrooted amino acid Bayesian tree. "lll" indicates the Likelihood absolute value; "*np*" indicates the number of parameters; "LRT" indicates the Likelihood Ratio Test value standing for significant when LRT < 0.05 (significant values indicated in bold)

	$\omega_{MT1/2}$	ω_{MT3}	$\omega_{amnioteMT4}$	$\omega_{tetrapodMT3}$	ω_{others}	lll	*np*	LRT
H0	0.1113/0.0942	$=\omega_{MT1/2}$	$=\omega_{MT1/2}$	$=\omega_{MT1/2}$	$=\omega_{MT1/2}$	5044.01/ 5173.31	134/129	–
H1	0.1447/ 0.1350	0.0909/ $=\omega_{others}$	$=\omega_{others}$	$=\omega_{others}$/ 0.0581	0.0884/ 0.0682	5040.57/ 5156.63	136/131	$\mathbf{1.2 \times 10^{-2}}$/ $\mathbf{2.4 \times 10^{-4}}$
H2	0.1443/ 0.1341	$=\omega_{others}$	0.0731/ 0.0519	$=\omega_{others}$	0.0990/ 0.0751	5039.86/ 5164.88	136/131	$\mathbf{1.6 \times 10^{-2}}$/ $\mathbf{2.2 \times 10^{-4}}$
H3	$=\omega_{others}$	0.0876 $=\omega_{others}$	0.0745/ 0.0537	$=\omega_{others}$/ 0.0610	0.1264/ 0.1211	5040.85/ 5164.85	136/131	$\mathbf{4.2 \times 10^{-2}}$/ $\mathbf{2.1 \times 10^{-4}}$

Orn_a_MT, Orn_b_MT, Orn_MT3, and all sequences of *P. sinensis* were removed from the dataset due to either missing data at the beginning of the sequences or very divergent amino acid sequences and unresolved phylogenetic placement (see dataset file, Fig. 1, and Online Resource 3). Sites with missing data would be excluded from the divergence analysis, thus reducing the amino acid sites for which the estimate of sequence divergence among the distinct MT clades would be calculated. Very distinct sequences for which phylogenetic placement was not well recovered (Orn_a_MT, Orn_b_MT, and Orn_MT3) were removed as they could interfere with divergence estimates. A DIVERGE analysis was run separately for the gene tree obtained with the Bayesian analyses on the nucleotide (one model of evolution) and amino acid, respectively. Because the software only operates with phylogenetic trees without polytomies, when present, these were solved following the optimal reconciliation obtained for each gene tree (Bayesian nucleotide and amino acid, see Online Resource 4). Analyses were run on both tree topologies to take into account the influence of different tree topologies recovered by our analyses. The three-dimensional (3D) structure of the MT2 protein of *R. norvegicus* (Uniprot protein database 2013, accession number P04355 and corresponding to Rat_MT2A in our dataset) was used as MT protein reference. Divergence was tested between (1) eutherian MT1 and MT2 versus eutherian or tetrapod MT3, (2) eutherian MT1 and MT2 versus amniote MT4, and (3) eutherian or tetrapod MT3 versus amniote MT4 (see "Results and Discussion" section and Fig. 1 for these clades). The coefficient of functional divergence, theta (θ), corresponding to the proportion of sites expected to be functionally divergent, was determined for all gap-free amino acid positions. θ is directly linked to the coefficient of rate correlation between the evolutionary rates of a given site within each gene cluster (Gu 1999; Wang and Gu 2001). It varies between 0 and 1, with $\theta = 0$ indicating no observed functional divergence. DIVERGE provides a ML statistical estimate of θ (ThetaML) (Gu 2001). The statistically significant functional divergence among clusters ($\theta > 0$) is evaluated by a LRT with $\theta = 0$ representing the null hypothesis. The LRT was used for each of the pairwise comparisons written above with the null hypothesis being rejected for $p < 0.05$. Once the statistical evidence for functional divergence after gene duplication is provided, sites that are likely to influence this divergence were identified by applying a cut-off value. The cut-off value corresponds to the posterior probability of functional divergence at a site (see Gu 1999 for further details). In our analyses, we applied a conservative cut-off value of 0.9 for all comparisons.

Results and Discussion

Dataset Assembly and Characteristics

The initial retrieved dataset contained 86 sequences. Based on the conservative approach used to build the dataset (see "Materials and Methods" section), after removal and replacement of sequences due to incorrect annotation or annotation problems (Online Resource 2), the final dataset consisted of 77 sequences (indicated in Online Resource 1). Approximately half of the sequences removed were intronless. In duplication events, intronless genes may be generated by retroposition when the mRNA is retrotranscribed into the genome and may still represent functional MT proteins. Despite this possibility, intronless sequences were removed from the dataset after the lack of introns was confirmed according to both databases used (Ensembl and NCBI) (Online Resource 2) due to the impossibility of assessing the functionality of these genes. In our dataset assembly, we recovered cases of incorrect annotation in the Ensembl database (Online Resource 2), which has been previously observed (e.g., McEwen et al. 2006).

In our study, a large diversity of MT genes was recovered among mammalian species, ranging from three genes in *M. domestica* to twelve in *P. troglodytes*, supporting what has previously been found in humans (e.g., Moleirinho et al. 2011; Tío et al. 2004). In mammals, multiple MT genes/proteins are associated with expression in distinct tissues and different metal affinities, with some MT genes being more ubiquitous and others more specific (reviewed in Guirola et al. 2012).

In the other vertebrates examined in our study, we found two MT genes for each bird species, and a total of two or three genes in reptiles (lizard and turtle), one gene in amphibian, and one or two genes in fish. Previous studies of MT genes in squamates and amphibians from cDNA recovered only a single MT gene per species (Riggio et al. 2003; Trinchella et al. 2008, 2012), confirming in amphibians what was already observed in *Xenopus* (Saint-Jacques et al. 1995). Conversely, the presence of additional MT genes in reptiles is a novel finding. In previous studies of MT genes in reptiles, which were limited to squamates, Riggio et al. (2003) and Trinchella et al. (2006, 2008) observed only one MT type in the different tissues studied (brain, liver, ovary) in the species *Podarcis sicula*. The same MT type was also expressed in the venemon glands of a snake species (Junqueira-de-Azevedo and Ho 2002).

A possible explanation for the additional reptilian MT genes recovered in our study can be found in MT research on chickens, in which a second MT gene copy was only recovered after the full genome of this species was released (Villarreal et al. 2006). Biochemical analyses suggest that the functional spectrum of the two chicken genes/proteins is intermediate between the mammalian MT1 and MT4 genes. Villarreal et al. (2006) proposed that the second chicken MT gene might have remained undiscovered until the full genome of this species was released due to restricted or limited expression in time or space or due to specific metal-induction mechanisms (see Nam et al. 2007). A similar hypothesis could also explain why our use of genomic data allowed for additional MT genes to be characterized in squamates. Our results, together with newly sequenced complete reptile genomes and biochemical studies, will provide a basis for testing differential expression and induction of MTs among distinct reptilian species, tissues, developmental stages, and metal responses.

Of the four fish species examined in our study, only one MT gene copy was found, except for *T. nigroviridis*, for which two gene copies were recovered. Two MT genes for fish were also reported in other studies (Bargelloni et al. 1999; Trinchella et al. 2008). Fully sequenced fish genomes (e.g., Howe et al. 2013) will provide further insight into this subject.

Most MT CDS in our study contained 162 nucleotides (61 amino acids), excluding the terminal stop codon. MT3 type has an additional seven amino acids in comparison to the other MTs, as previously observed (reviewed in Vašák and Meloni 2011). The majority of MT sequences included in our dataset consisted of three exons, following the classical structure of mammalian MTs (reviewed in Hidalgo et al. 2009) (see sequence alignment). Two exons encode the β-domain of the protein, while the third exon encodes the α-domain (Vašák and Meloni 2011). These thiol-rich domains bind with high affinity to different numbers and types of metal ions (e.g., Zn^{2+}, Cd^{2+}, and Cu^{2+}) consequently folding into two dumbbell-like shapes connected by a flexible region made up of lysine amino acids (reviewed in Hidalgo et al. 2009; see amino acid positions 33 and 34 in our alignment). The β- and α-domains are generally comprised of amino acids 1–30 and 31–61, respectively (Braun et al. 1992; Romero-Isart et al. 1999).

The combined MT sequence alignment with all species had 13 and 17 % of conserved nucleotides and amino acids (183 and 58 variable sites), respectively, with gaps considered as a different state. Approximately one-third of the conserved amino acids were within the β-domain and two-thirds in the α-domain. The higher number of conserved amino acids observed in the α-domain of the protein, a pattern previously reported in mammals for MT1 versus MT4 (Tío et al. 2004), is probably correlated with the higher structural constraints of this domain (reviewed in Hidalgo et al. 2009).

Within the main MT clades (as in Fig. 1), nucleotide and amino acid sequence identity was 72 and 21 sites (120 and

43 variable) for eutherian MT1 and MT2, 146 and 68 (58 and 16 variable) for eutherian MT3, 77 and 26 sites (127 and 42 variable) for tetrapod MT3, 78 and 29 (111 and 34 variable) for amniote MT4, 67 and 26 (122 and 37 variable) for amphibian, bird, and reptile MT, and 113 and 41 (67 and 19 variable) for fish MT, respectively.

Phylogenetic Analyses

The saturation test did not indicate evidence of saturation when assuming symmetrical and asymmetrical topology on the complete dataset, as well as for the third codon position alone (data not shown). Furthermore, distinct phylogenetic reconstruction methods and the nucleotide or amino acid datasets recovered largely similar tree topologies. Finally, the Bayes factor ratio between the partitioned and non-partitioned Bayesian analyses was 0.98, suggesting that both model strategies are appropriate.

All analyses identified the following distinct major MT clades: fish MT clade, eutherian MT1 and MT2 (indicated in all the figures and tables as MT1/2), and amniote MT4 (the ML analysis on amino acids only recovered a mammalian MT4) (Fig. 1). Analyses run on amino acid data recovered a tetrapod MT3 clade, while analyses run on nucleotide data recovered the sequences belonging to this clade separately as an amphibian/reptile/bird MT clade and a eutherian MT3 clade (Fig. 1 and Online Resource 3). ML analyses run using the aLRT–SH-like approach largely agree with topologies and branch supports obtained using bootstrap resampling. Differently from what obtained using a bootstrap resampling approach in the ML analysis on the amino acid dataset, the one ran using the aLRT–SH-like recovered the amniote MT4 clade (support above 0.8; trees and values not shown). According to our phylogenetic results, current nomenclature for distinguishing among different MT types is not necessarily accurate. In fact, we recovered a fish MT clade that contains MT1 and MT2 named sequences, an eutherian MT1 and MT2 clade in which the two MT types are not respectively monophyletic, a tetrapod MT3 (or a reptile/bird/amphibian MT) clade that includes MT2, MT3, and MT4 named sequences, and an amniote MT4 clade including MT1, MT3, and MT4 named sequences (Fig. 1, Online Resources 1; see also Synteny results below). In this paper, we will refer to main MT clades following the predominant MT types as delineated above.

The fish MT clade (Fig. 1 and node 53 in Online Resource 3) was always recovered with high support values. The distinct fish versus tetrapod MT clades were recovered in previous analyses based on a reduced dataset of MT type sequences (e.g., Nam et al. 2007; Trinchella et al. 2008, 2012).

Eutherian MT1 and MT2 (Fig. 1, and node 31 in Online Resource 3), the clade with largest number of sequences ($n = 39$), encompasses all eutherian MTs annotated as MT1, MT2, and some unknown sequences from *C. lupus familiaris*, *E. caballus*, *P. troglodytes*, and *S. scrofa*. This clade was recovered by all the trees (Fig. 1 and Online Resource 3). Our analyses did not recover a mammalian MT1 and MT2 clade including *Ornithorynchus* and *Monodelphis*, with the exception of the ML on amino acid data (Fig. 1 and Online Resource 3; see also Synteny results).

Eutherian MT3 (Fig. 1 and node 40 in Online Resource 3) includes MT3 sequences from eutherian mammals, since *O. anatinus* MT3 is outside this clade (see also Synteny results) and was recovered with high support values in all analyses (Fig. 1 and Online Resource 3). A previous study using a dataset with fewer species recovered a mammalian MT3 clade instead (Moleirinho et al. 2011).

The ML and Bayesian analyses of amino acid data recovered a tetrapod MT3 clade (eutherian MT3+bird/reptiles MT+amphibian MT) (Fig. 1). Analyses run on the nucleotide data recovered distinct eutherian MT3 and amphibian/reptile/bird MT clades including unknown MTs (Fig. 1 and Online Resource 3). Bayes factor calculations comparing alternative amino acid-based tree topologies (i.e., Bayesian trees obtained with constrained clades) suggest that a tree topology including a tetrapod MT3 clade, as obtained with the amino acid dataset, or separate eutherian MT3 and amphibian/reptile/bird MT clades, as obtained with the nucleotide dataset, are equally probable (BF = 1; data not shown). Trinchella et al. (2012) recovered a reptile/bird MT clade, but to the exclusion of amphibian sequences, using cDNA data from more species of squamate reptiles (but no other reptiles) and amphibians.

An amniote MT4 clade (Fig. 1 and node 49 in Online Resource 3) was recovered by all phylogenetic analyses except the ML amino acid analysis. This clade includes MT4 sequences from mammals, MT1, MT3, and MT4 sequences from birds and reptiles, as well as previously unidentified sequences. A potential mammal/bird MT4 clade was also previously recovered, although with no significant statistical support, using ML analyses of amino acid sequences from a reduced dataset by Trinchella et al. (2012). Based on phylogenetic analyses, the MT4 clade has been proposed to be of a more ancient origin than the rest of the MT types (e.g., mammalian MT4 in Moleirinho et al. 2011, and mammal/bird MT4 in Trinchella et al. 2012). Our results do not clearly resolve MT4 as the ancestral MT type (Figs. 1, 2, and Online Resource 3). In addition, independently of the dataset used (nucleotide or amino acid), phylogenetic relationships among main MT clades are generally poorly resolved to assess the polarity of the phylogenetic relationships among MT types (Fig. 1, Online Resource 3; see also Synteny results). The short length of the MT CDS

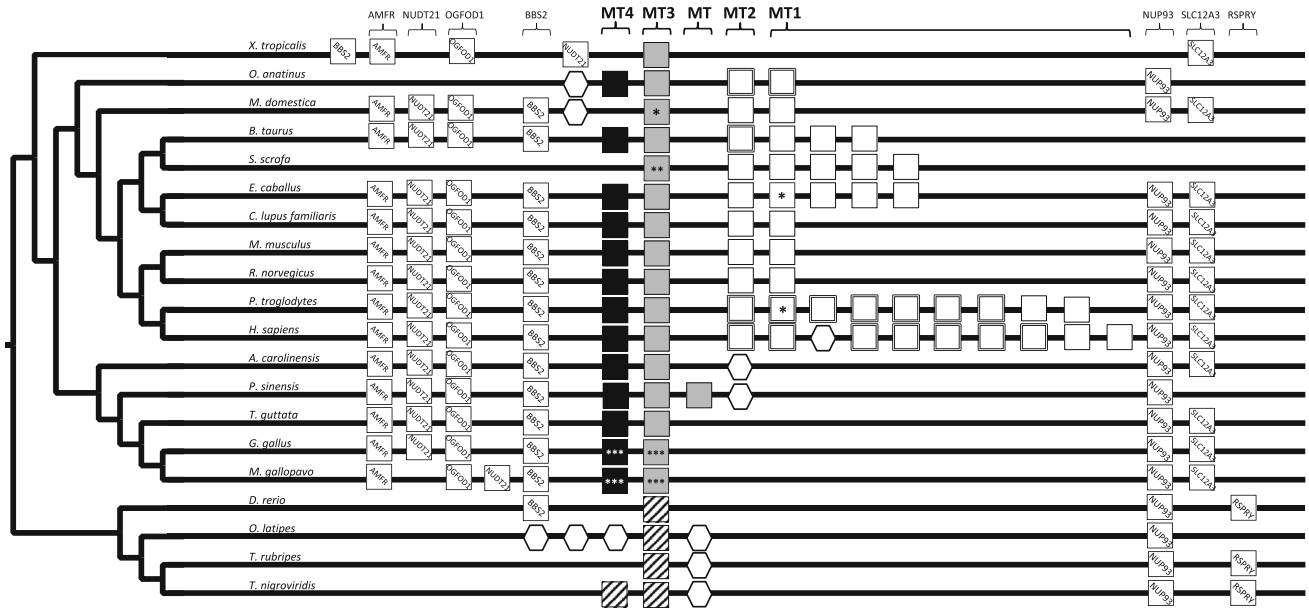

Fig. 2 Synteny analysis diagram. Species tree showing position of MT genes and neighboring genes on chromosomes. Each column represents a specific gene type (only for recognized ortholog genes). Neighbor genes: autocrine motility factor receptor (AMFR); nudix type motif 21 (NUDT21); 2-oxoglutarate and iron-dependent oxygenase domain containing 1 (OGFOD1); Bardet-Biedl syndrome 2 (BBS2); nucleoporin 93 (NUP93); solute carrier family 12-Sodium/Chloride transporters, Member 3 (SLC12A3); ring finger and SPRY domain containing 1 (RSPRY). Fish MT are represented by *squares* with *diagonal stripes*. Genes indicated with an hexagonal shape represent genes that are not orthologous to any genes in the other indicated species. Genes with double-line borders are paralogs of the genes of the same MT type. Neighboring genes for which correspondence between NCBI and Genomicus could not be found or for which certainty of their exact location could be not assessed are not indicated in the figure (e.g., *O. anatinus* and *Sus scrofa*). *Genes represented by NCBI annotation in our study instead of Ensembl annotation (as used in Genomicus). **Gene not used in this work due to annotation inconsistencies on Ensembl and NCBI (see Online Resource). ***Genes from chicken and turkey retrieved from NCBI (see Online Resource 2)

and the time of divergence among the different vertebrate groups do not permit higher phylogenetic resolution of the relationship among main MT clades.

Synteny Analysis

The amniote MT genes occur as a cluster of neighboring genes with a gene order that is generally well conserved in the amniote species we studied (Fig. 2). The synteny results suggest that the MT gene of *Sus scrofa,* which we did not include in our analysis due to the existence of a very large intron and uncertainty on the gene annotation (see Online Resource 2), would most likely correspond to MT3 genes. MT data from *M. gallopavo* show a gene inversion in comparison to other birds. The same MT gene neighbors in amniotes are also present in *Xenopus*, but show a different order on the chromosome (Fig. 2). Lack of conserved synteny between fish and tetrapods do not allow clear assessment of the orthology of the MT genes in fish (Fig. 2). The synteny analysis supports the existence of four main MT types occurring in vertebrates: a fish MT, an amniote MT4, a tetrapod MT3, and a mammalian MT1 and MT2 (Fig. 2). This result would support the tree topology obtained on the amino acid data, confirming the existence of tetrapod MT3 and amniote MT4 clades, which have never been identified before. Synteny analysis would also suggest that the lack of recovery of *O. anatinus* and *M. domestica* within mammalian MT1 and MT2 and MT3 clades was most likely due to lack of phylogenetic resolution (Figs. 1, 2). Mammals, *P. sinensis*, and *T. nigroviridis* MTs exhibit further gene duplications, while MT1 duplicated genes in mammals are known to be both functional and pseudogenes, but no data are so far available for other vertebrate species. Finally, as indicated by the phylogenetic results, current MT nomenclature across vertebrates does not reflect correct gene orthology (Figs. 1, 2). According to the synteny results, in tetrapods, genes recovered on the basis of their position on the chromosome as MT3 and MT4 include currently named MT1 and MT2 sequences from birds (Online Resource 1). *Xenopus* MT4 based on phylogenetic and synteny results is indicated as a MT3 type (Fig. 2).

Reconciliation Analysis

The "a priori" best outgroup chosen to root the phylogenetic vertebrate MT tree in previous studies (e.g.,

Moleirinho et al. 2011; Trinchella et al. 2012), the fish MT clade, was confirmed by our analyses (*D*/*L* score = 85, duplications = 32, losses = 37 using the Bayesian nucleotide gene tree, and *D*/*L* score = 105.5, duplications = 35, losses = 53 using the Bayesian amino acid gene tree). According to these results, the root in the fish clade would require fewer gene duplication and loss events than if the gene tree was rooted with another MT clade (data not shown).

The results obtained with the reconciliation analysis confirm the high turnover of gene duplication and loss predicted for this family (Online Resource 4). However, the lack of resolution of relationships among main MT clades (see also Phylogenetic analyses results) prevent a more precise assessment of the number of duplication and loss events occurring in vertebrates, as shown by the different number of duplication and losses estimated by the reconciliation analyses according to the different tree topologies used (see duplication and loss numbers above).

Analysis of Variation in Selective Pressure Among Lineages

An analysis was performed to determine whether differential selective pressure has occurred among main MT types. Generally, observed variation in selective pressure among coding sequences of gene duplicates may reflect an acceleration in non-synonymous substitutions. This could indicate functional divergence following a duplication event, eventually decreasing secondarily as an effect of purifying selection, which permits duplicated genes to maintain related but distinct functions (e.g., Gu 1999; Kondrashov et al. 2002; Li et al. 1985). Depending on when the functional divergence among paralog genes has occurred, different patterns of evolutionary rates may be detected immediately after the duplication event or among paralogs (see also Gu 1999 for further theoretical details).

In this study, estimates of ω, likelihood, and LRT values were generated by testing different hypotheses of variation in selective pressure among main MT types (Table 1b). The null hypothesis (H0) of constant selective pressure and mutation rate along the tree was rejected in all alternative hypotheses tested (H1-H3, $p < 0.05$, Table 1b). Our analysis also indicates that the evolution of MT genes is generally characterized by purifying selection ($\omega < 1$, Table 1b). However, since ω estimates are based on averages across all sites, our results do not exclude the possibility that positive selection and adaptation may have occurred at specific amino acid sites, as suggested by our functional analysis described in the following section. These results were obtained regardless of the type of input tree, Bayesian nucleotide or amino acid tree, used for the analyses.

Table 2 Hydropathic value results

MT clades	Max	Min	Average	GRAVY ± 2SD
Eutherian MT1/2	0.434	−0.0820	0.135	0.327 ± 0.0561
Eutherian MT3	−0.303	−0.469	−0.385	−0.250 ± 0.520
Tetrapod MT3	0.144	−0.506	−0.299	0.0108 ± 0.609
Mammal MT4	0.0758	−0.157	−0.022	0.145 ± 0.18
Amniote MT4	0.0758	−0.167	−0.0650	0.135 ± 0.233
Amphibian, bird, and reptile MT	0.144	−0.506	−0.211	0.106 ± 0.566
Fish MT	0.0767	−0.192	−0.0687	0.103 ± 0.243

GRAVY index calculated for main MT clades (see "Materials and Methods" section for additional information). "Max", "Min", "Average", and "SD" indicate, respectively, the maximum, minimum, average, and standard deviation GRAVY indeces obtained for sequences within a given clade

Independently of the input tree topology used for the analyses, our results indicate an increase of ω in eutherian MT1 and MT2 of at least one and a half times in comparison to the rest of the tetrapod MT types (H1 and H2, Table 1b). This result seems to be in agreement with the large number of duplication events occurring within this clade and the differential tissue and temporal expression of distinct MT1 genes observed in human and mouse (e.g., Moleirinho et al. 2011; Schmidt and Hamer 1986). Eutherian/tetrapod MT3 and amniote MT4 show similar values of ω suggesting similar mutation rates within each of these clades (H3, Table 1b).

Functional Analyses

Minimum, maximum, average, and standard deviation of the hydropathic GRAVY index for the main MT clades are provided in Table 2. MT4 clades (amniote MT4 and mammalian MT4) have similar slightly negative average GRAVY scores and hydrophobic profiles across the sequences, different from what has been observed for the other main MT clades (Table 2; Fig. 3). This would indicate different biochemical properties of the MT4 proteins from the other MTs. Our results confirm a negative value and large variation in the hydropathic GRAVY index (Capasso et al. 2003; Trinchella et al. 2008, 2012) (Table 2). Average hydropathic GRAVY indices for MT1 and MT2 and MT3 were at opposite extremes, with MT1 and MT2 being the only MT clade showing a positive average hydropathic value, while MT3 exhibited the most negative average value obtained for any MT type (Table 2). This could indicate more divergent biochemical properties, including metal-binding affinities, of these two MT types.

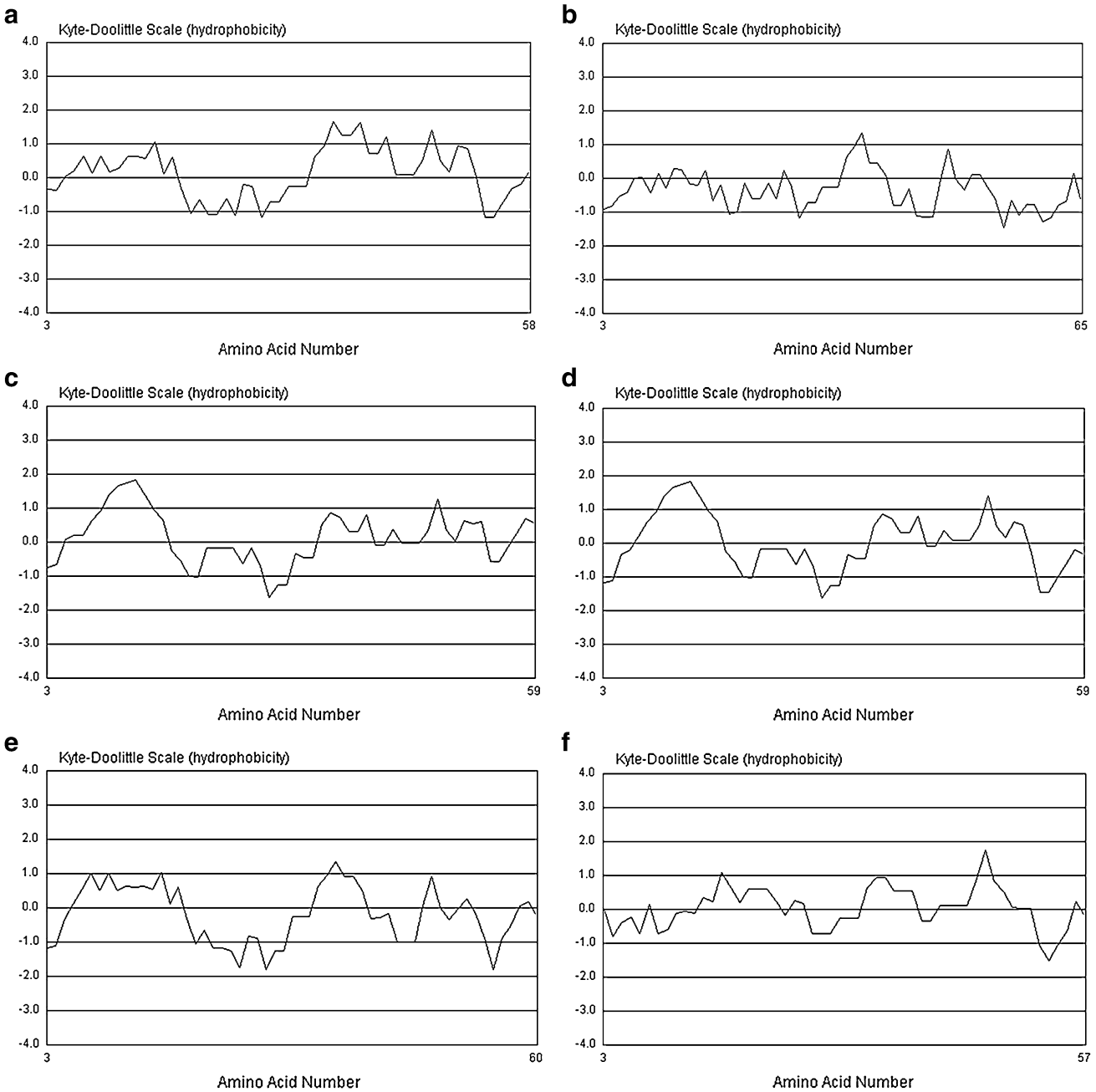

Fig. 3 Hydrophobicity plots using consensus amino acid sequences obtained for the following clades (see "Materials and Methods" section for additional explanations): **a** Eutherian MT1 and MT2; **b** Eutherian MT3; **c** Mammalian MT4; **d** Amniote MT4; **e** Amphibian, bird, and reptile MT; and **f** fish MT. *Y*-axes indicates hydrophobicity values, whereas *x*-axes indicates amino acid positions

The hydrophobicity plots for the main MT clades (Fig. 3) showed higher variability among clades in the beginning of the amino acid sequence, corresponding to the β-domain, which is the domain most involved in the functional divergence among MT types (e.g., Hidalgo et al. 2009; Tío et al. 2004, see also below). While the GRAVY index gives indications about the hydropathic character of a protein, and therefore, about its solubility and its biochemical properties, and is correlated with a higher capacity for undertaking conformational changes, the hydrophobicity plot permits a visualization of how the hydropathicity of the protein varies along its sequence. Therefore, while variation among MT types for the GRAVY index may suggest functional divergence, the hydropathicity plot may highlight domains of the proteins most likely associated with this divergence. Our results would, therefore, suggest a higher functional divergence among MT types in the β-domain of the MT protein,

Table 3 Type I and II divergence test results for nucleotide/amino acid topologies

	$\theta_I \pm$ SE	LRT	p	Pp cut-off = 0.9	$\theta_{II} \pm$ SE	P	Pp cut-off = 0.9
Eutherian MT1/2 vs. eutherian (or tetrapod) MT3	$7.99 \times 10^{-1} \pm 0.310$/	7.16/	**0.007**	4/	$8.24 \times 10^{-2} \pm 0.160$/	0.20/	8/
	$4.44 \times 10^{-1} \pm 0.170$	6.73	**0.009**/	1	$6.76 \times 10^{-2} \pm 0.185$	0.16	6
Eutherian MT1/2 vs. amniote MT4	$3.75 \times 10^{-1} \pm 0.141$/	7.09/	**0.008**/	2/	$2.178 \times 10^{-2} \pm 0.185$/	0.25/	1/
	$3.66 \times 10^{-1} \pm 0.140$	6.83	**0.008**	2	$5.070 \times 10^{-2} \pm 0.180$	0.20	0
Eutherian (or tetrapod) MT3 vs. amniote MT4	$1.21 \times 10^{-1} \pm 0.595$/	0.0412/	0.8/	0/	$-1.893 \times 10^{-1} \pm 0.157$/	0.35/	0/
	$2.09 \times 10^{-1} \pm 0.204$	1.045	0.3	0	$1.109 \times 10^{-1} \pm 0.186$	0.25	0

"θ_I" indicates the coefficent of functional divergence; "SE" indicates the standard error; "LRT" corresponds to the 2 log-likelihood ratio against the null hypothesis of $\theta_I = 0$; "p" indicates the p value; "Pp cut-off" represents the posterior probability cut-off for specific amino acid sites

Significant p-values ($p < 0.05$) are indicated in bold

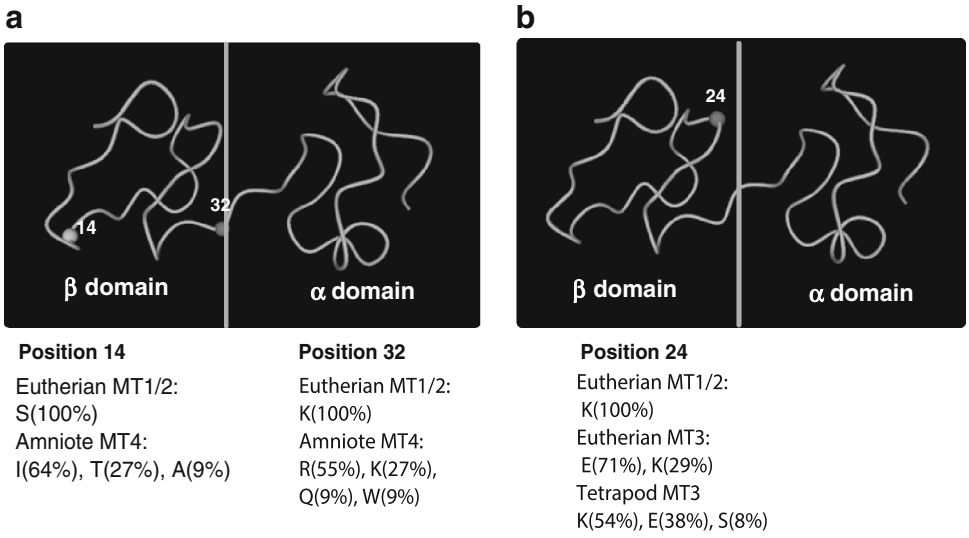

Fig. 4 MT 3D structure as obtained from DIVERGE using the RatMT2 3D protein as a model. Figure shows amino acid sites that were recovered involved in functional divergence independently of the tree topology used (see "Materials and Methods" section for further information). Numbered spheres on the protein structure indicate Type I divergent amino acids and position (cut-off value = 0.9) between: **a** eutherian MT1 and MT2 versus eutherian/tetrapod MT3, **b** eutherian MT1 and MT2 versus amniote MT4 clades. The type of amino acid change at that position for the compared clades is given below the figure by the Amino acid code: "**K**"—lysine, "**E**"—glutamic acid, "**S**"—serine, "**I**"—isoleucine, "**T**"—threonine, "**A**"—alanine, "**R**"—arginine, "**Q**"—glutamine, and "**W**"—tryptophan

further supporting the more constrained functional role of the α-domain (Hidalgo et al. 2009).

The DIVERGE analysis was performed to further study the protein functional divergence among main MT types. The results of this analysis were the same independently of the tree topology used (Bayesian nucleotide or amino acid trees) and indicated sites involved in Type I, but not Type II, functional divergence between eutherian MT1 and MT2 versus MT3 (eutherian or tetrapod) and amniote MT4, but not between MT3 and MT4 (Table 3). These results further support our findings on the different selective pressure between MT1 and MT2 versus MT3 and MT4, but similar mutation rates between MT3 and MT4. One site at position 24 and two sites at positions 14 and 32 (amino acid positions as in Online Resource 5) were recovered by the analyses run on both tree topologies as being involved in functional divergence between MT1 and MT2 versus MT3 and between MT1 and MT2 versus MT4, respectively (Fig. 4). These sites occur within the β-domain and in the flexible region connecting the two protein domains, further confirming the higher impact on functional divergence among MT types of the β-domain versus the α-domain. Some of the sites indicated to be involved in functional divergence among MT types have been reported to be associated with different protein functions (reviewed in Hidalgo et al. 2009). The amino acid site 14 occurs next to

the metal-binding cysteine, conserved in all main clades. Intercalating residues among the conserved cysteines in the β-domain are highly dissimilar and associated with functional divergence between MT4 and MT1 types (Tío et al. 2004). Site 32 occurs within the flexible region of lysines connecting the two domains. MT1 and MT2 clade have a conserved lysine in this position, while MT4 has mostly an arginine (Fig. 4). Both amino acids have similar biochemical characteristics; however, a change from lysine to arginine could interfere with the folding of the two domains, consequently modifying the function or metal-binding affinities between MT1 and MT2 versus MT4. Functional studies could further focus on how amino acid changes at the position recovered in our work (Fig. 4) could produce changes in functional activity of these proteins.

Conclusions

The dataset used in this work was based on representative vertebrate species with complete genome annotations that could be confirmed by more than one genomic database to improve data quality. As demonstrated here, this approach improves detections of ortholog and paralog genes, improving resolution of the molecular evolution of the MT gene family in vertebrates. Using this dataset, we were able to recover multiple MT types in all amniotes, suggesting that duplication and functional divergence in MTs is not limited to mammals and birds. Furthermore, our results suggest the existence of an amniote MT4 clade, a mammalian MT1 and MT2 clade, and a tetrapod MT3 clade. Our results, together with the analyses of functional divergence between main MT clades, suggest a likely association between MT functional divergence and duplication events in vertebrates and a marked functional distinction between MT1 and MT2 versus MT3 and MT4 in vertebrates.

In humans, MT1 and MT2 are inducible and expressed in almost every tissue. MT3 and MT4 are, on the other hand, relatively unresponsive to inducers that stimulate MT1 and MT2 expression and are mostly located in the central nervous system and in the stratified squamous epithelium, respectively (reviewed in Vašák and Meloni 2011). The limited data available on the expression and induction of distinct MT types in non-mammalian vertebrates and non-vertebrate chordates (e.g., Guirola et al. 2012; Nam et al. 2007) suggest the existence of two MT types in non-mammalian vertebrates, one that is more ubiquitous, and the other that is more specialized. The poor resolution of phylogenetic relationships among the main MT types does not allow a full interpretation of the evolutionary process of functional divergence in this gene family and to infer whether a more generalized MT type evolved into one or more functionally specialized MT types in amniotes. More biochemical and expression data are needed to understand the underlying mechanisms of functional divergence after gene duplication in MTs, especially between mammals and other vertebrates.

Our results on the number of ortholog genes occurring in each studied species and the sites potentially involved in functional divergence among MT types can help design future MT functional studies in other vertebrates, besides humans and mice. In fact, while a large body of biochemical and molecular work is currently available for mammalian model species, similar data are currently lacking for the distinct MT types recovered in non-mammalian vertebrates. Furthermore, in vertebrates, MT expression and concentrations are often used in toxicological and metal homeostasis studies (e.g., Andreani et al. 2007; Kim et al. 2013; Riggio et al. 2003). As studies of mammalian model species reveal, not all MT types are equally involved in the same function or expressed in the same tissue and at the same time, nor do they show the same metal affinity. Therefore, the lack of knowledge on similar potential differences among MT types in other vertebrates possessing multiple MT genes may be misleading or provide incomplete conclusions. Comparative genomic and biochemical studies will help fill this knowledge gap and contribute to our understanding of both MT evolution and functional divergence following gene duplication in vertebrates.

Acknowledgments We are thankful to M Fonseca, N. Galtier, JM Lourenço, B Nabholz, S Rocha, D Salvi, and Z Yang for feedback on this work. We are grateful to J-P Doyon for his help with the reconciliation analysis and his comments on this part. We are thankful to an anonymous reviewer for comments on an early version of this manuscript. YC was partially financially supported by a FCT (Fundação para Ciência e Tecnologia, Portugal) postdoctoral fellowship SFRH/BDP/73515/2010.

Disclaimer This study was not sponsored by the U.S. EPA, and the views expressed by the authors in this publication do not necessarily represent the views of the U.S. EPA or the United States.

References

Andreani G, Santoro M, Cottignoli S, Fabbri M, Carpenè E, Isani G (2007) Metal distribution and metallothionein in loggerhead (*Caretta caretta*) and green (*Chelonia mydas*) sea turtles. Sci Total Environ 390:287–294. doi:10.1016/j.scitotenv.2007.09.014

Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol 55:539–552

Bargelloni L, Scudiero R, Parisi E, Carginale V, Capasso C, Patarnello T (1999) Metallothioneins in antarctic fish: evidence for independent duplication and gene conversion. Mol Biol Evol 16:885–897

Binz P-A, Kägi JHR (1999) Classification of metallothionein. http://www.bioc.uzh.ch/mtpage/classif.html. Accessed 6 Feb 2013

Blindauer CA, Leszczyszyn OI (2010) Metallothioneins: unparalleled diversity in structures and functions for metal ion homeostasis and more. Nat Prod Rep 27:720–741. doi:10.1039/B906685N

Braun W, Vašák M, Robbins AH, Stout CD, Wagner G, Kägi JHR, Wüthrich K (1992) Comparison of the NMR solution structure and the X-ray crystal structure of rat metallothionein-2. Proc Natl Acad Sci USA 89:10124–10128

Brown CJ, Todd KM, Rosenzweig RF (1998) Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. Mol Biol Evol 15:931–942

Capasso C, Carginale V, Scudiero R, Crescenzi O, Spadaccini R, Temussi PA, Parisi E (2003) Phylogenetic divergence of fish and mammalian metallothionein: relationships with structural diversification and organismal temperature. J Mol Evol 57:S250–S257. doi:10.1007/s00239-003-0034-z

Capasso C, Carginale V, Crescenzi O, Di Maro D, Spadaccini R, Temussi PA, Parisi E (2005) Structural and functional studies of vertebrate metallothioneins: cross-talk between domains in the absence of physical contact. Biochem J 391:95–103. doi:10.1042/BJ20050335

Capdevila M, Atrian S (2011) Metallothionein protein evolution: a miniassay. J Biol Inorg Chem 16:977–989. doi:10.1007/s00775-011-0798-3

Carpenè E, Andreani G, Isani G (2007) Metallothionein functions and structural characteristics. J Trace Elem Med Biol 21:35–39. doi:10.1016/j.jtemb.2007.09.011

Chang D, Duda TF (2012) Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. Mol Biol Evol 29:2019–2029. doi:10.1093/molbev/mss068

Chen K, Durand D, Farach-Colton M (2000) NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol 7:429–447. doi:10.1089/106652700750050871

Chiari Y, Cahais V, Galtier N, Delsuc F (2012) Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol 10:65. doi:10.1186/1741-7007-10-65

Dallinger R, Höckner M (2013) Evolutionary concepts in ecotoxicology: tracing the genetic background of differential cadmium sensitivities in invertebrate lineages. Ecotoxicology 22:767–778. doi:10.1007/s10646-013-1071-z

Dallinger R, Berger B, Hunziker P, Kagi JH (1997) Metallothionein in snail Cd and Cu metabolism. Nature 388:237–238

Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165. doi:10.1093/bioinformatics/btr088

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772. doi:10.1038/nmeth.2109

Davis SR, Cousins RJ (2000) Recent advances in nutritional sciences metallothionein expression in animals: a physiological perspective on function. J Nutr 1:1085–1088

Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW (2006) The evolution of mammalian gene families. PLoS ONE 1(1):e85. doi:10.1371/journal.pone.0000085

Doyon J-P, Ranwez V, Daubin V, Berry V (2011) Models, algorithms and programs for phylogeny reconciliation. Brief Bioinf 12:392–400. doi:10.1093/bib/bbr045

Durand D, Halldórsson BV, Vernot B (2006) A hybrid micro-macroevolutionary approach to gene tree reconstruction. J Comput Biol 13:320–335. doi:10.1089/cmb.2006.13.320

Ensembl database (2013) www.ensembl.org. Accessed 2 May 2013

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Claphan P, Coates G, Fairley S, Fitzgerald S, Gordon L et al (2011) Ensembl 2011. Nucleic Acids Res 39:D800–D806. doi:10.1093/nar/gkq1064

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545

Fowler BA, Hildebrand CE, Kojima Y, Webb M (1987) Nomenclature of metallothionein. Exp Suppl 52:19–22

Garrett SH, Somji S, Todd JH, Sens MA, Sens DA (1998) Differential expression of human metallothionein isoform I mRNA in human proximal tubule cells exposed to metals. Environ Health Perspect 106:825–831

GRAVY Calculator (2013) www.gravy-calculator.de. Accessed 11 Oct 2013

Gu X (1999) Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664–1674

Gu X (2001) Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol 18:453–464

Gu X (2006) A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. Mol Biol Evol 23:1937–1945. doi:10.1093/molbev/msl056

Gu X, Vander Velden K (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics 18:500–501. doi:10.1093/bioinformatics/18.3.50

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704. doi:10.1080/10635150390235520

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321. doi:10.1093/sysbio/syq010

Guirola M, Pérez-Rafael S, Capdevila M, Palacios O, Atrian S (2012) Metal dealing at the origin of the Chordata phylum: the metallothionein system and metal overload response in *Amphioxus*. PLoS ONE 7(8):e43299. doi:10.1371/journal.pone.0043299

Hidalgo J, Chung R, Penkowa M, Vašák M (2009) Structure and function of vertebrate metallothioneins. Met Ions Life Sci 5:279–317

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffatto M, Collins JE, Humphray S, McLaren K, Matthews L et al (2013) The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503. doi:10.1038/nature12111

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Junqueira-de-Azevedo IDLM, Ho PL (2002) A survey of gene expression and diversity in the venom glands of the pitviper snake *Bothrops insularis* through the generation of expressed1 sequence tags (ESTs). Gene 299:279–291. doi:10.1016/S0378-1119(02)01080-6

Kent WJ (2002) BLAT—The BLAST-like alignment tool. Genome Res 12:656–664. doi:10.1101/gr.229202

Kim M, Park K, Park JY, Kwak I-S (2013) Heavy metal contamination and metallothionein mRNA in blood and feathers of black-tailed gulls (*Larus crassirostris*) from South Korea. Environ Monit Assess 185:2221–2230. doi:10.1007/s10661-012-2703-0

Kondrashov FA (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc R Soc 279:5048–5057. doi:10.1098/rspb 2012.1108

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. Genome Biol 3:1–0008. doi:10.1186/gb-2002-3-2-research0008

Koonin EV (2009) Darwinian evolution in the light of genomics. Nucleic Acids Res 37:1011–1034. doi:10.1093/nar/gkp089

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Lenormand T, Guillemaud T, Bourguet D, Raymond M (1998) Appearance and sweep of a gene duplication: adaptive response and potential for new functions in the mosquito *Culex pipiens*. Evolution 52:1705–1712

Li W, Luo C, Wu C (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Li C, Ortí G, Zhang G, Lu G (2007) A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evol Biol 7:44. doi:10.1186/1471-2148-7-44

Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257:342–358

Louis A, Muffato M, Roest Crollius H (2012) Genomicus: five genome browsers for comparative genomics in eukaryota. Nucleic Acids Res 41(D1):D700–D705. doi:10.1093/nar/gks1156

McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H, Elgar G (2006) Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. Genome Res 16:451–465. doi:10.1101/gr.4143406

Miles AT, Hawksworth GM, Beattie JH, Rodilla V (2000) Induction, regulation, degradation, and biological significance of mammalian metallothioneins. Crit Ver Biochem Mol Biol 35:35–70

Moleirinho A, Carneiro J, Matthiesen R, Silva RM, Amorim A, Azevedo L (2011) Gains, losses and changes of function after gene duplication: study of the metallothionein family. PLoS ONE 6(4):e18487. doi:10.1371/journal.pone.0018487

Muffato M, Louis A, Poisnel C-E, Roest Crollius H (2010) Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. Bioinformatics 26(8):1119–1121. doi:10.1093/bioinformatics/btq079

Nam D-H, Kim E-Y, Iwata H, Tanabe S (2007) Molecular characterization of two metallothionein isoforms in avian species: evolutionary history, tissue distribution profile, and expression associated with metal accumulation. Comp Biochem Physiol 145:295–305. doi:10.1016/j.cbpc.2006.10.012

NCBI database (2013) www.ncbi.nlm.nih.gov. Accessed 5 June 2013

Nordberg GF (1989) Modulation of metal toxicity by metallothionein. Biol Trace Elem Res 21:131–135

Nordberg M, Kojima Y (1979) Metallothionein. In: Kägi JWR, Nordberg M (eds) Proceedings of the first international meeting on metallotionein and other low molecular weight metal-binding proteins. Birkhäuser, Switzerland, pp 41–117

Nordberg M, Nordberg GF (2009) Metallothioneins: historical development and overview. In: Sigel A, Sigel H, Sigel RKO (eds) Metal ions in life sciences. The Royal Society of Chemistry, Cambridge, UK, pp 1–29

Nylander J, Ronquist F, Huelsenbeck J, Nieves-Aldrey J (2004) Bayesian phylogenetic analysis of combined data. Syst Biol 53:47–67. doi:10.1080/10635150490264699

Ohno S (1970) Evolution by gene duplication. Springer, London

Palacios O, Atrian S, Capdevila M (2011) Zn- and Cu-thioneins: a functional classification for metallothioneins? J Biol Inorg Chem 16:991–1009. doi:10.1007/s00775-011-0827-2

Palmiter RD (1998) The elusive function of metallothioneins. Proc Natl Acad Sci 95:8428–8430

Prince VE, Pickett FB (2002) Splitting pairs: the diverging fates of duplicated genes. Nat Rev Gen 3:827–837

Protein Hydrophobicity Plots Generator (2013) www.vivo.colostate.edu/molkit/hydropathy. Accessed 2 May 2013

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ et al (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 19:1316–1323. doi:10.1101/gr.080531.108

Rambaut A, Drummond AJ (2009) Tracer, MCMC Trace Analysis Tool, v1.5.0. http://beast.bio.ed.ac.uk/Tracer. Acessed in November 2012

Riggio M, Trinchella F, Filosa S, Parisi E, Scudiero R (2003) Accumulation of zinc, copper, and metallothionein mRNA in lizard ovary proceeds without a concomitant increase in metallothionein content. Mol Reprod Dev 66:374–382. doi:10.1002/mrd.10365

Romero-Isart N, Cols N, Termansen MK, Gelpí JL, González-Duarte R, Atrian S, Capdevila M, González-Duarte P (1999) Replacement of terminal cysteine with histidine in the metallothionein alpha and beta domains maintains its binding capacity. Eur J Biochem 259:519–527

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574. doi:10.1093/bioinformatics/btg180

Saint-Jacques E, April MJ, Séguin C (1995) Structure and metal-regulated expression of the gene encoding *Xenopus laevis* metallothionein-A. Gene 160:201–206

Schmidt CJ, Hamer DH (1986) Cell specificity and an effect of *ras* on human metallothionein gene expression. Proc Natl Acad Sci USA 83:3346–3350

Shaw JR, Coulbourne JK, Davey JC, Glaholt SP, Hampton TH, Chen CY, Folt CL, Hamilton JW (2007) Gene response profiles for *Daphnia pulex* exposed to the environmental stressor cadmium reveals novel crustacean metallothioneins. BMC Genomics 8:477. doi:10.1186/1471-2164-8-477

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739. doi:10.1093/molbev/msr121

Tío L, Villarreal L, Atrian S, Capdevila M (2004) Functional differentiation in the mammalian metallothionein gene family: metal binding features of mouse MT4 and comparison with its paralog MT1. J Biol Chem 279:24403–24413. doi:10.1074/jbc.M401346200

Tree of Life Web Project (2012) http://tolweb.org/tree/. Accessed 9 Oct 2012

Trinchella F, Riggio M, Filosa S, Volpe MG, Parisi E, Scudiero R (2006) Cadmium distribution and metallothionein expression in lizard tissues following acute and chronic cadmium intoxication. Comp Biochem Physiol C 144:272–278. doi:10.1016/j.cbpc.2006.09.004

Trinchella F, Riggio M, Filosa S, Parisi E, Scudiero R (2008) Molecular cloning and sequencing of metallothionein in squamates: new insights into the evolution of the metallothionein genes in vertebrates. Gene 423:48–56. doi:10.1016/j.gene.2008.06.027

Trinchella F, Esposito MG, Scudiero R (2012) Metallothionein primary structure in amphibians: insights from comparative evolutionary analysis in vertebrates. C R Biol 335:480–487. doi:10.1016/j.crvi.2012.05.003

Uniprot protein database (2013) www.uniprot.org/uniprot. Accessed 22 April 2013

Valls M, Bofill R, Gonzalez-Duarte R, Gonzalez-Duarte P, Capdevila M, Atrian S (2001) A new insight into metallothionein (MT) classification and evolution. The in vivo and in vitro metal binding features of *Homarus americanus* recombinant MT. J Biol Chem 276:32835–32843. doi:10.1074/jbc.M102151200

Vašák M, Armitage I (1986) Nomenclature and possible evolutionary pathways of metallothionein and related proteins. Environ Health Perspect 65:215–216

Vašák M, Meloni G (2011) Chemistry and biology of mammalian metallothioneins. J Biol Inorg Chem 16:1067–1078. doi:10.1007/s00775-011-0799-2

Vernot B, Stolzer M, Goldman A, Durand D (2008) Reconciliation with non-binary species trees. J Comput Biol 15:981–1006. doi:10.1089/cmb 2008.0092

Villarreal L, Tío L, Capdevila M, Atrian S (2006) Comparative metal binding and genomic analysis of the avian (chicken) and mammalian metallothionein. FEBS J 273:523–535. doi:10.1111/j.1742-4658.2005.05086.x

Wang Y, Gu X (2001) Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics 158:1311–1320

Xia X (2013) DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol 1–9. doi:10.1093/molbev/mst064

Xia X, Xie Z, Salemi M, Chen L, Wang Y (2003) An index of substitution saturation and its application. Mol Phylogenetic Evol 26:1–7. doi:10.1016/S1055-7903(02)00326-3

Yanai I, Camacho CJ, DeLisi C (2000) Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Phys Rev Lett 85:2641–2644. doi:10.1103/PhysRevLett.85.2641

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591. doi:10.1093/molbev/msm088

Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298. doi:10.1016/S0169-5347(03)00033-8

Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA 95:3708–3713