GATA family of transcription factors of vertebrates: phylogenetics and chromosomal synteny

Chunjiang He, Hanhua Cheng* and Rongjia Zhou*

Department of Genetics and Center for Developmental Biology, College of Life Sciences, Wuhan University, Wuhan 430072, P R China

**Corresponding authors (Fax, 86-27-68756253; Email, rjzhou@whu.edu.cn, hhcheng@whu.edu.cn)*

GATA genes are an evolutionarily conserved family, which encode a group of important transcription factors involved in the regulation of diverse processes including the development of the heart, haematopoietic system and sex gonads. However, the evolutionary history of the *GATA* family has not been completely understood. We constructed a complete phylogenetic tree with functional domain information of the *GATA* genes of both vertebrates and several invertebrates, and mapped the *GATA* genes onto relevant chromosomes. Conserved synteny was observed around the *GATA* loci on the chromosomes. *GATAs* have a tendency to segregate onto different chromosomes during evolution. The phylogenetic tree is consistent with the relevant functions of *GATA* members. Analysis of the zinc finger domain showed that the domain tends to be duplicated during evolution from invertebrates to vertebrates. We propose that the balance between duplications of zinc finger domains and *GATA* members should be maintained to exert their physiological roles in each evolutionary stage. Therefore, evolutionary pressure on the *GATAs* must exist to maintain the balance during evolution from invertebrates to vertebrates. These results reveal the evolutionary characteristics of the *GATA* family and contribute to a better understanding of the relationship between evolution and biological functions of the gene family, which will help to uncover the *GATA*s' biological roles, evolution and their relationship with associated diseases.

[He C, Cheng H and Zhou R 2007 GATA family of transcription factors of vertebrates: phylogenetics and chromosomal synteny; *J. Biosci.* **32** 1273–1280]

1. Introduction

The (T/A)GATA(A/G) structure was first found in the globin gene promoter of chicken (Evans *et al* 1988). Proteins binding to the (T/A)GATA(A/G) structure were named GATA, which have been identified in mammals, fish, Aves, insects, fungi and plants. The GATA family belongs to the zinc finger superfamily, with the zinc finger $CX_2CX_{17-20}CX_2C$; this is highly conserved within members of the family in different organisms. As a conserved gene family, it was established that GATA played important roles in several developmental processes. Lowry and Atchley (2000) constructed a phylogenetic tree of *GATA* genes, suggesting that the ancestral GATA protein contained only a single zinc finger and a single tandem duplication event prior to the divergence of the fungal and metazoan lineages (Lowry and Atchley 2000). Reyes *et al* (2004) analysed the *GATA* genes of *Arabidopsis* and rice, and defined a model of zinc finger domain in plants (Reyes *et al* 2004). Patient and McGhee (2002) divided *GATA* genes into two groups according to their functions (Patient and McGhee 2002). *GATA1/2/3* were classified into a group that was mainly expressed in the haematopoietic system, and *GATA4/5/6* into another group, which was mainly expressed in endodermally derived tissues (heart, lung, stomach, intestine, ovary, blood vessels, etc.) and has a close relationship with heart development and diseases.

Work has been done on the phylogenetic analysis of the GATA family in some organisms, including plants and animals (Lowry and Atchley 2000; Reyes *et al* 2004). However, there are not enough gene sources in these earlier studies to sufficiently reveal the evolutionary characteristics

Keywords. Phylogeny; transcription factor; vertebrates; zinc finger domain

of this gene family in both vertebrates and invertebrates. While we know well the expression patterns and functions of genes for many of these proteins, evolutionary models of the *GATA* family have not been completely understood. With the completion of genome sequencing, more *GATA* genes have been identified in different species, and further phylogenetic analyses of this gene family will facilitate our understanding of its evolutionary and functional significance. Based on the resources of open databases, we collected available full-length sequences of *GATA* genes in known vertebrates and several invertebrates belonging to different evolutionary groups, constructed a phylogenetic tree and performed evolutionary relationship analysis of the domains and gene members. We also mapped *GATA* members onto chromosomes to reveal the relations of conserved synteny and member segregation. These results will supply new information to understand the origin, evolution and classification of the *GATA* family.

Table 1. All proteins included in our analyses including source, length in amino acids and accession number

Sequence	Organism	Accession No.	Length	Sequence	Organism	Accession No.	Length
GATA1-Human	*Homo sapiens*	NP_002040	413	xGATA5a-*X. laevis*	*Xenopus laevis*	NP_001081962	390
GATA1-Zebrafish	*Danio rerio*	NP_571309	418	xGATA5b-*X. laevis*	*Xenopus laevis*	NP_001079831	388
GATA1-Mouse	*Mus musculus*	NP_032115	413	GATA5-Chicken	*Gallus gallus*	NP_990752	391
GATA1-Cow	*Bos taurus*	XP_873448	413	GATA6-Human	*Homo sapiens*	NP_005248	595
xGATA1-*X. laevis*	*Xenopus laevis*	NP_001079109	359	GATA6-Zebrafish	*Danio rerio*	NP_571632	383
GATA2-Human	*Homo sapiens*	NP_116027	480	GATA6-Mouse	*Mus musculus*	NP_034388	589
GATA2-Zebrafish	*Danio rerio*	NP_571308	456	xGATA6-*X. laevis*	*Xenopus laevis*	NP_001083725	502
GATA2-Mouse	*Mus musculus*	NP_032116	480	GATA6-Chicken	*Gallus gallus*	NP_990751	387
GATA2-Cow	*Bos taurus*	XP_583307	480	SpGATAc	*Strongylocentrotus purpuratus*	NP_999704	431
xGATA2-*X. laevis*	*Xenopus laevis*	NP_001084043	453	SpGATAe	*Strongylocentrotus purpuratus*	NP_001005725	567
GATA2-Chicken	*Gallus gallus*	NP_001003797	466	Ci-GATAa	*Ciona intestinalis*	BAE06471	641
GATA3-Human	*Homo sapiens*	NP_001002295	444	Ci-GATAb	*Ciona intestinalis*	BAE06472	553
GATA3-Zebrafish	*Danio rerio*	NP_571286	438	dGATAe	*Drosophila melanogaster*	NP_650516	746
GATA3-Mouse	*Mus musculus*	NP_032117	443	dGATAd	*Drosophila melanogaster*	NP_609383	842
GATA3-Cow	*Bos taurus*	NP_001070272	443	dGATAa	*Drosophila melanogaster*	P52168	540
xGATA3-*X. laevis*	*Xenopus laevis*	NP_001084335	435	dGATAb	*Drosophila melanogaster*	P52172	1264
GATA3-Chicken	*Gallus gallus*	NP_001008444	444	dGATAc	*Drosophila melanogaster*	P91623	486
GATA4-Human	*Homo sapiens*	NP_002043	442	ELT1	*C. elegans*	NP_001033435	488
GATA4-Zebrafish	*Danio rerio*	NP_571311	338	ELT2	*C. elegans*	NP_509755	433
GATA4-Mouse	*Mus musculus*	NP_032118	441	ELT3	*C. elegans*	AAD33964	226
GATA4-Cow	*Bos taurus*	XP_616466	442	ELT4	*C. elegans*	NP_741888	72+
GATA4-Chicken	*Gallus gallus*	XP_420041	410	ELT5	*C. elegans*	AAK32716	376
xGATA4-*X. laevis*	*Xenopus laevis*	AAB05647	392	ELT6	*C. elegans*	NP_500144	367
GATA5-Human	*Homo sapiens*	NP_536721	397	ELT7	*C. elegans*	AAC17756	198
GATA5-Zebrafish	*Danio rerio*	NP_571310	383	END1	*C. elegans*	NP_506475	221
GATA5-Mouse	*Mus musculus*	NP_032119	404	END3	*C. elegans*	NP_506480	242
GATA5-Cow	*Bos taurus*	NP_001029393	403				

+, partial sequence

2. Materials and methods

2.1 *Datasets*

Datasets of the amino acid sequences of *GATA* genes were collected from the GenBank database. Fifty-three full-length sequences of *GATA* genes of 6 vertebrates and 4 invertebrates (*Drosophila melanogaster*, *Caenorhabditis elegans*, *Ciona intestinalis* and *Strongylocentrotus purpuratus*) were obtained by BLAST and GenBank Entrez (table 1).

2.2 *Phylogenetic tree construction and alignment*

ClustalX software 1.81 was used to process multiple alignments. The matrix was set as the Gonnet series, and the parameters were set as follows. Gap opening penalty: 10; Gap extention penalty: 0.20; Delay divergent sequences: 30%. Phylogenetic trees were constructed by PHYLIP using the neighbour-joining (NJ) and maximum likelihood (ML) methods. The phylogenetic tree was analysed by the Treeview software. Alignment of C2C2 zinc finger domains were analysed by the Genedoc software.

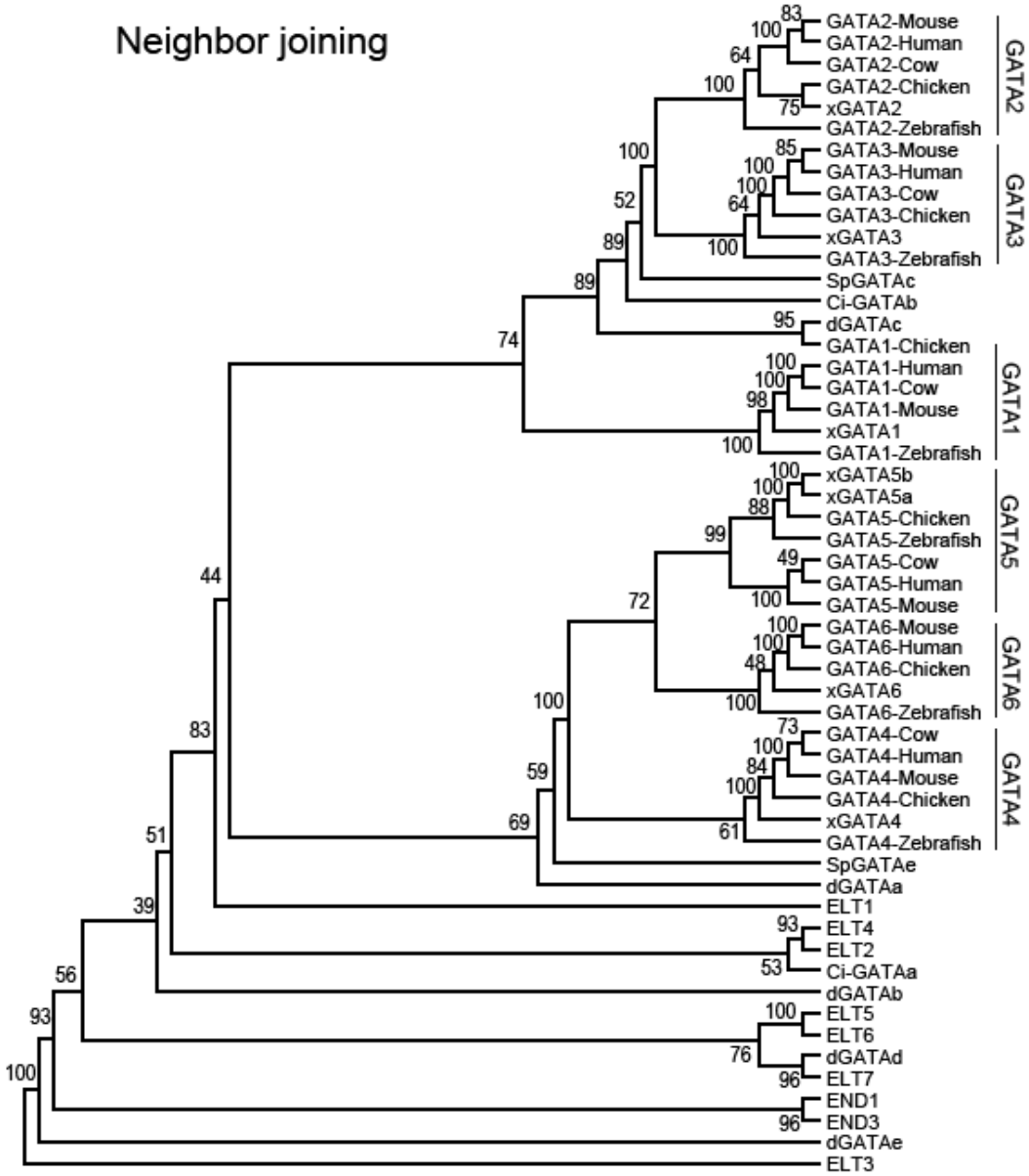


Figure 1. Phylogenetic analysis of vertebrate *GATA* family using the neighbour-joining (NJ) method. The NJ tree is constructed by PHYLIP. Numbers represent the bootstrap values (100 runs).The *GATA* gene family was divided into two subfamilies in vertebrates. Detailed information about each protein including GenBank accession numbers is listed in table 1.

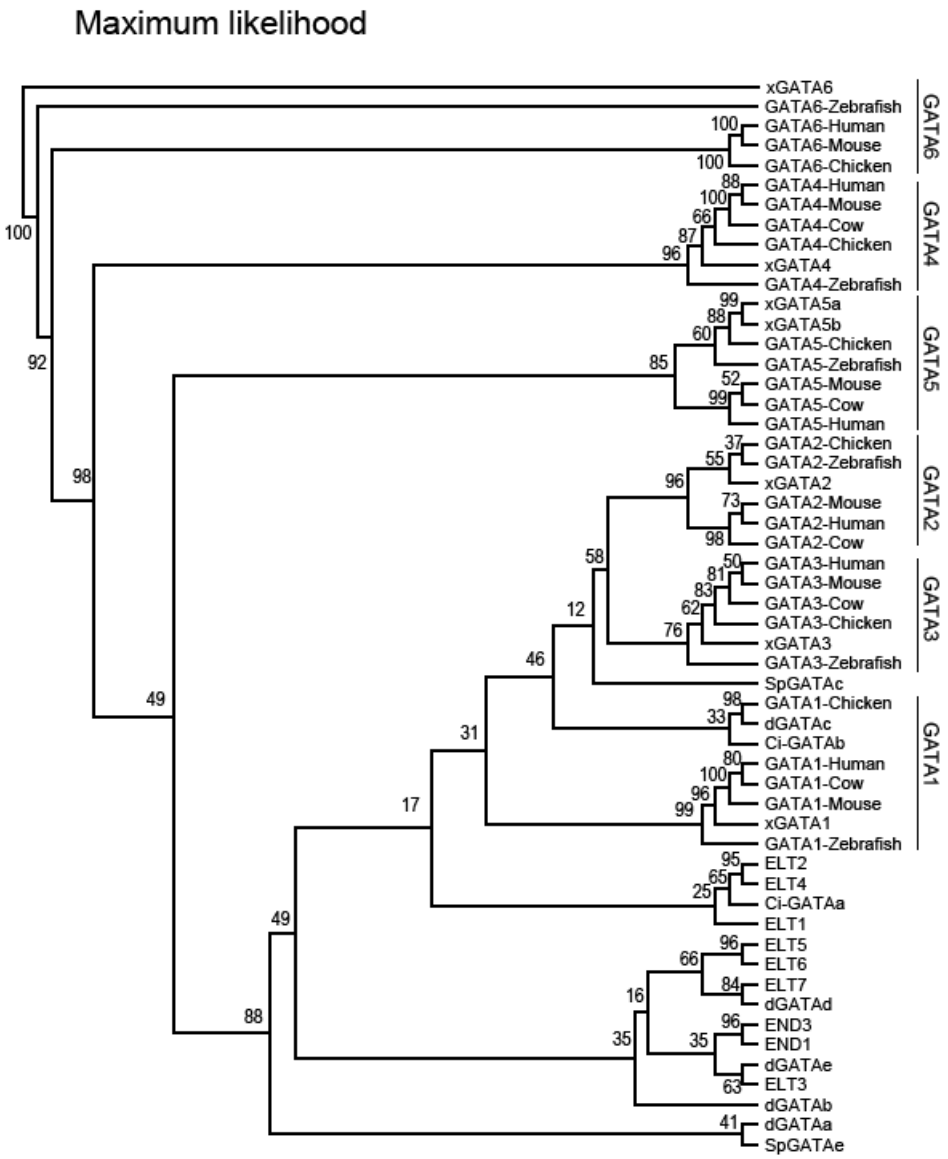


Figure 2. Phylogenetic analysis of vertebrate *GATA* family using the maximum likelihood (ML) method. The ML tree is constructed by PHYLIP. Numbers represent the bootstrap values (100 runs). The ML tree is basically consistent with the NJ tree. Detailed information about each protein including GenBank accession numbers is listed in table 1.

2.3 *Chromosome mapping*

Genes were mapped onto chromosomes based on the available genome resources of diverse species in the present databases (GenBank, Ensembl and UCSC). TBLASTN was used to align the amino acid and genomic sequences, and relevant genes were determined onto the chromosomes. To validate their locations, we searched gene information in GenBank (*http://www.ncbi.nlm.nih.gov/gene*). The results were confirmed by BLAST. Finally, the distribution of *GATA* genes on chromosomes in different organisms was analysed comparatively.

3. Results

3.1 *Phylogenetic tree of GATA genes in both vertebrates and invertebrates*

GATA genes were clustered into 6 groups, from *GATA1* to *GATA6* in vertebrates (figures 1 and 2). This result is consistent with their classification in human *GATAs*. According to the phylogenetic tree, *GATA* genes are divided into two subfamilies. Subfamily I contains *GATA1/2/3*, and subfamily II contains *GATA4/5/6*. The results of both the NJ and ML trees were basically identical. In invertebrates,

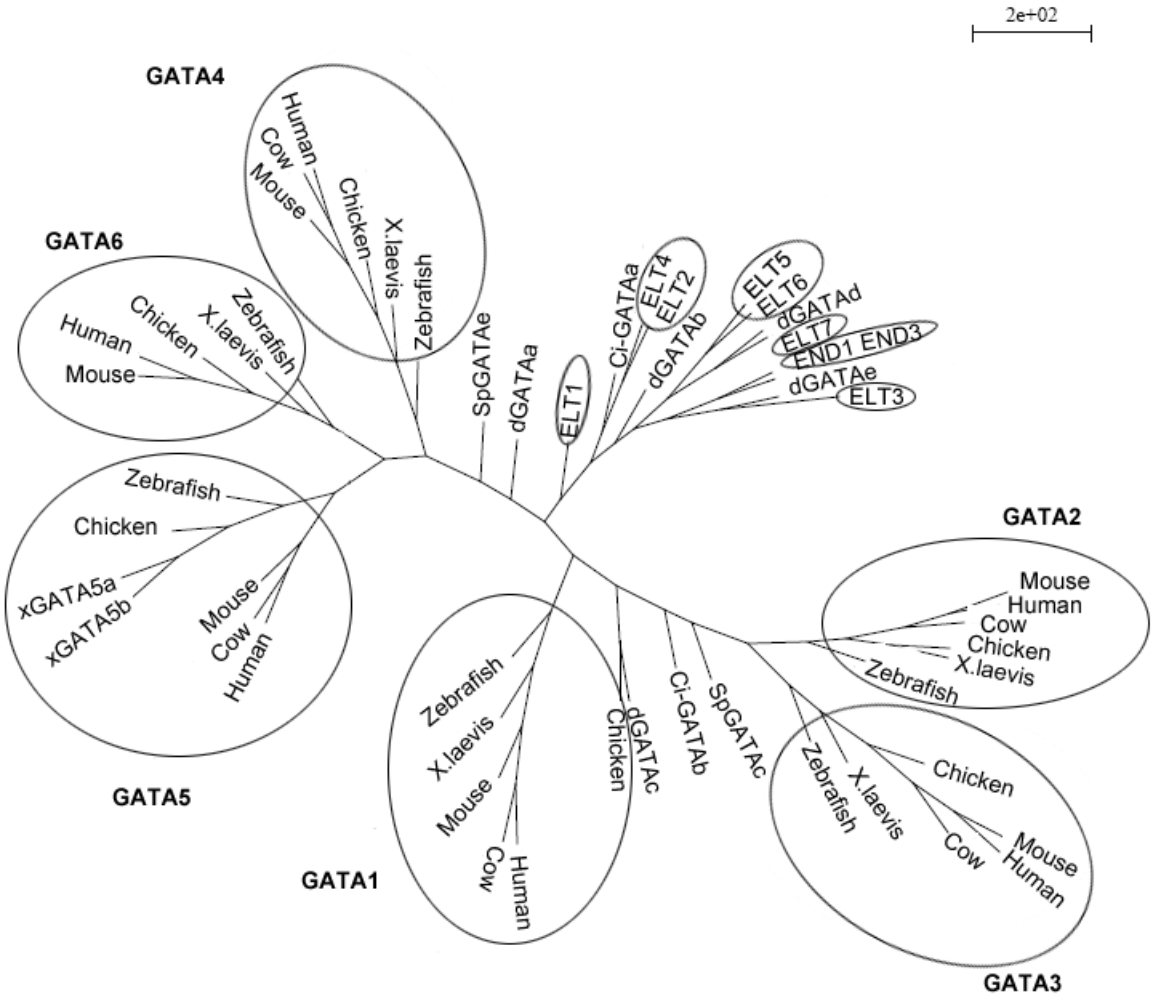


Figure 3. Unrooted tree of neighbour-joining method. The *GATAs* of vertebrates were grouped into six clusters (circled). Six protein groups of *C. elegans* were also circled. The unrooted tree represents the evolutionary distance among all the organisms.

GATAc of sea urchin (*SpGATAc*), *GATAc* of *Drosophila* (*dGATAc*) and *GATAb* of *C. intestinalis* (*Ci-GATAb*) were close to *GATA1* of vertebrates. *GATAe* of sea urchin (*SpGATAe*) and *GATAa* of *Drosophila* (*dGATAa*) were clustered with *GATA4* of high vertebrates. According to the evolutionary relationship, *GATAs* of *C. elegans* may be clustered into 6 groups of genes (figure 3).

3.2 *Duplications of* GATA *zinc finger domains*

Multiple alignments of zinc finger domains of 6 members of the GATA family of vertebrates revealed that two fingers have the same consensus: CXNCX4TX2WRX7*Φ*CNXC (*Φ*=V,L). Only a few bases had substitutions, which further determined their division into two subfamilies. For example, aa 720 of the N-finger in GATA1/2/3 is Q, but in GATA4/5/6, it is I, V or LS. aa 728 in GATA1/2/3 is K, but Q in GATA4/5/6 (figure 4). These two domains were present in all vertebrates, sea urchin and *C. intestinalis*. Most of GATAs of *Drosophila* and *C. elegans* did not contain the N-finger except *dGATAc*, *dGATAa* and *ELT1*. The C-finger domain was highly conserved and original. These results indicate that the duplication events of domains occurred during the evolutionary history from invertebrates to vertebrates and the N-finger was duplicated from the C-finger.

3.3 GATAs *segregated onto different chromosomes during evolution and conserved synteny of* GATAs *on chromosomes*

In invertebrates, *GATAs* were linked together on chromosomes (figure 5). However, in vertebrates, *GATAs*

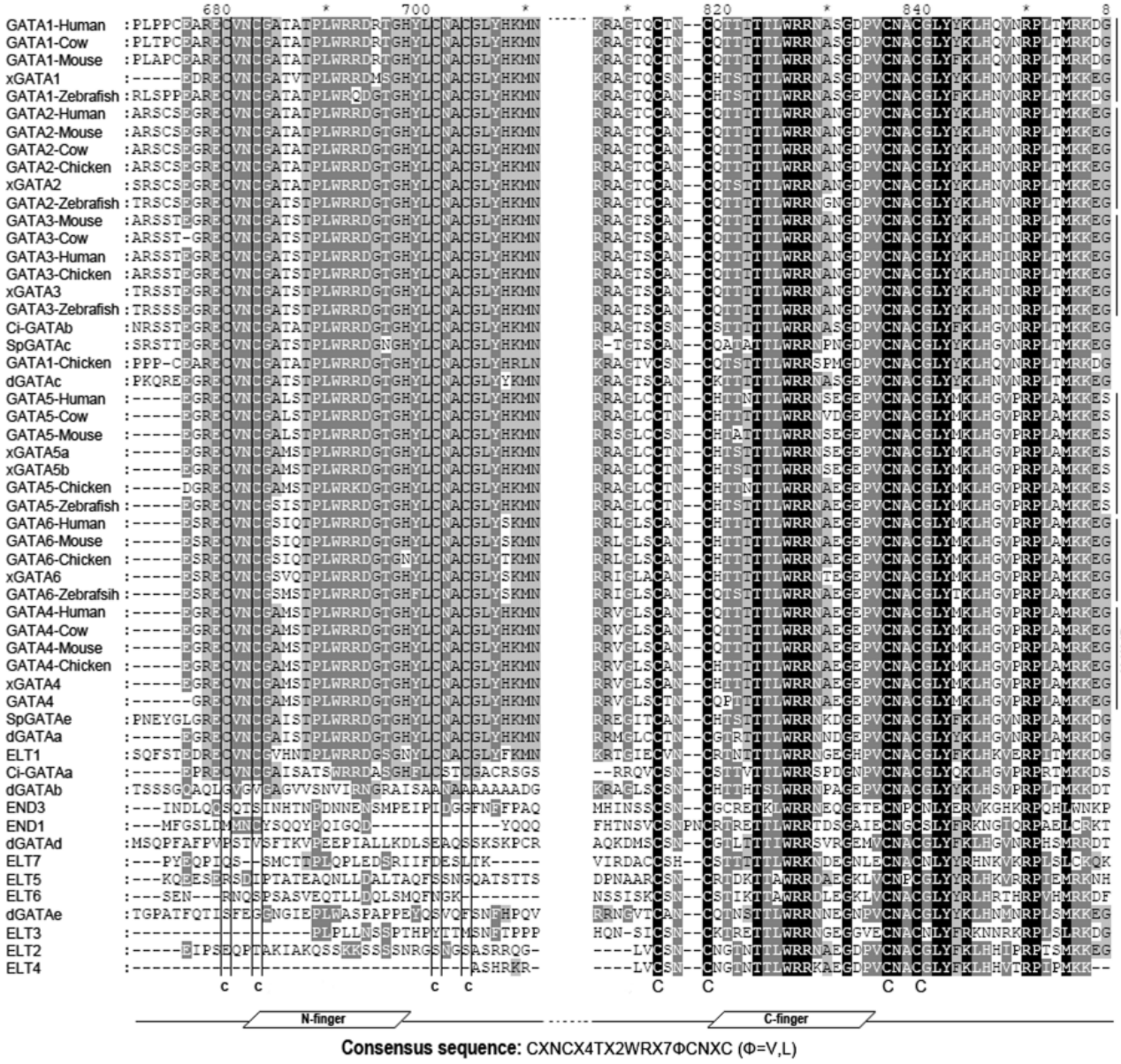


Figure 4. Amino acid sequence alignment of GATA zinc finger domains. Alignment of amino acid sequences of the GATA proteins showed a high level of sequence identities. Two zinc finger domains of each GATA have the consensus sequence: $CXNCX_4TX_2WRX_7$ *Φ*CNXC (*Φ*=V,L). GenBank accession numbers are the same as given in table 1.

tended to segregate onto different chromosomes. Especially in zebrafish and chicken, *GATA1* and *GATA2* were linked together on a chromosome, whereas in mammals, *GATA1* was located on chromosome X, and *GATA2* was assigned to an autosome. Six members of human *GATAs* were completely segregated onto different chromosomes. In addition, based on the positions of *GATA* and its flanking genes on a chromosome, a significant conserved synteny was observed among chromosomal regions around the *GATA* loci in vertebrates (figure 5).

4. Discussion

The *GATA* family is an evolutionarily conserved family of genes both in vertebrates and invertebrates. In vertebrates it contains six members. We collected all *GATA* genes of vertebrates and some genes of invertebrates available from the GenBank databases. Compared with a previous study (Lowry and Atchley 2000), our datasets are more abundant and the partial sequences used previously were replaced by full-length sequences available in the present databases.

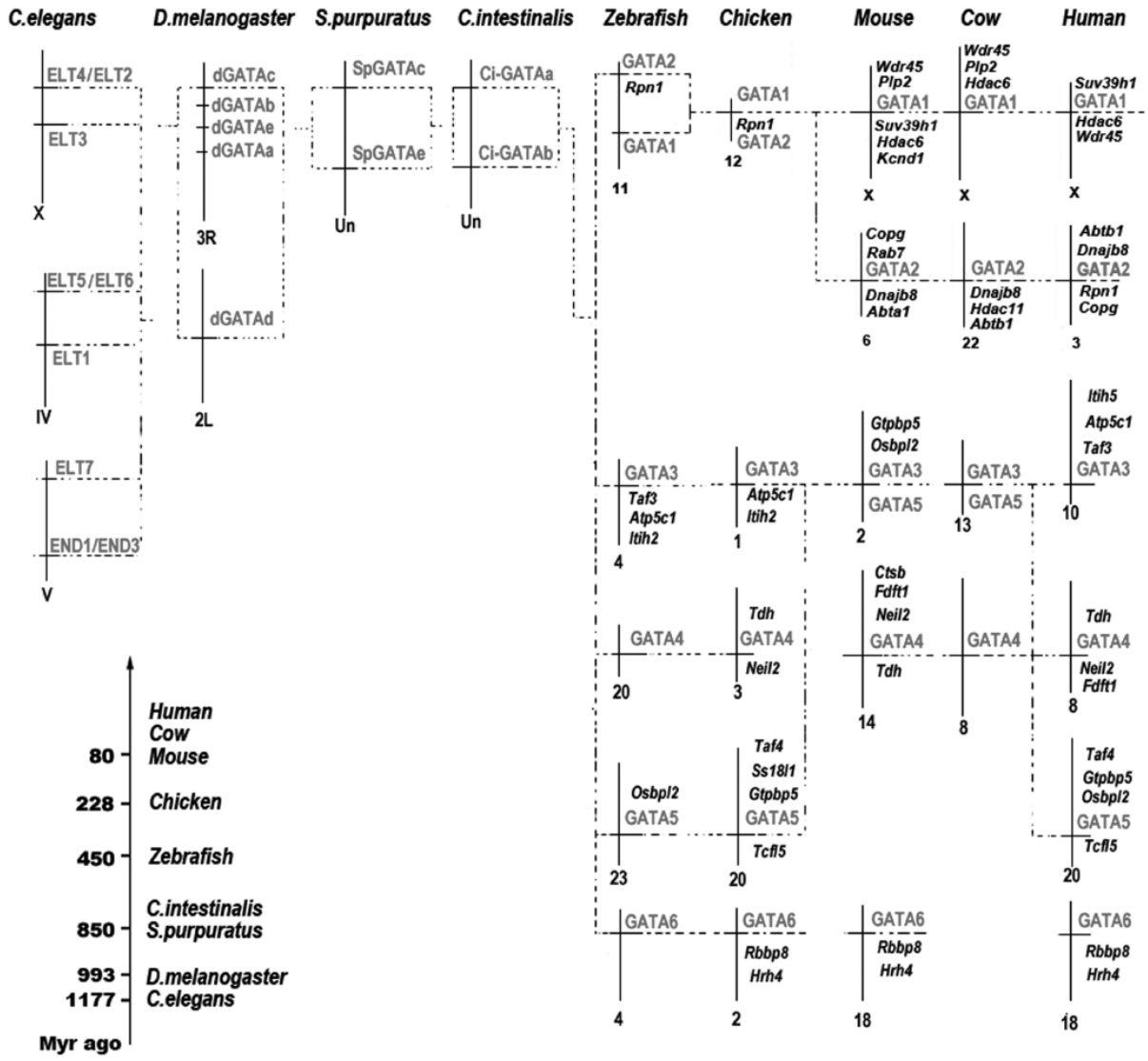


Figure 5. Comparative mapping of *GATAs* on chromosomes from invertebrates to mammals. The position of each *GATA* gene was from the NCBI Entrez. *GATA1* is on chromosome X, while *GATA2* is on autosomes. In chicken and zebrafish, both *GATA1* and *GATA2* are linked on one chromosome. Conserved synteny among *GATAs* and several close genes was seen. The evolutionary relationship of these species in million years (myr) are shown in the left panel.

A phylogenetic tree was constructed based on full-length amino acid alignments of the GATAs. The evolutionary relationship of the branches of the tree is credible because of high bootstrap values and consistency of topology structures by the two methods. The phylogenetic tree reveals that all members are clustered into two subfamilies (*GATA1/2/3* and *GATA4/5/6*) and this classification is consistent with their functions. *GATA1/2/3* mainly function in the haematopoietic system and *GATA4/5/6* play a role mainly in the cardiac system (Laverriere *et al* 1994; Patient and McGhee 2002; Yin and Herring 2005). The classification situation is also consistent with that of invertebrates such as *Drosophila* (Lowry and Atchley 2000; Fossett and Schulz 2001).

Analysis of zinc finger domains of the *GATA* family reveals that two fingers have the same consensus: CXNCX4TX2WRX7*Φ*CNXC (*Φ*=V,L), which is highly conserved in animals. However in *Arabidopsis* and rice, the form is CX2CX17-20CX2C (Reyes *et al* 2004). Plants and most invertebrates have only one finger. In *C. elegans*, only *ELT1* has two fingers and in *Drosophila*, two fingers exist only in two members, *dGATAa* and *dGATAc*. Nevertheless, all the *GATA* members of sea urchin and *C. intestinalis* have

two fingers. These results suggest that the zinc finger domain tends to duplicate during evolution from invertebrates to vertebrates. Furthermore, the fact that the C-finger domain exists in both invertebrates and vertebrates, but the N-finger in all members of vertebrates, sea urchin and *C. intestinalis*, and a few members of the genus *Drosophila* and *C. elegans* indicate that the C-finger domain is highly conserved and original, and the N-finger was duplicated from the C-finger domain. A more likely explanation of the finger domains is that the primitive *GATA* gene (with one finger) duplicated in early evolution to give two genes, one of which then duplicated the finger. The single-finger genes (*ELT2–7* genes in *C. elegans*, *Drosophila GATA b/d/e*) were derived from one of these, while the double-finger genes were derived from the other following another duplication giving the *GATA1/2/3* and *GATA4/5/6* groups. Therefore, lineage-specific expansions are responsible for much of the diversity of *GATA* in different animal genomes. It also appears that the original duplications occurred in tandem, as linkage is still seen in *Drosophila* and some other species.

The two zinc finger domains have functional differences. Earlier studies indicate that their DNA-binding sequences are also different. The C-finger mainly binds the consensus sequence (T/A)GATA(A/G), while the N-finger may bind with the consensus sequence (T/A)GATC(A/G) (Newton *et al* 2001; Trainor *et al* 2000). The functional difference may come from the divergence in evolution of the *GATA* gene family, which reflects the line of evolution and adaptation.

Although the GATA zinc finger may have duplicated during evolutionary history, the number of GATA members does not seem to have increased. Six member groups already existed in *C. elegans*. Furthermore, some members were lost during evolution, as only five members exist in *Drosophila* and two in Urochordates (*Ciona intestinalis*) and Echinoderms (sea urchin).

Although the possibility that some members are still to be identified among Urochordates and Echinoderms cannot be excluded, we infer that the *GATA* family may lose some members during early duplication events from *C. elegans* to Urochordates. These results suggest that evolutionary pressure on the GATAs must exist to balance duplications between the zinc fingers and the genes themselves. Even in the lineage of fish, evolutionary pressure is still on to keep the GATA members constant, although a third whole genome duplication event occurred in a branch of the teleost fish before 450 million years. Further work on the *GATA* family will of course help in understanding its biological functions, evolution and its relationship with associated diseases.

Acknowledgments

We thank one of the reviewers for suggesting some of the interpretations of these results. The work was supported by the National Natural Science Foundation of China, the National Key Basic Research Project (2006CB102103), the Program for New Century Excellent Talents in University and the 111 project #B06018. There are no financial conflicts of interest.

References

Evans T, Reitman M and Felsenfeld G 1988 An erythrocyte-specific DNA-binding factor recognizes a regulatory sequence common to all chicken globin genes; *Proc. Natl. Acad. Sci. USA* **85** 5976–5980

Fossett N and Schulz R A 2001 Functional conservation of hematopoietic factors in *Drosophila* and vertebrates; *Differentiation* **69** 83–90

Laverriere A C, MacNeill C, Mueller C, Poelmann R E, Burch J B and Evans T 1994 *GATA-4/5/6*, a subfamily of three transcription factors transcribed in developing heart and gut; *J. Biol. Chem.* **269** 23177–23184

Lowry J A and Atchley W R 2000 Molecular evolution of the *GATA* family of transcription factors: conservation within the DNA-binding domain; *J. Mol. Evol.* **50** 103–115

Newton A, Mackay J and Crossley M 2001 The N-terminal zinc finger of the erythroid transcription factor GATA-1 binds GATC motifs in DNA; *J. Biol. Chem.* **276** 35794–35801

Patient R K and McGhee J D 2002 The *GATA* family (vertebrates and invertebrates); *Curr. Opin. Genet. Dev.* **12** 416–422

Reyes J C, Muro-Pastor M I and Florencio F J 2004 The *GATA* family of transcription factors in *Arabidopsis* and rice; *Plant Physiol.* **134** 1718–1732

Trainor C D, Ghirlando R and Simpson M A 2000 GATA zinc finger interactions modulate DNA binding and transactivation; *J. Biol. Chem.* **275** 28157–28166

Yin F and Herring B P 2005 GATA-6 can act as a positive or negative regulator of smooth muscle-specific gene expression; *J. Biol. Chem.* **280** 4745–4752

MS received 15 May 2007; accepted 24 September 2007

*e*Publication: 10 October 2007

Corresponding editor: Stuart A Newman