GENOMICS **40,** 262–266 (1997)
ARTICLE NO. GE964591

Newly Identified Paralogous Groups on Mouse Chromosomes 5 and 11 Reveal the Age of a T-Box Cluster Duplication

ILYA RUVINSKY AND LEE M. SILVER[1]

Lewis Thomas Laboratory, Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544-1014

Received October 7, 1996; accepted December 26, 1996

A novel family of ancient transcription factors, the T-box family, involved in embryonic development in metazoans, was described recently. Four members of this family are grouped in two tightly linked pairs within the mouse genome. This arrangement can be explained by an original cluster formation followed by an *en masse* duplication. Here we demonstrate that this duplication event also included several closely linked genes. Using data obtained from linked paralogous genes, we show that the T-box cluster duplication occurred prior to the divergence between bony fish and tetrapods around 400 million years ago. This work facilitates our understanding of the status of the T-box gene family in different vertebrate lineages and also defines a novel paralogy group within the mouse genome.

INTRODUCTION

The vertebrate genome is composed of a large number of paralogous regions. This is a consequence of major steps in genome evolution such as multiple large-scale chromosomal duplication events as well as local duplications with subsequent chromosomal rearrangements. Recent advances in genomic mapping in mammals have revealed a number of chromosomal regions related by descent (Nadeau, 1991; Nadeau and Kosowsky, 1991; Lundin, 1993).

A novel gene family of putative transcription regulators sharing a conserved homology domain with the classical mouse gene *Brachyury* (*T*) was recently discovered and named the T-box family (Bollag *et al.*, 1994). Three genes, *Tbx1–Tbx3*, were originally described and found to be expressed at different stages of embryonic development. Later two additional genes were discovered, *Tbx4* and *Tbx5*, which mapped in close linkage with *Tbx2* and *Tbx3*, respectively (Agulnik *et al.*, 1996). Considering the close evolutionary relatedness of these two pairs of genes, a model of tandem duplication followed by cluster dispersion was proposed to account for the observed arrangement.

[1] To whom correspondence should be addressed. Telephone: (609) 258-5976. Fax: (609) 258-3345. E-mail: Lsilver@molbiol.princeton.edu.

If indeed such a dispersion did take place, it is expected that the other genes linked to the original cluster would be duplicated as well. If this were true, one should be able to uncover extended regions containing paralogous genes on chromosomes 5 and 11.

Here we present evidence that at least five pairs of related genes in addition to the T-box clusters demonstrate a genomic distribution pattern consistent with a proposed *en masse* duplication event. These genes, transcription factors (*Tcf*), acidic beta-crystallins (*Cryba*), nitric oxide synthases (*Nos*), LIM homeobox proteins (*Lhx*), and cAMP dependent regulatory protein kinases (*Prkar*), define a novel pair of paralogy groups within the vertebrate genome.

MATERIALS AND METHODS

Search Strategy

Candidates for paralogous pairs had to satisfy the following criteria: a gene had to be linked to the T-box cluster on one of the chromosomes, it had to be a member of a gene family with other members sequenced in the mouse (or rat) genome, and, finally, one of those paralogues had to map in the vicinity of the second T-box cluster. First, all sequenced genes mapping to the vicinity of one of the *Tbx* gene clusters were identified. Next, the protein sequences of these genes were used to perform TBLASTN searches (http://www.ncbi.nlm.nih.gov/BLAST). Finally, we determined map positions of mouse genes that showed high sequence identity to the query using the Mouse Genome Database (MGD; Jackson Laboratory, 1996).

Data Used

The following sequences were used in the present study (accession numbers are listed in parentheses):

Tcf genes. Chicken HNF1 (X67689), hamster HNF1 (M95297), human *TCF1* (M57732, J04771), mouse *Tcf1* (M57966), and *Xenopus laevis* HNF1 (X64759); human *TCF2* (X58840), mouse *Tcf2* (X55842), pig vHNF1 (X69675), rat vHNF1 (X56546), and *X. laevis* LFB3 (X76052); and salmon HNF1 (X79486).

Cryb genes. Chicken A1 (M15658), human *CRYBA1* (M14306), *Rana catesbeiana* A3 (X87761) (same as A1), and rat A1 (X15143); chicken A2 (U28145) and cow A2 (M60329); chicken A4 (U18260), cow A4 (M60328), and human *CRYBA4* (U59057); chicken B1 (M11619), cow B1 (X01808, M11850), human *CRYBB1* (U35340), and rat B1 (M13534, M13535); cow B2 (M22466), human *CRYBB2* (L10035), mouse *Crybb2* (M60559), rat B2 (X16072), and *R. catesbeiana* Bp (X91989); and chicken B3 (U28146) and rat B3 (M15901).

0888-7543/97 $25.00

Nos genes. Human *NOS1* (U17327), mouse *Nos1* (D14552), and rat NOS1 (X59949); human *NOS2* (L09210), mouse *Nos2* (U43428), rat NOS2 (D44591), and salmon NOS2 (X97013); cow NOS3 (M89952) and human *NOS3* (D26607); and *Drosophila melanogaster* (U25117) and *Rhodnius prolixus* (U59389).

Lhx genes. Chicken LIM1 (L35569), human *LHX1* (U14755), mouse *Lhx1* (Z27410), *X. laevis* LIM1 (X63889), and zebrafish LIM1 (L37802); and mouse *Lhx5* (U61155), *X. laevis* LIM5 (L42546), and zebrafish LIM5 (L42547).

Prkar genes. Human *PRKAR1A* (M18468), pig Prkar1a (X05942), and rat Prkar1a (M17086); human *PRKAR1B* (M65066) and mouse *Prkar1b* (M20473); and *Aplysia californica* PKA (X62382).

Phylogenetic Analyses

Amino acid sequences were aligned by eye using the ESEE sequence editor (Cabot and Beckenbach, 1989). Regions of uncertain alignment due to high variability or extensive length variation were omitted from the study. We analyzed amino acid sequences rather than nucleotide sequences because the latter become saturated faster with substitutions when distant evolutionary comparisons are performed. We have constructed neighbor-joining trees (Saitou and Nei, 1987) using Poisson-corrected distances. Statistical confidence of internal nodes was accessed by an interior-branch test of Rzhetsky and Nei (1992). Phylogenetic reconstructions were performed using the METREE program (Rzhetsky and Nei, 1994). We have used commonly accepted dates for the separation of major vertebrate lineages (Carroll, 1988).

RESULTS AND DISCUSSION

We have performed an exhaustive search of the cloned genes known to map within a 40-cM region surrounding the *Tbx3/Tbx5* cluster on chromosome 5 according to the scheme outlined under Materials and Methods. We have identified seven unrelated genes within this region that show homology to sequences mapping within the 40-cM region that surrounds the *Tbx2/Tbx4* cluster on chromosome 11. Five of these are discussed in detail below. The other two present more ambiguous cases and will be treated only briefly.

Murine Chromosomes 5 and 11 Contain Paralogous Groups

Figure 1 shows the maps of the relevant regions of chromosomes 5 and 11.

Tcf family. Two members of this family, *Tcf1* and *Tcf2*, map within 4 and 5 cM from the *Tbx3/5* and *Tbx2/4* clusters, respectively. They are POU- and homeodomain-containing transcription factors, which are also known as hepatic nuclear factors (HNF) and variant HNF or LFB3. It should be noted that Lundin (1993) has previously reported that members of this family are found on both chromosomes 5 and 11.

Cryb family. This family is composed of genes encoding major structural proteins of the vertebrate eye lens. *Cryba4* maps about 10 cM proximal to the *Tbx3/5* cluster, while *Cryba1* appears to be 5 cM away from the *Tbx2/Tbx4* cluster. These genes clearly satisfy the criteria described above.

Nos family. There are three genes for nitric oxide synthases identified in the mammalian genome. In the mouse two genes map to chromosome 5 (*Nos3* proximally at 9.0 cM and *Nos1* distally at 70.0 cM) and one (*Nos2*) to chromosome 11 (46.0 cM). Thus, *Nos1* and *Nos2* probably represent a pair of genes that trace their origin to the same gene duplication, which generated two T-box clusters.

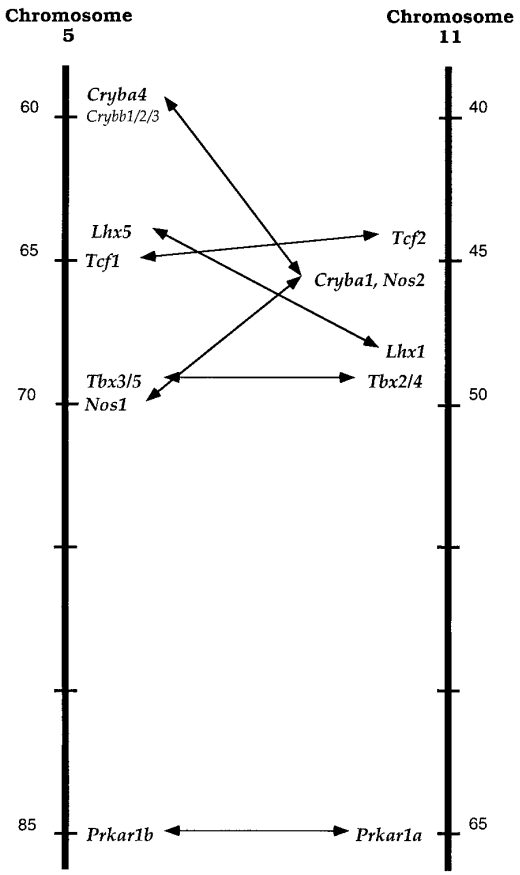

FIG. 1. The map of paralogous groups on mouse chromosomes 5 and 11. Arrows indicate paralogous loci. The map is drawn to scale. Only the genes discussed in the text are shown.

Lhx family. This is a family of LIM-containing homeobox transcription factors. *Lhx1* was mapped just 1 cM proximal to the *Tbx2/Tbx4* cluster on chromosome 11, and *Lhx5* was recently localized 5.8 cM distal to *Cryba4* and 8.5 cM proximal to the *Gus* gene, hence around 64 cM on chromosome 5 (Bertuzzi *et al.*, 1996).

Prkar family. Genes for α and β subunits of cAMP dependent regulatory protein kinase type 1 map 16 cM distal from the *Tbx2/Tbx4* and *Tbx3/Tbx5* clusters, respectively.

In addition to the genes described above, there also are two gene families that are potential candidates for belonging to the paralogy group surrounding the T-box genes.

One of them, the zinc finger protein family, has one gene (*Zfp38*) mapping to 78.5 cM on chromosome 5 and four genes (*Zfp2, Zfp3, Zfp147,* and *Zfp179*) located in the vicinity of the *Tbx2/Tbx4* cluster on chromosome 11. However, only two of these were mapped precisely

in the mouse (*Zfp2* and *Zfp3;* 45.0 and 41.0 cM, respectively), while the other two were mapped only in the human (17q21.3–q22 and 17p11.2), suggesting that murine homologues of these genes could be linked to the T-box cluster on chromosome 11. Due to the large size of the *Zfp* gene family (over 100 members), paralogy relations are obscure. Here we can only speculate that such a relationship could exist between *Zfp38* and one (or more) of the *Zfp* genes on chromosome 11, and it can be attributed to the same duplication event that created two T-box clusters.

The other gene family consists of the *Rpo2-1* (chromosome 5, 72.0 cM) and *Rpo2-2* (chromosome 11, 37.0 cM) genes encoding RNA polymerase II-1 and RNA polymerase II-2, which possibly represent a paralogous gene pair. However, paucity of information about *Rpo2-2* precludes more detailed analysis of these genes.

The accumulated data demonstrate the presence of a novel pair of paralogy groups covering at least 5–10 cM on murine chromosomes 5 and 11 and clustering around tandemly duplicated T-box genes on each chromosome. If the linkage between each paralogous group and a corresponding *Prkar1* gene is indeed not coincidental, it will increase the size of the original duplicated fragment to at least 20–25 cM.

It should be noted that although blocks of conserved linkage exist on two chromosomes, the order of genes is not strictly conserved. In addition there are some cases when a gene from one of the blocks does not have a paralogue in the other one. This may be explained by both limited mapping accuracy and local rearrangements that occurred subsequent to the initial duplication event. Also not all the genes between the flanking markers of these clusters have been mapped and/or sequenced, while some genes may have been lost during evolution.

Conservation of Paralogous Groups in Genomes of Other Vertebrates

Human homologues of the genes from the paralogy group on mouse chromosome 11 map to chromosome 17, while homologues of the genes from the mouse chromosome 5 cluster are found on human chromosomes 12 (*Tcf1, Nos1*), 22 (*Cryba1*), and 7 (*Prkar1b*). Although only limited information is available concerning other species, large chromosomal fragments encompassing the region of interest on mouse chromosome 11 show linkage conservation in owl monkey (chromosome 23), cattle (19), sheep (11), pig (12), and rat (10). In contrast, little can be said about the chromosome 5 cluster. These data are derived from the report of Wakefield and Graves (1996).

The Age of Paralogous Groups

Tcf family. The phylogenetic tree of 12 vertebrate *Tcf* genes based on an alignment of 375 amino acids is shown in Fig. 2A. It is an unrooted tree, as we lacked an appropriate outgroup, therefore it is difficult to conclude whether the split between the *Tcf1* and the *Tcf2* subfamilies occurred before or after the separation between fish and tetrapods around 400 million years ago (MYA). It is indeed possible that due to a higher rate of evolution among *Tcf1* subfamily members (see the tree), salmon HNF1 has accumulated a large number of amino acid replacements that forced its placement outside *Tcf1* and *Tcf2* subfamilies in the UPGMA analysis (not shown). Also the gene structure of salmon HNF1 is more similar to those of the rest of the genes in the *Tcf1* subfamily than it is to those of the genes of the *Tcf2* subfamily. It can be argued that this may serve as evidence of orthology between the salmon HNF1 gene and the other *Tcf1* genes, although alternative interpretations are possible. In any case, divergence between the two subfamilies has taken place prior to amniote–amphibian separation around 365 MYA.

Cryb family. The phylogenetic analysis (based on 185 amino acids) presented in Fig. 2B clearly demonstrates subdivision of the family into the two subgroups, the acidic and basic crystallins. Unfortunately, no complete sequence of acidic crystallin is known from a fish, therefore complicating the establishment of the age of this gene duplication. Nevertheless, within the acidic beta-crystallins, it is evident that *Cryba1* and *Cryba4* diverged prior to the split between the amphibians and the amniotes around 365 MYA.

Nos family. A phylogenetic tree of this family is presented in Fig. 2C. It can be seen that even though the relationships among the three genes are not well resolved, all three originated prior to the separation between bony fish and land vertebrates, hence more than 400 MYA. Unfortunately, only a partial sequence of fish nitric oxide synthase is available, therefore, the number of amino acids used in the analysis was reduced to 240. Also it should be noted that the two invertebrate genes were used as outgroups, since they were placed most basally in a UPGMA analysis (not shown) when complete sequences (without salmon NOS2) were analyzed.

Lhx family. The phylogenetic analysis of these genes (Fig. 2D) based on 412 amino acids clearly separates them into two distinct groups—*Lhx1* and *Lhx5.* Since each subgroup contains a fish gene it is clear that the divergence between *Lhx1* and *Lhx5* had to have occurred prior to the separation of the fish and tetrapod lineages approximately 400 MYA.

Prkar family. The utility of the phylogenetic analysis involving these genes is limited as the only vertebrate sequences available are the mammalian ones, thus making it impossible to date the divergence precisely. However, based on the tree topology (not shown), it is possible to conclude the it occurred subsequent to the divergence between the mollusks (represented by *Aplysia*) and the vertebrates on one hand and mamma-

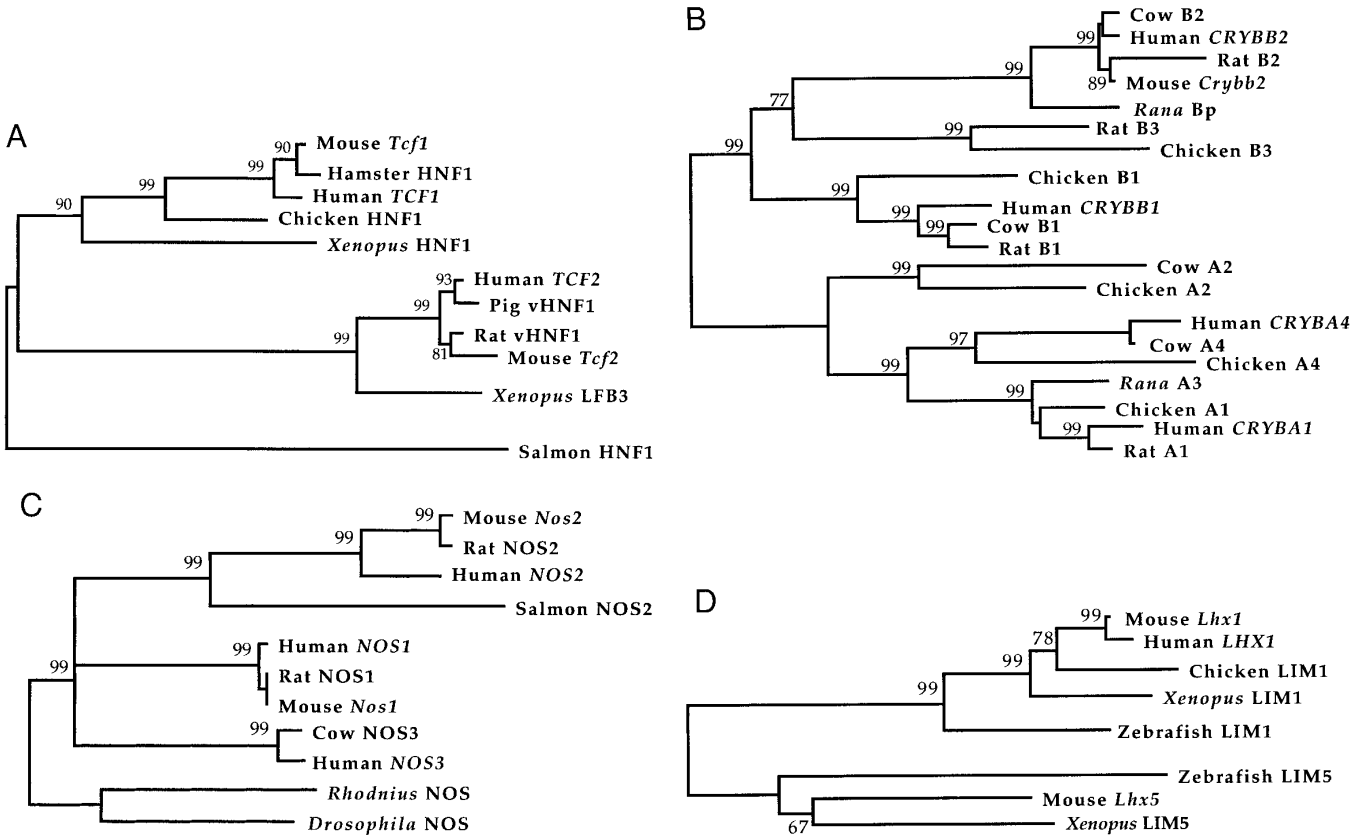

FIG. 2. Phylogenetic relationships within the gene families represented in paralogous groups on chromosomes 5 and 11. (**A**) *Tcf* gene family; (**B**) *Cryb* gene family; (**C**) *Nos* gene family; (**D**) *Lhx* gene family. All trees except for **C** are unrooted. Values above nodes indicate confidence levels. See text for more details.

lian radiation on the other hand and, perhaps, much closer to the former.

Present-Day Gene Function and Evolution of Developmental Complexity

It is intriguing that our time estimate of the origin of two T-box clusters derived purely from genomic considerations is in remarkable correspondence with the date proposed previously based on developmental data (Gibson-Brown *et al.*, 1996). Two of the T-box genes, *Tbx2* and *Tbx3*, are expressed in similar patterns in both anterior and posterior mesenchyme during development of fore- and hindlimb buds. If these appendages of modern tetrapods are a result of a rostral homeotic transposition as proposed by Tabin and Laufer (1993), involvement of a cognate gene pair (*Tbx2* and *Tbx3*) in similar expression domains in both appendages would necessitate its origin to be more ancient than that of tetrapods, i.e., over 365 MYA. Subsequently these genes were probably recruited in development of an autopod, a tetrapod-unique structure, as reflected in their complementary expression patterns in digit tips. Members of the second gene pair, *Tbx5* and *Tbx4*, have acquired unique roles in fore- and hindlimb development, respectively. Close linkage within the pairs provides a potential for coordinate regulation, which may be a selective force that maintained this linkage through over 600 million years of evolution (Agulnik *et al.*, 1996).

CONCLUSIONS

We have reported the discovery of a novel pair of paralogous groups within the mouse (vertebrate) genome, which spans no less than 5–10 cM (perhaps even 20–25 cM) and is located around the T-box gene clusters on mouse chromosomes 5 and 11. This discovery confirms our previous hypothesis that an *en masse* duplication was responsible for the creation of two T-box clusters after an initial formation of a single cluster. We have estimated the age of this duplication event to be over 400 million years, i.e., before the separation of the lineages leading to bony fish and tetrapods. If correct, this estimate predicts that the genomic arrangement of the two T-box clusters should be similar in all vertebrates. It is also in concord with the apparent roles that the *Tbx2–Tbx5* genes play during embryonic development. Isolation and characterization of these genes from a bony fish is under way and should definitively settle the question regarding the age of their origin.

ACKNOWLEDGMENTS

The authors are grateful to Anatoly Ruvinsky for providing an excellent working atmosphere and invaluable discussions during the early stages of this project. We have benefited from helpful conversations with Sergei Agulnik and Jeremy Gibson-Brown. This work was supported by a National Institutes of Health grant (HD-20275) to L.M.S.

REFERENCES

Agulnik, S. I., Garvey, N., Hancock, S., Ruvinsky, I., Chapman, D. L., Agulnik, I., Bollag, R., Papaioannou, V., and Silver, L. M. (1996). Evolution of mouse T-box genes by tandem duplication and cluster dispersion. *Genetics* **144:** 249–254.

Bertuzzi, S., Sheng, H. Z., Copeland, N. G., Gilbert, D. J., Jenkins, N. A., Taira, M., Dawid, I. B., and Westphal, H. (1996). Molecular cloning, structure, and chromosomal localization of the mouse LIM/homeobox gene *Lhx5*. *Genomics* **36:** 234–239.

Bollag, R. J., Siegfried, Z., Cebra-Thomas, J., Garvey, N., Davison, E. M., and Silver, L. M. (1994). An ancient family of embryonically expressed mouse genes sharing a conserved protein motif with the *T* locus. *Nature Genet.* **7:** 383–389.

Cabot, E. L., and Beckenbach, A. T. (1989). Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. *Comput. Appl. Biosci.* **5:** 233–234.

Carroll, R. L. (1988). "Vertebrate Paleontology and Evolution," Freeman, New York.

Gibson-Brown, J. J., Agulnik, S. I., Chapman, D. L., Alexiou, M., Garvey, N., Silver, L. M., and Papaioannou, V. E. (1996). Evidence of a role for T-box genes in the evolution of limb morphogenesis and the specification of forelimb/hindlimb identity. *Mech. Dev.* **56:** 93–101.

Jackson Laboratory. (1996). MGD: The Mouse Genome Database, *available* http://www.informatics.jax.org/mgd.html.

Lundin, L. G. (1993). Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* **16:** 1–19.

Nadeau, J. H. (1991). Genome duplication and comparative gene mapping. *In* "Advanced Techniques in Chromosome Research" (K. W. Adolph, Ed.), pp. 269–296, Dekker, New York.

Nadeau, J. H., and Kosowsky, M. (1991). Mouse map of paralogous genes. *Mamm. Genome* **1:** S433–460.

Rzhetsky, A., and Nei, M. (1992). A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol.* **9:** 945–967.

Rzhetsky, A., and Nei, M. (1994). METREE: A program package for inferring and testing minimum-evolution trees. *Comput. Appl. Biosci.* **10:** 409–412.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.

Tabin, C., and Laufer, E. (1993). *Hox* genes and serial homology. *Nature* **361:** 692–693.

Wakefield, M. J., and Graves, J. A. M. (1996). Comparative maps of vertebrates. *Mamm. Genome* **7:** 715–716.