Complex History of a Chromosomal Paralogy Region: Insights from Amphioxus Aromatic Amino Acid Hydroxylase Genes and Insulin-Related Genes

School of Animal and Microbial Sciences, University of Reading, Whiteknights, U.K.

Aromatic amino acid hydroxylase (AAAH) genes and insulin-like genes form part of an extensive paralogy region shared by human chromosomes 11 and 12, thought to have arisen by tetraploidy in early vertebrate evolution. Cloning of a complementary DNA (cDNA) for an amphioxus (Branchiostoma floridae) hydroxylase gene (AmphiPAH) allowed us to investigate the ancestry of the human chromosome 11/12 paralogy region. Molecular phylogenetic evidence reveals that AmphiPAH is orthologous to vertebrate phenylalanine (PAH) genes; the implication is that all three vertebrate AAAH genes arose early in metazoan evolution, predating vertebrates. In contrast, our phylogenetic analysis of amphioxus and vertebrate insulin-related gene sequences is consistent with duplication of these genes during early chordate ancestry. The conclusion is that two tightly linked gene families on human chromosomes 11 and 12 were not duplicated coincidentally. We rationalize this paradox by invoking gene loss in the AAAH gene family and conclude that paralogous genes shared by paralogous chromosomes need not have identical evolutionary histories.

Introduction

The past decade has witnessed a rapid expansion in the number of mammalian genes mapped to chromosomal locations. As data have accumulated, patterns have become apparent. Chromosomal paralogy regions are sets of linked genes, each having homologues linked on another chromosome (Lundin 1993). For example, the genes for phenylalanine hydroxylase, FGF6, insulin-like growth factor 1 (IGF-1), lactate dehydrogenase B, and two myogenic bHLH proteins, K-ras and parathyroid hormone-like hormone, map to human chromosome 12; all these genes have close homologues on human chromosome 11. The simplest explanation for this pattern is common ancestry: human chromosomes 11 and 12 probably arose by duplication of an ancestral chromosome, perhaps by tetraploidy of the genome (Lundin 1993).

There is now considerable evidence that tetraploidy did occur during early vertebrate evolution. Susumo Ohno was the first to seriously suggest this notion in his seminal book (Ohno 1970). More recently, analysis of gene family complexity in vertebrates and in the closely related invertebrate, amphioxus, has given strong support to the notion that the genome underwent at least one, and possibly two, rounds of tetraploidy in early vertebrate evolution (Holland et al. 1994; Sharman and Holland 1996; Sharman, Hay-Schmidt, and Holland 1997).

If chromosomal paralogy regions arose by tetraploidy during early vertebrate evolution, it seems reasonable to assume that each pair of related genes arose simultaneously at this time. This assumption has never been adequately tested, since it requires DNA sequence information from a close outgroup for each of two or more linked gene families within a paralogy group. Amphioxus may be the ideal outgroup for such analyses since it is the sister group of the vertebrates (defined here as synonymous with craniates) and is thought to have branched from the chordate lineage just before the putative tetraploidy events. Within the chromosome 11/12 paralogy group, the physical linkage between aromatic amino acid hydroxylase (AAAH) genes and insulin-like genes is particularly close. Chromosome 11 has tryptophan hydroxylase (TPH) and tyrosine hydroxylase (TH), while chromosome 12 has phenylalanine hydroxylase (PAH). Chromosome 11 has insulin and IGF-2, while chromosome 12 has IGF-1 (fig. 1). An amphioxus homologue of the vertebrate insulin-like genes (insulin, IGF-1, and IGF-2) has already been cloned (Chan, Cao, and Steiner 1990) and named insulin-like peptide (ILP). The physically closest genes to IGF-1 on human chromosome 12 and to IGF-2 and to insulin on human chromosome 11 are members of the AAAH gene family. We therefore elected to clone an amphioxus member of this gene family. Molecular phylogenetic analyses of the amphioxus AAAH gene and ILP gene sequences, in relation to vertebrate and other metazoan homologues, revealed the dates of gene duplication in each gene family. We find that the gene duplication yielding AAAH genes on chromosomes 11 and 12 occurred at a vastly different time from the duplication that yielded IGF genes on chromosomes 11 and 12.

Materials and Methods

Isolation and Characterization of cDNA Clones for AAAH Genes

An amphioxus cDNA library in λ-ZAPII (made from 5- to 24-h embryos and a gift from J. Langeland, University of Kalamazoo, Mich.) was screened with a probe (p299) containing exons 6 to 11 of a genomic cosmid clone for amphioxus PAH (Patton and Holland, unpublished data). The genomic clone was originally found by library screening with a polymerase chain re-
action fragment amplified with primers designed from vertebrate AAAH genes. Hybridizations were carried out in 7% SDS and 250 mM sodium phosphate buffer pH 7.2 at 65°C, and hybridizing clones were plaque purified. cDNA was rescued into pBSII SK+ (Stratagene), restriction mapped, and sequenced in both directions by a combination of subcloning and walking with specific oligonucleotide primers.

Molecular Phylogenetic Analysis of AAAH Genes and Insulin-Related Genes

Aromatic amino acid hydroxylase gene or cDNA sequences were obtained through text and sequence similarity searches against GenBank databases including data from the Nematode Genome Sequencing Project (Sanger Centre, Hinxton, Cambridge, U.K.). The sequences used were the following: human TH type 1 (Grima et al. 1987; GenBank X05290); human TPH (Boularand et al. 1990; GenBank X52836); human PAH (Kwok et al. 1985; GenBank K03020); bovine TH (D’Mello et al. 1988; SwissProt P17289); rabbit TPH (Grenett et al. 1987; GenBank M17250); mouse TH (Ichikawa, Sasaoka, and Nagatsu 1991; GenBank M69200); mouse TPH (Stoll, Kozak, and Goldman 1990; GenBank J04758); mouse PAH (Ledley et al. 1990; GenBank X51942); quail TH (Fauquet et al. 1988; GenBank M24778); chick TPH (Florez et al. 1996; GenBank U26428); frog TPH (Green and Besharse 1994; GenBank L20679); fly TH (Neckameyer and Quinn 1989; SwissProt P18459); fly PAH (Neckameyer and White 1992; SwissProt P17276); and nematode TH, TPH, and PAH (unpublished sequences citing Wilson et al. 1994; WEBACE B0432.5, C36H8.3, and K08F8.4). Vertebrate, amphioxus, Drosophila, and Caenorhabditis AAAH protein sequences were aligned using the CLUSTAL V program (Higgins, Bleasby, and Fuchs 1991). A distance matrix was constructed from the confidently alignable regions at all sequences (323 sites), excluding regions of dubious alignment, and the phylogeny was deduced by the neighbor-joining method of the PHYLIP package, version 3.572c (Felsenstein 1993). Confidence in each node of the phylogeny was assessed by performing 100 bootstrap resamplings of the data. Neighbor joining does not assume a molecular clock or a root position.

An identical strategy was undertaken for the insulin-related genes, except that a greater diversity of species could be included. We included human INS (Bell et al. 1980; GenBank J00265), human IGF-1 (Rotwein et al. 1986; SwissProt P05019), human IGF-2 (Dull et al. 1984; SwissProt P01344), rat insulin (Lomedico et al. 1979; SwissProt P01323), rat IGF-1 (Shimatsu and Rotwein 1987; SwissProt P08025), mouse IGF-2 (Rotwein and Hall 1990; GenBank U71085), rabbit insulin (Devaskar et al. 1994; GenBank U03610), chick insulin (Perler et al. 1980; SwissProt P01332), chick IGF-1 (Kajimoto and Rotwein 1989; SwissProt P18254), frog insulin (Shuldiner et al. 1989; SwissProt P12707), frog IGF-1 (Kajimoto and Rotwein 1990; SwissProt P16501), salmon insulin (Koval, Petrenko, and Kavsan 1989; SwissProt P04667), trout IGF-1 and IGF-2 (Shambrott and Chen 1992; SwissProt Q02815 and Q02816), dogfish insulin (Bajaj et al. 1983; SwissProt P12704, dogfish IGF-1 and IGF-2 (Duguay et al. 1995; GenBank Z50081 and Z50082), hagfish insulin (Chan el al. 1981; SwissProt P01342), hagfish IGF (Nagamatsu et al. 1991; SwissProt P22618), amphioxus ILP (Chan, Cao, and Steiner 1990; SwissProt P22334), and ascidian insulin and IGF (McRory and Sherwood 1997). Analysis was restricted to the confidently alignable A and B domains (50 sites); large regions of extreme length variability outside these domains (or inside for ascidian IGF) were deliberately excluded to avoid problems with gap weighting. Both analyses were performed without outgroups, since alignments suggested that prokaryotic AAAH genes (Zhao et al. 1994) and relaxin-like genes (Hudson et al. 1983) were too distantly related to provide reliable root positions.

Results

Cloning of an Amphioxus AAAH

Screening an amphioxus cDNA library with a genomic AAAH probe identified six hybridizing clones. Sequencing revealed that all clones derived from the same gene: all six clones started from the same 5' end. Alignment to known AAAH protein sequences suggests that the clone is missing approximately 14 codons (3%) of the protein. Since the extreme N-terminus of AAAH proteins is highly variable, it would be uninformative for our phylogenetic analyses in any case, and the near-full-length clone obtained is

Fig. 1—Chromosomal paralogy region shared by human chromosomes 11 and 12. Examples of genes mapping to 11p15 are shown; the dotted lines indicate gene family relationships to genes mapping to 12p12 or to 12q21–22. Chromosome 12 is drawn upside down to facilitate comparison. Map positions are from Online Mendelian Inheritance in Man at http://www.ncbi.nlm.nih.gov/Omim/.
likely to contain ample phylogenetic information. A putative polyadenylation signal was found starting at nucleotide 2837, followed 17 bp later by the poly (A) tail. Database searches using the deduced protein sequence revealed close similarity to known eukaryote AAAH genes, with a particularly strong similarity to PAH genes. On the basis of similarity, we designated this cDNA *AmphiPAH*; molecular phylogenetics provide a test of this hypothetical assignment.

**Molecular Phylogeny of the AAAH Gene Family**

To determine the phylogenetic relationship of *AmphiPAH* to other aromatic amino acid hydroxylase genes, we first aligned it with the deduced protein sequences of all eukaryote AAAH genes. This included three unpublished nematode AAAH sequences deposited in GenBank by the *Caenorhabditis elegans* genome-sequencing project. All sequences aligned extremely well, with a single exception: the putative polyadenylation signal

---

**Fig. 2**—Nucleotide and deduced amino acid sequence of amphioxus aromatic amino acid hydroxylase (AAAH) cDNA, *AmphiPAH*. Amino acid residues shown in bold cover the region used in phylogenetic analyses, after optimal alignment to 16 AAAH sequences. Asterisk, stop codon; underline, polyadenylation signal. GenBank/EMBL AJ001677.
Fig. 3—Phylogenetic relationships of eukaryote aromatic amino acid hydroxylase (AAAH) proteins, inferred by the neighbor-joining method. Figures on branches signify the percentage of times the sequences to the right of that node were found together following bootstrap resampling of the data. Nodes receiving below 60% bootstrap support were collapsed, leaving only well-supported aspects of the tree. Branch lengths are to scale; the scale bar denotes 0.1 inferred mutations per site. The tree is not rooted because of the absence of a suitable outgroup; nonetheless, TH, TPH, and PAH clearly fall as distinct genes across all metazoan taxa analyzed.

well across the C-terminal two-thirds of each enzyme. Seventy residues were absolutely conserved in all eukaryote hydroxylase proteins. A phylogenetic tree constructed by neighbor joining is shown in figure 3 with confidence assessed by bootstrapping.

The results show that each of the three hydroxylase gene types (TH, TPH, and PAH) forms a monophyletic group supported by very high bootstrap values. The position of the amphioxus hydroxylase gene confirms that it is directly orthologous to human PAH; the same conclusion applies to the *Drosophila* gene previously described as a possible precursor of PAH and TPH (Neckameyer and White 1992). The three nematode AAAH genes sequenced by the nematode genome project are clearly orthologous to TH, TPH, and PAH. The tree is consistent with the hypothesis of two separate phases of gene duplications in this gene family giving rise to three genes (TH, TPH, and PAH), but we cannot firmly deduce the relative order of the two duplications because of the absence of an outgroup. The clear conclusion is that all three gene types (TH, TPH, and PAH) originated by duplication prior to the divergence of nematodes, ar-
Fig. 4—Phylogenetic relationships of insulin-related proteins, inferred by the neighbor-joining method. Percentage bootstrap support is shown only where this reaches 60% or above; other nodes are collapsed. Scale denotes 0.1 mutations per site.

Molecular Phylogeny of the Insulin Gene Family

An amphioxus gene for insulin-like peptide (ILP) gene was cloned by Chan, Cao, and Steiner (1990); its phylogenetic relationships were analyzed by Ellsworth, Hewett-Emmett, and Li (1994). We have refined this analysis by incorporating the recently published sequences of insulin gene family members from the elasmobranch *Squalus acanthias* (Duguay et al. 1995) and the ascidian *Chelyosoma productum* (McRory and Sherwood 1997), since these give further insight into the timing of gene duplication in the insulin gene family. Unlike the case with the AAAH genes, we find that the metazoan insulin-related genes do not all fall into the three groups characteristic of mammals (fig. 4). For example, the hagfish IGF gene is found to be a possible outgroup to the IGF-1 and IGF-2 groups of genes. Furthermore, our phylogenetic analysis conclusively demonstrates that neither amphioxus ILP nor ascidian IGF genes are direct orthologues of either IGF-1 or IGF-2 genes; they are most probably direct descendants of an ancestral IGF. This implies that the gene duplication that gave rise to IGF-1 and IGF-2 occurred after the origin of vertebrates. The phylogenetic position of ascidian insulin is more confusing; our analyses suggest it could be a descendant of the precursor to insulin and both IGF genes. Further insight into the relationships between these genes comes from examination of their gene structures, including presence of particular protein domains. The deduced protein products of amphioxus ILP and ascidian IGF each have a C-terminal extension (domains D and E) characteristic of vertebrate IGF proteins, not present in insulin. This argues that amphioxus ILP and ascidian IGF are true IGF genes, not insulin genes. Both
lines of evidence point to one conclusion: IGF-1 and IGF-2 arose within the vertebrates.

Discussion

We chose to analyze AAAH genes and insulin-related genes as test cases to investigate the hypothesis that paralogy regions are reflections of tetraploidy in early vertebrate evolution. When a set of linked genes on one chromosome has linked relatives on another, it seems reasonable to assume that each paralogous pair arose simultaneously, by chromosome duplication. However, our characterization of an amphioxus AAAH gene and our molecular phylogenetic analyses of AAAH genes and insulin-related genes have yielded an unexpected result. We find that the AAAH genes on both human chromosomes 11 and 12 arose very early in metazoan evolution. The gene duplication events yielding TH, PAH, and TPH predate the divergence of nematodes, arthropods, and chordates, well back into the Precambrian era. In contrast, the neighboring IGF genes on the same chromosomes arose much later, during chordate radiation (postdating the hagfish/gnathostome divergence but predating the chondrichthyan/ray-finned fish/tetrapod divergences). The IGF duplication is dated to the time of the putative tetraploidy events in early vertebrate evolution, for which much supporting data have now accumulated (Holland et al. 1994; Holland 1996; Sharman and Holland 1996). This apparent paradox raises important questions concerning the nature and origin of chromosomal paralogy regions. The resolution that we propose below suggests that paralogy regions have a more complex history than previously recognized and also highlights the factors that affect the rates of gene loss in evolution.

The PAH gene is on chromosome 12 at 12q21.1, whereas TH and TPH map to 11p15.5. The IGF-1 gene maps adjacent to PAH, whereas IGF-2 is close to TH and TPH. The genes lie within a well-characterized and much-cited chromosomal paralogy group, with over 10 different gene families mapping to both chromosome regions (Lundin 1993; fig. 1). Considering the large number of gene families involved, it seems an inescapable conclusion that a substantial part of human chromosome 11 is evolutionarily related to part of chromosome 12, probably through chromosome duplication due to tetraploidy (Brissenden, Ullrich, and Francke 1984; Craig et al. 1986; Morton et al. 1986; Ledley et al. 1987; Lundin 1993). The discrepancies in the locations of some genes are likely to be due to pericentric inversion and/or translocations altering gene order on chromosome 12 compared with that on chromosome 11 (Brissenden, Ullrich, and Francke 1984; fig. 1). If we accept that human chromosomes 11 and 12 are evolutionarily related, how can the different duplication dates of AAAH and IGF genes be explained?

We propose that in the genome of an early chordate, a linked array of genes existed comprising TH, PAH, TPH, insulin, and IGF (fig. 5). Of these, at least TH, PAH, and TPH can be traced further back into early metazoan evolution, although when they became physically linked to insulin-related gene(s) cannot yet be deduced. During the early stages of vertebrate evolution, the chromosome containing this cluster of genes duplicated to yield two paralogous chromosomes, each containing linked copies of TH, PAH, TPH, insulin, and IGF. We propose that this occurred during tetraploidy of the whole genome, and we date this event to soon after the divergence of jawed vertebrates from the ancestral jawless vertebrates. This equates to the second of the two bouts of gene duplication during vertebrate evolution proposed by Holland et al. (1994) and Sharman and Holland (1996). After tetraploidy, the two IGF genes diverged in function, yielding IGF-1 and IGF-2. By contrast, the duplicate copies of TH, PAH, TPH, and insulin were all purged from the genome. Presumably, deletions, frameshifts, and nonsense mutations corroded duplicate copies to pseudogenes and ultimately deleted them. If loss of function occurred relatively soon after tetraploidy, as predicted by theoretical considerations (Marshall, Raff, and Raff 1994), it is not surprising that all traces of the duplicate genes have been removed from the genome during the 450 million years since tetraploidization. An interesting corollary of our proposed model is that some of the duplicate genes must have been deleted from each of the two daughter chromosomes. Thus, the genomic region that is now human chromosome 11 has lost a duplicate copy of the PAH gene; the loss has been more extensive on chromosome 12, which has been purged of duplicate TH, TPH, and insulin genes (fig. 5).

These conclusions have several implications for genome evolution. Most important, this example demonstrates that the net result of tetraploidy is not always gene duplication. Instead, as in the case of the AAAH
genes, the net result can be transfer of a gene from one chromosome to another. Hence, arrays of paralogous genes on two chromosomes are likely to include some genes related by duplication during tetraploidy and others related by gene transfer during tetraploidy. This implies that attempts to date chromosome or genome duplications using dates of gene duplications must proceed with caution. Certainly, it is not valid to use congruence of gene duplication dates within paralogous chromosomes as a test of the tetraploidy hypothesis. Likewise, caution must be exercised in extrapolating gene duplication dates from one gene family to its neighboring gene family (Bailey et al. 1997). More generally, our findings highlight differential gene loss as another factor contributing to the complex evolution of multigene families, alongside gene duplication, gene conversion, rate differences (Williams and Holland 1998), and convergent or parallel evolution (Stewart, Schilling, and Wilson 1987).

Finally, we note that there is a wide variance in the degree of gene loss following tetraploidy. All duplicated AAAH genes have been lost, in stark contrast to a large number of developmentally expressed genes in which duplicate genes have been retained (e.g., Hox gene clusters, Msx homeobox genes, myogenic genes, and IGF; Holland 1996). We speculate that the difference in extent of gene loss is related to the ease with which genes may acquire novel, selectively advantageous roles after duplication. Ubiquitous or broadly expressed genes, particularly those with enzymatic functions, may require rare and multiple substitutions in the protein-coding sequence to alter function. These may be so infrequent that genes are likely to be lost before novel functions arise. In contrast, genes with very localized sites of expression, particularly developmentally expressed genes, may be functionally adapted more simply, by loss or gain of enhancer sequences, creating novel sites of expression. We suggest that these genes are more likely to acquire new roles and to be retained after duplication.

Acknowledgments

P.W.H.H. thanks a critical seminar audience at Cambridge University, U.K., for help in rationalizing apparently contradictory results; Richard Sandford for advice at the earliest stages of this project; Seb Shimeld and Hiroshi Wada for critical comment; and Lars-G. Lundin for sparking an interest in paralogy. This work was funded by the Medical Research Council, U.K., and the University of Reading.

LITERATURE CITED


CLAUDIA KAPPEN, reviewing editor

Accepted July 23, 1998