MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 4, No. 2, June, pp. 139–149, 1995

A Phylogenetic Analysis of Vertebrate and Invertebrate *Notch*-Related Genes

ELEANOR M. MAINE,[1] JAMES L. LISSEMORE,[2] AND WILLIAM T. STARMER

Department of Biology, Syracuse University, 108 College Place, Syracuse, New York 13244

Received May 23, 1994; revised September 13, 1994

Members of the *Notch* gene family are thought to mediate inductive cell–cell interactions during development of a wide variety of vertebrates and invertebrates. These genes encode transmembrane proteins that appear to act as receptors and contain three repeated sequence motifs. Two of these motifs (an epidermal growth factor-like sequence and a *cdc10/SWI6*/ankyrin sequence) have been found in a large number of unrelated proteins, while the third motif (a *lin-12/Notch/glp-1* sequence) is unique to proteins of the *Notch* family. We present a phylogenetic analysis of 17 *Notch*-related genes from eight species that has implications as to the origins and relative functions of these genes in different species. Several independent gene duplications have occurred and at least one such duplication in the vertebrate lineage preceded the avian/mammalian divergence. Significantly, the overall organization of individual members of each internally repeated motif appears to have been conserved among species, suggesting that each repeat plays a unique role in protein function. Yet, where sequence divergence does occur among genes in vertebrate, dipteran, and nematode lineages, it may signify functional differences for specific regions in *Notch*-related proteins. © 1995 Academic Press, Inc.

INTRODUCTION

Notch-related genes comprise a growing family of vertebrate and invertebrate genes. *Notch* was originally identified by genetic analysis as a gene important for determining cell fate during development in *Drosophila melanogaster* (see reviews by Artavanis-Tsakonas *et al.*, 1991; Artavanis-Tsakonas and Simpson, 1991). Independently, the *lin-12* and *glp-1* genes were identified genetically as being essential for certain inductive interactions during development of *Caenorhabditis elegans* (Greenwald *et al.*, 1983; Austin and Kimble, 1987; Priess *et al.*, 1987; Schnabel, 1994); as shown by subsequent molecular work, the nematode LIN-12 and GLP-1 proteins are structurally very related to each other and to NOTCH (Yochem *et al.*, 1988; Yochem and Greenwald, 1988). More recently, two vertebrate oncogenes, mouse *int-3* and human *TAN-1*, have been shown to encode *Notch*-related proteins (Jhappan *et al.*, 1992; Robbins *et al.*, 1992; Ellisen *et al.*, 1991). Additional *Notch*-related genes have now been isolated from rat (Weinmaster *et al.*, 1991, 1992), mouse (Franco del Amo *et al.*, 1992, 1993; Reaume *et al.*, 1992; Kopan and Weintraub, 1993; Lardelli and Lendahl, 1993, 1994), *Xenopus laevis* (Coffman *et al.*, 1990), humans (Stifani *et al.*, 1992), chicken (D. Henrique and D. Ish-Horowicz, personal communication), and *Lucilia cuprina* (sheep blowfly) (P. Batterham, personal communication) on the basis of sequence similarity to *Drosophila Notch*. *Notch*, *glp-1*, and *lin-12* appear to encode membrane-associated receptor proteins that mediate a variety of cell–cell interactions during development (Greenwald *et al.*, 1983; Austin and Kimble, 1987; Priess *et al.*, 1987; Seydoux and Greenwald, 1989; Heitzler and Simpson, 1993; Crittenden *et al.*, 1994; Evans *et al.*, 1994; Mello *et al.*, 1994). Vertebrate genes likewise have vital roles during development: gene truncations of *TAN-1*, *int-3*, and *Xotch* are associated with hyperplasia (Ellisen *et al.*, 1991; Jhappan *et al.*, 1992; Robbins *et al.*, 1992; Coffman *et al.*, 1993) and null mutations cause embryonic lethality (Swiatak *et al.*, 1994).

Notch-related genes encode transmembrane proteins with three repeated sequence motifs as shown in Fig. 1 (Wharton *et al.*, 1985; Yochem *et al.*, 1988; Yochem and Greenwald, 1989; Coffman *et al.*, 1990; Weinmaster *et al.*, 1991, 1992; Franco del Amo *et al.*, 1992, 1993; Reaume *et al.*, 1992; Kopan and Weintraub, 1993; Lardelli and Lendahl, 1993). The extracellular domains contain 10–36 copies of an ~40-amino-acid epidermal growth factor-like (EGFL) sequence motif and 3 copies of an ~40-amino-acid *lin/Notch/glp* (LNG) sequence motif. The intracellular domains contain 6–7 copies of a *cdc10/SWI6*/ankyrin (CDC/ANK) sequence motif known in other systems to mediate protein–protein in-

[1] To whom correspondence should be addressed. Fax: (315) 443-2156. E-mail: "emmaine@mailbox.syr.edu".
[2] Present address: Department of Biology, John Carroll University, University Heights, OH 44118.

1055-7903/95 $6.00

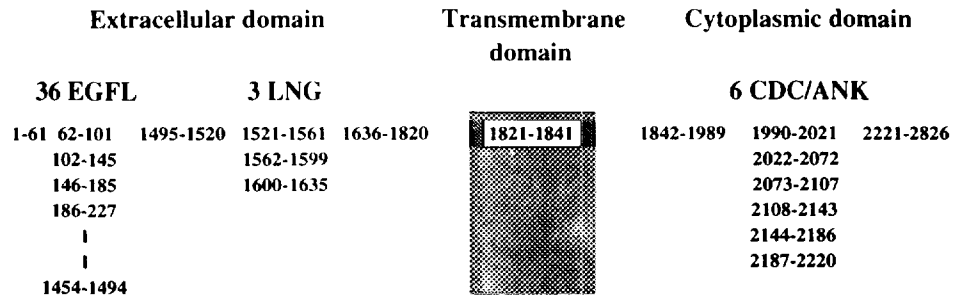

FIG. 1. Structure of a "generic" *Notch*-related gene in vertebrates and insects. Numbers refer to amino acid residues in the aligned set of genes. The extracellular domain contains 36 copies of the cysteine-rich EGFL motif closely followed by 3 copies of the cysteine-rich LNG motif; this region is followed by a stretch of nonrepeated sequence. A predicted transmembrane domain of 20 amino acids is present. The cytoplasmic domain contains 6 copies of the CDC/ANK motif flanked by stretches of nonrepetitive sequence. The *C. elegans* genes differ in three major ways from the gene diagrammed here: *lin-12* and *glp-1* contain only 13 or 10 EGFL repeats, respectively; each gene has a seventh, degenerate CDC/ANK repeat (Lissemore *et al.*, 1993); the carboxy-terminal nonrepeated sequence is much shorter (~150 amino acids). It is not yet known whether partially sequenced *Notch*-related genes used in this study (see below) have the same numbers of repeated motifs as do fully sequenced genes.

teractions (Davis and Bennett, 1990; Davis *et al.*, 1991; Thompson *et al.*, 1991; Inoue *et al.*, 1992; Wulczyn *et al.*, 1992; Bours *et al.*, 1993; Hatada *et al.*, 1993; reviewed by Blank *et al.*, 1992).

Here we present a phylogenetic analysis of 17 *Notch*-related gene sequences from eight species as well as an analysis of the sequence similarity among individual repeats of each type (EGFL, LNG, and CDC/ANK). The presence of 3 different repeated sequences (EGFL, LNG, and CDC/ANK) as well as the large number of EGFL repeats characteristic of most *Notch*-related proteins provides a unique opportunity for studying the origin of internally repeated motifs. Our results have implications as to the origins and relative functions of these genes in different species and of individual repeats within each gene.

MATERIALS AND METHODS

Accession numbers and/or sources for the sequences used in this study are as follows: *Notch,* M11664 (Wharton *et al.*, 1985); *lin-12,* M21478 (Yochem *et al.*, 1988); *glp-1,* M25580 (Yochem and Greenwald, 1989); *Xotch,* M33874 (Coffman *et al.*, 1990); *Notch1,* X57405 (Weinmaster *et al.*, 1991); *Notch2,* M93661 (Weinmaster *et al.*, 1992); *TAN-1,* M73980 (Ellisen *et al.*, 1991); *hN,* M99437 (Stifani *et al.*, 1992 and C. Blaumueller and S. Artavanis-Tsakonas, personal communication); *Notch-1,* Z11886 (Franco del Amo *et al.*, 1993); *Motch,* L02610, L02611, L02612, L02613, and L02614 (Reaume *et al.*, 1992); *mNotch,* Z21925 (Kopan and Weintraub, 1993); *MotchA,* X68278 (Lardelli and Lendahl, 1993); *MotchB,* X68279 (Lardelli and Lendahl, 1993); *int-3,* M80456 (Robbins *et al.*, 1992); *Scalloped wing,* P. Batterham, personal communication; *C-Notch-1, C-Notch-2,* D. Henrique and D. Ish-Horowicz, personal communication.

Phylogenetic analyses were performed by first aligning the amino acid sequences with the multiple alignment program Clustal V (Higgins *et al.*, 1992) and then constructing consensus trees using programs from the tree-building package Phylip 3.5c (Felsenstein, 1989). Rates of sequence divergence were analyzed by constructing distance matrices and fitting them to two models that make different assumptions about evolutionary rates; the FITCH model assumes constant rates of divergence whereas the KITCH model makes no such assumption (Felsenstein, 1989). See Results for details.

RESULTS

Phylogeny of Notch-Related Proteins

Phylogenetic relationships were examined among 17 *Notch*-related genes whose sequences are available to us. The complete amino acid sequences are available for 9 genes [*Drosophila Notch* (Wharton *et al.*, 1985), mouse *Notch-1,* human *TAN-1* (Ellisen *et al.*, 1991) and *hN* (Stifani *et al.*, 1992; C. Blaumueller and S. Artavanis-Tsakonas, personal communication), rat *Notch1* and *Notch2* (Weinmaster *et al.*, 1991, 1992), *Xenopus Xotch* (Coffman *et al.*, 1990), and *C. elegans lin-12* (Yochem *et al.*, 1988) and *glp-1* (Yochem and Greenwald, 1989)], while only partial sequences are available for 8 others (listed below). Some partial sequences do not overlap; therefore, we constructed two trees using sets of genes for which we have overlapping sequence, as presented in Fig. 2. Genes for which complete sequence is available were used for both analyses, as was mouse *Motch,* for which sequences from noncontiguous regions spanning the entire gene are available. Incomplete gene sequences included only in the analysis presented in Fig. 2 (left) are as follows: for mouse *MotchA* and *MotchB,* portions of the extracellular domain including some EGFL repeats and the

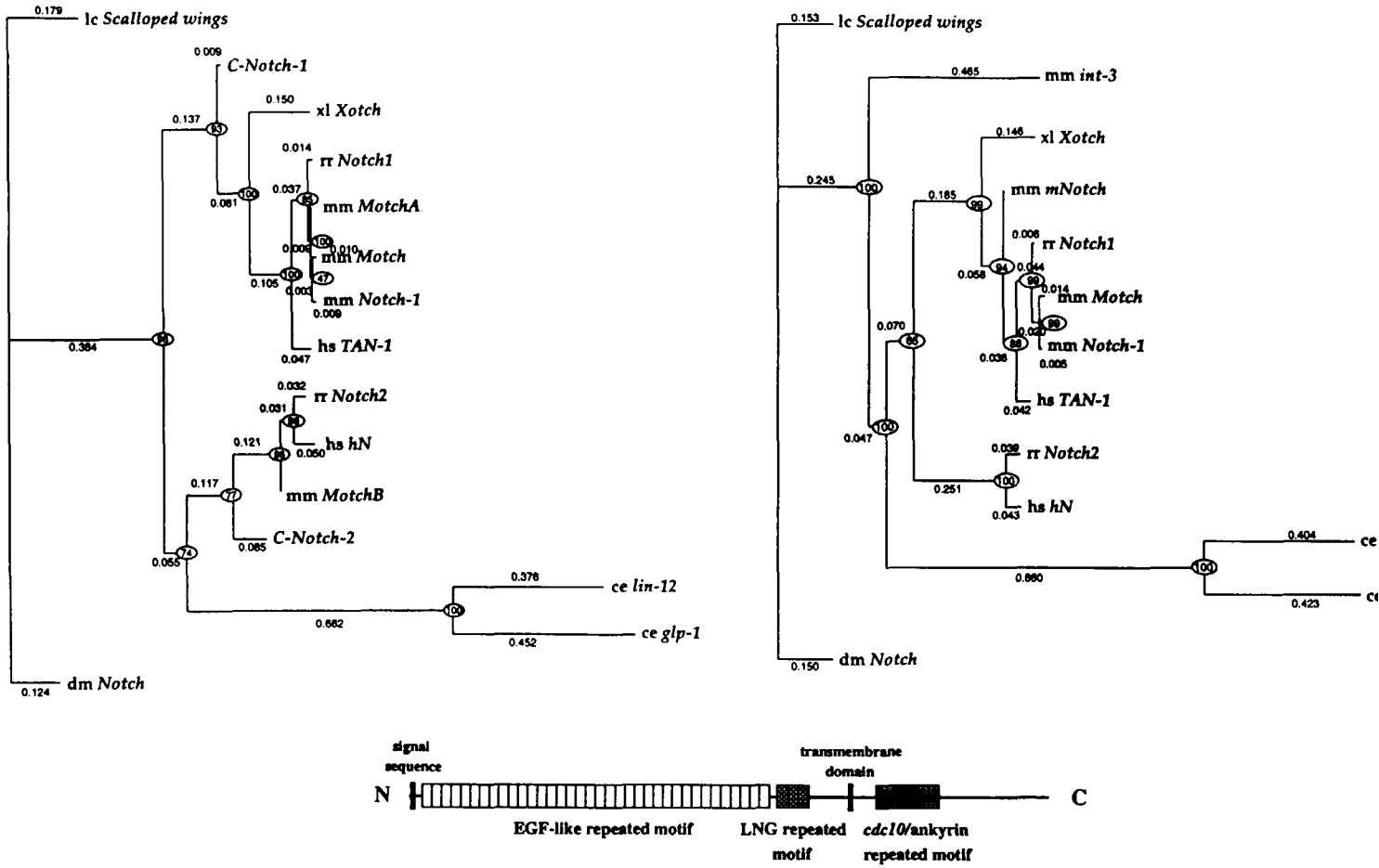

FIG. 2. Phylogenetic relationships among *Notch*-related genes. (Left) Phylogeny of 15 genes based on extracellular sequences. (Right) Phylogeny of 13 genes based on intracellular sequences. Two separate trees were constructed to allow use of partial sequence data. (Trees are not rooted.) Horizontal branch lengths reflect the number of amino acid substitutions; a value of 0.1 indicates 1 amino acid substitution per 10 amino acids. The circled number indicates the bootstrap value out of 100 resamples. For clarity, the organism is indicated before each gene name (except for chicken, where *C-* is included in the gene name): mm, mouse; rr, rat; hs, human; lc, *Lucilia cuprina* (sheep blowfly); dm, *Drosophila melanogaster*; ce, *Caenorhabditis elegans;* xl, *Xenopus laevis.* (Below) Schematic representation of predicted *Drosophila Notch* protein (Wharton *et al.*, 1985) with the three sets of repeated sequence motifs indicated as boxed regions; vertebrate genes have essentially the same structure. *C. elegans* genes have fewer EGFL repeats and a seventh, degenerate CDC/ANK repeat.

entire LNG region (Lardelli and Lendahl, 1993); for chicken *C-Notch-1* and *C-Notch-2*, the LNG regions as well as some EGFL repeats from *C-Notch-1* (D. Henrique and D. Ish-Horowitz, personal communication). Incomplete gene sequences included only in the analysis presented in Fig. 2 (right) are as follows: for mouse *int-3*, the intracellular and transmembrane domains (Robbins *et al.*, 1992); for mouse *mNotch*, the CDC/ANK repeats and a unique portion of the cytoplasmic domain (Kopan and Weintraub, 1993); and for blowfly *Scalloped wing*, the entire gene except the EGFL repeats and the amino terminus (P. Batterham, personal communication).

Gene phylogenies were examined by first aligning the amino acid sequences of each of 17 gene products with the multiple alignment program Clustal V (Higgins *et al.*, 1992; alignment of published sequences available in EMBL under Accession No. DS19782). Consensus gene trees were constructed by (1) bootstrapping (100 times) using the BOOT program, (2) estimating distances between sequences for each resample using PROTDIST (Kimura correction), (3) forming each tree by the neighborhood-joining method of Saitou and Nei (1987), and (4) forming the majority rule consensus tree by the method CONSENSE. All four programs are from the tree-building package Phylip 3.5c (Felsenstein, 1989). Unsequenced or deleted regions were excluded only in pairwise comparisons.

Our phylogenetic analyses of both sets of genes gave compatible results, placing the 17 genes into five groups: (1) *D. melanogaster Notch* and *L. cuprina* (sheep blowfly) *Scalloped wing;* (2) mouse *MotchA, Motch, Notch-1,* and *mNotch,* human *TAN-1,* rat *Notch1,* chicken *C-Notch-1,* and *X. laevis Xotch*; (3) mouse *MotchB,* human *hN,* rat *Notch2,* and chicken *C-Notch-2*; (4) *C. elegans lin-12* and *glp-1*; and (5) mouse *int-3*. Groups 2 and 3 must have been produced by a gene duplication prior to the divergence of avian and mammalian lineages 300 Myr B.P. We will refer to these two groups of genes as Type 1 (mouse *Motch, MotchA, Notch-1,* and *mNotch,* rat *Notch1,* human *TAN-1,* and chicken *C-Notch-1*) and Type 2 (mouse*MotchB,* rat *Notch2,* human *hN,* and chicken *C-Notch-2*) vertebrate genes. The *int-3* gene presumably arose by a separate duplication event within the vertebrate lineage. The *C. elegans* genes *glp-1* and *lin-12* appear to have been produced by an independent duplication event.

The four mouse Type 1 sequences (*Notch-1, mNotch, Motch,* and *MotchA*) may all correspond to a single gene, with minor sequence differences reflecting polymorphisms; some of these genes may have been recovered from different wild-type strains. With the exception of *Notch-1*, available sequence for mouse Type 1 genes is incomplete, as mentioned above; in particular, very small portions of *mNotch* and *MotchA* are available. In general, the placement of genes for which we have partial sequence may be less reliable than those for which we have complete sequence because comparisons are made over a relatively small portion of these genes.

Sequence Relationships among Internally Repeated Motifs

Repeated amino acid motifs presumably arise by repeated internal duplications, starting from one original sequence of each type. We examined the phylogenetic relationships among individual EGFL, LNG, and CDC/ANK repeats to determine whether unique components of different repeats have been conserved. Using those *Notch*-related genes for which a complete set of repeats has been sequenced, we compared each repeat within individual genes and among subsets of genes. Individual repeats of each type were defined using the published amino acid consensus sequences for each gene.

EGFL repeats. The amino acid sequences of the 36 EGFL repeats of human, mouse, and *Xenopus* Type 1 genes (*TAN-1, Notch-1,* and *Xotch,* respectively) were analyzed by several methods in order to gain insight into their origin and differences in their evolutionary rates. The rat Type 1 gene, *Notch1,* was not included in each of these analyses because it is very closely related to mouse *Notch-1*. A subset of these analyses was also performed to compare the 36 EGFL repeats of (1) human and rat Type 2 genes (*hN* and *Notch2*) and (2) Types 1 and 2 human and rat genes (*TAN-1, hN, Notch1,* and *Notch2*). In each case, our first approach was to construct a parsimonious phylogenetic tree for all of the sequences ($36 \times 2 = 72$, $36 \times 3 = 108$, or $36 \times 4 = 144$). The sequences were aligned and subjected to bootstrapping (100 resamples) and a majority rule consensus parsimonious tree was constructed using the PROTPARS program (Felsenstein, 1989).

The results from these analyses showed that the integrity of nearly all EGFL repeat units has been conserved among Type 1 vertebrate genes and between Type 2 vertebrate genes. For example, all No. 8 EGFL repeats are more similar to one another than they are to any other repeat 1–36. Therefore, the 36 EGFL repeats originated before the vertebrate lineage separated because divergence in the vertebrate sequences occurred after the divergence of each repeated unit. We have found two exceptions to this general rule: (1) EGFL-16 (EGFL repeat 16) in *Xenopus* is more similar to EGFL-31 (*Xenopus* and mammalian) than to mammalian EGFL-16, and (2) EGFL-23 has diverged in the two Type 2 genes analyzed. In most instances the mouse and human repeats (from Type 1 genes) are more similar to each other than either is to the corresponding *Xenopus* repeat. However, in four cases (EGFL-27, -28, -30, and -31) we find different patterns; for EGFL-27 and -28, the *Xenopus* and mouse se-

quences are most similar, and for EGFL-30 and -31 the *Xenopus* and human sequences are most similar.

Sequence divergence among the 36 repeats is so extensive that it is not possible to determine higher order relationships among them (i.e., the pattern of duplications that produced the final 36 repeats). For example, the cluster of EGFL-8 repeats is joined to the cluster of EGFL-30 repeats with a bootstrap value of 31%; the average bootstrap value between sets of repeats is 16.1%, with a range of 2.0 to 53.6%.

The integrity of each EGFL unit has been maintained within all vertebrate genes, regardless of whether they are Type 1 or Type 2. Comparison of rat and human Type 1 and Type 2 genes with each other reveals that all but seven EGFL repeats have maintained their similarity between the two sets of genes. Six repeats (EGFL-1, -2, -3, -4, -35, and -36) have diverged significantly between the two groups while remaining conserved within each group (data not shown). One additional repeat, EGFL-23, has also diverged between Type 1 and 2 genes as well as between Type 2 genes (as mentioned above). Therefore, the 36 repeats were established before duplication of the ancestral gene to produce Types 1 and 2. Sufficient sequence divergence has occurred between EGFL repeats in *Drosophila Notch* and both the vertebrate genes and the *C. elegans* genes to prevent reliable detection of relatedness (data not shown).

To investigate further the evolution of the individual EGFL repeats, we examined the relative rates of change of different EGFL repeats in three Type 1 genes, human *TAN-1*, mouse *Notch-1*, and *Xenopus Xotch*. For each of the 36 repeats, the branches connecting the three species (as shown in Fig. 5) were estimated by the maximum likelihood method. Branch lengths were then summed to estimate the total amount of evolution (i.e., $d_a + d_b + d_c$) for a given repeat since the three vertebrates diverged. Next, the variance and standard error for divergence distance in each repeat was estimated from the corresponding branch length variances. This calculation allows for a statistical comparison among repeats. Although these comparisons are not all independent, they did reveal repeats which either have changed considerably or have remained relatively unchanged in each of the three genes (Fig. 5).

There is no general trend among the three genes in terms of which repeats are divergent or well conserved (Fig. 5). In *Xotch*, four repeats at the ends of the region, EGFL-1, -2, -35, and -36, as well as two internal repeats (EGFL-5 and -16), are most divergent. In general, EGFL repeats in *Notch-1* and *TAN-1* are less divergent than in *Xotch*, and in particular the terminal repeats are fairly well (*TAN-1*) or especially well (*Notch-1*) conserved. In *Notch-1*, three internal repeats (EGFL-5, -13, and -29) are most divergent; in *TAN-1*, four different internal repeats (EGFL-2, -16, -22, and -27) are most divergent. Note that three repeats are especially divergent in more than one gene: EGFL-5 in *Xotch* and *TAN-1*, and EGFL-2 and -16 in *Xotch* and *Notch-1*. Each gene has several well-conserved repeats: 4 in *Xotch* (EGFL-8, -27, -30, and -31), 11 in *TAN-1* (EGFL-3, -8, -10, -11, -13, -14, -15, -18, -32, -33, and -34), and 9 in *Notch-1* (EGFL-4, -12, -14, -16, -17, -20, -27, -35, and -36). Again, note that three repeats are especially well conserved in more than one gene: EGFL-8 in *Xotch* and *TAN-1*, EGFL-14 in *TAN-1* and *Notch-1*, and EGFL-27 in *TAN-1* and *Notch-1*.

A second analysis was performed to compare the rates of change within different lineages for any given repeat. To do so, the distance matrices of the three vertebrate sequences and the *Drosophila Notch* sequence (as an outgroup) were fitted to two models, one assuming constant evolutionary rates (KITCH) and one without assumption about evolutionary rates (FITCH). These two procedures resulted in residual sums of squares that were compared via an F-statistic for an approximate test of the validity of the rate constancy assumption (Felsenstein, 1989). This procedure was performed for each repeat and revealed that two repeats (EGFL-19 and EGFL-23) have significantly violated the constancy assumption. In other words, EGFL-19 and EGFL-23 have changed at different rates in the three lineages (*Xenopus*, mouse, and human). This result could be used as an indicator of sequences which may have taken on different functions.

LNG and CDC/ANK repeats. The amino acid sequences of the LNG and CDC/ANK repeats were analyzed by several means as described above for EGFL repeats. According to these analyses, the identity of individual LNG and CDC/ANK repeats as well as their overall organization has been maintained over a longer evolutionary period than that of the EGFL repeats. Each distinct CDC/ANK repeat has been conserved in the eight vertebrate and insect genes for which we have data (Fig. 3), while divergence has occurred in the nematode lineage (data not shown). Each distinct LNG repeat has been conserved in all 13 genes for which we have data regardless of species (Fig. 4).

Analysis of the relative divergence of LNG and CDC/ANK repeats in the Type 1 genes *TAN-1*, *Notch-1*, and *Xotch* reveals striking differences among both sets of repeats (Fig. 5). CDC/ANK-1 and -3 are essentially invariant among all three genes. In contrast, CDC/ANK-2, -4, -5, and -6 are quite variable, being well conserved in some genes and relatively variable in others. In particular, CDC/ANK-6 has diverged extensively in *Notch-1* relative to *TAN-1* and *Xotch* (Fig. 5). Among the LNG repeats there is a similar variability. LNG-3 is best conserved in all three genes. In contrast, LNG-1 and -2 are relatively divergent in *Xotch* yet well conserved in *Notch-1* and *TAN-1* (Fig. 5).

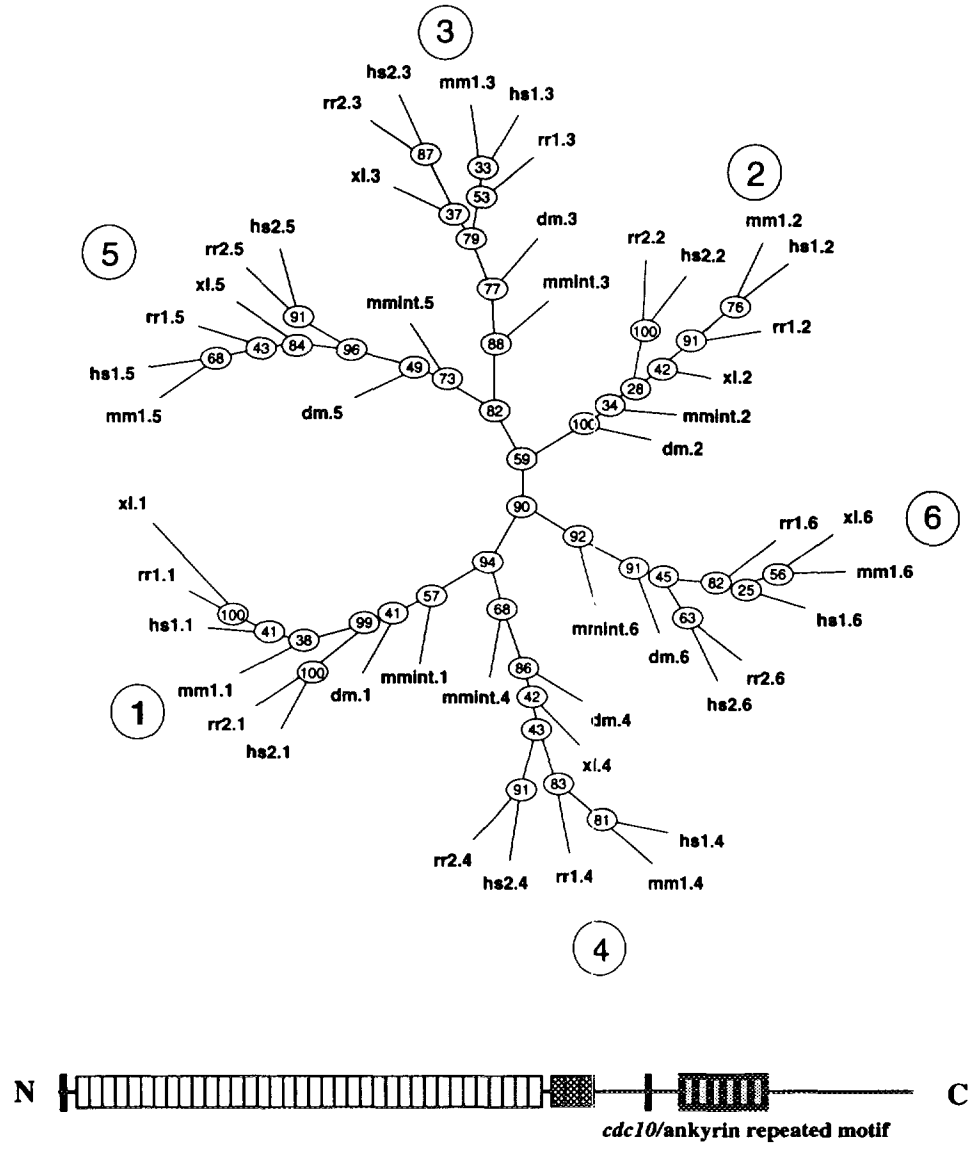

FIG. 3. Phylogenetic relationships among CDC/ANK repeats. Genes included in the analysis are rat (*Notch1*, rr1), mouse (*Notch-1*, mm1), human (*TAN-1*, hs1), and *Xenopus* (*Xotch*, xl) Type 1 genes, rat (*Notch2*, rr2) and human (*hN*, hs2) Type 2 genes, *Drosophila Notch* (dm), and mouse *int-3* (mmint). Repeat number is designated; abbreviated gene names are used to save space. The circled number at each branchpoint indicates bootstrap value out of 100 resamples. Note that each distinct repeat (1–6) is maintained across taxa. (Below) Representation of *Notch* protein with the CDC/ANK repeats highlighted.

DISCUSSION

We have presented a study of the sequence relationships among 17 genes of the *Notch* family and among individual repeats of each three internally repeated sequence motifs (EGFL, LNG, and CDC/ANK) within those genes. By exploring the evolution of the gene family and of internally repeated sequence motifs, we aimed to provide information useful for the study of gene function. In particular, we hoped to identify specific internal regions that might be involved in gene-specific function(s) or general function(s) common to all or a subset of *Notch*-like proteins. The results of our analysis of gene phylogeny suggest the following: (1) a gene duplication occurred in the vertebrate lineage at least as far back as the avian–mammalian divergence 300 Myr B.P.; (2) the *int-3* gene arose by a separate duplication within the vertebrate lineage; (3) the pair of nematode genes, *lin-12* and *glp-1*, were produced by an independent duplication event.

In our analysis of repeated sequences, we saw remarkable conservation of distinct repeats across taxa. The LNG repeats are best conserved, with each of three repeats found in nematodes, diptera, and vertebrates.

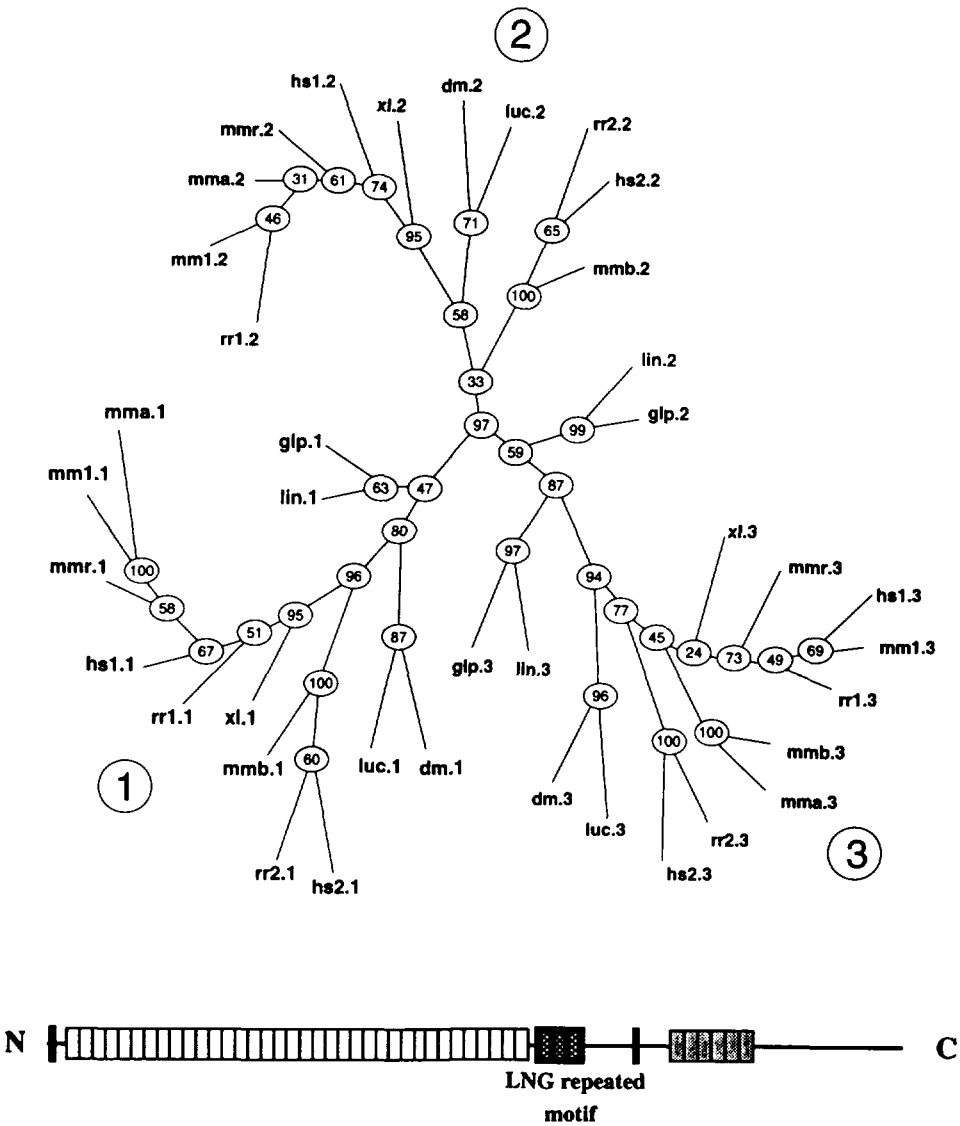

FIG. 4. Phylogenetic relationships among LNG repeats. Genes included in the analysis are six Type 1 genes [rat *Notch1* (rr1), mouse *Notch-1* (mm1), *MotchA* (mma), *Motch* (mmr), human *TAN-1* (hs1), and *Xenopus Xotch* (xl)], three Type 2 genes [rat *Notch2* (rr2), mouse *MotchB* (mmb), and human *hN* (hs2)], *Drosophila Notch* (dm), *Lucilia Scalloped wing* (luc), and *C. elegans glp-1* (glp), and *lin-12* (lin). Repeat number is designated; abbreviated gene names are used to save space. The circled number at each branchpoint indicates bootstrap value out of 100 resamples. Note that each distinct repeat (1–3) is maintained across taxa. (Below) Representation of *Notch* protein with the LNG repeats highlighted.

CDC/ANK repeats are somewhat less well conserved, with each of six repeats conserved in diptera and vertebrates; CDC/ANK repeats found in nematode genes are distinct from each other, but have diverged too much to correlate with specific repeats in other taxa. The organization of 36 distinct EGFL repeats has been maintained in diptera and vertebrates, but sufficient sequence divergence has occurred to prohibit direct correlation of each dipteran repeat with a specific vertebrate repeat. Furthermore, the vertebrate EGFL repeats appear to have been organized before any duplication of *Notch*-related genes occurred. Once each repeat type (EGFL, LNG, or CDC/ANK) was organized, no subsequent gene conversion event occurred in the lineage leading to any species. In general, we interpret this high conservation of repeats to mean that each repeat plays a unique role in gene function.

The tree (Fig. 3) representing the possible evolutionary relationship of the six CDC/ANK repeated units found in diptera and vertebrates can be used to postulate the sequence of events that produced the present linear structure. One possibility is that a series of duplications/deletions originating from a single unit was responsible. Although there are several hypothetical

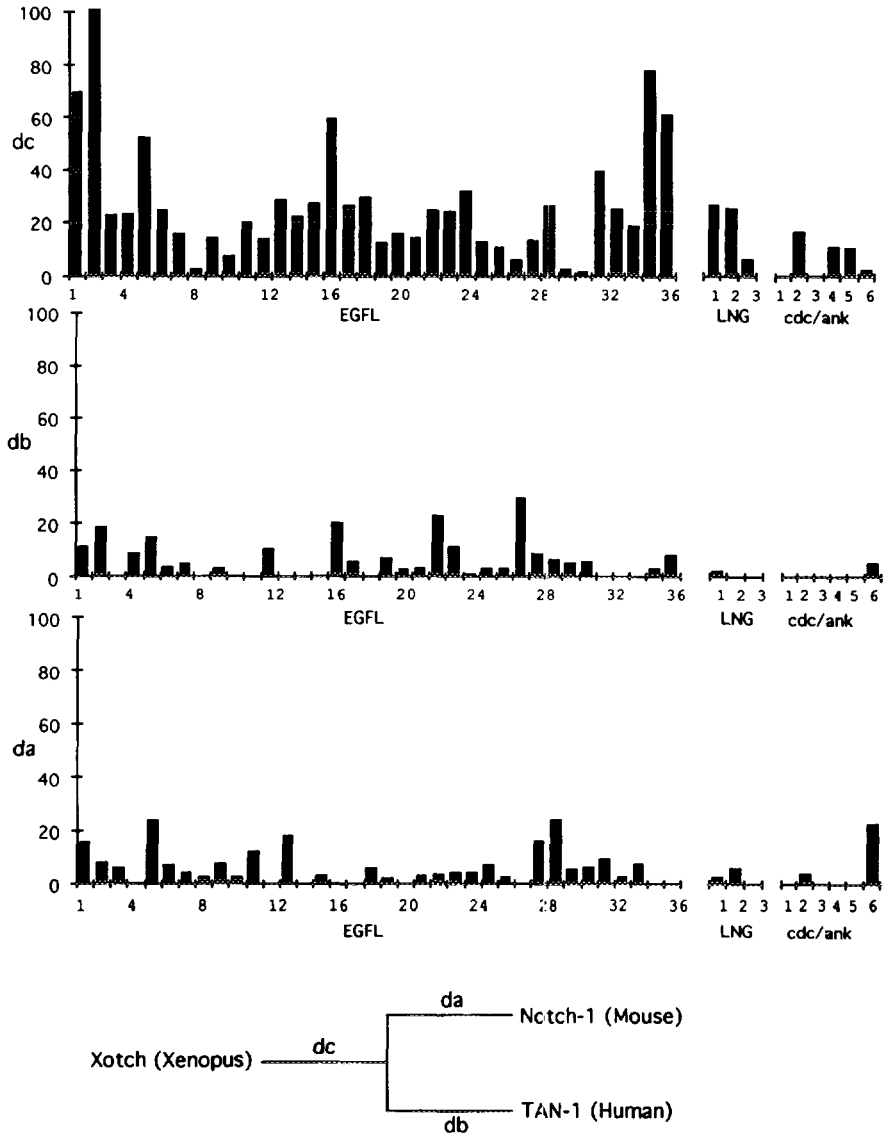

FIG. 5. Divergence among EGFL, LNG, and CDC/ANK repeats in the Type 1 vertebrate genes *Notch-1* (mouse), *TAN-1* (human), and *Xotch* (*Xenopus*). Each histograph represents one gene; each bar represents one EGFL, LNG, or CDC/ANK repeat. Divergence, d, is calculated as the number of amino acid substitutions per 100 amino acids of sequence. Each distance, d_a, d_b, and d_c, is indicated on the tree. Branch lengths were estimated by the maximum likelihood method (see text).

pathways that could produce the tree (Fig. 3), one is as follows. (1) A series of three single unit tandem duplications produces a four-unit structure; (2) the first three units then duplicate in tandem, producing a set of seven repeat units; (3) the fifth unit is deleted. This hypothetical sequence of events is depicted in Fig. 6, where the growing tree of relationships is on the left and the linear sequence of the units is on the right. The extent of sequence divergence among the large number of EGFL repeats precludes any proposal about higher order relationships among them.

In carrying out a phylogenetic analysis, we are interested in identifying amino acid sequences that might be involved in either highly conserved or gene-specific functions. Several of our analyses address this question. First, seven EGFL repeats (EGFL-1, -2, -3, -4, -23, -35, and -36) vary between Type 1 and 2 vertebrate genes. Second, EGFL-23 also varies between different Type 2 genes. Third, EGFL-19 and -23 have evolved at significantly different rates in different vertebrate lineages. Fourth, a number of EGFL, LNG, and CDC/ANK repeats are particularly well conserved or divergent among all or a subset of Type 1 vertebrate genes. Assuming that all *Notch*-related proteins function as receptors in signal transduction pathways, we expect many protein functions to be conserved among them. However, it is not unexpected that differences might exist as well between genes in different species (e.g.,

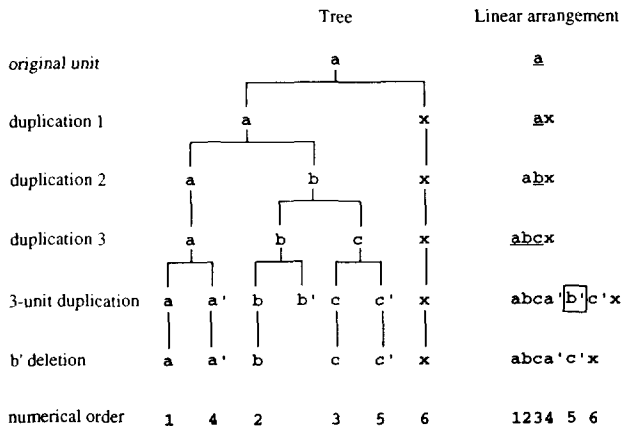

FIG. 6. Hypothetical sequence of duplications and a deletion to produce the series of six CDC/ANK repeats found in dipteran and vertebrate genes. (Left) The growing set of CDC/ANK units. (Right) The linear sequence of units; underlined units are subsequently duplicated, and the boxed unit is subsequently deleted.

Drosophila Notch and rat *Notch1*) and, perhaps, between duplicated genes within any given species (e.g., Type 1 and Type 2 vertebrate genes). Recent studies have shown that disruption of mouse Type 1 gene expression (*Notch-1* and *Motch*) causes embryonic lethality (Swiatek *et al.*, 1994), meaning that the Type 1 and Type 2 genes in mouse cannot be completely redundant for function. The variable repeats identified in our analyses may be responsible, at least in part, for functional differences between Type 1 and Type 2 genes. Furthermore, sequence variation in EGFL-23 may reflect a difference in function for that repeat between Type 2 genes in different species as well. Clearly, it is of interest to know how variable EGFL-23 is among other Type 2 genes (e.g., *C-Notch-2*).

Molecular analysis of mutations can also be helpful in elucidating the function of specific amino acid residues or regions in a protein. There is a wealth of information about the locations of mutations that produce a variety of mutant phenotypes in *Drosophila Notch* (including Hartley *et al.*, 1987; Kelley *et al.*, 1987; Lieber *et al.*, 1992; Xu *et al.*, 1990, 1992; De Celis *et al.*, 1993; Rebay *et al.*, 1993), *glp-1* (Mango *et al.*, 1991; Kodoyianni *et al.*, 1992; Roehl and Kimble, 1993), and *lin-12* (Greenwald and Seydoux, 1990; Struhl *et al.*, 1993). However, given the divergence among genes found in diptera, nematodes, and vertebrates, it may be difficult to correlate function of individual EGFL repeats in *Notch, glp-1,* and *lin-12* with those in the vertebrate genes. For example, several dominant, gain-of-function *Notch* mutations map to the highly divergent EGFL-23 (Hartley *et al.*, 1987; Kelley *et al.*, 1987), but it is not clear that similar mutations in the vertebrate genes would likewise produce gain-of-function phenotypes. In contrast, given the conservation of LNG repeats among dipteran, nematode, and vertebrate genes and of CDC/ANK repeats among dipteran and vertebrate genes, the functions of individual LNG and CDC/ANK repeats may be maintained within these genes. If so, then the effects on protein function of mutations in specific LNG repeats in vertebrates may be similar to their effects on NOTCH, GLP-1, or LIN-12 function; similarly, the effect of mutations in specific CDC/ANK repeats in vertebrates may be similar to their effects on NOTCH function. Unfortunately, few molecularly characterized *Notch* mutations cause specific alterations of a CDC/ANK or LNG repeat, and relatively few mutations in *glp-1* or *lin-12* alter a specific LNG repeat. However, it has been shown that an in-frame deletion of almost the entire LNG region in *glp-1* results in a null phenotype (Kodoyianni *et al.*, 1992); presumably the loss of this region in vertebrate *Notch*-related genes would eliminate gene function as well.

Our understanding of phylogenetic relationships among *Notch*-related genes is still incomplete. It remains to be seen whether *Xenopus* contains a Type 2 gene and whether vertebrates besides mouse contain an *int-3*-like gene. A fourth mouse gene, *mNotch3,* recently has been isolated (Lardelli and Lendahl, 1994) and a human *int-3*-like gene has been identified (D. Gallahan, personal communication). Analysis of additional vertebrate and invertebrate genes will broaden our understanding of the origins and functional relationships among *Notch*-related genes. In particular, given the large distance between insects and nematodes, analysis of *Notch*-related genes from other classes of invertebrates may be especially useful.

ACKNOWLEDGMENTS

We thank Phil Batterham, Christine Blaumueller, Spyros Artavanis-Tsakonas, Daniel Gallahan, Robert Callahan, Domingos Henrique, David Ish-Horowicz, and Ronald Conlon for providing DNA sequences and communicating other results prior to publication. We also thank Raphael Kopan for suggesting a comparison of EGF repeats. Shozo Yokoyama and David Sullivan provided valuable discussion of the data and comments on the manuscript. This work was supported by National Science Foundation grants to E.M.M. and W.T.S.

REFERENCES

Artavanis-Tsakonas, S., and Simpson, P. (1991). Choosing a cell fate: A view from the *Notch* locus. *Trends Genet.* **7:** 403–408.

Artavanis-Tsakonas, S., Delidaks, C., and Fehon, R. G. (1991). The *Notch* locus and the cell biology of neuroblast segregation. *Annu. Rev. Cell Biol.* **7:** 427–452.

Austin, J., and Kimble, J. (1987). *glp-1* is required in the germ line for regulation of the decision between mitosis and meiosis in *C. elegans. Cell* **51:** 589–599.

Blank, V., Kourilsky, P., and Israel, A. (1992). NF-κB and related proteins: Rel/dorsal homologies meet ankyrin-like repeats. *Trends Biochem. Sci.* **17:** 135–140.

Bours, V., Franzoso, G., Azarenko, V., Park, S., Kanno, T., Brown, K., and Siebenlist, U. (1993). The oncoprotein Bcl-3 di-

rectly transactivates through κB motifs via association with DNA-binding p50B homodimers. *Cell* **72:** 729–739.

Coffman, C., Harris, W., and Kintner, C. (1990). *Xotch*, the *Xenopus* homolog of *Drosophila Notch*. *Science* **249:** 1438–1441.

Coffman, C., Skoglund, P., Harris, W., and Kintner, C. (1993). Expression of an extracellular deletion of *Xotch* diverts cell fate in Xenopus embryos. *Cell* **73:** 659–671.

Crittenden, S. L., Troemel, E. R., Evans, T. C., and Kimble, J. (1994). GLP-1 is localized to the mitotic region of the *C. elegans* germ line. *Development* **120:** 2901–2911.

Davis, L. H., and Bennett, V. (1990). Mapping the binding sites of human erythrocyte ankyrin for the anion exchanger and spectrin. *J. Biol. Chem.* **265:** 10589–10596.

Davis, L. H., Otto, E., and Bennett, V. (1991) Specific 33-residue repeat(s) of erythrocyte ankyrin associate with the anion exchanger. *J. Biol. Chem.* **266:** 11163–11169.

De Celis, J. F., Barrio, R., del Arco, A., and Garcia-Bellido, A. (1993). Genetic and molecular characterization of a Notch mutation in its delta- and serrate-binding domain in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **90:** 4037–4041.

Ellisen, L. W., Bird, J., West, D. C., Soreng, A. L., Reynolds, T. C., Smith, S. D., and Sklar, J. (1991). *TAN-1*, the human homolog of the Drosophila *Notch* gene, is broken by chromosomal translocations in T lymphoblastic neoplasms. *Cell* **66:** 649–661.

Evans, T. C., Crittenden, S. L., Kodoyianni, V., and Kimble, J. (1994). Translational control of maternal *glp-1* mRNA establishes an asymmetry in the *C. elegans* embryo. *Cell* **77:** 183–194.

Felsenstein, J. (1989). PHYLIP—Phylogeny inference package (version 3.2). *Cladistic* **5:** 164–166.

Franco del Amo, F., Smith, D. E., Swiatek, P. J., Gendron-Maguire, M., Greenspan, R. J., McMahon, A. P., and Gridley, T. (1992). Expression pattern of *Motch*, a mouse homolog of *Drosophila Notch*, suggests an important role in early postimplantation mouse development. *Development* **115:** 737–744.

Franco del Amo, F., Gendron-Maguire, M., Swiatek, P. J., Jenkins, N. A., Copeland, N. G., and Gridley, T. (1993). Cloning, analysis, and chromosomal localization of *Notch-1*, a mouse homolog of *Drosophila Notch*. *Genomics* **15:** 259–264.

Greenwald, I., and Seydoux, G. (1990). Analysis of gain-of-function mutations of the *lin-12* gene of *Caenorhabditis elegans*. *Nature* **346:** 197–199.

Greenwald, I., Sternberg, P., and Horvitz, H. R. (1983). The *lin-12* locus specifies cell fates in Caenorhabditis elegans. *Cell* **34:** 435–444.

Hartley, D. A., Xu, T., and Artavanis-Tsakonas, S. (1987). The embryonic expression of the *Notch* locus of *Drosophila melanogaster* and the implications of point mutations in the extracellular EGF-like domain of the predicted protein. *EMBO J.* **6:** 3407–3417.

Hatada, E. N., Naumann, M., and Scheidereit, C. (1993). Common structural constituents confer IκB activity to NF-κB p105 and IκB/MAD-3. *EMBO J.* **12:** 2781–2788.

Heitzler, P., and Simpson, P. (1993). Altered epidermal growth factor-like sequences provide evidence for a role of *Notch* as a receptor in cell fate decisions. *Development* **117:** 1113–1121.

Higgins, D. G., Bleasby, A. J., and Fuchs, R. (1992). *CLUSTAL V:* Improved software for multiple sequence alignment. *Comput. Appl. Biosci.* **8:** 189–191.

Inoue, J., Kerr, L. D., Rashid, D., Davis, N., Bose, H. R., Jr., and Verma, I. (1992). Direct association of pp40/IκBβ with rel/NF-κB transcription factors: Role of ankyrin repeats in the inhibition of DNA binding activity. *Proc. Natl. Acad. Sci. USA* **89:** 4333–4337.

Jhappan, C., Gallahan, D., Stahle, C., Chu, E., Smith, G. H., Merlino, G., and Callahan, R. (1992). Expression of an activated *Notch*-related *int-3* transgene interferes with cell differentiation and induces neoplastic transformation in mammary and salivary glands. *Genes Dev.* **6:** 345–355.

Kelley, M. R., Kidd, S., Deutsch, W. A., and Young, M. W. (1987). Mutations altering the structure of epidermal growth factor-like coding sequences at the Drosophila *Notch* locus. *Cell* **51:** 539–548.

Kodoyianni, V., Maine, E. M., and Kimble, J. (1992). The molecular basis of loss-of-function mutations in the *glp-1* gene of *Caenorhabditis elegans*. *Mol. Biol. Cell* **3:** 1199–1213.

Kopan, R., and Weintraub, H. (1993). Mouse Notch: Expression in hair follicles correlates with cell fate determination. *J. Cell Biol.* **121:** 631–641.

Lardelli, M. and Lendahl, U. (1993). *MotchA* and *MotchB*—Two mouse *Notch* homologues coexpressed in a wide variety of tissues. *Exp. Cell Res.* **204:** 364–372.

Lardelli, M., and Lendahl, U. (1994). The novel *Notch* homologue mouse *Notch 3* lacks specific epidermal growth factor-repeats and is expressed in proliferating neuroepithelium. *Mech. Dev.* **46:** 123–136.

Lieber, T., Wesley, C. S., Alcamo, E., Hassel, B., Krane, J. F., Campos-Ortega, J. A., and Young, M. W. (1992). Single amino acid substitutions in EGF-like elements of Notch and Delta modify Drosophila development and affect cell adhesion in vitro. *Neuron* **9:** 847–859.

Lissemore, J. L., Currie, P. D., Turk, C. M., and Maine, E. M. (1993). Intragenic dominant suppressors of *glp-1*, a gene essential for cell-signaling in *Caenorhabditis elegans*, support a role for *cdc10/SW16*/ankyrin motifs in GLP-1 function. *Genetics* **135:** 1023–1034.

Mango, S. E., Maine, E. M., and Kimble, J. (1991). A carboxy-terminal truncation activates the *glp-1* protein to specify vulval fates in *Caenorhabditis elegans*. *Nature* **352:** 811–815.

Mello, C. C., Draper, B. W., and Priess, J. R. (1994). The maternal genes *apx-1* and *glp-1* and establishment of dorsal–ventral polarity in the early C. elegans embryo. *Cell* **77:** 95–106.

Priess, J. R., Schnabel, H., and Schnabel, R. (1987). The *glp-1* locus and cellular interactions in early *C. elegans* embryos. *Cell* **51:** 601–611.

Reaume, A. G., Conlon, R. A., Zirngibl, R., Yamaguchi, T. P., and Rossant, J. (1992). Expression analysis of a *Notch* homologue in the mouse embryo. *Dev. Biol.* **154:** 377–387.

Rebay, I., Fehon, R. G., and Artavanis-Tsakonis, S. (1993). Specific truncations of *Drosophila* Notch define dominant activated and dominant negative forms of the receptor. *Cell* **74:** 319–329.

Robbins, J., Blondel, B. J., Gallahan, D., and Callahan, R. (1992). Mouse mammary tumor gene *int-3*: A member of the *notch* gene family transforms mammary epithelial cells. *J. Virol.* **66:** 2594–2599.

Roehl, H., and Kimble, J. (1993). Control of cell fate in *C. elegans* by a GLP-1 peptide consisting primarily of ankyrin repeats. *Nature* **364:** 632–635.

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

Schnabel, R. (1994). Autonomy and non-autonomy in cell fate specification of muscle in the *Caenorhabditis elegans* embryo: A reciprocal induction. *Science* **263:** 1449–1452.

Seydoux, G., and Greenwald, I. (1989). Cell autonomy of *lin-12* function in a cell fate decision in C. elegans. *Cell* **57:** 1237–1245.

Stifani, S., Blaumueller, C. M., Redhead, N. J., Hill, R. E., and Artavanis-Tsakonas, S. (1992). Human homologs of a *Drosophila Enhancer of Split* gene product define a novel family of nuclear proteins. *Nature Genet.* **2:** 119–127.

Struhl, G., Fitzgerald, K., and Greenwald, I. (1993). Intrinsic activity of the Lin-12 and Notch intracellular domains in vivo. *Cell* **74:** 331–345.

Swiatak, P. J., Lindsell, C. E., Franco del Amo, F., Weinmaster, G., and Gridley, T. (1994). *Notch1* is essential for postimplantation development in mice. *Genes Dev.* **8:** 707–719.

Thompson, C. C., Brown, T. A., and McKnight, S. L. (1991). Convergence of Ets- and Notch-related structural motifs in a heteromeric DNA binding complex. *Science* **253:** 762–768.

Weinmaster, G., Roberts, V. J., and Lemke, G. (1991). A homolog of *Drosophila Notch* expressed during mammalian development. *Development* **113:** 199–205.

Weinmaster, G., Roberts, V. J., and Lemke, G. (1992). *Notch2*: A second mammalian *Notch* gene. *Development* **116:** 931–941.

Wharton, K. A., Johanson, K. M., Xu, T., and Artavanis-Tsakonas, S. (1985). Nucleotide sequence from the neurogenic locus *Notch* implies a gene product that shares homology with proteins containing EGF-like repeats. *Cell* **43:** 567–581.

Wulczyn, F. G., Naumann, M., and Scheidereit, C. (1992). Candidate proto-oncogene *bcl-3* encodes a subunit-specific inhibitor of transcription factor NF-κB. *Nature* **358:** 597–599.

Xu, T., Rebay, I., Fleming, R. J., Scottgale, T. N., and Artavanis-Tsakonas, S. (1990). The *Notch* locus and the genetic circuitry involved in early *Drosophila* neurogenesis. *Genes Dev.* **4:** 464–475.

Xu, T., Caron, L. A., Fehon, R. G., and Artavanis-Tsakonas, S. (1992). The involvement of the *Notch* locus in *Drosophila* oogenesis. *Development* **115:** 913–922.

Yochem, J., Weston, K., and Greenwald, I. (1988). The *Caenorhabditis elegans lin-12* gene encodes a transmembrane protein with overall similarity to *Drosophila Notch*. *Nature* **335:** 547–550.

Yochem, J., and Greenwald, I. (1989). *glp-1* and *lin-12*, genes implicated in distinct cell–cell interactions in C. elegans, encode similar transmembrane proteins. *Cell* **58:** 553–563.