BMC Evolutionary Biology

BioMed Central

Research article



# The evolution of Runx genes I. A comparative study of sequences from phylogenetically diverse model organisms

Jessica Rennert[1,2], James A Coffman[1], Arcady R Mushegian[1] and Anthony J Robertson*[1]

Address: [1]Stowers Institute for Medical Research, 1000 E. 50th Street, Kansas City, MO 64110, USA and [2]Computational Biosciences Program, Arizona State University, Tempe, AZ 85287, USA

Email: Jessica Rennert - jessrennert@msn.com; James A Coffman - jac@stowers-institute.org; Arcady R Mushegian - arm@stowers-institute.org; Anthony J Robertson* - ajr@stowers-institute.org

* Corresponding author

Published: 24 March 2003

*BMC Evolutionary Biology* 2003, **3**:4

Received: 16 October 2002
Accepted: 24 March 2003

This article is available from: http://www.biomedcentral.com/1471-2148/3/4



## Abstract

**Background:** Runx genes encode proteins defined by the highly conserved Runt DNA-binding domain. Studies of Runx genes and proteins in model organisms indicate that they are key transcriptional regulators of animal development. However, little is known about Runx gene evolution.

**Results:** A phylogenetically broad sampling of publicly available Runx gene sequences was collected. In addition to the published sequences from mouse, sea urchin, *Drosophila melanogaster* and *Caenorhabditis elegans*, we collected several previously uncharacterised Runx sequences from public genome sequence databases. Among deuterostomes, mouse and pufferfish each contain three Runx genes, while the tunicate *Ciona intestinalis* and the sea urchin *Strongylocentrotus purpuratus* were each found to have only one Runx gene. Among protostomes, *C. elegans* has a single Runx gene, while *Anopheles gambiae* has three and *D. melanogaster* has four, including two genes that have not been previously described. Comparative sequence analysis reveals two highly conserved introns, one within and one just downstream of the Runt domain. All vertebrate Runx genes utilize two alternative promoters.

**Conclusions:** In the current public sequence database, the Runt domain is found only in bilaterians, suggesting that it may be a metazoan invention. Bilaterians appear to ancestrally contain a single Runx gene, suggesting that the multiple Runx genes in vertebrates and insects arose by independent duplication events within those respective lineages. At least two introns were present in the primordial bilaterian Runx gene. Alternative promoter usage arose prior to the duplication events that gave rise to three Runx genes in vertebrates.

## Background

Runx genes encode the sequence-specific DNA binding subunit of a heterodimeric transcription factor, the defining feature of which is the Runt domain, a highly conserved 128 amino acid sequence involved in DNA binding, heterodimerization, nucleotide binding, and nuclear localization [1,2]. The Runt domain is named after the first member of the family to be discovered, the regulatory gene *runt* from *Drosophila melanogaster*. Runx genes have also been discovered and functionally characterized in mammals, sea urchins and nematodes, and in general are involved in the transcriptional control of

developmental processes [3,4]. In humans, mutations in each of the three Runx genes are associated with disease caused by defective control of cell proliferation and/or differentiation [4,5]. Most studies of Runx gene function and regulation have been carried out in mammals and in *D. melanogaster*, each of which has multiple Runx genes. It is not currently known how the Runx gene family evolved, nor is it known how many Runx genes the first animal possessed. Answering these questions will facilitate identification of primitive (general) and derived (specialized) aspects of Runx gene structure, function and regulation.

## Results and discussion

### *The collection of Runx gene sequences from phylogenetically diverse model organisms*

Sequence similarity searches of public databases using the BLAST program [6] were used to collect a set of currently available Runx gene sequences (listed in Supplemental Table 1) from phylogenetically diverse species with complete or nearly complete genome sequences. Previously undescribed Runx genes collected in our BLAST searches include two genes from the puffer fish (*Takifugu rubripes*) genome, a single gene from the sea squirt (*Cionia intestinalis*) genome, three genes from the mosquito (*Anopheles gambiae*) genome, and two new Runx genes from the fruit fly (*Drosophila melanogaster*)genome. Also collected in the set for comparison were the previously-described Runx genes from *D. melanogaster* (*runt* and *lozenge* [7]), *Caenorhabditis elegans* (*run* [8]), sea urchin (*SpRunt* [9,10]), *T. rubripes* (*TrRunx2* [11]), and mouse (*Runx1*, *Runx2*, and *Runx3*; the three human RUNX genes, which have been extensively characterized previously, are highly similar to their mouse orthologues [12,13] and hence were not included in the collected set). The collected set of sequences thus contains representatives of several major bilaterian phyla within both deuterostomes (Chordata and Echinodermata) and protostomes (Nematoda and Arthropoda). Runx sequences from lophotrochozoans and from basal (non-bilaterian) metazoans (e.g., Cnidaria and Porifera) are not found in the database, and it is not yet known whether non-bilaterian Runx genes exist. Extensive BLAST and PSI-BLAST searches of the public databases failed to recover any sequences with significant similarity to the Runt domain from non-metazoan organisms with completed or nearly completed genomes, suggesting that Runx genes may be a metazoan invention. The closest amino acid sequence similarity of any part of the Runt domain to sequences outside of metazoa is found in the motif GRSGRGKSF [2] (see Fig. 2), which matches the nucleotide binding P-loop sequence in various bacterial and archaeal ATP-binding proteins with diverse functions. This Runx sequence is functionally involved in minor groove interactions and DNA bending by the Runt domain [14]. An important difference between it and *bona fide* P-loops is that the latter are invariably followed by an α-helix, whereas the structure of the Runt domain is all beta (see below). Thus, although the Runx P-loop-like motif is capable of binding ATP [2], it is unclear whether this local sequence similarity is indicative of homology to the prokaryotic P-loop.

### *Runx gene copy number*

Two of the invertebrates in our collection with completely sequenced genomes, the protostome *C. elegans* and the deuterostome *C. intestinalis*, each were found to contain a single Runx gene. Screens of genomic libraries using the Runt domain sequence as probe indicate that the same is also apparently true of the sea urchin *S. purpuratus* [10], although final verification of this awaits the completion of the sea urchin genome. In contrast, the model organisms in which most functional studies of Runx genes have been performed (i.e. *D. melanogaster* and mouse) each contain multiple Runx genes. Both of the vertebrate species surveyed (mouse and *T. rubripes*) were each found to contain three Runx genes, as was the mosquito *A. gambiae*, whereas *D. melanogaster* contains four.

The simplest explanation of such a gene distribution is that possession of a single Runx gene is the primitive condition for bilaterians, i.e., that the multiplicity of Runx genes in the vertebrate and arthropod lineages resulted from independent gene duplications within those respective lineages. The results of a phylogenetic analysis of Runt domain sequences are consistent with this hypothesis (see below). Since fruit flies and mice each contain multiple Runx genes, it is probable that each of those genes has acquired specialized (e.g., region- or tissue-specific) functions that were derived subsequent to the duplications, and hence peculiar to the taxonomic group to which each of those model organisms belong. Experimental studies of invertebrates that contain only a single Runx gene (sea urchin, tunicate, and nematode) are therefore likely to highlight primitive (general) functions of this family of transcription factors in the control of cell fate, proliferation, and differentiation during metazoan development.

### *Exon and intron positions among Runx genes*

The gross structural features of the collected Runx genes, including positions of exons and introns (excepting those upstream of the proximal promoter of the vertebrate Runx genes – see below), are depicted schematically in Figure 1 (see also Supplemental Table 2 and 3). To locate exon-intron positions in previously uncharacterised genes (or predicted genes) that have not yet been associated with cDNAs, we performed spliced alignment, maintaining maximum similarity between homologues, as described in Methods. As can be seen in Figure 1, all of the Runx genes contain a central, highly conserved exon, which encodes the C-terminal third of the Runt domain (black box in Fig. 1). The N-terminal end of this exon is bounded by

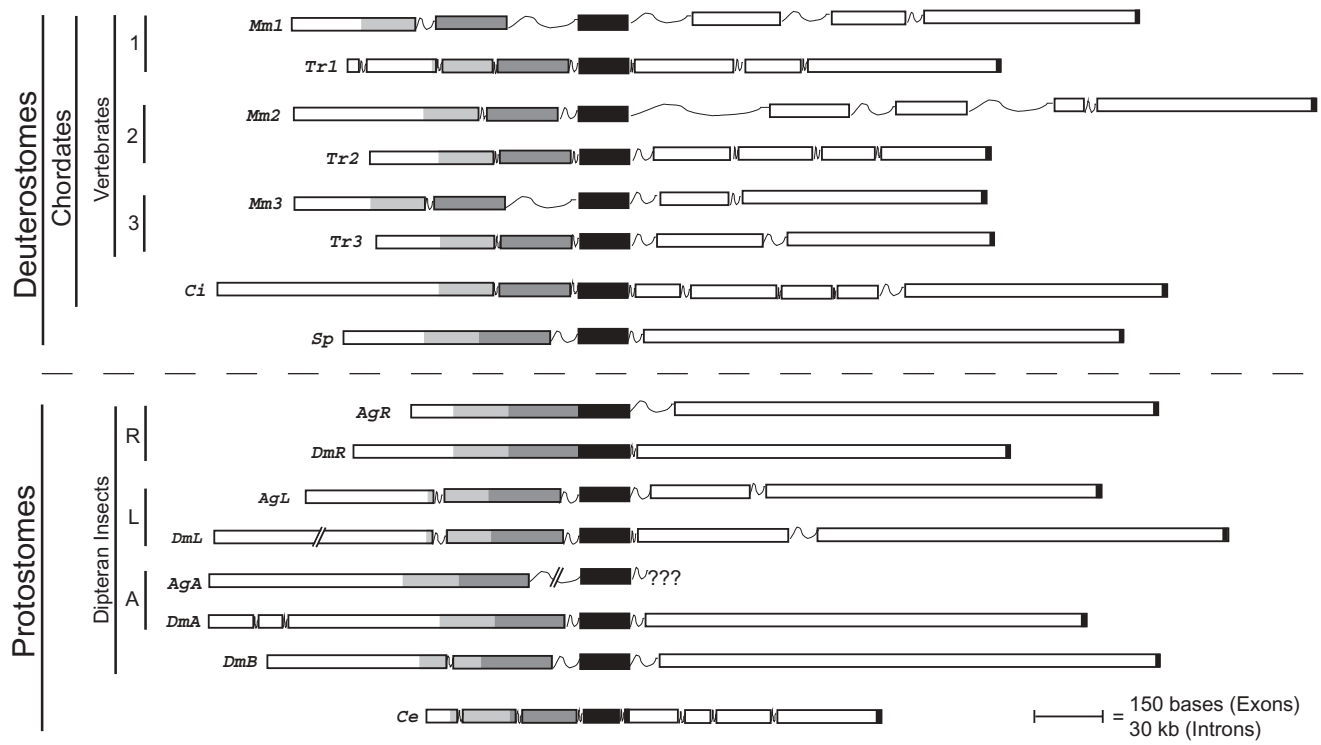


**Figure 1**

**Comparative structure of Runx genes collected from genome sequences of phylogenetically diverse model organisms.** Diagram of Runx genes from selected species. Boxes indicate exons and curved lines indicate introns, which are scaled approximately to the scale bar as indicated. Absolute exon and intron sizes are listed in Supplemental Table 2 and 3. The Runt domain is represented as filled boxes ranging from light grey (N-terminus) to black (C-terminus) in order to facilitate comparison to the chordate Runt domain, which is encoded by three different exons. The black bar at the end of all the genes represents the VWRPY (or IWRPF in the case of *CeRun*) Groucho recruitment motif that is at the C-terminus of proteins encoded by all Runx genes, which facilitated mapping of the previously uncharacterised genes. Exon sequences representing 5' and 3' untranslated regions are not depicted in the diagram. For the vertebrate Runx genes, only exons downstream of the proximal promoters (P2) are shown. Human *RUNX1* contains 3 additional alternatively spliced exons between those encoding the Runt domain and the C-terminal VWRPY motif [12], which may or may not be present in the mouse orthologue (*Mm1*) and are not shown here. Because of sequence gaps in the genomic sequence of the *Anopheles gambiae RunxA* gene (*AgA*), the exon structure 3' to the Runt domain of this gene is not known, as indicated by question marks. Abbreviations: *Mm1, M. musculus Runx1*, etc.; *Tr1, T. rubripes Runx1*, etc.; *Ci, C. intestinalis; Sp, S. purpuratus; AgA, A. gambiae RunxA; DmA, D. melanogaster RunxA; DmB, D. melanogaster RunxB; AgL, A. gambiae Lozenge; DmL, D. melanogaster Lozenge; AgR, A. gambiae Runt; DmR, D. melanogaster Runt; Ce, C. elegans*.

an intron whose position is absolutely conserved among all Runx genes except *DmRunt* and *AgRunt* (which have apparently lost the intron), while the C-terminal end is bounded by an intron that is conserved among all deuterostome species in our collection, and shifted but a few nucleotides downstream in the insect genes and upstream in the *C. elegans* gene. The existence and locations of all of the other introns are variable, but some of them are characteristic of specific clades.

All of the chordate Runx genes (including *CiRunt*) contain an intron within the N-terminal half of the Runt domain, the position of which is invariant with respect to nucleotide sequence. This intron is missing in the other Runx genes, including the sea urchin representative (*SpRunt*), and is therefore likely to be a chordate-specific feature of Runx genes that arose prior to the Runx gene duplications in the vertebrate lineage. Upstream of the Runt domain, *TrRunx1* is predicted to have two short introns that are not found in its mouse orthologue (*MmRunx1*) or in any of the other chordate Runx genes. The mammalian Runx

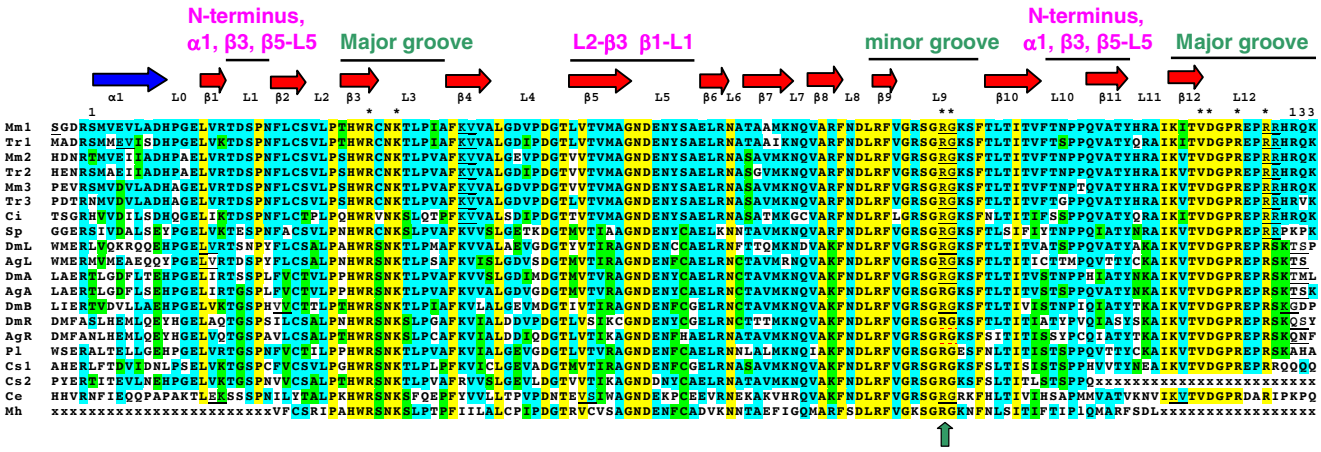


**Figure 2**

**Multiple sequence alignment of the Runt domain from collected sequences.** Alignment of Runx amino acid sequences created using Clustal W [16]. The number 1 denotes the conventional beginning of the Runt domain, which is 128 amino acids long. The shading highlights areas of amino acid conservation, with yellow indicating absolute conservation among all members, and blue or green indicating high conservation of a residue among several members. The domain structure (alpha helix and beta sheet elements) determined for *MmRunx1* [14,17], is shown as arrows above the alignment, and the surfaces involved in DNA contact (major and minor groove, shown in green) or interaction with specific structural motifs within the beta subunit (shown in pink) are denoted by lines above the structural motifs [14]. In cases where it is known, amino acid pairs split by introns in the primary transcripts are underlined and in boldface (i.e., all but the two spider sequences, the crayfish, and the nematode *Meloidogyne hapla*, for which only cDNAs are available). The arrow indicates a highly conserved intron that falls within sequence that encodes the purine nucleotide-binding consensus. Note that the Runt domains from the spider gene *Cs2* and the nematode gene *Mh* are partial, as indicated by the x's; these partial sequences were not used in the alignment used to generate trees (Table 1 and Figure 3). Abbreviations are as in Figure 1, with the addition of *Pl, Pacifastacus leniusculus; Cs1, Cupiennius salei Run-1, etc.; Mh, Meloidogyne hapla*.

genes have longer introns than their counterparts in *Takifugu*, as might be expected based on relative genome sizes. With the exception of *DmRunt*, all of the *Drosophila* Runx genes contain one (*DmLozenge* and *DmRunxB*) or two (*DmRunxA*) introns separating exons N-terminal to the Runt domain, but all of these have different positions, and were thus likely to have been incorporated subsequent to the divergence of these genes. The position of the first intron in *DmLozenge* is conserved in one of the mosquito genes (*AgLozenge*), indicating that these two genes are orthologues, which is confirmed by a phylogenetic analysis of sequences (see below). *CeRun* contains an intron within the N-terminal half of the Runt domain that is not found in any other Runx gene, and contains the largest number of exons of any of the genes in our collection. In general, the structure of the *C. elegans* gene is the most divergent of all of the Runx genes.

A variable number of exons are found 3' of the Runt domain among different species, and these reveal group-specific patterns. Except for the *lozenge* orthologues, which contain two such exons, all of the insect genes for which complete sequences are available contain only a single exon downstream of the Runt domain. The same is apparently true for the single sea urchin gene (although this is provisionally based on the fragmentary evidence of a single small BAC clone [10]), suggesting that a single exon downstream of the Runt domain may be the primitive condition for bilaterians. In vertebrates (from the proximal promoter), the *Runx3* orthologues each contain 2 exons downstream of the Runt domain, while the *Runx2* orthologues contain 4 and the *Runx1* orthologues contain 3. *CiRunt* contains 5 exons downstream of the Runt domain, while *CeRun* contains 4. In vertebrates at least, the multiplicity of downstream exons is reflected in a large variety of alternatively spliced transcripts that give rise to multiple protein isoforms that differ in the C-terminal sequences appended to the Runt domain [12,13]. It is important to note that the C-terminal exon of all Runx genes identified to date encodes the amino acid sequence VWRPY (or the functionally equivalent IWRPF in the case of *C. elegans*), which acts as a recruitment motif for the co-repressor Groucho/TLE [3]. This motif is at the C-terminus within the 3'-most exon of all of the genes, possibly with

the exception of *CiRunt* (where the open reading frame apparently continues beyond the VWRPY in the genome sequence), and was used in our analysis to identify the C-terminal exons of previously uncharacterised Runx genes.

### *Runt domain sequences across phylogeny*

The amino acid sequences of the Runt domains of all each of the collected genes, together with two previously described sequences from the spider *Cupiennius salei* [15], a sequence from the crayfish *Pacifastacus leniusculus*, and one from the nematode *Meloidogyne hapla*, were aligned using Clustal W [16], as shown in Figure 2. Functionally, the Runt domain is required both for DNA binding and for interaction with its heterodimeric partner (the beta subunit), which serves to allosterically enhance the DNA binding of the Runt domain [1,14]. Superposition of the alignment and the known crystal structure of mouse Runx1 [14,17] reveals that residues that make either direct or indirect contact with DNA are invariant (marked by asterisks in Fig. 2). Sequence motifs that interact with the beta subunit are also highly conserved.

The alignment of Runt domain amino acid sequences was used to construct unrooted trees by several complementary approaches (neighbour-joining, maximum-likelihood, and maximum-parsimony, described in Methods). We also employed Bayesian inference of phylogeny [18]. This analysis was performed using 137 aligned amino acids with no gaps (Fig. 2), and included all of the sequences shown in Figure 2 except for the incomplete sequences from a spider (*C. salei run-2*) and from the root-knot nematode (*M. hapla*). The consensus topology produced by each method was largely consistent with accepted phylogenies for all nodes with significant bootstrap values (Fig. 3); the main observations derived from each method are summarized in Table 1. The consensus of phylogenetic approaches strongly indicates the monophyly of the deuterostome Runx genes in the collection, with the expected relationships (echinoderms(urochordates(vertebrates))). The trees also suggest that the gene duplications that produced the three vertebrate Runx genes occurred early in the chordate lineage leading to vertebrates, but subsequent to its divergence from urochordates. All of the trees place *Runx1* basal to *Runx3* and *Runx2* (Fig. 3 and Table 1), arguing against an earlier proposal for the basal position of *Runx3* [8,13] among the vertebrate genes, which was inferred from relative gene size, exon-intron structure, and pattern of expression.

While the consensus of trees summarized in Table 1 is for the most part consistent with the expected monophyly of the protostome genes in our collection (nematodes, arthropods), the branching order is ambiguous. Among the insect representatives, each of the mosquito genes can be confidently assigned orthologues from *Drosophila* (Fig. 3), as was also suggested by intron positions (Fig. 1). Beyond that, low bootstrap values prohibit a confident assessment of evolutionary branching order of the arthropod Runx paralogues, and it is not clear whether they are monophyletic (although the position of the crayfish gene suggests a deep duplication event within in the insect/crustacean clade at least). Resolution of the evolutionary relationships among the protostome Runx genes will require a broader sampling of Runx genes from Arthropods as well as from protostomes in general.

The occurrence of three Runx genes in vertebrates, each on a different chromosome, is consistent with a scenario of multiple gene duplications, or two successive rounds of genome duplications that may have occurred near the base of the vertebrate branch of chordates [19]. In contrast, the multiple Runx genes in *Drosophila* are all physically linked to *runt* on the X chromosome, and were almost certainly generated by local gene duplication events (for the location of the *Drosophila* genes, see FlyBase Genome Browser: http://www.bdgp.org/cgi-bin/annot/gbrowse). The two newly discovered *Drosophila* Runx genes both contain a complete open reading frame encoding the Runt domain (Fig. 2) and conserved exon-intron structure (Fig. 1 and 3), suggesting that they are probably not pseudogenes. This is further supported by our finding that each of the three mosquito Runx genes has an orthologue in *Drosophila* (Figs. 1, 2, and 3).

### *Alternative promoter usage among vertebrate runx genes*

All of the vertebrate Runx genes are transcribed from two alternative promoters, P1 (distal) and P2 (proximal). We found no evidence (as would be indicated by cDNA clones or ESTs) in the database for alternative promoter usage among any of the invertebrate Runx genes examined, suggesting that alternative promoters may have evolved in the chordate lineage leading to vertebrates, subsequent to its divergence from urochordates. All known invertebrate Runx gene products begin with long 5'UTRs and N-terminal amino acid sequences that are more like those of the products from the proximal promoter (P2) of the vertebrate genes, suggesting that P2 is the primitive promoter. Definitive resolution of these issues awaits a more comprehensive database of genomes and expressed sequences in appropriate invertebrate systems. Figure 4A depicts the transcript structures generated by alternative promoter usage and splicing in the vertebrate Runx genes, and Figure 4B shows the N-terminal amino acid sequences derived from each promoter. It is clear from these data that the usage of alternative promoters evolved prior to the gene duplication events in the vertebrate lineage, since all three genes in both mouse and *Takifugu* utilize alternative promoters that give rise to similar alternative N-termini within the respective orthologues.

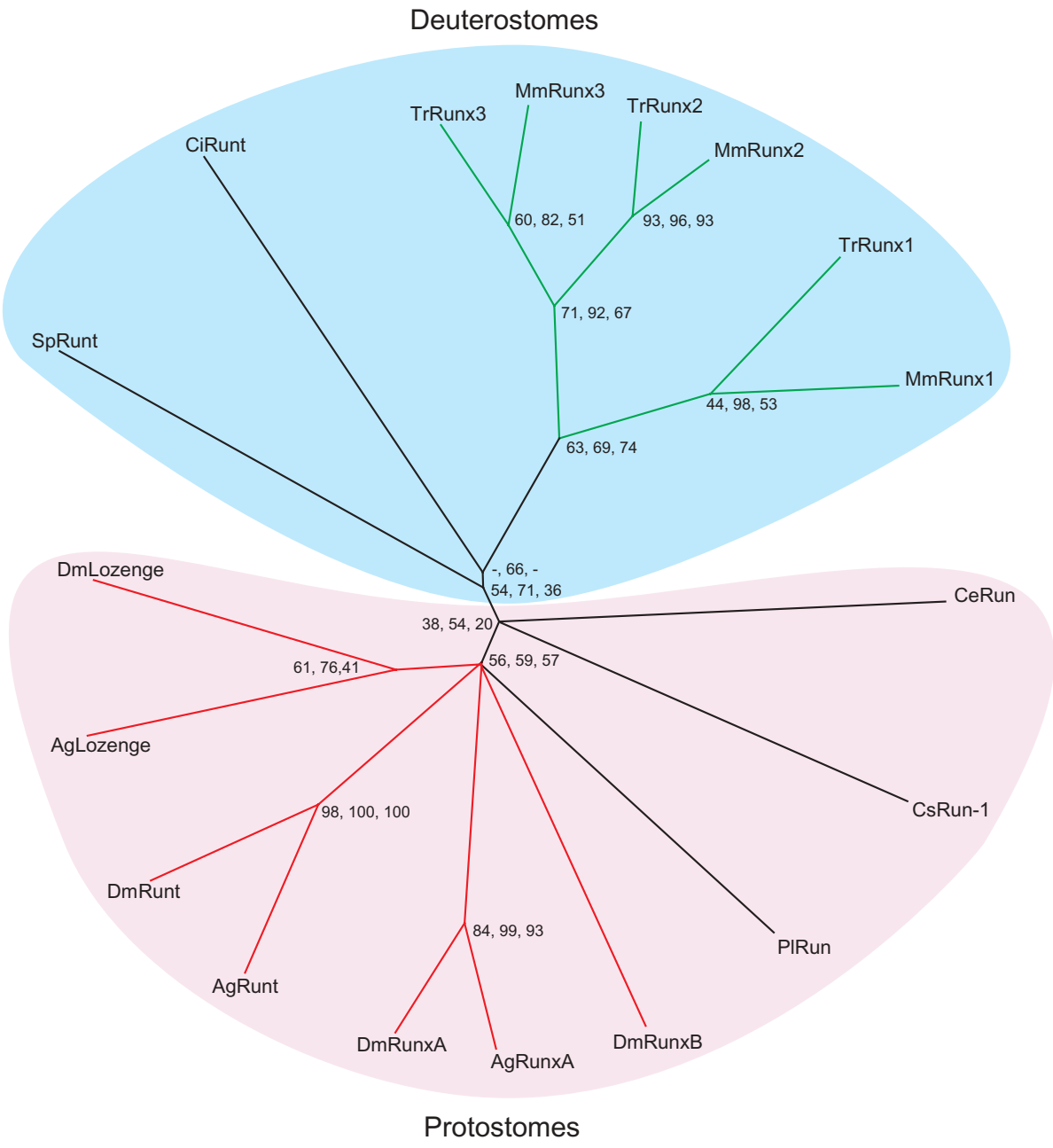


**Figure 3**

**Phylogenetic tree of the Runx protein family.** Maximum likelihood tree calculated with the assumption of a molecular clock from the multiple amino acid sequence alignment shown in Figure 2. The numbers at each node are the bootstrap support values obtained by maximum likelihood (no molecular clock assumed), neighbour joining, and maximum parsimony, in that order. The dashes (-) by the branch leading to the *Ciona gene* (*CiRunt*) indicate that the maximum likelihood and maximum parsimony methods did not support that node (these methods each place *CiRunt* on a branch with *TrRunx1*, with support values of 50 and 52, respectively). Branches with unresolved topology were collapsed. Coloured branches are used to highlight the vertebrate lineage (green) and the insect lineage (red). The shading highlights the deuterostome (blue) and protostome (pink) representatives in the sample. Abbreviations for species names are the same as in Figures 1 and 2.

**Table 1: Inferences from phylogenetic analysis of Runt domain sequences***

| | NEIGHBOR | PROML | PROTPARS | MRBAYES |
|---|---|---|---|---|
| Monophyly of deuterostome homologues? | Yes (71 % bootstrap support for the deepest branch) | Yes | Yes for vertebrates and urochordates (74%), ambiguous for echinoderms | Yes (93 % posterior probability) |
| *Runx3* basal in vertebrates? | No | No | No | No |
| Deepest deuterostome branch is a single-copy gene? | Yes | Yes | Yes | Yes |
| Deepest protostome branch is a single-copy gene? | Unclear (unresolved topology of the nematode and spider paralogues) | Unclear (unresolved topology of the nematode and spider paralogues) | Unclear (unresolved topology of the nematode and spider paralogues) | Unclear (unresolved topology of the nematode and spider paralogues) |

*Inferences from trees constructed using neighbour-joining (NEIGHBOR), maximum likelihood (PROML), maximum parsimony (PROTPARS), and Bayesian inference of phylogeny (MRBAYES) algorithms, as described in Methods. The PROML and PROMLK programs produced identical results.

## Conclusions

(1) In the public databases, the Runt domain is currently found only in sequences from bilaterians. This suggests that it may be a metazoan (and possibly a bilaterian) invention, and that the genomic regulatory networks through which Runx genes control the fate, proliferation and differentiation of cells are unique to animals.

(2) The primitive condition in bilaterians is most likely a single Runx gene, represented in the deuterostomes of our collection by the sea urchin *S. purpuratus* and the tunicate *C. intestinalis*, and in the protostomes of our collection by *C. elegans*. Runx gene duplications appear to have occurred independently in the lineages leading to vertebrates (which have at least three Runx genes) and insects (which have three or four Runx genes), suggesting that Runx genes in these latter organisms have probably acquired a number of specialized, taxon-specific regulatory functions (e.g., segmentation in arthropods [15], bone development in vertebrates [20], etc.). Thus, studies of Runx genes in sea urchins, tunicates, and nematodes are likely to highlight primitive, pan-bilaterian regulatory functions of this family of genes in the cell biology of animal development.

(3) The ancestral bilaterian Runx gene was apparently assembled from three exons and contained two introns, one within the sequence that encodes the Runt domain, and a second that borders the sequence encoding the C-terminal end of the Runt domain.

(4) The ancestor of all three vertebrate Runx genes utilized two alternative promoters that generate alternative N-terminal sequences.

## Methods

### *Collection and assembly of previously uncharacterised runx genes*

The *Ciona intestinalis* Runx gene was assembled with sequence files obtained from the Trace Archive database. A "seed" sequence was found in *C. intestinalis* gDNA, Ti 119616831 (Supplemental Table 1), by a MegaBLAST search using nucleotide sequence encoding exon 3 of *SpRunt* in the NCBI Trace server site http://www.ncbi.nlm.nih.gov/blast/mmtrace.html. A translation of the seed sequence also showed significant similarity to the Runt domain. The seed sequence showing Runx homology was in turn used as the query to obtain overlapping sequence and those obtained replaced the query. This repetitive BLAST process was continued such that the seed sequence was extended 25 kb in both the 5' and 3' direction relative to the coding orientation of the seed. The intron position assignments were aided by an alignment of the acquired *C. intestinalis* Runx gene with partial cDNA sequences obtained from a TBLASTN search using the *S. purpuratus* SpRunt-1 protein sequence to query the *C. intestinalis* cDNA project site at http://ghost.zool.kyoto-u.ac.jp/indexr1.html. Introns in those regions of the gene upstream of the Runt domain, not found in any cDNA, were manually assigned by spliced alignment, maintaining maximum homologue similarity. Specifically, putative introns were located by searching for the termini of open reading frames, and then positioned in such a way that the resulting open reading frame had maximum similarity to that of a homologous gene while the intron-exon junctions had the expected consensus 5'GT/-AG-3' sequence. A known intron in a similar position in an orthologous gene was considered a validation of correct position assignment. Highly conserved sequence blocks or residues, such as the Runt domain or the C-terminal Groucho-binding site VWRPY, served as anchoring reference points. Multiple overlapping sequences in the *C. intestinalis* contig provided a confidence in the final sequence assembly. Close inspection of the sequence alignments showed no variation among high confidence regions of the sequence where high confidence regions means at least 40–50 nucleotides from the end of the read or with few unassigned base calls near the region. The complete absence of sequence variation, the multiple sequence coverage, and the high sequence quality all indi-

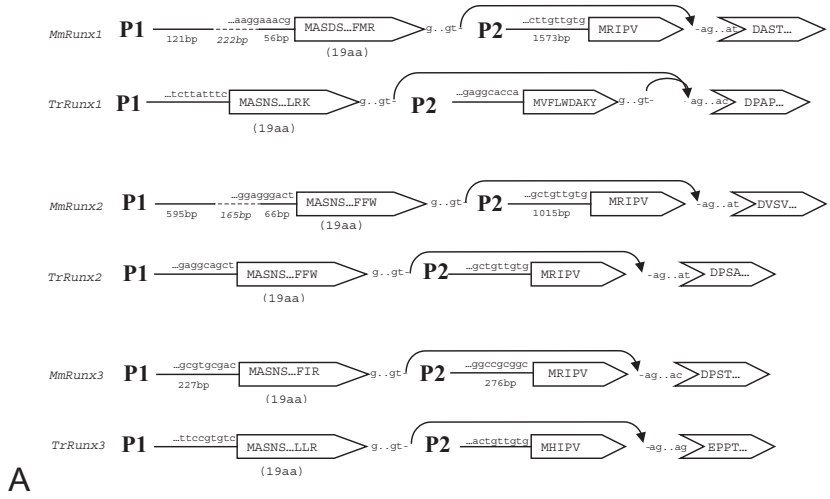


**Figure 4**

**Alternative transcripts and N-termini produced by transcription from proximal and distal promoters of vertebrate runx genes.** (A) Diagram of alternative transcripts from distal (P1) and proximal (P2) promoters in vertebrate Runx genes. Solid lines represent untranslated regions of exons, while dashed lines represent introns. *TrRunx1* distal transcript is theoretical and based on its similarities to other distal transcripts (i.e., its open reading frame begins with a sequence encoding MASNS), location (~14 kb from proximal transcript) and size (19 aa). *TrRunx2* and *TrRunx3* transcripts from P1 were derived from similarities to *MmRunx2* and *Danio rerio runtb*, respectively. *TrRunx2* was recently published [11]. *M. musculus* distal transcripts were gathered from Genebank entries and verified with publications (see Supplemental Table 1). Note: Unlike all of the other vertebrate Runx genes in our collection, *TrRunx1* transcript from P2 has a small exon (encoding 9 amino acids) followed by a 1 kb intron (see also B, below, and 3). This small exon splices into the same location within the downstream coding sequence as does the exon from the distal transcript. (B) Alternative N-terminal amino acid sequences of various vertebrate Runx genes analysed in our survey. N-termini from the distal promoters are shown in green, while the N-termini from the proximal promoters are shown in red. The aspartate or glutamate to which the N-terminal sequences from the distal promoter are joined are highlighted in yellow. The positions and nucleotides at the termini of introns are indicated in lower case.

cate that no additional Runx genes are present in the *C. intestinalis* genome, which was confirmed by inspection of the recently completed genome at http://genome.jgi-psf.org/ciona4/ciona4.home.html.

Three Runt domain genes in the *A. gambiae* genome and two previously unidentified Runt domain genes in the *D. melanogaster* genome were found by TBLASTN search using the *S. purpuratus* Runt protein sequence as the query. The newly identified *Drosophila* Runx gene sequences (CG1379 and CG15455) are present in a single sequence file (see Supplemental Table 1) and with opposite coding orientation and were named for this study *RunxA* (CG1379) and *RunxB* (CG15455). No cDNAs specific for the mosquito Runx genes or for the *Drosophila RunxA* or *RunxB* genes was found despite an extensive search, so putative intron positions were manually located in these genes by spliced alignment as described above. Unlike all of the deuterostome genes, the positions of introns among the insect genes were not conserved to the nucleotide, and were thus assigned with less confidence and should hence be considered provisional.

Three Runx genes were identified in the recently completed *T. rubripes* genome sequence by a TBLASTN search using the *S. purpuratus* Runt protein sequence as the query sequence at the *Fugu* BLAST Server of the UK Human Genome Mapping Project Resource Centre http://fugu.hgmp.mrc.ac.uk/blast/. cDNA sequences from the zebrafish (*D. rerio*) were used to place the start of the coding region in the distal promoter of *TrRunx1* and *TrRunx3*, and cDNA sequences from the teleost *O. laptipes* were used to place the start of the coding region in the distal promoter of *TrRunx2* (Supplemental Table 1). The promoter structure of *TrRunx2* published by Eggers et al. [11] during the course of our study, was in agreement with our results.

In addition to the gene sequences, two published partial cDNA sequences from a spider (*Cupiennius salei*) and an EST from a nematode (*Meloidogyne hapla*) were obtained from the NCBI database and used in the multiple sequence alignment. Since the nematode sequence and one of the spider sequences (*run-2*) only contain part of the runt domain, these were not used in the phylogenetic analysis.

#### *Multiple sequence alignments and phylogenetic tree construction*

The alignment of the amino acid sequences of all Runt domains used in this study was performed using the modified Clustal W [16] program in the AlignX® module of Vector NTI (InforMax, Inc.). The trees were calculated using programs from the PHYLIP package [Felsenstein, J. 1993–2002. PHYLIP (Phylogeny Inference Package) version 3.6a3. *Distributed by the author. Department of Genetics, University of Washington, Seattle*] or by the MrBayes program [18]. Specifically, 100 bootstrap replicates of the alignment were constructed using the SEQBOOT program, and the distances between sequences were computed by the PROTDIST program using the PAM substitution model. Neighbour joining trees were built using the NEIGHBOR program, and the consensus tree was derived using the CONSENSE program. For the maximum-likelihood trees, the experimental versions of the programs PROML (no assumption of molecular clock) and PROMLK (molecular clock assumed) from PHYLIP version 3.6a3 were used with the JTT evolutionary model and with the assumption of constant change rate between sites. The maximum parsimony trees were constructed using the PROTPARS program.

## Authors' contributions

J.R. performed most of the sequence analysis and drafted early versions of the Figures under the direction of A.J.R. J.A.C. assisted with the interpretation of the data and drafted the manuscript and final Figures. A.R.M performed statistical evaluation of the phylogenetic trees and wrote the relevant sections of the manuscript. A.J.R. conceived and designed the project, assisted with the bioinformatics, and developed the protocol for locating introns without cDNA sequence information.

## Additional material

### Supplemental Table 1

*The collection of Runx genes used in this study. This table indicates the sources of Runx gene sequences from selected species, and additional references that were used in the analysis. The* C. intestinalis *Runx gene (*CiRunt*) was derived from a contig assembled using shot-gun sequence segments queried from NCBI's trace archive. The accession number given can be used as a seed for querying the database start the retrieval of necessary sequence. Vector NTI ContigXpress was used to assemble the sequence. Verification of* CiRunt *was performed by comparing the gDNA to cDNA available on-line at the Nori Satoh Laboratory web page. Only the first cDNA sequence accession number is given and can be used as a seed to extract the remaining sequences. There was 2 × coverage of the cDNA. A special note needs to be made about the sequence upstream from the start of the Runt domain. The sequence from the start of the runt gene to the start of the Runt domain is theoretical. No cDNA could be found to verify this segment. The* T. rubripes *Runx genes were derived from scaffolds available on-line at the MRC HGMP-RC site.* M. musculus Runx2 *and* D. rerio runtb *were used to identify* T. rubripes Runx2 *and* Runx3, *respectively.* T. rubripes Runx2 *was verified by a published report [11].* D. rerio runta *was used to identify* T. rubripes Runx1. *The sea urchin gene SpRunt has not yet been completely sequenced, and sequence is currently available only for the SpRunt-1 cDNA [9] and for the termini of the exons and introns [10]. ND, not determined.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-3-4-S1.doc]

### Supplemental Table 2

*Sequence length of Runx genes from selected species. This table indicates the total and intron length in nucleotides of the sequences of each of the collected Runx genes from the proximal promoters. Exon lengths (bold numbers) are given in amino acids. Note that the nucleotide lengths of the* S. purpuratus *gene and introns are approximate (based on the lengths estimated for PCR products by gel electrophoresis), as these have not yet been completely sequenced.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-3-4-S2.doc]

### Supplemental Figure 1

*Comparison of relative exon-intron positions and splice sites among Runx genes. Details of exon-intron structure of genes analysed in this study. Exons are represented as boxes, and the number in the box indicates the number of amino acids. The central bold boxes represent the C-terminus of the Runt domain, which are contained in a single compact exon in all of the genes except for* DmRunt *and* AgRunt, *which lack the intron that borders the N-terminal end of this exon that is found in all of the other genes. The nucleotide sequences of the splice site termini are indicated: the junctions between exon and intron are represented by two dots (..), while the intervening intron sequence is indicated by a dash (-). The lengths of the introns in nucleotide number are indicated in parentheses. For the vertebrate Runx genes, only exons downstream of the proximal promoters are shown. The gene and intron lengths for the sea urchin gene* SpRunt *are approximate and based on PCR products obtained from a BAC clone [10] that has not yet been completely sequenced.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-3-4-S3.doc]

## Acknowledgements

This research was made possible by a Stowers Institute Functional Genomics Fellowship awarded to J.R., and was funded entirely by the Stowers Institute for Medical Research. We thank Dr. Galina Glazko (Pennsylvania State University) for assistance with the phylogenetic analysis, and anonymous reviewers for providing a number of constructive criticisms that improved the manuscript.

## References

1. Kagoshima H, Shigesada K, Satake M, Ito Y, Miyoshi H, Ohki M, Pepling M and Gergen P **The Runt domain identifies a new family of heteromeric transcriptional regulators** *Trends Genet* 1993, **9:**338-341
2. Crute BE, Lewis AF, Wu Z, Bushweller JH and Speck NA **Biochemical and biophysical properties of the core-binding factor alpha2 (AML1) DNA-binding domain** *J Biol Chem* 1996, **271:**26251-26260
3. Wheeler JC, Shigesada K, Gergen JP and Ito Y **Mechanisms of transcriptional regulation by Runt domain proteins** *Semin Cell Dev Biol* 2000, **11:**369-375
4. Coffman JA **Runx transcription factors and the developmental balance between cell proliferation and differentiation** *Cell Biology International* 2003, **In Press:**
5. Lund AH and van Lohuizen M **RUNX: a trilogy of cancer genes** *Cancer Cell* 2002, **1:**213-215
6. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ **Basic local alignment search tool** *J Mol Biol* 1990, **215:**403-410
7. Canon J and Banerjee U **Runt and Lozenge function in Drosophila development** *Semin Cell Dev Biol* 2000, **11:**327-336
8. Nam S, Jin YH, Li QL, Lee KY, Jeong GB, Ito Y, Lee J and Bae SC **Expression pattern, regulation, and biological role of runt domain transcription factor, run, in Caenorhabditis elegans** *Mol Cell Biol* 2002, **22:**547-554
9. Coffman JA, Kirchhamer CV, Harrington MG and Davidson EH **SpRunt-1, a new member of the runt domain family of transcription factors, is a positive regulator of the aboral ectoderm-specific CyIIIA gene in sea urchin embryos** *Dev Biol* 1996, **174:**43-54
10. Robertson AJ, Dickey CE, McCarthy JM and Coffman JA **The expression of SpRunt during sea urchin embryogenesis** *Mech Dev* 2002, **117:**327-330
11. Eggers JH, Stock M, Fliegauf M, Vonderstrass B and Otto F **Genomic characterization of the RUNX2 gene of Fugu rubripes** *Gene* 2002, **291:**159-167
12. Levanon D, Glusman G, Bangsow T, Ben-Asher E, Male DA, Avidan N, Bangsow C, Hattori M, Taylor TD, Taudien S, Blechschmidt K, Shimizu N, Rosenthal A, Sakaki Y, Lancet D and Groner Y **Architecture and anatomy of the genomic locus encoding the human leukemia-associated transcription factor RUNX1/AML1** *Gene* 2001, **262:**23-33
13. Bangsow C, Rubins N, Glusman G, Bernstein Y, Negreanu V, Goldenberg D, Lotem J, Ben-Asher E, Lancet D, Levanon D and Groner Y **The RUNX3 gene--sequence, structure and regulated expression** *Gene* 2001, **279:**221-232
14. Tahirov TH, Inoue-Bungo T, Morii H, Fujikawa A, Sasaki M, Kimura K, Shiina M, Sato K, Kumasaka T, Yamamoto M, Ishii S and Ogata K **Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta** *Cell* 2001, **104:**755-767
15. Damen WG, Weller M and Tautz D **Expression patterns of hairy, even-skipped, and runt in the spider Cupiennius salei imply that these genes were segmentation genes in a basal arthropod** *Proc Natl Acad Sci U S A* 2000, **97:**4515-4519
16. Thompson JD, Higgins DG and Gibson TJ **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice** *Nucleic Acids Res* 1994, **22:**4673-4680
17. Backstrom S, Wolf-Watz M, Grundstrom C, Hard T, Grundstrom T and Sauer U **The RUNX1 Runt Domain at 1.25A Resolution: A Structural Switch and Specifically Bound Chloride Ions Modulate DNA Binding** *J Mol Biol* 2002, **322:**259
18. Huelsenbeck JP and Ronquist F **MRBAYES: Bayesian inference of phylogenetic trees** *Bioinformatics* 2001, **17:**754-755

19. Holland PW, Garcia-Fernandez J, Williams NA and Sidow A **Gene duplications and the origins of vertebrate development** *Dev Suppl* 1994, 125-133
20. Komori T, Yagi H, Nomura S, Yamaguchi A, Sasaki K, Deguchi K, Shimizu Y, Bronson RT, Gao YH, Inada M, Sato M, Okamoto R, Kitamura Y, Yoshiki S and Kishimoto T **Targeted disruption of Cbfa1 results in a complete lack of bone formation owing to maturational arrest of osteoblasts** *Cell* 1997, **89:**755-764