Dev Genes Evol (2008) 218:567–574
DOI 10.1007/s00427-008-0244-x

SEQUENCE CORNER

Conserved cluster organization of insect Runx genes

Riyue Bao · Markus Friedrich

Received: 14 March 2008 / Accepted: 29 July 2008 / Published online: 29 August 2008
© Springer-Verlag 2008

Abstract Runx gene family transcription factors play important regulatory roles in metazoan development. The *Drosophila* genome contains four Runx genes, two of which are well studied (*runt* and *lozenge*) and two of which have not been explored in depth yet (*CG42267* and *CG34145*). In the absence of paralog-specific orthologs in vertebrates, we investigated the evolutionary conservation of the *Drosophila* Runx genes in the genomes of mosquito, red flour beetle, honeybee, and distantly related Bilateria. All four *Drosophila* Runx genes are conserved in other endopterygote insects. Strikingly, their genetic linkage and genomic arrangement is also highly conserved despite past recombination events in the same genomic region. The preservation of Runx gene cluster organization for at least 300 million years resembles that of insect Hox and selected Wnt genes. We propose a model for the conserved cluster organization of developmental gene family paralogs based on differential coduplication of regulatory elements that act over long distances.

Electronic supplementary material The online version of this article (doi:10.1007/s00427-008-0244-x) contains supplementary material, which is available to authorized users.

Communicated by: P. Simpson

R. Bao · M. Friedrich (✉)
Department of Biological Sciences, Wayne State University,
5047 Gullen Mall,
Detroit, MI 48202, USA
e-mail: friedrichm@wayne.edu

M. Friedrich
Department of Anatomy and Cell Biology,
Wayne State University, School of Medicine,
540 East Canfield Avenue,
Detroit, MI 48201, USA

Keywords Developmental gene family evolution · Synteny · Gene cluster · Wnt · Hox · Intron evolution

Introduction

The Runt box (Runx) genes encode a family of heteromeric transcription factors, which are characterized by the highly conserved DNA- and protein-binding Runt domain of around 130 amino acids (Kagoshima et al. 1993). Runx genes have been discovered in mammals, nematodes, lancelet fish, and sea urchins (Bae and Lee 2000; Robertson et al. 2002; Stricker et al. 2003). Studies in *Drosophila* and mouse revealed that Runx transcription factors play important roles in metazoan development (for review, see Coffman 2003). Genetic depletion of the primordial Runx gene *runt* (*run*) in *Drosophila* leads to periodic deletions of larval segments (Gergen and Butler 1988; Gergen and Wieschaus 1985). In addition, *Drosophila run* plays roles in sex determination and neurogenesis (Duffy et al. 1991; Torres and Sanchez 1992). A second *Drosophila* Runx gene, *lozenge* (*lz*), is required for patterning in the developing antenna and the eye as well as in hematopoiesis for the differentiation of crystal cells (Bataille et al. 2005; Daga et al. 1996; Gupta and Rodrigues 1995).

The *Drosophila* genome project revealed two additional Runx genes, which are genetically linked with *run* and *lz* on the X chromosome (Rennert et al. 2003): *CG42267* and *CG34145*. These genes have not yet been studied in depth, but a genome-wide RNA interference (RNAi) screen identified an isoform of *CG42267* to be involved in the control of cell survival (Boutros et al. 2004).

Orthologs of *Drosophila run* have been described in the red flour beetle *Tribolium castaneum* and the honeybee *Apis mellifera* (Choe et al. 2006; Dearden et al. 2006). In

addition, the African malaria mosquito *Anopheles gambiae* was found to possess at least three Runx gene homologs (Rennert et al. 2003). To clarify the evolutionary conservation of *Drosophila* Runx genes, we searched for orthologs in the genomes of *Aedes aegypti* (yellow fever mosquito), *T. castaneum*, *A. mellifera*, *Nasonia vitripennis* (parasitoid wasp), *Strongylocentrotus purpuratus* (sea urchin), and *Nematostella vectensis* (sea anemone). Our analysis revealed that all *Drosophila* Runx paralogs are highly conserved in the insect species. Strikingly, the genomic arrangement of insect Runx genes is preserved as well, despite evolutionary turnover of unrelated genes in the same genomic region. These findings imply that insect Runx genes have been conserved as a gene cluster for more than 300 million years, pointing to underlying regulatory constraints. A model of developmental gene cluster evolution is proposed assuming the inherited, and thus shared, need for regulation by a single ancestral enhancer element.

Materials and methods

Sequence retrieval and ortholog search

The *Drosophila melanogaster lz*, *CG42267*, *CG34145*, and *run* sequences (accession numbers FBgn0002576, FBgn0259162, FBgn0083981, and FBgn0003300) were retrieved from Flybase and used as queries in searches against the genome sequence databases of yellow fever mosquito (*A. aegypti* genome database version 1.0), red flour beetle (*T. castaneum* Georgia GA2 genome database version 1.1), honeybee (*A. mellifera* DH4 genome database version 4.0), parasitic wasp (*N. vitripennis* genome assembly 1.1), sea urchin (*S. purpuratus* genome assembly 2.1), sea anemone (*N. vectensis* genome assembly 1.0), and mouse (*Mus musculus* genome database version 37.1) with BLASTP or TBLASTX (Altschul et al., 1997). The *Branchiostoma lanceolatum* (lancelet fish) *Runt* sequence was obtained through keyword search in the National Center for Biotechnology Information (NCBI) protein database (accession number AAN08565; Stricker et al. 2003).

Multiple sequence alignment

Multiple protein sequence alignments were generated with CLUSTAL W (Thompson et al. 1994) and inspected by eye. Alignment sites with gaps in the Runx domain were eliminated for tree reconstruction except for sites that included gaps in *A. gambiae CG42267* and *N. vectensis Runt*, in which case gaps were due to missing sequence in the open reading frame prediction rather than evolutionary divergence of the protein sequence.

Gene tree reconstruction

Neighbor joining analysis with JTT amino acid substitution model was run in Phylip 3.66 (Felsenstein 2005; Jones et al. 1992). Protein maximum parsimony tree reconstruction was carried out in MEGA version 4.0 (Saitou and Nei 1987; Tamura et al. 2007). TREE-PUZZLE analysis was run with TREE-PUZZLE version 4.0 (Schmidt et al. 2002). Branch support in maximum parsimony and neighbor joining trees was assessed by nonparametric bootstrap on 100 pseudoreplicates (Felsenstein 1986).

Genomic investigation of linkage

Physical gene positions were determined using the NCBI Entrez Map Viewer of the *D. melanogaster*, *T. castaneum*, and *A. mellifera* genomes (http://www.ncbi.nlm.nih.gov/mapview/). In *A. aegypti*, because of the fragmented status of the available genome (Waterhouse et al. 2008), physical gene positions were retrieved from supercontigs in the NCBI nucleotide database. Supercontigs NW_001810917 and NW_001810728 were linked based on the updated gene model of the *A. aegypti CG42267* ortholog, which extends into both contigs (data in Electronic supplementary material).

Results and discussion

Conservation of insect Runx paralogs

To explore the origin of the four *Drosophila* Runx genes, we performed protein sequence BLAST searches in the genome databases of *A. aegypti*, *T. castaneum*, *A. mellifera*, *N. vitripennis*, *S. purpuratus*, and *N. vectensis*. Outside the insects, our search discovered a new *S. purpuratus* Runx gene (*Spur_Runt2*) in addition to the previously published *Spur_Runt1* gene (Rennert et al. 2003). Gene tree analysis provided strong support that the mammalian and sea urchin Runx paralogs dated back to independent gene-duplication events (Fig. 1). Only one Runx homolog was found in the sea anemone and the lancelet fish (Stricker et al. 2003). Consistent with previous conclusions (Rennert et al. 2003), these findings demonstrated that the late ancestor of Metazoa possessed a single Runx gene, which experienced independent duplications in at least vertebrates, echinoderms, and arthropods.

Four Runx family member genes were found in all investigated insect species. To clarify the phylogenetic relationships between the insect Runx sequences, we generated a multiple alignment of the conserved sequence regions that encompass the Runx domain (data in Electronic supplementary material). Phylogenetic tree estimation

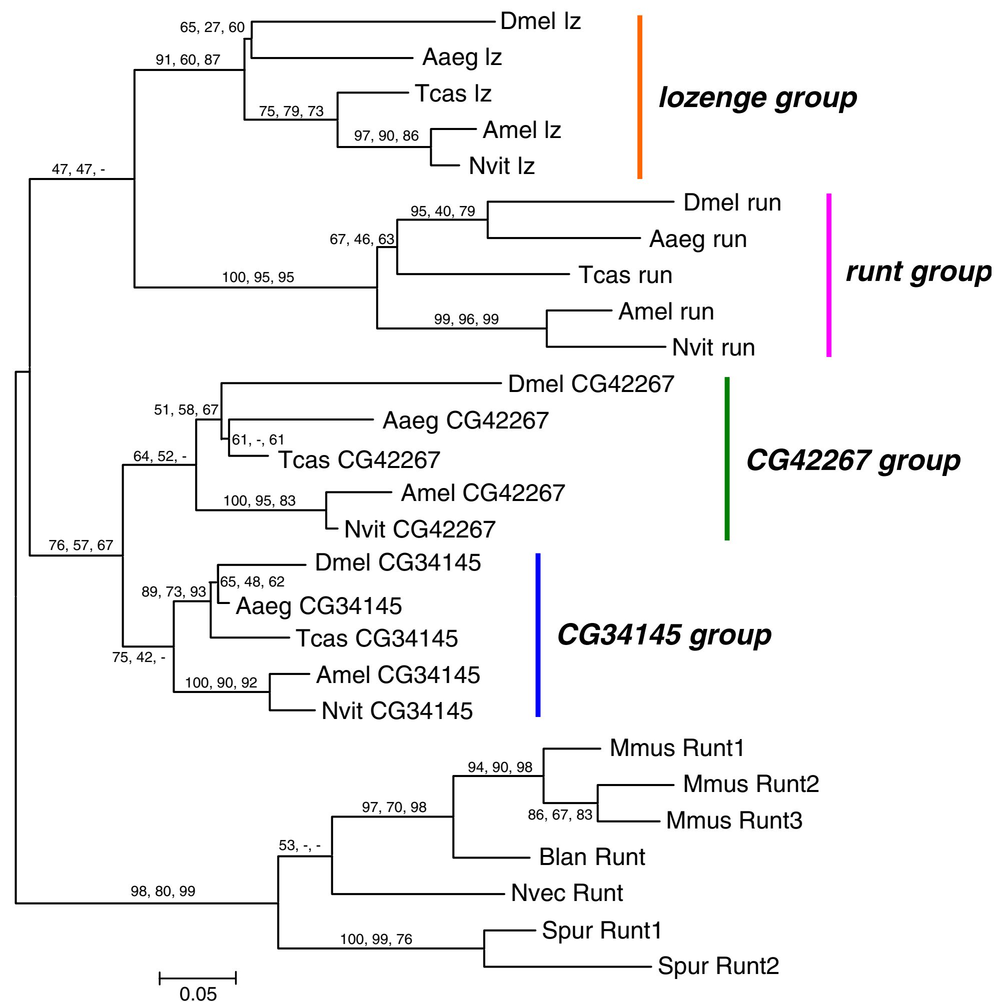

Fig. 1 Phylogenetic analysis of insect Runx genes. Neighbor-joining tree based on 142 conserved amino acid sites including the Runx domain. Tree branch length was adjusted by likelihood mapping in TREE-PUZZLE version 4.0 with neighbor-joining tree topology as the input. *Left branch support numbers* reflect neighbor-joining support, *middle branch support numbers* reflect maximum parsimony support, and *right branch support numbers* reflect TREE-PUZZLE support. *Bar* represents 0.05 amino acid substitutions per site. *Aaeg*, *A. aegypti*; *Amel*, *A. mellifera*; *Blan*, *B. lanceolatum*; *Dmel*, *D. melanogaster*; *Mmus*, *M. musculus*; *Nvec*, *N. vectensis*; *Nvit*, *N. vitripennis*; *Spur*, *S. purpuratus*; *Tcas*, *T. castaneum*

strongly supported four orthology groups, each of which included one of the four *Drosophila* Runx genes (Fig. 1). The deuterostome and cnidarian homologs rooted the insect Runx orthology groups such that the *run* and *lz* groups formed a metacluster that was sister to a second metacluster composed of the *CG42267* and *CG34145* orthology groups. This topology implied that the four insect Runx orthology groups originated by parallel duplications of two ancestral sister paralogs. However, considering the low support of the branch uniting the *run* and *lz* groups metacluster (<47), it is also possible that *lz*, *CG42267*, and *CG34145* originated through consecutive duplications, in which case *run* is the oldest paralog. Interestingly, the latter hypothesis is also weakly supported in trees in which the Runx homolog of *Caenorhabditis elegans* (*Runt* related family member, *rnt-1*) is included (data in Electronic supplementary material). The *C. elegans rnt-1* sequence, however, has experienced extreme substitution rate acceleration and breaks up the monophyly of the Runx cluster, possibly due to a long branch attraction artifact.

Dynamic intron evolution in insect Runx paralogs

Comparative gene structure analysis confirmed the presence of an approximately 40 amino acid long exon at the C-terminal region of the Runx domain flanked by two highly conserved introns (ancient introns 1 and 2 in Fig. 2) (Rennert et al. 2003). The *run* orthologs of *Drosophila*, mosquito, red flour beetle, and wasp, however, lacked ancient intron 1, which is still present in the honeybee. Considering the closer relationship of Coleoptera to Diptera than to Hymenoptera (Savard et al. 2006), it is possible that ancient intron 1 was lost in the ancestor of Coleoptera and Diptera. However, this scenario also implies a second loss in the parasitic wasp *Nasonia*. Further, we noted that the *Apis* intron-corresponding ancient intron 1 is very small (123 bp). It is therefore also possible that ancient intron 1was lost before the diversification of the endopterygote insect lineages and that *Apis* convergently regained a small intron at the same position. Requiring two evolutionary changes, this scenario is equally parsimonious as the first,

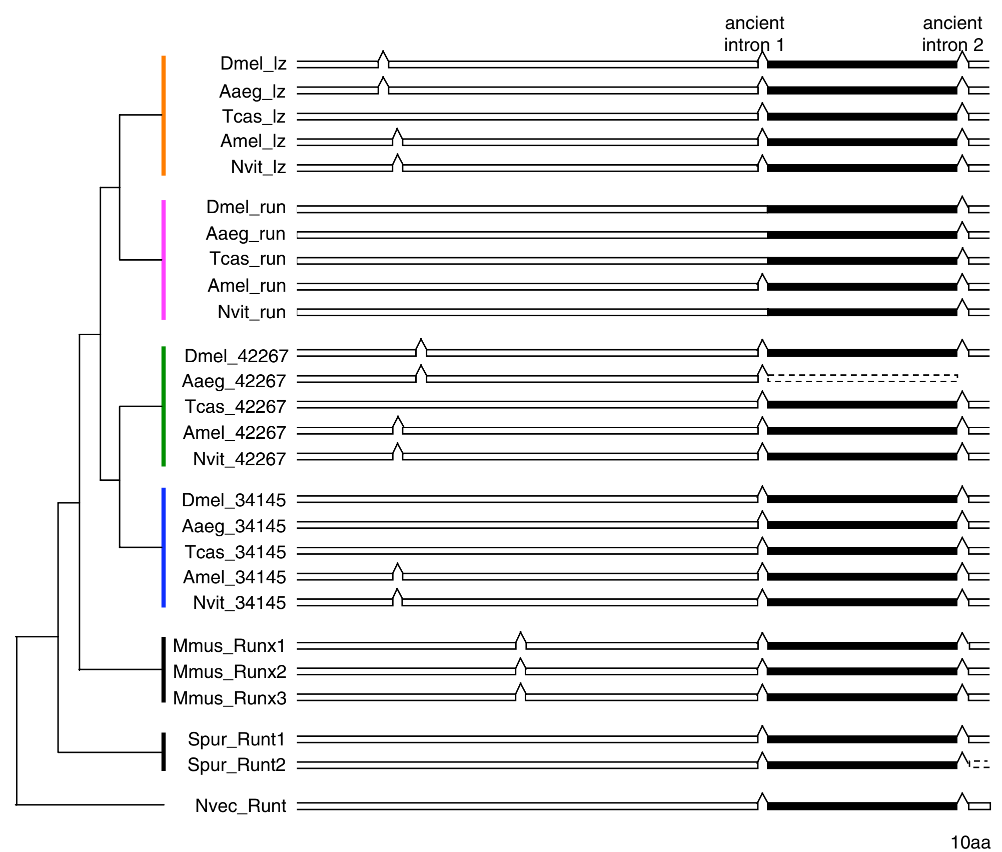

Fig. 2 Intron conservation and divergence in the Runx domain. *Boxes* indicate exons and *connecting lines* introns. Exons drawn relative to the scale bar. *White boxes* represent N-terminal Runx domain with variable N-terminal introns. *Black boxes* represent the highly conserved exon in the C-terminal Runx domain. *Dashed line* indicates region of incomplete gene model prediction. Gene and species name abbreviations same as in Fig. 1. See text for details

which requires at least two losses (Coleoptera + Diptera, *Nasonia*). Data from a wider range of Hymenoptera should resolve this issue.

Also, the N-terminal region of the Runx domain is characterized by evidence of dynamic intron gain or loss. Most strikingly, in the Hymenoptera *lz*, *CG42267* and *CG34145* gene share an intron at the N-terminal end of the Runx domain (Fig. 2). The dipteran *lz* and *CG42267* genes likewise contain introns in this region. Their exact position, however, is only conserved between orthologs but not paralogs. Moreover, no comparable introns are present in the *Tribolium* Runx homologs. Considering the most consistently supported Runx family tree (Fig. 1), the dipteran *lz* and *CG42267* specific introns were most likely acquired after the diversification of Diptera from Coleoptera and Hymenoptera. The scattered distribution of the position-conserved hymenopteran introns in *lz*, *CG42267* and *CG34145*, on the other hand, could have been caused by multiple losses of an ancestral intron in Diptera and Coleoptera. Alternatively, concerted evolution may have caused the spread of an intron that originated during hymenopteran evolution. Yet another possibility is that, assuming that *run* is the oldest paralog, the acquisition of the position-conserved hymenopteran intron could have occurred in the precursor gene to *lz*, *CG42267*, and *CG34145*. Consistent with this, an early divergence of *run* is the second strongest supported hypothesis in gene tree reconstruction (data in Electronic supplementary material). However, also this scenario implies multiple intron losses in Coleoptera and Diptera after the diversification of major endopterygote lineages. At this point, the sequence of intron loss and gain in the insect Runx genes can only be tentatively reconstructed. The data are unambiguous, however, in highlighting evolutionary turnover of gene structure in the N-terminal Runx domain region.

Conservation of insect Runx paralog linkage

In *Drosophila*, all four Runx genes map within a region of 11.4 Mb on the X chromosome (Fig. 3). To explore if the genetic linkage of *Drosophila* Runx genes was also conserved, we investigated the genomic location of Runx genes in *Tribolium*, *Apis*, and *Aedes* (Honeybee Genome Sequencing Consortium 2006; Tribolium Genome Sequencing Consortium 2008). Remarkably, in both *Tribolium* and *Apis*, all Runx genes were located within less than 4 Mb on the same chromosome. The region spanning orthologs of *CG42267*, *CG34145*, and *run* was less than 0.16 Mb long in *Drosophila*, *Tribolium*, and *Apis*. Moreover, the arrangement and orientation of *CG42267*, *CG34145*, and *run* orthologs were identical in all four species. Differences in the genomic organization of the Runx paralogs stemmed only from the

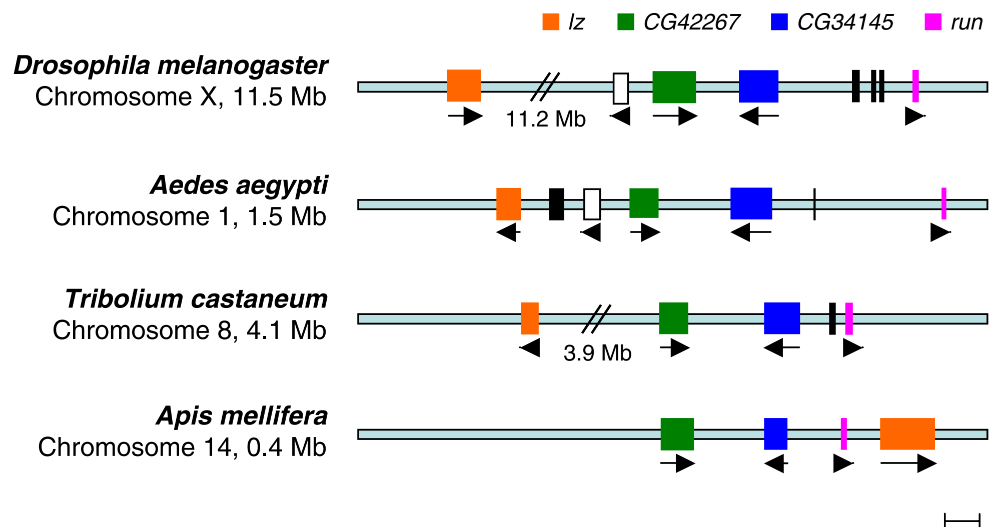

Fig. 3 Conserved linkage of insect Runx genes. The *number next to the chromosome name under each species name* represents the length of region shown in the figure. Runx loci are indicated by *blocks* colored consistent with Fig. 1. Open reading frame directions indicated by *arrows*. Non-Runx genes indicated by *black boxes* except for *Tak1*, which is indicated by *white box*. Scale reflects 20 kb in *D. melanogaster*, *T. castaneum*, and *A. mellifera* but 80 kb in *A. aegypti*. See text for details

more variable position of *lz*. In *Drosophila*, *lz* was located 11.2 Mb distally of *CG42267* relative to the centromere. In *Tribolium*, *lz* was less strongly linked to the other Runx paralogs being separated from *CG42267* by 3.86 Mb. In addition, *Tribolium lz* was inverted compared to the orientation of *Drosophila lz*. *Aedes* also possesses an inversion in *lz*; however, *lz* is tightly linked to *CG42267* (0.25 Mb in distance). In *Apis*, *lz* was likewise tightly linked to the other Runx paralogs but proximal to *run*, thus implying a relocation from one end of the cluster to the other (Fig. 3). Taken together, these data revealed that the close linkage of *run* with *CG42267* and *CG34145* in *Drosophila* is an ancestral aspect of the genomic organization of insect Runx genes. Second, the comparatively weak linkage of *lz* to the rest of the clustered Runx genes in *Drosophila* is likely derived as *lz* is more closely linked in the other insect genome model species including *Aedes*. However, *lz* in general is of more variable position than the rest of the Runx genes, even within different *Drosophila* species (data not shown).

Evidence of past gene rearrangements in the genomic environment of the Runx paralog arrays

Considering the evidence for conserved microsynteny in the Runx gene-containing genomic region, we investigated the possibility of conserved linkage of additional protein-coding genes (Fig. 3). In the region between *CG42267* and *run*, we found no evidence for further linkage conservation besides that of the Runx paralogs. *Apis* was unique in containing no additional gene models in the entire Runx gene region. The region between *CG42267* and *CG34145* was free of further gene models in all species. In the region between *CG34145* and *run*, *Tribolium* contained the ortholog of the *Drosophila yellow-f2* (XP_969206). This arrangement was unique to *Tribolium* with *yellow-f2* located on different chromosomes than the Runx genes in *Aedes* and *Drosophila*. In *Drosophila*, three different genes are contained in the segment between *CG34145* and *run* (Fig. 3). Two of these genes (accession numbers NP_608404 and NP_608405) were *Drosophila* orphan genes based on the lack of significant BLAST hits in other insect genome models. The third gene was a member of the Cytochrome P450 family (accession number NP_608403). In *Aedes*, only single gene model (XP_001657549) was found between the *CG34145* and *run* orthologs. This gene model lacked significant similarity to known genes and therefore may likewise be an orphan gene. A second candidate *Aedes* orphan gene was present between *lz* and *CG42267* together with the ortholog of *Drosophila TGF-beta activated kinase 1* (*Tak1*) (Fig. 3). In both *Drosophila* and *Aedes*, the *Tak1* gene was located in the Runx gene cluster, distal to *CG42267* (Fig. 3). *Tak1* is therefore the sole position-conserved gene in the genomic environment of the clustered Runx genes that is not related to the Runx gene family. This situation, however, occurred only to dipteran genomes in the present sample of species. In the four-way genome comparison of *Drosophila*, *Aedes*, *Tribolium*, and *Apis*, the conservation of microsynteny applied exclusively to the Runx gene paralogs.

Similar degrees of genetic linkage in the Runx and Hox gene clusters

The finding that the close linkage of insect Runx paralogs remained strongly conserved despite rearrangements of other genes in the same region suggested that the Runx genes evolved as a cluster under the impact of linkage-enforcing constraints. The mammalian Runx paralogs *Runx1*, *Runx2*, and *Runx3* are located on different chromosomes and therefore cannot serve as reference paradigm for assessing the cluster status of insect Runx genes. The

paradigm example of conserved microsynteny of closely related members of a developmental gene family is the Hox gene cluster (reviewed in Garcia-Fernandez 2005). To assess if the conserved linkage of insect Runx genes represented similar features, we compared the genomic evolution of the Runx genes with that of the Hox gene cluster in *Drosophila*, *Aedes*, *Tribolium*, and *Apis*.

We first asked if the members of the Hox and Runx regions exhibited a similar degree of linkage. *Aedes* was excluded from this analysis because the *Aedes* Hox gene cluster has not yet been described in detail. Not counting the more extensively rearranged *lz* paralog, the average intergenic distances between Runx paralogs are 59,122, 35,657, and 36,293 bp in *Drosophila*, *Tribolium*, and *Apis*, respectively, compared to 47,699, 84,141, and 115,890 bp in the Hox complexes of the same species (Brown et al. 2002; Dearden et al. 2006; Drysdale and Crosby 2005; Lewis et al. 2003). Thus taken together, *CG42267*, *CG34145*, and *run* are similarly closely linked like Hox genes in Diptera and even more closely linked in Coleoptera and Hymenoptera. Next, we noted that, similar to the case of the Runx gene array, the Hox gene cluster of some species (such as *Apis* and *Drosophila*) is colonized by genes that are not members of the Hox gene family (Dearden et al. 2006; Negre and Ruiz 2007; Shippy et al. 2008). Also similarly, in no case is linkage of these genes conserved (Shippy et al. 2008). This parallel suggested that, in the case of both the Hox and Runx gene regions, constraints act on preserving linkage specifically between the developmental gene family members in spite of recombination events that lead to rearrangements of unrelated genes.

Relaxed developmental gene cluster conservation in *Drosophila*

The Hox gene cluster of *Drosophila* is characterized by a higher number of modifications compared to that of *Apis* and *Tribolium*. In *Drosophila*, the Hox genes are divided into the *Antennapedia* and *Bithorax* gene complexes (*ANT*-C and *BX*-C), while the Hox genes remained preserved in a single array in *Anopheles*, *Tribolium*, and *Apis* (Brown et al. 2002; Dearden et al. 2006; Drysdale and Crosby 2005; Lewis et al. 2003). Second, five transpositions of protein coding genes other than Hox gene family transcription factors were discovered within the Hox cluster of *Drosophila*, while no transposition events occurred in the mosquito and *Tribolium* Hox cluster and presumably only one in *Apis* (Dearden et al. 2006; Negre and Ruiz 2007; Shippy et al. 2008). In combination, these data revealed a striking parallel in the substantially more dramatic diversification of the Hox and Runx gene clusters in *Drosophila* compared to *Tribolium* and *Apis*.

It has been recently concluded that the higher similarity of the *Tribolium* Hox cluster organization to that in other metazoans indicates that the *Tribolium* Hox cluster represents a more ancestral organization than the partially dissociated *Drosophila* Hox cluster (Shippy et al. 2008). As outgroup data are lacking for the Runx gene cluster, we suggest by argument of analogy that the Runx gene clusters of *Apis* and *Tribolium* likewise represent more ancestral states of genomic organization. In summary, considering that the insect Runx and Hox gene clusters are consistently similar in their evolutionary dynamics, we conclude that the insect Runx gene array represents a conserved gene family paralog cluster.

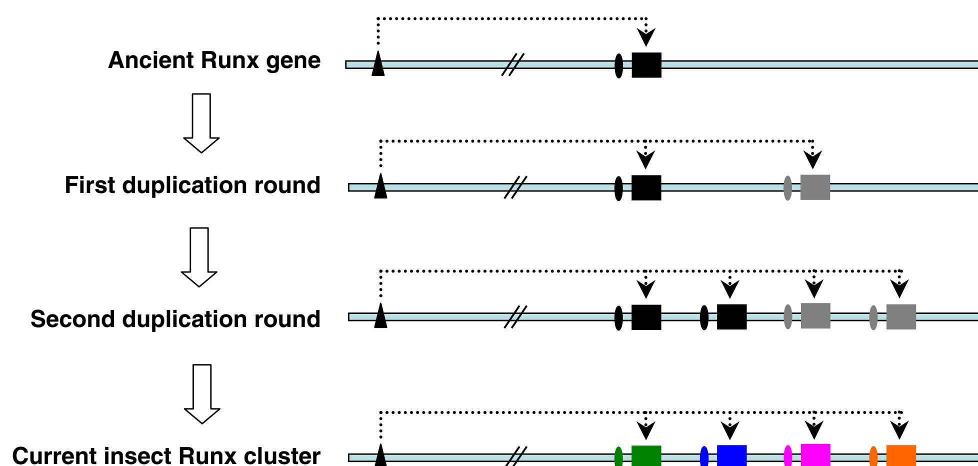

Fig. 4 Developmental gene family cluster conservation by differential coduplication of cis-regulatory elements. *Triangle* represents putative long-distance enhancer. *Ellipses* represent proximal *cis*-regulatory elements upstream of the gene. *Boxes* represent protein-coding regions. *Dashed lines* indicate regulatory interaction between the putative shared long-distance enhancer and members of the expanding developmental gene cluster. Ancestral and intermediate Runx genes are indicated by *black and grey boxes*. Subfunctionalized paralogs in the current Runx cluster are colored consistent with Fig. 3. See text for details

A model for the evolution of developmental gene family clusters: differential coduplication of *cis*-regulatory elements

Our findings lead to the question of whether the clustering of insect Runx genes is of functional significance and, if so, which functional constraints may be responsible. The nature of functional constraints that lead to the preservation of cluster organization has been most extensively pursued in the case of the Hox gene complex (Kmita and Duboule 2003). While a number of factors have been identified in vertebrates, the forces underlying cluster conservation in invertebrates, and thus the most ancestral type of cluster, is still elusive. Indeed, the very existence of such constraints has been questioned (Negre and Ruiz 2007). Recent analysis of the genome of the sea anemone *N. vectensis* discovered a significant degree of synteny with vertebrate genome structure (Putnam et al. 2007). As long as the reasons for this surprising degree of linkage conservation remain unclear, these data may indicate that the conservation of synteny could reflect rarity of chromosomal rearrangements. This model has been referred to as “phylogenetic inertia” (Negre and Ruiz 2007). However, the deeply conserved synteny in *Nematostella* applies to genes maintained on corresponding chromosomal scaffolds despite major rearrangements (macrosynteny) rather than the conservation of precise local gene order (microsynteny; Zdobnov and Bork 2007). Further, the genome sequence comparison between *Anopheles* and *Drosophila* uncovered only a very limited fraction of genes in conserved synteny (~30%; Bolshakov et al. 2002; Zdobnov and Bork 2007; Zdobnov et al. 2002). Even more dramatically, only 7% of protein-coding genes were located in conserved microsynteny groups in the comparison of *Drosophila*, *Anopheles*, and *Apis* (Honeybee Genome Sequencing Consortium 2006). These numbers may be conservative in that annotation gaps or mistakes in the more recently published genome drafts may prevent the identification of microsynteny in a significant number of cases. Indeed, the Runx cluster is not included in previous large-scale studies on genome synteny in insects (Bolshakov et al. 2002; Severson et al. 2004; Zdobnov and Bork 2007; Zdobnov et al. 2002). Nonetheless, the available data show that, while macrosynteny is indeed stable over long periods of time, the preservation of precise local gene order as in the case of the Hox and Runx gene clusters is exceptional and hence more likely to be the result of conserving constraints. Of note, evidence for a similar case of developmental paralog cluster conservation has recently been described for the *Wnt1*, *Wnt6*, *Wnt9*, and *Wnt10* paralogs in insects (Bolognesi et al. 2008). This suggests that developmental gene cluster organization is a more widespread phenomenon than currently appreciated.

Based on the Hox gene cluster paradigm, chromatin regulation or global enhancers may lead to the conservation of the Runx gene cluster (Kmita and Duboule 2003). In agreement with the second mechanism, recent studies discovered a significant correlation between microsynteny conservation and the presence of highly conserved noncoding sequence elements in insect genomes (Engstrom et al. 2007). Moreover, these genomic regions are significantly enriched with transcriptional regulators of development and thus form genomic regulatory blocks (Engstrom et al. 2007). Unfortunately, we were unable to assess if this is the case for the Runx gene cluster because the orthology of genes sampled in synteny blocks is not yet documented in an accessible manner (Engstrom et al. 2007; Waterhouse et al. 2008). Nonetheless, long-distance regulatory elements like enhancers are the most likely constraining force responsible for this cluster. Functional and comparative genomic studies of *CG42267* and *CG34145* in *Drosophila* and other insect models have the potential to reveal *cis*-regulatory ties between these two genes and *run*.

We note that future studies of Hox, Wnt, and Runx gene cluster conservation hold the promise to identify general factors of developmental gene cluster conservation. Importantly, the insect Runx cluster is of more recent origin and lower complexity. It should therefore be easier to understand the evolutionary origin of cluster conservation constraints. As a first step in this direction, we propose a model in which the inheritance of regulatory dependence on an ancestral long-distance *cis*-regulatory element may enforce cluster preservation (Fig. 4). That is, the primordial protostome Runx gene was regulated by a long-distance enhancer element. Mutational events of a large enough scale to duplicate the coding region and closely linked *cis*-regulatory information but too small to coduplicate the long-distance enhancer would lead to a situation in which the correct transcriptional regulation of the resulting new duplicates would remain dependent on the same long-distance regulatory element. The continued occurrence of such events can be imagined to lead to a string of tandem duplicated developmental gene paralogs, whose recombinatorial flexibility is limited by the reach of the essential enhancer element.

Acknowledgments We are grateful to Susan St.Pierre for prompt Flybase gene annotation correction, Jason Caravas for comments on the manuscript, and the two anonymous reviewers for stimulating suggestions. Riyue Bao was supported by a WSU College of Arts and Sciences Rumble Research Fellowship. Markus Friedrich was supported by a WSU Career Development Chair Award.

References

Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Bae SC, Lee J (2000) cDNA cloning of *run*, a *Caenorhabditis elegans* Runt domain encoding gene. Gene 241:255–258

Bataille L et al (2005) Resolving embryonic blood cell fate choice in *Drosophila*: interplay of GCM and RUNX factors. Development 132:4635–4644

Bolognesi R et al (2008) *Tribolium* Wnts: evidence for a larger repertoire in insects with overlapping expression patterns that suggest multiple redundant functions in embryogenesis. Dev Genes Evol 218:193–202

Bolshakov VN et al (2002) A comparative genomic analysis of two distant diptera, the fruit fly, *Drosophila melanogaster*, and the malaria mosquito, *Anopheles gambiae*. Genome Res 12:57–66

Boutros M et al (2004) Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. Science 303:832–835

Brown SJ et al (2002) Sequence of the *Tribolium castaneum* homeotic complex: the region corresponding to the *Drosophila melanogaster* antennapedia complex. Genetics 160:1067–1074

Choe CP et al (2006) A pair-rule gene circuit defines segments sequentially in the short-germ insect *Tribolium castaneum*. Proc Natl Acad Sci U S A 103:6560–6564

Coffman JA (2003) Runx transcription factors and the developmental balance between cell proliferation and differentiation. Cell Biol Int 27:315–324

Daga A et al (1996) Patterning of cells in the *Drosophila* eye by Lozenge, which shares homologous domains with AML1. Genes Dev 10:1194–1205

Dearden PK et al (2006) Patterns of conservation and change in honeybee developmental genes. Genome Res 16:1376–1384

Drysdale RA, Crosby MA (2005) FlyBase: genes and gene models. Nucleic Acids Res 33:D390–D395

Duffy JB et al (1991) Expression and function of the *Drosophila* gene *runt* in early stages of neural development. Development 113: 1223–1230

Engstrom PG et al (2007) Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res 17:1898–1908

Felsenstein J (1986) Jackknife, bootstrap and other resampling methods in regression analysis discussion. Ann Stat 14:1304–1305

Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle

Garcia-Fernandez J (2005) The genesis and evolution of homeobox gene clusters. Nat Rev Genet 6:881–892

Gergen JP, Butler BA (1988) Isolation of the *Drosophila* segmentation gene *runt* and analysis of its expression during embryogenesis. Genes Dev 2:1179–1193

Gergen JP, Wieschaus EF (1985) The localized requirements for a gene affecting segmentation in *Drosophila*: analysis of larvae mosaic for *runt*. Dev Biol 109:321–335

Gupta BP, Rodrigues V (1995) Distinct mechanisms of action of the Lozenge locus in *Drosophila* eye and antennal development are suggested by the analysis of dominant enhancers. J Neurogenet 10:137–151

Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443:931–949

Jones DT et al (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Kagoshima H et al (1993) The Runt domain identifies a new family of heteromeric transcriptional regulators. Trends Genet 9:338–341

Kmita M, Duboule D (2003) Organizing axes in time and space; 25 years of colinear tinkering. Science 301:331–333

Lewis EB et al (2003) Evolution of the homeobox complex in the Diptera. Curr Biol 13:R587–R588

Negre B, Ruiz A (2007) HOM-C evolution in *Drosophila*: is there a need for Hox gene clustering. Trends Genet 23:55–59

Putnam NH et al (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94

Rennert J et al (2003) The evolution of Runx genes I. A comparative study of sequences from phylogenetically diverse model organisms. BMC Evol Biol 3:4

Robertson AJ et al (2002) The expression of *SpRunt* during sea urchin embryogenesis. Mech Dev 117:327–330

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

Savard J et al (2006) Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of holometabolous insects. Genome Res 16:1334–1338

Schmidt HA et al (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Severson DW et al (2004) Comparative genome analysis of the yellow fever mosquito *Aedes aegypti* with *Drosophila melanogaster* and the malaria vector mosquito *Anopheles gambiae*. J Hered 95:103–113

Shippy TD et al (2008) Analysis of the *Tribolium* homeotic complex: insights into mechanisms constraining insect Hox clusters. Dev Genes Evol 218:127–139

Stricker S et al (2003) A single amphioxus and sea urchin runt-gene suggests that runt-gene duplications occurred in early chordate evolution. Dev Comp Immunol 27:673–684

Tamura K et al (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599

Thompson JD et al (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Torres M, Sanchez L (1992) The segmentation gene runt is needed to activate *Sex-lethal*, a gene that controls sex determination and dosage compensation in *Drosophila*. Genet Res 59:189–198

Tribolium Genome Sequencing Consortium (2008) The genome of the model beetle and pest *Tribolium castaneum*. Nature 452:949–955

Waterhouse RM et al (2008) The *Aedes aegypti* genome: a comparative perspective. Insect Mol Biol 17:1–8

Zdobnov EM, Bork P (2007) Quantification of insect genome divergence. Trends Genet 23:16–20

Zdobnov EM et al (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. Science 298:149–159