Adaptive evolution of Hox-gene homeodomains after cluster duplications
Vincent J Lynch*1, Jutta J Roth1,2,3 and Günter P Wagner1

Address: 1Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, Connecticut 06551, USA, 2Department of Genetics and General Biology, University of Salzburg, Hellbrunnerstrasse 34, 5020 Salzburg, Austria and 3National Institute for Medical Research, Division of Developmental Biology, The Ridgeway, London, NW7 1AA, UK

Email: Vincent J Lynch* - vincent.j.lynch@yale.edu; Jutta J Roth - jroth@nimr.mrc.ac.uk; Günter P Wagner - gunter.wagner@yale.edu

* Corresponding author

Abstract

Background: Hox genes code for homeodomain-containing transcription factors that function in cell fate determination and embryonic development. Hox genes are arranged in clusters with up to 14 genes. This archetypical chordate cluster has duplicated several times in vertebrates, once at the origin of vertebrates and once at the origin of gnathostomes, an additional duplication event is associated with the origin of teleosts and the agnaths, suggesting that duplicated Hox cluster genes are involved in the genetic mechanisms behind the diversification of vertebrate body plans, and the origin of morphological novelties. Preservation of duplicate genes is promoted by functional divergence of paralogs, either by subfunction partitioning among paralogs or the acquisition of a novel function by one paralog. But for Hox genes the mechanisms of paralog divergence is unknown, leaving open the role of Hox gene duplication in morphological evolution.

Results: Here, we use several complementary methods, including branch-specific $d_\text{s}/d_\text{S}$ ratio tests, branch-site $d_\text{s}/d_\text{S}$ ratio tests, clade level amino acid conservation/variation patterns, and relative rate ratio tests, to show that the homeodomain of Hox genes was under positive Darwinian selection after cluster duplications.

Conclusion: Our results suggest that positive selection acted on the homeodomain immediately after Hox clusters duplications. The location of sites under positive selection in the homeodomain suggests that they are involved in protein-protein interactions. These results further suggest that adaptive evolution actively contributed to Hox-gene homeodomain functions.

Background

The homeobox codes for a highly conserved 60 amino acid DNA-binding motif (the homeodomain) found in transcription factors [1]. One class of homeobox-containing transcription factor genes are the Hox genes, which are homologous to the genes in the Drosophila homeotic (HOM) gene cluster, that specify cell fate during embryonic development [11] and have derived functions in other tissues [2]. Multiple Hox genes located in tightly linked clusters have been identified in all animal phyla examined, with the archetypical chordate cluster having 14 genes (Hox1–Hox14) [3]. The number of Hox clusters has increased several times in vertebrate evolution: the cluster duplicated twice in early vertebrates leading to four clusters (HoxA-D) with 42 genes [4,5] and additional cluster duplications in teleost fish led to 7–8 clusters with 45–47
genes [6,7]. Independent duplications have also occurred in the jawless vertebrates hagfish [8] and lamprey [9].

Models of duplicate gene preservation predict functional differentiation of paralogs based on protein sequence or regulatory divergence [10,11]. Although numerous models of duplicate gene divergence have been proposed, four different mechanisms of functional divergence are likely to explain preservation of duplicate Hox genes: acquisition of novel functions by one paralog (neo-functionalization) [12], passive erosion of functional redundancy due to complementary degenerative mutations, (sub-functionalization) [11], models that predict the accumulation of neutral mutations, which later acquire functional constraints because the environment or genetic background changes (the Dykhuizen-Hartl effect) [13] or divergent adaptive selection of both paralogs (adaptive diversification) [14]. This list has recently been expanded by the introduction of the subneofunctionalization [15] and the adaptive radiation [16] models that predict rapid subfunctionalization after duplication followed by a prolonged period of neofunctionalization and adaptive divergence of duplicate genes in a process analogous to species radiations, respectively. Here, we are interested in testing whether positive selection acted immediately after cluster duplications to promote functional divergence and identify which mechanisms discussed above most adequately explain the preservation of Hox duplicates in vertebrates.

How paralogous Hox genes have been retained is not known, although evidence suggestive of positive selection after cluster duplication has been identified in Hox7 [17], Hox5 and Hox6 [18] paralogs. In these studies, however, it is not clear whether directional selection was responsible for the maintenance of the duplicated genes or other mechanisms promoted the maintenance of duplicates [19]. In addition, evidence for positive selection immediately after Hox cluster duplications has recently been identified in teleost fish for HoxA-11 and HoxB-5 [20]. These data suggest that, in the evolution of ray-finned fishes, some duplicate Hox genes have been preserved by functional differentiation through the action of positive Darwinian selection immediately following the gene duplication. This suggests that Hox genes may have also experienced adaptive evolution following the cluster duplications earlier in vertebrate evolution.

Hox cluster duplication and gene diversification has been proposed to be one of the genetic mechanisms behind the diversification of vertebrates and body plans and the origin of morphological novelties [21-23]. This association, however, is difficult to reconcile with the perceived degree of sequence conservation between the homeodomains of Hox genes and the numerous examples of functional equivalence of Hox/Hom genes from strikingly divergent organisms [24-28]. Mouse HoxA-5, for example, is able to activate the same target genes as its Drosophila homolog, Sex combs reduced (Src), in axis determination indicating strong conservation of function over 500 to 600 million years [29], but counter examples also exist, showing functional non-equivalence of Ubx orthologs from fairy shrimp, velvet worm and Drosophila [30] and non-equivalence of homeodomains from HoxA-4, HoxA-10, HoxA-11 and HoxA-13 paralogs from mouse [31,32].

In this paper we investigate the sequence divergence in homeodomains from the four gnathostome Hox clusters, including genes from basal vertebrates and sarcopterygians like shark and coelacanth, respectively. This is the first study of homeodomain divergence with extensive taxon sampling allowing us to identify the relative phylogenetic age of substitution events in vertebrate phylogeny. We use three different, but complementary, approaches to test for functional divergence among paralogs: comparison of patterns of amino acid sequence conservation/variation among paralog clades, \( d_{s}/d_{r} \) ratio tests to detect directional selection and identify positive sites, and comparison of clade level polymorphisms/divergence rates. Our results indicate that after cluster duplication positive Darwinian selection acted on the homeodomain of Hox proteins prior to the divergence of the modern gnathostome and bony fish lineages. We find amino acid substitutions at sites that are not involved in structural constraints and are located on the molecular surface where they are available for protein-protein interactions were targets of positive selection. We suggest that the action of positive selection at a subset of sites not constrained by ancestral (plesiomorphic) functions after cluster duplications led to the emergence of novel protein interactions while maintaining ancestral ones. This model can help reconcile the role of Hox genes in morphological diversification and innovation with their extreme sequence conservation.

**Results and discussion**

**Functional divergence of paralog-group homeodomains**

We compiled a database of Hox genes with 4–5 species for each gene (155 sequences in total) and compared conserved and variable sites between paralog group members to identify if there are characteristic residues that distinguish which cluster a paralog belongs to (for example, see Figure 1). This analysis identified many sites that are conserved among species but variable between genes in the same paralog group (‘cluster-specific’ residues; Figure 2). Although the homeodomain is a highly conserved motif, it is not invariant; in fact only 17 residues are absolutely conserved between all vertebrate Hox genes in our alignment, suggesting that variable sites could be functionally divergent. Many of these variable sites have been previ-
ously shown to be 'characteristic residues' that distinguish paralog groups from each other and have been suggested to be engaged in protein-protein interactions [33].

To test if 'cluster-specific' residues are functionally divergent we estimated the coefficient of functional divergence (θ), which measures the difference in evolutionary rate at amino acid sites between gene clusters. Rejection of the null hypothesis (θ = 0) is strong evidence for altered functional constraints after gene duplication (or speciation) [34]. We found significant evidence of type-I functional divergence for comparisons between HoxA, HoxB, and HoxD clusters (θI = 0.24–0.37, p < 0.05; Figure 3) under the ((AD)(BC)) topology and under the (B(A(CD))) topology (θI = 0.271–0.396, p < 0.05; Figure 3) at sites that differentiate paralog groups. We also found evidence of functional divergence of the HoxB cluster from the protoHoxACD cluster after the initial duplication event predicted under the (B(A(CD))) model (θI = 0.221 ± 0.07, p < 0.05; Figure 3). Results were not significant for comparisons between HoxC and any other cluster (θI = 0.001–0.029) under either duplication model. Comparisons between HoxB and HoxA to protoBC under the (B(A(CD))) model and protoAD to protoBC under the ((AD)(BC)) model, θI((AD)(BC)) = 0.138–134 and θI[(AD)(BC)] = 0.001, respectively, were not significant. Type-I sites are defined as those with an amino acid that is conserved in one clade but variable in the sister clade, implying that the site is under structural/functional constraints in the first clade that is absent in the variable clade [34]. Type-I sites are located in the amino- and carboxy terminal ends of the homeodomain (outside of the 3 helices and in regions not predicted to be well structured) and in loop connecting helix 2 and 3.

Recently, a method has been developed to test for type-II functional divergence [35]. Type-II sites are those that are highly conserved in both clades but are fixed for amino acids with different biochemical properties between sister clades, implying these residues are responsible for functional differences between these groups. Although all paralog groups had at least one site with evidence of type-II divergence, all of which were radical amino acid substitutions (defined as a change in polarity or charge, but not size) of surface residues, the θII values are extremely small (θII = 0.001–0.062), highlighting the conservation between homeodomains from different clusters. These results are not unexpected given that this method calculates θ across all sites in an alignment and thus effectively averages site-wise θ values. With only ~4% of sites/cluster showing a pattern of type-II divergence in our concatenated alignment, it is not likely that the ~46 possible type-II sites have θII values high enough to compensate for the extremely low θII values of the over 900 sites with θII effectively equal to zero.

Accelerated evolution of homeodomains after cluster duplication

Several models of molecular evolution have been proposed to account for the preservation of duplicate genes. Although the details of each model can vary (see introduction), they differ in their predictions regarding the pattern of sequence evolution following gene duplication. The neo-functionalization and divergent selection models

---

**Figure 1**

An Example of Functional Divergence in Hox6 Paralogs. The phylogeny of the genes is shown on the left with the location of the cluster duplication indicated with an open circle and speciation events indicated with a closed circle. Post-cluster-duplication branches (PCD) and post speciation branches (PS) are highlighted blue and gray, respectively. These branch types were used in the calculation of θPCD and θPS in two ratio analyses for all paralog groups. The amino acid sequence of Hox6 genes from human (Hsa), chicken (Gga), frog (Xtr), coelacanth (Lme) and shark (Hfr) are shown on the left with divergent sites highlighted in red. Below the alignment sites are identified as type-I (1), type-II (2) or both (3). Amino acid substitutions are classified as conservative (C), moderate (M) or radical (R).
predict that the nonsynonymous substitution rate will be increased following gene duplication because of positive Darwinian selection in the gene acquiring the new function, while the Dykhuizen-Hartl and DDC models predict an increase in the substitution rate because of relaxed purifying selection. It is possible to distinguish between these models by comparing nonsynonymous ($d_{N}$) to synonymous ($d_{S}$) substitution rate ($d_{N}/d_{S} = \omega$) with $\omega = 1$, <1, and >1 indicating neutral evolution, purifying selection and directional selection.

Unlike the functional divergence methods developed by Gu [36,37], estimating selection using the $d_{N}/d_{S}$ ratio is, by definition, dependent on the degree of divergence of the sequences under study. Thus, short sequences with a high degree of amino acid conservation but substantial synonymous site divergence may not contain enough signal to reliably obtain estimates of $d_{N}$ and $d_{S}$. We assessed whether homeodomain sequences contained sufficient information for reliable rate estimates by examining the tree length statistic $S$, the number of nucleotide substitutions per codon. For individual paralog groups $S$ range from 6.3–13.3 (average = 10.00), with tree length $d_{N}$ averaging 1.2 substitutions per nonsynonymous site and tree length $d_{S}$ averaging 18 substitutions per synonymous site along the tree. Interestingly, simulation studies [38] have shown that at levels of sequence divergence similar to our datasets, use of the $\chi^2$ made the likelihood ratio test statistic (LRT) extremely conservative such that the type-I error rate is very small. Similarly, the power of the LRT to reject the null hypothesis even when it is false (type-II error) was found to be conservative even at medium to high levels of sequence divergence [38]. The power of the LRT increases as the number of sequences increases such that at 17 taxa
Figure 3
Relative Rate Ratio Tree and Coefficient of Functional Divergence. Numbers of replacement and silent, invariant and variant substitutions are shown above branches (RI/RV, SI/SV) for lineage with significant results indicating adaptive evolution. Coefficients of functional divergence ($\theta$) estimated from DIVERGE are shown on the right; $\theta$ is shown on the internal branch separating HoxB from protoHoxACD for the divergence between HoxB and protoHoxACD. Results are shown for both the ((AD)(BC)) (A.) and (B(A(CD))) (B) topologies. *, p < 0.05; **, p < 0.001.
power is nearly 100% [38], suggesting that our inclusion of at least 8–12 sequences (depending on the paralog group) helped alleviate loss of power from short conserved sequences. The simulation study results indicated that the optimal sequence divergence depends on the dataset and appears to be within the medium-to-high range [38]. Our data indicate that results based on estimates of \( d_{sy} \) and \( d_{s} \) from this homeodomain dataset are reliable, if conservative.

To estimate the strength and kind of selection acting on Hox gene homeodomains, we used maximum likelihood methods to estimate the nonsynonymous (\( d_{sy} \)) to synonymous (\( d_{s} \)) substitution rate ratio [39,40]. The one ratio model is the simplest and provides a measure of the average strength and direction of selection acting on the gene throughout its history and can test if there was an increase in the rate of evolution after Hox cluster duplications. As expected, the \( d_{sy}/d_{s} \) ratio for the homeodomains of all paralog groups is much less than 1 (0.0033–0.0359) highlighting the dominant role purifying selection plays on Hox gene evolution. To test if there was an increase in the nonsynonymous substitution rate following Hox cluster duplication we used a two-ratios model that estimated separate \( \omega \)'s for post cluster duplication (\( \omega_{PCD} \)) and post speciation (\( \omega_{PS} \)) branches. Post cluster duplication branches evolved significantly faster (3–27x) than post speciation branches for 10 of 13 paralog groups; the remaining 3 paralog groups had \( \omega_{PCD} > \omega_{PS} \) but the results were not significant (Table 1). A more complex model that allowed each post duplication lineage to have separate \( d_{sy}/d_{s} \) ratios from each other (the paralog 6 group for example: \( \omega_{PCD-B6} \) \( \omega_{PCD-C6} \) and \( \omega_{PCD-A6} \)) and post speciation (\( \omega_{PS} \)) branches was not better than the simple two-ratio model indicating that paralogs experienced similar selective forces after cluster duplication. These results are consistent with previous data from Hox5, Hox6 and Hox7 and indicate there was a period of rapid evolution of the homeodomain after Hox cluster duplication that could have been the result of either positive Darwinian selection or relaxed purifying selection.

**Adaptive evolution of homeodomains after cluster duplication: Relative rate ratio tests**

Although positive selection at the molecular level is most often tested using the \( d_{sy}/d_{s} \) ratio, this method has several inherent limitations. The most problematic of which is when positive selection is acting at a limited number of sites while the majority are under strong purifying selection. Under these conditions \( d_{sy} \) will never become larger than \( d_{s} \) and the signal for positive selection will be masked. In addition, when there is a large amount of sequence divergence between two nodes in a tree (site saturation) the accuracy of \( d_{sy} \) and to a lesser extent \( d_{s} \) is greatly reduced. These two limitations of the \( d_{sy}/d_{s} \) ratio to detect positive selection are particularly important for studying selective forces after Hox cluster duplications since very few sites (less than 15%) changed after duplication and the duplication events are relative ancient (about 560 MYA; ref), leading to substantial synonymous site divergence. Thus, even though we found evidence of accelerated rates of sequence evolution post cluster duplication, it is unlikely that the \( d_{sy}/d_{s} \) ratio tests used above would be able to detect positive selection (\( \omega > 1 \)).

One complementary method that has been developed to compensate for some of limitations of the \( d_{sy}/d_{s} \) ratio is the relative rate ratio test of Creevey and McInerney [41], which is an extension of the contingency test of neutrality proposed by Templeton [42] and McDonald and Kreitman [43]. Briefly, this method reconstructs ancestral sequences for each node in a phylogenetic tree using parsimony and identifies all substitutions that result in non-synonymous and synonymous changes for each node. Substitutions are classified as replacement invariable (RI, i.e. nonsynonymous substitutions that are not substituted again in descend lineages), replacement variable (RV, i.e. nonsynonymous substitutions that are substituted again in descend lineages), silent invariable (SI, i.e. synonymous substitutions that are not substituted again in descend lineages) and silent variable (SV, i.e. synonymous substitutions that are substituted again in descend lineages).

Under neutral evolution the ratio of RI/RV will not be significantly different from SI/SV. Similarly, a period of relaxed purifying selection may increase RI/RV relative to SI/SV, but RI/RV will never be significantly greater than the neutral expectation given by SI/SV since the rate of replacement substitution can only exceed the rate of silent (neutral) substitution under positive selection. During an episode of positive selection, advantageous substitutions will become fixed in a lineage and remain invariant in descend lineages, elevating the ratio of RI/RV relative to the neutral expectation given by SI/SV. Thus, when lineages are identified with a significantly greater RI/RV than SI/SV positive selection is indicated.

Using the relative rate ratio test to examine selective forces after cluster duplications identified that post-duplication lineages under the ((AD)(BC)) and the (B(A,CD)) models had significantly larger RI/RV than SI/SV (Figure 3 and Tables 2 and 3), indicating these duplication events were followed by adaptive evolution and supporting the results obtained with the \( d_{sy}/d_{s} \) ratio tests and further suggesting that the increase in rates identified from the \( d_{sy}/d_{s} \) ratio were due to positive selection.
Adaptive evolution of homeodomains after cluster duplication: dN/dS

The lineage-specific dN/dS model utilized above has been extended to account for variable dN/dS between sites and can detect positive selection at specific sites in specific lineages under appropriate conditions [44,45]. These branch-site models are ideal for detecting short episodes of positive selection that acted on a few sites while the majority of sites in the protein remained under purifying selection, as is likely to have occurred in the homeodomain after Hox cluster duplication. Applying branch-site models to post cluster duplication (ωPCD) branches identified sites under positive selection after cluster duplication (Figure 3) in paralog groups 1–6, 9 and 13 (Table 4). Positive Sites were identified with posterior probabilities (PP) greater than 0.90 using the both the liberal Neive Empircal Bayes (NEB) and the more conservative Bayes Empirical Bayes (BEB) methods implemented in PAML3.15, although only the result of the BEB method is shown. In addition, two genes in the Hox3, 5, 6, and 13 paralog groups have evidence of positive selection, but the results are not statistically significant. The sites identified under positive selection are the same as those that show evidence of type-II functional divergence.

Table 1: Likelihood parameter estimates under the lineage-specific models.

| Model    | ϵ    | PS-dN | PS-dS | PD-dN | PD-dS | ω0   | ωPCD | Sig.   |
|----------|------|-------|-------|-------|-------|------|------|--------|
| Hox1     |      |       |       |       |       |      |      |        |
| One ratio| -707.88 | 0.0046 | 0.0025 | 0.5130 | 0.0008 | 0.0096 | 0.1250 | P < 0.05 |
| Two ratio| -705.14 |       |       |       |       |      |      |        |
| Hox2     |      |       |       |       |       |      |      |        |
| One ratio| -800.02 | 0.0037 | 0.6313 | 0.0090 | 0.6700 | 0.0058 | 0.1334 | 0.05    |
| Two ratio| -798.08 |       |       |       |       |      |      |        |
| Hox3     |      |       |       |       |       |      |      |        |
| One ratio| -989.99 | 0.0042 | 1.1187 | 0.0003 | 0.0718 | 0.0037 |       |        |
| Two ratio| -984.72 | 0.0034 | 1.3424 | 0.0050 | 0       | 0.0025 |       |        |
| Hox4     |      |       |       |       |       |      |      |        |
| One ratio| -1573.85 | 0.0046 | 0.4348 | 0.0081 | 0.7736 | 0.0105 |       |        |
| Two ratio| -1570.91 | 0.0034 | 0.4518 | 0.0151 | 0.5811 | 0.0076 | 0.0259 | P < 0.05 |
| Hox5     |      |       |       |       |       |      |      |        |
| One ratio| -1015.97 | 0.0028 | 0.8522 | 0.0266 | 8.1561 | 0.0033 |       |        |
| Two ratio| -1014.86 | 0.0025 | 0.9172 | 0.0279 | 1.4231 | 0.0027 | 0.0196 | n.s.   |
| Hox6     |      |       |       |       |       |      |      |        |
| One ratio| -1309.19 | 0.0110 | 0.7089 | 0.0242 | 1.5604 | 0.0155 |       |        |
| Two ratio| -1297.90 | 0.0096 | 0.7323 | 0.0337 | 0.5095 | 0.0131 |       |        |
| Hox7     |      |       |       |       |       |      |      |        |
| One ratio| -701.50  | 0.0032 | 0.8238 | 0       | 0      | 0.0039 |       |        |
| Two ratio| -696.90  | 0.0025 | 1.0950 | 0.0037 | 0       | 0.0023 | (0.5/0) | P < 0.01 |
| Hox8     |      |       |       |       |       |      |      |        |
| One ratio| -1092.36 | 0.0055 | 0.3889 | 0.0129 | 0.9167 | 0.0141 |       |        |
| Two ratio| -1089.88 | 0.0041 | 0.4206 | 0.0204 | 0.4981 | 0.0098 | 0.0409 | P < 0.05 |
| Hox9     |      |       |       |       |       |      |      |        |
| One ratio| -1389.70 | 0.0042 | 0.8239 | 0.0112 | 2.1879 | 0.0101 |       |        |
| Two ratio| -1377.10 | 0.0004 | 0.6179 | 0.0110 | 0.0950 | 0.0054 | 0.1156 | P < 0.01 |
| Hox10    |      |       |       |       |       |      |      |        |
| One ratio| -1790.54 | 0.0118 | 0.3296 | 0.0204 | 0.5690 | 0.0359 |       |        |
| Two ratio| -1783.60 | 0.0124 | 0.5521 | 0.0225 | 0       | 0.0224 | (4.5/0) | P < 0.01 |
| Hox11    |      |       |       |       |       |      |      |        |
| One ratio| -1210.27 | 0.0054 | 0.5091 | 0.0132 | 1.2578 | 0.0105 |       |        |
| Two ratio| -1206.53 | 0.0037 | 0.5339 | 0.0244 | 0.8452 | 0.0070 | 0.0288 | P < 0.01 |
| Hox12    |      |       |       |       |       |      |      |        |
| One ratio| -1092.80 | 0.0141 | 0.6274 | 0.0545 | 2.3793 | 0.0229 |       |        |
| Two ratio| -1092.72 | 0.0141 | 0.6378 | 0.0563 | 1.8156 | 0.0222 | 0.031   | n.s.   |
| Hox13    |      |       |       |       |       |      |      |        |
| One ratio| -1745.34 | 0.0171 | 0.7425 | 0.0419 | 1.8193 | 0.0230 |       |        |
| Two ratio| -1745.61 | 0.0160 | 0.7516 | 0.0496 | 0.9426 | 0.0212 | 0.0526 | 0.063   |

Parameters that indicate rate accelerations are in bold. ℵ, likelihood of the model. Sig., significance of the model.
and map onto the molecular surface of the homeo-
domain, facing away from the DNA and in an orientation
that would facilitate protein-protein interactions (Figure
2).

While no sites under positive selection were identified in
paralog groups 7, 8 and 10–12, a class of sites in each was
identified with $\omega = 1$ (Table 4). Given that the ability of
likelihood models to detect sites with $\omega > 1$ is an
extremely difficult computational problem, it is possible
that these sites actually experienced positive selection, but
that the models are not able to identify $\omega > 1$. An equally
likely explanation that does not invoke positive selection
is that the $\omega = 1$ is an accurate estimate for the rate at this
sites, and is actually indicative of relaxed functional con-
straints after duplication, that the sites have not been sub-
stituted again indicates they under strong purifying
selection in post-speciation lineages, supporting a Dykhu-
izen-Hartl mechanism for their evolution.

The structural basis of homeodomain evolution
To gain a better understanding of how functional con-
straints on the homeodomain relate to sequence diver-
gence, we generated a sequence logo [46,47] from the
multiple sequence alignment of Hox-gene homeodo-
mains and mapped the location of sites under positive
Table 2: Results of the Creevey-McInerney test unde the ((AD)(BC)) topology.

| Branch | RI | RV | SI | SV | G-Value | G-test | Sig.  |
|--------|----|----|----|----|---------|--------|------|
| branch 0 | 17 | 57 | 134 | 370 | G = 0.443499 | Gtest:0.900000 | P > 0.500000 |
| branch 1 | 31 | 105 | 187 | 583 | G = 0.140809 | Gtest:0.900000 | P > 0.500000 |
| branch 2 | 72 | 175 | 120 | 798 | G = 7.496351 | Gtest:0.005000 | P > 0.000000 |
| branch 3 | 129 | 270 | 224 | 1008 | G = 33.356056 | Gtest:0.005000 | P > 0.000000 |
| branch 4 | 28 | 37 | 75 | 230 | G = 8.448683 | Gtest:0.005000 | P > 0.000000 |
| branch 5 | 44 | 84 | 128 | 391 | G = 4.740880 | Gtest:0.050000 | P > 0.025000 |
| branch 6 | 82 | 168 | 153 | 621 | G = 17.139877 | Gtest:0.005000 | P > 0.000000 |
| branch 7 | 227 | 486 | 379 | 1704 | G = 54.900509 | Gtest:0.005000 | P > 0.000000 |
| branch 8 | 9 | 86 | 57 | 334 | G = 1.802721 | Gtest:0.200000 | P > 0.100000 |
| branch 9 | 35 | 137 | 102 | 470 | G = 0.542717 | Gtest:0.500000 | P > 0.200000 |
| branch 10 | 96 | 197 | 118 | 642 | G = 36.213978 | Gtest:0.005000 | P > 0.000000 |
| branch 11 | 2 | 44 | 57 | 289 | G = 5.863089 | Gtest:0.025000 | P > 0.010000 |
| branch 12 | 12 | 57 | 85 | 412 | G = 0.003849 | Gtest:0.990000 | P > 0.950000 |
| branch 13 | 59 | 93 | 102 | 562 | G = 37.628811 | Gtest:0.005000 | P > 0.000000 |
| branch 14 | 172 | 337 | 222 | 1259 | G = 77.613304 | Gtest:0.005000 | P > 0.000000 |
| branch 15 | 446 | 117 | 602 | 3266 | G = 101.06102 | Gtest:0.005000 | P > 0.000000 |

RI, replacement invariant. RV, replacement variant. SI, synonymous invariant. SV, synonymous variant. The G-value and results of the G-test are shown along with the significant of the results.

Table 3: Results of the Creevey-McInerney test unde the ((AD)(BC)) topology.

| Branch | RI | RV | SI | SV | G-Value | G-test | Sig.  |
|--------|----|----|----|----|---------|--------|------|
| branch 0 | 10 | 83 | 60 | 333 | G = 1.304401 | Gtest:0.500000 | p-value > 0.200000 |
| branch 1 | 24 | 136 | 110 | 465 | G = 1.477014 | Gtest:0.500000 | p-value > 0.200000 |
| branch 2 | 63 | 188 | 130 | 649 | G = 8.360223 | Gtest:0.005000 | p-value > 0.000000 |
| branch 3 | 20 | 57 | 132 | 369 | G = 0.005322 | Gtest:0.950000 | p-value > 0.900000 |
| branch 4 | 39 | 110 | 183 | 586 | G = 0.376969 | Gtest:0.900000 | p-value > 0.500000 |
| branch 5 | 57 | 172 | 206 | 797 | G = 2.035316 | Gtest:0.200000 | p-value > 0.100000 |
| branch 6 | 103 | 232 | 220 | 1017 | G = 25.201559 | Gtest:0.005000 | p-value > 0.000000 |
| branch 7 | 9 | 42 | 60 | 286 | G = 0.003351 | Gtest:0.990000 | p-value > 0.950000 |
| branch 8 | 12 | 71 | 94 | 399 | G = 1.045448 | Gtest:0.500000 | p-value > 0.200000 |
| branch 9 | 48 | 99 | 111 | 547 | G = 17.036821 | Gtest:0.005000 | p-value > 0.000000 |
| branch 10 | 29 | 37 | 79 | 229 | G = 8.251780 | Gtest:0.005000 | p-value > 0.000000 |
| branch 11 | 43 | 87 | 139 | 386 | G = 2.192354 | Gtest:0.200000 | p-value > 0.100000 |
| branch 12 | 64 | 150 | 158 | 608 | G = 7.819194 | Gtest:0.005000 | p-value > 0.000000 |
| branch 13 | 134 | 305 | 275 | 1231 | G = 28.853481 | Gtest:0.005000 | p-value > 0.000000 |
| branch 14 | 263 | 608 | 495 | 2337 | G = 61.973667 | Gtest:0.005000 | p-value > 0.000000 |
| branch 15 | 373 | 1093 | 626 | 3300 | G = 60.691624 | Gtest:0.005000 | p-value > 0.000000 |

RI, replacement invariant. RV, replacement variant. SI, synonymous invariant. SV, synonymous variant. The G-value and results of the G-test are shown along with the significant of the results.
selection and residues with known functions onto the logo and the crystal structure of the homeodomain bound to DNA (Figure 4). Adaptive/functionally divergent sites are grouped into three discrete regions of the homeodomain: the extreme amino and carboxy terminal arms just outside of the homeodomain proper and in the C-terminal end of helix-2 extending into the loop connecting helix-2 and helix-3.

The repressor domain, where the majority of protein interactions have been found, and helix-3 are free from positive sites, likely reflecting conserved functions shared by all Hox genes. Several proteins have been shown to bind in the repressor domain including the CREB binding protein (CBP) [48], high mobility group protein 1 (HMG1) [49], members of the Maf family of basic-leucine zipper (bZip) activators [50], and geminin [51].

| Model | Model | Parameters | Positive Sites | Sig. |
|-------|-------|------------|----------------|------|
| Hox1  | M1a   | -707.37    | $p_0 = 0.982, p_1 = 0.018; \omega_0 = 0.0017, \omega_1 = 1$ | Not Allowed |
|       | MA    | -697.04    | $P_{+1} = 0.935, p_2 = 0.065; \omega_{01} = 0.0034/1, \omega_2 = 999$ | 3 (PP > 0.99) |
| Hox1  | MA    | -697.04    | $P_{+1} = 0.935, p_2 = 0.065; \omega_{01} = 0.0034/1, \omega_2 = 999$ | P << 0.001 |
| Hox2  | M1a   | -800.91    | $p_0 = 0.985, p_1 = 0.015; \omega_0 = 0.0059, \omega_1 = 1$ | Not Allowed |
|       | MA    | -794.72    | $P_{+1} = 0.981, p_2 = 0.019; \omega_{01} = 0.0062/1, \omega_2 = 17.98$ | 1 (PP > 0.95) |
| Hox3  | M1a   | -1684.44   | $p_2 = 0.979, p_1 = 0.021; \omega_0 = 0.0223, \omega_1 = 1$ | Not Allowed |
|       | MA    | -1647.20   | $P_{+1} = 0.833, p_2 = 0.167; \omega_{01} = 0.0116/1, \omega_2 = 170.6$ | 12 (PP > 0.95) |
| Hox4  | M1a   | -1547.08   | $p_0 = 0.987, p_1 = 0.013; \omega_0 = 0.0057, \omega_1 = 1$ | Not Allowed |
|       | MA    | -1538.64   | $P_{+1} = 0.987, p_2 = 0.013; \omega_{01} = 0.0046/1, \omega_2 = 3.70$ | 1 (PP > 0.99) |

Parameters that indicate positive selection are in bold. $\ell$, likelihood of the model. Sig., significance of the model. Positive sites were identified using the Bayes empirical Bayes (BEB) method.
region also overlaps with the Sp1 transactivation region [52]. In addition to characterized protein-protein interactions, the repressor domain also contains the majority of ‘characteristic-residues’ that distinguish cognate groups from each other indicating the majority of sites in this region were already under functional constraints after the tandem duplications which created the Hox gene cluster and were not available to be targets of adaptive selection after cluster duplication. Interestingly, the first 3 sites of the geminin-binding region were under directional selection in different paralog groups, including radical amino acid substitutions, suggesting selection to modulate geminin binding between paralog group members.

Mapping functionally divergent amino acid sites and sites under positive selection in and around helix-2 onto the logo and crystal structure shows that purifying selection has acted to preserve hydrophobic/aliphatic residues critical for the nuclear exportation signal [53] and positive selection has acted exclusively on sites that occur at the molecular surface. These sites form a small cluster at the posterior end of helix-2 in a prime location for protein-
protein interactions. Beyond the ultra-conserved helix-3, which also contains the nuclear localization signal [54], sites under positive selection have been identified in an unstructured connecting loop leading to additional structures in the carboxy terminus. The amino-terminal arm of the homeodomain, which confers functional specificity on Hox proteins, contains the majority of sites under positive selection suggesting that selection has acted to modify functional specificity between paralogs. This region appears to be unstructured and is a prime target for protein-protein interaction sites. This pattern of purifying and positive selection suggests that after Hox cluster duplications, selection acted on protein-protein interaction sites in such a way that ancestral functions were maintained while the acquisition of novel protein interaction partners driven was driven by selection on non-constrained amino acids. These derived interactions could be those responsible for novel Hox gene functions in vertebrates.

Conclusion
The homeodomain serves multiple functions in addition to DNA-binding, including containing nuclear localization and export signals, transcriptional activation and repression domains and other protein-protein interaction sites [33,54]. These functions combine to impose severe limitations on the degree of sequence divergence that can be accommodated by the homeodomain of Hox genes. Even with these constraints, however, the relatively small set of amino acids that were free to diverge after cluster duplication were subject to positive selection. Although the Hox cluster duplications are relatively ancient (450 MYA), complicating the detection of positive selection, we find congruence between multiple methods a strong indicating that our results are reliable. These results support an important role for the action of positive Darwinian selection in the divergence of Hox genes after cluster duplications, particularly at sites that distinguish paralog groups (‘cluster-specific’ residues).

Nearly all ‘cluster-specific’ residues map onto the molecular surface of the homeodomain, similar to the paralog group specific sites [33], suggesting changes in amino acid properties could influence interaction of the homeodomain with other proteins. Cofactor associations are important for Hox proteins and most other transcription factor functions; these protein-protein interactions occur at the molecular surface through the formation of hydrophobic and ionic bonds and other intermolecular interactions such as salt bridges and van der Waals forces. Thus, changes in the physicochemical properties of amino acids participating in these bonds could disrupt preexisting interactions and/or lead to new interactions. These changes could provide a selective advantage for maintaining duplicate genes through the origin of novel protein-protein interactions (effectively reducing degeneracy between paralogs) leading to new gene functions.

Methods
The homeodomain of Hox genes was identified from BLAST searches of the nr database at NCBI. At least four members of each gene from diverse taxa were included in the dataset. The sequences were aligned based on the translated amino acid sequences with Se-Al v2.0, alignments were simple given the high degree of sequence conservation within paralog groups. Regions of ambiguous alignment just outside of the homeodomain but within exon 2 were excluded. Most alignments ranged from 70–82 amino acids. The alignment is available from V.J.L. and has been deposited in TREBASE.

We used codon-based maximum likelihood models of coding sequence evolution implemented in CODEML in the PAML package of programs (version 3.15) to test for lineages and amino acid sites under positive selection. Sites were classified as being under positive selection if they were identified from the Bayes Empirical Bayes (BEB) method with a posterior probability of greater than 0.90. The branching order of the Hox cluster duplications is still debated (refs), but our analyses suggest that the most likely topologies are ((AD)(BC)) and (B(A(CD))) (a detailed analysis of Hox cluster duplication history is beyond the scope of this paper and will be presented elsewhere). We used 2 alternate trees to test for selection: ((AD)(BC)) and (B(A(CD))) and found no significant differences between the results of these different topologies. Functional divergence was tested with DIVERGE alpha1.2 (obtained from X. Gu). We also used the relative rate ratio test of Creevey and McInerny [41] implemented in the program CRANN to test for adaptive evolution. Both DIVERGE and CRANN analyses used the 2 alternate topologies discussed above.

Authors’ contributions
VJL designed and carried out the project, and wrote the manuscript with contributions from JJR and GPW. JJR provided information on protein-protein interaction sites and GPW provided biological insights and guidance during the course of this study.

Acknowledgements
The authors would like to thank the members of GPW lab for reading and commenting on earlier version of this manuscript. Financial support for this study was provided by the National Science Foundation (NSF) grant IBN#0321470 to GPW, and intramural funds from Yale University.

References
1. Duboule D: Guidebook to the Homeobox Genes. Oxford, Oxford University Press; 1994.
2. Kobayashi A Behringer, R. R.: Developmental Genetics of the Female Reproductive Tract in Mammals. Nature Reviews Genetics 2003:1225–1237.
22. Wagner GP Amemiya, C.T., Ruddle, F.: Evidence for a Hox14 paralog group in vertebrates. Current Biology 2004, 14:R183-R184.

23. Schugart K K, Ruddle FH: Mammalian homeobox-containing genes: genome organization, structure, expression and evolution. British Journal of Cancer 1988, 9:9-13.

24. Wagner GP Amemiya, C.T., Ruddle, F.: Hox Cluster Duplications and the Opportunity for Evolutionary Novelties. Proceedings of the National Academy of Sciences USA 2000, 107:14603-14606.

25. Amores A, Force, A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH: Zebrafish hox Clusters and Vertebrate Genome Evolution. Science 1998, 282:1711-1714.

26. Amores A, Suzuki T, Yan YL, Pomero J, Singer A, Amemiya C, Postlethwait JH: Developmental Roles of Pufferfish Hox Clusters and Genome Evolution in Ray-Fin Fish. Genome Res 2004, 14:1-10.

27. Studler PF Fried, C., Prohaska, S.J., Bailey, W.Y., Ruddle, F., Farey, J.M., Beumer, C.G.: Evidence for independent hox gene duplications in the hagfish lineage: a PCR-based gene inventory of Eptatretus stoutii. Molecular Phylogenetics and Evolution 2004, 32:686-694.

28. Fried C Prohaska, S.J., Stadler, P.F.: Independent Hox-cluster duplications in lampreys. J Exp Zool B Mol Dev Evol 2003, 299:18-24.

29. Kimura M: The neutral theory of molecular evolution. Cambridge, Cambridge University Press: 1983.

30. Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait J: Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetcs 1999, 151:1531-1545.

31. Goodman M, Moore, G.W., Matsuda, G: Darwinian evolution in the genealogy of haemoglobin. Nature 1975, 253:603-608.

32. Dylkholt D, Hardt DL: SELECTIVE NEUTRALITY OF 6PGD ALLOZYMES IN E. COII AND THE EFFECTS OF GENETIC BACKGROUND. Genetics 1980, 96:801-817.

33. Hughes AL: Adaptive Evolution of Genes and Genomes. New York, Oxford University Press: 2000.

34. He X Zhang, J: Rapid subfunctionalization accompanied by prolonged and substantial nonfunctionalization in duplicate gene evolution. Genetics 2005, 169:1157-1164.

35. Francino MP: An adaptive radiation model for the origin of new gene functions. Nature Genetics 2005, 37:573-577.

36. Dyer MA, Moy a, Marin I: Selection in Coding Regions Determined Hox7 Genes Evolution. Mol Biol Evol 2003, 20:2104-2112.

37. Van de Peer Y Taylor, J.S., Braaich, I., Meyer, A.: The ghost of selection past: Rates of evolution and functional divergence of ananclytically duplicated genes. Journal of Molecular Evolution 2001, 53:636-644.

38. Lynch VJ Roth, J.J., Takahashi, K., Dunn, C., Nonaka, D.F., Stopper, G., and the opportunity for evolutionary novelties. BMC Evolutionary Biology 2006, 6:86

39. Schughart K KC Ruddle FH.: Molecular phylogenetic analysis by codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular biology and evolution 2002, 19:908-917.

40. Yang Z Nielsen, R.: The ghost of selection present: Rates of molecular evolution and the cell-cycle regulator geminin inhibits hox function through direct and interdependent regulatory pathways. Mol Biol Evol 2005, 22:1107-1118.

41. Crooks GE Hooi, G., Chandoni, J.M., Brenner, S.E.: WebLogo: A sequence logo generator. Genome Research 2004, 14:188-190.

42. Schneider TDS R.M.: Sequence Logos: A new way to display consensus sequences. Nucleic Acids Research 1990, 18:6097.

43. Shen W, Chrobok D, Krishnan K, Lawrence Hj, Largman C: HOXB6 Protein Is Bound to CREB-binding Protein and Represses Gliobin Expression in a DNA Binding-dependent, PBX Interaction-independent Process. J Biol Chem 2004, 279:39895-39904.

44. Sakazke J Gerard, M., Favier, B., Potter, S.S., Duboule, D.: Functional equivalence and rescue among group 11 Hox gene products in vertebral patterning. Developmental Biology 2004, 276:925-328.

45. Zappavigna V Falciola, L., Helmer -Citterich, M., Mavilo, F., Bianchi, M.E.: HMGI interacts with HOX proteins and enhances their DNA binding and transcriptional activation. EMBO Journal 1996, 15:4981-4991.

46. Kataoka K Yoshihito-Nakagawa, K., Shioda, S., Nishizawa, M.: A set of Hox proteins interact with the Maf oncprotein to inhibit its DNA binding, transactivation, and transforming activities. Journal of Biological Chemistry 2001, 276:19199-19206.

47. Luo J, Yang X, Takahara, Y., Snoeck, H., Kessel, M.: The cell-cycle regulator geminin inhibits hox function through direct and polycymediated interactions. Nature 2004, 427:749-753.

48. Suzuki M Ueno. N., Kuroiwa, A.: Hox proteins functionally cooperate with the GC box-binding protein system through distinct domains. Journal of Biological Chemistry 2003, 278:30148-30156.

49. Maizel A Benasude, O., Prochiantz, A., Joliot, A.: A short region of its homeodomain is necessary for engrailed nuclear export. Development 1999, 126:3183-3190.
54. Roth JJ Breitenback, M., Wagner, G.P.: Multiple functions contribute to the evolutionary conservation of the homeodomain of Hoxa-11. Journal of Experimental Zoology part B: Molecular and Developmental Evolution in press.