Developmental Biology 383 (2013) 175–185

Contents lists available at ScienceDirect

Developmental Biology

journal homepage: www.elsevier.com/locate/developmentalbiology

Resource

Analysis of central Hox protein types across bilaterian clades: On the diversification of central Hox proteins from an Antennapedia/Hox7-like protein

Stefanie D. Hueber [a,c,*], Jens Rauch [c], Michael A. Djordjevic [a], Helen Gunter [b], Georg F. Weiller [a], Tancred Frickey [a,c]

[a] ARC Centre of Excellence (CILR), Research School of Biology, College of Medicine, Biology and the Environment, RN Robertson Building (Bldg 46) Biology Place, The Australian National University, Acton, Canberra ACT 0200, Australia
[b] Lehrstuhl fuer Zoologie und Evolutionsbiologie, Department of Biology, Universität Konstanz, D-78457 Konstanz, Germany
[c] Applied Bioinformatics, Department of Biology, Universität Konstanz, D-78457 Konstanz, Germany

ARTICLE INFO

Article history:
Received 17 May 2013
Received in revised form
30 August 2013
Accepted 5 September 2013
Available online 18 September 2013

Keywords:
Hox proteins
Molecular function
Sequence similarity
Hox protein types
Antennapedia (Antp)
Hox7

ABSTRACT

Hox proteins are among the most intensively studied transcription factors and represent key factors in establishing morphological differences along the anterior–posterior axis of animals. They are generally regarded as highly conserved in function, a view predominantly based on experiments comparing a few (anterior) Hox proteins. However, the extent to which central or abdominal Hox proteins share conserved functions and sequence signatures remains largely unexplored.

To shed light on the functional divergence of the central Hox proteins, we present an easy to use resource aimed at predicting the functional similarities of central Hox proteins using sequence elements known to be relevant to Hox protein functions. We provide this resource both as a stand-alone download, including all information, as well as via a simplified web-interface that facilitates an accurate and fine-tuned annotation of novel Hox sequences. The method used in the manuscript is, so far, the only published sequence-based method capable of differentiating between the functionally distinct central Hox proteins with near-identical homeodomains (such as the *Drosophila* Antp, Ubx and Abd-A Hox proteins). In this manuscript, a pairwise-sequence-similarity based approach (using the bioinformatics tool CLANS) is used to analyze all available central Hox protein sequences. The results are combined with a large-scale species phylogeny to depict the presence/absence of central Hox sequence-types across the bilaterian lineage. The obtained pattern of distribution of the Hox sequence-types throughout the species tree enables us to infer at which branching point a specific type of central Hox protein was present.

Based on the Hox sequences currently available in public databases, seven sequence-similarity groups could be identified for the central Hox proteins, two of which have never been described before (Echi/Hemi7 and Echi/Hemi8). Our work also shows, for the first time, that Antp/Hox7-like sequences are present throughout all bilaterian clades and that all other central Hox protein groups are specific to sub-lineages in the protostome or deuterostome branches only.

Introduction

Hox proteins play a decisive role in patterning the main body axis of bilaterians and unraveling the relationships between sequence, structure and function of Hox proteins (and their associated effects on the body plan patterning) represents one of the most exciting puzzles in developmental biology (Lewis, 1978). Early on, it was discovered that Hox proteins encoded within a Hox gene cluster differ in their amino acid sequence and protein function (Krumlauf, 1992; Hueber et al., 2007). In contrast, some Hox proteins from different species were identified as remarkably similar in sequence and function (presumably because a shared set of biochemical properties allow them to induce the same morphological structures and/or regulate the same set of downstream genes when expressed under the same conditions (same model organism, expression pattern and time frame)) (McGinnis et al., 1990; Zhao et al., 1993; Lutz et al., 1996).

* Correspondence to: University of Konstanz, Faculty of Biology, Applied Bioinformatics Lab, PO Box 645, D-78457 Konstanz, Germany.
E-mail addresses: stefanie.hueber@uni-konstanz.de (S.D. Hueber), jens.rauch@uni-konstanz.de (J. Rauch), michael.djordjevic@anu.edu.au (M.A. Djordjevic), helen.gunter@uni-konstanz.de (H. Gunter), georg.weiller@anu.edu.au (G.F. Weiller), tancred.frickey@uni-konstanz.de (T. Frickey).

0012-1606/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ydbio.2013.09.009

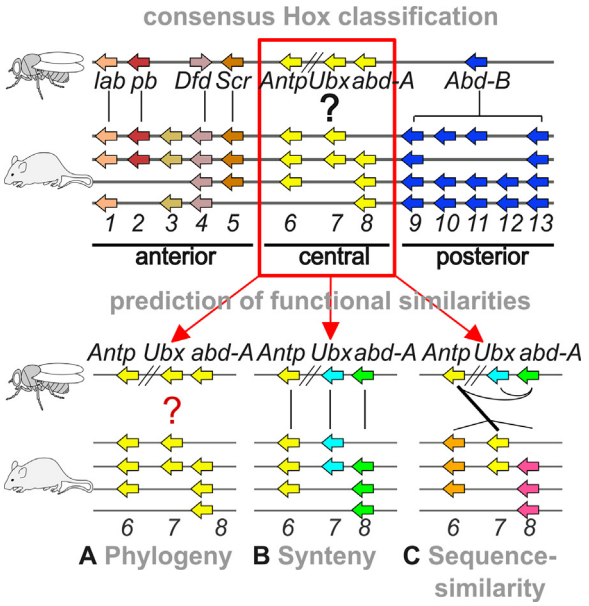

Fig. 1. Classification of central Hox genes. Top: Current consensus classification of the mouse *Mus musculus* and the fruit fly *D. melanogaster* Hox genes according to phylogenetic inference of the encoded protein sequence homeodomain. Relationships between the Antp, Ubx, Abd-A and Hox6, Hox7 and Hox8 proteins cannot be fully resolved. Bottom: Inferring functional similarity of the encoded proteins for the central Hox genes. (A) Phylogeny does not fully resolve the relationships of the encoded central Hox proteins and therefore does not provide any statement regarding functional similarity of these proteins (Gehring et al., 2009; Samadi and Steiner, 2009). (B) The Synteny based classification postulates that the relative location of the Hox genes within the Hox cluster reflects their ancestry and function. It predicts a one-to-one orthology scenario with Antp being orthologous to Hox6, Ubx to Hox7 and Abd-A to Hox8 and, consequently, these protein pairs also as most similar in function (Gehring et al., 2009; Choo and Russell, 2011; Michaut et al., 2011). (C) Pairwise-sequence-similarity identifies the most sequence-similar proteins in the phylogenetically unresolved central group as *Drosophila spp.* Antp and vertebrate Hox7 proteins. The observed sequence-similarity pattern is compatible with a scenario assuming co-orthology of the proteins, but not with the postulated synteny classification. Based on the sequence similarities, the highest functional similarity across these proteins is predicted for Antp and Hox7 proteins (Hueber et al., 2010).

Definition Box

Central Hox proteins: While current classifications tend to term the vertebrate Hox4–8 as "central", it should be noted that the definition of what constitutes "central" or "middle" Hox proteins used to vary wildly (compare (De Rosa et al., 1999) with (Carroll et al., 2005) and (Garcia-Fernàndez, 2005)). Assignment of Hox4 and Hox5 group proteins are largely undisputed, whereby the ancestral deuterostome Hox4 is considered to be orthologous and most similar in function to Deformed (Dfd) and the ancestral deuterostome Hox5 as orthologous and most similar in function to Sex combs reduced (Scr) (see Fig. 1, Top). As the aim of this manuscript is to provide insights into the sequence-relationships of specifically those central Hox proteins for which no clear orthologs can be assigned via phylogenetic methods (see Fig. 1, Bottom), this manuscript uses the original definition of "central" Hox proteins, encompassing only Hox6–8 in vertebrates and Antennapedia (Antp), Ultrabithorax (Ubx) and Abdominal-A (Abd-A) in arthropods (Hueber et al., 2010; De Rosa et al., 1999; Balavoine et al., 2002; Veraksa et al., 2000; Hueber and Lohmann, 2008; Merabet et al., 2009).

Monophyletic: A grouping of leaves/taxa/species in a phylogenetic tree that are derived from a common branch/ancestor (e.g. a grouping of bats and shrews based on e.g. body features).

Polyphyletic: A grouping that combines leaves/taxa/species from different branches in a phylogenetic tree to the exclusion of closer leaves/taxa/relatives (e.g. a grouping of bats and birds based on e.g. their ability to fly).

Due to the ability of certain Hox proteins to induce similar developmental effects across distantly related bilaterian species and the presence of common features in their protein sequences, the last common ancestor of all bilateria is assumed to have already possessed a differentiated set of multiple Hox proteins (numbers vary, e.g. depending on the assumed number of central Hox proteins, (see Hueber et al., 2010)). Current classification schemes are often used as a basis to predict which Hox proteins from different organisms are likely to carry out similar molecular/biochemical functions (Gehring et al., 2009; Mann et al., 2009; Choo and Russell, 2011). For anterior Hox proteins (e.g. vertebrate Hox1–5) these predictions of functional equivalency are consistent (see Fig. 1, Top) and supported by sequence-similarity studies (Hueber et al., 2010), phylogenetic analyses (Graham et al., 1989; De Rosa et al., 1999; Balavoine et al., 2002) as well as experimental evidence (McGinnis et al., 1990; Zhao et al., 1993; Lutz et al., 1996).

For central and abdominal Hox proteins (e.g. vertebrate Hox 6–13), however, the story is not quite as simple (Hueber et al., 2010; Feiner et al., 2011). Despite intensive research efforts and the accumulation of considerable amounts of sequence and experimental data, the characterization of the functional similarities across the central and abdominal Hox proteins has been hampered by technical difficulties (see below). For the central Hox proteins (e.g. vertebrate Hox6–8) (see Definition Box), the main difficulty in predicting the functional similarities of proteins, using phylogenetic inference as a proxy, lies in the near identity of the homeodomain sequences. The DNA-binding homeodomain is the only sequence region unambiguously alignable across all Hox proteins (unambiguous multiple sequence alignments are a requirement for high-quality phylogenetic inference). Unfortunately, the DNA-binding domain of the central Hox proteins does not contain sufficient sequence variation to allow phylogenetic approaches to reproducibly differentiate between central Hox proteins. However, these proteins are known to have distinct developmental functions (e.g. in *Drosophila*) and show considerable variation in their sequences outside the homeodomain. Unfortunately, these additional, potentially functionally informative, sequence regions cannot (or should not) be employed for phylogenetic inference as these regions are not readily alignable across all Hox proteins and may represent features that arose after their multiplication and differentiation of Hox-proteins in the Hox cluster (i.e. represent non-homologous features) (Gehring et al., 2009; Hueber et al., 2010).

Due to the above difficulties in gaining clear predictions of functional similarity across the central Hox proteins based on phylogenetic methods, a different surrogate approach has been widely employed. This approach is based on postulating a syntenic organization of the Hox genes, in which the last common ancestor of protostomes and deuterostomes is assumed to have possessed three differentiated central-type Hox genes that have since remained in the same relative genomic positions. This synteny-based classification is often presented in the literature as diagrams and has been employed as the basis for experimental designs and functional inferences (Gehring et al., 2009; Mann et al., 2009; Choo and Russell, 2011; Michaut et al., 2011; Pick and Heffer, 2012). However, this one-to-one synteny-based assignment of functionally equivalent Hox proteins (i.e. Antp to Hox6, Ubx to Hox7 or Abd-A to Hox8) is neither supported by experimental evidence nor by phylogenetic methods and the protein sequence-similarities contradict these assignments (e.g. Antp is more similar to Hox7 than Hox 6 (Malicki et al., 1990; Hueber et al., 2010)).

One paper has often been cited as showing functional equivalence of Antp and Hox6 proteins and this has been taken as

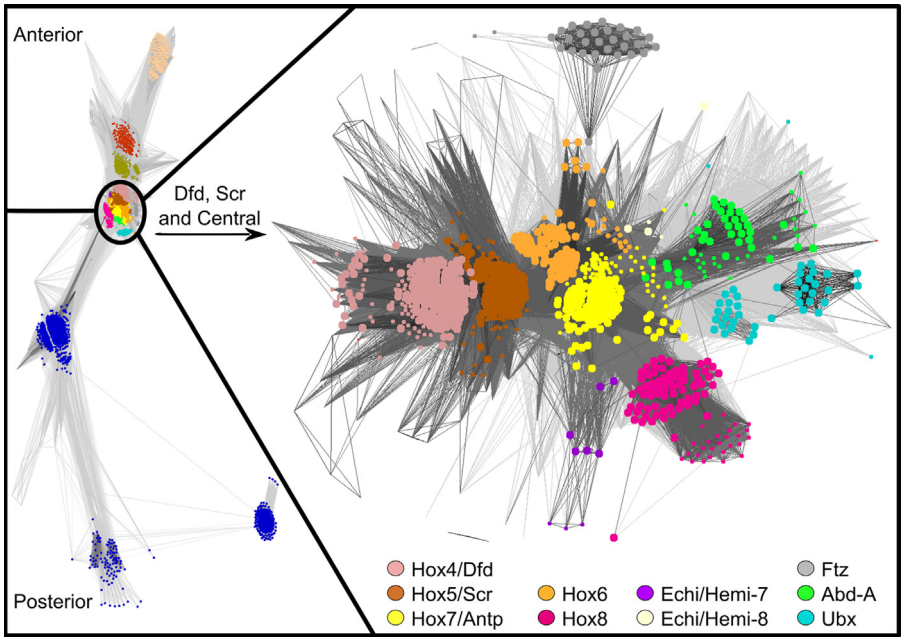

Fig. 2. 2D representation of the Hox protein sequence-similarities. CLANS representation of pairwise sequence-similarities for the 2629 sequences identified as belonging to the core Hox-group. Sequences are represented as dots and lines connecting the dots represent their pairwise similarities. The more similar the sequences, the darker the line, the stronger the attraction between dots, and the more closely they are located to each other in the map. Left side: CLANS map derived for the core group of Hox and Hox-like proteins (2629 sequences, similarity *p*-value cutoff=$1e^{-23}$). Coloring of the dots is according to the classification scheme shown in Fig. 1. Right side: CLANS map focusing on the Hox4, Hox5, Hox6, Hox7, Hox8, Dfd, Scr, Antp, Ftz, Abd-A, Ubx and sequence similar Hox proteins that formed one compact cluster in the map displayed on the left (1095 sequences, similarity *p*-value cutoff=$1e^{-30}$). Ten separate clusters are identifiable and color-coded according to the types of sequences present in each group. Large colored dots represent sequences containing the YPWM-motif, linker region and homeodomain, while small dots represent truncated or fragment sequences missing one or more of these elements. Dots without any coloring represent sequences that could not be unambiguously assigned to any specific sequence similarity group. The depicted maps are shown in 2D for technical reasons. The sequence groups were assigned based on a 3D-view of the dataset providing an additional discriminatory dimension. The 2D and the 3D versions of the sequence maps are provided in the supplementary materials, as is an overview of the similarities and differences between the vertebrate and arthropod sequence similarity groups (supplementary Fig. S3).

evidence supporting the functional predictions based on a syntenic classification of Hox genes (Malicki et al., 1990). The suggestion of functional equivalence is based on the similarity in phenotypes induced by ectopic expression of the Antp and Hox6 proteins (generic leg induction and bristle pattern in T1). It should be noted that Malicki et al. are cautious in their conclusions, explicitly noting that the type of leg induced by Hox6 is not the same type of leg induced by Antp and also note, that Hox7 is in fact more similar in sequence to Antp. Both comments are largely ignored in the literature, even though additional experiments have since indicated that Antp and Hox6 are not as similar in function as previously assumed. Follow-up experiments performed by Percival-Smith et al. (Percival-Smith et al., 2005) showed, for example, that induction of generic legs is a feature common to most ectopically expressed Hox proteins and does not represent a feature specific to Antp or Hox6. The second analyzed phenotype, the ability of other Hox proteins to induce a T1 bristle pattern, has never systematically been examined. The classical claim that Antp and Hox6 proteins display more functional similarities to each other than to other central-type Hox proteins therefore lacks adequate experimental support and raises questions about the accuracy of the synteny-classification-based functional predictions. The question therefore remains whether functional subtypes can be predicted for the central Hox proteins across the bilaterian clades and, if so, which of the central Hox proteins should be regarded as most similar in function.

Aim of this manuscript is to predict which of the central Hox proteins from different organisms are likely to exhibit similar biochemical properties, resulting in similar abilities to interact e.g. with DNA enhancer elements or protein interaction partners, leading to the corresponding ability of the proteins to induce similar developmental effects when expressed under the same experimental conditions. The underlying assumption of the approach we use, is that similarity of Hox proteins across the sequence elements we employ, which are known to be important to Hox protein function, can be used as a proxy to predict putative similarity in the biochemical properties or molecular mechanism of action of the corresponding proteins. In this manuscript, we specifically focus on the group of Hox proteins that exhibits the most errors in annotation: the central Hox proteins. The central Hox proteins include the Antennapedia (Antp), Ultrabithorax (Ubx) and Abdominal-A (Abd-A) proteins from *Drosophila* as well as the Hox6, 7, and 8 proteins from vertebrates. The method used in the manuscript is, so far, the only published sequence-based method that has proved capable of differentiating between central Hox proteins known to be functionally distinct (such as the *Drosophila* Antp, Ubx and Abd-A Hox proteins) (Hueber et al., 2010). The analysis is based on sequence regions spanning the YPWM-motif, 'linker'-region and homeodomain, which represent sequence elements known to be involved in determining the molecular function of Hox proteins (Fig. S1, (Merabet et al., 2003; Joshi et al., 2007)). We restricted our analysis to the region spanning the YPWM-linker-homeodomain of Hox proteins, as inclusion of additional sequence regions increases the prevalence of species specific sequence-elements, which would not provide us with information about sequence features that have remained conserved across distantly related species. Addition of species specific, but also clade specific sequence-elements that are not shared across all proteins to be compared (the linker regions can be clade specific in sequence length and composition, but every Hox protein possesses a linker region) can induce clustering biases. In the simplest case, proteins with species specific or clade specific sequence signatures may simply not be assignable because they contain a high content of additional sequence signatures that

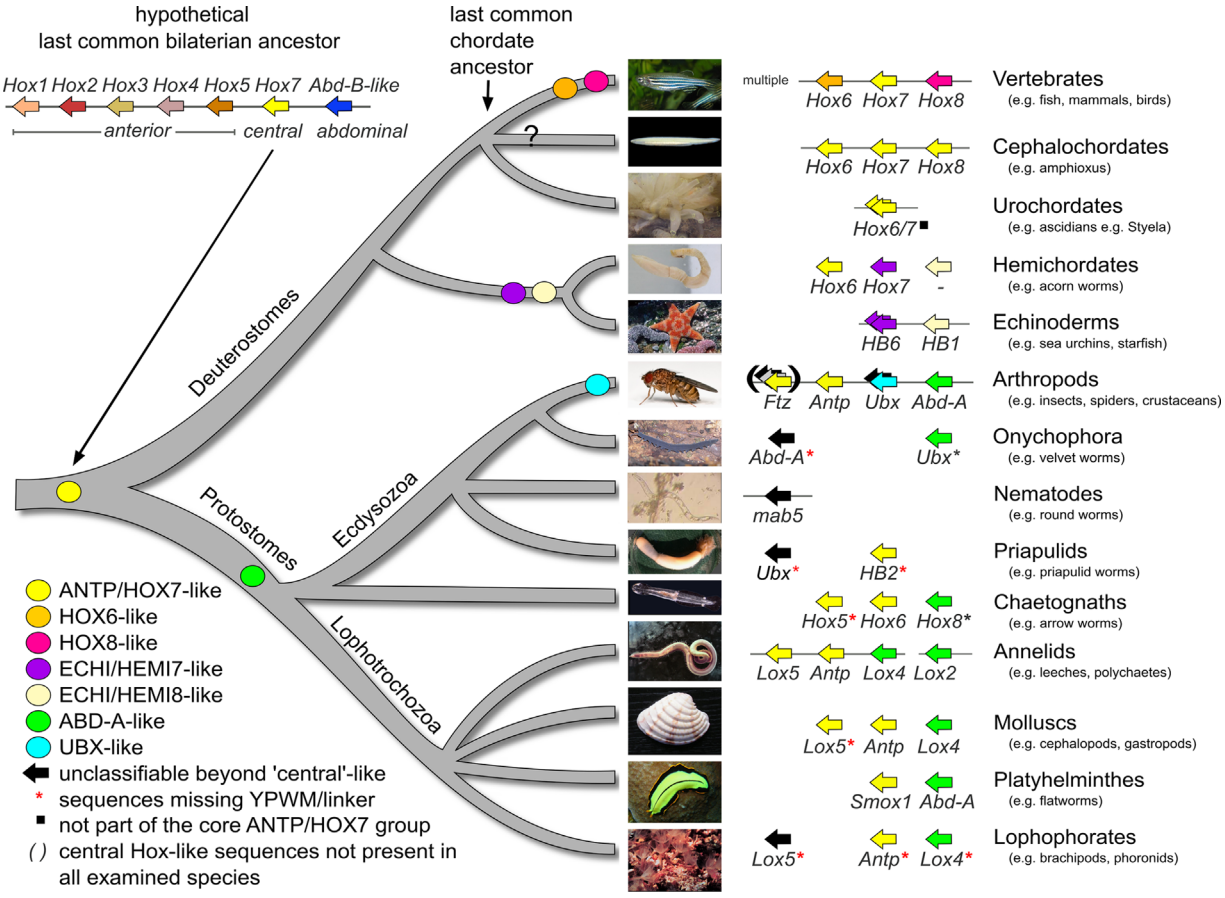

Fig. 3. Distribution of central Hox proteins across protostome and deuterostome lineages. The left side depicts a cladogram approximating the major taxonomic divisions in the protostome and deuterostome lineages, adapted from Dunn et al., 2008 (Dunn et al., 2008) and the presumed first occurrence of the different central Hox protein types (circles, color-coded as in Fig. 2). The right side depicts the genes (arrows), chromosomal arrangement and annotated names for the different types of central Hox proteins present in the various clades (consensus depicted) (sources for the pictures are listed in Supplementary Table S1). Horizontal lines connecting Hox genes indicate that these are represented according to their relative locations on the chromosome. Missing horizontal lines indicate an absence of data regarding the relative location of the genes in the genome. Positional information was determined by examining the genomic location of the Hox genes in sequenced organisms (vertebrates, arthropods, cephalochordates and nematodes). Additional positional information was received for cephalochordates (Garcia-Fernàndez and Holland, 1994), urochordates (Spagnuolo et al., 2003), echinoderms (Popodi et al., 1996), arthropods (Negre and Ruiz, 2007) and annelids (Fröbius et al., 2008). Identifiable isoforms (e.g. via FlyBase) are not shown. Sequences with a near identity to other sequences in the same species yet differing in more than three amino acids, i.e. potential splice variants or recently duplicated genes, are represented as slightly shifted arrows of the same color. No central-like Hox proteins could be identified in species outside the bilaterian clade.

are not shared with other proteins (resulting in either sequences placed outside of clusters or the formation of artificially separated clusters). In other cases, proteins may be grouped due to the presence of additional sequence motifs not found across all Hox proteins (e.g. the UbdA motif, which is specific to protostome Ubx and Abd-A proteins). In the worst case, such clustering biases can lead to proteins being clustered together, if they share low complexity regions such as polyA or polyQ regions, even though these regions may be acquired independently or exhibit differing functions depending on the precise location and composition of the poly amino acid regions within the protein. Employing a shorter sequence region rather than the YPWM-linker-homeodomain, e.g. using only the homeodomains, would abolish our ability to distinguish between the different types of central Hox proteins known to have different functions (see Hueber et al., 2010). Therefore, we base our analysis on the region of Hox protein sequences that is known to be relevant to DNA-binding, contains sufficient sequence-variation to enable us to distinguish between the different types of central Hox proteins known to have different functions, yet has remained identifiable across all Hox proteins being examined (and is conserved between those Hox proteins from arthropods and vertebrates that are known to exhibit similar phenotypes or can rescue each other in KO-studies (e.g. Dfd and Hox4 or Scr and Hox5)). The bioinformatics tool “CLANS” (Frickey and Lupas, 2004) is used to visualize the sequence-similarity patterns across the set of all available central Hox proteins in a 3D space to detect groups of sequences showing greater than average similarity to one another. This approach, using only the YPWM-linker-homeodomain regions, was previously successfully employed to resolve the phylogenetically unresolvable central-type Hox proteins (see Fig. 1C) into separate sequence similarity groups (i.e. sequence-types) (Hueber et al., 2010). This previous work focused on a restricted number of taxa (the model organisms mouse, zebrafish, amphioxus, *C. elegans* and *Drosophila*) and did not encompass the full breadth of variability present in the Hox protein sequences throughout the protostome and deuterostome lineages, nor did it allow any prediction regarding when the different central Hox protein sequence types might have arisen.

Primary aim of this work is to assess the extent to which the observed sequence patterns coincide with the branching pattern of the species tree and therefore are likely to be biologically meaningful, e.g. suitable to act as an indicator of a putative conserved molecular function, or whether the pattern is more randomly dispersed over the species tree, e.g. likely due to random or independent loss or acquisition of these sequence patterns. To address this point, we generated a resource providing a 3D visualization of the all-against-all pairwise similarities of the central Hox protein sequences present in the NCBI non redundant database (NCBI nr) (see Figs. 2 for a 2D

picture and supplementary materials for the corresponding 3D file). This resource includes a number of analysis features, for example allowing the visualization and highlighting or removal of sequences from a given taxonomic or similarity group (see suppl. Fig. S2). To assess the coincidence of the sequence-similarity groupings we observe with the presumed evolutionary history of the bilaterian lineage, we overlayed the presence and absence of the identified Hox sequence-types on to a published tree for bilaterian species (Dunn et al., 2008) (see Fig. 3). The groups we identified reflected monophyletic branches in the tree, indicating that the identified Hox sequence-types likely correspond to either a functional or evolutionary constraint acting on the Hox sequences. A random or polyphyletic placement of the identified sequence-types throughout the tree would, in contrast, have indicated that the sequence similarity-types identified were mainly based on random similarities in the sequences or due to sequence-features that arose convergently. Control tests were carried out to assess the dependence of the identified sequence-similarity groups on absence or presence of specific sequences. To this effect we removed sequences corresponding to known functional subgroups (see suppl. Fig. S4) as well as entire taxonomic clades (see suppl. Fig. S5). In addition, we also examined the reproducibility of the identified sequence-similarity groups under different similarity *p*-value cutoffs (suppl. Fig. S6). In all cases, the observed sequence-similarity groups were reproducible and indicated the existence of seven central Hox sequence-types in the bilaterian lineage.

Over the course of this analysis it became apparent that a considerable amount of sequences from non-standard organisms had highly inconsistent (or plainly incorrect) annotations. As central Hox proteins had the most errors in annotation, an additional aim of this manuscript is to facilitate a more accurate annotation of new sequences for researchers working with central Hox proteins. The idea is to replace the use of single or reciprocal BLAST searches and their corresponding pitfalls with a more accurate means of determining which of the central Hox protein sequence similarity groups best match a given novel unannotated sequence (or vice-versa). Furthermore, the resource retains and depicts potentially conflicting data and allows identification of derived sequences (or, if enough sequences are deposited, potentially the identification of novel central Hox sequence groups). All of this is provided via an online-service that allows the easy addition of new sequences into the set of Hox proteins used in our analysis. These new sequences are placed in 3D space in relation to their similarity to the Hox sequence-similarity groups, thereby providing a quick and easy prediction of which functional subgroups of the Hox family the new sequences are likely to belong to (http://bioinformatics.uni-konstanz.de/HueberHox/cHoxViewer/).

Materials and methods

Retrieval of Hox proteins

A flow-chart overview of the approach is depicted in Supplementary Fig. S7. To identify all central Hox protein sequences present in the NCBI-nr database (National Centre for Biotechnology Information non-redundant GenBank protein database, May 20th 2010, ftp.ncbi.nlm.nih-.gov/blast/db/FASTA/nr.gz), we used an iterated PSI-BLAST search (blastpgp version 2.2.22) (Altschul et al., 1997) (inclusion value of 10^{-30}, results returned up to *e*-values of 10). The eight *Drosophila melanogaster* Hox proteins were used as independent queries against the NCBI-nr database and run to convergence (Lab: 5 iterations, Pb: 7 iterations, Dfd: 13 iterations, Scr: 11 iterations, Antp: 12 iterations, Ubx: 19 iterations, Abd-A: 16 iterations and Abd-B: 2 iterations). From these searches, all high-scoring segment pairs (HSPs) with *e*-values up to 10 were taken and the corresponding full-length sequences were extracted from the database. This approach ensured the inclusion of all sequences that might be relevant to our analysis (50585 non-redundant sequences).

Aim was to start from a well-defined set of Hox proteins in the protostome lineage and see whether, and to which extent, we could identify the well-described sets of known Hox proteins from the vertebrate lineage, thereby giving us an estimate for how likely we were to miss other Hox protein sequences of putative interest. The above approach proved highly successful, as the search broadened to a point that we retrieved many homeodomain protein sequences from well outside the Hox protein family (NK, Paired/Pax, Wox, TALE, Lim, etc.). It is therefore unlikely that, using this approach, we missed any of the Hox protein family sequences present in the database.

Identifying YPWM, linker+homeodomain regions

The sequence region that previously provided the highest resolution classification for the central type Hox proteins is the region containing the YPWM (or FPWM), linker and homeodomain (YPWM-linker-homeodomain) (Hueber et al., 2010). To extract this region from the full-length sequences, we derived a Profile-Hidden-Markov-Model (HMM) from an alignment of the YPWM-linker-homeodomains for the eight *D. melanogaster* Hox proteins (programs used: AlnEdit (Frickey, 2005) and Muscle (Edgar, 2004)). The alignment was manually curated and a global HMM was derived using HMMer (Eddy, 1998). The resulting HMM, calibrated with 5000 replicates, was used to identify the corresponding YPWM-linker-homeodomains in the Hox-related full-length sequences (13282 sequences provided hits to the HMM with e-values better than 10). All Hox proteins known in the major model organisms (*Mus musculus*, *Danio rerio*, *Branchiostoma floridae*, *Caenorhabditis elegans*, *Drosophila melanogaster*), could be recovered using this *Drosophila*-centric approach, indicating that our HMM was general enough to adequately identify the homeodomains of all Hox proteins. The very relaxed *e*-value cutoff (10) was chosen to minimize the chance of excluding false-negative sequences from our subsequent analyses. The NCBI-nr database contained a number of Hox protein concatamers that are unlikely to represent any Hox proteins present in nature, as Hox proteins contain only one homeodomain. We therefore removed sequences containing more than a single homeodomain from the dataset (13049 sequences remained). The remaining sequences were subsequently analyzed using CLANS (Frickey and Lupas, 2004).

CLANS clustering (identification of Hox protein sequence similarity groups):

CLANS provides a visual representation, in 3D, of the pairwise similarities of all sequences to each other using a force-directed layout/clustering approach (Fruchterman-Reingold (Fruchterman and Reingold, 1991)). Sequences are represented as dots and the pairwise sequence similarities are visualized as lines connecting the respective dots. Higher similarities are represented by darker lines and correspond to higher pairwise attractive forces. The higher the pairwise similarity between two sequences, the closer these two sequences tend to be located in 3D space. Chance similarities have a negligible effect in such a large map as they are averaged out by the sheer number of pairwise similarities being analyzed. Only groups of sequences that exhibit a systematic degree of pairwise similarity across many of their sequences/members are pulled together into clusters. This facilitates the visual selection of groups of sequences with higher than average similarity to one another. As shown in Fig. 2 (left side), most of the Hox protein sequences form well separated clusters that are easily identifiable (e.g. Hox1, Hox2, Hox3, Hox9–13). In comparison, Dfd,

Scr and the central Hox proteins are much more similar to one another than any of the other groups and therefore are treated as one large cluster in this view. However, when examining this cluster in greater detail (Fig. 2, right side), Dfd+Hox4, Scr+Hox5, and the central Hox protein sequences readily resolve into separate clusters (in the 3D view and, for the core members of each cluster, also apparent under more stringent *P*-value cutoffs in the 2D views (Supplementary Fig. S6)). The sequence-based groupings produced by CLANS are inherently stable and robust and reproducible groupings are formed even when specific sequence groups or all sequences of a given taxonomic clade are removed (see Supplementary Fig. S4 and S5, respectively). Similarly, the relative location of the various sequence groups changes only little over a wide range of different *p*-value cutoffs (see Supplementary Fig. S6). The *p*-value cutoff chosen for the analysis ($P=1e^{-30}$) was the one providing a good visual separation of all sequence clusters, while excluding as little data as possible. As a control, we have also assessed the effects on Hox protein sequence groups by clustering full-length sequences (see Supplementary Fig. S8). The full length clustering provides the same basic sequence-similarity groupings.

Comparison of clustering results and species phylogeny

All sequences of Dfd+Hox4, Scr+Hox5 and central-like Hox proteins were examined in detail and the overwhelming majority could be assigned to one of the ten sequence-similarity clusters shown in Fig. 2 (right side). Supplementary Fig. S9 depicts profiles derived for each of the Hox sequence similarity clusters, thereby providing a visual representation of the sequence signatures that led to the respective clusters. A phylogeny of the species (Dunn et al., 2008) was used as a basis to map the presence of the respective central Hox protein types onto a cladogram depicting the major protostome and deuterostome lineages (Fig. 3). Specifically, we marked the earliest branch-point in the cladogram at which subsequent lineages contained a specific type of Hox protein. Multiple representative species were analyzed for each of the protostome and deuterostome clades to avoid potential artefacts arising from sparse sampling. The relative position of Hox genes in their respective genomic Hox gene cluster(s) was determined by examining the available genomes for sequenced organisms of the various lineages (vertebrates, arthropods, cephalochordates and nematodes). Additional information was retrieved from the literature for cephalochordates (Garcia-Fernàndez and Holland, 1994), urochordates (Spagnuolo et al., 2003), echinoderms (Popodi et al., 1996), arthropods (Negre and Ruiz, 2007) and annelids (Fröbius et al., 2008). The major taxonomic groupings we used in our analysis correspond to the NCBI taxonomic clades: Chordata, Cephalochordata, Urochordata, Hemichordata, Echinodermata, Arthropoda, Onychophora, Nematoda, Priapulida, Chaetognatha, Annelida, Mollusca, Platyhelminthes and Lophophorata (see Supplementary Fig. S2). The complete set of species that were manually validated regarding their taxonomic assignment and presence of Hox protein types are available in the CLANS save-files (see link below). Fig. 3 summarizes the above data by depicting which Hox protein types could be identified in which of the bilaterian lineages.

Web-resource and annotation of novel sequences

The online-resource http://bioinformatics.uni-konstanz.de/HueberHox/cHoxViewer/ provides a 3D visualization of the pairwise similarities across all central Hox sequences used in this analysis. Due to memory limitation imposed by the web-interface, only the relative positions of the sequences in 3D space are shown. The closer two dots are located in 3D space, the more similar their respective sequences were. Any subset of sequences present in the 3D map can be selected from a listing of sequence names provided via a tab below the graph window. Sequences selected by the user immediately appear highlighted in the map.

In addition, novel sequences can be added to the 3D sequence-similarity map of central and central-like Hox proteins. First, new sequences are compared to pre-calculated HMM's derived from the respective sequence-similarity groups. These HMM's were derived from high-quality alignments of the YPWM-linker-homeodomain-containing sequences present within each group. Novel sequences are then placed in the map so that their distance from the centers of gravity of the groups best reflects their relative similarities to the HMM's derived from the respective groups (closer=more similar). In addition to displaying these similarities graphically in the map, a selectable text-based listing of recently added sequences and their respective similarities to each group is provided in an additional tab.

The 3D overview of the similarities between each added sequence to the different sequence similarity groups and the associated text listing both provide an important resource to help annotate new central-like Hox proteins. The 3D overview provides a quick and easy way of placing novel sequences in relation to the identified sequence similarity groups, while the text based listing provides more detailed information as to how each sequence resembles each of the respective groups in cases where the 3D assignment may seem unclear or ambiguous.

Data files

The CLANS files on which this analysis was based, the version of CLANS used for the analysis and the alignments used in the Supplementary are available for download from http://bioinformatics.uni-konstanz.de/HueberHox/centralHox/ and as additional files included as part of this manuscript. The web-resource itself, providing a 3-D view of the location of the sequences in the CLANS map and the ability to classify novel sequences is accessible under http://bioinformatics.uni-konstanz.de/HueberHox/cHoxViewer/ (requires a WebGL enabled browser e.g. Firefox (version 4 or above) or Google Chrome). Due to the amount of data involved and the limitations on memory, graphics and GPU/CPU when displaying data via a web-browser, only the positions of the sequences/dots are shown. For full viewing and analysis capabilities please use the CLANS program [Additional File 10] (http://bioinformatics.uni-konstanz.de/HueberHox/centralHox/) in combination with the provided save-files.

Results

The CLANS dataset

The program CLANS was used to perform an all-against-all pairwise sequence similarity analysis of the set of all Hox protein sequences deposited in the NCBI non-redundant protein database "nr" (Fig. 2). The left side provides an overview of the similarities across the entire family of Hox proteins, while the right side depicts a more detailed representation of the cluster combining proteins of the Hox4–8, Dfd, Scr, Antp, Abd-A and Ubx sequence types. Sequences are depicted as dots (2629 (left side) and 1095 (right side)) and lines connecting the dots represent the corresponding pairwise sequence similarities. The more similar two sequences are to each other, the darker the line connecting them and the more closely they will be located in relation to each other in the map.

Our dataset also included numerous partial sequences that do not span the entire sequence region upon which this analysis is based (Supplementary Figure S1: YPWM-motif, linker region and homeodomain). For example, sequences that lack the YPWM+linker region are missing one of the most informative regions for classifying

central Hox proteins across deuterostome and protostome clades. Consequently, these partial sequences cannot be assigned to a specific Hox type with the same confidence as sequences containing the full YPWM-linker-homeodomain region. We therefore manually identified the sequences containing the complete region of interest (YPWM+linker+Homeodomain) and highlighted these with large colored dots in Fig. 2. Partial or fragment sequences are represented by smaller colored dots. Sequences we could not assign to a specific sequence similarity group were left uncolored.

Based on visual inspection, 10 sequence-similarity groups were identified (Fig. 2): two groups consisting of classical "anterior" Hox protein types (see Fig. 1) (De Rosa et al., 1999; Balavoine et al., 2002): Dfd/Hox4 (light red) and Scr/Hox5 (brown); a group combining *Drosophila* Ftz-like (gray) proteins (often regarded as recently derived from a central Hox protein ancestor (Telford, 2000; Löhr et al., 2001; Damen, 2002)); and seven distinct sequence similarity groups for the central Hox proteins. Five of the seven groups encompass the vertebrate and *Drosophila spp.* central-group Hox proteins, including Abd-A (green), Ubx (blue), Hox6 (orange), Hox8 (dark red) and Antp+Hox7 (yellow) (Hueber et al., 2010). The remaining two sequence-similarity groups have not previously been described and consist of central Hox protein types found only in the echinoderm and hemichordate lineages (Echi/Hemi7 colored in beige and Echi/Hemi8 in purple). The presence or absence of species within each identified sequence similarity grouping was subsequently mapped onto a cladogram depicting the phylogenetic relationships of the major protostome and deuterostome lineages (Fig. 3). By interpreting the presence and absence of Hox sequence types across the tree of the bilaterian lineage, we observe the following:

Four central Hox protein types are specific to deuterostomes

Three separate types of central Hox proteins were recognizable in all examined vertebrate groups, including deeply-branching vertebrates such as the lamprey *Petromyzon marinus* or the horn shark *Heterodontus francisci* as well as in "higher" vertebrates such as the mouse *Mus musculus*, chicken *Gallus gallus* and zebrafish *Danio rerio*: Hox6 (orange), Hox7 (yellow) and Hox8 (dark red) (Fig. 2 and Supplementary Fig. S2). Sequences grouping with the vertebrate Hox6 and Hox8 proteins could not be detected in any invertebrate species, including the cephalochordate *Branchiostoma lanceolatum*, the urochordate *Ciona intestinalis* as well as the hemichordates *Saccoglossus kowalevskii* and *Balanoglossus simodensis*, and the echinoderms *Strongylocentrotus purpuratus* and *Metacrinus rotundus* (Fig. 3 and Supplementary Fig. S2). All three central Hox proteins from *Branchiostoma lanceolatum* are most similar in sequence to the Antp/Hox7 group proteins (yellow) and, as such, cannot be regarded as either of the Hox6 or Hox8 sequence types. *Ciona intestinalis* contains a single central-type Hox protein which is most similar to the Antp/Hox7-like sequences. Hemichordates and echinoderms possess two previously undescribed lineage-specific types of central Hox proteins, that form two separate sequence-similarity groups: a group that lies peripheral to the Antp/Hox7 group (Echi/Hemi7) (purple) and a second, more derived, version of an Antp/Hox7 type sequence (Echi/Hemi8) (beige). A third central-type Hox protein present in hemichordates remains clearly identifiable as an Antp/Hox7 type sequence.

In summary, four central Hox protein types are specific to deuterostomes: Hox6 and Hox8 type proteins, which are only found in the vertebrate lineage, and Echi/Hemi7 and Echi/Hemi8 type proteins, which are only found in the echinoderm and hemichordate lineages.

Two central Hox protein types are specific to protostomes

In the arthropod clade, we could assign central Hox protein sequences from insects (*Drosophila melanogaster*, *Tribolium castaneum*), chelicerates (*Cupiennus salei*, *Parasteatoda tepidariorum*, *Ixodes scapularis*) and crustaceans (*Procambarus clarkii*, *Daphnia magna*) to each of three sequence-similarity groups: Antp/Hox7-like (yellow), Abd-A-like (green) and Ubx-like (blue) (Fig. 2 and Supplementary Fig. S2). While only a single Antp-like protein could be identified per species for the crustacean lineage, the species in the chelicerate and insect lineages contain two Antp-like proteins. In the chelicerates both proteins are highly similar to Antp type proteins, while in the insect lineage one of the proteins is clearly recognizable as an Antp type protein, referred to as *Drosophila* Antp, while the second has diverged in sequence and function to a greater extent and is referred to as *Drosophila* Ftz. This Ftz protein (gray) was previously shown to no longer have a Hox-like expression and function in *Drosophila* embryos (Ftz is a pair-rule protein previously predicted to have been derived from a central Hox protein ancestor (Telford, 2000; Löhr et al., 2001; Damen, 2002)).

In each of the three well-described nematode model organisms, *Caenorhabditis elegans*, *Caenorhabditis briggsae* and *Pristionchus pacificus*, we could identify only a single, central-like Hox protein (MAB-5 protein type), which could not be assigned to any specific sequence similarity group within the central Hox-proteins. The lophotrochozoa, represented in Fig. 3 by the annelid *Capitella teleta*, the cephalopod *Euprymna scolopes* and the platyhelminths *Schistosoma mansoni* and *Girardia tigrina,* all contain at least one Antp/Hox7-like and one Abd-A-like protein, but no *Drosophila* Ubx-like proteins.

In summary, two central Hox protein types are specific to protostomes: Abd-A type proteins can be found across all protostome clades, while *Drosophila* Ubx type sequences are restricted to the arthropod lineage.

The central Hox protein type Antp/Hox7 is common to both protostomes and deuterostomes

No Hox proteins resembling Dfd/Hox4, Scr/Hox5 or the above central Hox sequence types were identifiable in species outside the bilaterian clade.

Within the central Hox proteins, only a single sequence type (Antp/Hox7) could be identified as present across both the protostome and deuterostome lineages. With the exception of the *Drosophila* Ubx type, which is most similar to the Abd-A type proteins found in protostomes, all other central Hox protein types (vertebrate Hox6 & Hox8, echinoderm/hemichordate Echi/Hemi7 and Echi/Hemi8 and protostome Abd-A) are consistently more similar to Antp/Hox7-like proteins than to any of the other central Hox protein groups.

The central Hox Webserver

All the above mentioned results are derived from a pairwise sequences similarity analysis performed in the CLANS software. The central Hox Web-resource, presented in this manuscript, was developed to provide a means for novel sequences to be classified in relation to the known Hox and central Hox proteins that constituted our dataset. Classification of novel sequences is performed based on a comparison of the sequences to Profile-Hidden-Markov-Models (HMM's) derived from multiple sequences alignments of the sequence similarity groups identified in the CLANS analysis (the respective alignments are available in the supplementary materials).

New sequences can be added via the "add sequences" tab in the viewer and thereafter automatically appear in the 3D map. Each sequence is placed in the map so that distance between itself and the center-of-gravity of the respective sequence-similarity groups best reflects the differences between the E-values of the sequence-to-HMM comparisons. The better the match, the closer the sequence is placed to the center of gravity of the given group. In addition, the precise *E*-values of the sequence-to-HMM comparisons are shown as bar-graphs in a further tab.

For most sequences this approach allows a clear classification into one of the sequence-similarity groups described above. However, this approach has some drawbacks and two exceptions are worth mentioning. In cases where only the homeodomain region of central Hox proteins is submitted, the corresponding sequence-to-HMM comparisons across all groups produce very similar *E*-values, resulting in the sequences being placed almost equidistantly to the groups and no clear assignment being possible. This highlights the importance of including the complete YPWM and linker region when attempting to classify central and central-like Hox proteins. In another case (gi|2352536 [Hellobdella triseralis]), use of the full YPWM+linker+homeodomain sequence failed to result in a clear assignment, with the sequence appearing as equally similar to the Dfd/Hox4 and Scr/Hox5 groups (the *E*-value difference between the first hit, Scr/Hox5, and the second hit, Dfd/Hox4, is negligible). Depending on the amino acids used to exhibit a given function, we would expect this protein, annotated as Lox20, to act either more similarly to Hox5/Scr (e.g. due to the threonine in position 13 of the homeodomain) or to Hox4/Dfd (e.g. absence of a tyrosine in position 11 of the homeodomain). Should the linker region be relevant to a given Dfd/Hox4 or Scr/Hox5 specific function, this particular Lox20 protein may not share said function with either Hox4 or Hox5. In cases where the YPWM+Linker+homeodomain region does not provide a clear assignment, we strongly recommend to look at sequences outside this region as, depending on the Hox sequence-type being examined, their may exist additional sequence motifs suited to providing a clear-cut assignment.

Both examples are used to point out that the classification provided by the web-interface is highly dependent on the precise sequence elements used. Truncated or atypical sequences may lead to suboptimal results and results beyond the best-hit (e.g. the second-best hit showing a similar *E*-value to the first hit) should be taken into account to gauge the confidence of any given assignment.

Discussion

We provide the first large-scale sequence-comparison based resource for the Hox protein family that is capable of resolving the various types of central Hox proteins present across protostomes and deuterostomes. Previous work by Thomas-Chollier et al. (Thomas-Chollier et al., 2007), resulted in a clear classification of the central Hox proteins in the vertebrate lineage. However, in their later work (Thomas-Chollier et al., 2010), which incorporated both protostomes and deuterostomes sequences, they no longer resolved the vertebrate central Hox proteins, instead, grouping them into a single large Hox6/7/8 group of paralogs. It is noteworthy to remark that although we utilize a very different approach (visualization of all-against-all pairwise sequence similarities versus HMM models with yes or no answers to whether a given sequence belongs to a HMM defined by Thomas-Chollier), our classification displays similarities to their latter work, in which the Antp/Ftz/Lox5 proteins in the protostome lineage were also identified as most similar to the vertebrate central Hox proteins. In addition, we identify vertebrate Hox7 proteins as the most similar to the protostome Antp/Lox5 type proteins and determine approximately when the various central Hox protein sequence-types diversified. That our visualization approach retains and displays conflicting data is important as the presence or absence as well as the quantity of conflicting data provides crucial information to gauge how reliable each assignment of a sequence to a given sequence-similarity group is likely to be.

As expected, the sequence similarity based groupings of the anterior Hox proteins Hox1/Lab Hox2/Pb, Hox3, Hox4/Dfd, Hox5/Scr are consistent with previous groupings based on phylogenetic inference, synteny data as well as studies comparing Hox protein functions. Phylogenetic approaches as well as the available functional data cannot resolve the 'central' and 'abdominal' type proteins and surrogate classification methods employed to resolve the central Hox group, i.e. synteny or sequence-similarity based, provide conflicting results (Fig. 1 and Supplementary Fig. 3). Our analysis indicates that the conflict between sequence-similarity and synteny is not due to Antp and Hox7 type proteins having convergently evolved similarities in the arthropod and vertebrate lineages, as all major protostome and deuterostome clades contain a central Hox protein of the Antp/Hox7 type. Instead, the taxonomic distribution indicates that the sequence pattern between Antp and Hox7 is due to a conserved sequence element (see Supplementary Figure 8 and compare to Supplementary Figure 3). Moreover, nearly all clade specific central Hox protein types (Hox6, Hox8 Echi/Hemi7, Echi/Hemi8 and Abd-A) are most similar in sequence to the Antp/Hox7 type. We therefore conclude that Antp/Hox7 type sequences best represent the central Hox protein sequence type present in the last common ancestor of the protostome and deuterostome lineage. The other six central Hox protein types identified are specific to clades branching after the protostome-deuterostome divide and therefore represent more recently derived members of the central Hox group.

Deuterostome lineage

The most deeply branching point at which we can identify well differentiated Hox6 and Hox8 type sequences is within the vertebrate lineage after its split from the urochordate and cephalochordate lineages. Hox6 and Hox8 type proteins appear to have diverged further from the ancestral-type central Hox protein in sequence, and presumably in function, than Hox7. Similarly, the most deeply branching point at which we can identify well differentiated Echi/Hemi7 and Echi/Hemi8 sequences is within the echinoderm and hemichordate lineages after their split from the chordates, which also indicates a divergence in sequence and function of these proteins from the ancestral-type central Hox protein. In summary, all major deuterostome clades possess proteins of the Antp/Hox7 type in addition to various clade specific central Hox protein types (Hox6, Hox8 Echi/Hemi7 and Echi/Hemi8), which are likely to have been derived from duplications/triplications of an ancestral Antp/Hox7 type protein.

Protostome lineage

The most deeply branching point at which well differentiated *Drosophila* Ubx type proteins can be identified is within the arthropod lineage. Even among the Ubx-type sequences, two distinct sequence-similarity groups can be discerned – the first includes proteins of the Ubx-IA isoform (isoform E), specific to dipterans, and the second includes proteins of the Ubx type-IVA (isoform B), which is found in insects, chelicerates and crustaceans. Balavoine et al. present two evolutionary scenarios for Lox2, Lox4, Abd-A and Ubx proteins (Balavoine et al., 2002). In the first, the annelid Lox2 is regarded as an ortholog of *Drosophila* Ubx, however, our analyses assign Lox2 and Ubx to different sequence-similarity groups and, instead, assign Lox2 as similar to Abd-A

(Fig. 3). The clustering of Lox2 with arthropod Abd-A type proteins is more consistent with the second evolutionary scenario proposed by Balavione et al., i.e. that Lox2, Lox4, Abd-A and Ubx proteins were derived from a single gene present in the last common ancestor of protostomes. Our observation that Ubx is arthropod specific and that dipterans have a lineage specific isoform of this protein also supports the claim that Ubx is a rapidly evolving Hox protein (Ronshaugen et al., 2002). Moreover, our observation that the Ubx isoforms Ia and IVA form distinct sequence-similarity groups fits well with previous comparative studies that demonstrate divergent functions for these protein isoforms in *Drosophila* (Liu et al., 2009; Reed et al., 2010). It should be noted that we cannot rule out the presence of further Ubx type sequences in the genomes of other protostomes as, in some lineages, sequence data were only available from single species at the time of our analyses (most notably for the onychophoran and lophophorate species *Acanthokara kaputensis* and *Lingula anatina*). In contrast, sequences of the Antp/Hox7 and Abd-A central Hox types are present in nearly all protostome lineages including the ecdysozoan and lophotrochozoan clades, indicating that the last common protostome ancestor likely possessed at least two differentiated central type Hox proteins, one of the Antp/Hox7 type and a second of the Abd-A type.

Our analyses cluster Lox4 amongst the Abd-A group sequences and Lox2 alongside, but not within the Lox4/Abd-A group, which is consistent with the evolutionary scenario proposed by Balavoine et al. (Balavoine et al., 2002) that the Ubd-A peptide containing group of Hox proteins may have arisen from a central-type Hox protein after the protostome/deuterostome split. Based on our data, this ancestral protostome specific protein would best be represented by proteins of the Abd-A group. However, our data does not provide insights as to how or when the different UbdA-motif containing sequences Abd-A, Ubx, Lox2 or Lox4 might have arisen as the UbdA-motifs are specific to only a subset of the central Hox and therefore were not taken into account in this analysis (no sequence elements that are not present across the majority of Hox proteins were taken into account).

The ancestral central type Hox protein

Our analyses suggest that it is unlikely the last common ancestor of protostomes and deuterostomes possessed an even partially differentiated triplet of central Hox proteins, as only a single central Hox protein is common to all protostome and deuterostome lineages (Antp/Hox7). All other central type Hox sequence similarity groups are specific to either the protostomes or deuterostome lineages and, with the single exception of the rapidly evolving Ubx proteins, display the highest level of sequence similarity to sequences of the Antp/Hox7 type. This indicates that each of these groups was independently derived from an ancestral protein which is now best represented by sequences of the Antp/Hox7 type.

We propose two possible explanations for these observations: (1) First, the last common ancestor of protostomes and deuterostomes possessed only a single Antp/Hox7 type sequence and subsequent duplication and divergence of this protein gave rise to the lineage-specific forms. (2) Second, the last common ancestor of these lineages already possessed multiple copies of an Antp/Hox7 type protein that thereafter were subject to very different selective constraints and therefore evolved divergently to form the lineage-specific forms. In either case, functional comparisons of central Hox proteins between protostome and deuterostome species can only be expected to yield information about features that have remained conserved since the protostome-deuterostome split. To gain insights into the link between conserved sequence elements and conserved function of Hox proteins, the examination of lineage specific Hox proteins types is not to be recommended (e.g. the proteins Hox8 and Ubx, predicted to be orthologous by synteny but not phylogeny or sequence similarity, shared next to no sequence signatures that indicate the presence of a putative shared/conserved function). Based on analysis of the pairwise sequence similarities, the highest functional similarity across the central Hox proteins is predicted for comparisons of the Antp and Hox7 proteins. However, one should keep in mind that lineage specific Hox protein types may still contain additional conserved sequence signatures specific to the central Hox proteins. Depending on the precise molecular mechanism or Hox-function being examined and the precise sequence elements involved in this protein function, it is also possible that Antp appears most similar to Hox7 in one function (e.g. if the RSXXXD linker or near identical homeodomains are of predominant importance), while in another function Antp may appear most similar to Hox8 (e.g. if an RXQ sequence in the linker is of predominant importance and small variations in the homeodomain are negligible for this function). The ideal experimental setup to elucidate which sequence signatures are necessary for a given function, would therefore include all those Hox proteins from the species to be compared (e.g. *Drosophila* and mouse) that contain central Hox specific sequence signatures that are conserved across the protostome/deuterostome split.

The old and the new classifications

Previously, central Hox proteins from protostomes and deuterostomes were either classified as a single phylogenetically unresolvable group or, alternatively, classified by synteny as three groups of orthologous and functionally equivalent sequences, i.e. the orthologous pairs being Hox6 and Antp, Hox7 and Ubx, Hox8 & Abd-A (Gehring et al., 2009; Mann et al., 2009; Choo and Russell, 2011; Michaut et al., 2011; Pick and Heffer, 2012) (see Fig. 1). Based on our pairwise-sequence-similarity clustering, we identified seven distinct sequence types for the central Hox proteins: one present in both protostomes and deuterostomes (Antp/Hox7), one present in all protostomes (Abd-A), one specific to arthropods (Ubx), two specific to vertebrates (Hox6, Hox8) and two previously undescribed types of Hox proteins specific to the echinoderm and hemichordate lineages (Echi/Hemi7, Echi/Hemi8) (see Fig. 3). The most deeply branching point at which we can identify Hox protein sequences of the Dfd/Hox4, Scr/Hox5 and central Hox types is within the bilateria. This observation is consistent with the hypothesis that only bilaterians possess central and central-like Hox proteins (Kamm et al., 2006; Chourrout et al., 2006; Ryan et al., 2006; Amemiya and Wagner, 2006; Ryan et al., 2007). However, it should be noted that this analysis represents a snapshot of the currently available data. Not all bilaterian clades are equally well represented in the public databases and, with the accumulation of more sequence data for the poorly sampled clades, it is possible that further lineage specific Hox sequence types remain to be discovered. In addition, we would like to point out that the sequences we analyzed are known to be relevant to the function of Hox proteins and are conserved in clearly defined functionally similar proteins across protostomes (usually *Drosophila*) and deuterostomes (usually vertebrates). This does not mean that the YPWM-motif, 'linker'-region and homeodomain are the only sequences relevant to the function of Hox proteins. Species specific elements nearly always exist, but this analysis focused specifically on those sequence elements present across the whole Hox protein family.

Conclusion

For Hox proteins, the accuracy of the often depicted "synteny-based" classification scheme should be re-evaluated in light of our

findings. While the more anterior Hox proteins are indeed highly conserved in both homeodomain sequence and protein function, the analysis we present indicates that the central Hox proteins from protostomes and deuterostomes are more diverse than previously assumed (Gehring et al., 2009; Mann et al., 2009; Choo and Russell, 2011; Michaut et al., 2011; Pick and Heffer, 2012) and that the Hox7/Antp-like proteins are likely to represent the least derived form of a putative ancestral central-type Hox protein.

Additional implications arise from the results of our analysis: On the one hand, the ability to differentiate between conserved and more recently evolved sequence types can provide new supplementary sets of synapomorphic traits to classify species by. On the other hand, the ability to accurately identify which proteins diverged in sequence and which retained ancestral sequence features is crucial to elucidating the link between protein sequence and protein function for the Hox proteins. Lastly, there are other cases where a high-quality multiple sequence alignment is not feasible for some of the sequence regions known to be relevant to protein function (such as the YPWM-linker-homeodomain region) and where the regions that can be unambiguously multiply aligned, such as the homeodomain on its own, do not provide sufficient phylogenetic signal to resolve the protein family. In such cases, we regard the use of an all-against-all pairwise similarity based approach coupled with the information provided by a species phylogeny as an extremely useful resource to help identify further sequence elements conserved across taxonomic clades.

The advantages of using an all-against-all pairwise similarity based approach for classification and annotation and their corresponding 3D visualization via tools such as CLANS is that by depicting the all-against-all similarities more information can be utilized (e.g. additional sequence information) and retained (e.g. more information is displayed than in classification tools providing a "black box" that simply assigns a given sequence to a classification group without providing information of how similar the sequence is to the best similarity group in comparison to how similar the sequences of a given group are to each other) and conflicting data is incorporated into the visualization and remains available to the user (e.g. how similar the sequence is to the 2nd or 3rd best similarity groups). Employing all-against-all similarities, rather than the more common single or reciprocal best blast hits often used for annotating new sequences, avoids many of the biases that plague BLAST-based annotation approaches. In addition, the ability to retain conflicting information can greatly improve annotation confidence as, e.g. a chimeric sequence derived from multiple Hox proteins are shown as similar to multiple sequence-similarity groups, but are not clearly assigned to any single one of the sequence/functional group it was derived from. Highly derived sequences are, similarly, clearly identifiable as not part of any known functional or sequence-similarity group. Both our central Hox protein classification resources, the downloadable version with an encompassing software toolkit as well as the web-based version, will hopefully allow a more detailed analysis and annotation of novel central Hox protein sequence in the future.

Authors' contributions

SDH conceived the study, carried out the sequence-similarity analysis and drafted the manuscript. JR participated in the analysis and provided materials and analysis tools. MAD provided material and analysis tools, participated in the study design, coordination and manuscript preparation. HG participated in the analysis and manuscript preparation. GFW provided material, participated in the study design and manuscript preparation. TF provided material and analysis tools, participated in the study design, coordination, analysis and manuscript preparation. All authors have read and approved the final manuscript.

Acknowledgments

We would like to thank Michael Schubert for contributing to Fig. 3.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.ydbio.2013.09.009.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
Amemiya, C.T., Wagner, G.P., 2006. Animal Evolution: When Did the "Hox System"-Arise? Curr. Biol. 16, R546–R548.
Atkinson, H.J., Morris, J.H., Ferrin, T.E., Babbitt, P.C., 2009. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PloS one 4, e4345.
Balavoine, G., De Rosa, R., Adoutte, A., 2002. Hox clusters and bilaterian phylogeny. Mol. Phylogenet Evol. 24, 366.
Choo, S.W., Russell, S., 2011. Genomic approaches to understanding Hox gene function. Adv. Genetics 73, 55.
Chourrout, D., Delsuc, F., Chourrout, P., Edvardsen, R.B., Rentzsch, F., Renfer, E., Jensen, M.F., Zhu, B., De Jong, P., Steele, R.E., 2006. Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. Nature 442, 684–687.
Carroll, S.B., Grenier, J.K., Weatherbee, S.D., 2005. From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design. Wiley-Blackwell.
Damen, W.G., 2002. fushi tarazu: a Hox gene changes its role. BioEssays 24, 992.
De Rosa, R., Grenier, J.K., Andreeva, T., Cook, C.E., Adoutte, A., Akam, M., Carroll, S.B., Balavoine, G., 1999. Hox genes in brachiopods and priapulids and protostome evolution. Nature 399, 772–776.
De Rosa, R., Grenier, J.K., Andreeva, T., Cook, C.E., Adoutte, A., Akam, M., Carroll, S.B., Balavoine, G., 1999. Hox genes in brachiopods and priapulids and protostome evolution. Nature 399, 772–776.
DeLano, W.L., 2002. The PyMOL molecular graphics system. DeLano Scientific.
Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745–749.
Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14, 755–763.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Feiner, N., Ericsson, R., Meyer, A., Kuraku, S., 2011. Revisiting the origin of the vertebrate Hox14 by including its relict sarcopterygian members. J. Exp. Zool. B Mol. Evol. 316, 515–525.
Frickey, T., Lupas, A., 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics 20, 3702–3704.
Frickey T. 2005. AlnEdit at the RSBS bioinformatics server. ⟨http://bioinformatics.uni-konstanz.de/programs/alnedit⟩ (accessed 23.12.09).
Frickey T. 2010. MotifDraw: Create a graphical representation of a motif (for example identified by Mclip). ⟨http://bioinformatics.uni-konstanz.de/utils/index.php#0⟩ (accessed 09.01.12).
Fröbius, A.C., Matus, D.Q., Seaver, E.C., 2008. Genomic organization and expression demonstrate spatial and temporal Hox gene colinearity in the lophotrochozoan Capitella sp I. PLoS ONE 3, 4004.
Fruchterman, T.M., Reingold, E.M., 1991. Graph drawing by force-directed placement Software: Practice and experience 21, 1129–1164.
Garcia-Fernàndez, J., Holland, P.W.H., 1994. Archetypal organization of the amphioxus Hox gene cluster. Nature 370, 563–566.
Garcia-Fernàndez, J., 2005. The genesis and evolution of homeobox gene clusters. Nat. Rev. Genet. 6, 881–892.
Gehring, W.J., Kloter, U., Suga, H., 2009. Evolution of the Hox gene complex from an evolutionary ground state. Curr. Top Dev. Biol. 88, 35–61.
Graham, A., Papalopulu, N., Krumlauf, R., 1989. The murine and Drosophila homeobox gene complexes have common features of organization and expression. Cell 57, 367–378.
Hueber, S.D., Bezdan, D., Henz, S.R., Blank, M., Wu, H., Lohmann, I., 2007. Comparative analysis of Hox downstream genes in Drosophila. Development 134, 381–392.
Hueber, S.D., Lohmann, I., 2008. Shaping segments: Hox gene function in the genomic age. Bioessays 30, 965–979.
Hueber S. D. 2009. Identification and functional analysis of Hox downstream genes in Drosophila. Identification and functional analysis of Hox downstream genes in Drosophila, Eberhard Karls University of Tuebingen, Biology Department ⟨http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-38027⟩.

Hueber, S.D., Weiller, G.F., Djordjevic, M.A., Frickey, T., 2010. Improving Hox protein classification across the major model organisms. PLoS ONE 5, e10820.
Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B., Mann, R.S., 2007. Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell 131, 530–543.
Kamm, K., Schierwater, B., Jakob, W., Dellaporta, S.L., Miller, D.J., 2006. Axial patterning and diversification in the cnidaria predate the Hox system. Curr. Biol. 16, 920–926.
Krumlauf, R., 1992. Evolution of the vertebrate Hox homeobox genes. Bioessays 14, 245–252.
Lewis, E.B., 1978. A gene complex controlling segmentation in Drosophila. Nature 276, 565–570.
Liu, Y., Matthews, K.S., Bondos, S.E., 2009. Internal regulatory interactions determine DNA binding specificity by a Hox transcription factor. J. Mol. Biol. 390, 760–774.
Löhr, U., Yussa, M., Pick, L., 2001. Drosophila fushi tarazu: a gene on the border of homeotic function. Curr. Biol. 11, 1403–1412.
Lutz, B., Lu, H.C., Eichele, G., Miller, D., Kaufman, T.C., 1996. Rescue of Drosophila labial null mutant by the chicken ortholog Hoxb-1 demonstrates that the function of Hox genes is phylogenetically conserved. Genes Dev. 10, 176–184.
Malicki, J., Schughart, K., McGinnis, W., 1990. Mouse Hox-2.2 specifies thoracic segmental identity in Drosophila embryos and larvae. Cell 63, 961–967.
Merabet, S., Hudry, B., Saadaoui, M., Graba, Y., 2009. Classification of sequence signatures: a guide to Hox protein function. Bioessays 31, 500–511.
Mann, R.S., Lelli, K.M., Joshi, R., 2009. Hox Specificity: unique roles for cofactors and collaborators. Curr. Top Dev. Biol. 88, 63.
McGinnis, N., Kuziora, M.A., McGinnis, W., 1990. Human Hox-4.2 and Drosophila deformed encode similar regulatory specificities in Drosophila embryos and larvae. Cell 63, 969–976.
Merabet, S., Kambris, Z., Capovilla, M., Bérenger, H., Pradel, J., Graba, Y., 2003. The hexapeptide and linker regions of the AbdA Hox protein regulate its activating and repressive functions. Dev. Cell 4, 761–768.
Michaut, L., Jansen, H.J., Bardine, N., Durston, A.J., Gehring, W.J., 2011. Analyzing the function of a hox gene: an evolutionary approach. Dev. Growth Differ. 53, 982–993.
Negre, B., Ruiz, A., 2007. HOM-C evolution in Drosophila: is there a need for Hox gene clustering? Trends Genet. 23, 55–59.
Percival-Smith, A., Teft, W.A., Barta, J.L., 2005. Tarsus determination in Drosophila melanogaster. Genome 48, 712–721.
Pick, L., Heffer, A., 2012. Hox gene evolution: multiple mechanisms contributing to evolutionary novelties. Ann. N.Y. Acad. Sci. 1256, 15–32.
Popodi, E., Kissinger, J.C., Andrews, M.E., Raff, R.A., 1996. Sea urchin Hox genes: insights into the ancestral Hox cluster. Mol. Biol. Evol. 13, 1078.
Reed, H.C., Hoare, T., Thomsen, S., Weaver, T.A., White, R.A.H., Akam, M., Alonso, C. R., 2010. Alternative splicing modulates Ubx protein function in Drosophila melanogaster. Genetics 184, 745.
Ronshaugen, M., McGinnis, N., McGinnis, W., 2002. Hox protein mutation and macroevolution of the insect body plan. Nature 415, 914–917.
Ryan, J.F., Burton, P.M., Mazza, M.E., Kwong, G.K., Mullikin, J.C., Finnerty, J.R., 2006. The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: evidence from the starlet sea anemone, Nematostella vectensis. Genome Biol. 7, 1–20.
Ryan, J.F., Mazza, M.E., Pang, K., Matus, D.Q., Baxevanis, A.D., Martindale, M.Q., Finnerty, J.R., 2007. Pre-bilaterian origins of the Hox cluster and the Hox code: evidence from the sea anemone, Nematostella vectensis. PLoS One, 2.
Samadi, L., Steiner, G., 2009. Involvement of Hox genes in shell morphogenesis in the encapsulated development of a top shell gastropod (Gibbula varia L.). Dev. Genes Evol. 219, 523–530.
Spagnuolo, A., Ristoratore, F., Di Gregorio, A., Aniello, F., Branno, M., Di Lauro, R., 2003. Unusual number and genomic organization of Hox genes in the tunicate Ciona intestinalis. Gene 309, 71–79.
Telford, M.J., 2000. Evidence for the derivation of the Drosophila fushi tarazu gene from a Hox gene orthologous to lophotrochozoan Lox5. Curr. Biol. 10, 349–352.
Thomas-Chollier, M., Leyns, L., Ledent, V., 2007. HoxPred: automated classification of Hox proteins using combinations of generalised profiles. BMC Bioinformatics 8, 247–258.
Thomas-Chollier, M., Ledent, V., Leyns, L., Vervoort, M., 2010. A non-tree-based comprehensive study of metazoan Hox and ParaHox genes prompts new insights into their origin and evolution. BMC Evolutionary Biology 10, 73–84.
Veraksa, A., Del Campo, M., McGinnis, W., 2000. Developmental patterning genes and their conserved functions: from model organisms to humans. Mol. Genet. Metab. 69, 85–100.
Zhao, J.J., Lazzarini, R.A., Pick, L., 1993. The mouse Hox-1.3 gene is functionally equivalent to the Drosophila Sex combs reduced gene. Genes Dev. 7, 343–354.