Current Biology, Volume *22*

Supplemental Information

Ghost Loci Imply Hox and ParaHox Existence in the Last Common Ancestor of Animals

Olivia Mendivil Ramos, Daniel Barker, David E.K. Ferrier

Supplemental Inventory

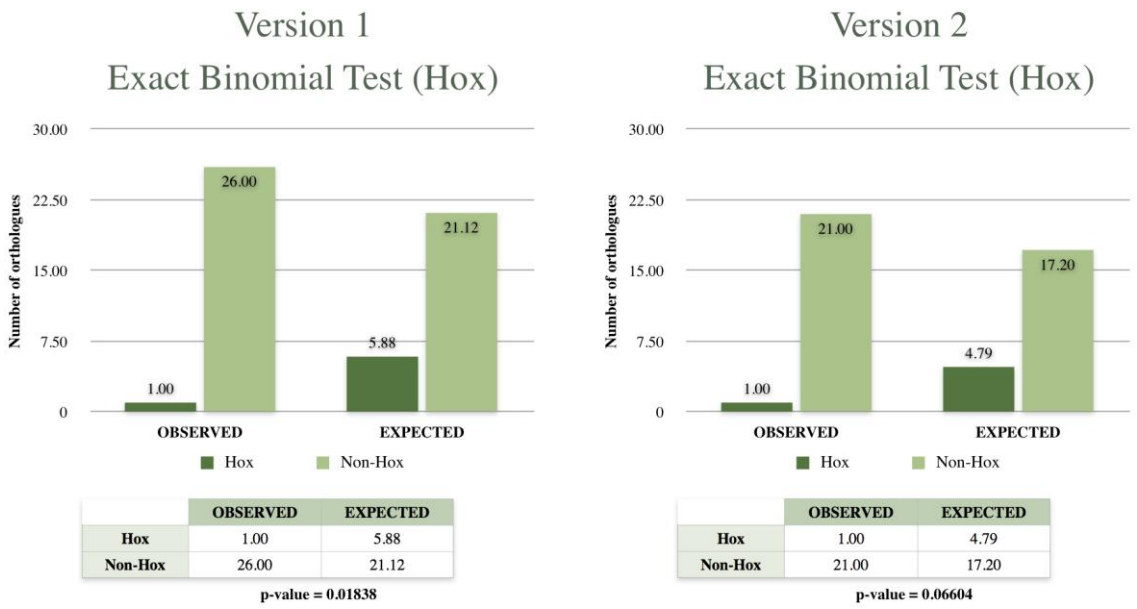


Figure S1A. Hox Exact Binomial Tests (Version 1 and Version 2), Related to Figure 2
* statistically significant at the 5% level

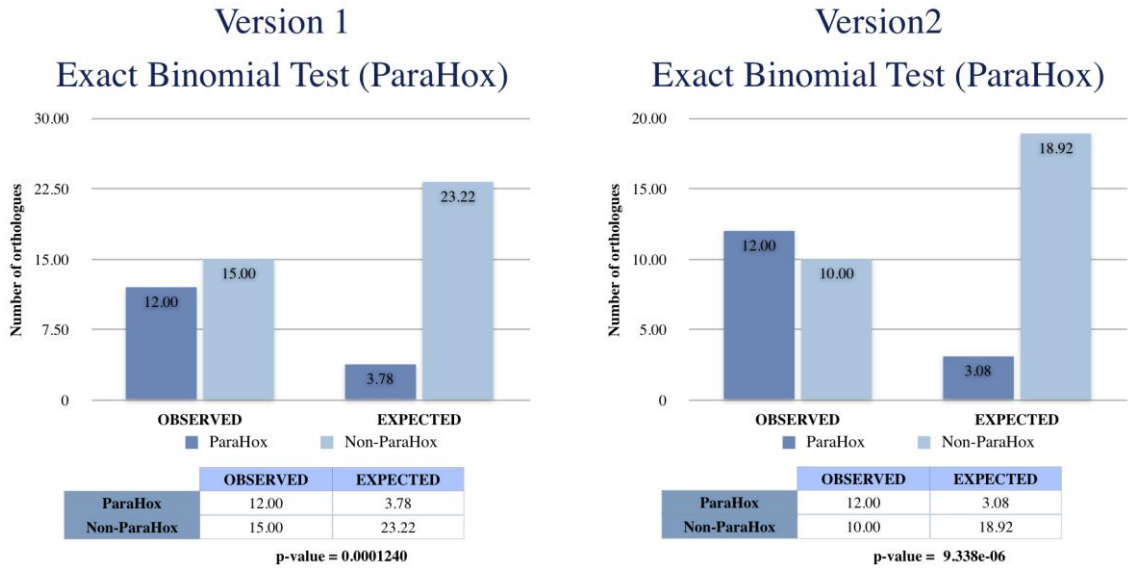


Figure S1B. ParaHox Exact Binomial Tests (Version 1 and Version 2) , Related to Figure 2
** statistically significant at the 1% level

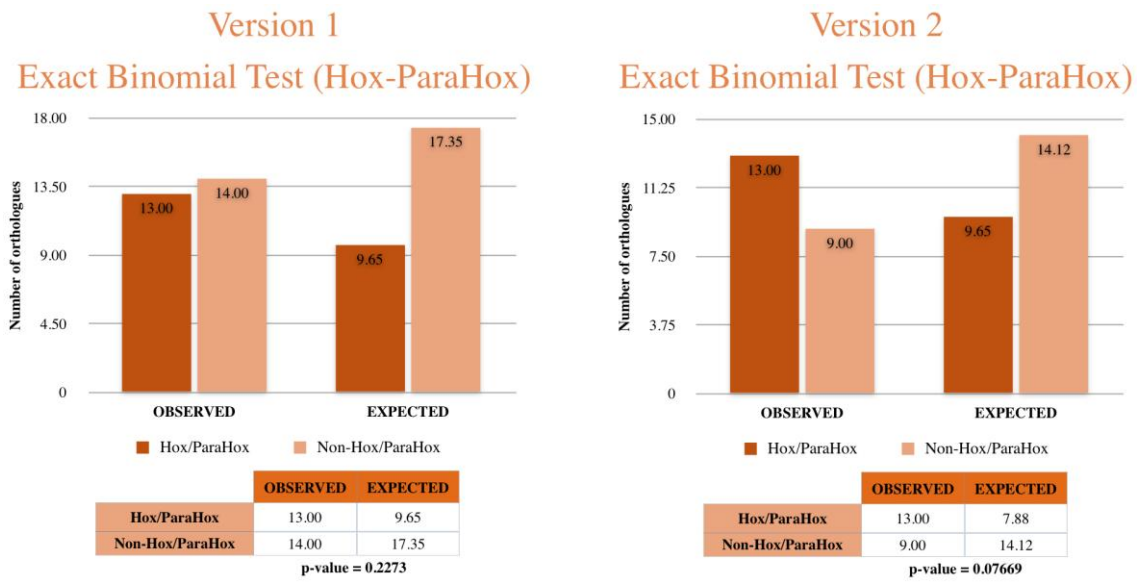


Figure S1C. Hox/ParaHox Exact Binomial Tests (Version 1 and Version 2), Related to Figure 2

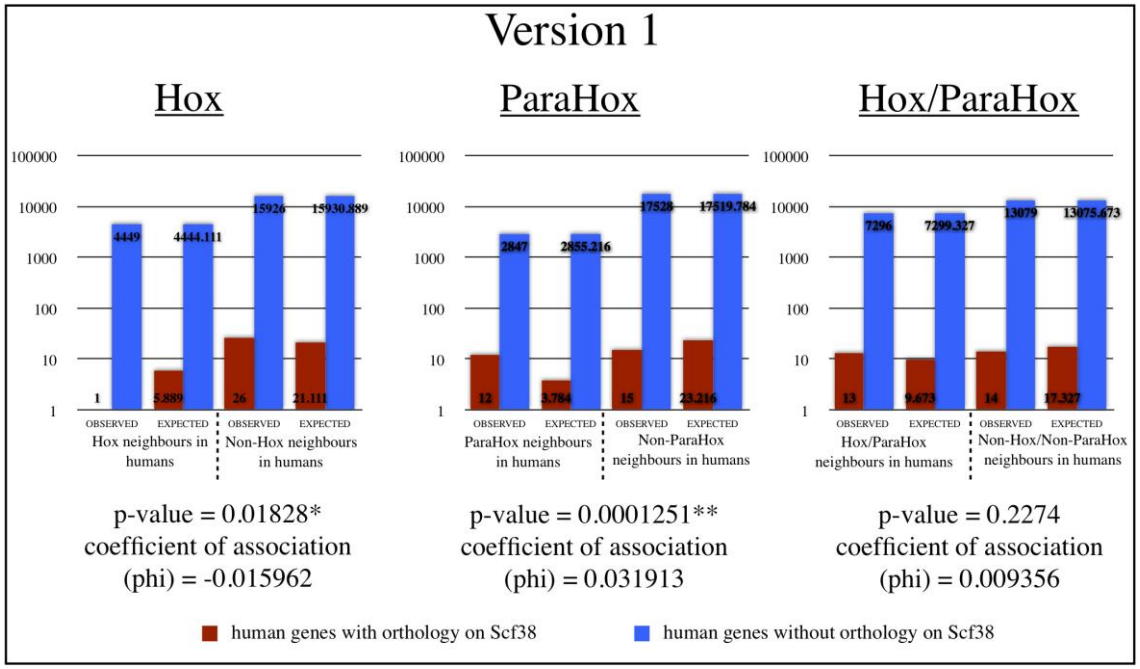


Figure S1D. Fisher's Exact Tests (Version 1), Related to Figure 2

* statistically significant at the 5% level

** statistically significant at the 1% level

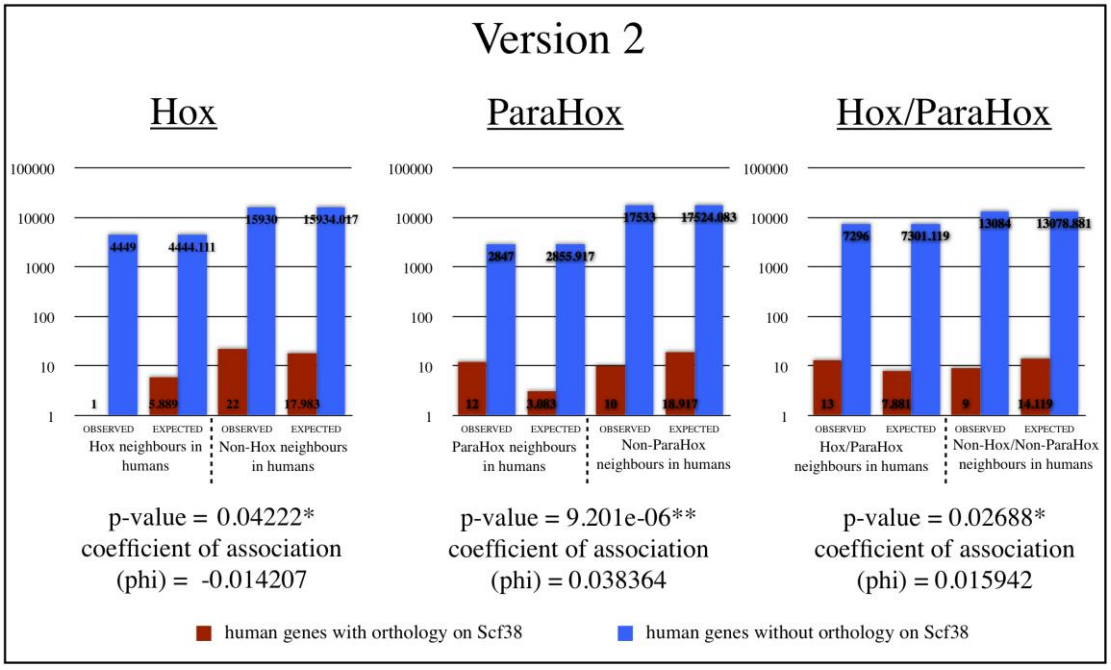


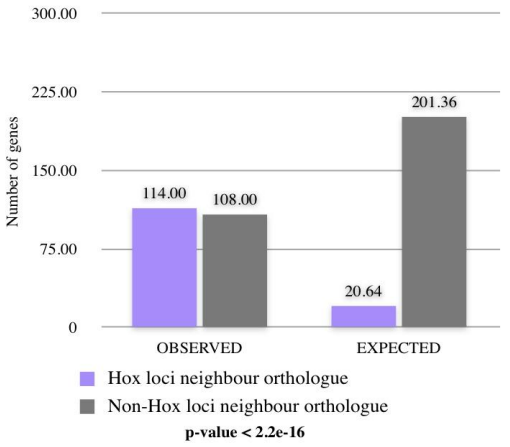
Figure S1E. Fisher's Exact Tests (Version 2), Related to Figure 2

* statistically significant at the 5% level

** statistically significant at the 1% level

Exact Binomial Test (Ghost Hox)

	OBSERVED	EXPECTED
Hox loci neighbour orthologue	114.00	20.64
Non-Hox loci neighbour orthologue	108.00	201.36



Probability of being a Hox loci neighbour in Scaffold 3 PSc3	0.092968750
Probability of not Hox loci neighbour in Scaffold 3 QSc3	0.907031250
Total	1

Figure S1F. Ghost Hox Exact Binomial Test, Related to Figure 2

** statistically significant at the 1% level

A

	Capitella teleta	*Lottia gigantea*
Probability of a gene being Hox scaffold P_H	0.003208391176924	0.015093706762819
Probability of a gene being ParaHox scaffold P_{PH}	0.000863797624557	0.007001802859419
Probability of a gene being NK scaffold P_{NK}	0.007373129723893	0.052240996184646

B

Capitella teleta	scaffold	capacity	neighbouring orthologues
NK p-value = 2.501e-5 **	815	14	0
	493	16	0
	315	26	2
	725	20	0
	95	63	2
	33020	1	0
	31	89	3
	694	10	2
ParaHox p-value = 1	760	10	0
	444	18	0
Hox p-value = 0.4809	33	62	1
	70	29	0
	292	13	0

Lottia gigantea	scaffold	capacity	neighbouring orthologues
NK p-value = 2.3e-10 **	122	44	0
	19	277	9
	72	97	0
	40	168	4
	88	88	5
	21	245	8
	9	321	9
	263	6	0
ParaHox p-value = 0.0508	85	82	3
	80	85	1
Hox p-value = 0.3799	12	360	1

** statistically significant at the 1% level

C

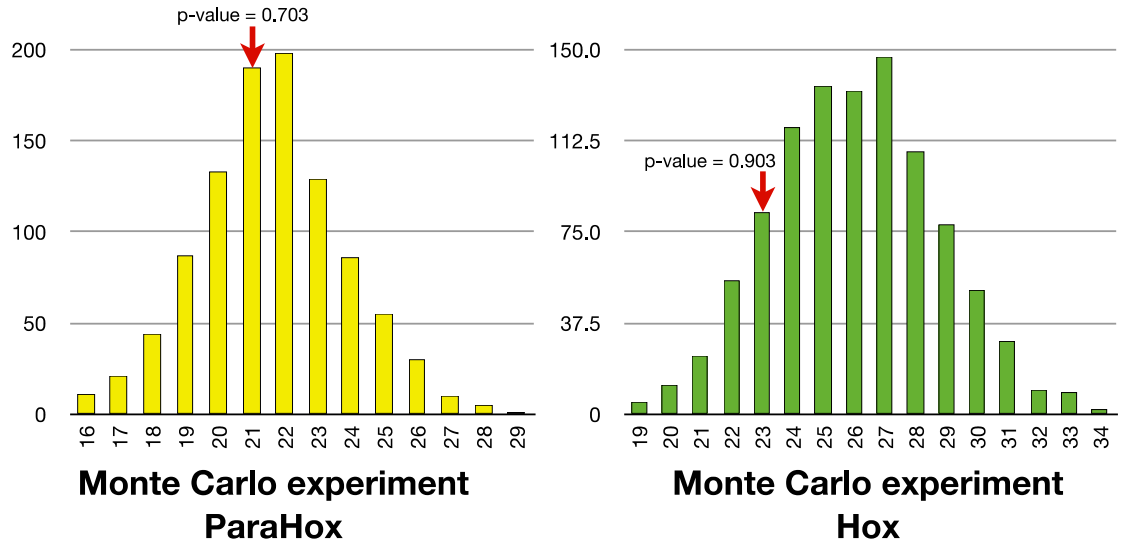


Figure S2. Related to Figure 3

(A) Probabilities of a gene being in Hox, ParaHox and NK scaffolds in *Capitella teleta* and *Lottia gigantea* genomes.

(B) Summary of gene numbers of Hox, ParaHox and NK-bearing scaffolds and p-values of Binomial Exact Tests in *Capitella teleta* and *Lottia gigantea* genomes.

(C) Histograms of the Monte Carlo experiments of Hox and ParaHox PAL genes found in *Monosiga brevicollis*. Simulation of randomized location of *M. brevicollis* orthologues of bilaterian-cnidarian Hox neighbours (green bins) and orthologues of placozoan ParaHox neighbours (yellow bins) across the *M. brevicollis* scaffolds. Red arrow indicates observed number of scaffolds with Hox/ParaHox neighbour orthologues in *M. brevicollis*.

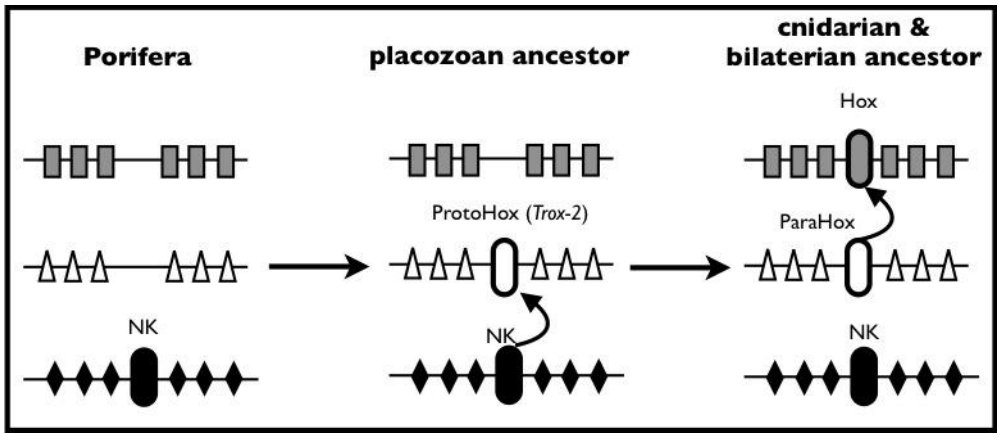


Figure S3. Less Parsimonious Alternative to the Ghost Locus Hypothesis, Related to Figure 4

Summary of duplication events occurring via retrotransposition or DNA-based transposition in the basal animal lineages. In the placozoan ancestor the duplication of the NK gene(s) via a transposition event that did not include non-homeobox neighbour genes gives rise to the ProtoHox (of which *Trox-2* is a direct descendant). In the cnidarian and bilaterian ancestor the ProtoHox duplicates via another transposition event that did not include neighbouring genes. One of the copies evolves into ParaHox gene(s) and the other gives rise to the Hox gene(s). Note, this scenario also requires asymmetrical evolution after the ProtoHox/*Trox-2* state, such that the ParaHox descendant gene Gsx retains greater similarity with the ProtoHox/*Trox-2* gene than do any other descendant Hox and ParaHox genes.

Supplemental Experimental Procedures with further Results and Discussion

Identification of Orthologues in *Trox-2* Scaffold (Scaffold 38) of *Trichoplax adhaerens*

Trichoplax adhaerens scaffold 38 (GenBank accession number: DS985276.1) contains 37 genes besides *Trox-2*. To identify human orthologues we performed reciprocal BLASTp searches against the human genome (build 37 patch 2). This helped to establish whether each *T. adhaerens* gene had a one-to-one, one-to-many or many-to-many relationship with human genes, or no orthology at all (i.e. *Trichoplax* specific gene). *T. adhaerens* protein sequences and their candidate human orthologues with their respective family members (if they had them) were aligned using MAFFT (v6.846b --einsi --Blossum62 --ep 0.0) and viewed in Jalview (version 2.6.1) to edit alignments for phylogenetic tree building. Alignment editing was refined by cross-comparison with multiple alignments built by GBLOCKS (http://molevol.cmima.csic.es/castresana/Gblocks_server.html, default settings). In cases without *T. adhaerens* or human family members or duplicates (i.e. putative one-to-one relationship), orthologue sequences of other chordates and *Nematostella vectensis* were included to help identify conserved domains and motifs and *T. adhaerens* gene identity. SMART (smart.embl-heidelberg.de) was used to help confirm these conserved domains and motifs (SMART_ACC.xls is available on http://biology.st-andrews.ac.uk/cegg/downloads.aspx). In cases with family members (one-to-many or many-to-many), phylogenetic trees were constructed using Modelgenerator (v0.85) followed by Neighbour Joining in PHYLIP (v3.69), Maximum Likelihood in PhyML (v3.0) and Bayesian Markov chain Monte Carlo in MrBayes (v3.1.2). Node support for NJ trees was estimated from 1000 bootstrap replicates. Node support for ML were estimated from 100 bootstrap replicates and for Bayesian trees we used 1000000 generations; 5000 for sample frequency; burn-in of 50; two runs of four chains each. This tree building helped to resolve some of the one-to-many and many-to-many relationships as one-to-one. Orthologous gene locations in the human genome were noted.

The details of each *T. adhaerens* scaffold 38 gene are as follows:

1) TRIADDRAFT_62201 (Accession number: XP_002118187.1)

Reciprocal BLASTp searches show that this protein has similarity to the human kinesin-3 family. According to the current classification based on the Kinesin motor (KISc), human kinesins are comprised of 14 families (plus a collection of 'orphan' genes), divided into 28 subfamilies [1]. The human kinesin-3 family is composed of eight members (KIF16B, StarD9, KIF1A, KIF1B, KIF1C, KIF13A, KIF13B and KIF14). Apart from the KISc motif, the FHA motif is characteristic of this family [1]. *T. adhaerens* also contains four further kinesin-3 family sequences. A multiple alignment with all human kinesin genes and the putative *T. adhaerens* kinesin-3 genes showed no obvious affinity of TRIADDRAFT_62201 with a particular human kinesin subfamily. Moreover, the SMART analysis showed that TRIADDRAFT_62201 has no FHA motif, consistent with its very divergent nature. A neighbour-joining tree (JTT, 1000 bootstraps) showed that this protein is a divergent member of the kinesin family. Due to the divergent nature of TRIADDRAFT_62201 we discarded it from the synteny analysis.

2) TRIADDRAFT_62202 (Accession number: XP_002118164.1)

Reciprocal BLASTp searches show that this protein is a putative orthologue to human pericentriolar material 1 or PCM1. BLASTp searches using the human PCM1 and TRIADDRAFT_62202 sequences against their own genomes revealed no other family members. Chordate PCM1 sequences have a GTP/ATP binding site motif with the consensus [A,G]-X4-G-K-[S,T] and various motifs rich in aspartic acid and glutamic acid (EDDEx6AEx3, DEx6QD and

EDENEDEEMEEFEE) [2]. The *Trichoplax* orthologue does not show any of these motifs but does have extensive sequence similarity at the C-terminus end (data not shown), which is also the case for the cnidarian putative PCM1 sequences from *Nematostella vectensis* and *Hydra magnipapillata*. Hence, we named this protein *Tad_PCM1*, and include this protein in the synteny analysis as a "one-to-one" orthologue relationship.

3) TRIADDRAFT_62203 (Accession number: XP_002118165.1)

The results from the reciprocal BLASTp searches indicate that this protein has no significant match with any human protein. Consequently this protein is not informative for synteny analysis.

4) TRIADDRAFT_33759 (Accession number: XP_002118188.1)

Reciprocal BLASTp searches indicate that this protein is a putative orthologue of human Torsin 1A. The human torsin family is composed of five members: TORSIN 1A, TORSIN 1B, TORSIN 2A, TORSIN3A and C9orf167. Also, the reciprocal BLASTp searches indicated another putative *Trichoplax* torsin, TRIADDRAFT_58752. The torsin family belongs to the superfamily AAA+. The torsins have four short motifs: Walker A, Walker B, SN, sensor IV. These motifs are all present in the *T. adhaerens* sequences. The ClpB heat shock protein family is closely related to the torsins [3,4]. Torsins and Clpbs are characterized by six conserved cysteines. In Torsin sequences the cysteine closest to the C-terminus is embedded in the motif GCK. In ClpB sequences the sequence is instead GAR [3,4]. A molecular phylogenetic analysis, including some ClpB genes as an outgroup, shows that TRIADDRAFT_58752 and TRIADDRAFT_33759 form a sister group to the torsins of humans and other animals. We thus classify the orthologue relationship as "many-to-many" and accommodate this in the statistical analyses as described below.

5) TRIADDRAFT_64406 (Accession number: XP_002118166.1)

Reciprocal BLASTp searches indicate that this protein is a putative orthologue to human neurochondrin. Neurochondrin is a leucine-rich protein [5]. No further family members in *T. adhaerens* or in human were found. The multiple alignment shows extensive conservation of leucine-rich motifs in TRIADDRAFT_64406 (data not shown), confirmed by SMART. Hence, we named this protein *Tad_NCDN*, and include this protein in the synteny analysis as a "one-to-one" orthologue relationship.

6) TRIADDRAFT_9204 (Accession number: XP_002118167.1)

Reciprocal BLASTp searches indicate that this protein is a putative orthologue to human matrilins, human fibrillins and human fibulins. These proteins share the following domains: epidermal growth factor-like domain (EGF), calcium-binding EGF-like domain (EGF_CA) and von Willebrand factor type A domain (VWA). The distinction amongst these families is by a characteristic combination of these domains [6-12]. A multiple alignment of TRIADDRAFT_9204 with human matrillins, fibulins and fibrillins showed that single EGF and EGF_CA motifs are present in TRIADDRAFT_9204. This result was confirmed by SMART analysis. This *T. adhaerens* sequence is thus relatively short, with very few motifs, and its classification is consequently poorly resolved. Hence, we discarded this protein from the analysis.

7) TRIADDRAFT_51183 (Accession number: XP_002118168.1)

Reciprocal BLASTp searches identify TRIADDRAFFT_51183 as a putative orthologue to human Hydroxysteroid (17-beta) dehydrogenase 10. Other family members were retrieved

from the *Trichoplax* and human genomes. The hydroxysteroid 17-beta dehydrogenase (HSD17B) family belongs to the short chain dehydrogenase/reductase (SDR) superfamily. The HSD17B family shares several amino acid sequence motifs with the SDR superfamily: TGXXXGXG (part of the Rossman fold), NAG (structural stabilization), YXXK (active centre) and PGXXXT (C-terminal to active site) [13-16]. The multiple alignment and SMART analysis showed the conservation of these motifs in TRIADDRAFFT_51183. To clarify specific orthology and paralogy relationships a phylogenetic analysis was done. For this analysis the rest of the members of the HSD17B family were used as an outgroup. From this analysis we identified a one-to-one relationship between TRIADDRAFFT_51183 and HSD17B10. Also, we identified a recent paralogue of this particular *T. adhaerens* sequence (TRIADDRAFT_22420) also orthologous to human HSD17B10, indicating a possible lineage-specific duplication in *T. adhaerens*. Hence, we named this protein *Tad_HSD17B10A*, and include this protein in the synteny analysis as a "one-to-one" orthologue relationship.

8) TRIADDRAFT_33760 (Accession number: XP_002118168.1)

Reciprocal BLASTp searches indicate that this protein is a putative orthologue to a human fibrillin or fibulin protein. These two families share domains (see TRIADDRAFT_9204 above). A multiple alignment of TRIADDRAFT_33760 with human fibulins and human fibrillins, along with a SMART analysis, revealed conservation of some motifs. However, the TRIADDRAFT_33760 gene model is very short, which contributes to it lacking a clear affinity to a particular human gene(s). Therefore, due to this difficulty in identifying TRIADDRAFT_33760 orthology, we discarded this gene from the synteny analysis.

9) TRIADDRAFT_33711 (Accession number: XP_002118189.1)

Reciprocal BLASTp searches indicate that TRIADDRAFT_33711 is a putative orthologue to human vacuolar protein sorting 36 (VPS36). No further family members were found in the *T. adhaerens* or human genomes. The human protein is characterised by a split pleckstrin-homology domain Φ XKX(G/A/S/P)X…(K/R)…X(R/K)XRX(F/L) also known as the glue domain [17,18]. Multiple alignment of TRIADDRAFT_33711 with chordate orthologues of VPS36, and SMART analysis, showed that TRIADDRAFT_33711 also conserves this motif. Therefore, we named this protein *Tad_VPS36*, and include this protein in the synteny analysis as a "one-to-one" orthologue relationship.

10) TRIADDRAFT_33740 (Accession number: XP_002118170.1)

Reciprocal BLASTp searches indicate that TRIADDRAFT_33740 is a putative orthologue of human Aminoadipate-semialdehyde-dehydrogenase-phosphopantetheinyl transferase (AASDHPPT). No further family members were found in the *T. adhaerens* and human genomes. Human ASSDHPPT is characterised by the phosphopantetheinyl transferase motif GXD…E…(W/F/L)XX(K/R)E(A/S)XXK [19]. Multiple alignment of TRIADDRAFT_33740 with several other chordate orthologues, along with SMART analysis, showed the conservation of the phosphopantetheinyl transferase motif in TRIADDRAFT_33740. Therefore, we named this protein *Tad_AASDHPPT*, and include this protein in the synteny analysis as a "one-to-one" orthologue relationship.

11) TRIADDRAFT_33763 (Accession number: XP_002118190.1)

Reciprocal BLASTp searches indicate that TRAIDDRAFT_33763 is a putative orthologue of human Apoptosis Induction Factors (AIFM's). The human AIFM family contains three members, characterised by a FAD or NAD(P) binding Rossmann motif ((V/I)XGX(1-2)GXXGXXX(G/A)) [20]. A second member of this family was retrieved from the *T. adhaerens*

genome (TRIADRAFT_59728). Multiple alignment and SMART analysis of the human and *T. adhaerens* sequences showed conservation of the Rossman motif. Orthology and paralogy relationships were investigated with molecular phylogenetic analyses. This identified a one-to-one orthology relationship of TRAIDDRAFT_33763 with human AIFM1 (and a one-to-one orthology relationship between TRIADDRAFT_59728 and human AIFM3). Therefore we included TRAIDDRAFT_33763 in the synteny analysis.

12) TRIADDRAFT_33726 (Accession number: XP_002118171.1)

Reciprocal BLASTp searches indicate that TRIADDRAFT_33726 is a putative orthologue to human tumor suppressor candidate 3 (TUSC3) and magnesium transporter protein (MAGT1). Both human proteins share the thioredoxin-like motif (CXXC) and oligosaccharyl transferase motif [21]. A multiple alignment of the human proteins, other chordate orthologues and TRIADDRAFT_33726, along with a SMART analysis, confirmed conservation of the motifs. Phylogenetic analysis showed that TRIADDRAFT_33726 is a proto-orthologue to both chordate proteins.

13) TRIADDRAFT_62215 (Accession number: XP_002118191.1)

Reciprocal BLASTp searches indicate that TRIADDRAFT_62215 is a putative orthologue of a human G-protein couple receptor (GPCR). The current consensus based on the common functional unit, the seven α-helical transmembrane motif (7TM) divides the human GPCRs superfamily into five classes (Glutamate, Rhodopsin, Adhesion, Frizzled/Taste2 and Secretin (GRAFS)) [22]. Beyond this basic level of classification more fine-scale affinities are very poorly resolved in phylogenetic trees. Consistent with this no unambiguous orthology of TRIADDRAFT_62215 with any particular human gene(s) can be determined. Therefore we discarded this protein from the synteny analysis.

14) TRIADDRAFT_62216 (Accession number: XP_002118172.1)

Reciprocal BLASTp searches indicate that this hypothetical protein is also a putative GPCR. For the same reasons as outlined for TRIADDRAFT_62215 we discarded this protein from the synteny analysis.

15) TRIADDRAFT_62217 (Accession number: XP_002118173.1)

Reciprocal BLASTp searches indicate that TRIADDRAFT_62217 is a putative orthologue of human thioredoxin. Thioredoxins are characterized by a cysteine-rich sequence motif (W-C-G-P-C-K and three cysteines after the previous motif). The thioredoxin superfamily is divided into families by a common thioredoxin fold encoded by the two residues in between of the two cysteines of the actives sites (thioredoxin: C-G-P-C; glutaredoxin: C-P-T-C; DbsA: C-P-H-C) [23-25]. Further paralogues of TRIADDRAFT_62217 were retrieved from the *T. adhaerens* genome. Surprisingly three out of the four *T. adhaerens* family members were in scaffold 38 (TRIADDRAFT_62226 and TRIADDRAFT_62227). Multiple alignment of these three proteins demonstrated that they are unlikely to be recent duplicates as there are extensive differences between the sequences (data not shown). Multiple alignment of the four putative *T. adhaerens* thioredoxins and the human thioredoxins showed the conserved motif W-C-G-P-C-K, but outside this motif no cysteine conservation was observed (data not shown). Phylogenetic analyses with the entire coding sequences do not reveal a clear orthologue identification. Therefore, we excluded these proteins from the synteny analysis.

16) TRIADDRAFT_33746 (Accession number: XP_002118192.1)

The results from the reciprocal BLASTp searches indicated that this hypothetical protein is an orthologue to human Yip family member 6 (YIPF6). The human Yip family is composed of seven members and characterized by the motif DLYGP and GY [26,27]. Further members of the human Yip family as well as four putative members of this family from the *T. adhaerens* genome were retrieved. One of the *T. adhaerens* genes is also in scaffold 38 (TRIADDRAFT_5826). The SMART analysis confirmed the presence of conserved motifs. The molecular phylogenetic analysis helped identify a "one-to-one" orthologue relationship between human YIPF6 and TRIADDRAFT_33746, which we named *Tad_YIPF6,* and a "one-to-many" orthologue relationship between TRIADDRAFT_5826 (which we named *Tad_YIPF5/7*) and human YIPF5 and YIPF7. Therefore, we included both proteins in the synteny analysis.

17) TRIADDRAFT_33724 (Accession number: XP_002118174.1)

The results from the reciprocal best-hit BLASTp search indicated that this hypothetical protein is a putative orthologue to Glucosamine-6-phosphate deaminase 2 (GNPDA2). The human GNPDA family is composed of two members (GNPDA1 and GNPDA2) and the motif that characterizes this family is (L/I/V/M)3XGX(L/I/T)X(L/I/V/M)XG(L/I/V/M)GX(D/E/I)3XGX(I)X(L)X(V)XG(I)GX(D)H [28,29]. No further family members were found in the *T. adhaerens* genome. The multiple alignment with TRIADDRAFT_33724 and other chordate orthologues, along with SMART analysis, confirmed conservation of the family-characterizing motif. The phylogenetic analysis helped identify a "one-to-many" orthologue relationship between TRIADDRAFT_33724, GNPDA1 and GNPDA2. Therefore, we named this protein *Tad_GNPDA* and included it in the synteny analysis.

18) TRIADDRAFT_62220 (Accession number: XP_002118175.1)

The reciprocal best hit BLASTp search indicated that this hypothetical protein is a putative orthologue of the human FERM and PDZ domain containing proteins: collectively the FRMPDs. There are four FRMPDs (FRMPD1, FRMPD2, FRMPD3 and FRMPD4) in the human genome and they are distinguished by the order of appearance of the FERM and PDZ domains in their sequences [30]. Another putative FRMPD member in the *T. adhaerens* genome was found. The multiple alignment and SMART analysis of TRIADRAFT_62220 with human FRMPD domain containing proteins showed conservation of the FERM domain and conserved tryptophan motifs. Phylogenetic analysis helped to identify a "one-to-many" orthologue relationship between TRIADDRAFT_62220 and human FRMPD1, FRMPD3 and FRMPD4. The other *T. adhaerens* sequence, TRIADDRAFT_64201 showed affinity with human FRMPD2. Hence, we named TRIADDRAFT_62220 *Tad_FRMPD1/3/4* and included it in the synteny analysis.

19) TRIADDRAFT_62221 (Accession number: XP_002118193.1)

The reciprocal best-hit BLASTp searches indicated that this hypothetical protein is a putative orthologue to the BTB/POZ domain containing proteins. The human genome possesses 16 BTB/POZ domain containing proteins. No further BTB/POZ domain containing proteins were found in the *T. adhaerens* genome. The multiple alignment of TRIADDRAFT_62221 with the human BTB/POZ domain containing proteins showed affinity with the human BTB/POZ domain containing protein 12 [31]. SMART analysis confirmed the conserved motifs. Molecular phylogenetic analysis confirmed a "one-to-one" orthologue relationship of TRIADDRAFT_62221 with human BTB/POZ domain containing protein 12. Hence, we named this protein *Tad_BTB/POZ12* and included it in the synteny analysis.

20) TRIADDRAFT_62222 (Accession number: XP_002118176.1)

A GPCR, which was excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

21) TRIADDRAFT_62223 (Accession number: XP_002118194.1)

The reciprocal best-hit BLASTp search indicated that this protein is a putative orthologue to members of the human intracellular membrane-associated calcium independent phospholipase (PNPLA) family. The patatin domain characterizes the human PNPLA family, which is composed of nine members [32]. The human PNPLA genes differ in their motif content besides the patatin domain. Further putative *T. adhaerens* PNPLA genes were retrieved. The patatin motif is conserved in TRIADDRAFT_62223 according to the SMART analysis, and a multiple alignment showed affinity with human PNPLA8 and PNPLA9. However, molecular phylogenetic analysis revealed that TRIADDRAFT_62223 does not have a clear affinity with any particular human PNPLA gene, whilst other *T. adhaerens* PNPLA genes do have clearer orthologue relationships. Therefore, we excluded it from the synteny analysis analysis.

22) TRIADDRAFT_62224 (Accession number: XP_002118177.1)

The results from the reciprocal BLASTp searches indicated that this protein does not have a significant match with any human protein, and so it is excluded from the synteny analysis.

23) TRIADDRAFT_33732 (Accession number: XP_002118195.1)

The results from the best-hit reciprocal BLASTp searches indicated that this hypothetical protein is a putative orthologue of human Chloride channel proteins. The human Chloride channel protein family is composed of seven members, characterized by seven very well conserved transmembrane helices [33]. Further members of this family were retrieved from the *T. adhaerens* genome. The multiple alignment and SMART analysis showed conservation of the transmembrane helices in TRIADDRAFT_33732 and affinity for human CLCN3, 4, 5 genes. Molecular phylogenetic analysis helped identify a "one-to-many" orthologue relationship between TRIADDRAFT_33732 and CLCN3, CLCN4 and CLCN5. Therefore, we named this protein *Tad_CLCN3/4/5* and included it in the synteny analysis.

24) TRIADDRAFT_62226 (Accession number: XP_002118196.1)

A putative thioredoxin, excluded from the analysis as discussed for TRIADDRAFT_62217.

25) TRIADDRAFT_62227 (Accession number: XP_002118197.1)

A putative thioredoxin, excluded from the analysis as discussed for TRIADDRAFT_62217.

26) TRIADDRAFT_5826 (Accession number: XP_002118178.1)

Discussed in 16) TRIADDRAFT_33746

27) TRIADDRAFT_62229 (Accession number: XP_002118198.1)

Reciprocal BLASTp searches indicated that this protein has no significant match with any human protein, and so we exclude it from the synteny analysis.

28) TRIADDRAFT_62230 (Accession number: XP_002118179.1)

A GPCR, which was excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

29) TRIADDRAFT_7464 (Accession number: XP_002118199.1)

A GPCR, which was excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

30) TRIADDRAFT_5463 (Accession number: XP_002118200.1)

A GPCR, which was excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

31) TRIADDRAFT_64407 (Accession number: XP_002118180.1)

The BLASTp search indicated no BLAST hits at all.

32) TRIADDRAFT_62233 (Accession number: XP_002118181.1)

The results from the best-hit reciprocal BLASTp searches indicated that this hypothetical protein is a putative orthologue of human sterol regulatory element-binding transcription factors (SREBF1). This family belongs to a higher-order group B of the basic helix-loop-helix (bHLH) superfamily [34]. The human sterol regulatory element-binding transcription factors family is composed of two members and as for other bHLH superfamily members is characterized by a DNA-binding basic region followed by two α-helices. No further members of this family were retrieved from the *T. adhaerens* genome. The multiple alignment showed conservation of the α-helices and DNA-binding basic region and affinity for the orthologues of SREBF1. The SMART analysis confirmed the conserved motifs. Molecular phylogenetic analysis helped identify a "one-to-many" orthologue relationship between TRIADDRAFT_62233 and human SREBF1 and SREBF2. Thus, we named this protein *Tad_SREBF1/2* and included it in the synteny analysis.

33) TRIADDRAFT_33728 (Accession number: XP_002118201.1)

Trox-2

34) TRIADDRAFT_62235 (Accession number: XP_002118182.1)

A GPCR, which was excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

35) TRIADDRAFT_62236 (Accession number: XP_002118183.1)

A GPCR, which was excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

36) TRIADDRAFT_62237 (Accession number: XP_002118184.1)

A GPCR, which was excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

37) TRIADDRAFT_62238 (Accession number: XP_002118185.1)

A GPCR, which was excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

38) TRIADDRAFT_62239 (Accession number: XP_002118186.1)

A GPCR, which was excluded from the synteny analysis, as discussed for TRIADDRAFT_62215.

For a summary of the orthologue identification see Table S1.

Statistical Significance of the Observed Synteny Conservation of *Trox-2* Scaffold

After identifying *T. adhaerens*-human orthologues we classified them into Hox loci neighbour orthologues, ParaHox loci neighbour orthologues and Non-Hox/ParaHox loci neighbour orthologues. Hox loci neighbour orthologues are those *T. adhaerens* genes with human orthologues located on any of the human chromosomes bearing a Hox cluster (Chromosomes 2, 7, 12 and 17). ParaHox loci neighbour orthologues are those *T. adhaerens* genes with human

orthologues located on any of the human chromosomes bearing ParaHox loci (Chromosomes 4, 5, 13 and X). Non-Hox/ParaHox orthologues are those *T. adhaerens* genes with human orthologues located on chromosomes other than 2, 7, 12, 17, 4, 5, 13 or X. Also, we performed two sets of tests to accommodate tandem or segmental duplications on the human lineage which result in co-linkage of multiple members of a particular gene family. One version included the single location of each of the human orthologues and the second version included the collapsed location of the human paralogues (e.g., in the case of the torsins four out of the five members are located on human chromosome 9, and in this case we counted just one location in chromosome 9 within the second set of tests; see Table S1).

The observed synteny conservation was statistically tested with two tests: Exact Binomial test and Fisher's Exact test. These two tests were conducted in R (codes are available on request). The numbers used in our tests are based on human genome version 37 patch 2 and are derived as follows:

1) The total number of protein-coding genes (pcg) in chromosomes 1 to 23 and X (20447 pcg, Table S4 C12). From this number we subtracted the number of protein-coding genes in the Hox clusters (39 pcg = C2) and ParaHox 'clusters' (6 pcg = C6), leaving a total number of protein-coding genes without Hox and ParaHox genes (20402 pcg = C11).
2) We made the distinction of type of orthologues according to their location in the human genome. Hox loci neighbours (4450 pcg = C3) include the total number of protein-coding genes on chromosomes 2 (1275 pcg), 7 (942 pcg), 12 (1055 pcg) and 17 (1217 pcg), excluding the genes from the Hox clusters (39 pcg). ParaHox loci neighbours (2859 pcg = C7) include the total number of protein-coding genes on chromosomes 4 (781 pcg), 5 (899 pcg), 13 (333 pcg) and X (852 pcg) excluding the genes from the ParaHox 'clusters' (6 pcg). Non-Hox loci neighbours (15952 pcg = C4) are those excluding the Hox loci neighbours (4450 pcg), Hox clusters (39 pcg) and ParaHox 'clusters' (6 pcg) from the total number of protein-coding genes, and non-ParaHox loci neighbours (17543 pcg = C8) are those excluding the ParaHox loci neighbours (2859 pcg) from the total number of protein-coding genes without Hox and ParaHox genes (20402 pcg). Hox/ParaHox loci neighbours (7309 pcg = C10) are the sum of Hox loci neighbours (4450 pcg) and ParaHox loci neighbours (2859 pcg), and non-Hox/ParaHox loci neighbours (13093 pcg = C9) are those excluding ParaHox (2859 pcg) and Hox (4450 pcg) loci neighbours as well as Hox (39 pcg) and ParaHox (6 pcg) from the total number of protein-coding genes (20447 pcg).

These numbers are summarized as follows.

C1= number of genes in Hox chromosomes = 4480

C2 = number of genes in the Hox clusters = 39

C3 = number of genes that are Hox loci neighbours = 4450

C4 = number of genes that are non-Hox loci neighbours = 15997

C5 = number of genes in ParaHox chromosomes = 2865

C6 = number of genes in the ParaHox 'clusters' = 6

C7 = number of genes that are ParaHox loci neighbours = 2859

C8 = number of genes that are non-ParaHox loci neighbours = 17588

C9 = number of genes in non-(Hox/ParaHox) loci neighbours = 13093

C10 = number of genes in Hox/ParaHox loci neighbours = 7309

C11 = total number of genes in genome minus Hox and ParaHox clusters = 20402

C12 = total number of genes in genome = 20447

From these numbers we calculated the probabilities of a randomly chosen human gene being a Hox locus neighbour, ParaHox locus neighbour and Non-Hox/ParaHox neighbour. These probabilities are used to perform the Binomial Exact Test, with the following values:

P1 = Probability of being a Hox locus neighbour Ph = 0.217635839 (= C3/C11)

P2 = Probability of not being a Hox locus neighbour Qh = 0.782364161 (= C4/C11)

P3 = Probability of being a ParaHox locus neighbour Pph = 0.139824913 (= C7/C11)

P4 = Probability of not being a ParaHox locus neighbour Qph = 0.860175087 (=C8/C11)

P5 = Probability of being a Hox/ParaHox neighbour Pnhph = 0.357460752 (=C10/C11)

P6 = Probability of being a non-Hox/ParaHox neighbour Qnhph = 0.641750809 (=C9/C11)

The Exact Binomial Test was used to test departure of observed numbers of Hox neighbour orthologues (or ParaHox neighbour orthologues or Hox/ParaHox neighbour orthologues) on scaffold 38 from those expected on the basis of the frequency of Hox neighbours (or ParaHox neighbours or Hox/ParaHox neighbours) in the human genome. We plotted the observed and expected number of genes in scaffold 38 for each one of the tests. For all the plots the expected number of orthologues is calculated by multiplying the total observed number of orthologues (i.e. 27 genes in version 1 and 22 genes in version 2) by the category probabilities (P1 – P6 above) (Fig. S1A-C).

Fisher's exact test is used to test whether there is a statistically significant association between the apparent concentration of human orthologues in *T. adhaerens* scaffold 38 and the human ParaHox loci neighbours or human Hox loci neighbours or human Hox/ParaHox loci neighbours. Also, we calculated the coefficient of association, which measures the directionality of this association [35]. We represented the contingency tables (see Table S1) used for computing the statistical test in bar charts, comparing the number of human orthologues observed with the number of human orthologues expected under the null hypothesis, for each one of the tests (Fig. S1D and Fig. S1E). The expected numbers of human orthologues were calculated for a particular cell by multiplying its row by its column totals and dividing the product by the grand total [35].

Identification of orthologues and Synteny Analysis of Scaffold 3 of *Trichoplax adhaerens* Genome

Since *T. adhaerens* has a ParaHox locus with a ParaHox gene, we wanted to test whether there is a Hox locus in *T. adhaerens* but lacking a Hox gene, that is a 'ghost' Hox locus. We used the Hox Putative Ancestral Linkage (PAL) gene list from *N. vectensis* [36]. The Hox PAL gene list accommodates orthologues into groups that have conserved linkage across bilaterian Hox-bearing chromosomes and *N. vectensis* scaffolds. We use this gene list to do BLASTp searches against the *T. adhaerens* genome, using the reciprocal best-hit criteria to compile the list of *Trichoplax* orthologues that could be part of the bilaterian-cnidarian-placozoan (BCP) Hox PAL

(see Table S2). Starting from 267 *N. vectensis* genes in the list we found 222 orthologues in *T. adhaerens*. From these 222 orthologues 114 are in *T. adhaerens* scaffold 3.

Statistical Significance of the Observed Synteny Conservation of Scaffold 3 of *Trichoplax adhaerens* Genome

In order to test whether the apparent concentration of Hox loci neighbour orthologues found in scaffold 3 of *T. adhaerens* is significantly different from a random distribution in the *T. adhaerens* genome, we performed an Exact Binomial test. This test was conducted in R (code is available on request). For this test we calculated the probability of a gene being in scaffold 3 of *T. adhaerens*, which is the number of genes annotated in scaffold 3 (1071) divided by the total number of genes annotated in all *T. adhaerens* scaffolds (11520). The probability of a gene not being somewhere in scaffold 3 is one minus the probability of a gene being in scaffold 3 (Fig. S1F).

Identification of Orthologues in *Amphimedon queenslandica* Using the Bilaterian-Cnidarian-placozoan (BCP) Hox PAL Gene List

Using the same logic as we did for the *T. adhaerens* ghost Hox locus, we wanted to first find if there are orthologues of human Hox loci neighbours in the *A. queenslandica* genome and then infer whether these orthologues are clustered.

In order to accomplish the first aim we used the BCP Hox PAL gene list to do BLASTp searches against *A. queenlandica*'s genome. We followed the reciprocal best-hit criteria to find putative orthologues to the Hox loci neighbours in *A. queenslandica*. This produced a list of 187 *A. queenslandica* genes orthologous to the BCP Hox PAL genes deduced above (see Table S2).

Monte Carlo-Based Test for Synteny Conservation of the BCP Hox PAL Genes in the *A. queenslandica* Genome

The *A. queenslandica* genome is assembled to a subchromosomal level (i.e. scaffold level). In order to test whether there is clustering of the Hox neighbour orthologues in *A. queenslandica*, we obtained an empirical null distribution of the number of *A. queenslandica* scaffolds expected to be occupied by the 187 genes, in the absence of any conservation of synteny, based on 1000 simulations [37]. In each simulation, each of the 30,060 genes of *A. queenslandica* (including the 187 Hox neighbour orthologues) was allocated to a scaffold, with the scaffold randomly selected, with replacement, with probability proportional to its observed gene content. This simulated genome is then compared to the actual genome scaffold by scaffold. The comparison is between the Hox neighbour orthologues placed at random and the expected frequency of Hox neighbour orthologues for each scaffold. If the content of Hox neighbour orthologues in a scaffold from the simulated genome exceeds the expected frequency of Hox neighbour orthologues of that scaffold for that cycle the "exceeded frequency" would increase by one. This comparison was performed for all the simulated scaffolds. The cycle ends once this comparison is finished. Each cycle is stored in a relational database table called amphisimulation. This cycle is repeated 1,000 times. In practice the "exceeded frequency" always equalled the number of scaffolds occupied by one or more of the 187 genes.

The empirical *P* value for a one-tailed test of the alternative hypothesis of clustering versus the null hypothesis of no clustering may be calculated as the proportion of simulations in

which the number of scaffolds occupied by the 187 genes is less than or equal to the actual number observed. This simulation was implemented in Python (code available on http://biology.st-andrews.ac.uk/cegg/downloads.aspx).

Creation of *T. adhaerens* Localized-ParaHox PAL

There is no putative ancestral linkage gene list for the ParaHox loci in *N. vectensis*. This is due to the fact that the *N. vectensis* ParaHox synteny is more localized than the scale of analysis used by Putnam et al [36,38]. However, *T. adhaerens* scaffold 5 has significant synteny with the close, localized neighbourhoods of the human ParaHox genes (see Tables S8.2 and S8.3 in [39]). These close neighbourhoods were described by Srivastava et al [39] as chromosomal segments with particular coordinates (Chromosome 5, segment name 5.4, molecular coordinates 139835480-167951722; Chromosome X, segment name X.6, molecular coordinates 70406305-106924338; Chromosome 13, segment name 13.1, molecular coordinates 1-41837067; Chromosome 4, segment name 4.2, molecular coordinates 25986602-57101698; and see Table S2). Their annotation is dated for the version of the human genome corresponding to build 36. We checked whether the coordinates annotated for that genome build have changed in the current build used in this study (human build 37 patch 2). We confirmed that no relevant change had occurred and so used these segments to build up a localized-ParaHox PAL gene list from *T. adhaerens*. This localized-ParaHox PAL gene list was used to test for a ghost ParaHox locus in the *A. queenslandica* genome.

We constructed the localized-ParaHox PAL list as follows. First, we gathered the number of genes (the protein coding genes, pcg) for each human segment (Segment 5.4 has 210 pcg, Segment X.6 has 157 pcg, Segment 13.1 has 125 pcg and Segment 4.2 has 103 pcg). Second, with each gene of the human segments we did a BLASTp search against the *T. adhaerens* genome. We then applied a filter to the BLASTp search outputs, retaining a gene if it is a top hit and has a bit score greater than 70 and an e-value less than 10^{-10} and is also located in *T. adhaerens* scaffold 5. This resulted in 67 *T. adhaerens* genes from human segment 5.4 searches, 68 genes from segment X.6, 68 genes from 13.1 and 46 genes from 4.2 searches. These *T. adhaerens* genes were next used for BLASTp searches against the human genome, filtering the outputs for genes that were a top hit and had a bit score greater than 70 and an e-value less than 10^{-10} and were located in human chromosomal segments 5.4, X.6, 13.1 or 4.2. This resulted in 70 pairs of orthologues. Within these pairs were five GPCR pairs. We discarded these due to the ambiguity in their classification and the difficulty in assigning orthology with confidence, as discussed for TRIADDRAFT_62215 above, which left 65 gene pairs in our localized-ParaHox PAL (see Table S2).

It is noteworthy that scaffold 5 has the clear ParaHox neighbourhood synteny signal in the analyses of Srivastava et al [39], and not scaffold 38, which contains *Trox2*. This is because scaffold 38 is too small, with too few genes, to be included in the *T. adhaerens* synteny analysis of Srivastava et al [39]. We predict that *T. adhaerens* scaffold 5 and 38 are closely linked in the placozoan genome.

Identification of Orthologues in *A. queenslandica* Genome Using *T. adhaerens* Localized-ParaHox PAL (l-ParaHox PAL) Gene List

We used the same procedure as we did for finding the *T. adhaerens* ghost Hox locus, to first find if there are orthologues of human ParaHox neighbours in *A. queenslandica* and second deduce whether these orthologues are clustered. We found 44 l-ParaHox PAL in *A. queenslandica* genome (see Table S2).

Monte Carlo-Based Test for Synteny Conservation of the l-ParaHox PAL Genes in the *A. queenslandica* Genome

We performed the same simulations as for the Hox loci neighbours, but incorporating the number of ParaHox neighbour orthologues determined in the previous section.

Determining whether *A. queenslandica* Genome Has a Ghost ProtoHox Locus or Ghost Hox and ParaHox Loci

In order to infer whether the clustered Hox and ParaHox neighbour orthologues in *A. queenslandica* are coincident, as would be expected for a ProtoHox locus, or instead they are distinct, independent ghost loci, we used the output of both Hox and ParaHox simulations above. For each cycle of both experiments we recorded how many scaffolds had an overlap of at least one orthologue of a Hox neighbour and at least one orthologue of a ParaHox neighbour. The empirical *P* value for a test of the alternative hypothesis of clustering versus the null hypothesis of no clustering was calculated as the proportion of simulations in which the number of scaffolds with both kinds of orthologue was greater than or equal to the observation. This implies *A. queenslandica* has separate ghost Hox and ParaHox loci, as opposed to a ProtoHox condition which would have entailed the overlap of Hox and ParaHox neighbours occurring with a frequency in the upper tail of the empirical null distribution.

Synteny Analysis of NK loci of *A. queenslandica* and Statistical Significance of Observed Synteny

As a further test of whether the Hox and ParaHox loci are already distinct from the NK locus in *A. queenslandica* (as implied above) or instead the NK locus acted as the source of the ProtoHox/Hox/ParaHox loci (as inferred by Larroux et al [40]), we analysed the neighbouring genes of the NK cluster-bearing scaffold in *A. queenslandica* (scaffold 13506). We performed BLASTp searches against the lophotrochozoan genomes of *Capitella teleta* and *Lottia gigantea*. We did not use ecdysozoan genomes due to their extensive genome rearrangements, particularly with respect to the linkage patterns of the ANTP-class genes [41]. Also, vertebrate genomes cannot be used for this particular NK-versus-ParaHox/Hox linkage analysis because in vertebrates some NK clusters have become secondarily linked with some ParaHox loci. It is known that these linkages do not reflect the ancestral chordate condition from the data from amphioxus [41,42].

We used the reciprocal BLAST best-hit criteria to identify orthologues of the *A. queenslandica* NK cluster neighbours. We then determined which of these genes localised to either NK cluster gene-bearing scaffolds, Hox gene-bearing scaffolds, or ParaHox gene-bearing scaffolds in both *C. teleta* and *L. gigantea*. In *C. teleta* nine orthologues are located on NK-cluster gene scaffolds, which themselves have a total number of 239 genes (excluding the

homeobox genes themselves). In *L. gigantea* 35 orthologues are on NK-cluster gene scaffolds, which contain a total of 1,246 genes. For the ParaHox scaffolds *C. teleta* has zero orthologues of sponge NK neighbours from a total of 28 genes, whilst *L. gigantea* has four out of 167. For the Hox scaffolds *C. teleta* has one orthologue out of 104 genes, and *L. gigantea* has one orthologue out of 360 genes. We used an Exact Binomial Test to test whether this distribution of orthologues of sponge NK neighbours in lophocotrozoan NK, ParaHox and Hox scaffolds represents statistically significant synteny with the *A. queeslandica* NK cluster scaffold. We calculated the probability of a gene being on a Hox scaffold as the total number of annotated genes in the Hox scaffolds (*C. teleta* 104, *L. gigantea* 360), divided by the total number of annotated genes for the genome (*C. teleta* 32415, *L. gigantea* 23851). The probability of a gene being on a ParaHox scaffold is the total of the annotated genes on ParaHox scaffolds (*C. teleta* 28, *L. gigantea* 167), divided by the total number of annotated genes for the genome (*C. teleta* 32415, *L. gigantea* 23851). Finally, the probability of a gene being on an NK scaffold is the total of the annotated genes in the NK scaffolds (*C. teleta* 104, *L. gigantea* 360), divided by the total number of annotated genes for the genome (*C. teleta* 32415, *L. gigantea* 23851) (Fig. S2A). In order to test whether the apparent concentration of NK loci neighbours found in *C.teleta* and *L. gigantea* genomes are significantly similar to the one in the *A. queenslandica* NK bearing scaffold (Contig13506) we performed a Binomial Exact Test. Also, in order to test whether the apparent concentration of Hox and ParaHox loci neighbours found in *C. teleta* and *L. gigantea* are significantly not similar to the one in the *A. queenslandica* NK bearing scaffold we performed a Binomial Exact Test. The results from these tests showed that the probability of finding the apparent concentration of NK loci neighbour orthologues found in the NK bearing scaffolds of *L. gigantea* and *C. teleta* is like the one in the *A. queenslandica* NK bearing scaffold (Fig. S2B).

Identification of Orthologues in *Monosiga brevicollis* Using the Bilaterian-Cnidarian Hox PAL Gene List and Localized ParaHox PAL (l-ParaHox PAL) and *T. adhaerens* Scaffold 38 Gene List

We wanted to test whether the linkage of the Hox and ParaHox neighbour orthologues is exclusive to metazoans, as would be predicted from the complete lack of ANTP-class homeobox genes from non-metazoan lineages and if the ProtoHox condition evolved with the origin of the Metazoa. For this purpose we used the genome of *Monosiga brevicollis,* as a representative from the choanoflagellate sister group of metazoans [43]. Using the same logic as we did for the *T. adhaerens* and *A. queenslandica* ghost Hox loci, we wanted to first find if there are orthologues of bilaterian-cnidarian Hox loci neighbours in *M. brevicollis* and then infer whether these orthologues are clustered. Also, we wanted to find if there is a clustering of orthologues of the l-ParaHox PAL genes that we deduced from comparisons between *T. adhaerens, N. vectensis* and humans (see above). We also included a search for *Monosiga* orthologues for the genes in *T. adhaerens* scaffold 38, which contains the ParaHox gene *Trox-2*.

In order to accomplish the first aim we used the BC Hox PAL gene list to do BLASTp searches against the *M. brevicollis* genome. We followed the reciprocal best-hit criteria to find putative orthologues to the Hox loci neighbours in *M. brevicollis*. This produced a list of 139 *M. brevicollis* genes (see Table S2). Similarly the search for putative orthologues to the ParaHox loci neighbours in *M. brevicollis* produced a list of 52 *M. brevicollis* genes orthologous to l-ParaHox PAL (41 orthologues) and to *T.adhaerens* scaffold 38 (11 orthologues) (see Table S2).

Monte Carlo-Based Test for Synteny Conservation of the Hox and the ParaHox Loci Neighbours in the *M. brevicollis* Genome

We performed the same simulations as for the Hox and ParaHox loci neighbour analyses in *A. queenslandica*, but incorporating the number of Hox (139) and ParaHox (52) neighbour orthologues in *M. brevicollis*. Also, we used the total number of genes for *M. brevicollis*, 9196, and the total number of scaffolds, 218. We calculated the empirical value for a one-tailed test of the alternative hypothesis of clustering versus the null hypothesis of no clustering as the proportion of the simulation in which the number of scaffolds occupied by the 139 (Hox case) or 52 (ParaHox case) genes is less than or equal to the actual number observed.

The observed distribution of Hox and ParaHox neighbour orthologues in *M. brevicollis* does not differ from the null simulated distributions that represent random distributions of these genes across the choanoflagellate genome (Fig. S2C). This lack of clustering of these genes reveals that there are no ghost Hox and ParaHox loci in *M. brevicollis*. The Hox and ParaHox loci thus appear to be specific to the Metazoa, as expected.

Alternative to the Ghost Locus Hypothesis

Our conclusion that the 'ghost' Hox and ParaHox loci of Porifera and Placozoa reflect the loss of these homeobox genes from these lineages is the most parsimonious explanation for our data. We hypothesize that the ancestral ProtoHox locus, containing one or more homeobox genes along with a variety of neighbouring non-homeobox genes, duplicated to generate two loci that became the Hox and ParaHox loci. Evidence for the ProtoHox duplication being a relatively large-scale, multi-gene event comes from two sources. Firstly, both Hox and ParaHox clusters are flanked by collagen and tyrosine kinase receptor genes, which is taken as evidence that the ProtoHox duplication included the homeobox genes of the cluster as well as some neighbouring genes [44]. Secondly, Lanfear and Bromham [45] statistically tested the likelihoods of the alternative ProtoHox models (individual genes through to 2-, 3- and 4-gene clusters [46-50]), and found support for either of the 3- or 4-gene models. Since the Hox and ParaHox clusters are on separate chromosomes, this large, multi-gene duplication presumably happened in one of two ways.

The first possible route would be a whole genome (WGD) or whole chromosome duplication, such that the Hox and ParaHox clusters were on distinct chromosomes at their point of origin, and following this duplication extensive gene loss occurred along the daughter chromosomes such that the distinctive sets of Hox and ParaHox neighbours that we have characterized here remained. Such levels of gene loss following duplication are to be expected judging from the observation that less than 30% of paralogues (ohnologs) have been retained in humans after the 2R WGD at the origin of vertebrates [51], and there are ever increasing numbers of examples of polyploidy and WGD in animals as people look more closely [52-54]. Whole chromosome or whole genome duplication followed by extensive paralogue loss, such that the bulk of duplicated genes are returned to the single-copy state in a mutually exclusive pattern is thus not unusual.

The second possible route would be a large, multi-gene segmental duplication within a single chromosome. This intra-chromosomal event would then have been followed by translocation of one of the daughter homeobox clusters (along with neighbouring genes) to another chromosome, either via a chromosome arm exchange or via chromosome fission. These scenarios would have resulted in some non-homeobox neighbours from the ProtoHox locus

ending up with the descendant ParaHox locus and some with the descendant Hox locus. Either of these two routes would have resulted in the pattern of gene distribution summarized in Figure 1, and hence form the underlying logic for our analyses.

An alternative to our hypothesis would require that the duplication into Hox and ParaHox did not involve any ProtoHox neighbour genes. This could happen via a retrotransposition or a small-scale, inter-chromosomal DNA-based transposition (Fig. S3). We consider a retrotransposition event unlikely because such an event only involves a single coding sequence, and the ProtoHox duplication is most likely to have involved a cluster of genes [44,45,50,55]. Also, retrogenes adopt the regulatory elements of the locus into which they insert, and so are likely to have a very different expression profile from their parent gene. Both the Hox and ParaHox genes have anterior-posteriorly staggered expression patterns in the nervous system and other tissues in bilaterians. This may well be indicative of the ancestral ProtoHox duplication having involved a segment of DNA that included both the coding sequence and regulatory elements.

With regards to DNA-based transpositions it is very rare for them to be long enough to encompass entire coding sequences and even rarer for such events to include associated regulatory elements as well. These small-scale transpositions can occur either during the process of segmental duplication (SD), or when a gene transposes without duplication (as a Positionally Relocated Gene (PRG) [56]). With regards to segmental duplications the median size of the duplication is significantly smaller than the average size of a gene in nematodes, humans and flies. In *Caenorhabditis elegans* the median size of duplication is 1.4kb and the average gene length is 2.5kb [57]. In humans the average SD sizes are 18,564bp and 14,759bp for intra-chromosomal and inter-chromosomal duplications respectively [58], whilst the average gene length is 55,970bp [51]. In *Drosophila pseudoobscura* duplication sizes are rarely more than a few kilobases [59]. Furthermore, these estimates of gene sizes are, by necessity, based on analysis of exons due to the inevitable lack of information about regulatory elements for entire gene sets across whole genomes. The consequent high probability of SDs and PRGs not taking all of the ancestral regulatory regions to the new location, even if they do take all of the coding sequence, has been suggested as a route to evolutionary innovation, of new genes with new regulation and expression patterns determined by their novel genomic location (e.g. [56]). This contrasts with the similarity in the expression of the Hox and ParaHox genes already mentioned above.

The greater likelihood of the Ghost Locus hypothesis being the explanatory basis for our observations rather than the alternative of small-scale DNA-based transpositions can also be seen from not just the sizes and arguments detailed above, but also from a consideration of the relative rates of different events. SDs tend to be intra-chromosomal rather than inter-chromosomal, and of SDs as a whole, tandem duplications are the most common with, for example, 75-90% of SDs being tandem in mammals like the cow [60]. Also, in *D. melanogaster* 86% of SDs are in the intra-chromosomal category [61]. Consequently, the most likely mode of duplication that gave rise to the Hox and ParaHox genes from the ProtoHox state (if it was not via a whole chromosome or whole genome duplication) was via an intra-chromosomal duplication, such that the Hox and ParaHox genes were linked at their point of origin and hence surrounded by the same neighbouring genes. Following this likely intra-chromosomal origin for Hox and ParaHox the two types of gene were subsequently relocated to separate chromosomes. The issue now becomes whether the separation onto distinct chromosomes involved a small-scale transposition that took only the homeobox gene(s) and no/few neighbours or instead entailed a large-scale

translocation involving either the Hox or ParaHox genes as well as a number of neighbouring genes. A small-scale transposition is what would be required for the 'alternative' hypothesis whilst a large-scale translocation would be required for the Ghost Locus hypothesis to be supported. Large-scale translocations are common. For example, one in 500 newborn humans carry reciprocal translocations of large segments of chromosome arms [62-65], probably mediated via interchromosomal Low-Copy Repeats (LCRs) which number several hundred in the human genome [66]. Although large-scale inter-chromosomal translocations are thought to be less common in non-mammalian models such as fruit flies and nematodes [67], several examples have nevertheless been documented. Using data from the 12 sequenced *Drosophila* genomes Bhutkar et al [68] described the relocation of a large segment of Muller element A to Muller element D in *D. pseudoobscura*, as well as the fusion of Muller element F to D in *D. willistoni*. In some other insects, such as ants, the rates of translocations and chromosome fissions (and fusions) is much more dramatic [69]. By contrast small-scale DNA transpositions that move intact genes, or even small gene clusters such as that hypothesized for the transposition of a Hox or ParaHox cluster, are rare. Bhutkar et al [56] characterized the Positionally Relocated Genes (PRGs) in the 12 sequenced *Drosophila* genomes, distinguishing such genes from those moved by large-scale events such as reciprocal chromosome arm translocations or chromosome fusion and fissions. These authors found only 514 PRGs across all 12 *Drosophila* genomes that had translocated to a non-syntenic chromosome arm. Nearly all of these events involved only a single gene (478 out of 514), leaving only 18 events involving the remaining 36 genes. However, these authors caution that at least some of these potential multi-gene events could be artefacts due to genome mis-assembly [56]. Also, a proportion of the PRGs are due to retrotransposition events (24 – 40%), which reduces the proportion of events that are compatible with the Hox/ParaHox translocation scenario still further (as discussed above). Thus, in contrast to the relatively common chromosome arm exchanges and translocations involving large numbers of genes, the transposition of small, but multi-genic, DNA segments between chromosomes is very rare. Clearly more data is desirable to confirm that the findings from the 12 *Drosophila* genomes can be extended more widely across the animals, but with the current understanding of genome rearrangements it seems clear that the types of rearrangements underpinning the Ghost Locus hypothesis (either whole chromosome or whole genome duplications, or intra-chromosomal duplications followed by large-scale chromosome arm translocations) are much more frequent than the types of rearrangements required for the alternative hypothesis (small scale DNA-based transposition, possibly including several genes).

Whilst we cannot exclude these alternative small-scale events as a route to the separation of the Hox and ParaHox loci, we note that they are relatively rare. Although such rearrangements as SDs are themselves quite common, the occurrence of SD-like events that are suitable for explaining the Hox/ParaHox situation are in fact very rare. In this case these events are required to have been inter-chromosomal (which is rarer than intra-chromosomal [56-59]) as well as likely being multi-genic, including several homeobox genes and all of their associated regulatory elements (and maybe even a few non-homeobox neighbours) if the hypotheses of Lanfear and Bromham [45], Brooke et al [55] and Minguillón and Garcia-Fernàndez [44] are not to be rejected. We thus favour our hypothesis of whole locus split or duplication including homeobox genes and non-homeobox neighbours, followed by differential gene loss of descendant neighbours (gene loss being a common occurrence [51, 70-74]), and in some cases loss of the homeobox genes themselves which results in ghost loci.

List of Genomes Used in This Study

Nematostella vectensis v1.0, http://www.ncbi.nlm.nih.gov/nuccore/ABAV00000000

Trichoplax adhaerens Grell-BS-99 v1.0, http://www.ncbi.nlm.nih.gov/nuccore/ABGP00000000

Homo sapiens GRCh37.p2, http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606

Amphimedon queenslandica v1.0, http://spongezome.metazome.net/cgi-bin/gbrowse/amphimedon/

Lottia gigantea v1.0, http://genome.jgi-psf.org/Lotgi1/Lotgi1.info.html

Capitella teleta v1.0, http://genome.jgi-psf.org/cgi-bin/browserLoad?db=Capca1&position=scaffold_1:25000-125000

Author Contributions

O.M.R. performed research and analysed data; O.M.R., D.B. and D.E.K.F designed research; O.M.R., D.B. and D.E.K.F. wrote the paper.

Supplemental References

1. Wickstead, B., Gull, K., and Richards, T.A. (2010). Patterns of kinesin evolution reveal a complex ancestral eukaryote with a multifunctional cytoskeleton. BMC Evol Biol *10*: 110.

2. Balczon, R., Bao, L., and Zimmer, W.E. (1994). PCM-1, A 228-kD centrosome autoantigen with a distinct cell cycle distribution. J. Cell Biol. *124*, 783–793.

3. Zhu, L., Wrabl, J.O., Hayashi, A.P., Rose, L.S., and Thomas, P.J. (2008). The torsin-family AAA+ protein OOC-5 contains a critical disulfide adjacent to Sensor-II that couples redox state to nucleotide binding. Mol. Biol. Cell *19*, 3599–3612.

4. Ozelius, L.J., Page, C.E., Klein, C., Hewett, J.W., Mineta, M., Leung, J., Shalish, C., Bressman, S.B., de Leon, D., Brin, M.F. et al. (1999). The TOR1A (DYT1) gene family and its role in early onset torsion dystonia. Genomics *62*, 377–384.

5. Mochizuki, R., Ishizuka, Y., Yanai, K., Koga, Y., and Fukamizu, A. (1999). Molecular cloning and expression of human neurochondrin-1 and -2. Biochim. Biophys. Acta *1446*, 397–402.

6. Sicot, F.X., Tsuda, T., Markova, D., Klement, J.F., Arita, M., Zhang, R.Z., Pan, T.C., Mecham, R.P., Birk, D.E., and Chu, M.L. (2008). Fibulin-2 Is Dispensable for Mouse Development and Elastic Fiber Formation. Mol. Cell. Biol. *28*, 1061–1067.

7. Timpl, R., Sasak, T., Kostka, G., and Chu, M-L. (2003). Fibulins: a versatile family of extracellular matrix proteins. Nat Rev Mol Cell Biol *4*, 479–489.

8. Whittaker, C.A., and Hynes, R.O. (2002). Distribution and evolution of von Willebrand/integrin A domains: widely dispersed domains with roles in cell adhesion and elsewhere. Mol. Biol. Cell *13*, 3369–3387.

9. Frank, S., Schulthess, T., Landwehr, R., Lustig, A., Mini, T., Jenö, P., Engel, J., and Kammerer, R.A. (2002). Characterization of the matrilin coiled-coil domains reveals seven novel isoforms. J. Biol. Chem. *277*, 19071–19079.

10. Kielty, C.M., Baldock, C., Lee, D., Rock, M.J., Ashworth, J.L., and Shuttleworrth, C.A. (2002). Fibrillin: from microfibril assembly to biomechanical function. Phil. Trans. R. Soc. B *357*, 207–217.

11. Handford, P.A., Downing, A.K., Reinhardt, D.P., and Sakai, L.Y. (2000). Fibrillin: from domain structure to supramolecular assembly. Matrix Biology *19*, 457–470.

12. Deák, F., Wagener, R., Kiss, I., and Paulsson, M. (1999). The matrilins: a novel family of oligomeric extracellular matrix proteins. Matrix Biol. *18*, 55–64.

13. Kavanagh, K.L., Jörnvall, H., Persson, B., and Oppermann, U. (2008). Medium- and short-chain dehydrogenase/reductase gene and protein families. Cell. Mol. Life Sci. *65*, 3895–3906.

14. Mindnich, R., Möller, G., and Adamski, J. (2004). The role of 17 beta-hydroxysteroid dehydrogenases. Mol. Cell. Endocrinol. *218*, 7–20.

15. Kleiger, G., and Eisenberg, D. (2002). GXXXG and GXXXA Motifs Stabilize FAD and NAD(P)-binding Rossmann Folds Through Cα–H···O Hydrogen Bonds and van der Waals Interactions. J. Mol. Biol. *323*, 69–76.

16. Baker, M.E. (2001). Evolution of 17 [beta]-hydroxysteroid dehydrogenases and their role in androgen, estrogen and retinoid action. Mol. Cell. Endocrinol. *171*, 211–215.

17. Lemmon, M.A., and Ferguson, K.M. (2001). Molecular determinants in pleckstrin homology domains that allow specific recognition of phosphoinositides. Biochem. Soc. Trans. *29*, 377–384.

18. Alam, S.L., Langelier, C., Whitby, F.G., Koirala, S., Robinson, H., Hill, C.P., and Sundquist, W.I. (2006). Structural basis for ubiquitin recognition by the human ESCRT-II EAP45 GLUE domain. Nat. Struct. Mol. Biol. *13*, 1029–1030.

19. Joshi, A.K., Zhang, L., Rangan, V.S., and Smith, S. (2003). Cloning, Expression, and Characterization of a Human 4'-Phosphopantetheinyl Transferase with Broad Substrate Specificity. J. Biol. Chem. *278*, 33142–33149.

20. Susin, S.A., Lorenzo, H.K., Zamzami, N., Marzo, I., Snow, B.E., Brothers, G.M., Mangion, J., Jacotot, E., Costantini, P., Loeffler, M., et al. (1999). Molecular characterization of mitochondrial apoptosis-inducing factor. Nature *397*, 441-449.

21. Zhou, H., and Clapham, D.E. (2009). Mammalian MagT1 and TUSC3 are required for cellular magnesium uptake and vertebrate embryonic development. Proc. Natl. Acad. Sci. U.S.A. *106*, 15750–15755.

22. Fredriksson, R., Lagerström, M.C., Lundin, L-G., and Schiöth, H.B. (2003). The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol. Pharm. *63*, 1256–1272.

23. Atkinson, H.J., and Babbitt, P.C. (2009). An Atlas of the Thioredoxin Fold Class Reveals the Complexity of Function-Enabling Adaptations. PLoS Comp. Biol. *5*: e1000541.

24. Carvalho, A.P., Fernandes, P.A., and Ramos, M.J. (2006). Similarities and differences in the thioredoxin superfamily. Prog. Biophys. Mol. Biol. *91*, 229–248.

25. Martin, J.L. (1995). Thioredoxin--a fold for all reasons. Structure *3*, 245–250.

26. Calero, M., Winand, N.J., and Collins, R.N. (2002). Identification of the novel proteins Yip4p and Yip5p as Rab GTPase interacting factors. FEBS Letters *515*, 89–98.

27. Yang, X., Matern, H.T., and Gallwitz, D. (1998). Specific binding to a novel and essential Golgi membrane protein (Yip1p) functionally links the transport GTPases Ypt1p and Ypt31p. EMBO J. *17*, 4954–4963.

28. Zhang, J., Zhang, W., Zou, D., Chen, G., Wan, T., Li, N., and Cao, X. (2003.) Cloning and functional characterization of GNPI2, a novel human homolog of glucosamine-6-phosphate isomerase/oscillin. J. Cell. Biochem. *88*, 932–940.

29. Wolosker, H., Kline, D., Bian, Y., Blackshaw, S., Cameron, A.M., Fralich, T.J., Schnaar, R.L., and Snyder, S.H. (1998). Molecularly cloned mammalian glucosamine-6-phosphate deaminase localizes to transporting epithelium and lacks oscillin activity. FASEB J. *12*, 91–99.

30. Stenzel, N., Fetzer, C.P., Heumann, R., and Erdmann, K.S. (2009). PDZ-domain-directed basolateral targeting of the peripheral membrane protein FRMPD2 in epithelial cells. J. Cell Sci. *122*, 3374–3384.

31. Andersen, S.L., Bergstralh, D.T., Kohl, K.P., LaRocque, J.R., Moore, C.B., and Sekelsky, J. (2009). Drosophila MUS312 and the Vertebrate Ortholog BTBD12 Interact with DNA Structure-Specific Endonucleases in DNA Repair and Recombination. Mol. Cell *35*, 128–135.

32. Wilson, P.A., Gardner, S.D., Lambie, N.M., Commans, S.A., and Crowther, D.J. (2006). Characterization of the human patatin-like phospholipase family. J Lipid Res *47*, 1940–1949.

33. Mindell, J.A., and Maduke, M. (2001). CIC chloride channels. Genome Biol. *2*: 3003.1–3003.6.

34. Simionato, E., Ledent, V., Richards, G., Thomas-Chollier, M., Kerner, P., Coornaert, D., Degnan, B., and Vervoort, M. (2007). Origin and diversification of the basic helix-loop-helix gene family in metazoans: insights from comparative genomics. BMC Evol Biol *7*: 33.

35. Sokal, R.R., and Rohlf, F.J. (1994). Biometry. Third Edition, (W. H. Freeman).

36. Putnam, N.H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V.V., et al. (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science *317*, 86-94.

37. Manly, B.F.J. (1991). Randomization and Monte Carlo Methods in Biology. First Edition, (Chapman and Hall).

38. Hui, J.H.L., Holland, P.W.H., and Ferrier, D.E.K. (2008). Do cnidarians have a ParaHox cluster? Analysis of synteny around a *Nematostella* homeobox gene cluster. Evol Dev. *10*, 725-730.

39. Srivastava, M., Begovic, E., Chapman, J., Putnam, N.H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M.L., et al. (2008). The *Trichoplax* genome and the nature of placozoans. Nature *454*, 955-960.

40. Larroux, C., Fahey, B., Degnan, S.M., Adamski, M., Rokhsar, D.S., and Degnan, B.M. (2007). The NK homeobox gene cluster predates the origin of Hox genes. Curr Biol *17*, 706-710.

41. Hui, J.H.L., McDougall, C., Monteiro, A.S., Holland, P.W.H., Arendt, D., Balavoine, G., and Ferrier, D.E.K. (2012). Extensive chordate and annelid macrosynteny reveals ancestral homeobox gene organisation. Mol. Biol. Evol. *29*, 157-165.

42. Castro, L.F.C., and Holland, P.W.H. (2003). Chromosomal mapping of ANTP class homeobox genes in amphioxus: piecing together ancestral genomes. Evol. Dev. *5*, 459–465.

43. King, N., Westbrook, M.J., Young, S.L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., et al. (2008). The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature *451*, 783-788.

44. Minguillón, C., and Garcia-Fernàndez, J. (2003). Genesis and evolution of the *Evx* and *Mox* genes and the extended Hox and ParaHox gene clusters. Genome Biology *4*:R12.

45. Lanfear, R., and Bromham, L. (2008). Statistical tests between competing hypotheses of Hox cluster evolution. Syst. Biol. *57*, 708-718.

46. Ryan, J.F., Mazza, M.E., Pang, K., Matus, D.Q., Baxevanis, A.D., Martindale, M.Q., and 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*:e153.

47. Chourrout, D., Delsuc, F., Chourrout, P., Edvardsen, R.B., Rentzsch, F., Renfer, E., Jensen, M.F., Zhu, B., de Jong, P., Steele, R.E. et al. (2006). Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. Nature *442*, 684-687.

48. Garcia-Fernàndez, J. (2005). Hox, ParaHox, ProtoHox: facts and guesses. Heredity *94*, 145-152.

49. Finnerty, J.R. and Martindale, M.Q. (1999). Ancient origins of axial patterning genes: Hox genes and ParaHox genes in the Cnidaria. Evol Dev *1*, 16-23.

50. Ferrier, D.E.K., and Holland, P.W.H. (2001). Ancient origin of the Hox gene cluster. Nat. Rev. Genet. *2*, 33-38.

51. Makino, T., and McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. PNAS *107*, 9270-9274.

52. Mable, B.K. (2004). 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. Biol J Linn Soc *82*, 453-466.

53. Le Comber, S.C., and Smith, C. (2004). Polyploidy in fishes: patterns and processes. Biol J Linn Soc *82*, 431-442.

54. Gallardo, M.H., Kausel, G., Jiménez, A., Bacquet, C., González, C., Figueroa, J., Köhler, N., and Ojeda, R. (2004). Whole-genome duplications in South American desert rodents (Octodontidae). Biol J Linn Soc *82*, 443-451.

55. Brooke, N.M., Garcia-Fernàndez, J., and Holland, P.W.H. (1998). The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. Nature *392*, 920-922.

56. Bhutkar, A., Russo, S.M., Smith, T.F., and Gelbart, W.M. (2007). Genome-scale a

nalysis of positionally relocated genes. Genome Res *17*, 1880-1887.

57. Katju, V., and Lynch, M. (2003). The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. Genetics *165*, 1793-1803.

58. Zhang, L., Lu, H.H.S., Chung, W., Yang, J., and Li, W-H. (2005). Patterns of segmental duplication in the human genome. Mol Biol Evol *22*, 135-1441.

59. Meisel, R.P. (2009). Evolutionary dynamics of recently duplicated genes: selective constraints on diverging paralogs in the *Drosophila pseudoobscura* genome. J Mol Evol *69*, 81-93.

60. Liu, G.E., Ventura, M., Cellamare, A., Chen, L., Cheng, Z., Zhu, B., Li, C., Song, J. and Eichler, E.E. (2009) Analysis of recent segmental duplications in the bovine genome. BMC Genomics *10*:571.

61. Fiston-Lavier, A.S.D., Anxolabehere, D. and Quesneville, H. (2007) A model of segmental duplication formation in *Drosophila melanogaster*. Genome Res *17*, 1458-1470.

62. Oliver-Bonet, M., Navarro, J., Carrera, M., Egozcue, J. and Benet, J. (2002) Aneuploid and unbalanced sperm in two translocation carriers: evaluation of the genetic risk. Molecular Human Reproduction *8*, 958-963.

63. Ogilvie, M.C. and Scriven, P.N. (2002) Meiotic outcomes in reciprocal translocation carriers ascertained in 3-day human embryos. Eur J Hum Genet *10*, 801-806.

64. Chang, E.M., Han, J.H., Kwak, I.P., Lee, W.S., Yoon, T.K. and Shim, S.H. (2012) Preimplantation genetic diagnosis for couples with a Robertsonian translocation: practical information for genetic counselling. J Assist Reprod Genet *29*, 67-75.

65. Anton, E., Blanco, J., Egozcue, J. and Vidal, F. (2004) Sperm FISH studies in seven male carriers of Robertsonian translocation t(13;14)(q10;q10). Human Reproduction *19*, 1345-1351.

66. Ou, Z., Stankiewicz, P., Xia, Z., Breman, A.M., Dawson, B., Wiszniewska, J., Szafranski, P., Cooper, M.L., Rao, M., Shao, L. et al. (2011) Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. Genome Res *21*, 33-46.

67. Hillier, L.W., Miller, R.D., Baird, S.E., Chinwalla, A., Fulton, L.A., Kobolbt, D.C. and Waterston, R.H. (2007) Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. PLoS Biol *5*:e167.

68. Bhutkar, A., Schaeffer, S.W., Russo, S.M., Xu, M., Smith, T.F. and Gelbart, W.M. (2008) Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. Genetics *179*, 1657-1680.

69. Lorite, P. and Palomeque, T. (2010) Karyotype evolution in ants (Hymenoptera: Formicidae), with a review of the known ant chromosome numbers. Myrmecol News *13*, 89-102.

70. Hughes, A.L., and Friedman, R. (2004). Differential loss of ancestral gene families as a source of genomic divergence in animals. Proc. R. Soc. Lond. B Suppl. *271*, S107-S109.

71. Danchin, E.G.J., Gouret, P., and Pontarotti, P. (2006). Eleven ancestral gene families lost in mammals and vertebrates while otherwise universally conserved in animals. BMC Evol Biol *6*:5.

72. Miller, D.J., Hemmrich, G., Ball, E.E., Hayward, D.C., Khalturin, K., Funayama, N., Agata, K., and Bosch, T.C.G. (2007). The innate immune repertoire in Cnidaria – ancestral complexity and stochastic gene loss. Genome Biol. *8*:R59.

73. Wyder, S., Kriventseva, E.V., Schröder, R., Kadowaki, T., Zdobnov, E.M. (2007). Quantification of ortholog losses in insects and vertebrates. Genome Biol. *8*:R242.

74. Takahashi, T., McDougall, C., Troscianko, J., Chden, W-C., Jayaraman-Nagarajan, A., Shimeld, S.M., and Ferrier, D.E.K. (2009). An EST screen from the annelid *Pomatoceros lamarckii* reveals patterns of gene loss and gain in animals. BMC Evol. Biol. *9*:240.