Sea anemone genomes reveal ancestral metazoan chromosomal macrosynteny

Bob Zimmermann[1,2], Sofia M.C. Robb[3], Grigory Genikhovich[1], Whitney J. Fropf[3], Lukas Weilguny[1], Shuonan He[3], Shiyuan Chen[3], Jessica Lovegrove-Walsh[1], Eric M. Hill[3], Cheng-Yi Chen[3], Katerina Ragkousi[3,4], Daniela Praher[1], David Fredman[1], Yehu Moran[1,5], Matthew C. Gibson[3]*, Ulrich Technau[1,2]*

[1]Department of Neurosciences and Developmental Biology, Faculty of Life Sciences, University of Vienna, Althanstrasse 14, 1090 Vienna, Austria
[2]Research platform SinCeReSt, University of Vienna, Althanstrasse 14, 1090 Vienna, Austria
[3]Stowers Institute for Medical Research, Kansas City, MO 64110 USA
[4]Department of Biology, Amherst College, Amherst, MA 01002 USA
[5]*Current address:* The Alexander Silberman Institute of Life Sciences, Faculty of Science, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel
*To whom correspondence should be addressed: mg2@stowers.org, ulrich.technau@univie.ac.at

Abstract

Cnidaria (sea anemones, jellyfish, corals and hydra) form a close sister group to Bilateria. Within this clade, the sea anemone *Nematostella vectensis* has emerged as a slow evolving model for investigating characteristics of the cnidarian-bilaterian common ancestor, which diverged near the Cambrian explosion. Here, using long read sequencing and high throughput chromosome conformation capture, we generate high quality chromosome-level genome assemblies for *N. vectensis* and the closely related edwardsiid sea anemone, *Scolanthus callimorphus.* In both cases we find a robust set of 15 chromosomes comprising a stable linkage group detectable within all major clades of sequenced cnidarian genomes. Further, both genomes show remarkable chromosomal conservation with chordates. In contrast with Bilateria, we report that extended Hox and NK gene clusters are chromosomally linked but do not retain a tight spatial conservation. Accordingly, there is a lack of evidence for topologically associated domains, which have been implicated in the evolutionary pressure to retain tight microsyntenic gene clusters. We also uncover ultra-conserved noncoding elements at levels previously undetected in non-chordate lineages. Both genomes are accessible through an actively updated genome browser and database at https://simrbase.stowers.org

Introduction

Cnidaria (corals, sea anemones, jellyfish and hydroids) constitutes a large clade of basally branching Metazoa, dating back 550-650 Mya [1–3]. Its robust evolutionary position as sister to Bilateria makes Cnidaria a key group to study the evolution of bilaterian features, such as axis organization, mesoderm formation and central nervous system development [4]. The list of cnidarian model systems established in the laboratory is ever-expanding [5].

The edwardsiid sea anemone *Nematostella vectensis* was first recognized as a potential laboratory model in 1992 [6]. The system provides access to thousands of embryos in a single spawn, and has since become widely adopted [7,8]. The initial sequencing of its genome revealed uncanny conservation of synteny and gene content to bilaterians [9], which has led to several surprising discoveries about the genome of the cnidarian-bilaterian common ancestor [2,9,10]. Although conservation signals of recently diverged taxa have long been a cornerstone to genomic inquiry [11], no second genome within the family *Edwardsiidae* has been sequenced. In addition, high quality, chromosome-level assemblies have been a crucial tool for establishing model organisms, not only for understanding gene organization and content, but also for investigation of variation and regulatory elements. Nevertheless, the current assembly of the *N. vectensis* genome remains at scaffold-level and is derived from four haplotypes [9].

To address these issues and lend further insight into the chromosomal linkage and organization of Cnidaria, we generated high quality genomes of two edwarsiid sea anemones, *Nematostella vectensis* and *Scolanthus callimorphus*. While *N. vectensis* is mostly found in the brackish water estuaries of the North American coast, the larger but morphologically similar “worm anemone” *S. callimorphus* is found in seawater on European coasts [12,13]. Using PacBio sequencing in combination with high-throughput chromosome conformation capture, we built pseudo-chromosomes for both species and established the ancestral linkage groups of the *Edwardsiidae*. Notably, the comparison of these two genomes with fully sequenced genomes of a sponge, other cnidarians and several bilaterians further allowed the reconstruction of the history of metazoan chromosomes through macrosynteny conservation. We demonstrate that while the chromosomes of Cnidaria are quite robust, rapid changes in local synteny are pervasive, including the extended Hox and NK cluster. Additionally, we report the first ultra-conserved noncoding elements outside of Bilateria, some of which are linked to extended Hox cluster genes. These genomes are served on an actively maintained genome browser.

Results

High Quality Chromosome-Level Assemblies of Two Edwardsiid Genomes

High molecular weight genomic DNA was extracted from individual edwardsiid sea anemones *N. vectensis* and *S. callimorphus* (Figure 1a,b) and subjected to short-read sequencing. Using a *k*-mer coverage model, we estimated the genome lengths for each species (EDF1). At 244 Mb,

the estimate of *N. vectensis* is substantially shorter than previously suggested [9], however this discrepancy could be attributed to the previous use of four haplotypes in sequencing. Interestingly, the *S. callimorphus* genome is substantially longer at 414 Mb, making it the largest sequenced actiniarian genome. The genome of the sea anemone *Exaiptasia pallida* is closer in length to *N. vectensis* [14], despite branching earlier, indicating that the genome lengths among Actiniaria may be more dynamic than previously indicated (Figure 1c).

N. vectensis and *S. callimorphus* PacBio long-read libraries were constructed from the same DNA used to estimate the genome size (*S. callimorphus*) or individuals from the same clonal population (*N. vectensis*). Self-corrected PacBio sequences were assembled into initial contigs which were already highly contiguous after an initial pass (nv2contigs, sc1contigs, EDF2a-c). The initial assembly showed indications of redundancy as indicated by BUSCO scores (EDF3), likely caused by heterozygous alleles assembling into separate contigs. The greater number of duplicate BUSCOs in *S. callimorphus* corresponds to its higher heterozygosity as indicated by the *k*-mer model (EDF1). After removal of these redundant contigs, the duplicate BUSCOs were reduced 3.4-fold in *S. callimorphus* and 8-fold in *N. vectensis* (EDF3).

Compared to the published *N. vectensis* assembly, the contiguity of both genomes in terms of contig-level N50 was over 25 fold higher (EDF2a-c,f). The *N. vectensis* assembly was further scaffolded by generating libraries using the Dovetail Chicago *in vitro* proximity ligation platform (see Materials and Methods for details). The Dovetail scaffolded genome both improved the N50 contiguity 2-fold (EDF2f) and united a single fragmented BUSCO (EDF3).

We were additionally able to validate the order and correctness of the *N. vectensis* intra-scaffold sequence using the REAPR pipeline [15]. We extracted DNA fragments from two individuals, measured the insert size distribution and sequenced paired ends. REAPR was used to break the genome where a substantial portion of paired read mappings on the contigs conflicted with the expected distance. The contiguity of the broken assemblies, as compared to the initial scaffolded assemblies, was much higher than that of the original *N. vectensis* assembly, and also relatively higher as a fraction of the raw N50 (EDF2d). Additionally, the fraction of the scaffolded genome considered to be error-free (in terms of both sequence and contiguity) was significantly improved over the previous *N. vectensis* scaffolds (EDF2e). Taken together, not only are the nv2scaffolds substantially more contiguous than the previous nv1scaffolds, the sequence within the scaffolds has fewer misassemblies and errors.

In order to obtain a chromosome-level assembly of *N. vectensis* and *S. callimorphus*, we performed high throughput chromosomal conformation capture (Hi-C) on a single individual from each species. After automated assembly followed by assembly review (see Materials and Methods for details) both contact maps showed evidence for 15 chromosomes (Figure 2a,b). This is in line with the previous estimates based on the number of *N. vectensis* metaphase plates [9] and the analysis of *N. vectensis* chromosome spreads [16]. We were also able to validate the dovetail scaffolding by performing a second, independent assembly of the purged contigs using our Hi-C data. As shown in EDF4, the assemblies were nearly identical,

confirming the robustness of both the scaffolding and the assembly method. The minor differences in these assemblies were inspected in further manual review and ties were broken according to the Hi-C contact signal in the Dovetail assembly.

Edwardsiid genomes show extensive macrosynteny conservation across metazoans

We next sought to determine whether the pseudo-chromosomes identified above corresponded to ancestral linkage groups (ALGs) common to the two edwardsiids in gene content and order. Indeed, each of the 15 pseudo-chromosomes of both species have a single corresponding pair (Figure 3a). We observed that 8117 of 8692 mutual best BLAST hits between the edwardsiids were retained on their respective ALG. However, gene order was largely lost from the most recent common ancestor (MRCA), which we estimate to have diverged between 150 and 174 Mya (EDF13). Most pseudo-chromosomes, according to their ALG pairings, corresponded in length but are much larger in *S. callimorphus* (Figure 4a-1). This is accounted for by a large fraction of unclassified, potentially lineage-specific repeat classes (EDT1). Remarkably, the longest two pseudo-chromosomes in *S. callimorphus* correspond to the 7th shortest and the shortest pseudo-chromosome in *N. vectensis*, respectively. Both pseudo-chromosomes were rich in repetitive sequences in both species (Figure 4a-3). In particular the LTR/Pou repeat class, a pan-metazoan repeat class abundant in *Drosophila* genomes, but absent from mammalian genomes [17], was enriched in both pseudo-chromosomes relative to others (*z*-scores 1.3 and 1.4, respectively, Extended Data Table 2-5) as well as their counterparts in *N. vectensis* (*z*-scores 1.8 and 1.5).

In order to assess the level of conservation of chromosome-scale gene linkages with non-anthozoan cnidarians, we next mapped the ALGs observed in the edwardsiid species onto the genome of a scyphozoan jellyfish. In *Rhopilema esculentum,* recent Hi-C analysis demonstrates the existence of 21 chromosomes, in line with independently observed chromosome spreads [18,19]. While many of the edwardsiid ALGs were split in the jellyfish, we observed a clear many-to-one mapping of all *R. esculentum* chromosomes (Figure 3b). This is remarkable considering the age of the Anthozoa-Medusozoa divergence time, which is estimated at more than 550 Mya by calibrated molecular clocks (EDF13, [20]). Comparing the edwardsiid ALGs to other scaffold-level assemblies, we found similar divisions in the genomes of the anthozoan *E. pallida,* the scleractinian coral *Acropora millepora* and the medusazoans *Clytia hemisphaerica* and *Hydra magnipapillata* (EDF5-8). We were unable to fully confirm all 15 groups in the Hydrozoa, however, our orthology detection is hindered by the technical issue that these genomes are in draft quality and, additionally, the genomes have evolved more rapidly [21]. On the other hand, the hydra clade has been consistently shown to have a 15-chromosome karyotype [22]. Taken together, this indicates that a putatively minimal set of 15 cnidarian ALGs is conserved in the *Edwardsiidae*.

We carried this analysis one step further and compared the edwardsiid ALGs with a bilaterian–the lancet *Branchiostoma floridae*–an early branching chordate that lacks the two rounds of whole-genome duplication found in vertebrates [23] (Figure 3c). Strikingly, the *B.*

floridae pseudo-chromosomes exhibit retention of many ALGs from the bilaterian-cnidarian MRCA. We observed that 6 of the edwardsiid chromosomes correspond to a single ALG with the lancelet while the remaining groups appear to have arisen from fission and translocation events.

We also compared the edwardsiid and *B. floridae* ALGs to those in the chromosomes of the sponge *Ephydatia muelleri,* a member of Porifera, which branched off prior to the cnidarian-bilaterian split [24]. Strikingly, many ALGs are clearly retained between them, giving us a glimpse at the common ancestral chromosomes Porifera and Planulozoa (the group of Cnidaria and Bilateria). Fewer genes were retained between *Edwardsiidae* and *E. muelleri*, but despite diverging earlier, we identified fewer apparent chromosomal translocation events in the MRCA than between *B. floridae* and *N. vectensis* (Figure 3d). Interestingly, the primary Hox-bearing chromosome in the *Edwarsdiidae* corresponds to several chromosomes in both *E. muelleri* and *B. floridae*. Macrosyntenic analysis between the chordate and the sponge reveals that these groups were likely present in the common ancestor of Metazoa and underwent fusion events in the *Edwardsiidae* lineage (EDF9).

Conversely, vertebrate chordates such as the early branching teleost fish *Lepisosteus oculatus* and humans show additional translocation events from the MRCA (Figure 3e, EDF10). Ecdysozoan genomes, on the other hand, while bearing many detectable orthologs, exhibit no apparent retention of chromosomal linkage in the case of *Drosophila melanogaster* (Figure 3f), or weak retention in the case of *Caenorhabditis elegans* (EDF11). Rapid intrachromosomal gene shuffling is a well-known phenomenon in *Drosophila,* however within the drosophilid clade, chromosomes are retained on well-established linkage groups [25].

Chromosomal Organization of the NK and extended Hox gene clusters

The chromosome-level assembly of the *N. vectensis* genome allowed us not only to follow the evolution of gene linkages by comparing macrosyntenies at the genome-wide scale, but also to re-address the evolution of specific gene clusters. Prominent examples of clusters of homeodomain transcription factor coding genes ancestral for Bilateria include the SuperHox cluster (*Hox* genes, *Evx*, *Dlx*, *Nedx*, *Engrailed*, *Mnx*, *Rough*, *Hex*, *Mox* and *Gbx*), the ParaHox cluster (*Gsx*, *Xlox*, *Cdx*), and the NK/NK-like cluster (*NK5*, *NK1*, *Msx*, *NK4*, *NK3*, *Ladybird*, *Tlx-like*, *NK7*, *NK6*) as well as *NK2* group genes located separately [26–28]. It has been hypothesized that each of these originated from a single gene cluster, which then disintegrated during evolution (for review see [28]). Genomic analysis revealed that, similar to Bilateria, *N. vectensis* possesses a separate ParaHox cluster of two genes, *Gsx* and a mixed identity *Xlox*/*Cdx* on chromosome 10, and a SuperHox cluster on chromosome 2 containing *Hox*, *Evx*, *Mnx*, and *Rough*, as well as more distant *Mox* and *Gbx* [29] (Figures 4 and 5). We identified an *NK* cluster on chromosome 5 containing *NK1*, *NK5*, *Msx*, *NK4*, *NK3*, *NK7*, *NK6*, and more distant *Ladybird*, a *Tlx-like* gene and, intriguingly, *Hex,* which is also linked to the NK cluster in the hemichordate *Saccoglossus kowalevskii* [30] and in the cephalochordate *Branchiostoma floridae*. Similar to Bilateria, the *NK2* genes were clustered separately and found on

chromosome 2 (Figures 4 and 5). In contrast, in earlier branching sponges, neither ParaHox nor extended Hox genes exist, and only the NK cluster is present with a single *NK2/3/4* gene, two *NK5/6/7* genes, an *Msx* ortholog, as well as possible *Hex* and *Tlx* orthologs [31], EDF12. Taken together, this allows us to propose that the last common ancestor of Cnidaria and Bilateria possessed an NK-cluster on a chromosome different from the one carrying the SuperHox cluster, and a separate NK2 cluster, which might have been on the same chromosome as the SuperHox cluster (Figure 5). The hypothesized SuperHox-NK Megacluster [26], if it ever existed, must have both formed and broken apart during the time after the separation of the sponge lineage, but before the origin of the cnidarian-bilaterian ancestor (Figure 5, Supplementary Text).

Sea anemone genomes lack evidence for topologically associating domains

In the past decade, high resolution chromosome conformation capture has increased interest in topologically associating domains (TADs), recurring chromosomal folding motifs evidenced by signals in Hi-C contact maps [32]. Flanking regions of TADs are positively correlated with CCCTC-binding factor (CTCF) binding sites. Interestingly, no CTCF ortholog has been detected among non-bilaterian animals [33,34]. Notably, we do not detect any evidence for TADs in our edwardsiid pseudo-chromosomes in the form of differential enrichment of contact density, exemplified in the contact map shown in Figure 2c,d. By comparison, analyses of other datasets generated using the identical protocol at similar levels of resolution have identified clear signals of TADs [35].

To our knowledge, only one other study has addressed the higher-order chromosomal organization in a non-bilaterian [24]. In this work, Kenny and colleagues searched for TADs in their chromosome-level assembly of the sponge *E. muelleri.* The authors were able to detect bioinformatic evidence for the presence of TADs, which would make it the first CTCF-less species for which this is reported. While the contact data pass the software threshold, the contact maps resemble patterns we observe in our assemblies at the boundary of scaffolds or contigs, which can be the result of differential mappability from repetitive content or assembly issues. We therefore deliberately do not report any results from a TAD finder, since, after multiple rigorous rounds of manual assembly update, we can assert that the data we have generated do not qualitatively represent TAD boundaries per se, and most results would be likely false positives. While the precise definition of a TAD is still evolving [36,37], both data sets lack many characters of TADs identified in CTCF-containing genomes: hierarchical compartments, mammalian-specific “corner peaks” indicating strong interactions, and in our case, loop peaks and inter-contig compartments. This suggests that the presence of CTCF is necessary for the formation of TADs, however, we still cannot exclude the possibility that performing the experiment with a more homogenous cell population, or sequencing at a higher resolution, would reveal a signal on a smaller scale.

Genomic Distribution of Ultraconserved Non-coding Elements

Longer stretches of unusually highly conserved non-coding genomic regions have been studied thoroughly in vertebrates with varying methodologies [38–43]. These ultraconserved non-coding elements (UCNEs) are generally posited to be *cis*-regulatory regions, but the precise mechanistic requirement for such high conservation is not known. While some previous studies have argued that the degree of conservation observed in UCNEs is specific to vertebrates [44], a more recent study has asserted that UCNEs are present in smaller numbers within *Drosophila* [45]. We were therefore interested to find out how many, and to what extent, UCNEs are present in *Edwardsiidae*. We adopted criteria previously used to detect UCNEs between chicken and humans [43] and found 145 regions in the *N. vectensis* genome that were highly conserved with *S. callimorphus* (EDT 7). 116 of these regions fell into 37 syntenic clusters of at most 500 kb intervening gaps, and the remaining 29 were singletons. Several such clusters were close to NK paralogs, such as one containing 12 UCNEs and spanning 70 kb surrounding the Nk3-4 cluster on chromosome 5 and three UCNEs upstream of the Nkx2.2 cluster on chromosome 2, a pattern previously reported in vertebrates [41]. Additionally, we detected a single UCNE by the PaxC gene, and while UCNEs have been previously reported in vertebrates, this neural developmental gene appears to have arisen from a cnidarian-specific duplication [46]. On the other hand, no UCNEs are found near the edwardsiid *Irx* gene, despite their ubiquity among Bilateria [47–49]. The *Irx* gene cluster is one of the most striking examples of convergent evolution, with the drosophilids, cephalochordates, polychaetes and vertebrates each having an independently evolved cluster of three genes exhibiting staggered expression [48]. The authors suggested that the function of the convergent UCNEs observed in amphioxus and vertebrates is maintenance of the cluster. Its absence in the edwardsiids would support this notion. Furthermore, a more recent study implicated the presence of conserved TADs, pre-dating the duplication of the *Irx* gene in vertebrates, were necessary for their complex expression pattern [49]. Thus, this suggests the combination of distal enhancers encoded in UCNEs with dynamic higher-order chromosomal organization are key elements in the transition to a more complex regulatory landscape.

Discussion

The assembly of two high quality, chromosome-level edwardsiid genomes has illuminated several intriguing aspects about chromosomal structure, the NK and extended Hox clusters, the conservation of non-coding elements and the status of topologically associated domains in the common ancestor to cnidarians and bilaterians. In addition, the highly improved *N. vectensis* genome will prove to be an invaluable resource in future studies of both coding and non-coding regions, structural variants among populations and continued development of functional tools for this model organism.

As previously reported, a wide array of ANTP class homeobox genes can be observed in the *N. vectensis* genome [50], however orthology and initial linkage analysis suggested independent diversification of the anterior and non-anterior Hox genes in Cnidaria and Bilateria [29]. In this work, we revisited previous syntenic analyses of the NK and extended Hox cluster organization

[50,51] in the context of our pseudo-chromosomes. Nearly all members of the extended Hox cluster were distributed among distant, isolated microsyntenic blocks on pseudo-chromosome 2, with the single exception of *HoxF/Anthox1*, located on pseudo-chromosome 5 (Figures 4 and 5). This indicates a lack of proximity constraint on the Hox genes, contrasting with the situation Bilateria. In addition, while a staggered spatiotemporal pattern of Hox expression along the secondary, directive axis of the *N. vectensis* larva and polyp can be observed [52], unlike Bilateria, there is no correlation between expression and cluster position [53].

The dispersed NK and extended Hox cluster may be due to the apparent lack or diminished higher-order chromosome organization. In line with this, it was recently observed that the hoxd cluster boundaries in mouse are marked by TAD boundaries [54], and the cluster's intra-TAD gene order is deemed to be under selective pressure [55]. Moreover, the absence of CTCF in cnidarians, ctenophores and sponges suggests that the existence of TADs might represent a bilaterian-specific feature.

While microsynteny analyses reveal little conservation of the local gene order in the genomes of *N. vectensis* and *S. callimorphus*, macrosyntenic analysis of the edwardsiid chromosomes compared to available cnidarian genomes revealed a high level of conservation. We show evidence that nearly all major clades exhibit a complement of the same 15 groups present in *S. callimorphus* and *N. vectensis*, with little evidence of translocation since the common cnidarian ancestor split an estimated 580 Mya. This stands in stark contrast to the history of, for example, the 326 Mya amniota ancestral genome estimated to have 49 distinct units, whose extant taxa consist of multiple translocated segments and variable chromosomes [56]. However, by far more remarkable is that macrosynteny is maintained between the edwardsiids, the early branching chordate *B. floridae*, and the sponge *E. muelleri*. Our analyses reveal clear one-to-one, one-to-few or few-to-one conservation of the chromosome-level linkages between cnidarians, sponges and early chordates, which suggests a striking conservation of macrosyntenies during early animal evolution. In contrast, the conservation becomes weaker when cnidarian linkage groups are compared with the ones from more derived bilaterians such as vertebrates and is absent in ecdysozoans. It is tempting to speculate that the emergence of the TADs in Bilateria may have restricted local rearrangements and released the constraints on maintaining the ancestral macrosyntenies conserved all the way back to the origin of multicellular animals.

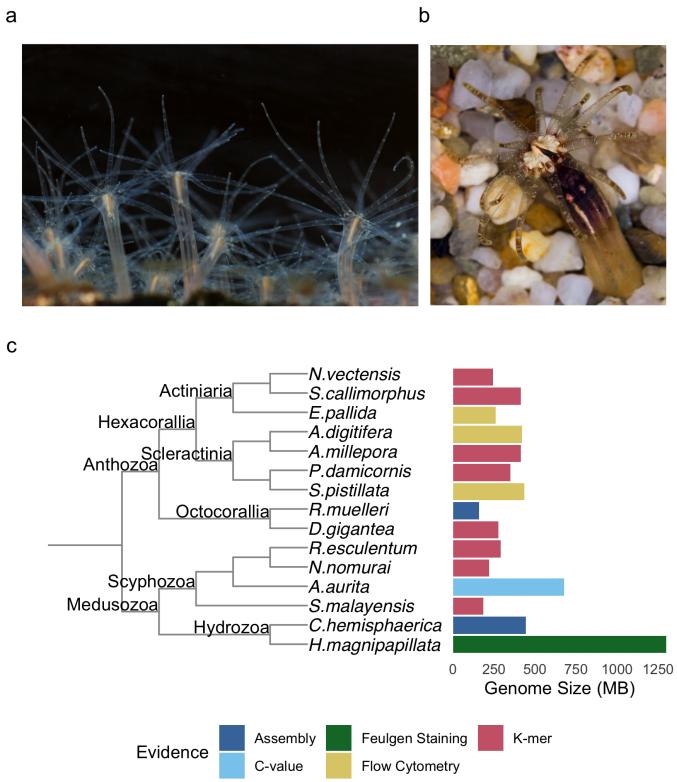

Figure 1. a) *Nematostella vectensis*. photo credit: Johanna Kraus and Patrick RH Steinmetz b) *Scolanthus callimorphus*. photo credit: Robert Reischl c) Genome size estimates of cnidarian genomes taken from [10,14,18,19,57–65].

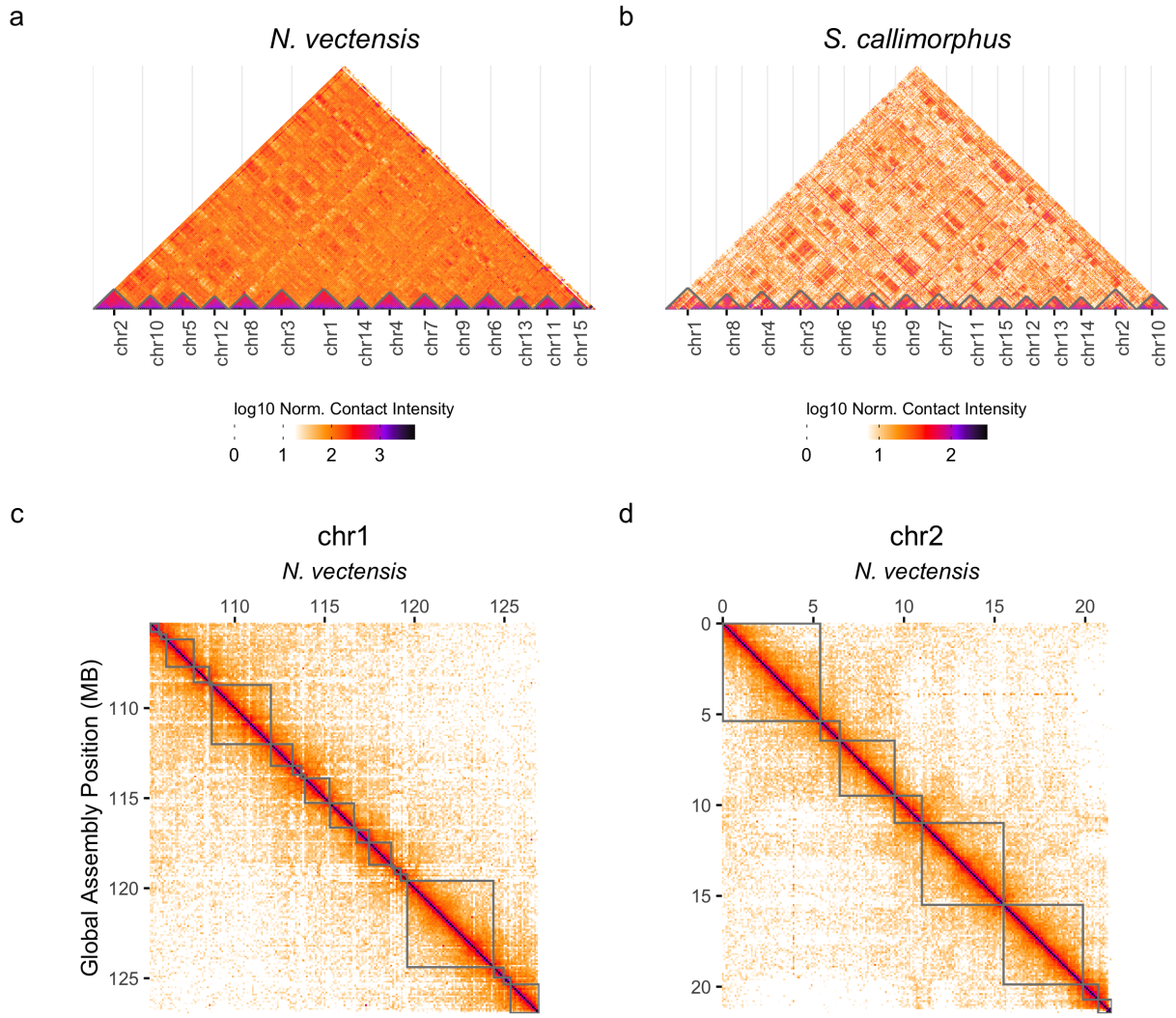

Figure 2 Hi-C contact maps. Each point represents a binned, normalized intensity of chromosomal contact as measured by the number of ligated fragments sequenced. (a,b) Contact maps of *N. vectensis* scaffolds (a) and *S. callimorphus* contigs (b). (c,d) Contact maps for the two largest chromosomes of the assembly. Grey boxes represent either pseudo-chromosomes (a,b) or dovetail scaffolds (c,d). Color scales for (a), (c) and (d) are identical.

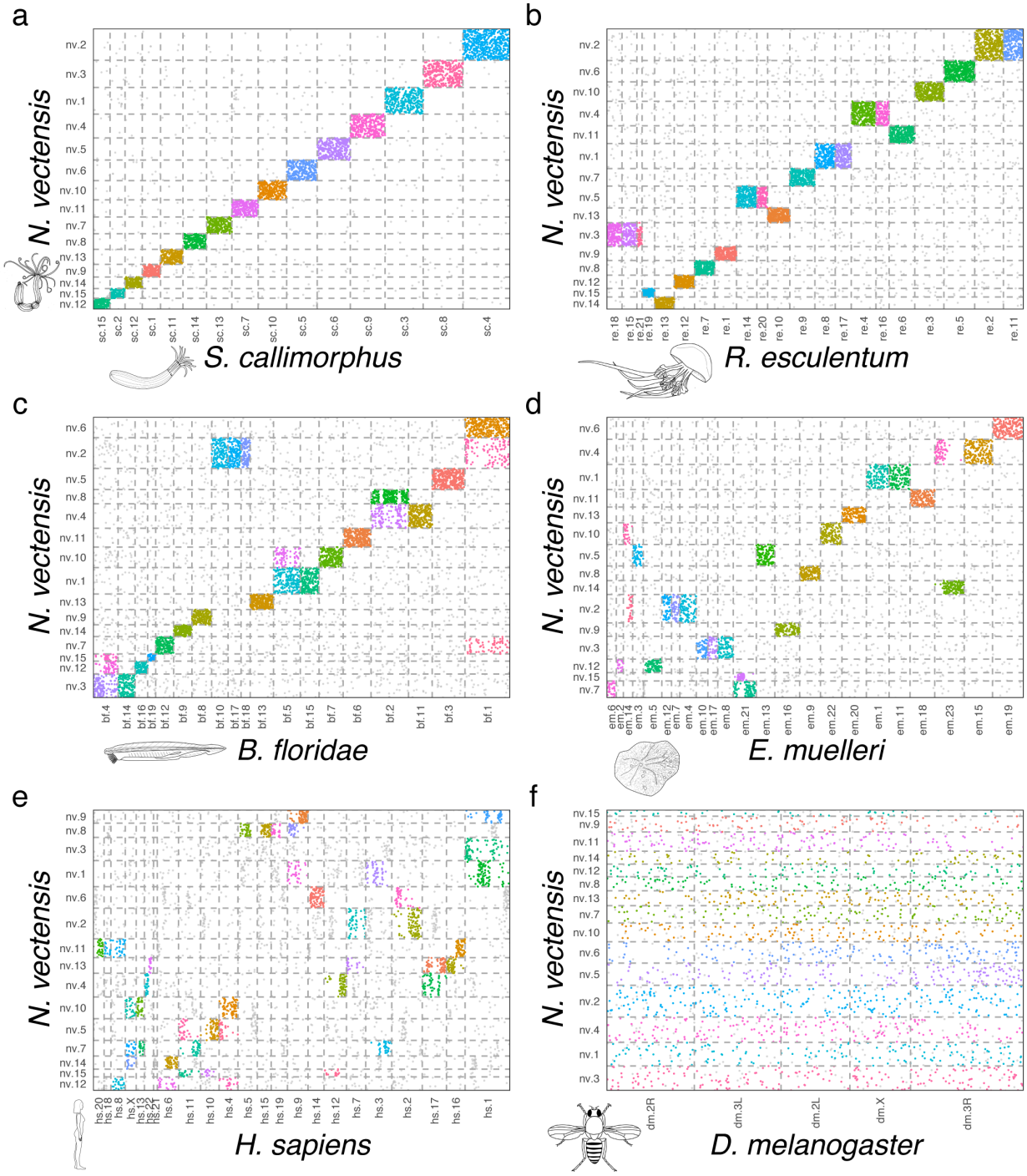

Figure 3. Oxford plots of the macrosyntenic relationships of chromosomes between species. The color coding represents inferred ancestral linkage groups between the two species represented in each figure, except for (f), which is colored by the edwardsiid ancestral linkage groups shown in (a). See materials and methods for details.

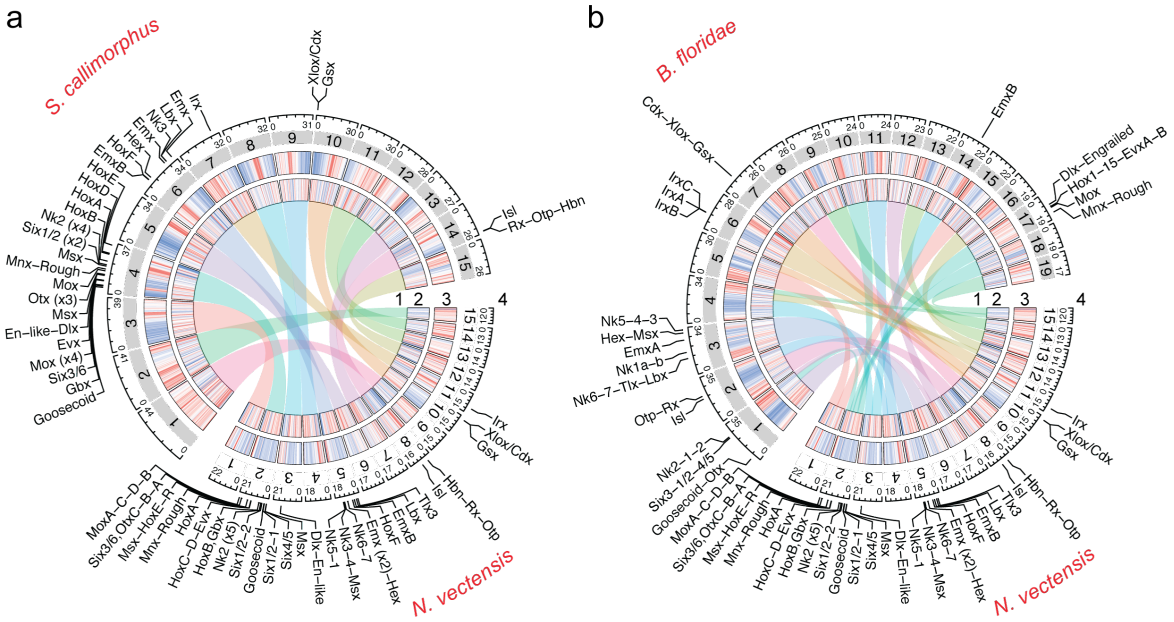

Figure 4. Chromosomal relationships, genomic content and locations of NK and extended Hox cluster genes of (a) *N. vectensis* and *S. callimorphus* and (b) *N. vectensis* and *B. floridae*. 1) Chord diagram of macrosyntenic relationships of chromosomes based on the inferred ancestral linkage groups. Chords represent the proportion of the chromosome's gene content in common with the compared genome, not the positional information. 2) Scaled and centered gene density relative to the respective genome. Red is high and blue is low. 3) Scaled and centered density of interspersed repeat elements relative to the respective genome. Red is high and blue is low. 4) Locations of extended Hox and NK cluster genes.

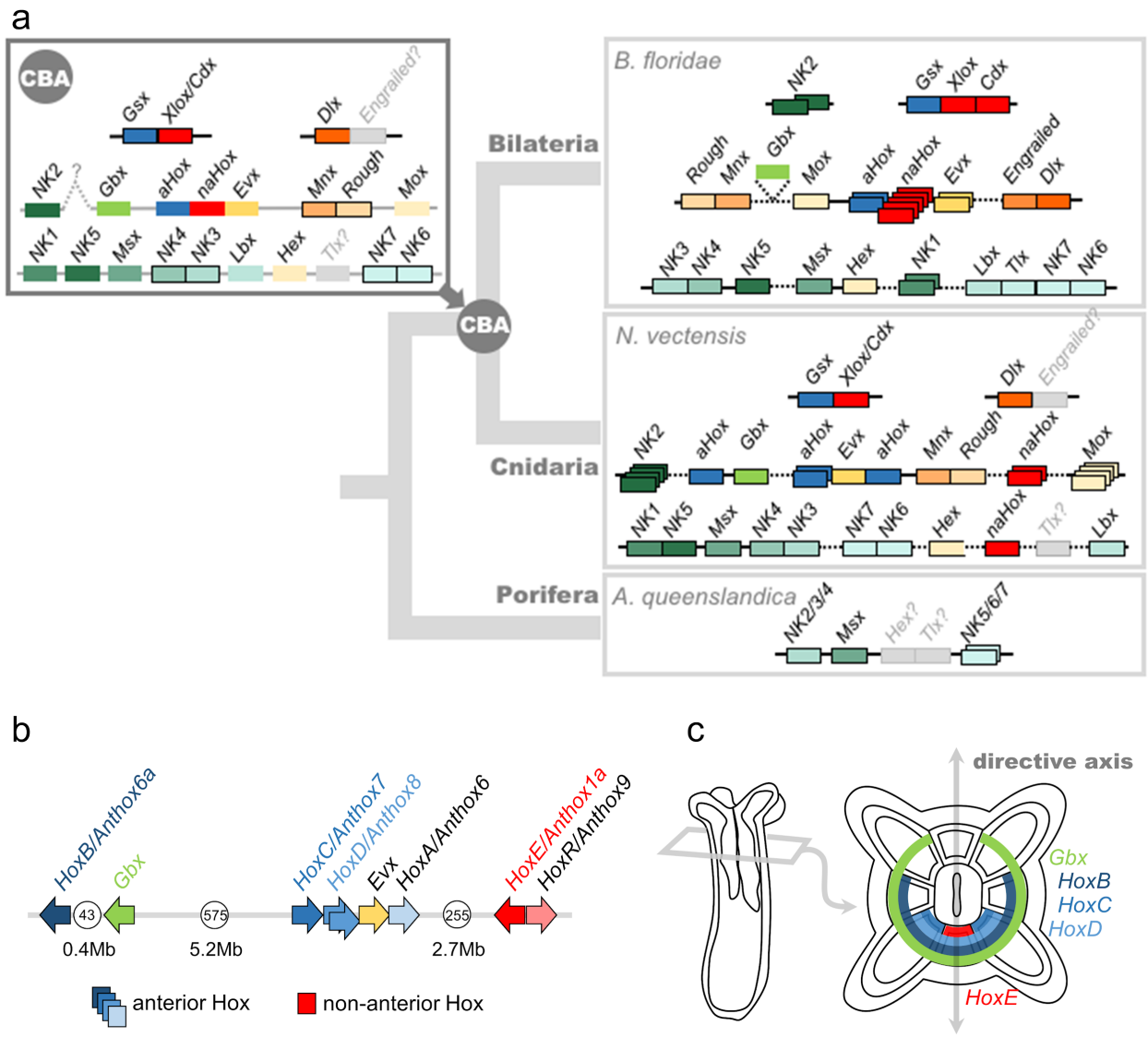

Figure 5. Evolution of a selection of Antennapedia class homeobox gene clusters. a) Composition of the extended Hox, ParaHox and NK clusters in the sponge *Amphimedon queenslandica*, the sea anemone *N. vectensis* and in a chordate *B. floridae*, and the deduced cluster composition of the cnidarian-bilaterian ancestor (CBA). Greyed-out genes with question marks have uncertain orthology. Genes shown as borderless boxes have uncertain position relative to neighboring cluster members. Stacked boxes represent clusters of paralogs of the indicated ancestral gene. Genes in immediate proximity are indicated by abutting boxes. Linked genes of the same class separated by 1 to 50 intervening genes are connected with solid lines, over 50, with dashed lines. Grey intergenic connectors in the CBA indicate that the distances and the number of the intervening genes between the cluster members cannot be estimated. *NK2* of the CBA may be linked to the extended Hox cluster. Since *B. floridae Gbx* remained unplaced in the chromosome-level assembly, its position was taken from the scaffold-level assembly of *Branchiostoma lanceolatum* [66]. b,c) Organization of the *Nematostella* Hox cluster (b) and staggered expression of *Gbx* and Hox genes along the directive axis of the animal (c).

Arrows show the direction of transcription for each of the genes. The number of intervening genes is indicated in white circles. The spatial collinearity of *Gbx* and Hox expression is not observed.

Acknowledgements

Special thanks to Oleg Simakov for extensive discussions and suggestions. We thank Matthew Nicotra for providing us with the HMW DNA extraction protocol used for *S. callimorphus.* We thank Robert Reischl for the photo of *S. callimorphus* and Patrick R.H. Steinmetz and Hanna Kraus for the photo of *Nematostella vectensis* (Fig. 1). Special thanks to Tatiana Bagaeva for the cartoon drawings of animals used in this study. We are grateful to the Stowers Institute Molecular Biology Core facility, particularly Amanda Lawlor, Michael Peterson and Anoja Perera. This work was supported by grants of the Austrian Science Fund FWF (P24858; P21108) to U.T., support from the Stowers Institute for Medical Research to M.G. and an NIH Ruth L. Kirschstein NRSA (F32 GM131522) to E.M.H.. We are also grateful for the support of the CNRS Marine Station in Roscoff and the Assemble grant 227799 to U.T. for collecting *S. callimorphus*.

Data Availability

All raw data is available via the National Center for Biotechnology Information under the accession PRJNA667495. The assembled genomes can be downloaded, browsed and searched at on publicly available browsers at https://simrbase.stowers.org/starletseaanemone and https://simrbase.stowers.org/wormanemone. Code used to generate the analyses is available from the authors upon request.

Materials and Methods

Animal Care and Source

Nematostella vectensis animals were cultured as previously described [67] at the University of Vienna and the Stowers Institute. Adult male and female individuals were verified by induction in isolation. Scolanthus callimorphus animals were collected at the Île Callot, Carantec, France. After transport, they were kept in seawater at 20°C and fed freshly hatched Artemia salina weekly or biweekly.

Sequencing

Short Read DNA-Seq

Genomic DNA samples were extracted from both adult male and female individual *N. vectensis* adults using the DNeasy Blood and Tissue Kit (Qiagen). After purification, approximately 5 ug of genomic DNA was recovered from each sample. Following DNA extraction, samples were sheared and size selected for ~500 bp using a Blue Pippin Prep machine (Sage Science).

Following size selection, sequencing libraries were created using a KAPA HTP Library Prep kit (Roche) and subjected to paired-end sequencing on an Illumina NextSeq 500. *S. callimorphus* DNA samples for library preparation were aliquoted from high molecular weight extractions, described below.

High Molecular Weight DNA Extraction and Library Prep

N. vectensis high molecular weight DNA was extracted at Dovetail Genomics. Samples were quantified using Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). The PacBio SMRTbell library (~20kb) for PacBio Sequel was constructed using SMRTbell Template Prep Kit 1.0 (PacBio, Menlo Park, CA, USA) using the manufacturer recommended protocol. The pooled library was bound to polymerase using the Sequel Binding Kit 2.0 (PacBio) and loaded onto PacBio Sequel using the MagBead Kit V2 (PacBio). Sequencing was performed on the PacBio Sequel SMRT cell, using Instrument Control Software Version 5.0.0.6235, Primary analysis software Version 5.0.0.6236 and SMRT Link Version 5.0.0.6792.

High molecular weight DNA from a single *Scolanthus callimorphus* adult animal was extracted using a modified Urea-based DNA extraction protocol [68,69]. A whole animal was flash frozen and ground with mortar and pestle. While frozen, drops of buffer UEB1 (7M Urea, 312.5 mM NaCl, 50 mM Tris-HCl pH 8, 20 mM EDTA pH 8, 1 % w:v N-Lauroylsarcosine sodium salt) were added and crushed with the tissue. Tissue was incubated in a final volume of 10 mL UEB1 at RT for 10 minutes. Three rounds of phenol-chloroform extraction were performed followed by DNA precipitation by addition of 0.7 volume isopropanol. The pellet was transferred to a fresh tube and washed twice in 70 % EtOH and twice more in 100 % EtOH, dried, and resuspended in TE buffer.

A library for PacBio sequencing was then prepared from the high molecular weight sample using the SMRTbell® Express Template Prep Kit v1. The libraries were then sequenced on a PacBio Sequel machine over 3 SMRT Cells, yielding a total of 22.85 Gb over 1,474,285 subreads. An aliquot of the same sample was used to prepare a library using the NEBNext® Ultra™ II DNA Library Prep Kit for Illumina. This was then subjected to 50 cycles of single-end sequencing in one flow cell lane using an Illumina HiSeq 2500 system.

Chicago libraries

2 Chicago libraries were prepared as described previously [70]. Briefly, for each library, ~500 ng of HMW gDNA (mean fragment length = 100 kbp) was reconstituted into chromatin in vitro and fixed with formaldehyde. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq 2500 (rapid run mode). The number and length of read pairs produced for each

library was: 116 million, 2x101 bp for library 1; 35 million, 2x101 bp for library 2. Together, these Chicago library reads provided 125 x sequence coverage of the genome (1-100 kb pairs).

Chromatin was extracted from a single *Nematostella vectensis* adult male and *Scolanthus callimorphus* adult (unknown sex) nuclei using the Phase Genomics Proximo Hi-C animal protocol. After proximity ligation and purification, 16 ng and 9 ng of DNA was recovered, respectively. For library preparation 1 µl of Library Reagent 1 was added 12 PCR cycles were performed. The final library was subjected to 150 total cycles of paired-end sequencing using an Illumina NextSeq 550 machine yielding a total of 13.5 gigabases.

Hi-C sequencing, *Scolanthus callimorphus* PacBio library preparation and sequencing, Scolanthus Illumina DNA library preparation and sequencing and adult *Nematostella vectensis* RNA library preparation and sequencing was performed at the VBCF NGS Unit (https://www.viennabiocenter.org/facilities). *Nematostella vectensis* DNA size selection, library preparation, and sequencing were performed by the Molecular Biology Core at the Stowers Institute for Medical Research.

Developmental and adult *N. vectensis* RNA sequencing was performed as follows. *N. vectensis* were spawned and eggs were de-jellied and fertilized as previously described [67]. Spawning and embryo development took place at 18 °C. Eggs and embryos from different stages were collected (300 per sample) in duplicate as indicated: eggs (within 30min of spawn), blastula (7.5hpf), gastrula (23.5hpf) and planula (72hpf). Eggs and embryos were collected in eppendorf tubes and centrifuged to a pellet at 21,000 x g for 1 min. All seawater was quickly removed and pellets were resuspended in 150ml lysis buffer (RLT buffer supplied by the Qiagen RNeasy kit (#74104), supplemented with β-mercaptoethanol). The samples were homogenized with an electric pestle (1 min continuous drilling) and further supplemented with 200 ml of the above lysis buffer. Homogenized samples were then transferred into QIAshredder columns (Qiagen #79654) and centrifuged at 21,000 x g for 2 min. The flow throughs were supplemented with 1 ml 70 % ethanol and transferred to RNeasy columns and were processed according to the Qiagen RNeasy protocol. Quality and integrity of the RNA was evaluated using the Agilent RNA 600 pico kit (Agilent Technologies) and RNA samples were stored at -80 °C until further processing. cDNA libraries were then constructed for polyA stranded sequencing. The resulting libraries were sequenced on Illumina HiSeq using paired end runs (RapidSeq- 2x150bp).

Genome Assembly

Size estimates for *Nematostella vectensis* and *Scolanthus callimorphus* were derived using Genomescope [71], taking the result of the highest *k* (56 and 18) which converged under the model.

Initial assemblies based on PacBio sequencing of *N. vectensis* and *S. callimorphus* were generated using canu version 1.8 [72] with the parameters `rawErrorRate=0.3 correctedErrorRate=0.045`.

N. vectensis haplotigs were removed using Purge Haplotigs [73]. First, the source PacBio reads were aligned onto the canu assembly using minimap2 [74] using the parameters `-ax map-pb --secondary=no`. Following this a coverage histogram was generated using the Purge Haplotigs script `readhist`. Per the documented Purge Haplotigs protocol, lower, mid and high coverage limits were found by manual inspection of the plotted histogram to be 12, 57 and 130 respectively. All initial contigs marked as suspect or artifactual were removed from further analysis with the Purge Haplotigs script `purge`.

The input de novo assembly, shotgun reads, and Chicago library reads were used as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies (Putnam et al, 2016). Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After scaffolding, shotgun sequences were used to close gaps between contigs.

Repetitive DNA

Repetitive DNA was found using two strategies. First, known repeats found in repbase [75] were searched in the assemblies using RepeatMasker [76] using the parameters `-s -align -e ncbi` in addition to `-species nematostella` for *N. vectensis* and `-species edwardsiidae` for *S. callimorphus*. Second, novel repeat sequences were found using RepeatModeler version 2.0 [77]. After generating the repeat library, genomes' repeat regions were detected with the corresponding library using the same parameters in RepeatMasker.

Due to lower sequencing coverage, diploid per-scaffold coverage could not be deconvolved from haploid, and therefore Purge Haplotigs could not be used. Initial removal of redundant contigs was performed with Redundans version 0.14a [78] using the parameters `--noscaffolding --nogapclosing --overlap 0.66`. Only contigs marked in the reduced version of the genome were used in further analysis.

Hi-C sequences were aligned to the reduced genomes of *N. vectensis* and *S. callimorphus* using `bwa mem` [79,80] using the parameters `-5SP`. Candidate assemblies were generated by mapping the Hi-C sequences to the initial contigs subjected to repbase library repeat masking (contigs_standardmask) and novel repeats (contigs_aggressivemask). For *N. vectensis,* an additional candidate assembly was generated by mapping Hi-C sequences to the Chicago library scaffolded sequences using repbase masking (dovetail_standardmask). 2 candidate assemblies, using rebase masking (contigs_standardmask) and repeatmodeler masking (contigs_aggressivemask). These mappings were used to generate initial chromosomal assemblies using Lachesis [81], specifying the restriction site GATC. Assemblies were manually reviewed using Juicebox Assembly Tools version 1.11.08 [82]. Candidate assemblies were compared using the nucmer aligner with default parameters and visualized using mummerplot

[83]. Assemblies were converted over to Juicebox format using juicebox_scripts (https://github.com/phasegenomics/juicebox_scripts). In the case of *S. callimorphus*, duplicate regions were clipped, and the resulting contigs were subjected to another round of alignment, assembly and review.

Genome and gene model set assembly and completeness was assessed using BUSCO version 3.0.2 [84], using the gene set `metazoa_odb9` as the standard.

N. vectensis scaffold correctness was assessed using REAPR [15]. *N. vectensis* assembly nemVec1 was downloaded from the JGI website [9]. Sequences from the adult male and adult female (see Sequencing) were aligned to nemVec1 and the *N. vectensis* genome after scaffolding with Chicago libraries using SMALT as well as the REAPR tool perfectmap using an expected insert size of 400, as determined from fragment analysis. Error-free bases and contiguity after breaking the genome were extracted from the results.

Gene Models

N. vectensis, S. callimorphus and *M. senile* paired end sequences obtained from a previous study [85] and data available on BioProject PRJNA430035 were used to generate *de novo* assembled transcripts.

Trinity version 5.0.2 [86] was run on each library using the flags `--min_contig_length 200 --min_kmer_cov 2`. For those which had a strand-specific library preparation, the flag `--SS_lib_type RF` was applied. To reduce redundancy, cd-hit version 4.6.8 [87,88] was applied with the flags `-M 0 -c 1`. Transdecoder version 5.0.6 [89] was used to detect open reading frames in the resulting reduced set of transcripts. Transcript abundance was quantified using salmon version 1.2.1 [90] using the flags `--seqBias --useVBOpt --discardOrphansQuasi --softclip`.

For PacBio Iso-seq, 12 *N. vectensis* RNA samples were collected over the course of multiple developmental stages, adult tissues and regeneration time points. For developmental stages, zygotes spawned by a single batch of wildtype colony were kept at 22°C, and collected at 0 hpf, 24 hpf, 48 hpf, 72 hpf and 7 dpf. Adult tissues were collected from sex-sorted, sexually mature wildtype individuals kept at 22°C. The male and female mesenteries were harvested separately by surgically opening the body column and carefully peeling off the attached body column tissues. Adult oral discs were collected by surgical removal of tentacles as well as the attached pharyngeal regions. Regeneration was induced by amputating the oral part of a sexually mature individual at the mid-pharyngeal level. Regenerating tissues close to the wound were collected at 4 hpa and 12 hpa, respectively. All the samples were deep frozen and lysed using TRIzolTM reagent (Invitrogen). Phenol-chloroform extraction was performed to remove undissolved mesoglea from adult tissues. DirectzolTM RNA Miniprep Plus Kit (Zymo) was then used to purify total RNA from the aqueous phase. For each sample, 2 µg of total RNA with RIN > 7 was submitted to UC Berkeley for Iso-seq library construction.

RNA Libraries were sequenced at UC Berkeley using PacBio Sequel-II system. Raw subreads bams were processed and demultiplexed using PacBio's isoseq v3.2 conda pipeline. The steps include consensus generation, primer demultiplexing, polyA refinement and data clustering using default parameters. This resulted in the generation of 406,317 high quality HIFI reads and used to build Nvec200 transcriptome.

HIFI reads were mapped to the *N. vectensis* genome using minimap2 [74] using parameters (`-ax splice -uf --secondary=no`) to obtain the primary best alignments. Reads were then grouped and collapsed down to potential transcripts using PacBio's cDNA_Cupcake toolkit. Based on PacBio's guideline, transcripts with degraded 5' reads and have less than 10 FL counts were removed. Chimeric transcripts were then analyzed to find potential fusion genes. For reads that didn't map to the genome, de novo transcriptome assembly was performed using graph-based tool Cogent with kmer size equals 30. In total, we were able to capture 17817 genes and 81999 transcripts from the mapped reads, as well as 12781 de novo transcripts from the unmapped reads.

Evidence for *S. callimorphus* gene models were taken from RNA-sequencing and repeats. *S. callimorphus* RNA-seq reads (see Sequencing) were mapped to the *S. callimorphus* contigs using STAR version 2.7.3a [91]. These mappings were used as evidence for intron junctions to generate putative gene models and estimating hidden Markov model parameters using BRAKER2 [92,93]. Gene models were then refined using Augustus version 3.3.3 [94] using extrinsic evidence from STAR splice junctions and the location of repeats from RepBase (see Genome Assembly) as counter-evidence for transcription. These models were filtered with the following criteria: 1) genes completely covered by RepeatModeler repeats (see Genome Assembly) were removed 2) predicted gene models were required to be either supported by external RNA-seq evidence as reported by Augustus or have a predicted ortholog as reported by Eggnog-mapper [95]. This resulted in a set of 24,625 gene models. Transcription factor identity was inferred by aligning the predicted protein sequences to Pfam A domains version 32.0 [96] using hmmer version 3.3 [97]. Transcription factor families were based on domains curated in a previous work [98].

Extended Hox cluster, NK cluster and ParaHox genes were found with BLAT [99] matches of published models [29,46,50,51,100–103] to the nv1 genome, taking the best hits. If an NVE gene model [85] corresponded to the matched genomic region, its location in the nv2 genome was then determined for macrosynteny analysis. In cases where no published gene was known, reciprocal BLAST hits between the bilaterian and cnidarian counterpart were taken as evidence for orthology.

Divergence Estimates

Single copy orthologs were detected by collecting common complete and duplicated BUSCO genes present in the *S. callimorphus* and *N. vectensis* genomes. Where duplicated BUSCOs

were present, the transcript with the highest score was taken. This resulted in a total of 541 orthologs. BUSCOs found in genomes obtained from previous studies [10,14,19,23,24,64,65,104–109] were used to generate multiple alignments. Genes were aligned with mafft version 7.427 using the E-INS-i model and a maximum 1000 refinement iterations [22]. Alignments were trimmed using trimAl version 1.4.rev15 using the “gappyout” criteria [22]. A maximum likelihood tree was inferred using iqtree version 2.0.6, using the model finder partitioned on each gene, constrained to nuclear protein models (89). Divergence estimates were determined using r8s version 1.8.1 using the Langley-Fitch likelihood method [111]. Age ranges were estimated by fixing the split between Bilateria and Cnidaria at 595.7 and 688.3 Mya [112].

Ultraconserved Elements

In order to determine noncoding elements conserved between *S. callimorphus* and *N. vectensis*, genomes repeat-masked from both *de novo* and repbase repeats were blasted using NCBI+ version 2.10.0 [113], using the flags `-evalue 1E-10 -max_hsps 100000000 -max_target_seqs 100000000 -task megablast -perc_identity 0 -template_length 16 -penalty -2 -word_size 11 -template_type coding_and_optimal`. Additionally, the `-dbsize` parameter was set to the estimated genome size. Candidate hits were then filtered using criteria loosely based on previous work [43]: for each high-scoring pair, a sliding window method was used to determine subsections of the alignment with at least 95 % identity, and extending these windows as long as the identity remains at this level. *N. vectensis* elements mapping to more than one locus in the *S. callimorphus* genome were reduced to the longest locus pair in both genomes. Elements mostly mapping to coding sequence were removed, and the remaining elements were classified as intron or non-coding, depending on location. Recurring UCE sequences that were not identified by RepeatModeler or RepeatMasker were detected with blastclust version 2.2.26 using requiring the length of hit to cover at least 90 % of either sequence for linkage.

Macrosynteny Analysis

Branchiostoma floridae gene models and sequences were retrieved from the recently published study [23]. Gene orthology between *S. callimorphus, N. vectensis* and *B. floridae* were determined pairwise using reciprocal best matches. All against all comparisons were performed with NCBI+ blastp version 2.10.0 [113] using an e-value threshold of 1e-5. Reciprocal best matches were determined using match bit scores. Genomes were downloaded from previous studies [10,14,58,64,65,107–109,114].

Ancestral linkage groups were determined following an adapted version of the statistical test previously described [23]. For every pair of chromosomes in two genomes *x* and *y*, contingency tables for Fisher’s exact test consisting of four cells were constructed: a) the shared mutual hits between those chromosomes, b) the shared mutual hits in all other chromosomes in genome *y* and the same chromosome in *x*, c) the shared mutual hits in all other chromosomes in genome

x and the same chromosome in y, and d) the shared mutual hits in all other chromosomes in genome *x* and *y*. Fisher's *p*-values were Bonferroni corrected for the total number of pairs of chromosomes, and chromosome pairs with adjusted *p*-values less than 0.05 were considered significant ALGs.

References

1. Cartwright P, Halgedahl SL, Hendricks JR, Jarrard RD, Marques AC, Collins AG, et al. Exceptionally Preserved Jellyfishes from the Middle Cambrian. PLOS ONE. 2007;2: e1121. doi:10.1371/journal.pone.0001121
2. Technau U, Genikhovich G, Kraus JEM. Cnidaria. In: Wanninger A, editor. Evolutionary Developmental Biology of Invertebrates 1. Vienna: Springer Vienna; 2015. pp. 115–163. doi:10.1007/978-3-7091-1862-7_6
3. Zapata F, Goetz FE, Smith SA, Howison M, Siebert S, Church SH, et al. Phylogenomic Analyses Support Traditional Relationships within Cnidaria. PLOS ONE. 2015;10: e0139068. doi:10.1371/journal.pone.0139068
4. Technau U, Steele RE. Evolutionary crossroads in developmental biology: Cnidaria. Development. 2011;138: 1447–1458. doi:10.1242/dev.048959
5. He S, Grasis JA, Nicotra ML, Juliano CE, Schnitzler CE. Cnidofest 2018: the future is bright for cnidarian research. EvoDevo. 2019;10: 20. doi:10.1186/s13227-019-0134-5
6. Hand C, Uhlinger KR. The Culture, Sexual and Asexual Reproduction, and Growth of the Sea Anemone *Nematostella vectensis*. The Biological Bulletin. 1992;182: 169–176. doi:10.2307/1542110
7. Layden MJ, Rentzsch F, Röttinger E. The rise of the starlet sea anemone *Nematostella vectensis* as a model system to investigate development and regeneration: Overview of starlet sea anemone *Nematostella vectensis*. Wiley Interdisciplinary Reviews: Developmental Biology. 2016;5: 408–428. doi:10.1002/wdev.222
8. Rentzsch F, Technau U. Genomics and development of Nematostella vectensis and other anthozoans. Curr Opin Genet Dev. 2016;39: 63–70. doi:10.1016/j.gde.2016.05.024
9. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, et al. Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. Science. 2007;317: 86–94. doi:10.1126/science.1139158
10. Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, et al. The dynamic genome of Hydra. Nature. 2010;464: 592–596. doi:10.1038/nature08830
11. Wilson TJ, Lazner F, Kola I, Hertzog PJ. The Mouse and the Genomic Era. In: Clark MS, editor. Comparative Genomics. Boston, MA: Springer US; 2000. pp. 97–121. doi:10.1007/978-1-4615-4657-3_5
12. Gosse P. Scolanthus callimorphus. World List of Actiniaria, accessed through : World Register of Marine Species. 2020. Available: http://www.marinespecies.org/aphia.php?p=taxdetails&id=100910 on 2020-08-16
13. Wilson E. Scolanthus callimorphus Worm anemone. Marine Life Information Network: Biology and Sensitivity Key Information Reviews, [on-line]. 2005. Available: https://www.marlin.ac.uk/species/detail/1226
14. Baumgarten S, Simakov O, Esherick LY, Liew YJ, Lehnert EM, Michell CT, et al. The genome of Aiptasia, a sea anemone model for coral symbiosis. Proc Natl Acad Sci USA. 2015;112: 11893–11898. doi:10.1073/pnas.1513318112
15. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool

for genome assembly evaluation. Genome Biology. 2013;14: R47. doi:10.1186/gb-2013-14-5-r47
16. Guo L, Accorsi A, He S, Guerrero-Hernández C, Sivagnanam S, McKinney S, et al. An adaptable chromosome preparation methodology for use in invertebrate research organisms. BMC Biology. 2018;16: 25. doi:10.1186/s12915-018-0497-4
17. de la Chaux N, Wagner A. BEL/Pao retrotransposons in metazoan genomes. BMC Evolutionary Biology. 2011;11: 154. doi:10.1186/1471-2148-11-154
18. Li Y, Gao L, Pan Y, Tian M, Li Y, He C, et al. Chromosome-level reference genome of the jellyfish Rhopilema esculentum. Gigascience. 2020;9. doi:10.1093/gigascience/giaa036
19. Nong W, Cao J, Li Y, Qu Z, Sun J, Swale T, et al. Jellyfish genomes reveal distinct homeobox gene clusters and conservation of small RNA processing. Nature Communications. 2020;11: 3051. doi:10.1038/s41467-020-16801-9
20. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017;34: 1812–1819. doi:10.1093/molbev/msx116
21. Steele RE, David CN, Technau U. A genomic view of 500 million years of cnidarian evolution. Trends Genet. 2011;27: 7–13. doi:10.1016/j.tig.2010.10.002
22. Zacharias H, Anokhin B, Khalturin K, Bosch TCG. Genome sizes and chromosomes in the basal metazoan Hydra. Zoology. 2004;107: 219–227. doi:10.1016/j.zool.2004.04.005
23. Simakov O, Marlétaz F, Yue J-X, O'Connell B, Jenkins J, Brandt A, et al. Deeply conserved synteny resolves early events in vertebrate evolution. Nature Ecology & Evolution. 2020;4: 820–830. doi:10.1038/s41559-020-1156-z
24. Kenny NJ, Francis WR, Rivera-Vicéns RE, Juravel K, de Mendoza A, Díez-Vives C, et al. Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge Ephydatia muelleri. Nature Communications. 2020;11: 3676. doi:10.1038/s41467-020-17397-w
25. Ranz JM, Maurin D, Chan YS, Grotthuss M von, Hillier LW, Roote J, et al. Principles of Genome Evolution in the Drosophila melanogaster Species Group. PLOS Biology. 2007;5: e152. doi:10.1371/journal.pbio.0050152
26. Pollard SL, Holland PWH. Evidence for 14 homeobox gene clusters in human genome ancestry. Current Biology. 2000;10: 1059–1062. doi:10.1016/S0960-9822(00)00676-X
27. Butts T, Holland PWH, Ferrier DEK. The Urbilaterian Super-Hox cluster. Trends in Genetics. 2008;24: 259–262. doi:10.1016/j.tig.2007.09.006
28. Ferrier DEK. Evolution of Homeobox Gene Clusters in Animals: The Giga-Cluster and Primary vs. Secondary Clustering. Frontiers in Ecology and Evolution. 2016;4. doi:10.3389/fevo.2016.00036
29. Chourrout D, Delsuc F, Chourrout P, Edvardsen RB, Rentzsch F, Renfer E, et al. Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. Nature. 2006;442: 684–687. doi:10.1038/nature04863
30. Simakov O, Kawashima T, Marlétaz F, Jenkins J, Koyanagi R, Mitros T, et al. Hemichordate genomes and deuterostome origins. Nature. 2015;527: 459–465. doi:10.1038/nature16150
31. Larroux C, Fahey B, Degnan SM, Adamski M, Rokhsar DS, Degnan BM. The NK Homeobox Gene Cluster Predates the Origin of Hox Genes. Current Biology. 2007;17: 706–710. doi:10.1016/j.cub.2007.03.008
32. Friedman N, Rando OJ. Epigenomics and the structure of the living genome. Genome Res. 2015;25: 1482–1490. doi:10.1101/gr.190165.115
33. Heger P, Marin B, Bartkuhn M, Schierenberg E, Wiehe T. The chromatin insulator CTCF

and the emergence of metazoan diversity. PNAS. 2012;109: 17507–17512. doi:10.1073/pnas.1111941109
34. Schwaiger M, Schonauer A, Rendeiro AF, Pribitzer C, Schauer A, Gilles AF, et al. Evolutionary conservation of the eumetazoan gene regulatory landscape. Genome Research. 2014;24: 639–650. doi:10.1101/gr.162529.113
35. Sieber KB, Batorsky A, Siebenthall K, Hudkins KL, Vierstra JD, Sullivan S, et al. Integrated Functional Genomic Analysis Enables Annotation of Kidney Genome-Wide Association Study Loci. JASN. 2019;30: 421–441. doi:10.1681/ASN.2018030309
36. Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. Nature Reviews Genetics. 2018;19: 789–800. doi:10.1038/s41576-018-0060-8
37. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. Science Advances. 2019;5: eaaw1668. doi:10.1126/sciadv.aaw1668
38. Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, et al. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature. 2002;420: 578–582. doi:10.1038/nature01251
39. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. Ultraconserved elements in the human genome. Science. 2004;304: 1321–1325. doi:10.1126/science.1098119
40. Sandelin A, Bailey P, Bruce S, Engström PG, Klos JM, Wasserman WW, et al. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics. 2004;5: 99. doi:10.1186/1471-2164-5-99
41. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, et al. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 2005;3: e7. doi:10.1371/journal.pbio.0030007
42. Elgar G, Vavouri T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends in Genetics. 2008;24: 344–352. doi:10.1016/j.tig.2008.04.005
43. Dimitrieva S, Bucher P. UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks. Nucleic Acids Res. 2013;41: D101–D109. doi:10.1093/nar/gks1092
44. Retelska D, Beaudoing E, Notredame C, Jongeneel CV, Bucher P. Vertebrate conserved non coding DNA regions have a high persistence length and a short persistence time. BMC Genomics. 2007;8: 398. doi:10.1186/1471-2164-8-398
45. Kern AD, Barbash DA, Chang Mell J, Hupalo D, Jensen A. Highly Constrained Intergenic Drosophila Ultraconserved Elements Are Candidate ncRNAs. Genome Biol Evol. 2015;7: 689–698. doi:10.1093/gbe/evv011
46. Matus DQ, Pang K, Daly M, Martindale MQ. Expression of Pax gene family members in the anthozoan cnidarian, Nematostella vectensis: Pax gene expression in Nematostella vectensis. Evolution & Development. 2007;9: 25–38. doi:10.1111/j.1525-142X.2006.00135.x
47. Calle-Mustienes E de la, Feijóo CG, Manzanares M, Tena JJ, Rodríguez-Seguel E, Letizia A, et al. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. Genome Res. 2005;15: 1061–1072. doi:10.1101/gr.4004805
48. Irimia M, Maeso I, Garcia-Fernàndez J. Convergent Evolution of Clustering of Iroquois Homeobox Genes across Metazoans. Mol Biol Evol. 2008;25: 1521–1525. doi:10.1093/molbev/msn109

49. Tena JJ, Alonso ME, de la Calle-Mustienes E, Splinter E, de Laat W, Manzanares M, et al. An evolutionarily conserved three-dimensional structure in the vertebrate Irx clusters facilitates enhancer sharing and coregulation. Nature Communications. 2011;2: 310. doi:10.1038/ncomms1301
50. Ryan JF, Mazza ME, Pang K, Matus DQ, Baxevanis AD, Martindale MQ, et al. Pre-Bilaterian Origins of the Hox Cluster and the Hox Code: Evidence from the Sea Anemone, Nematostella vectensis. Fay J, editor. PLoS ONE. 2007;2: e153. doi:10.1371/journal.pone.0000153
51. Ryan JF, Burton PM, Mazza ME, Kwong GK, Mullikin JC, Finnerty JR. The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: evidence from the starlet sea anemone, Nematostella vectensis. Genome Biology. 2006;7: R64. doi:10.1186/gb-2006-7-7-r64
52. He S, Viso F del, Chen C-Y, Ikmi A, Kroesen AE, Gibson MC. An axial Hox code controls tissue segmentation and body patterning in Nematostella vectensis. Science. 2018;361: 1377–1380. doi:10.1126/science.aar8384
53. McGinnis W, Krumlauf R. Homeobox genes and axial patterning. Cell. 1992;68: 283–302. doi:10.1016/0092-8674(92)90471-N
54. Rodríguez-Carballo E, Lopez-Delisle L, Zhan Y, Fabre PJ, Beccari L, El-Idrissi I, et al. The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. Genes Dev. 2017;31: 2264–2281. doi:10.1101/gad.307769.117
55. Lazar NH, Nevonen KA, O'Connell B, McCann C, O'Neill RJ, Green RE, et al. Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. Genome Res. 2018;28: 983–997. doi:10.1101/gr.233874.117
56. Sacerdot C, Louis A, Bon C, Berthelot C, Roest Crollius H. Chromosome evolution at the origin of the ancestral vertebrate genome. Genome Biology. 2018;19: 166. doi:10.1186/s13059-018-1559-1
57. Voolstra CR, Li Y, Liew YJ, Baumgarten S, Zoccola D, Flot J-F, et al. Comparative analysis of the genomes of Stylophora pistillata and Acropora digitifera provides evidence for extensive differences between species of corals. Sci Rep. 2017;7: 1–14. doi:10.1038/s41598-017-17484-x
58. Moya A, Huisman L, Ball EE, Hayward DC, Grasso LC, Chua CM, et al. Whole transcriptome analysis of the coral Acropora millepora reveals complex responses to CO_2-driven acidification during the initiation of calcification. Mol Ecol. 2012;21: 2440–2454. doi:10.1111/j.1365-294X.2012.05554.x
59. Cunning R, Bay RA, Gillette P, Baker AC, Traylor-Knowles N. Comparative analysis of the Pocillopora damicornis genome highlights role of immune system in coral evolution. Scientific Reports. 2018;8: 16134. doi:10.1038/s41598-018-34459-8
60. Jiang JB, Quattrini AM, Francis WR, Ryan JF, Rodríguez E, McFadden CS. A hybrid de novo assembly of the sea pansy (Renilla muelleri) genome. Gigascience. 2019;8. doi:10.1093/gigascience/giz026
61. Jeon Y, Park SG, Lee N, Weber JA, Kim H-S, Hwang S-J, et al. The Draft Genome of an Octocoral, Dendronephthya gigantea. Genome Biol Evol. 2019;11: 949–953. doi:10.1093/gbe/evz043
62. Kim H-M, Weber JA, Lee N, Park SG, Cho YS, Bhak Y, et al. The genome of the giant Nomura's jellyfish sheds light on the early evolution of active predation. BMC Biology. 2019;17: 28. doi:10.1186/s12915-019-0643-7
63. Ohdera A, Ames CL, Dikow RB, Kayal E, Chiodin M, Busby B, et al. Box, stalked, and

upside-down? Draft genomes from diverse jellyfish (Cnidaria, Acraspeda) lineages: Alatina alata (Cubozoa), Calvadosia cruxmelitensis (Staurozoa), and Cassiopea xamachana (Scyphozoa). Gigascience. 2019;8. doi:10.1093/gigascience/giz069
64. Gold DA, Katsuki T, Li Y, Yan X, Regulski M, Ibberson D, et al. The genome of the jellyfish Aurelia and the evolution of animal complexity. Nature Ecology & Evolution. 2019;3: 96. doi:10.1038/s41559-018-0719-8
65. Leclère L, Horin C, Chevalier S, Lapébie P, Dru P, Peron S, et al. The genome of the jellyfish Clytia hemisphaerica and the evolution of the cnidarian life-cycle. Nat Ecol Evol. 2019;3: 801–810. doi:10.1038/s41559-019-0833-2
66. Marlétaz F, Firbas PN, Maeso I, Tena JJ, Bogdanovic O, Perry M, et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. Nature. 2018;564: 64–70. doi:10.1038/s41586-018-0734-6
67. Fritzenwanker JH, Technau U. Induction of gametogenesis in the basal cnidarian Nematostella vectensis (Anthozoa). Development Genes and Evolution. 2002;212: 99–103. doi:10.1007/s00427-002-0214-7
68. Chen J, Dellaporta S. Urea-based Plant DNA Miniprep. In: Freeling M, Walbot V, editors. The Maize Handbook. New York, NY: Springer; 1994. pp. 526–527. doi:10.1007/978-1-4612-2694-9_85
69. Sanders SM, Ma Z, Hughes JM, Riscoe BM, Gibson GA, Watson AM, et al. CRISPR/Cas9-mediated gene knockin in the hydroid Hydractinia symbiolongicarpus. BMC Genomics. 2018;19: 649. doi:10.1186/s12864-018-5032-z
70. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 2016;26: 342–350. doi:10.1101/gr.193474.115
71. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33: 2202–2204. doi:10.1093/bioinformatics/btx153
72. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. Genome Research. 2017;27: 722–736. doi:10.1101/gr.215087.116
73. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19: 460. doi:10.1186/s12859-018-2485-7
74. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100. doi:10.1093/bioinformatics/bty191
75. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015;6: 11. doi:10.1186/s13100-015-0041-9
76. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013. Available: http://www.repeatmasker.org
77. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. PNAS. 2020;117: 9451–9457. doi:10.1073/pnas.1921046117
78. Pryszcz LP, Gabaldón T. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 2016;44: e113–e113. doi:10.1093/nar/gkw294
79. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio]. 2013 [cited 3 Jun 2020]. Available: http://arxiv.org/abs/1303.3997
80. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

81. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature Biotechnology. 2013;31: 1119–1125. doi:10.1038/nbt.2727
82. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Systems. 2016;3: 99–101. doi:10.1016/j.cels.2015.07.012
83. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biology. 2004; 9.
84. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. New York, NY: Springer; 2019. pp. 227–245. doi:10.1007/978-1-4939-9173-0_14
85. Fredman D, Schwaiger M, Rentzsch F, Technau U. Nematostella vectensis transcriptome and gene models v2.0. 2013. doi:10.6084/m9.figshare.807696.v1
86. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011;29: 644–652. doi:10.1038/nbt.1883
87. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152. doi:10.1093/bioinformatics/bts565
88. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22: 1658–1659. doi:10.1093/bioinformatics/btl158
89. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc. 2013;8. doi:10.1038/nprot.2013.084
90. Srivastava A, Malik L, Sarkar H, Zakeri M, Soneson C, Love MI, et al. Alignment and mapping methodology influence transcript abundance estimation. bioRxiv. 2019; 657874. doi:10.1101/657874
91. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21. doi:10.1093/bioinformatics/bts635
92. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32: 767–769. doi:10.1093/bioinformatics/btv661
93. Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: Incorporating Protein Homology Information into Gene Prediction with GeneMark-EP and AUGUSTUS. 2018; 1.
94. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 2006;7: 62. doi:10.1186/1471-2105-7-62
95. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol Biol Evol. 2017;34: 2115–2122. doi:10.1093/molbev/msx148
96. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47: D427–D432. doi:10.1093/nar/gky995
97. Eddy SR. Accelerated Profile HMM Searches. PLOS Computational Biology. 2011;7: e1002195. doi:10.1371/journal.pcbi.1002195
98. de Mendoza A, Sebe-Pedros A, Sestak MS, Matejcic M, Torruella G, Domazet-Loso T, et

al. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proceedings of the National Academy of Sciences. 2013;110: E4858–E4866. doi:10.1073/pnas.1311818110
99. Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genome Res. 2002;12: 656–664. doi:10.1101/gr.229202
100. Mazza ME, Pang K, Martindale MQ, Finnerty JR. Genomic organization, gene structure, and developmental expression of three Clustered otx genes in the sea anemone Nematostella vectensis. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution. 2007;308B: 494–506. doi:10.1002/jez.b.21158
101. Matus DQ, Thomsen GH, Martindale MQ. Dorso/Ventral Genes Are Asymmetrically Expressed and Involved in Germ-Layer Demarcation during Cnidarian Gastrulation. Current Biology. 2006;16: 499–505. doi:10.1016/j.cub.2006.01.052
102. Hudry B, Thomas-Chollier M, Volovik Y, Duffraisse M, Dard A, Frank D, et al. Molecular insights into the origin of the Hox-TALE patterning system. Tautz D, editor. eLife. 2014;3: e01939. doi:10.7554/eLife.01939
103. Mazza ME, Pang K, Reitzel AM, Martindale MQ, Finnerty JR. A conserved cluster of three PRD-class homeobox genes (homeobrain, rx and orthopedia) in the Cnidaria and Protostomia. EvoDevo. 2010;1: 3. doi:10.1186/2041-9139-1-3
104. Ying H, Hayward DC, Cooke I, Wang W, Moya A, Siemering KR, et al. The Whole-Genome Sequence of the Coral Acropora millepora. Genome Biol Evol. 2019;11: 1374–1379. doi:10.1093/gbe/evz077
105. Shinzato C, Shoguchi E, Kawashima T, Hamada M, Hisata K, Tanaka M, et al. Using the *Acropora digitifera* genome to understand coral responses to environmental change. Nature. 2011;476: 320–323. doi:10.1038/nature10249
106. Consortium* TC elegans S. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science. 1998;282: 2012–2018. doi:10.1126/science.282.5396.2012
107. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al. The Release 6 reference sequence of the Drosophila melanogaster genome. Genome Res. 2015;25: 445–458. doi:10.1101/gr.185579.114
108. Hydra 2.0 Web Portal. [cited 11 Jun 2020]. Available: https://research.nhgri.nih.gov/hydra
109. Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, et al. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nature Genetics. 2016;48: 427–437. doi:10.1038/ng.3526
110. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37: 1530–1534. doi:10.1093/molbev/msaa015
111. Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics. 2003;19: 301–302. doi:10.1093/bioinformatics/19.2.301
112. dos Reis M, Thawornwattana Y, Angelis K, Telford MJ, Donoghue PCJ, Yang Z. Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. Current Biology. 2015;25: 2939–2950. doi:10.1016/j.cub.2015.09.066
113. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421. doi:10.1186/1471-2105-10-421
114. Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Cho J, et al. WormBase: a modern Model Organism Information Resource. Nucleic Acids Res. 2020;48: D762–D767.

doi:10.1093/nar/gkz920