Robust 3D modelling reveals spatiosyntenic properties of animal genomes

Tereza Clarence

The Francis Crick Institute

Nicolas Robert

University of Vienna <https://orcid.org/0000-0002-7367-2241>

Fatih Sarigol

University of Vienna <https://orcid.org/0000-0002-5842-6114>

Xiao Fu

The Francis Crick Institute <https://orcid.org/0000-0002-8611-891X>

Paul Bates

The Francis Crick Institute <https://orcid.org/0000-0003-0621-0925>

Oleg Simakov (✉ oleg.simakov@univie.ac.at)

University of Vienna <https://orcid.org/0000-0002-3585-4511>

Brief Communication

Keywords: 3D modelling, animal genomes

Posted Date: August 10th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-757454/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Title: Robust 3D modelling reveals spatiosyntenic properties of animal genomes

Authors: Tereza Clarence^{1,^}, Nicolas Serge Matthieu Robert^{2,^}, Fatih Sarigol^{2,^}, Xiao Fu¹, Paul A. Bates^{1,*}, Oleg Simakov^{2,*}

- 1 Biomolecular Modelling Laboratory, The Francis Crick Institute, 1 Midland Rd, London NW1 1AT, UK
- 2 Department of Neuroscience and Developmental Biology, University of Vienna, Austria

* Correspondence to Oleg Simakov (oleg.simakov@univie.ac.at) and Paul A. Bates (paul.bates@crick.ac.uk)

Abstract:

Animal genomes are organized into chromosomes that are remarkably conserved in their gene content, forming distinct evolutionary units (macrosynteny). We developed a novel three-dimensional chromosomal modelling approach to show that syntenic signals are reflected in conserved three-dimensional networks, encompassed within interaction spheres. We show evidence for evolutionary constraints that could not be surmised by genomic sequence alone, thereby underlining the importance of three-, rather than just two-, dimensional organization.

Main text:

Gene order, defined solely based on the one-dimensional chromosomal location, is largely preserved at local and chromosomal genomic levels across vast evolutionary distances in animals (600 million years¹⁻³) - termed micro and macrosynteny, respectively. Previous reports show evidence for functional linkages of genes in microsynteny^{1,4,5}. However, little is known about their three-dimensional (3D) organization across animals and whether there is any selective pressure to maintain the very visibly preserved chromosomal organization (macrosynteny) across species⁶. This observation of high preservation of macrosyntenic linkages in most animal species hints at some constraints, but any functional inference has been lacking due to the absence of chromosomal structural information⁷. Here we describe a novel method, 3D-HiChrom (Figure 1a), that utilizes available Hi-C information to reconstruct 3D models of chromosomes. We use it to profile 3D property of genes that constitute macro (and microsyntenic) groups revealing that conventional syntenic is reflected by contact-rich spatial organization, as well as revealing genomic regions that are “spatiosyntenic” i.e., genomically distant but physically close.

Our approach utilizes a model consisting of beads on a string, with bead diameters dependent upon the resolution for which the Hi-C data relates (Figure 1a). Beads are initially randomly dispersed and after multiple cycles imposed by the algorithm converge to stable chromosome topologies. During each cycle, beads slowly adjust their positions dependent upon simple physical forces in conjunction with the Hi-C constraints (see Figure 1 and Methods for algorithmic details). As Hi-C quality is highly variable between species, we developed a partitioning approach where we subdivide the chromosomes into interaction spheres (Figure 1a). This partitioning approach allows us to confidently profile in a species-dependent manner interactions within each ‘interaction sphere’ (IntSph) (Supplementary Figure 1C-E). Outliers on a per-chromosome basis may reflect actual chromosomal dynamics or assembly issues.

With the models and orthology information (Methods) we have asked whether any of the orthologous co-localization at gene pair, macro- and microsyntenic levels are reflected in the three-dimensional structures. For this, we profiled the Euclidean (3D) and genomic (2D)

distances between orthologous bins along the chromosomes of two evolutionary divergent species - the chordate amphioxus and the mollusc blood clam (Supplementary Figure 2). The two species span the ancient bilaterian split in the animal phylogeny (over 500 million years ago⁸) providing for the highest evolutionary conservation signal. They also show relatively good quality of Hi-C data, which is still lacking for many of the invertebrate clades and will become crucial for further elaboration of the results presented here. To this end, we provide chromosomal models for other species in our analysis (Supplementary Figure 3), however where Hi-C resolution can still be improved.

We find that ortholog gene pairs, if they are located on the same chromosome in both species, tend to localize within a partial 3D vicinity of around 200-300nm. While this reflects overall chromosomal folding, in randomly assigned orthologies such mid-distance interactions (when normalized by the genomic distance) were depleted (Supplementary Figure 4). To further investigate this result in the context of chromosomal homologies, we identified homologous chromosomes between the two distantly related animals (Supplementary Figure 5a). Previous studies have shown patterns of largely one-to-one correspondence between animal chromosomes, with individual genes retaining their chromosomal identity (yet in a scrambled order) in multiple species^{3,9}. Conversely, around half of orthologous genes move and disperse across multiple animal chromosomes, resulting in an off-diagonal dotted pattern. Genomes can thus be partitioned into genes which stay within the homologous chromosomes (ancestral linkage groups, ALGs) and those that disperse or move to other chromosomes (losing their ancestral linkage group identity, nonALG). We tested whether genes that are maintained in their ALG identity, and thus define the macrosyntenic pattern, are more likely to come into interactions with each other than dispersed genes.

We investigated contact density between bins within each so-called ALG IntSph (consisting of at least one ALG) and nonALG IntSph (lacking the presence of any ALG genes). Similarly to orthologous gene pairs, we found that the ALG IntSphs display more contacts between the ALG gene containing bins (Figure 1b) than nonALG IntSphs and were forming more contacts within 200-300nm range (Figure 1c). This observation is striking in the context that ALG-ALG contacts were less abundant than nonALG-nonALG contacts (altogether 32,352 ALG-ALG contacts compared to 92,632 nonALG-nonALG in amphioxus, Supplementary Figure 5b,c). Interestingly, we find that while this pattern is overall conserved between homologous chromosomes (Supplementary Figure 5c), some chromosomes stand out. In particular, chromosomes 3 and 4 in amphioxus are particularly enriched for nonALG-nonALG contacts (contact score 28 and 25 for chromosome 3 and 4, respectively) while having a very low contact score for ALG-ALG contacts (12 and 10, respectively). These chromosomes have been shown to undergo recent fusions⁶. Together, this suggests presence of evolutionary constraints to maintain global chromosomal organization.

To further understand properties of local genome organization, we assessed 3D contacts between known microsyntenic regions (defined here as local conserved gene clusters of 3 or more genes, Methods). We observed that IntSphs consisting of at least one microsynteny tend to have significantly higher number of contacts (mean 7.68 and median 6.0 interacting partners) compared to IntSphs with only randomly sampled microsynteny⁵ (mean 6.74 and median 5.0 interacting partners, Figure 2a - p-value 6.5×10^{-4} and 3.9×10^{-10} for amphioxus and blood clam, respectively). This suggests that local microsyntenic linkages also help form 3D interaction hubs.

While little is known about the functionality of the majority of observed microsyntenic linkages, we investigated the folding around well studied and often syntenic homeobox genes¹⁰⁻¹². While the Hox cluster itself forms a tight microsynteny in most animal genomes examined to-date, genes from the proposed ancestral SuperHox and NK clusters^{13,14} are usually dispersed along their respective chromosomes (Figure 2b,c). While amphioxus 3D

model found evidence for co-localization of the proposed SuperHox cluster genes within a single 'sphere' of 75-100nm, in blood clam the genes appeared more dispersed (Supplementary Figure 6). Nevertheless, our analysis showed that 14 genes were located in spatial proximity (within a single IntSph of radius 50nm) of the proposed SuperHox cluster genes in both blood clam and amphioxus and thus form shared conserved interactors (Figure 2b). The region of the mouse genome spanning three of these genes (*atp5g3*, *atf2* and the *jazf2* pseudogene) has been found in the local vicinity of and interacting with the promoters of *hoxd* genes and regulate their transcription¹⁵. This contrasts with their syntenic state in amphioxus and blood clam: while *atp5g3*, *atf2* and *jazf2* are located on the same chromosome as the *hox* genes in these animals, they are not in the genomic vicinity of the *hox* genes. Our inference of their interactions in invertebrates thus suggests importance of the retained topological, rather than local syntenic, interaction around this cluster¹¹. Similarly, we can also detect six shared genes in the NK chromosomal network (Figure 2c), however their function is largely unknown. In spite of the substantial gene scrambling and disruption of many microsyntenic linkages between homologous chromosomes, this analysis points to the existence of evolutionary 3D constraints upon chromosomal organization that can only be detected via 3D modelling approach.

We have presented a robust method to quantify shapes of animal chromosomes. Since our 'interaction sphere' approach allows for approximation to compensate for Hi-C quality, the method can be readily applied to an increasing number of species for which Hi-C data sets are accumulating. Our study focuses on implementation of this method for two phylogenetically divergent bilaterian species, as well as describing its application in other species, which is currently impeded by the Hi-C quality. Combined with comparative genomic analyses of gene linkages at several scales of genomic organization, we show evidence for the existence of conserved 3D constraints on genome folding. In particular, we show evidence that genes within both macro- and microsyntenies display more 3D contacts than non-syntenic (and randomly sampled) regions. Our novel methodology paves the way for further detection of topologically and evolutionarily conserved genomic regions (spatiosynteny), providing testable hypotheses for their functional profiling.

References

1. Irimia, M. *et al.* Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res.* **22**, 2356–2367 (2012).
2. Simakov, O. *et al.* Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**, 526–531 (2013).
3. Putnam, N. H. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071 (2008).
4. Kikuta, H. *et al.* Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17**, 545–555 (2007).
5. Zimmermann, B., Robert, N. S. M., Technau, U. & Simakov, O. Ancient animal genome architecture reflects cell type identities. *Nat. Ecol. Evol.* **3**, 1289–1293 (2019).
6. Simakov, O. *et al.* Deeply conserved synteny resolves early events in vertebrate evolution. *Nat. Ecol. Evol.* **4**, 820–830 (2020).
7. Schaeffer, S. W. Muller “Elements” in *Drosophila* : How the Search for the Genetic Basis for Speciation Led to the Birth of Comparative Genomics. *Genetics* **210**, 3–13 (2018).
8. Evans, S. D., Hughes, I. V., Gehling, J. G. & Droser, M. L. Discovery of the oldest bilaterian from the Ediacaran of South Australia. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 6 (2020).
9. Srivastava, M. *et al.* The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* **466**, 720–726 (2010).
10. Duboule, D. The rise and fall of Hox gene clusters. *Development* **134**, 2549–2560 (2007).
11. Acemel, R. D. *et al.* A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. *Nat. Genet.* **48**, 336–341 (2016).
12. Nong, W. *et al.* Jellyfish genomes reveal distinct homeobox gene clusters and conservation of small RNA processing. *Nat. Commun.* **11**, 3051 (2020).
13. Butts, T., Holland, P. W. H. & Ferrier, D. E. K. The Urbilaterian Super-Hox cluster. *Trends Genet.* **24**, 259–262 (2008).
14. Ferrier, D. E. K. Evolution of Homeobox Gene Clusters in Animals: The Giga-Cluster and Primary vs. Secondary Clustering. *Front. Ecol. Evol.* **4**, (2016).

15. Montavon, T. *et al.* A Regulatory Archipelago Controls Hox Genes Transcription in Digits. *Cell* **147**, 1132–1145 (2011).
16. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
17. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
18. Van Dongen, S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. & Appl.* **30**, 121–141 (2008).
19. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
20. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**, e9490 (2010).
21. Cannon, J. T. *et al.* Xenacoelomorpha is the sister group to Nephrozoa. *Nature* **530**, 89–93 (2016).
22. Dunn, C. W., Giribet, G., Edgecombe, G. D. & Hejnol, A. Animal Phylogeny and Its Evolutionary Implications. *Annu Rev Ecol Evol Syst* **45**, 371–395 (2014).
23. Kapli, P. *et al.* Lack of support for Deuterostomia prompts reinterpretation of the first Bilateria. *Sci. Adv.* **7**, eabe2741 (2021).
24. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
25. Zhong, Y. & Holland, P. W. H. HomeoDB2 : functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evol. Dev.* **13**, 567–568 (2011).
26. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
27. Leinonen, R., Sugawara, H., Shumway, M., & on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).

28. Guo, Y. *et al.* A chromosomal-level genome assembly for the giant African snail *Achatina fulica*. *GigaScience* **8**, giz124 (2019).
29. Bai, C.-M. *et al.* Chromosomal-level assembly of the blood clam, *Scapharca (Anadara) broughtonii*, using long sequence reads and Hi-C. *GigaScience* **8**, giz067 (2019).
30. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
31. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
32. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
33. Paulsen, J., Liyakat Ali, T. M. & Collas, P. Computational 3D genome modelling using Chrom3D. *Nat. Protoc.* **13**, 1137–1152 (2018).
34. Lando, D., Stevens, T. J., Basu, S. & Laue, E. D. Calculation of 3D genome structures for comparison of chromosome conformation capture experiments with microscopy: An evaluation of single-cell Hi-C protocols. *Nucleus* **9**, 190–201 (2018).
35. Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).

Acknowledgements:

This research was funded in whole, or in part, by the Wellcome Trust (FC001003). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

O.S. is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 945026).

For authors TC, XF and PAB this work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001003), the UK Medical Research Council (FC001003), and the Wellcome Trust (FC001003). N.S.M.R., F.S., and O.S. were supported by the Austrian Science Fund FWF grant P32190.

Figure 1. Hi-Chrom modelling pipeline and its utilization for macrosyteny interactions.

A) Schematic of 3D modelling tools and analysis pipeline; first row depicts the workflow of Hi-C pre-processing and modelling. Second row shows mapping of genomic features and their consequent structural analysis. B) Distribution of ALG or nonALG contact density within the interaction sphere (radius 50nm) for blood clam (left) and amphioxus (right), respectively. C) Density hexplot depicting relationship between genomic and Euclidean distance of ALG or nonALG contact pairs for amphioxus and blood clam, respectively.

Figure 2. Genes in microsynteny form local 3D interaction hubs. A) Distribution of observed vs random microsynteny contact number [defined as] within the interaction sphere (radius 50nm) for amphioxus (left) and blood clam (right), respectively. B, C) Interaction network of SuperHox (green, C) and NK (purple, D) clusters in amphioxus and blood clam, respectively. Circled numbers correspond to putative shared interactors of SuperHox/NK clusters between amphioxus and blood clam.

Supplementary Figure 1. Interaction profiling effectively reflects on chromosome geometry. A) Proportion of satisfied vs. unsatisfied constraints per chromosome model for amphioxus and blood clam, respectively. B) Cosine similarity between Hi-C contacts used for modelling and reconstructed model contacts per chromosome for amphioxus and blood clam, respectively. C) Varying sizes of IntSph and their impact on the capture of local interactions. *B. floridae* (D) and *A. broughtonii* (E), respectively, and their spatial interaction profiles utilizing distinct interaction sphere radii; 50 (red), 75 (ochre), 100 (green), 125 (blue) and 150 (purple). Chromosomes scaffolds were reconstructed with 150kb-bin resolution. A) Dark grey regions on *B. floridae* chromosomal scaffold 1 and 3 corresponds to low interaction valence and are located on the chromosomal periphery. On the contrary, dark region on scaffold 2 corresponds to a deeply nested fragment with high interaction valence. B) Dark grey regions on *A. broughtonii* chromosomal scaffold 4 (cyan) and 7 (purple) corresponds to low interaction valence and are located on the chromosomal surface, more isolated from the rest of the chromosome. Dark region on the scaffold 1 (light orange) points to genomic positions with highest interaction valence, despite being located more towards chromosomal surface. In this case, it is the indicative of the overall bending of the model where both sides of chromosomal model are coming into spatial proximity thus enhancing the interaction valence.

Supplementary Figure 2. 3D models of chromosomal scaffolds in amphioxus and blood clam. Circos plot depicting homology between chromosomes of amphioxus and blood clam.

Dark orange links correspond to 1:1 homology whilst light orange links show 1:many homology. 3D models of chromosomal scaffolds surround the circus plots; with each chromosome color-coded in spectrum with blue as a start of chromosome and red as the end of chromosome. 3D models of chromosomal scaffolds in blood clam adopt warm-like structures with lots of bending and twists present, while amphioxus models show more globular, bundle-like structures. Small circles next to the chromosomes indicate presence of clusters of interest.

Supplementary Figure 3. Examples of chromosomal scaffold models from Hi-Chrom.

Model representations of 8 largest chromosomes in *A. fulica* (A), *P. maximus* (B) and *R. esculentum* (C), respectively. Models are coloured based on spectrum (blue = start of chromosome, red = end of chromosome).

Supplementary Figure 4. The relationship between genomic and Euclidean distances in the sets of observed and randomized ortholog pairs. Orthology was defined by best reciprocal BLAST similarity. Only orthologous pairs that map to the same set of chromosomes were considered. A,D) Euclidean distances between ortholog pairs in observed and fully randomized orthology set, respectively, reflecting the overall shape of the chromosomes; B,E) Genomic distances between ortholog pairs in observed and fully randomized orthology set, respectively; C,F) Ratios between Euclidean and Genomic distances between orthologous gene pairs in observed and randomized orthologies, respectively. Observed orthologous gene pairs are more commonly found in 2×10^{-4} - 4×10^{-4} nm/bp region, indicating that more distant interactions between orthologs are present in the observed set. Zero distance (within the same bin) interactors are excluded.

Supplementary Figure 5. Macrosynteny conservation in 3D. A) Oxford dot plot depicting ALG homology between amphioxus and blood clam. B) Genomic position of ALGs (red) and nonALGs (blue) along chromosomes in amphioxus and blood clam, respectively. C) Mean ALG-ALG vs. nonALG-nonALG interaction pairs within IntSph (radius 50nm) per chromosome(left) together with mean ALG-empty vs. nonALG-empty interaction (right) for amphioxus and blood clam, respectively. D) Total number of ALG (coral) and nonALG (light cyan) genes per chromosome for amphioxus and blood clam, respectively.

METHODS

Comparative genomics

Orthology assignment

Genomes, annotations and protein sequences of 80 species were obtained from the databases summarized in Supplementary table 1. Only the longest isoform of each gene was retained. Orthogroups were identified using Orthofinder v2.4.1¹⁶, in conjunction with diamond 0.9.36¹⁷, and MCL 14.137¹⁸. Identification of Hierarchical Orthogroups at the root node (root node HOGs) is based on orthogroup gene trees and species trees that were built using both MAFFT 7.427¹⁹ and FastTree 2.1.11²⁰. For all the following analyses, we consider genes belonging to the same root node HOG to be orthologs.

Identification of microsyntenic blocks

Microsyntenic blocks were inferred using methods described in ⁵. In order to determine which microsyntenic blocks of *Achatina fulica*, *Anadara broughtonii* and *Branchiostoma floridae* were present in the Nephrozoan Last Common Ancestor (Nephrozoan blocks), we used two cladograms (Supplementary File X). One was based on the Nephrozoa hypothesis and monophyletic deuterostomes^{21,22}, while the other was based on the Xenacoelomorpha hypothesis with paraphyletic deuterostomes (after ref. ²³). Since both methods produced broadly similar sets of blocks (overlap coefficient of 85 to 98 % for the three species), only the first set was kept for subsequent analyses. Nephrozoan blocks were defined as blocks found in at least two species of each ingroup or of one ingroup and the outgroup. To obtain background information, we employed two block randomization methods. For each observed block, 100 random blocks were sampled either across the whole genome, as described in ref. ⁵, or only across the same chromosome bearing the observed block.

Macrosynteny dotplots and assignment of genes to ancestral linkage groups

Pairwise ortholog oxford plots for all possible species pairs between *A. fulica*, *A. broughtonii*, *B. floridae* and *Rhopilema esculentum* were generated, using the previously identified root node HOGs, but retaining only one-to-one orthologs for each species pair. Chromosome homologies were assessed with a Fisher exact test against a null model of gene permutation (i.e. if $p < 0.05$, chromosomes were considered to be homologous), with Benjamini-Hochberg correction for false discovery rate for the multiple tests done for each species pair.

Another set of ortholog oxford plots between all the possible species pairs between *A. fulica*, *A. broughtonii* and *B. floridae* were also built. Here, one-to-one orthologs were defined based on a reciprocal best hit approach, using *B. floridae* as a reference. Only the proteins of *B. floridae* with a reciprocal best hit in *A. fulica* and *A. broughtonii* were retained (4866 groups of 3 proteins each). Each group was assigned to the same BLG as the *B. floridae* protein they comprised according to ref. ⁶.

Annotation of homeodomain proteins

The homeodomain HMM profile from Pfam (<http://pfam.xfam.org/family/pf00046>) was used to query the proteomes of *B. floridae*, *A. broughtonii* and *A. fulica*, using hmmsearch from the HMMER 3.3 package²⁴ with a 0.1 e-value threshold. Homeodomain candidates were then used to query the HomeoDB2 database²⁵ using the blastp utility from the BLAST+ package. If at least 8 out of the 10 top hits of a query were of the *Antennapedia-class* (ANTP), they were considered as ANTP candidates.

ANTP candidates were then aligned to the homeodomains of the ANTP class from HomeoDB2 using Mafft 7.427 with default parameters. The alignment was trimmed using the gappyout algorithm of trimAl²⁶ and used to infer a phylogenetic tree with FastTree 2.1.11²⁰. This tree was used to isolate the *B. floridae*, *A. broughtonii* and *A. fulica* orthologs to the members of the SuperHox (*hox* genes, *dlx*, *en*, *evx*, *gbx*, *hhx*, *meox*, *mnx*, *nedx* and *ro*), the ParaHox (*cdx*, *gsx* and *pdx*), the NK (*nk* genes, *lhx*, *lhx*, *msx*, *tlx* and *ventx*) and NK2 cluster (*msxl*, *nk2.1*, *nk2.2*) (reviewed in ref. ¹⁴) In addition, a second round of search for orthologs was performed. To this end, all the proteins of *B. floridae*, *A. broughtonii* and *A. fulica* found in the same Orthofinder root node HOGs as the already isolated SuperHox, ParaHox, NK and NK2 cluster genes were recovered. These candidates were used as queries against the BLAST nr database. If the top hits were annotated members of the SuperHox, ParaHox, NK or NK2 clusters, they were added to the existing list of putative orthologs.

Hi-C analysis

Hi-C sequencing libraries were downloaded from NCBI Sequence Read Archive using SRA Toolkit²⁷. Hi-C data of *A. fulica*²⁸ (accession ID SRX5181756), *B. floridae*⁶ (accession ID SRX3274438) and *A. broughtonii*²⁹ (accession ID SRX5337861), together with their reference genomes were used for preprocessing. Sequence quality was evaluated using FastQC 0.11.8 (<https://github.com/s-andrews/FastQC>) and HiC-Pro 2.11.1 software³⁰ was used to generate interaction matrices between 150kb windows of each chromosome. Bowtie2³¹ and samtools 1.11³², as a part of HiC-Pro, were utilized to map the Hi-C reads to the reference genome assemblies. Specific parameters for HiC-Pro run are listed in Table 1.

Table 1. Parameter settings used for HiC-Pro pre-processing

BOWTIE2_GLOBAL_OPTIONS	--very-sensitive -L 30 --score-min L,-0.6,-0.2 --end-to-end --reorder
BOWTIE2_LOCAL_OPTIONS	--very-sensitive -L 20 --score-min L,-0.6,-0.2 --end-to-end --reorder
GET_ALL_INTERACTION_CLASSES	1
GET_PROCESS_SAM	1
RM_SINGLETON	1
RM_MULTI	1
RM_DUP	1
MATRIX_FORMAT	upper
MAX_ITER	100
FILTER_LOW_COUNT_PERC	0.02
FILTER_HIGH_COUNT_PERC	0
EPS	0.1

3D modelling and SASA analysis

3D model generation

The 3D structure of individual chromosomes was constructed using a home-built C++ software, motivated by studies *Chrom3D*³³ and *NucDynamics*^{34,35}. Each chromosome has a

beads-on-a-string representation and starts with a randomized conformation. Then, the time evolution of chromosome conformation is governed by the Newton equation of motion, with forces (detailed below) implemented to characterize the chromosome structural integrity ($\rightarrow F_i^{ten}$), volume exclusion between spatially overlapping genomic sites ($\rightarrow F_i^{rep}$), drag by nucleoplasm ($-\gamma \rightarrow v_i$), and genomically distant interactions suggested by Hi-C ($\vec{F}_i^{Hi-C}$).

The dynamics

The dynamics of a coarse-grained chromatin bead i is governed by the following Newtonian equation of motion:

$$m\vec{a}_i = -\gamma\vec{v}_i + \vec{F}_i^{rep} + \vec{F}_i^{ten} + \vec{F}_i^{Hi-C}$$

where $\vec{a}_i$ and $\vec{v}_i$ are the instantaneous acceleration and velocity of the bead, respectively; m is the mass of the bead; γ is the drag coefficient; $\vec{F}_i^{rep}$, $\vec{F}_i^{ten}$, and $\vec{F}_i^{Hi-C}$ are forces implemented in the model to characterise the mutual volume exclusion between beads, the interaction between genomically consecutive beads, and the interaction between genomically distant beads with high Hi-C frequency. Computationally, Verlet integration is applied to calculate the trajectories of chromosome beads over time.

The volume exclusion force

The volume exclusion between any two spatially overlapping beads is assumed linearly elastic. The contribution of this force to a bead i is described by the following equation:

$$\vec{F}_i^{rep} = \sum_{j \neq i}^N K^{rep} (d_{i,j} - d_{rep0}) \hat{u}_{i,j}, \text{ if } d_{i,j} < d_{rep0},$$

where K^{rep} is the spring constant reflecting the incompressibility of genetic content within the beads in contact; $d_{i,j}$ is the distance between the centre of two consecutively connected beads i and j ; d_{rep0} is the rest length of the linearly elastic spring (in our case $d_{rep0} = 2 * r_{bead}$); $\hat{u}_{i,j}$ is a unit vector pointing from bead i to bead j .

The chromatin tension force

The interaction between two genomically consecutive beads is assumed linearly elastic. The contribution of this force to a bead i is described by the following equation:

$$\vec{F}_i^{ten} = K^{ten} (d_{i,i-1} - c_2) \hat{u}_{i,i-1} + K^{ten} (d_{i,i+1} - c_2) \hat{u}_{i,i+1}$$

where K^{ten} is the spring constant of the inter-bead 'chromatin' linker, $d_{i,i+1}$ is the distance between the centre of two consecutively connected beads i and $i + 1$; c_2 is the rest length of the linearly elastic spring; $\hat{u}_{i,i+1}$ is a unit vector pointing from bead i to bead $i + 1$.

The Hi-C restraint force

The interaction between genomically distant beads is also assumed linearly elastic. The contribution of this force to a bead i is described by the following equation:

$$\vec{F}_i^{Hi-C} = \sum_{j \neq i}^M K^{Hi-C} (d_{i,j} - d_{Hi-C0}) \hat{u}_{i,j}, \text{ if } p_{i,j} > p_{rep0}$$

where K^{Hi-C} is a constant reflective of the constraint strength implied by HiC and applies to any pairs of coarse-grained beads that have pairwise Hi-C frequency greater than a threshold value, namely, $p_{i,j} > p_{rep0}$; d_{Hi-C0} is the rest length of the linearly elastic spring; $\hat{u}_{i,j}$ is a unit vector pointing from bead i to bead j .

Data preparation and modelling

Normalized HiC-Pro sparse matrix was parsed into matrices of separate chromosomes, containing only intra-chromosomal contacts without any scaffold interactions. We neglected all inter-chromosomal interactions due to the multi-cell nature of the Hi-C experiment. Such matrices were then used as an input to our chromatin reconstruction program with specific cut-offs to filter interaction frequencies used as reconstruction constraints.

High number of contact restraints results in very dense and compact chromatin models. Such structures might be lacking desired biological relevance, however there is no current knowledge about optimal number of contacts per selected genomic region, therefore we used a mean value of interaction frequency as a Hi-C cut-off and all interactions below selected Hi-C threshold were neglected. The total number of *cis*-contacts per chromosome together with number of used constraints for reconstruction with different Hi-C threshold is shown in Table 2,3 and 4, for blood clam, amphioxus and jellyfish, respectively. Structural measurements of mapped syntenic blocks (such as SASA, coverage, depth) can be, to some extent, affected by the number of constraints and thus compactness of chromatin model. We keep the selection criteria for Hi-C threshold consistent among individual chromosomes to mitigate this impact. We reconstructed chromosomal scaffolds of blood clam and amphioxus and jellyfish with different interaction frequency (IF) cut-off thresholds (Table 2-4) and compare the models. Overall, the geometry and fines topology of the models was very similar between different IF cut-off threshold (data not shown). The higher the number of constraints utilized to build a model, the more densely packed the chromosomal scaffold was. Since the quality of the Hi-C datasets utilized is variable, we selected models from mean IF cut-off to be further analysed, in order to capture majority of interactions obtained from Hi-C.

All chromosomes were reconstructed with three replicates and each model was initialized with different conformation based on principles of self-avoiding random walk (SAWR). The reconstruction process starts after initialization, when the pseudo-energy of 3D chromosome conformations, calculated as the sum of kinetic and potential energy in the system, is monitored throughout the simulation as an indicator of convergence of the system. This is then accompanied by root-mean square (RMSD) analysis across all the time-point structures towards the final structure. We ran the reconstruction algorithm for 10,000-time steps and the final chromosome structure of each replicate run was then taken for further analysis. In order to validate correlation of our models with Hi-C IF map, we calculated cosine similarity between IF contacts, which were selected as restraints for 3D modelling, and Euclidean distance of corresponding genomic position in the model (Supplementary Figure 1b). In addition, we calculated the proportion of satisfied/violated contacts (Supplementary Figure 1a).

Table 2. Hi-C constraints applied for single chromosome model reconstruction from mean/median and 35th percentile IF threshold cut-off in amphioxus.

Config	Mean IF	Median IF	35th percentile	mean IF %	median IF %	35th IF %
<i>Sc7u5tJ_339</i>	4.71	2	1.5	86	50.01	35
<i>Sc7u5tJ_1590</i>	4.58	1.98	1.47	85.76	50.01	35.01
<i>Sc7u5tJ_366</i>	4.79	1.99	1.48	86.56	50.01	35
<i>Sc7u5tJ_1568</i>	61.89	42.35	7.5	57.81	50	35.94
<i>Sc7u5tJ_1587</i>	5.94	2.37	1.83	86.58	50	34.99
<i>Sc7u5tJ_1571</i>	4.02	1.67	1.25	86.4	50	35
<i>Sc7u5tJ_1559</i>	3.83	1.58	1.17	85.79	50	35
<i>Sc7u5tJ_320</i>	6	2.26	1.71	88.29	50.01	35.01
<i>Sc7u5tJ_1517</i>	5.45	2.27	1.69	85.77	50	35
<i>Sc7u5tJ_350</i>	3.63	1.62	1.21	85.06	50	35
<i>Sc7u5tJ_1565</i>	7.11	2.5	1.89	89.12	50.01	35.01
<i>Sc7u5tJ_566</i>	5.46	2.12	1.58	87.64	50	35
<i>Sc7u5tJ_1552</i>	5	2.16	1.63	85.84	50	35.01
<i>Sc7u5tJ_1442</i>	3.48	1.55	1.16	84.24	50	35
<i>Sc7u5tJ_1579</i>	5.05	2.25	1.72	85.58	50	35
<i>Sc7u5tJ_190</i>	4.99	2.05	1.55	86.93	50	35
<i>Sc7u5tJ_1485</i>	5.75	2.45	1.87	86.56	50.01	35
<i>Sc7u5tJ_1398</i>	4.5	1.71	1.29	87.71	50	35
<i>Sc7u5tJ_417</i>	5.33	2.2	1.67	86.18	50	35

Table 3. Hi-C constraints applied for single chromosome model reconstruction from mean/median and 35th percentile IF threshold cut-off in blood clam.

Config	Mean IF	Median IF	35th percentile	Mean IF %	Median IF %	35th IF %
<i>Lachesis_group0</i>	7.45	3.02	2.04	78.26	50	35
<i>Lachesis_group1</i>	8.39	3.28	2.18	78.07	50	35
<i>Lachesis_group2</i>	8.31	3.27	2.13	77.65	50	35
<i>Lachesis_group3</i>	9.07	3.42	2.22	77.63	50	35
<i>Lachesis_group4</i>	8.17	3.51	2.41	77.23	50	35
<i>Lachesis_group5</i>	9.2	3.89	2.62	77.12	50	35
<i>Lachesis_group6</i>	8.75	3.77	2.58	78.44	50	35
<i>Lachesis_group7</i>	9.67	4.26	2.81	76.48	50	35
<i>Lachesis_group8</i>	9.58	4.06	2.68	77.06	50	35
<i>Lachesis_group9</i>	9.59	4.15	2.73	77.51	50	35
<i>Lachesis_group10</i>	9.56	4.1	2.67	77.47	50	35
<i>Lachesis_group11</i>	9.81	4.08	2.61	76.82	50	35
<i>Lachesis_group12</i>	9.96	3.84	2.49	76.73	50	35

<i>Lachesis_group13</i>	10.39	4.29	2.82	78.13	50	35
<i>Lachesis_group14</i>	11.32	4.83	3.17	76.8	50	35
<i>Lachesis_group15</i>	11.79	5.13	3.45	78.08	50	35
<i>Lachesis_group16</i>	12.9	5.55	3.73	78.02	50	35
<i>Lachesis_group17</i>	12.82	6.03	4.24	78.53	50	35
<i>Lachesis_group18</i>	14.42	6.68	4.74	79.12	50	35

Table 4. Hi-C constraints applied for single chromosome model reconstruction from mean/median and 35th percentile IF threshold cut-off in jellyfish.

Config	Mean IF	Median IF	35th percentile	Mean IF %	Median IF %	35th IF %
SWAQ01000795.1	39	17.9	9.73	76.8	50	35
SWAQ01001661.1	41.65	21.32	15.67	79.76	50.01	35
SWAQ01000001.1	39.12	18.78	12.89	78.51	50.01	35.01
SWAQ01001671.1	45.47	23.98	16.98	77.79	50.01	35
SWAQ01000010.1	44.38	22.56	17.05	78.97	50	35.01
SWAQ01000004.1	55.81	30.84	23.81	79.48	50	35
SWAQ01000217.1	42.96	19.53	12.51	78.88	50.01	34.99
SWAQ01000092.1	47.79	23.42	16.21	78.61	50	35.01
SWAQ01001676.1	50.99	26.12	20.25	80.51	50	35.01
SWAQ01000558.1	46.46	23.34	15.72	79.14	50	35
SWAQ01001658.1	50.39	26.24	15.48	79.23	50.02	35.02
SWAQ01000448.1	55.48	28.41	21.98	80.83	50.02	35
SWAQ01001681.1	48.13	25.24	18.4	79.19	50	35.02
SWAQ01001655.1	53.29	28.33	20.87	80.54	50.02	34.99
SWAQ01000359.1	65.35	35.24	25.45	77.43	50.02	35.01
SWAQ01000409.1	71.88	36.93	27.11	78.17	50	35
SWAQ01000005.1	68.9	35.15	26.22	77.38	50.03	35.01
SWAQ01001657.1	73.96	40.12	30.24	80.65	50	35
SWAQ01000107.1	88.98	46.51	36.26	77.87	50	34.98
SWAQ01000978.1	78.94	41.97	26.12	75.98	50	35.02
SWAQ01000045.1	155.3	5	83.63	62.92	75.27	50.11
	5	83.63	62.92	75.27	50.11	35.05

Model analysis and gene mapping

Genes and microsynteny locations were mapped onto the chromosome models within the 150kb-resolution beads. Due to low resolution, some 150kb-regions include multiple ALGs and nonALGs together. If at least one ALG is present per 150kb-bead, we treat this region as ALGs content only.

Interaction sphere (IntSph) analysis

In order to identify spatio-functional units within chromosome scaffolds, we performed sliding Interaction Sphere (IntSph) analysis; an imaginary sphere with specific radius moving along the chromosome scaffold and detecting spatial contacts within (Figure 1a, Supplementary Figure 1c-e).

The smaller the radius of IntSph, the more the linear contacts are dominant as interacting partners within IntSph. The larger the radius, the more genomically long-range interactions can be included as demonstrated in Supplementary Figure 1c. To measure the IntSph occupancy we defined ‘*contact density*’ as the number of interacting beads within defined radius of IntSph.

Supplementary Data

Supplementary Table 1.

TYPE	SPECIES
NEPHROZOA HYPOTHESIS, MONOPHYLETIC DEUTEROSTOMES	((CREFR,SPHAR)Ichtyosporea,(CAPOW,(SALRO,(((AMPQU,EPHMU)De mospongiae,SYCCI)Porifera,(PLEBA,MNELE)Ctenophora,(((TRIAD,HOIH O)Placozoa,((((HYDVU,CLYHE)Hydrozoa,(MORVI,(SANMA,(RHOES,AU RAU)Discomedusae)Scyphozoa)Acraspeda)Medusozoa,(DENG1,(ACRMI, (EXAPA,ACTEQ)Metridioidea,NEMVE)Actiniaria)Hexacorallia)Anthozoa) Cnidaria,(HOFMI,((((LINAN,(((HELRO,EISAN)Clitellata,CAPTE)Annelida, SCHME)'Parenchymia_Annelida')'Parenchymia_Annelida_Brachiopoda',(((OCTMI,(OCTBI,OCTSI))Octopus,((DORPE,ARCDU)Theutida,EUPSC)De capodiformes)Cephalopoda,(((ACHFU,POMCA)Apogastropoda,LOTGI)G astropoda,(((PECMA,MIZYE)Pectinida,CRAGI)Pteriomorpha,ANABR)Biva lvia)Pleistomollusca)Conchifera)Lophotrochozoa,ADIVA)Spiralia,(PRICA,((((PARTE,IXOSC)Chelicerata,((((DROME,ANOGA)Diptera,TRICA)Apara glossata,APIME)Holometabola,FOLCA)Hexapoda,((PENVA,PORTR)Deca poda,DAPMA)Altocrustacea)Pancrustacea,STRMA)Mandibulata)Arthropo da,RAMVA)Tactopoda,CAEEL))Ecdysozoa)Protostomia,(((ANNJA,((STR PU,APOJA)Asterozoa,(ACAPL,ASTRU)Echinozoa))Echinodermata,(PTYF L,SACKO)Hemichordata)Ambulacraria,(BRAFL,(CIOIN,(CALMI,(((HIPCO, (MAYZE,DANRE)Ovalentaria)Percomorpha,LEPOC)Neopterygii,(LATC H,(((MUSMU,HOMSA)Euarchontoglires,CANLUFA)Boreoeutheria,(((CORC OCO,GALGA)Aves,(CHEMY,PELSI)Testudines),PODMU)Diapsida)Amnio ta,XENLA)Tetrapoda)Sarcopterygii)Osteichthyes)Gnathostomata)Olfactor es)Chordata)Deuterostomia)Nephrozoa)Bilateria)Planulozoa)Parahoxozo a)Metazoa)Choanozoa)Filozoa)Holozoa
XENAMBULACRARIA HYPOTHESIS, PARAPHYLETIC DEUTEROSTOMES	((CREFR,SPHAR)Ichtyosporea,(CAPOW,(SALRO,(((AMPQU,EPHMU)De mospongiae,SYCCI)Porifera,(PLEBA,MNELE)Ctenophora,(((TRIAD,HOIH O)Placozoa,((((HYDVU,CLYHE)Hydrozoa,(MORVI,(SANMA,(RHOES,AU RAU)Discomedusae)Scyphozoa)Acraspeda)Medusozoa,(DENG1,(ACRMI, (EXAPA,ACTEQ)Metridioidea,NEMVE)Actiniaria)Hexacorallia)Anthozoa) Cnidaria,((((LINAN,(((HELRO,EISAN)Clitellata,CAPTE)Annelida,SCHME)'Parenchymia_Annelida')'Parenchymia_Annelida_Brachiopoda',(((OCTMI ,(OCTBI,OCTSI))Octopus,((DORPE,ARCDU)Theutida,EUPSC)Decapodif ormes)Cephalopoda,(((ACHFU,POMCA)Apogastropoda,LOTGI)Gastropo da,(((PECMA,MIZYE)Pectinida,CRAGI)Pectinida,ANABR)Bivalvia)Pleisto mollusca)Conchifera)Lophotrochozoa,ADIVA)Spiralia,(PRICA,((((PARTE,I XOSC)Chelicerata,((((DROME,ANOGA)Diptera,TRICA)Aparaglossata,A PIME)Holometabola,FOLCA)Hexapoda,((PENVA,PORTR)Decapoda,DAP MA)Altocrustacea)Pancrustacea,STRMA)Mandibulata)Arthropoda,RAMV A)Tactopoda,CAEEL))Ecdysozoa)Protostomia,(BRAFL,(CIOIN,(CALMI,(((HIPCO,(MAYZE,DANRE)Ovalentaria)Percomorpha,LEPOC)Neopterygii,(LATC H,(((MUSMU,HOMSA)Euarchontoglires,CANLUFA)Boreoeutheria,(((CORCOCO,GALGA)Aves,(CHEMY,PELSI)Testudines),PODMU)Diapsid a),XENLA)Tetrapoda)Sarcopterygii)Osteichthyes)Gnathostomata)Olfactor es)Chordata)Centroneuralia,(((ANNJA,((STRPU,APOJA)Asterozoa,(ACA PL,ASTRU)Echinozoa)Echinodermata),(PTYFL,SACKO)Hemichordata)A mbulacraria,HOFMI)Xenambulacraria)Bilateria))))))