universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Comparative structural genomics of *Octopus*

verfasst von / submitted by

Dalila Destanović

angestrebter akademischer Grad / in partial fulfillment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 220
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Joint-Masterstudium Evolutionary Systems Biology
Betreut von / Supervisor:	Assoz. Prof. Dr. Oleg Simakov

Acknowledgements

Working on this thesis has been a great learning opportunity and resulted in my growth as a scientist and a person. Throughout this journey, I received a lot of support from my colleagues and friends and this would not have been possible without them.

I would like to express my utmost gratitude to Darrin Schultz, as my advisor, who guided me through this process with his intellect, patience and empathy. His mentorship went beyond this thesis and prepared me for challenges to come.

I am also grateful to Oleg Simakov who accepted me in his lab, advised me, and gave me the opportunity to work on the topic I am interested in while being a part of an incredibly knowledgeable and friendly community.

On that note, I want to extend my heartfelt thanks to the Simakov group for their continuous support and genuine friendliness. I could not have joined a better group.

Lastly, I want to thank my family (mother and brother) and friends for being an endless well of emotional support throughout this journey, Special thanks to my roommate Natalija Milokovic for her day-to-day support over the past two years which have not been easy on either one of us.

Published works statement

A part of the research presented here was published through the pre-print server for biological sciences BioRxiv, operated by Cold Spring Harbor Laboratory: https://doi.org/10.1101/2023.05.16.540928. This genome report is currently under review in G3: Genes, Genomes, Genetics.

Funding statement

I would like to thank to Haus des Meeres Verein: Wissenschaft und Forschung: Rupert Riedl Preis for funding this work. I was also supported by ERC-H2020-EURIP grant to Oleg Simakov, grant number 945026.

Abstract

Coleoid cephalopods (octopuses, squids, cuttlefish) stand out for many morphological innovations, yet genomes have only recently become available for species in this clade. Studies of cephalopod genomes have hinted that whole-genome rearrangements could have played a crucial role in their evolution. Rearrangements were shown to impact gene regulation, by bringing together new genes into organized structures called topologically associating domains (TADs). Changes within TADs are expected to be deleterious, and are considered to be conserved throughout evolution. However, changes in TAD structure may also be a powerful mechanism of novelty evolution during speciation. Understanding if and how the genome-scale rearrangements relate to the emergence of unique cephalopod biology remains an open question. This thesis takes advantage of newly available octopus chromosome-scale genomes to investigate whether TAD borders are colocalized with rearrangement breakpoints. To do so, we assembled a high-quality-chromosome-scale *O. vulgaris* genome, created a tool that tests the hypothesis using a scale-free approach in TAD detection (Breakpointer2), and used it to compare *O. vulgaris* genome to three other octopus genomes. We observed 543 unique or shared breakpoints over 44 million years of divergence time. For each comparison, a one-tailed permutation test was performed on each chromosome and whole-genome data. By performing this test on whole-genome data, we found strong signal that chromosomal breakpoints happened at TAD borders (α<1e-05), while strong signal was observed for ~58% of chromosomes when the test was done on individual chromosomes (α<0.05). Therefore, we conclude that there is support to claim that chromosomal rearrangements are colocalized with TAD borders. This research sets the stage for further exploration of evolutionary history of individual coleoid cephalopod lineages.

Die Zusammenfassung

Koleoid-Kopffüßer zeichnen sich durch viele morphologische Innovationen aus, doch Genome sind für Arten dieser Gruppe erst seit kurzem verfügbar. Studien an Kopffüßergenomen haben darauf hingewiesen, dass die Umordnung des gesamten Genoms eine entscheidende Rolle bei ihrer Entwicklung gespielt haben könnte. Es wurde gezeigt, dass Umlagerungen Auswirkungen auf die Genregulation haben, indem sie neue Gene in organisierten Strukturen zusammenführen, die als topologisch assoziierende Domänen (TADs) bezeichnet werden. Man geht davon aus, dass Änderungen innerhalb von TADs schädlich sind und dass sie im Laufe der Evolution als konserviert gelten. Allerdings können Veränderungen in der TAD-Struktur auch ein wirksamer Mechanismus der Neuheitsentwicklung während der Artbildung sein. Es bleibt eine offene Frage, ob und wie die Umordnungen im Genommaßstab mit der Entstehung einer einzigartigen Kopffüßerbiologie zusammenhängen. Diese Masterarbeit nutzt neu verfügbare Oktopus-Chromosomengenome, um zu untersuchen, ob TAD-Grenzen mit Umlagerungsbruchpunkten kolokalisiert sind. Zu diesem Zweck haben wir ein qualitativ hochwertiges *O. vulgaris*-Genom im Chromosomenmaßstab zusammengestellt, ein Tool erstellt, das die Hypothese mithilfe eines skalenfreien Ansatzes bei der TAD-Erkennung testet (Breakpointer2), und es verwendet, um das *O. vulgaris*-Genom mit drei zu vergleichen andere Oktopusgenome. Wir haben über einen Divergenzzeitraum von 44 Millionen Jahren 543 einzigartige oder gemeinsame Haltepunkte beobachtet. Für jeden Vergleich wurde ein einseitiger Permutationstest für jedes Chromosom und die gesamten Genomdaten durchgeführt. Durch die Durchführung dieses Tests an Gesamtgenomdaten fanden wir ein starkes Signal dafür, dass chromosomale Bruchpunkte an den TAD-Grenzen auftraten (α<1e-05), während ein starkes Signal für ~58 % der Chromosomen beobachtet wurde, als der Test an einzelnen Chromosomen durchgeführt wurde (α). <0.05). Daher kommen wir zu dem Schluss, dass es Belege für die Behauptung gibt, dass chromosomale Umlagerungen mit TAD-Grenzen kolokalisiert sind. Diese Forschung schafft den Grundstein für die weitere Erforschung der Evolutionsgeschichte einzelner Koleoid-Kopffüßer-Linien.

1. Introduction

1.1. 3D genome structure

Understanding the genome 3D structure and how it changes throughout evolutionary time is an emerging topic in evolutionary biology (Eres et al., 2019; Hofstatter et al., 2022; Preger-Ben Noon & Frankel, 2022; M. Wang et al., 2018). To fully understand the functionality and regulation of particular regions of genomes, and the genome as a whole, it is necessary to consider the genome as a dynamic and complex 3D structure. Recent studies tackled the topic of macroevolutionary changes, synteny, and genomic rearrangements (such as large-scale inversions, indels, fusions, and fissions) in the emergence of novelty, speciation, and a general understanding of the evolutionary history over different time scales (Berdan et al., 2021; Cheng et al., 2020; Huang & Rieseberg, 2020; Porubsky et al., 2020). These large changes would impact the folding of the genome, but we have yet to understand how these changes in the 3D structure impact gene regulation and the general genome functionality (Álvarez-González et al., 2022; Vara et al., 2021).

Studying this question is challenging. To fit the total DNA in the nucleus (a size of 5-10 μm, depending on the species and cell type (Webster et al., 2009)) together with accompanying elements, it needs to be compressed and folded in a complex structure constituted of different levels. For example, the DNA found in the nucleus of one human cell is approximately 2 meters long. When the total number of cells in the human body is taken (about 50 trillion cells), each individual has DNA long enough to spread out from Earth to the Sun more than 300 times (Annunziato, 2008). For the research presented in this thesis, the relevant structural levels are the ones that are formed in the interphase (Figure 1.1). This is the phase in which the cell performs its function and spends most of its life. There, the chromosomes are in a diffused form and they occupy distinct territories within the nucleus (Cremer & Cremer, 2001) (Figure 1.1.1). These distinct territories are called chromosome territories (CTs). The partitioning limits the inter-chromosomal interactions within the nucleus (Rosin et al., 2019). Within the chromosome territories, one can find the “open” chromatin (euchromatin) with predominantly active genes, and the “closed” chromatin (heterochromatin) with inactive regions (Fritz et al., 2019; Lieberman-Aiden et al., 2009) (Figure 1.1.2). The lower level of structure is the topologically associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012) (Figure 1.1.3). This structural unit contains co-expressed genes, together with their enhancers and promoters. TADs are constituted of multiple loops (Figure 1.1.4), as the lowest level of structural organization (Acemel et al., 2017). On top of the general complexity presented at different levels, the 3D genome structure is dynamic and changes through the stages of cell differentiation and cell phases. This adds another level of complexity in understanding genome structure through time.

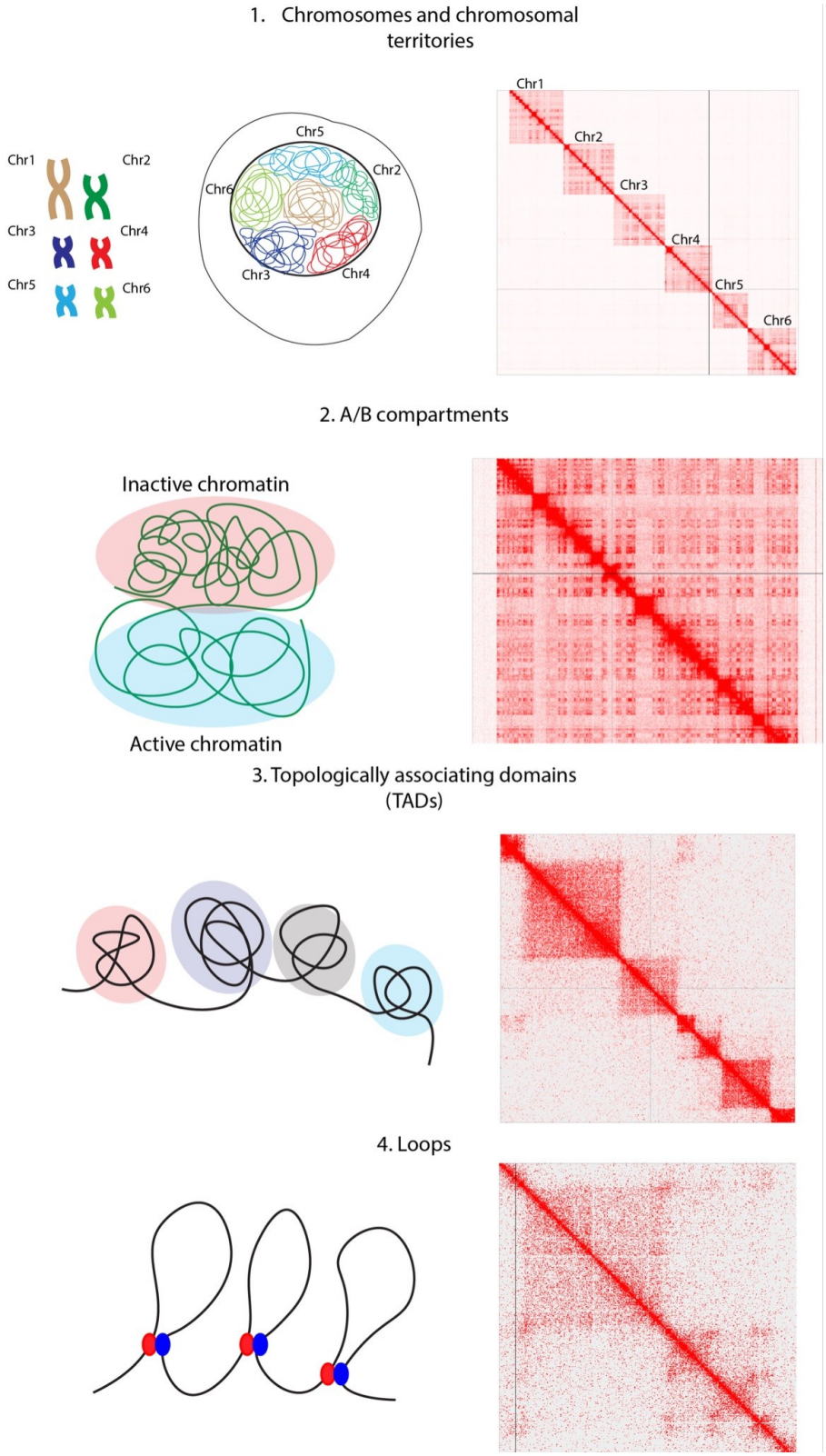


Figure 1.1 | Different levels of structural organization of the genome in the interphase: On the left side we show the sketch of the structures in the nucleus, and on the right side we show how these structures look like on the Hi-C map.

One can study genome structure through genome conformation capture techniques, out of which the most common today is Hi-C sequencing. The original Hi-C sequencing technique lets us see the contacts between distant parts of the genome. The cells are fixed with formaldehyde which cross-links adjacent DNA. Then the DNA is treated with restriction enzymes and marked with biotin. The fragments are then randomly ligated which leads to the bonding of distant genomic regions that were in physical contact (DNA in Hi-C libraries also forms connections with other haplotypes, but makes connections with the same haplotype at a much higher rate). The DNA is then purified, sheared, and captured for ligation events with biotin-capturing streptavidin beads, then amplified into a sequenceable library (Lieberman-Aiden et al., 2009).

The initial purpose of this technique was to explore the genome-wide 3D conformation of chromatin (Lieberman-Aiden et al., 2009), as seen in Hi-C heatmaps in Figure 1.1. The genome-wide long-range interaction data was later repurposed to generate chromosome-scale genome assemblies (Burton et al., 2013), as it was informative on the order and orientation of the contigs in a genome. This led to the wide usage of the Hi-C data to generate chromosome-scale assemblies (Kaplan & Dekker, 2013; F. Li et al., 2020; Q. Li et al., 2019; Nossa et al., 2014; Selvaraj et al., 2013).

This technology combined with long-read sequencing has enabled the generation of high-quality chromosome-scale assemblies (Ballare et al., 2023; Wilder et al., 2022). Currently, there are more advanced versions of Hi-C sequencing technology more suited for particular questions like DNAse Hi-C (DNase used instead of a restriction enzyme to generate equal coverage throughout the genome) (Ramani et al., 2016), micro-C (uses the micrococcal nuclease to cut at the level of nucleosomes and generate high-resolution data) (Hsieh et al., 2015), and single-cell Hi-C sequencing (developed to explore cell-to-cell variability in genome structure) (Nagano et al., 2013).

1.2. Topologically associating domains in animals

One of the discoveries about 3D genome structure facilitated by Hi-C data was topologically associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012). TADs are structural and functional units of the genome (Acemel et al., 2017). In the Hi-C heatmap, these structures are seen as square-like patterns with high interaction scores surrounded by low interaction score regions (Figure 1.1.3). A TAD has borders enriched in insulator elements (Matharu & Ahanger, 2015; Van Bortle et al., 2014) (CTCF binding sites in mammals, for example (Gómez-Marín et al., 2015)), and one can find promoters, cis-regulatory regions, and co-expressed genes within a TAD. Thus, mutations within a TAD can disable or alter the expression of an entire block of genes (Huynh & Hormozdiari, 2019; Lupiáñez et al., 2015, 2016; Tena & Santos-Pereira, 2021a). Intuitively, one can assume that these units would be under selection pressure to be conserved. It has been found that changes in TADs have a deleterious effect (Krefting et al., 2018; Melo et al., 2020). Also, they are studied more in humans, because of their medical relevance, and it has been shown that changes in some TADs are correlated with problems in development (Lupiáñez et al., 2015; Tena & Santos-Pereira, 2021b). Similarly, mutations in the nucleotide sequences might cause abnormalities in an individual's development and overall fitness, but they are still one of the drivers of

evolution (Gates, 1921; Muller, 1918; Neel, 1949). Therefore, changes in TADs could act as a powerful mechanism of novelty evolution during speciation.

An additional problem with studying TADs is that there is no commonly accepted definition (Beagan & Phillips-Cremins, 2020; Rowley & Corces, 2018; Szabo et al., 2019). Furthermore, most of the animal groups are poorly investigated, with most of the research being carried out on vertebrates. For example, it is still unclear whether non-bilaterians have TADs (Cazet et al., 2022; Kenny et al., 2020; Schultz et al., 2021; Zimmermann et al., 2020). Therefore, their evolutionary origin is still not understood, as TADs were described only in *Hydra*, a cnidarian (Cazet et al., 2022). It was not found in other species/clades of non-bilaterians. Limited data show that bilaterians (not just chordates) have TADs, such as *Drosophila* (Sexton et al., 2012), *C. elegans* on X-chromosomes (Crane et al., 2015), while TADs are smaller on autosomes (Sawh & Mango, 2022). Furthermore, TADs were described in *Euprymna scolopes*, a squid species, and it was found that they are larger compared to the ones found in vertebrates (median size of 2.5Mb compared to human 1.2Mb) (Schmidbaur et al., 2022). Also, it was inferred that cephalopods might have a similar mechanism of TAD formation, as the cohesin complex proteins were found, as well as the CCCTC-like motif at the binding sites (Schmidbaur et al., 2022).

1.3. Evolution of TADs: open questions

One open question in the field of evolutionary genomics is the stability of TADs over time. Exploration of the conservation of TADs throughout evolution was done on some groups, such as vertebrates (Krefting et al., 2018). It was concluded that TAD borders are generally conserved in this clade. On the other hand, it was shown that in mice and humans, genome rearrangements can disrupt TAD boundaries and that this may lead to the generation of novel regulatory units on multiple examples (Gilbertson et al., 2022). Furthermore, when comparing gibbons and humans it was found that genome rearrangements within TADs lead to the generation of novel TADs, even though it is a general trend that TADs are conserved (Lazar et al., 2018). On the other hand, chimpanzees and humans have only around 43% of TADs conserved (Eres et al., 2019). There is a growing need in the community to perform further interspecific genome-wide comparisons, as there is insufficient data to claim the interspecific conservation of TADs (Eres & Gilad, 2021).

The evolution of TADs has been studied primarily in vertebrates, but that is just one clade in the evolutionary history of metazoans. Among the invertebrate bilaterians, this question was investigated in flies - different species of the genus *Drosophila* (Liao et al., 2021; Torosin et al., 2020, 2022). Comparisons of TADs in *D. melanogaster* and *D. pseudoobscura* (divergence time estimation: 49 million years ago) showed 30-40% of TAD conservation. However, the rearrangement breakpoints were enriched for TAD boundaries and depleted within TADs (Liao et al., 2021). Another finding was that different types of TADs evolve under different evolutionary forces in *Drosophila*, with those containing developmental genes evolving slower than those containing broadly expressed genes (Torosin et al., 2020, 2022).

Therefore, there is evidence that the chromosomes preferentially break at the TAD borders in vertebrates (Krefting et al., 2018, Lazar et al., 2018) and drosophilids (Liao et al., 2021), but to understand the evolution of TADs, it is necessary to investigate other animal clades. High-quality chromosome-scale assemblies today will enable the exploration of TAD evolution in different groups of animals. The abundance of this resource will facilitate research on animals with different evolutionary histories and will allow the exploration of genome evolution under different scenarios. One of the groups that could be used for this type of research is cephalopods. As mentioned previously, TADs were described in a cephalopod species, *Euprymna scolopes* (Schmidbaur et al., 2022). This implies that they should be present in other cephalopod species as well, and with a growing number of chromosome-scale assemblies available for octopus species, it is possible now to explore TADs in this lineage (Albertin et al., 2022; Jiang et al., 2022; F. Li et al., 2020).

1.4. Cephalopods in the context of evolutionary genomics

Coleoid cephalopods (octopus, squid, cuttlefish) have a complex evolutionary history, but the mechanisms of genome evolution remain largely unexplored. These animals have complex nervous systems (Fiorito et al., 2014; Z. Y. Wang & Ragsdale, 2019; Young, 1964), advanced learning abilities (Borrelli et al., 2020; Brown & Piscopo, 2013), exceptional camouflage abilities (Chiao & Hanlon, 2019), as well as morphological innovations compared to other mollusks. The investigation of the genetic basis of these novel traits has been impeded by the lack of high-quality genomic resources because cephalopod genomes are large in size and have a high percentage of repetitive regions (Albertin et al., 2015; Jiang et al., 2022; Kim et al., 2018; Schmidbaur et al., 2022).

The first cephalopod whole-genome assembly was of *O. bimaculoides* published in 2015 (Albertin et al., 2015). This resource enabled the first genomic analyses of the evolution of cephalopods. As the cephalopod genomes are much bigger than those of other mollusks, the initial hypothesis was that whole genome duplications have contributed to the evolution of novelties in this clade. Analyses done in the *O. bimaculoides* genome showed that this was not a likely scenario (Albertin et al., 2015). Instead, it was proposed that the main drivers of octopus genome evolution are genome-wide rearrangements mediated by expansions of transposable elements. This idea was supported as more genomes of octopus species became available because nearly half of the genome in each species contained repetitive sequences rich in transposable element families (Albertin et al., 2015, 2022; Kim et al., 2018; F. Li et al., 2020; Zarrella et al., 2019). With advances in sequencing technology, the genomic resources for cephalopod research were growing, and other species were sequenced as well (Kim et al., 2018, Zarrella et al., 2019). Not long after that, chromosome-scale assemblies of *O. sinensis* (F. Li et al., 2020), *O. bimaculoides* (Albertin et al., 2022), and *Amphioctopus fangsiao* (Jiang et al., 2022) became available.

Whole-genome assemblies, especially chromosome-scale ones, have enabled further investigation of the evolutionary history of cephalopods. It was found that whole-genome rearrangements happened in the coleoid ancestor (Ritschard et al., 2019). Afterwards, synteny analyses of available chromosome-scale genomes have shown that the smaller octopus karyotype is a result of several secondary fusions of the more ancestral squid chromosomal complement (Albertin et al., 2022). However, we still do not understand the

individual evolutionary histories of the squid and octopus lineages. It is known that a genome-wide rearrangement event mediated by expansions of transposable elements happened in the common ancestor of coleoid cephalopods, but we do not know if these events continued to happen in the individual lineages. The activity of the transposable elements can trigger large-scale changes in chromosome structure, such as duplications, deletions, or inversions (Bhat et al., 2022; Cáceres et al., 1999; Sharma et al., 2021; J. Zhang et al., 2006, 2013; J. Zhang & Peterson, 2004, 2005).

The high percentage of transposable elements in the coleoid cephalopod genomes (~40% (Albertin et al., 2015; Jiang et al., 2022; Kim et al., 2018; F. Li et al., 2020)) indicates it is possible that their genomes continued rearranging throughout evolutionary history. Additionally, octopus karyotype is a result of several secondary fusions (Albertin et al., 2022; Jiang et al., 2022). Chromosomal fusions change nuclear architecture during meiosis, in a way that it changes the recombination landscape in primary spermatocytes (Vara et al., 2021). Therefore, fusions in the common ancestor of octopuses might have increased the rates of rearrangement events in this lineage. If this is the case, they would make a substantial model system to explore TAD evolution. In this thesis, we explore intrachromosomal rearrangements among the species pairs (*O. vulgaris-O. bimaculoides*, *O. vulgaris-O. sinensis,* and *O. vulgaris-A. fangsiao*), and test the TAD border colocalization with these breakpoints.

2. Aims

The evolution of TADs is not understood in octopus species. As the resources for this type of analysis are available now (Albertin et al., 2022; Jiang et al., 2022; F. Li et al., 2020), we sought to explore the evolution of chromosomes in the octopus lineage and determine whether synteny breaks preferentially happen at the TAD borders. To explore this question, we designed our research through three aims presented here.

2.1. Aim 1: Generate a chromosome-scale genome of *O. vulgaris*

The first aim of the comparative analysis was to generate a reference-quality chromosome-scale octopus genome assembly. We needed a high-quality and high-completeness octopus genome that could be trusted as the starting point of our comparisons, and for that, we generated a chromosome-scale genome assembly from *O. vulgaris*. Also, this genome was aligned with other existing chromosome-scale genomes of octopus species to explore orthologous relationships among the chromosomes of species.

2.2. Aim 2: Scale-free investigation of rearrangements in the context of genome topology

As we want to look into the colocalization of rearrangement breakpoints and TAD borders in the octopus lineage, we need to conceive a TAD-caller-free paradigm to infer this. These tools, called TAD callers, are used to detect TADs in the genome by using different biological and statistical criteria. However, to do so, these tools require assumptions that can make them impractical to use in non-model organisms. This is because we are not familiar with the biological features of TADs in the octopus lineage. Therefore, we had to come up with a tool that uses the data on chromosomal breakpoints and uses a TAD caller-free method to identify TADs. We did this by using insulation scores to infer the position of TAD borders. The local insulation score minimum represents a region between the TADs. Because we do not need information on the size, or the number of TADs, for our research question it is enough to differentiate between regions between the TADs and within the TADs, and we can do this by looking at the insulation scores.

2.3. Aim 3: Topological signatures of syntenic breaks in octopus genomes

In our preliminary genome-genome comparisons of octopus species, we found intrachromosomal rearrangements. We hypothesize that topologically associated domains (TADs), as structural and functional units of the genome, should be preserved across the octopus species despite numerous intrachromosomal rearrangements, as the selection pressure should preserve these units (Figure 2.1). To test this hypothesis, we will use the tool generated in Aim 2 on the available chromosome-scale genomes of octopus species. To further explore our hypothesis, we will visualize the breakpoints and TADs of chosen regions.

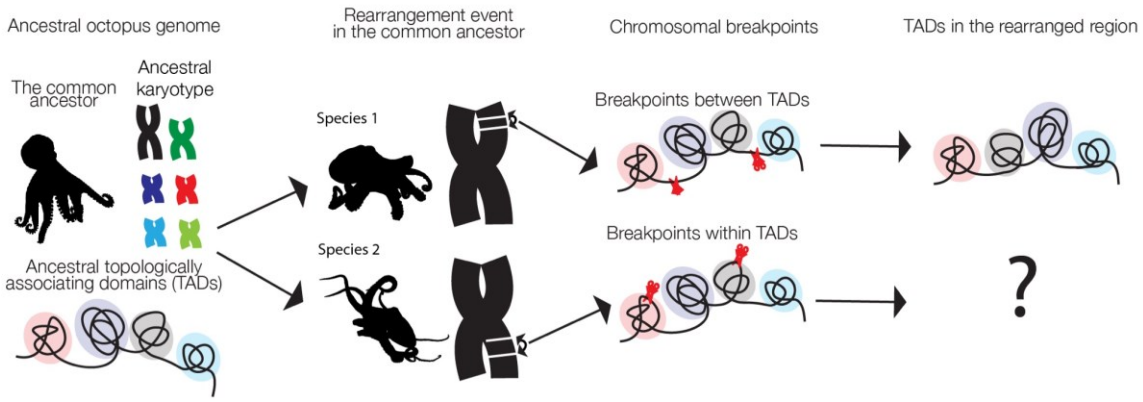


Figure 2.1 | Hypothesis overview: The common ancestor of all the octopods there was an existing ancestral karyotype. During the speciation events, different chromosomal rearrangements happen. Here we see two diverging octopus species in which two independent inversions occurred. We want to know if these synteny breaks preferentially happen between TADs. In this case, the outcome is that the order of TADs is reversed, but if a break happens within a TAD, we can only hypothesize what might have happened in the species' evolutionary history.

3. Generation of the chromosome-scale assembly of *O. vulgaris*

Content of this aim was published through the pre-print server for biological sciences BioRxiv, operated by Cold Spring Harbor Laboratory: https://doi.org/10.1101/2023.05.16.540928.

3.1. Introduction

O. vulgaris is one of the remaining key species in cephalopod research that misses a chromosome-scale genome. Although this species is an established model organism in neurobiology (Abbott et al., 1995; Hochner, 2012; Zarrella et al., 2015), there is more to understand about its diversity and biology. For example, *O. vulgaris* was considered a cosmopolitan species until recently (Amor et al., 2014, 2017, 2019; Avendaño et al., 2020; G. Gleadall, 2016; Leite et al., 2008). More detailed morphological and molecular studies have narrowed its distribution to the northeastern Atlantic and Mediterranean. These studies have also defined novel species, as well as types of *O. vulgaris* which are a part of *O. vulgaris* complex.

The initial genetic studies of *O. vulgaris* date back to the mid-twentieth century when the first karyotypes were described (Inaba, 1959; Vitturi et al., 1982). These studies suggested that *O. vulgaris* has a haploid karyotype of 28 chromosomes, while a later study showed that three octopus species have 30 chromosomes in the haploid genome (Gao & Natsukari, 1990). Since then, more comprehensive karyological studies have been done on octopus species (Adachi et al., 2014; J. Wang & Zheng, 2017). For all investigated octopus species, a haploid set of 30 chromosomes has been reported. What should be noted is that in all the mentioned studies above, except for Vitturi and associates (1982), *O. vulgaris* specimens were obtained in Japan. This is significant because as of 2017, *O. vulgaris* found in East Asia has been described as a separate species *O. sinensis* (Amor et al., 2017; G. Gleadall, 2016). Therefore, even if it states that *O. vulgaris* has been karyotyped in the paper, it is likely that these individuals were of O. sinensis species.

Therefore, we assembled the genome of *O. vulgaris sensu stricto*. This novel resource will enable a further and better understanding of the evolutionary history of the octopus lineage, as well as facilitate research in neurobiology and behavior.

3.2. Material and methods

Our role in this project was to manually curate the assembly, as well as assess the assembly quality. Therefore, the steps that led to the generation of the assembly will be briefly described. In contrast, the manual curation of the Hi-C heatmaps, and the analysis we performed will be described in greater detail.

3.2.1. Assembling of the chromosome-scale Octopus vulgaris genome

An adult specimen of octopus was captured by the fishermen in the Bay of Naples (40°48'04.1"N 14°12'32.7"E). Two types of tissues were sampled for the sequencing, the central nervous tissue was used for Omni-C sequencing (ERR11272680), while spermatophores were used to generate the ONT long-read sequences (ERR11272652, ERR11272653). Preparation of the samples was performed at Stazione Zoologica Anthon Dohrn, Naples, Italy while sequencing was done at the National Center for Genomic Analysis (CNAG), Barcelona, Spain, and SciLifeLab in Solna, Sweden. The group from CNAG also assembled the genome and scaffolded it, and then forwarded it to us for manual curation. For more details on the initial genome assembly see Destanović et al., 2023.

3.2.2. Generation of the Hi-C heatmaps and manual curation

We generated a Hi-C map to check for genome misassemblies and to curate the assembly manually. We used Omni-C paired-end reads (ERR11272680) generated from the spermatophore tissue of an adult specimen. To index the genome, and align the Omni-C reads to the genome, we used Chromap v0.2.3 (H. Zhang et al., 2021). Aligning the Omni-C reads to the genome resulted in a *.pairs* files (quality cutoffs: 2,10) converted with awk, version 4.2.1 (Aho et al., 1988) to *.longp* files, a format used by Juicebox Assembly Tools. We ran the script *run-assembly-visualizer.sh* from the 3D-DNA pipeline (Dudchenko et al., 2017) on *.longp* files and created *.hic* files. Also, we used the *generate-assembly-file-from-fasta.awk* script from the 3D-DNA pipeline (Dudchenko et al., 2017), and the *assembly-from-fasta.py* from the Artisanal pipeline (https://bitbucket.org/bredeson/artisanal) to generate *.assembly* files necessary to be able to manually curate the genome using the Juicebox Assembly Tools.

The resulting *.hic* files were visualized using the Juicebox visualization tool, version 1.11.08 (Durand et al., 2016). We corrected the order and orientation of scaffolds and placed a great portion of unplaced scaffolds within the pseudochromosomes. Corrections were done in five iterations on the adult Omni-C dataset. A new *.fasta* assembly was generated from the corrected *.assembly* file by using the *assembly-to-fasta.py* script from the Artisanal pipeline.

Also, the collaborators from the CNAG have done a contamination analysis and decontaminated the assembly.

3.2.3. Genome completeness assessment

The chromosome-scale assemblies ((Destanović et al., 2023), *O. vulgaris* pre-curation, GCA_001194135.2, GCA_006345805.1, (Jiang et al., 2022) were processed with tools from the Seqtk toolkit, version 1.3 (https://github.com/lh3/seqtk). First, we generated a list with the names of the first 30 scaffolds for each genomic *.fasta* file. Then, this was used to generate *.fasta* files only with chromosome-scale sequences by using the seqtk subseq tool. These files were then processed through the seqtk seq -l 60 option, so that each sequence line has 60 characters. The chromosome-scale assemblies of *O. bimaculoides* and *O. sinensis* were also processed through awk, version 4.2.1 (Aho et al., 1988), so that their sequences have upper case letters only, as these assemblies had both upper case and lower case characters. In this way, files of purely chromosome-level scaffolds were created for homology analysis and BUSCO. To calculate the percentage of bases in the chromosome-scale scaffolds (Tab. 1) in the pre-curated and manually curated genome assembly of *O. vulgaris*, we used bioawk, version 1.0 (https://github.com/lh3/bioawk).

The quality of the whole-genome nucleotide sequence was also assessed using BUSCO v5 (Manni et al., 2021) on the gVolante2 (Nishimura et al., 2017, 2019). The Metazoa odb10 database was used. For this analysis, only processed sequence files were used, containing only the chromosome-scale scaffolds.

3.2.4. Homology with other octopus species and nomenclature of the chromosome-scaled scaffolds

To look into the homology of the *O. vulgaris* chromosomes with closely related octopus species and base the nomenclature, we generated dot plots using the D-GENIES web application, version 1.4.0 (Cabanettes & Klopp, 2018). To do this, we have used chromosome-scale assemblies of *O. bimaculoides* (GCA_001194135.2), *O. sinensis* (GCA_006345805.1), and *A. fangsiao* (Jiang et al., 2022). To achieve this, we used Snakemake, version 7.19.1-3.11.1(Köster & Rahmann, 2012) to utilize the *GAP_dgenies_prep* script (https://github.com/conchoecia/genome_assembly_pipelines). This Snakemake script used Minimap2, version 2.24 (H. Li, 2018, 2021) in its workflow. It generated the *.paf* and index files that were uploaded to the web application (https://dgenies.toulouse.inra.fr). *O. vulgaris* assembly has been used as a reference, and *O. bimaculoides*, *O. sinensis*, and *A. fangsiao* assemblies were used as the query sequences. Scaffolds were renamed based on homology with *O. bimaculoides* chromosomes.

3.3. Results

3.3.1. Manual curation of the genome

Curation improved the quality of the final assembly, as 495 scaffolds were placed in the chromosome-scale scaffolds (Tab. 3.1), and 47 were removed in the contamination analysis. The final assembly in size of 2.80Gb has 30 chromosome-scale scaffolds (243 in total). The Hi-C heatmap of the final assembly is shown in Figure 3.1.

Table 3.1 | Assembly statistics: for contig-level assembly, scaffolded assembly and manually curated scaffolded assembly

Assembly	***Number of contigs or scaffolds***	***Scaffold sequence total***	***N50/L50***	***Number of scaffolds > 50 KB***	***% of the bases in chromosome-scaled scaffolds***
Contig-level assembly	2776	0	3.6 MB/214	0	0
Hi-C Scaffolded assembly	768	2801.6 MB	118.3 MB/9	296	95.82
Final version	226	2801.3 MB	118.9 MB/9	57	99.34
Chromosome-scale *O. bimaculoides* (Albertin et al., 2022)	145326	2342.5 MB	96.9 MB/9	85	95.46
Chromosome-scale *O. sinensis* (F. Li et al., 2020)	13516	2719.2 MB	105.9 MB/10	1800	86.09
Chromosome-scale *A. fangsiao* (Jiang et al., 2022)	6409	4341.1 MB	169.7 MB/10	1769	93.05

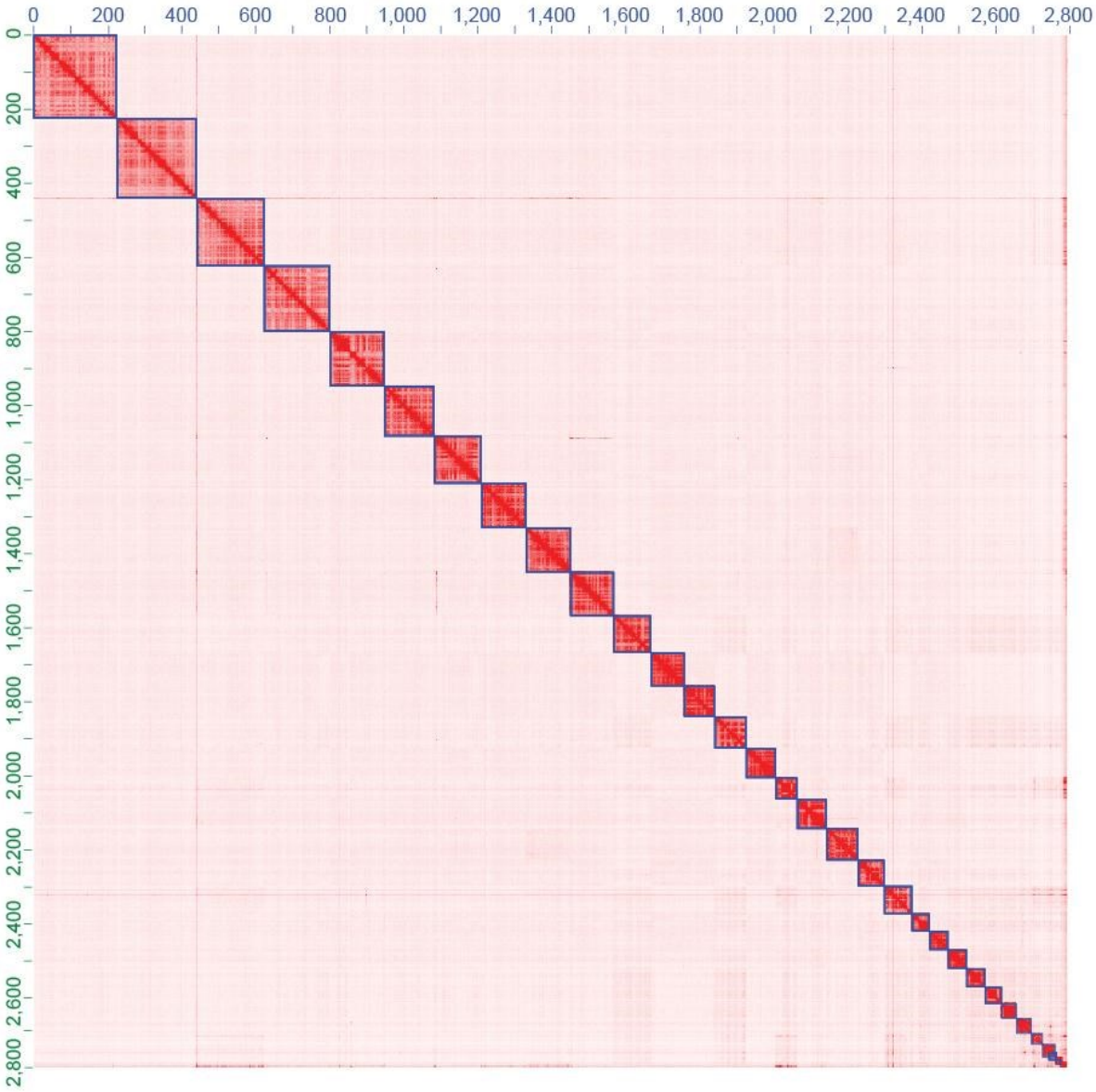


Figure 3.1 | Hi-C heatmap of the final version of *O. vulgaris*: This Hi-C heatmap shows 30 chromosome-scale scaffolds of *O. vulgaris* with just a few regions with off-diagonal signal.

3.3.2. Completeness assessment

The BUSCO score was calculated for two versions of *O. vulgaris* genome, *O. bimaculoides*, *O. sinensis*, and *A. fangsiao*. For the chromosome-scale *O. vulgaris* genome, the BUSCO score for a whole-genome nucleotide sequence using the metazoan reference dataset was 94.9% for complete plus fragmented genes (954 core genes). The full score can be seen in Tab. 2. This is an improvement considering the BUSCO score of the previous *O. vulgaris* genome assembly (GCA_003957725.1) for complete genes was 63.1% (Zarrella et al., 2019).

Table 3.2 | BUSCO scores for the available octopus genome for the metazoa_odb12 dataset: BUSCO stats of the previous and new *O. vulgaris* genome assembly version, and of other octopus species with available chromosome-scale assemblies

Genome	Complete BUSCO	Single BUSCO	Duplicated BUSCO	Fragmented BUSCO	Missing BUSCO
Chromosome-scale *O. vulgaris*	92.3% [881]	91.8% [876]	0.5% [5]	2.6% [25]	5.1% [48]
O. vulgaris (Zarrella et al., 2019)	63.1% [602]	62.6% [597]	0.5% [5]	24.8% [237]	12.1% [115]
Chromosome-scale *O. bimaculoides* (Albertin et al., 2022)	94.6% [903]	94.2% [899]	0.4% [4]	3.2% [31]	2.2% [20]
Chromosome-scale *O. sinensis* (F. Li et al., 2020)	92.0% [878]	91.5% [873]	0.5% [5]	2.9% [28]	5.1% [48]
Chromosome-scale *A. fangsiao* (Jiang et al., 2022)	93.5% [892]	91.6% [874]	1.9% [18]	3.5% [33]	3.0% [29]

3.3.3. Chromosome-scale homology among the available octopus genomes and nomenclature of O. vulgaris chromosome-scaled scaffolds

The final version of the *O. vulgaris* genome was aligned to the genomes of three octopus species, *O. sinensis*, *O. bimaculoides*, and *A. fangsiao*. These D-GENIES plots are shown in Figure 3.2. The alignments show that each chromosome in *O. vulgaris* is collinear to a chromosome in *O. sinensis*, *O. bimaculoides*, and *A. fangsiao*.

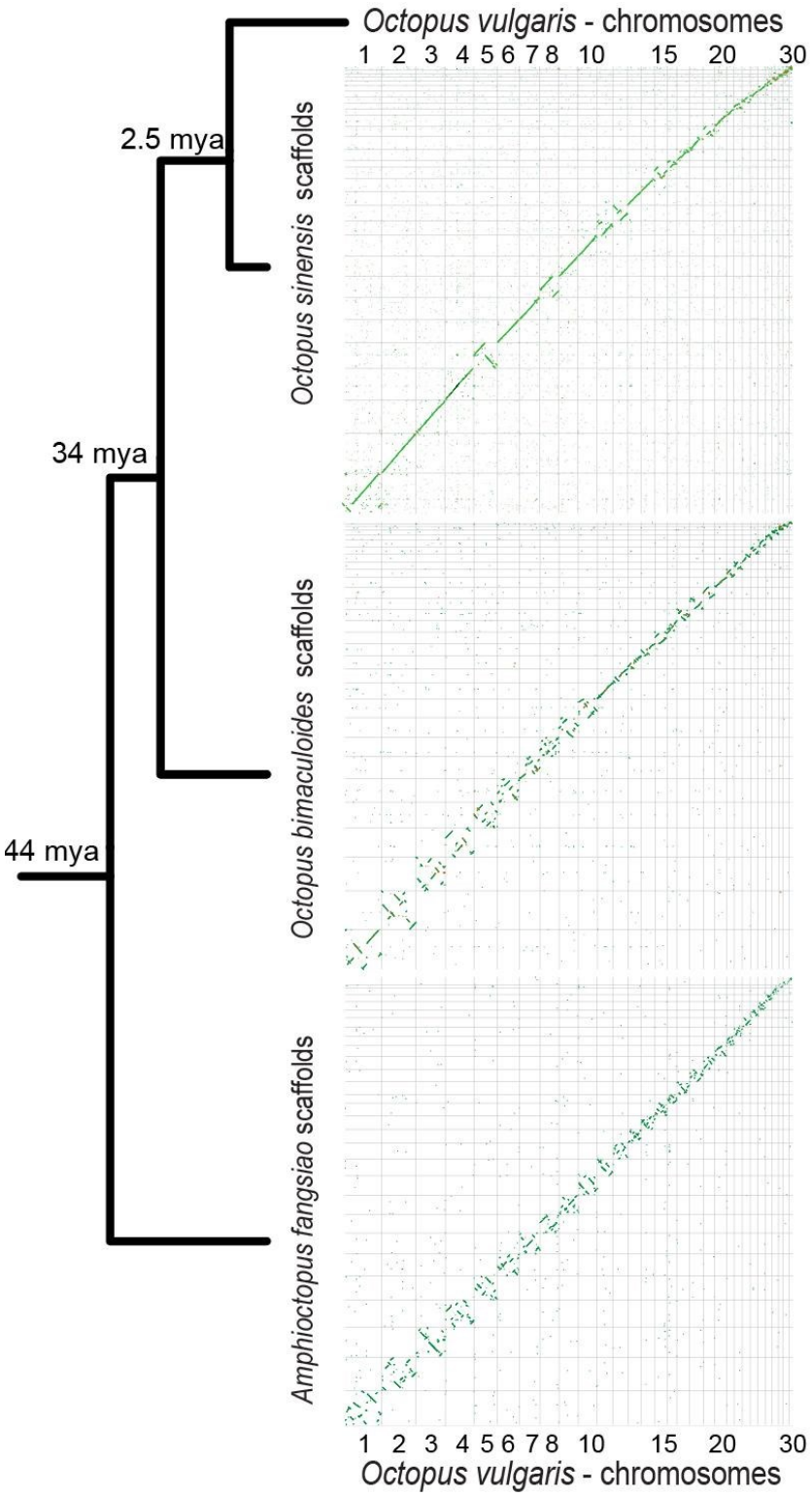


Figure 3.2 | Genome-genome alignments of *O. vulgaris* against other available octopus genomes: We see karyotype conservation (1n=30) with a gradient of intrachromosomal rearrangements occurrence as the divergence time increases.

3.4. Discussion

The haplotype of 30 chromosomes is found across octopus species (Adachi et al., 2014; Gao & Natsukari, 1990; Wang & Zheng, 2017). The signal across the Hi-C heatmap supports that *O. vulgaris* also has a haplotype of 30 chromosomes. With the expansion of molecular techniques and their application in evolutionary studies, it was found that a so-called cosmopolitan species *O. vulgaris* actually contains multiple species that form *O. vulgaris* species complex (Amor et al., 2014, 2017, 2019; Avendaño et al., 2020; G. Gleadall, 2016; Leite et al., 2008). Recently, the East Asian Common Octopus, previously considered to be *O. vulgaris*, was described as a separate species (*O. sinensis*). The spatial distribution of this species includes Japan, meaning that described *O. vulgaris* karyotypes on the specimen captured in Japan (Gao & Natsukari, 1990) probably belong to this species. However, we do not have a reason to suspect that chromosomal numbers would differ in *O. vulgaris* compared to this closely-related species. This is also confirmed by looking at Figure 3.2, as all the chromosomes present in *O. vulgaris* are collinear to a chromosome in tested species.

The first genome-genome alignment (Fig. 3.2) between *O. vulgaris* and *O. sinensis* shows that a number of chromosomes are completely collinear, while a few large-scale intrachromosomal rearrangements are observed. This is not surprising considering that *O. sinensis* belongs to the *O. vulgaris* species complex (Amor, 2016; Amor et al., 2017, 2019; Avendaño et al., 2020). A similar pattern is also observed in the comparison with *O. bimaculoides* (Fig. 3.2). However, a greater number of chromosomal rearrangements are present. This is expected considering that *O. bimaculoides* and *O. vulgaris* diverged around 34 million years ago (mya) (Jiang et al., 2022). *O. sinensis* and *O. vulgaris* have diverged over the last 2.5 mya (Amor et al., 2019), as estimated for the species of *O. vulgaris* complex. In Figure 3.2, the collinearity between the chromosomes of *O. vulgaris* and *A. fangsiao* is still maintained. As expected, as *A. fangsiao* is the most distant to *O. vulgaris* of the compared species, the genomes are even more rearranged. These species are estimated to have diverged 44 mya (Jiang et al., 2022).

Karyotype stability was described in the squid lineage, to be more precise, on loliginid and sepiolid squids (Albertin et al., 2022). This study has suggested that the smaller karyotype found in octopuses (N=30) compared to squids (N=46) is a result of secondary fusions of a more ancestral squid chromosomal complement. We infer from the genome-genome comparisons that a similar pattern of intrachromosomal rearrangements with the conservation of individual chromosomes is seen in octopus species, as described in squids (Albertin et al., 2022). However, the loliginids and sepiolids are estimated to have diverged 100 mya (Albertin et al., 2022), while the genera *Octopus* and *Amphioctopus* are estimated to have diverged 44 mya (Jiang et al., 2022). Therefore, to claim karyotype stasis in Octopodiformes, a more-distant species chromosome-scale genome is needed.

3.5. Conclusion

We generated a high-quality, high-completeness genome of *O. vulgaris* as a starting point for our comparisons. We will use this new genome and compare it to the other available chromosome-scale genome assemblies of species *O. sinensis*, *O. bimaculoides*, *A. fangsiao* (Albertin et al., 2022; Jiang et al., 2022; F. Li et al., 2020). When we explored the orthologous relationships among the chromosomes of the three species against chromosomes of *O. vulgaris*, the preliminary results showed numerous intrachromosomal rearrangements. This means that the set of chromosome-scale genomes makes a good system to study the evolutionary stability of TADs in this group.

4. Scale-free investigation of rearrangements in the context of genome topology

4.1. Introduction

As discussed in Chapter 1.3, TAD conservation was explored in particular clades, such as vertebrates (Eres et al., 2019; Gilbertson et al., 2022; Krefting et al., 2018; Lazar et al., 2018), and drosophilids (Liao et al., 2021; Torosin et al., 2020, 2022). TAD-callers were used to define TAD borders and identify TADs in these groups. A TAD-caller is a type of program that uses either feature-based, clustering-based methods or graph-partitioning algorithms to identify TADs (Liu et al., 2022; Sefer, 2022). They use different assumptions (number, size, hierarchy of TADs) and criteria (such as filtering methods, window size, etc.) to call TADs (Liu et al., 2022). It has been assessed that TA- callers can give varying results on the same data sets when using different resolutions of the Hi-C matrix (bin size), sequencing depths, and normalization methods (Liu et al., 2022; Sefer, 2022; Zufferey et al., 2018). Another problem is that some algorithms use a fixed window size which leads to a global loss of information, as the smaller window sizes lead to the identification of smaller TADs, and bigger windows lead to the identification of only bigger TADs (Liu et al., 2022).

Comparative studies of TAD-callers find that programs using biological feature-based algorithms are more accurate compared to clustering and graph-based methods (Sefer, 2022; Zufferey et al., 2018). For example, it is established that in the human genome, there is an enrichment of Alu SINE elements, insulator elements (like CTCF-binding sites (Ong & Corces, 2014)), housekeeping genes, transcription start sites, etc (Dixon et al., 2012; Shin et al., 2016). Thus, knowing the biological features of TADs in a particular species or clade enables more accurate TAD calling, or even the creation of their own tool, like for *Drosophila* (Liao et al., 2021). Therefore, most of the organisms in which TAD conservation was studied were model organisms (Eres et al., 2019; Lazar et al., 2018; Liao et al., 2021). If one works with non-model organisms, where biological characteristics of TADs have not been explored, there is a need to implement alternative approaches in TAD identification methods.

Given the limitations of TAD-callers, we opted to conceive a TAD-caller-free method to answer our research question. What we want to know is whether the breakpoints of synteny happen preferentially at the TAD borders. This means that we do not need information on the sizes or the number of observed TADs. The only necessary observation is to identify whether the found breakpoints fall between TADs, or within a TAD. We can infer the positions of TADs by using insulation scores (Crane et al., 2015). Insulation scores present an aggregation of contacts within a particular genomic interval. The regions with low insulation scores are “insulating”, and these represent the regions between the TADs (Kruse et al., 2020). Therefore, using this approach, we can infer whether the particular breakpoints in the genome happened in a high-score insulating region (within a TAD), or in a low-score insulating region (between TADs).

In this section, we present an analysis paradigm to test the correlation between TAD borders and synteny breakpoints. We created a new tool, named Breakpointer2, that uses data on genome-genome alignment to call breakpoints and test their colocalizations with TAD borders, by using the insulation scores.

4.2. Material and methods

4.2.1. Preparation of the input files

We wanted to create an algorithm that calls chromosomal breakpoints and tests their colocalization with TAD borders. To do so, we needed to define the input files first. The first step was to generate a genome-genome alignment *.paf* file, which was obtained in Aim 1, Chapter 3.2.4. Briefly, we used Minimap2, version 2.24 (H. Li, 2018, 2021) to align the genomes. The default output is in *.paf* format, a convenient format to parse, as it is organized as a *.tsv* file. The particular alignment file that was used for the generation of this tool was the *O. vulgaris* - *O. bimaculoides* alignment.

To generate the insulation scores we used the FAN-C Toolkit (Kruse et al., 2020). We used Chromap v0.2.3 (H. Zhang et al., 2021) to index the genome and to map the Hi-C reads to the genome (quality cutoff 0). Then, we used Pairix, v0.3.7 (Lee et al., 2022) to use the bgzip compression on the *.pairs* file. Also, a *.txt* file with chromosome sizes was generated using bioawk, v1.0 (https://github.com/lh3/bioawk). Then we used Cooler, version 0.9.1 (Abdennur & Mirny, 2020) to generate the *.cool* matrix with a bin size of 5000, and later a bin size of 100000. The *.cool* file was normalized using the HiCNormalization tool, a part of the HiCExplorer toolkit, version 3.7.2 (Ramírez et al., 2018; Wolff et al., 2018, 2020). Then, the normalized file was balanced using cooler balance (Abdennur & Mirny, 2020).

The normalized and balanced *.cool* file was converted into the *fanc* format as recommended by the FAN-C Toolkit authors (Kruse et al., 2020) using the `fanc from-cooler` command. This tool can set up multiple window sizes at once which might facilitate this process. Single window sizes might be prone to local matrix differences. Also, we needed to optimize for the values of window normalization and the offset value. The tool generates by default a FAN-C InsulationScores object, but the output can be specified in other formats as well, such as .bed or .gff. We used the *fanc* file to calculate insulation scores using `fanc insulation` command using the option of geometric mean (-g) and multiple window sizes of 100000 250000 500000 750000 1000000. We used the files in *.bed* format as the input in the tool.

With an alignment *.paf* file, species 1 insulation score *.bed* file, species 2 insulation score *.bed* file, we could now construct and optimize a tool that will parse these files in order to identify breakpoints, and their insulation scores.

4.2.2. Overview - Program to analyze insulation scores and synteny breakpoints

Because there are no publicly available tools that parse files containing the insulation scores in relation to chromosomal breakpoints in the species, we wrote a program that would do this. We named the program Breakpointer2 and implemented it in Python. We used the following packages in order to manage to do the task: Pandas v1.5.2. (Team, 2023), Matplotlib v3.6.3.(Hunter, 2007), numpy v1.24.1. (Harris et al., 2020), os, and statistics. The code was written using the Spyder environment (Raybaut, 2009), and tested on the Life Science Compute Cluster (LiSC) of the University of Vienna.

The workflow of the tool can be grouped into five tasks (Figure 4.1):

1. Finding the breakpoints and plotting the alignments with denoted breakpoints,
2. Finding the hypothetical location of the breakpoint (midpoint) and retrieving the insulation score for each,
3. Generation of the *.bed* file that can be used to visualize breakpoint windows,
4. Statistical test: comparing the median of breakpoint insulation scores to medians of randomly picked insulation scores (this is done for each chromosome in species, and for the entire genome),
5. Generation of the histogram to visualize the results for each chromosome and the entire genome.

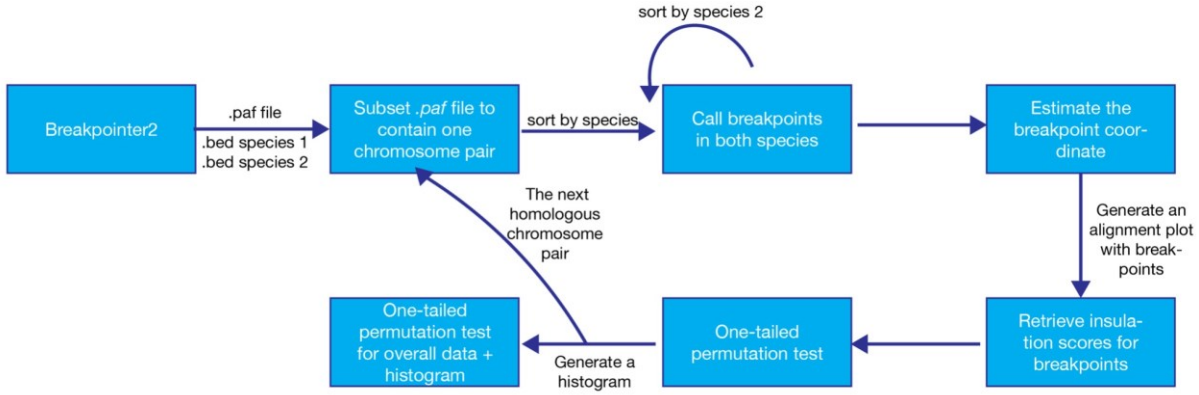


Figure 4.1 | Workflow presenting the organization of the *Breakpointer2* tool: Breakpointer2 takes the input files (*.paf* file, *.bed* files for both species), and subsets them for data based on the homologous chromosome pair. First, it calls breakpoints on the subset, then it estimates the midpoint and retrieves the insulation score for it, then performs a one-tailed permutation test and generates the histogram, and also plots the chromosome-chromosome alignment with breakpoints. When this is done for all the tested chromosomes in the comparison, the one-tailed permutation test and histogram plotting are done for the entire genome for each species.

The program input is three files: a genome-genome alignment *.paf* file comparing species 1 (`sp1`) and species 2 (`sp2`), a *.bed* file with insulation scores for `sp1`, and a *.bed* file with insulation scores for `sp2`. The *.paf* input file is filtered for alignments with a quality (`q`) lower than 30 and for the length of the alignment (`len_co`) 100000 while the *.bed* files are filtered for *nan* values. These values were set as default after testing, see Chapter 4.3.1 (Table 4.1). Also, irrelevant columns from the input are excluded.

Table 4.1 | Input variables in the Breakpointer2: Breakpointer2 requires some variables to be altered in the code itself. Also, some variables have a default value that was optimized on the *O. vulgaris* - *O. bimaculoides* comparison, but if the compared species are evolutionarily more distant, such as *A. fangsiao*, the strictness of these parameters might have to be modified.

Input variables	Meaning	Default
`--sp1`	Species 1 in *.paf* file	No default
`--sp2`	Species 2 in *.paf* file	No default
`--len_co`	The minimum length of the alignment in the *.paf* file (in 1 row)	100000 bp
`--q`	The minimum quality score of the alignment in the *.paf* file (in 1 row)	30
`--len_bp`	The minimum distance between a stop coordinate of an alignment (in 1 row) and the start coordinate (in the next row) to call it a breakpoint	5000000 bp
`--num_rounds`	Number of permutated samples for the one-tailed permutation test.	100000

The outputs of the program are an altered *.paf* file with data on breakpoints (Figure 4.2 A), a chromosome-chromosome alignment plot with breakpoints (Figure 4.3 D), *.bed* files for each species' breakpoint windows, an overview of statistical results for each chromosome (Figure 4.2 B) and the whole genome (Figure 4.2 C), a histogram used to compare the breakpoint insulation score median against medians of permuted samples for each chromosome (Figure 4.3 A-B), and the whole genome (FIgure 4.3 C). The altered *.paf* file is different from the original input, as it was filtered by the quality and length of the alignments. Also, it contains information on breakpoints, as well as insulation scores of the breakpoints. The chromosome-chromosome alignment plot is generated for each pair of homologous chromosome pairs in the initial dataset. The overview of statistical data includes information on the dataset used, chromosome name, number of breakpoint insulation scores, false detection rate, and the total number of breakpoints in the chromosome.

A

index	Osin_chr	Osin_chr_size	Osin_start	Osin_stop	ative_orientati	Ovu_chr	Ovu_chr_size	Ovu_start	Ovu_stop	nmatch	alen	qual_score	breakpoints	ikpoint_in_spe	midpoint_Osin	sul_score_Osi	midpoint_Ovu	isul_score_Ov
2815	NC_043807.1	77502973	23384825	23725680	-	Ovu11	99980488	84969852	85315130	185587	387564	60	nan	nan	nan	nan	nan	nan
2816	NC_043807.1	77502973	23725716	23999999	-	Ovu11	99980488	84683513	84944082	168787	287138	60	nan	nan	nan	nan	nan	nan
2817	NC_043807.1	77502973	24000006	24689339	-	Ovu11	99980488	84072337	84683506	360169	672614	60	nan	nan	nan	nan	nan	nan
2818	NC_043807.1	77502973	24632791	24999999	-	Ovu11	99980488	83711509	84071930	226099	397138	60	nan	nan	nan	nan	nan	nan
2819	NC_043807.1	77502973	25000014	25726970	-	Ovu11	99980488	82996761	83711466	412962	790456	60	nan	nan	nan	nan	nan	nan
2820	NC_043807.1	77502973	25727218	25999996	-	Ovu11	99980488	79419275	79690569	151593	304373	60	nan	nan	nan	nan	nan	nan
2821	NC_043807.1	77502973	26000003	26234763	-	Ovu11	99980488	79184983	79419268	159676	246378	60	True	Ovu	2.62697e+07	-2.8275	nan	nan
2822	NC_043807.1	77502973	26304562	26486933	-	Ovu11	99980488	80173308	80279488	59990	115810	60	nan	nan	nan	nan	nan	nan
2823	NC_043807.1	77502973	26407332	26510465	+	Ovu11	99980488	81376371	81472386	46349	113491	60	nan	nan	nan	nan	nan	nan
2824	NC_043807.1	77502973	26592261	26739753	+	Ovu11	99980488	82012307	82148332	67171	167265	60	True	Osin	2.68095e+07	-1.21977	8.25725e+07	-0.249992
2825	NC_043807.1	77502973	26879216	26999999	-	Ovu11	99980488	93743727	93850581	65008	129433	60	True	Osin	nan	nan	9.41634e+07	-0.347285
2826	NC_043807.1	77502973	27010438	27272997	-	Ovu11	99980488	93114281	93380432	158304	288278	60	nan	nan	nan	nan	nan	nan
2827	NC_043807.1	77502973	27331600	27443095	+	Ovu11	99980488	92891336	93001955	69510	120861	60	nan	nan	nan	nan	nan	nan
2828	NC_043807.1	77502973	27443278	27576936	-	Ovu11	99980488	92565790	92708358	76787	143912	60	nan	nan	nan	nan	nan	nan
2829	NC_043807.1	77502973	27577041	27948574	-	Ovu11	99980488	92158814	92518330	219736	390770	60	nan	nan	nan	nan	nan	nan
2830	NC_043807.1	77502973	28000001	28143569	-	Ovu11	99980488	91657958	91811043	79360	168558	60	True	Osin	2.81804e+07	-3.2695	nan	nan
2831	NC_043807.1	77502973	28233276	28388158	-	Ovu11	99980488	99804530	99956362	124115	157140	60	True	Osin	nan	nan	nan	nan
2832	NC_043807.1	77502973	28388483	28690485	-	Ovu11	99980488	99434603	99751161	268923	323614	60	nan	nan	nan	nan	nan	nan
2833	NC_043807.1	77502973	28698613	28999997	-	Ovu11	99980488	99125646	99411759	229892	321934	60	nan	nan	nan	nan	nan	nan
2834	NC_043807.1	77502973	29000006	29994895	-	Ovu11	99980488	98158976	99125637	742082	1048649	60	nan	nan	nan	nan	nan	nan
2835	NC_043807.1	77502973	30000004	30507855	-	Ovu11	99980488	97352728	97870773	341183	555984	60	nan	nan	nan	nan	nan	nan
2836	NC_043807.1	77502973	30507991	30659342	-	Ovu11	99980488	96996588	97147885	100149	162844	60	nan	nan	nan	nan	nan	nan
2837	NC_043807.1	77502973	30659477	30836038	-	Ovu11	99980488	96799802	96970614	110814	183086	60	nan	nan	nan	nan	nan	nan
2838	NC_043807.1	77502973	30836150	30999953	-	Ovu11	99980488	96311853	96463966	98641	171218	60	nan	nan	nan	nan	nan	nan
2839	NC_043807.1	77502973	31000029	31999994	-	Ovu11	99980488	95346327	96311777	645876	1055252	60	nan	nan	nan	nan	nan	nan
2840	NC_043807.1	77502973	32000001	32267202	-	Ovu11	99980488	95083531	95346320	185356	276501	60	nan	nan	nan	nan	nan	nan
2841	NC_043807.1	77502973	32327025	32808031	-	Ovu11	99980488	94476208	95035091	337692	601180	60	True	Osin	3.28807e+07	-5.39395	nan	nan
2842	NC_043807.1	77502973	32881321	32999727	-	Ovu11	99980488	76586100	76724960	68360	150302	60	True	Osin	nan	nan	7.67666e+07	0.115307
2843	NC_043807.1	77502973	33004330	33141858	-	Ovu11	99980488	76458428	76585981	79793	147158	60	nan	nan	nan	nan	nan	nan

B

Index	Dataset	Chromosome	Insulation_breakpoint_median	Num_of_Breakpoint_insul_scores	False_detection_rate	Number_of_smaller_permutation_medians	Number_of_bigger_permutation_medians	Estimated_number of breakpoints
0	sorted_by_Ovu	NC_042997.1	-1.1459	8	0.00025	25	99975	14
0	sorted_by_Osin	NC_042997.1	-1.1459	8	0.00017	17	99983	14
0	sorted_by_Ovu	NC_042998.1	-2.98807	4	<1e-05	0	100000	10
0	sorted_by_Osin	NC_042998.1	-2.98807	4	<1e-05	0	100000	10
0	sorted_by_Ovu	NC_042999.1	-2.33107	6	<1e-05	0	100000	12
0	sorted_by_Osin	NC_042999.1	-2.33107	6	<1e-05	0	100000	12
0	sorted_by_Osin	NC_043000.1	-2.51916	6	<1e-05	0	100000	8
0	sorted_by_Ovu	NC_043000.1	-2.51916	6	<1e-05	0	100000	8
0	sorted_by_Ovu	NC_043001.1	-1.49182	6	0.00014	14	99986	14
0	sorted_by_Osin	NC_043001.1	-1.49182	6	0.0002	20	99980	14

C

Index	Species	Insulation_breakpoint_median	Num_of_Breakpoint_insul_scores	False_detection_rate	Number_of_smaller_permutation_medians	Number_of_bigger_permutation_medians
0	Ovu	-0.469121	180	<1e-05	0	100000
0	Osin	-2.08599	156	<1e-05	0	100000
0	Ovu	-0.469121	180	<1e-05	0	100000
0	Osin	-2.08599	156	<1e-05	0	100000

Figure 4.2 | Raw data outputs in Breakpointer2: (A) The modified *.paf* file includes relevant alignment information from the original *.paf* file with additional information on the estimated breakpoint location and insulation scores for both species, (B) The results of the one-tailed permutation test for each chromosome of a species, (C) The results of the one-tailed permutation test for the whole genome of each species. These outputs are shown in Spyder IDE (Raybaut, 2009).

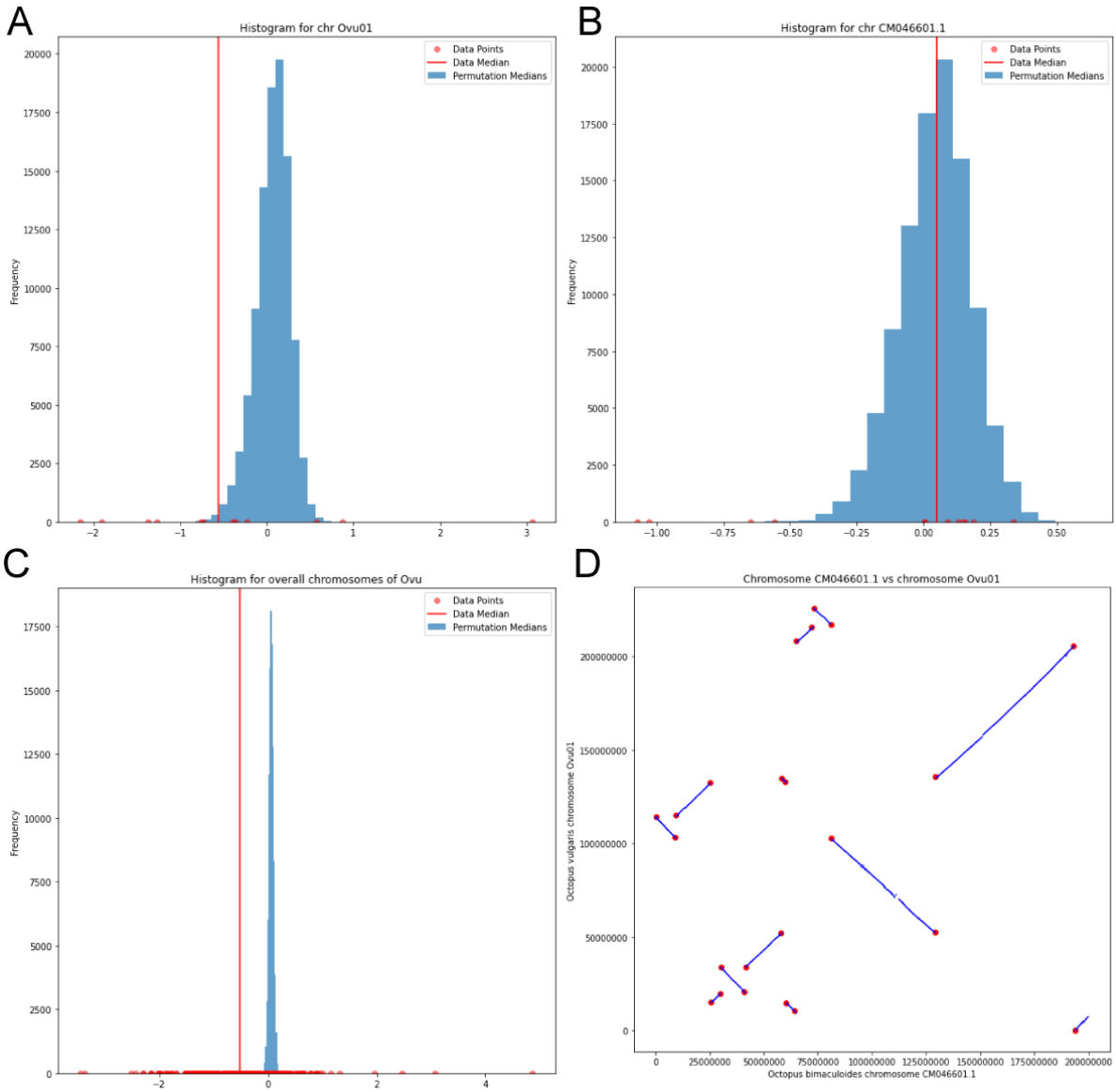


Figure 4.3 | Data visualization in Breakpointer2: In A-C we see histograms presenting the distribution of the medians of randomly sampled insulation scores against the median value of breakpoint insulation scores presented as a vertical line. The dots represent the actual values of insulation scores of estimated breakpoints. (A) Example of output for chromosome 1 for *O. vulgaris*, (B) Example of output for chromosome 1 for *O. bimaculoides*, (C) Example of output for whole-genome data for *O. vulgaris*, (D) Example of the chromosome-chromosome alignment plot with denoted breakpoints as red dots. In this particular example, chromosome 1 of *O. bimaculoides* (`sp1`, x-axis) is aligned against chromosome 1 of *O. vulgaris* (`sp2`, y-axis).

4.2.3. Detailed program description to analyze synteny breakpoints and insulation scores

The tool was named Breakpointer2 and argparse module was added to it so it can be run on the command line. Mandatory argparse arguments are: `--input_paf` (the input *.paf* file), `--input_ins_score_sp1` (insulation score table generated in FAN-C for species 1), `--input_ins_score_sp2` (insulation score table generated in FAN-C for species 2), `--output_fin_tab` (the name under which the modified *.paf* file is saved in the directory). Optional argparse variables are shown in Table 4.1.

Step 1 - Input file requirements: The names of the chromosome-scale scaffolds of both species are required. It is enough for the chromosome-scale scaffolds of species 1 to be inserted in the order present in the genome *.fasta* file. The names of the scaffold for species 2, however, have to be entered in the order of homology relative to species 1. Then, a dictionary is created that connects the homologous chromosomes of the two species. Because the program currently does not have a section that retrieves a list of chromosomes for each species, we used this command to retrieve the names of the scaffolds from a genome *.fasta* file quickly in the script: `grep '^>' genome.fasta | sed -E 's/^[[:space:]]*[^[:space:]]+/&,/' | sed 's/>//g' | sed 's/[^[:space:],]\+/"&"/g' | tr '\n' ' '`. This command formats the chromosome names from the *.fasta* file in a way that they can be directly put into the Python script.

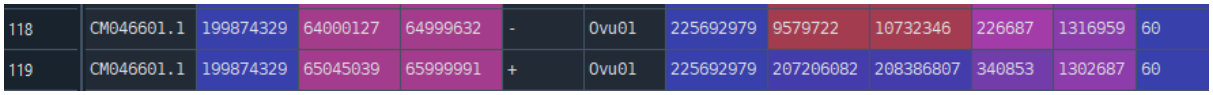
Breakpointer2 sorts the data frames (alignment file, both insulation score files) based on the alignment coordinates first from the sp1 and in the second iteration based on the coordinates of sp2. Then, the second for loop creates a subset for each pair in the dictionary (the key is `sp1` chromosome, and the value is homolog in `sp2`). The results are generated for a chromosome pair at the time and are joined at the end of the program.

Step 2 - Synteny Breakpoint detection: This data frame contains only one pair of chromosomes sorted first by coordinates in `sp1`. Then it iterates through each row of the subset and finds if there are significant jumps (of 5Mb, parameter `len_bp` in Table 4.1) in the stop coordinate in the row and the start coordinate of the next row. When this is done, the data frame is sorted by coordinates of sp2, and the process is repeated. Each breakpoint is defined by two points in both species. For example, in Figure 4.4, two rows of a data frame were sorted by coordinates for species 1 (the first 4 columns, presented by the first chromosome of *O. bimaculoides*). By looking at the coordinates of the second species (*O. vulgaris* in this case), there is a jump in the alignment from 10Mb to 207Mb (Figure 4.4), while the region is continuous in the chromosome of species 1. Then, the data frame is sorted by the coordinates in the second species, and the process is repeated. For a better understanding, the data frame sorted according to the second species is shown in Figure 4.5. This step produces a modified *.paf* file with information on breakpoints.

Step 3 - Calculation of the insulation scores of the breakpoints: Briefly, this function takes the modified *.paf* file with information on the breakpoints, the insulation score table for one of the species, and the species for which it calculates the midpoint. The midpoint is calculated as we do not know the exact place of the ancestral breakpoint, nor do we know in which species the break happened (Figure 4.3 D). So, for species 1 (Figure 4.4), we would

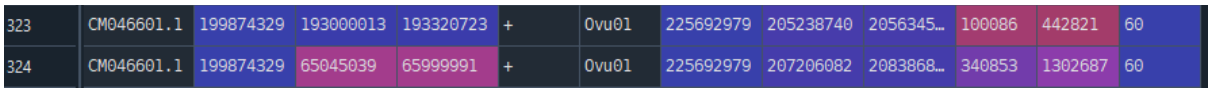
take the stop coordinate of the first row, and the start coordinate of the second row. Then, the function finds the midpoint between these two coordinates as the estimate of where the breakpoint happened. Then, it takes the midpoint coordinate and finds this coordinate in the range of the bin sizes of the insulation score data frame and adds both of this information in the data frame. Then, the data frame is sorted by the coordinates in the second species, and the midpoint is calculated for species 2 (See Figure 4.5).

The final subset for a chromosome pair now contains information on breakpoint coordinates, as well as the estimated midpoint, and insulation scores of those for both species.



118	CM046601.1	199874329	64000127	64999632	-	Ovu01	225692979	9579722	10732346	226687	1316959	60
119	CM046601.1	199874329	65045039	65999991	+	Ovu01	225692979	207206082	208386807	340853	1302687	60

Figure 4.4 | The original content of the *.paf* file is sorted by the chromosome, start, and stop coordinates of the alignments in species 1: When the alignment file is sorted by the data on species 1, we can detect breakpoints only in species 2, as seen in columns 8 and 9.



323	CM046601.1	199874329	193000013	193320723	+	Ovu01	225692979	205238740	2056345…	100086	442821	60
324	CM046601.1	199874329	65045039	65999991	+	Ovu01	225692979	207206082	2083868…	340853	1302687	60

Figure 4.5 | The original content of the *.paf* file sorted by the chromosome, start, and stop coordinates of the alignments in species 2: When the alignment file is sorted by the data on species 2, we can detect breakpoints only in species 1, as seen in columns 8 and 9.

Step 4 - Generation of a *.bed* file for breakpoint visualization: Because we want to be able to visualize the breakpoints on the Hi-C heatmap, and explore the breaks in the context of genome topology, we wrote a part of the code that will generate a *.bed* file for each species containing information on the coordinates of breakpoint windows. As explained in the previous step, we do not know where the breakpoint happened, so we have a window of estimation. Therefore, the *.bed* file for each species contains the information on the chromosome, the start coordinate of the breakpoint window, and the stop coordinate of the breakpoint window.

Step 5 - One-tailed permutation test on the insulation scores: We know that the local insulation score minimums are the insulating regions (the regions between the TADs) (Figure 4.6 A-B). So, if the synteny breaks in the genome tend to happen between TADs, we would observe lower insulation scores than what is expected by chance. Therefore, we implemented a one-tailed permutation test. To have a more robust result relative to the outlier, we chose to use the value of the median to perform this test. In Figure 4.3 C, the vertical line presenting the median of the breakpoints insulation scores is not skewed to the right in spite of the outliers. To understand the concept of the one-tailed permutation test easier refer to Figure 4.6.

The median of breakpoint insulation scores is calculated for each chromosome. Then, a random sample of insulation scores from the same chromosome is drawn. The random sample is of the same size as the sample of the real breakpoint insulation scores to reduce sample size bias. The median is calculated for the random sample. This is repeated for num_rounds (Table 4.1) times and each median is saved to a list. The random sample

medians are then plotted as a histogram, and breakpoint insulation score median is plotted together with the individual breakpoint insulation score data points (Figure 4.3 A-C). To understand how likely it is to observe the pattern seen in the histogram, we calculate false discovery rate (FDR) to interpret the results. We calculate FDR by counting the number of random sample insulation score medians that are lower than the breakpoint insulation score median and divide that by the number of trials. This statistical test is also performed on the whole-genome data.

Step 6 - Plotting of the alignment and breakpoints: This section of the code plots the best alignments in the *.paf* file with the breakpoints. A chromosome-chromosome alignment is generated for each homologous pair of chromosomes. Collinear regions are shown as blue lines, while the breakpoints are red dots (Figure 4.5 D).

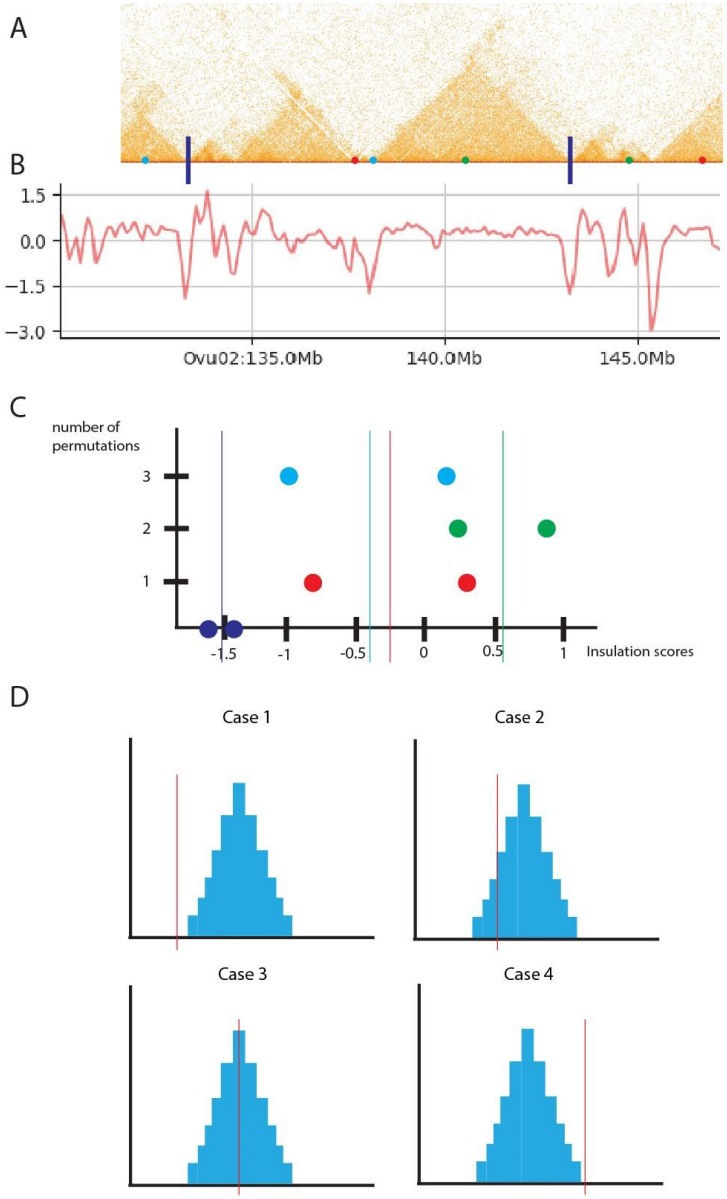


Figure 4.6 | Simplified explanation of the one-tailed permutation test: (A) A region in the genome in which some kind of chromosomal break happened. To test whether this break happened at the TAD borders, we will take the insulation scores of the breakpoint regions (dark blue lines). In the panel below (B) insulation scores for this region are observed. We take the insulation scores for these two data points and denote them as blue dots on the x-axis (Panel C). Then, the median is calculated and plotted as a vertical line. We then proceed to take random spots from the same region of the genome and take two data points at the time, calculate their median and plot them in Panel C. Each median from Panel C is then plotted as a histogram in Panel D. (D) The red vertical line is the observed median, and the bar is the distribution of randomly sampled insulation score medians for a set number of trials. Case 1: Observed median is lower than every median of randomly sampled insulation score trial. This suggests that the breakpoint median is unlikely to be observed by chance. This would be observed if chromosomal breaks happened at TAD borders preferentially. Case 2: Observed median is lower than about 70% medians in permutation trials. We can not exclude that the observed breakpoint median is a product of chance. Case 3: The observed median sits right in the middle of the distribution of medians generated through a permutation test. We can not exclude that the observed breakpoint median is a product of chance. Case 4: The observed median is greater than the medians calculated in the permutation test. This would also be unlikely to be observed by chance. This case is, however, unexpected since it would mean that genomes preferentially break in the highest interaction areas (mid-TADs).

4.3. Results

4.3.1. Breakpointer2 testing and optimization

The code was tested using the dataset of *O. vulgaris* and *O. bimaculoides*. The `sp1` in the alignment was *O. bimaculoides* and `sp2` was *O. vulgaris*. The same *.paf* file was used for the generation of the D-GENIES plot in Section 3.

The bin size we used for the generation of the matrix used for the calculation of the insulation score was 100kb. In the *.bed* file that is generated, insulation scores are reported for this range. Based on the way that our code works for this part, is that it takes a midpoint and looks for it within this 100kb window. This is broad enough to make up for the fact that we are not sure of the exact location of the breakpoint. When we attempted to retrieve information with a matrix with a bin size of 5kb, there were a lot of missing insulation scores in the insulation score files, so it did not work for our goal and the algorithm we came up with. Also, for the generation of the FAN-C insulation score files, we decided to use the files generated with a window size of 500kb as it had the least missing data (compared to windows of 100kb, 250kb, 750kb, 1Mb).

The default variables shown in Table 4.1 were also optimized on this dataset. For the breakpoint cutoff value, we optimized it to 5Mb. We chose this number because lower cutoffs (2-4Mb) would lead to false positives, which would be visible in the chromosome-chromosome plot generated in later steps. Also, higher cutoffs (6+Mb) would fail to detect some of the large-scale inversions. To estimate this, we looked at the chromosome-chromosome plots and used different cutoff values for breakpoint detection. An example of chromosome 9 in the *O. vulgaris* and *O. bimaculoides* is shown in Figure 4.7.

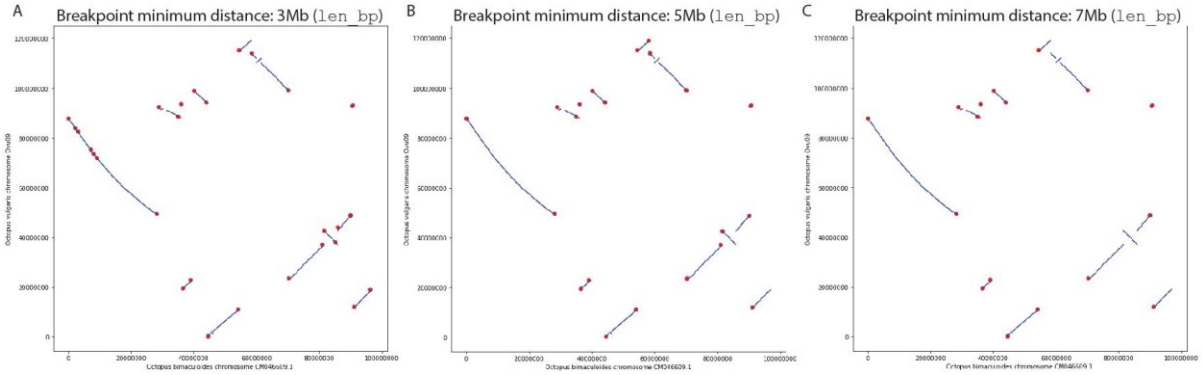


Figure 4.7 | Optimization of the `len_bp` variable– chromosome 9 example: (A) When the minimum cut off value is set to 3Mb, most of the real breakpoints are detected. However, there are 5 false positives. (B) When the cut off is set to 5 Mb, there are a few missed real breakpoints, but there are no false positives. (C) When this cutoff is set to 7Mb, the called breakpoints are accurate, but there are more missing real breakpoints than in case (B).

We set the alignment quality cutoff to 30, filtering all the alignments with quality scores less than 30. On our test data (*O. vulgaris* and *O. bimaculoides*), this worked well, as the majority of the relevant alignments had a maximum quality score, and poorly aligned fragments had low scores. Because of this contrast in the data set (either very high quality scores or very low), setting the quality cutoff to a higher value of 40, or 50 did not change the results significantly (60 is the maximum quality score). However, we thought that for comparisons of more distantly related species, it is important to not raise this cutoff value

higher. In this case, alignment quality scores that are higher than 30, but lower than 50 could be a reflection of the sequence divergence.

Another variable that we proposed the default value for is the minimum length of the alignment. We set this to 100kb when we tested the code on the *O. vulgaris* and *O. bimaculoides* comparison. This efficiently filtered the smaller alignments that would slow the program down without excluding important information on the collinearity of the two genomes. Filtering by the alignment length and quality efficiently reduced the *.paf* file from 133916 alignments to 4188 increasing the speed of processing (Figure 4.8). The step that takes the most time to be processed is the one-tailed permutation test, as it needs to be repeated for the set number of iterations. The code was tested first with 100 random sample permutations, and as that worked well timewise, we increased the number of permutations to 10000, and then to 100000.

Breakpointer2 was then also tested using the files obtained in *O. vulgaris* and *O. sinensis,* and *O. vulgaris* and *A. fangsiao* comparisons. The defaults we set in testing *O. vulgaris* and *O. bimaculoides* pair worked well for *O. vuglaris* and *O. sinensis* comparison. For the *O. vulgaris* and *A. fangsiao* comparison, the chromosome-chromosome alignments were too fragmented. So, we decided to reduce the minimum alignment length to 50000, then to 25000, and finally to 10000. With a higher `len_co` cutoff, there were too many false breakpoints, and the alignments had more gaps.

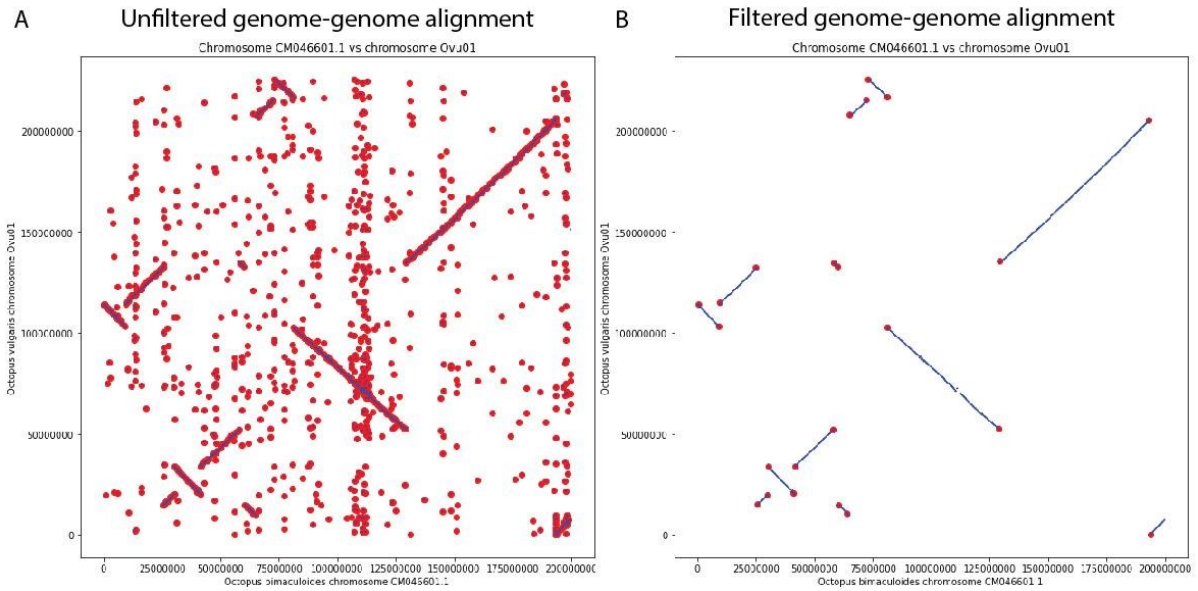


Figure 4.8 | Necessity of the filtering step – example of chromosome 1: (A) When the analysis is done on the unfiltered alignment file, there is too much noise for Breakpointer2 to be able to analyze the dataset. (B) The filtering step of genome-genome alignment with optimized parameters (`q`, `len_co`) is essential to proper functioning of the algorithm. In addition to reduction of noise, the relevant biological information is not filtered out.

4.4 Discussion

Breakpointer2 is a tool that can be used to infer the stability of TADs of closely related species, by using the genome-genome alignment data (*.paf* file) and insulation scores of both genomes. The advantage is that it can be used on non-model species, as it uses insulation scores as the input. This means that it can work without assumptions regarding the biological characteristics of the TADs (enrichment in particular motifs, types of genes, size, presence of transposable elements, etc.) in the given organism. Although we suggested a 100kb bin size for the Hi-C matrix used to generate the insulation score file, this window is flexible and depends completely on the data used in the Breakpointer2.

However, this tool is limited to analyses of intrachromosomal rearrangements, as it works by analyzing pairs of homologous chromosomes in a dictionary. Therefore, one should visualize the *.paf* file to infer this information. This is why making comparisons only among species that have a conserved karyotype is possible now. Further modification in the implementation of the code will enable the analyses of interchromosomal rearrangements as well.

Another caveat is that the results of the analysis depend on the quality of the assembly and the Hi-C data. If the chromosome-scale genome was not properly manually curated, it could have regions in false orientation (this leads to false detection of inversions), misplaced regions (false detection of translocations), or just missing regions of the genome (false detection of deletions). For this reason, we assembled a chromosome-scale genome of *O. vulgaris*, so we can have a genome of high completeness as a starting point for the comparisons.

The way the program is currently implemented does not allow us to detect interchromosomal rearrangements. Therefore, we can not distinguish regions that were translocated to another chromosome from deletions. This causes a problem with Breakpointer2's accuracy, as the breakpoint window becomes very large which can lead to inadequate estimation of the potential breakpoint spot and wrong retrieved insulation score. Therefore, further expansion of analyses Breakpointer2 can perform, can also lead to higher accuracy of the current analyses.

In the program, only one type of input file for genome alignments (*.paf)* is used, which is a file format produced exclusively by minimap2 (H. Li, 2018). One of the potential next steps could be to implement a conversion step for other widely used file formats, such as *.bam*, or *.sam*. However, minimap2 is relatively quick and accurate compared to other popular aligners such as BWA-mem (H. Li & Durbin, 2009), so *.paf* would stay the recommended format to use.

Another improvement to this tool would be to implement homology detection, and the ability to detect interchromosomal rearrangements. Then it could be applied to compare available chromosome-scale genomes of octopus to those of squid species or more distant species, like *Nautilus*.

4.5. Conclusion

In this section, we created a tool that can be used to test the hypothesis that TADs are preserved relative to chromosomal breakpoints. This tool enables us to explore the colocalization of TAD borders and chromosomal breakpoints accurately, by using just the genome-genome alignments and insulation scores as starting data. This enables us to go to our next aim, and use the data on chromosome-scale genomes of *O. vulgaris*, *O. sinensis*, *O. bimaculoides*, and *A. fangsiao* (Albertin et al., 2022; Destanović et al., 2023; Jiang et al., 2022; F. Li et al., 2020) to further explore the evolutionary history of the octopus lineage.

5. Topological signatures of syntenic breaks in octopus genomes

5.1. Introduction

The main hypothesis is that when chromosomal rearrangements happen, chromosomal breaks tend to occur between the TADs (at the TAD borders). We hypothesize this because in this case, the structural and functional units (TADs) remain intact, which should lead to the preservation of their regulation and function. Therefore, changes within the TADs should be selected against. The main and alternative hypotheses are shown in Figure 5.1. Scenario A (Figure 5.1) shows the reasoning behind the formed hypothesis. Also, under Scenario B (Figure 5.1) we offer possible outcomes of alternative hypotheses.

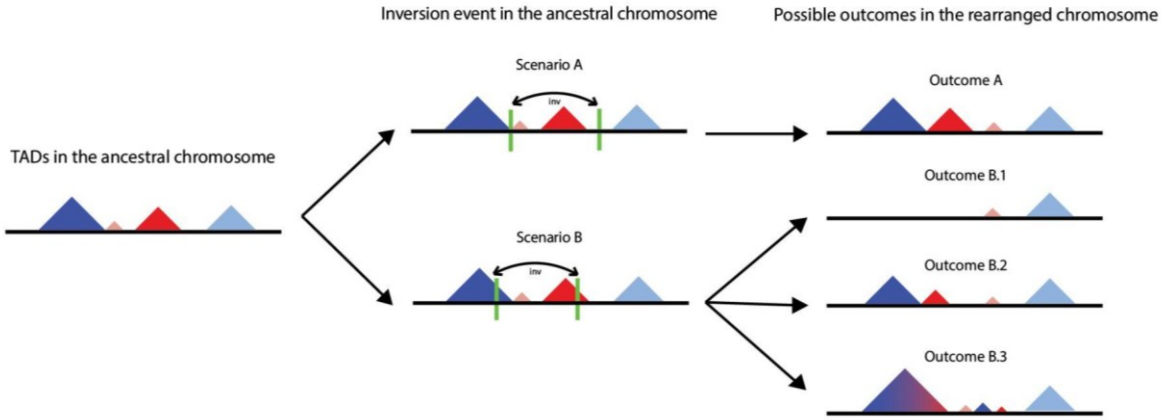


Figure 5.1: In this figure, we show topologically associating domains (TADs) in the ancestral genome. Depending on where the chromosomal breaks in the ancestral genome happened, there are two possible scenarios: Scenario A depicts breaks between the TADs which results in the conservation of TADs as units with changes in the order in the rearranged genome (Outcome A), while Scenario B depicts chromosomal breaks that happened within the TADs which can result in different outcomes in the rearranged genome. Scenario A supports our hypothesis, while Scenario B offers alternative hypotheses as to what might be the outcome of this type of chromosomal break. Outcome B1 depicts the complete disintegration of rearranged TADs, Outcome B2 depicts the partial disintegration of rearranged TADs, and Outcome B3 depicts the formation of novel TADs, either in the hybridization of two rearranged TADs (in gradient) or the formation of smaller independent TADs that were part of a bigger rearranged TAD.

We used the chromosome-scale genomes of available octopus species to test this hypothesis. In Aim 1 (Section 3), we saw that in the set of octopus species, we observed that there are intrachromosomal rearrangements. Here, we quantify these rearrangements and tested their colocalization with TAD boundaries. Additionally, we visualize some of the breaks together with the TADs to get a better understanding of the history of octopus TAD evolution.

5.2. Material and methods

Three pairs of species were tested for breakpoints: *O. vulgaris* against *O. sinensis*, *O. vulgaris* against *O. bimaculoides*, and *O. vulgaris* against *A. fangsiao*.

5.2.1. Running Breakpointer2 on species comparisons

The genome-genome alignment *.paf* files were generated in Chapter 3.2.4. Insulation scores were calculated as described in Chapter 4.2.1. We used the insulation score *.bed* files that were generated with 500kb windows. We used only chromosome-scale scaffolds for the calculation of insulation scores and the alignments to reduce noise, as the unplaced scaffolds could align with chromosome-scale scaffolds in the other species. Breakpointer2 uses data on homologous scaffolds in the species and excludes other data points. Therefore, the presence of the unplaced scaffolds in the alignment would introduce false detection of breakpoints.

The default settings (presented in Table 4.1) were used to generate the data on comparisons of *O. vulgaris* and *O. sinensis* and *O. vulgaris* and *O. bimaculoides*. The default values for the alignment length (`len_co`), and the quality cutoff (`q`) for filtering. The distance between the alignments was set to 5Mb, and the number of permutations for the statistical test was set to 100000. As *O. vulgaris* and *A. fangsiao* are more distantly related, the default value for filtering `len_co` was reduced to 10000.

5.2.2. Visualization of breakpoints and evolutionary inference

One of the outputs of the Breakpointer2 is the *.bed* file with breakpoint coordinates for each species. However, the original *.bed* file does not have a header, so when it is generated on the command line, we use the following bash commands to remove the header and the index information: `cat sp1_breakp.bed | tail -n +2 | cut -d " " -f 2,3,4 >bimac_breakp.bed` (for both species in each comparison). We chose to visualize the heatmaps and the breakpoints by using HiGlass v1.4 (Kerpedjiev et al., 2018). To ingest the relevant data to be able to generate a heatmap on HiGlass browser for each species, we used the command `docker exec higlass-container6 python higlass-server/manage.py ingest_tileset`. For this, we needed the chosen bin size for the matrix, the chromosome sizes file (generated by bioawk), the *.mcool* file generated in Aim 2. To visualize the breakpoints in HiGlass, we processed the *.bed* file with Clodius v0.20.1 (Kerpedjiev et al., 2018), and generated the *.beddb* file. We used the Hi-C matrix of *O. vulgaris* and visualized the breakpoints against all species.

To examine the evolutionary history of the chromosomes of the tested species, we defined cases that reflect the presence/absence of a breakpoint in a comparison. We defined seven cases (Table 5.1). Presence of the breakpoint is denoted with 1, absence with 0. Some regions in the genome are conserved (case 0), while some breaks might be present in just one species (case 1-3), two species (case 4-6) or three species (case 7).

Table 5.1 | Presence of breaks in species comparisons: We defined eight different scenarios based on the breakpoint presence. The presence of the same breakpoint is denoted with 1, while the absence is denoted with 0.

Case	0	1	2	3	4	5	6	7
O. sinensis	0	1	0	0	1	1	0	1
O. bimaculoides	0	0	1	0	1	0	1	1
A. fangsiao	0	0	0	1	0	1	1	1

Case 1 states the break is found only in *O. sinensis* meaning that the rearrangement happened in *O. sinensis* after the *O. vulgaris*/*O. sinensis* split. Case 2 presents the breaks that happened in the evolutionary history of *O. bimaculoides* after the split from the common ancestor of *O. vulgaris* and *O. sinensis*. Case 3 represents breaks that happened after the split of *A. fangsiao* common ancestor and the common ancestor of *Octopus* sp., but as we do not have an outgroup for this comparison, we can not know in which lineage this happened. Case 4 and 5 are unlikely scenarios. Case 4 is found when the break is detected in comparisons of *O. vulgaris* against *O. bimaculoides* and *O. sinensis*, but is not rearranged in *A. fangsiao* and *O. vulgaris*. The case 5 describes a similar scenario where the rearrangement is found in *O. sinensis* and *A. fangsiao*, but is missing in *O. bimaculoides*. Case 6 describes a scenario in which the breakpoint likely happened in the common ancestor of *O. vulgaris* species complex (meaning the break is present in *O. vulgaris* and *O. sinensis*). Case 7 presents breaks that likely happened in evolutionary history of *O. vulgaris* after splitting from the *O. vulgaris*/*O. sinensis* common ancestor.

5.3. Results

5.3.1. Colocalization of synteny breakpoints and TAD borders in O. vulgaris and O. bimaculoides

The *.paf* file contained 133916 aligned regions of the two genomes. After the quality filtering process, this was reduced to 4188 aligned regions for the 30 pairs of chromosomes. *O. bimaculoides* was denoted as `sp1`, while *O. vulgaris* was denoted as `sp2`. The *O. vulgaris* insulation file had 27716 insulation scores, while the *O. bimaculoides* had 22251 for the entire genome. The insulation score is shown in window sizes of 100kb, and the number of the insulation scores matches the difference in the genome size (with values missing for some windows).

The Breakpointer2 tool was then run on the input files. The comparison was done on all 30 chromosomes in both species. The estimated number of breakpoints for each chromosome ranged from 1 to 26, with the total estimation reaching the number of 305 breakpoints.

In *O. bimaculoides*, we found that in nine chromosomes there was significant signal for the presence of breakpoints between, not inside of, TADs (one-tailed permutation test, false discovery rate $\alpha < 0.05$). For the same set of breakpoints, we found 20 chromosomes with significant signal for breakpoints between TADs in *O. vuglaris* (one-tailed permutation test, false discovery rate $\alpha < 0.05$). What should be noted is that among the first 20 chromosomes in *O. vulgaris*, only three chromosomes do not have a strong signal for breakpoints between TADs. This could be due to smaller chromosomes having more misplaced or missing regions in chromosomes meaning that the observed breakpoints are not real.

Statistical results for one chromosome stand out in particular, and that is chromosome 10 in *O. vulgaris* (Ovu10) (Figure 5.2). The false discovery rate is 0.59939 (Figure 5.2 C), suggesting that there is no signal for preferential chromosomal breaks on TAD borders. In the alignment plot (Figure 5.2 A) the breakpoints were properly called, so that is an unlikely cause of observed difference. The statistical results retrieved for *O. bimaculoides* also suggested that there is no strong signal that synteny breaks happen at TAD borders in chromosome 10 (Figure 5.2 B).

Finally, when we test the insulation scores for the breakpoints found in the entire genomes (not individual chromosomes), we find a strong signal that breaks happen preferentially between TADs in both genomes (one-tailed permutation test, false discovery rate $\alpha < 1e^{-5}$) (Figure 5.3, Table 5.2).

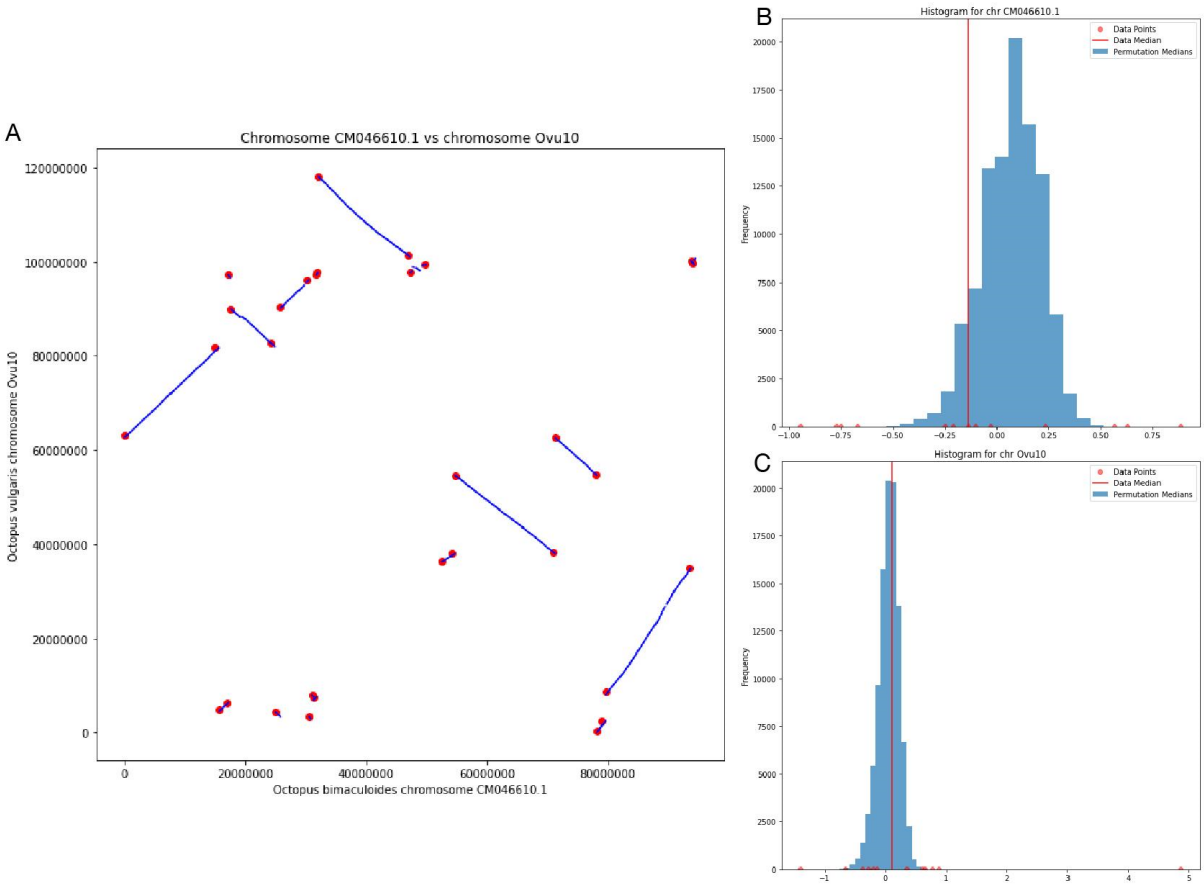


Figure 5.2 | Example of chromosome 10 (Ovu10): (A) Alignment of chromosome 10 of *O. bimaculoides* (x-axis) against chromosome 10 of *O. vulgaris* (y-axis). The red dots representing breakpoints show that the program detected the breakpoints efficiently. (B) One-tailed permutation test histogram for chromosome 10 in *O. bimaculoides*. (C) One-tailed permutation test histogram for chromosome 10 in *O. vulgaris*.

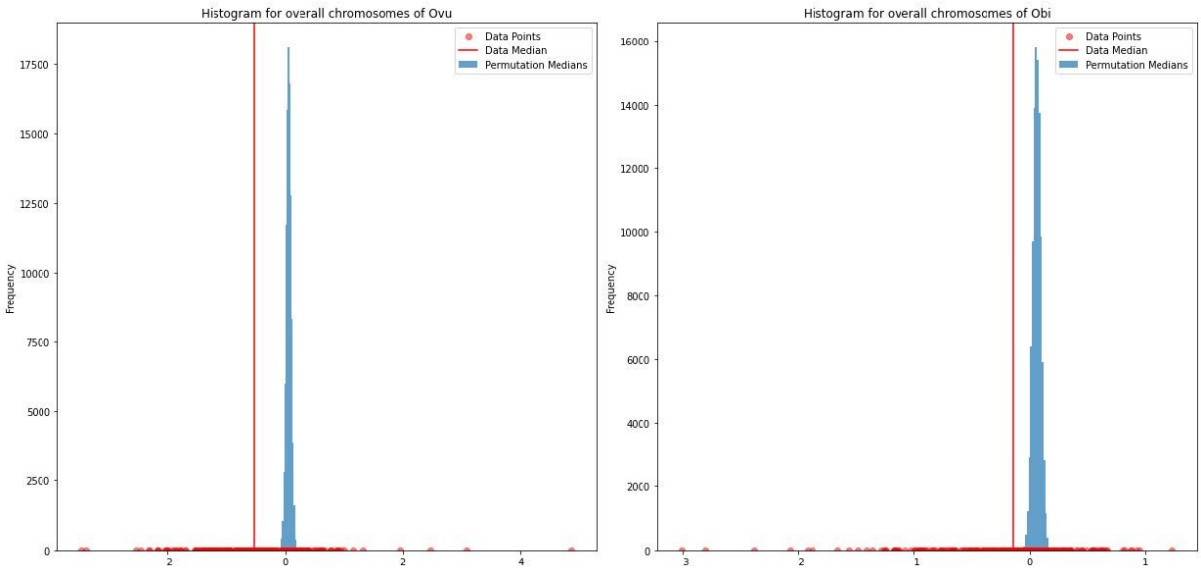


Figure 5.3 | Histograms of one-tailed permutation tests on whole-genome data of *O. vulgaris* (left) and *O. bimacucloides* (right): The breakpoint insulation score median falls left of the distribution of the medians of the permuted samples, showing that the median of breakpoint insulation scores is lower than what is expected by chance. This supports the hypothesis that breakpoints tend to happen between TADs.

5.3.2. Colocalization of synteny breakpoints and TAD borders in O. vulgaris and O. sinensis

The pre-filtering alignment file had 47336 alignments, and after filtering 4943. The *.bed* insulation score file for *O. vulgaris* was the same as for the *O. vulgaris- O. bimaculoides* comparison, while the *.bed* insulation score file for *O. sinensis* had 23304 data points for a genome of 2.7Gb, indicating missing data for around 0.4Gb of this genome.

The chromosome NC_043014.1 was excluded from the analysis, as no insulation scores were calculated for it. The length of this chromosome is around 55 Mb, and that's an additional missing portion of the information. Its homolog is chromosome 20 in *O. vulgaris* which was also excluded. In the final results of the analysis, the homologous pairs of NC_043006.1 and Ovu09, and NC_043009.1 and Ovu13 had no identified breakpoints. Therefore, 27 pairs were tested.

The number of breakpoints per chromosome varied from 1 to 13, with the total number of breakpoints being 180. For individual chromosome-chromosome comparisons, in *O. sinensis* genome, we found 23 chromosomes with a strong signal that synteny breakpoints happened between the TADs (one-tailed permutation test, false discovery rate $\alpha < 0.05$). The same was observed for 13 chromosomes in *O. vulgaris*. For *O. vulgaris*, a similar pattern of weaker signal in smaller chromosomes was observed. The statistical test on whole-genome data showed the same results as for the *O. vulgaris* and *O. bimaculoides* comparison (Table 5.1).

5.3.3. Colocalization of synteny breakpoints and TAD borders in O. vulgaris and A. fangsiao

The input *.paf* file had 1909289 alignments in total. After the filtering process, 17628 lines were kept in the *.paf* data frame.

The same insulation score table was used for *O. vulgaris* as in the previous two sections, and the *.bed* file for insulation scores for *A. fangsiao* had 40261 data points. Because smaller chromosomes were more fragmented in the assembly, the last 6 homologous pairs yielded no results. We analyzed 24 *A. fangsiao* chromosomes (LG01-LG24) and the 24 homologs in *O. vulgaris*.

We detected 493 breakpoints in total, with 11 to 34 breaks per chromosome. In *A. fangsiao*, we found significant signal suggesting that rearrangement breaks happen at TAD borders rather than within TADs in 15 chromosome (one-tailed permutation test, false discovery rate $\alpha < 0.05$), while the same was observed for 14 chromosomes in *O. vulgaris*. The results of the statistical test on whole-genome data showed the same pattern as in previous comparisons, suggesting that chromosomal breaks tend to happen at TAD borders (Table 5.1).

Table 5.2 | Results of the one-tailed permutation test for the whole genome and chromosome-level data: In this table we show insulation score medians, number of tested insulation scores, and false discovery rate as the result of the one-tailed permutation test The false discovery rate indicates that it is highly unlikely to observe the pattern we obtained by pure chance.

Comparison	Species	Number of insulation scores in the sample (whole-genome)	False discovery rate (whole-genome)	Number of chromosomes with $\alpha < 0.05$	Number of tested chromosomes
O. vulgaris vs *O. bimaculoides*	*O. bimaculoides*	250	<1e-05	9	30
O. vulgaris vs *O. bimaculoides*	*O. vulgaris*	259	<1e-05	20	30
O. vulgaris vs *O. sinensis*	*O. sinensis*	156	<1e-05	23	27
O. vulgaris vs *O. sinensis*	*O. vulgaris*	180	<1e-05	13	27
O. vulgaris vs. *A. fangsiao*	*A. fangsiao*	437	<1e-05	15	24
O. vulgaris vs. *A. fangsiao*	*O. vulgaris*	451	<1e-05	14	24

5.3.4. Visualization in HiGlass

We loaded the *.mcool* matrix of *O. vulgaris* Hi-C data in the HiGlass web browser, and viewed the *.bed* files containing the breakpoint windows for breaks against *O. sinensis*, *O. bimaculoides*, and *A. fangsiao* (Figure 5.4). Generally, the breakpoints called in bigger chromosomes looked real, as the breakpoint windows were narrow and there were occurrences of shared breakpoints among different comparisons (Figure 5.4 A). In Figure 5.4. B we show chromosome Ovu27 with breakpoint windows that span for regions larger than a megabase. These are likely not real breakpoints, but are a consequence of poorly assembled smaller chromosomes (misplaced or missing genomics regions would create such a pattern). Similar long breakpoint windows are observed in smaller chromosomes (Ovu20-Ovu30), with the exception of Ovu24 and Ovu26.

We also visualized the Hi-C matrix of *O. bimaculoides* together with the breakpoints, as we noticed that insulation score medians were higher compared to those retrieved in *O. vulgaris*, *O. sinensis*, and *A. fangsiao* genomes (Figure 5.4. C). By visualizing this genome on HiGlass, we saw that it was generally more noisy and TADs as structures were difficult to spot. The Hi-C data was visualized also for the *O. sinensis* and *A. fangsiao* genome, and it was of high resolution and had distinctive TADs, similar to what is observed in *O. vulgaris* (Figure 5.4 A).

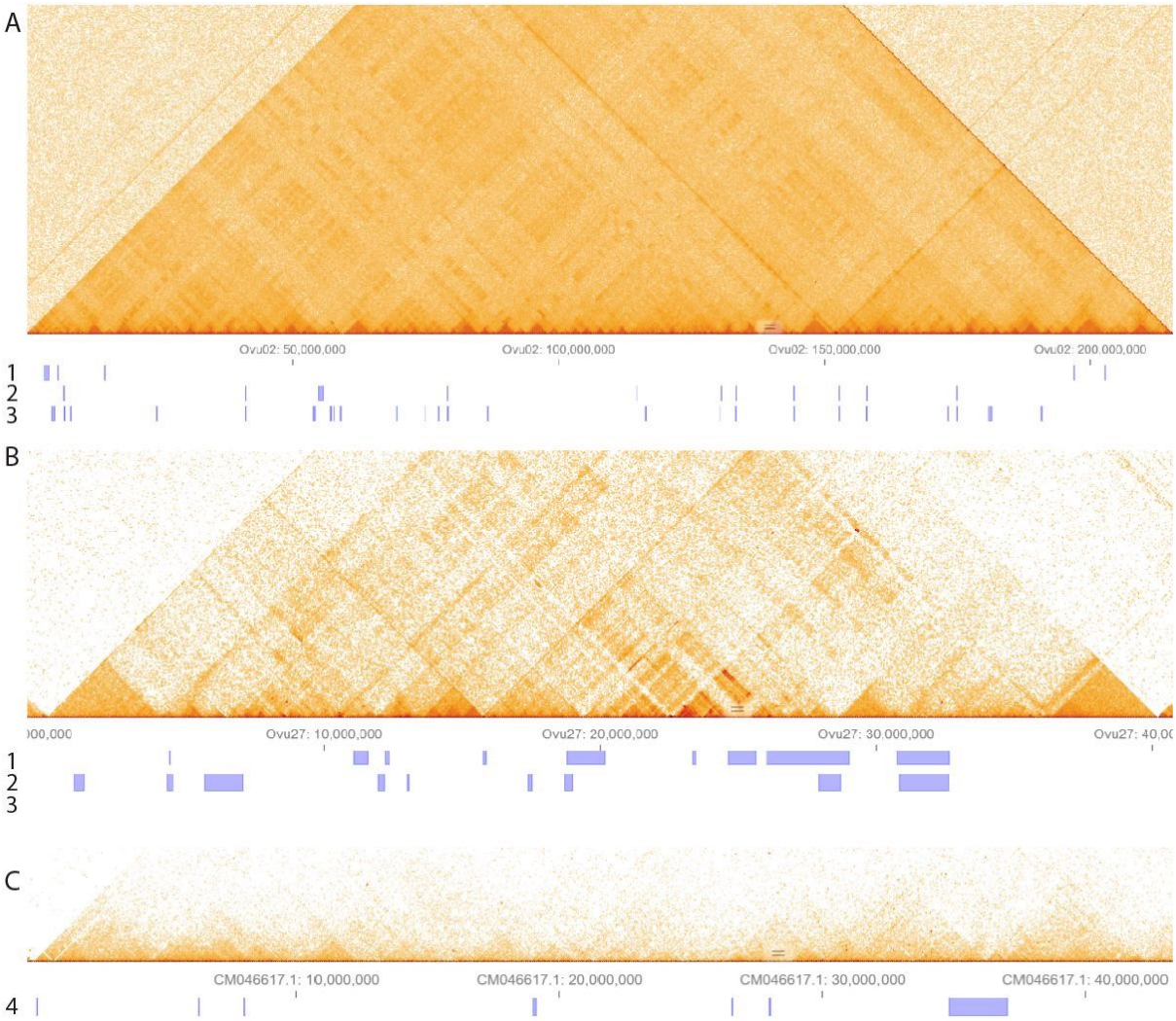


Figure 5.4 | Visualization of breakpoints in HiGlass web browser: (A) Chromosome 2 of *O. vulgaris* (Ovu02) with breakpoint tracks for each comparison (**1** - *O. vulgaris*-*O. sinensis,* **2**- *O. vulgaris* - *O. bimaculoides*, **3**- *O. vulgaris* - *A. fangsiao*) at 320k resolution. (B) An example of a chromosome (80k resolution) that was poorly assembled either in *O. vulgaris*, or *O. sinensis* and *O. bimaculoides*, or all of them. We see that breakpoint windows are wide which indicates that the region is actually missing, and that it was detected erroneously. The tracks are the same as in (A), but as the alignment between *O. vulgaris* and *A. fangsiao* was too fragmented to call breakpoints for these homologs, the track 3 is empty. (C) A Hi-C heatmap of a region of chromosome 17 in *O. bimaculoides* (80k resolution). We see some TADs, but the signal is weaker and more homogenous when compared to the Hi-C heatmap retrieved for *O. vulgaris*.

5.3.5. Exploration of shared and unique breakpoints across compared pairs

We found examples of breakpoint windows sitting right at the TAD borders, which is a visual support of our main hypothesis (Figure 5.5). We also observed that some of the breakpoints were shared in the datasets (Figure 5.5 B-C), and counted the specific cases for 24 chromosomes used in the comparison of *O. vulgaris* against *A. fangsiao* (Table 5.3).

We observe 29 shared breaks (case 7) in three comparisons, suggesting that these breaks happened in the evolutionary history of *O. vulgaris* after splitting from the common ancestor shared with *O. sinensis*. We also observed 135 breaks (case 6) that happened after the split of *O. bimaculoides* and *O. vulgaris* species complex, suggesting that these rearrangements had a role in the speciation and divergence of this species complex. We had the highest

number of observations for breaks seen only in the comparison of *O. vulgaris* against *A. fangsiao*. Because we do not have an outgroup for this comparison, we do not know whether the observed breaks happened in the evolutionary history of the common ancestor of *Octopus sp.* or in the ancestor of *Amphioctopus*. The rearrangements observed only in *O. vulgaris*-*O. bimaculoides* comparison (50) suggest that these events happened in the evolutionary history of *O. bimaculoides* after splitting from the ancestor shared with *O. vulgaris* and *O. sinensis*. The finding that there are more synteny breaks seen in the comparison of *O. vulgaris* and *O. sinensis* (73) than in the comparison against *O. bimaculoides* (50) is probably due to missing regions on the chromosome-scale scaffolds in the *O. sinensis* assembly. A low number of breakpoints matching cases 4 and 5 were detected (2, 13, respectively). This is likely due to failure of detection of the breakpoint in one of the comparisons, or a false detection of a breakpoint in one of the species.

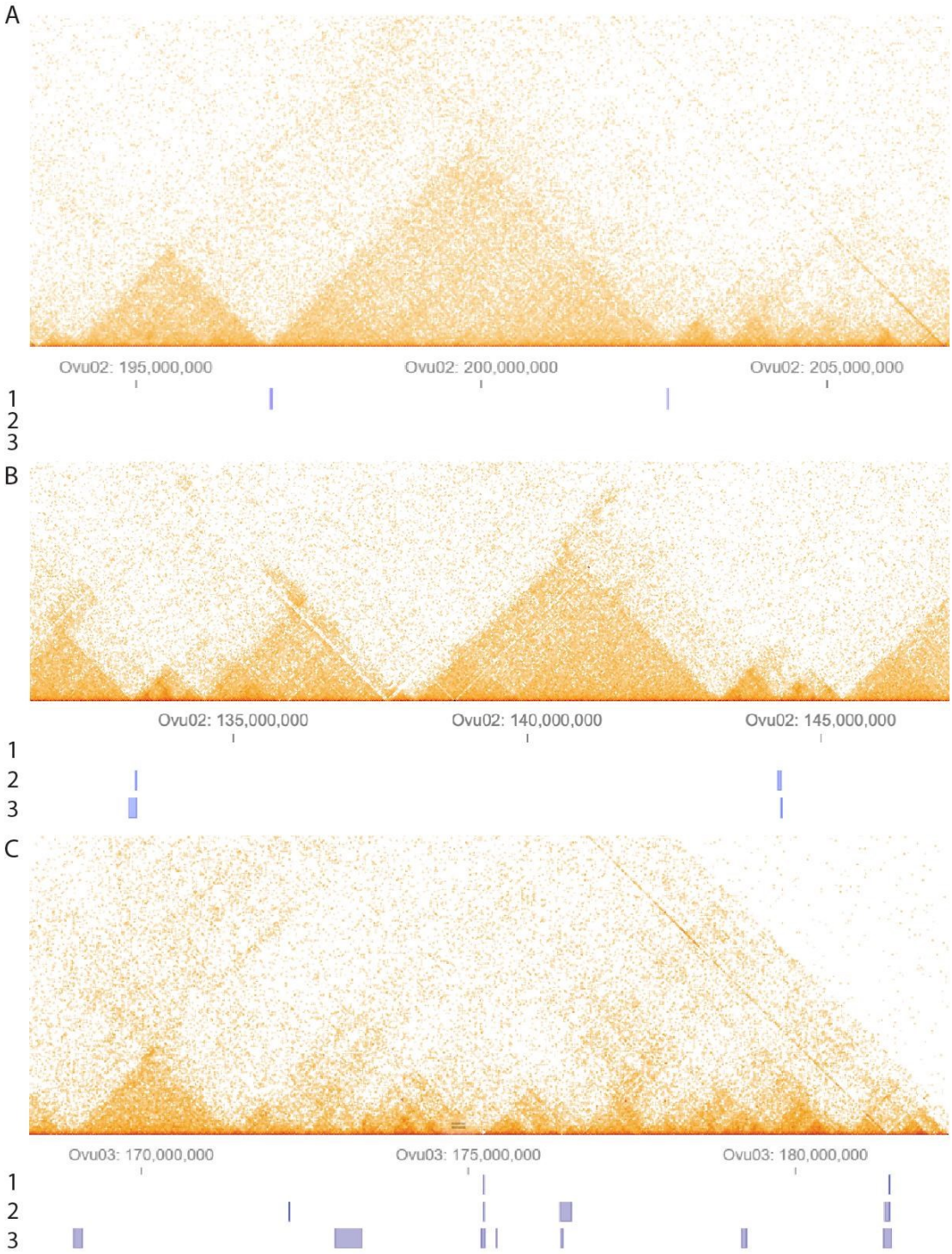


Figure 5.5: Breakpoint visualization in HiGlass: Breakpoint tracks: 1 - *O. vulgaris-O. sinensis,* **2**- *O. vulgaris - O. bimaculoides*, **3**- *O. vulgaris - A. fangsiao*. (A) Two breakpoints detected between *O. sinensis* and *O. vulgaris* at TAD borders. (B) Shared breakpoints in *O. vulgaris-O. bimaculoides* and *O. vulgaris- A. fangsiao* comparison, also between TADs. (C) Breakpoints detected in all three comparisons (ca. 175Mb and 181Mb).

Table 5.3 | An overview of the cases found in the datasets: For each chromosome, the number of different cases as defined in Table 5.1 were counted.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
Ovu01	2	1	21	0	0	8	3
Ovu02	5	2	16	0	0	9	0
Ovu03	3	4	17	0	1	10	2
Ovu04	3	1	11	0	0	9	1
Ovu05	3	4	13	0	0	6	2
Ovu06	3	9	7	0	0	8	2
Ovu07	2	5	10	0	1	9	1
Ovu08	8	6	8	1	0	10	1
Ovu09	0	1	16	0	0	10	0
Ovu10	2	4	10	0	0	7	0
Ovu11	8	1	6	0	1	5	1
Ovu12	5	1	11	0	1	3	2
Ovu13	0	3	11	0	0	9	0
Ovu14	2	0	10	0	0	7	0
Ovu15	4	0	10	1	0	3	2
Ovu16	2	3	6	0	0	1	6
Ovu17	4	1	8	0	1	5	1
Ovu18	2	0	9	0	1	6	1
Ovu19	7	1	6	0	0	2	0
Ovu20	0	1	8	0	3	0	0
Ovu21	3	0	8	0	0	3	0
Ovu22	2	0	4	0	1	4	2
Ovu23	3	1	6	0	3	0	0
Ovu25	0	1	9	0	0	1	2
Total	73	50	241	2	13	135	29
					Sum of cases		543

5.4. Discussion

We used four different octopus chromosome-scale genomes (Albertin et al., 2022; Destanović et al., 2023; Jiang et al., 2022; F. Li et al., 2020) to test the hypothesis that chromosomal rearrangement breakpoints tend to happen at TAD borders. As shown in Table 5.2, the chromosome-chromosome comparisons show varying results in supporting our hypothesis. In *O. bimaculoides* we observed significant signal for synteny breaks happening at TAD borders in only nine chromosomes, compared to 20 chromosomes for the same set of breakpoints in *O. vulgaris*. This can be explained by visualizing *O. bimaculoides* Hi-C data in HiGlass (Figure 5.4 C). There, we saw that TADs are not clearly defined in the Hi-C heatmap implying that insulation scores might be more homogenous throughout the genome, making it more difficult to distinguish areas between and within TADs (likely due to a poor Hi-C library). The results retrieved for the same breakpoints in the *O. vulgaris* genome that has a high resolution Hi-C matrix show significant results for 20 chromosomes supporting our hypothesis. Results for seven out of ten chromosomes that do not support the main hypothesis, are among the Ovu20-Ovu30 smaller chromosomes that have more gaps in the alignments. Therefore, we argue that results lacking support for our hypothesis might be because of a lower quality of Hi-C data in *O. bimaculoides*, and falsely detected breakpoint regions in the smaller chromosomes due to missing regions in the assemblies.

In the *O. sinensis* genome, data on 23 chromosomes supported our hypothesis, while for the same breakpoints, statistically significant patterns were observed for 13 chromosomes in *O. vulgaris*. *O. sinensis* was the first chromosome-scale octopus genome, and it has many unplaced scaffolds in the chromosome-scale scaffolds. This could have led to false detection of missing regions as breakpoints. In Table 5.3 we observe 73 breaks that are uniquely found in the comparison of *O. vulgaris* and *O. sinensis* genomes which is a high number considering that these two species split 2.5 mya. For example, eight unique breaks were observed in chromosome Ovu11 and its homolog in *O. sinensis*, compared to six unique breaks between Ovu11 and its homolog in *A. fangsiao* (divergence time: 44 mya).

In the comparison of *A. fangsiao* and *O. vulgaris* genomes, we obtained similar results. Data generated on 15 chromosomes in *A. fangsiao*, and 14 chromosomes in *O. vulgaris* provide support for our main hypothesis. We believe that inclusion of comparisons that are not limited to intrachromosomal rearrangements would make these tests more accurate. For example, any region that is present in *A. fangsiao*, but was translocated on a different chromosome in evolutionary history of either species, would be detected as a long breakpoint window, just as it is for missing regions. So, we would detect a real breakpoint, but the algorithm we use to estimate the potential breakpoint location and retrieve insulation score would not be accurate in this case.

Nevertheless, a one-tailed permutation test on overall data in each comparison supports the main hypothesis that synteny breaks tend to happen at TAD borders. This also shows that the algorithm for breakpoint detection works well overall. We also confirmed this by visualizing the breaks in HiGlass (Figure 5.4, Figure 5.5). By visual estimation, the breakpoint windows in most well-assembled bigger chromosomes are narrow and are likely real breakpoints (Ovu01-Ovu19, Ovu24, Ovu26).

This enables us to explore the breakpoints in the context of evolutionary history of the octopus lineage. Based on the data we have, we can pinpoint in which lineage some breakpoints happened and reconstruct the chromosomal evolutionary history to some degree. For example, we can say that occurrences of case 6 (breakpoints shared in *O. bimaculoides* and *A. fangsiao*) are rearrangements that likely happened in the common ancestor of *O. vulgaris* and *O. sinensis*. The most parsimonious explanation is that we observe the ancestral state in *O. bimaculoides* and *A. fangsiao*, and that the derived state is observed in *O. vulgaris* and *O. sinensis*. The most parsimonious explanation for the pattern seen in case 7 is that the ancestral state is present in all three compared species, but is derived only in *O. vulgaris*. Case 4 and 5 were detected in small numbers (2,13, respectively), but are still unlikely scenarios to be observed. For instance, case 4 indicates that the same state is present in *O. vulgaris* and *A. fangsiao*, and a different state in *O. sinensis* and *O. bimaculoides*. A biological explanation for this would be that two independent rearrangement events happened in *O. vulgaris* and *A. fangsiao* in the same place, or in *O. sinensis* and *O. bimaculoides*. Another scenario is that a rearrangement event happened in the *Octopus sp.* lineage, and that this event was reversed after the split from the common ancestor shared with *O. sinensis* in *O. vulgaris*. A similar explanation can be offered for case 5, but the pairs sharing a state are *O. vulgaris* and *O. bimaculoides*, and *O. sinensis* and *A. fangsiao*. However, these explanations are not parsimonious, and these observations are likely either due to a problem with the algorithm in Breakpointer2 (it is possible that the breakpoint was not detected in either *A. fangsiao* or *O. bimaculoides*), or an assembly problem (the same repetitive region can be missing in two genomes out of four making it seem like there is a breakpoint). When it comes to unique breakpoints (case 1-3), we can only say that breakpoints seen in the comparison with *O. bimaculoides* happened after the split from the common ancestor shared with *O. vulgaris*/*O. sinensis* lineage, while we can not pinpoint where the rearrangement happened in two other cases.

Analyzing breakpoints within the family Octopodidae on these four species gives insight into the dynamics of chromosome evolution over the 44 million years of evolution. Nevertheless, to understand the chromosomal evolution of the octopus lineage, we need to retrieve chromosome-scale genomes from other families within the order Octopodiformes. Even within the branch we studied (*Amphioctopus*-*Octopus*), we still do not understand the dynamics of chromosomal evolution. *A. fangsiao* has a genome size of 4.3 Gb (Jiang et al., 2022) which is closer to the size of genomes found in some squid species, such as *Doryteuthis (Loligo) pealeii* (4.6Gb) (Albertin et al., 2022), than 2.8Gb in *Octopus* species (Albertin et al., 2022; Destanović et al., 2023; F. Li et al., 2020). We speculate that the genomes were reduced in *Octopus* species, rather than reduced throughout evolutionary history of Octopodiformes, and then secondarily expanded in lineage of *Amphioctopus fangsiao*. Further comparative research is needed to understand this, and it would be ideal to have an octopus species from a different family as an outgroup to explore this.

5.5. Conclusion

We used Breakpointer2 to quantify rearrangements in octopus species pairs (*O. vulgaris*-*O. sinensis*, *O. vulgaris*-*O. bimaculoides*, *O. vulgaris*-*A. fangsiao*), and test whether the rearrangement breakpoints are collocalized with TAD borders. We found 543 different shared and unique breakpoints across 44 million years of evolutionary history of four tested species. Through a one-tailed permutation test on insulation scores of these breakpoints, we obtained a strong signal in support of the main hypothesis when we used whole-genome data, and for most chromosome-chromosome comparisons. Some inconsistencies in the results retrieved from chromosome-chromosome comparisons can be explained by limitations of the used chromosome-scale assemblies (misassemblies, misplaced, and missing regions), and of Breakpointer2. Nevertheless, this is the first analysis of TAD conservation on invertebrate non-model species and sets the stage for further research on evolutionary constraints on TADs, and evolution of coleoid cephalopods.

6. Future directions and outlook

In this research project, we assembled a chromosome-scale genome of *O. vulgaris*, created a tool that can be used to explore colocalization of TAD borders and chromosomal rearrangements (Breakpointer2), and ran this tool to explore evolutionary history and conservation of TADs in octopus lineage.

Although there were three chromosome-scale genomes of octopus species (Albertin et al., 2022; Jiang et al., 2022; Li et al., 2020), we wanted to generate a high-quality octopus genome that we can use as a reference in our comparisons and we assembled the *O. vulgaris* chromosome-scale genome (Destanović et al., 2023). We used Omni-C technique (DNase Hi-C) resulting in improved resolution in viewing genome structures, compared to the other three genomes that were scaffolded using the original Hi-C technique. This has proven one of the key steps in this research, as having a reference genome with a high resolution Hi-C matrix is essential to explore the genome structures. For example, limitations of *O. bimaculoides* genome (Albertin et al., 2022) is that the Hi-C data is of low resolution making it difficult to use to explore genome structures, and also it is missing around 0.4Gb of estimated genome size. *O. sinensis* (Li et al., 2020) and *A. fangsiao* (Jiang et al., 2022) genomes had a high enough resolution data to explore genomic structures. However, *O. sinensis* has missing regions in the main genome, which leads to false detection of breakpoints.

Due to the shortcomings of TAD-callers, we devised a tool (Breakpointer2) that uses a scale-free approach to differentiate regions within TADs and between TADs (TAD borders). Breakpointer2 calls and quantifies breakpoints, retrieves insulation scores for them, performs a one-tailed permutation test on each chromosome, and then on whole-genome data. Because the tool does not require assumptions, such as the size of TADs, biological features, etc., it can be easily used on other closely-related species without further improvements.

With changes in the implementation of Breakpointer2, we can also broaden the application of this tool. For now, it can only analyze closely-related species. One of the future improvements of this tool includes changing its algorithm so it does not analyze only homologous chromosomes. This will enable interchromosomal analyses and analyses of more distantly-related species. Also, we would like to implement a filter for breakpoint window length to prevent the missing regions from being detected as a breakpoint. Another goal would be to implement an algorithm that would differentiate between and count the types of genomic rearrangements.

By running Breakpointer2 on species pairs (*O. vulgaris*-*O. sinensis*, *O. vulgaris*-*O. bimaculoides*, *O. vulgaris*-*A. fangsiao*), we gathered evidence that synteny breaks tend to colocalize with TAD borders. Nevertheless, this does not mean that chromosomal rearrangements happen only at TAD boundaries, and further research is needed to inspect other potential cases. One of the potential future directions is to analyze types of genes that are located near these breakpoints. For example, in *Drosophila* it has been found that TADs containing developmental genes tend to be more conserved compared to other types of TADs (Torosin et al., 2022). However, to be able to make conclusions about chromosomal evolution in octopus lineage, further research is needed.

This thesis sets the stage for further investigation into how these ancient units are evolving within particular clades of coleoids. Breakpointer2 can be used on chromosome-scale genomes of squid as well, as they have a conserved karyotype of n=46 (Albertin et al., 2022; M. Li et al., 2022), so the algorithm based on homologous chromosomes would work well. The squid genomes are larger compared to octopus genomes, and other inversion dynamics can be expected. Also, it would be beneficial to expand the research in the light of cephalopod-specific microsyntenies (microsyntenies associated with cephalopod innovations (MACIs)). Presumably MACIs arose in evolutionary history when the events of major genomic reshuffling happened (Albertin et al., 2022; Ritschard et al., 2019; Schmidbaur et al., 2022). If the rearrangements in the common ancestor of coleoids lead to an emergence of novel microsyntenies, exploration of rearrangements we discovered in octopus lineage might have disrupted some of the ancient metazoan syntenies as well, and formed novel microsyntenies specific for the octopus lineage.

In general, more work is needed to understand how intrachromosomal rearrangements shaped octopod evolution. In reconstruction of evolutionary history of octopuses (by comparing genomes of *Nautilus pompilius*, *A. fangsiao*, and *O. sinensis*), it was suggested that the ancestral karyotype of the common ancestor of was n=26, and that 31 fissions, and 30 subsequent fusions happened in the common octopus ancestor, including loss of two chromosomes, and gain of six (Jiang et al., 2022). In the light of recent findings that chromosomal fusions impact recombination landscapes in primary spermatocytes (Vara et al., 2021), this means that octopus lineage might have had a higher rate of rearrangement events. A future challenge is understanding how intrachromosomal rearrangement events affected the evolution of the octopus lineage compared to squids.

References:

Abbott, N. J., Williamson, R., & Maddock, L. (1995). *Cephalopod NeurobiologyNeuroscience Studies in Squid, Octopus and Cuttlefish*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198547907.001.0001

Acemel, R. D., Maeso, I., & Gómez-Skarmeta, J. L. (2017). Topologically associated domains: A successful scaffold for the evolution of gene regulation in animals. *WIREs Developmental Biology*, *6*(3). https://doi.org/10.1002/wdev.265

Adachi, K., Ohnishi, K., Kuramochi, T., Yoshinaga, T., & Okumura, S.-I. (2014). Molecular cytogenetic study in *Octopus* (*Amphioctopus*) *areolatus* from Japan. *Fisheries Science*, *80*(3), 445–450. https://doi.org/10.1007/s12562-014-0703-4

Aho, A. V., Kernighan, B. W., & Weinberger, P. J. (1988). *The AWK programming language*. Addison-Wesley Pub. Co.

Albertin, C. B., Medina-Ruiz, S., Mitros, T., Schmidbaur, H., Sanchez, G., Wang, Z. Y., Grimwood, J., Rosenthal, J. J. C., Ragsdale, C. W., Simakov, O., & Rokhsar, D. S. (2022). Genome and transcriptome mechanisms driving cephalopod evolution. *Nature Communications*, *13*(1), 2427. https://doi.org/10.1038/s41467-022-29748-w

Albertin, C. B., Simakov, O., Mitros, T., Wang, Z. Y., Pungor, J. R., Edsinger-Gonzales, E., Brenner, S., Ragsdale, C. W., & Rokhsar, D. S. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, *524*(7564), 220–224. https://doi.org/10.1038/nature14668

Álvarez-González, L., Burden, F., Doddamani, D., Malinverni, R., Leach, E., Marín-García, C., Marín-Gual, L., Gubern, A., Vara, C., Paytuví-Gallart, A., Buschbeck, M., Ellis, P. J. I., Farré, M., & Ruiz-Herrera, A. (2022). 3D chromatin remodelling in the germ line modulates genome evolutionary plasticity. *Nature Communications*, *13*(1), 2608.

https://doi.org/10.1038/s41467-022-30296-6

Amor, M. D., Doyle, S. R., Norman, M. D., Roura, A., Hall, N. E., Robinson, A. J., Leite, T. S., & Strugnell, J. M. (2019). *Genome-wide sequencing uncovers cryptic diversity and mito-nuclear discordance in the* Octopus vulgaris *species complex* [Preprint]. Molecular Biology. https://doi.org/10.1101/573493

Amor, M. D., Norman, M. D., Cameron, H. E., & Strugnell, J. M. (2014). Allopatric Speciation within a Cryptic Species Complex of Australasian Octopuses. *PLoS ONE*, *9*(6), e98982. https://doi.org/10.1371/journal.pone.0098982

Amor, M. D., Norman, M. D., Roura, A., Leite, T. S., Gleadall, I. G., Reid, A., Perales-Raya, C., Lu, C., Silvey, C. J., Vidal, E. A. G., Hochberg, F. G., Zheng, X., & Strugnell, J. M. (2017). Morphological assessment of the *Octopus vulgaris* species complex evaluated in light of molecular-based phylogenetic inferences. *Zoologica Scripta*, *46*(3), 275–288. https://doi.org/10.1111/zsc.12207

Annunziato, A. (2008). DNA packaging: Nucleosomes and chromatin. *Nature Education*, *1*(1), 26.

Avendaño, O., Roura, Á., Cedillo-Robles, C. E., González, Á. F., Rodríguez-Canul, R., Velázquez-Abunader, I., & Guerra, Á. (2020). Octopus americanus: A cryptic species of the O. vulgaris species complex redescribed from the Caribbean. *Aquatic Ecology*, *54*(4), 909–925. https://doi.org/10.1007/s10452-020-09778-6

Ballare, K. M., Escalona, M., Barr, K., Seligmann, W., Sacco, S., Sahasrabudhe, R. M., Nguyen, O., Wyckoff, C., Smith, T. B., & Shapiro, B. (2023). A reference genome assembly of the declining tricolored blackbird, *Agelaius tricolor*. *Journal of Heredity*, *114*(1), 44–51. https://doi.org/10.1093/jhered/esac053

Beagan, J. A., & Phillips-Cremins, J. E. (2020). On the existence and functionality of topologically associating domains. *Nature Genetics*, *52*(1), 8–16.

https://doi.org/10.1038/s41588-019-0561-1

Berdan, E. L., Blanckaert, A., Butlin, R. K., & Bank, C. (2021). Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *PLOS Genetics*, *17*(3), e1009411. https://doi.org/10.1371/journal.pgen.1009411

Bhat, A., Ghatage, T., Bhan, S., Lahane, G. P., Dhar, A., Kumar, R., Pandita, R. K., Bhat, K. M., Ramos, K. S., & Pandita, T. K. (2022). Role of Transposable Elements in Genome Stability: Implications for Health and Disease. *International Journal of Molecular Sciences*, *23*(14), 7802. https://doi.org/10.3390/ijms23147802

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, *31*(12), 1119–1125. https://doi.org/10.1038/nbt.2727

Cabanettes, F., & Klopp, C. (2018). D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *6*, e4958. https://doi.org/10.7717/peerj.4958

Cáceres, M., Ranz, J. M., Barbadilla, A., Long, M., & Ruiz, A. (1999). Generation of a Widespread Drosophila Inversion by a Transposable Element. *Science*, *285*(5426), 415–418. JSTOR.

Cazet, J. F., Siebert, S., Little, H. M., Bertemes, P., Primack, A. S., Ladurner, P., Achrainer, M., Fredriksen, M. T., Moreland, R. T., Singh, S., Zhang, S., Wolfsberg, T. G., Schnitzler, C. E., Baxevanis, A. D., Simakov, O., Hobmayer, B., & Juliano, C. E. (2022). *New* Hydra *genomes reveal conserved principles of hydrozoan transcriptional regulation* [Preprint]. Developmental Biology. https://doi.org/10.1101/2022.06.21.496857

Cheng, Y., Shang, D., Luo, M., Huang, C., Lai, F., Wang, X., Xu, X., Ying, R., Wang, L., Zhao, Y., Zhang, L., Long, M., Cheng, H., & Zhou, R. (2020). Whole genome-wide

chromosome fusion and new gene birth in the Monopterus albus genome. *Cell & Bioscience*, *10*(1), 67. https://doi.org/10.1186/s13578-020-00432-0

Crane, E., Bian, Q., McCord, R. P., Lajoie, B. R., Wheeler, B. S., Ralston, E. J., Uzawa, S., Dekker, J., & Meyer, B. J. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, *523*(7559), 240–244. https://doi.org/10.1038/nature14450

Cremer, T., & Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, *2*(4), 292–301. https://doi.org/10.1038/35066075

Destanović, D., Schultz, D. T., Styfhals, R., Cruz, F., Gómez-Garrido, J., Gut, M., Gut, I., Fiorito, G., Simakov, O., Alioto, T. S., Ponte, G., & Seuntjens, E. (2023). *A chromosome-level reference genome for the common octopus,* Octopus vulgaris *(Cuvier, 1797)* [Preprint]. Genomics. https://doi.org/10.1101/2023.05.16.540928

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. https://doi.org/10.1038/nature11082

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., & Aiden, E. L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95. https://doi.org/10.1126/science.aal3327

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems*, *3*(1), 99–101. https://doi.org/10.1016/j.cels.2015.07.012

Eres, I. E., & Gilad, Y. (2021). A TAD Skeptic: Is 3D Genome Topology Conserved? *Trends in Genetics*, *37*(3), 216–223. https://doi.org/10.1016/j.tig.2020.10.009

Eres, I. E., Luo, K., Hsiao, C. J., Blake, L. E., & Gilad, Y. (2019). Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLOS Genetics*, *15*(7), e1008278. https://doi.org/10.1371/journal.pgen.1008278

Fritz, A. J., Sehgal, N., Pliss, A., Xu, J., & Berezney, R. (2019). Chromosome territories and the global regulation of the genome. *Genes, Chromosomes and Cancer*, *58*(7), 407–426. https://doi.org/10.1002/gcc.22732

G. Gleadall, I. (2016). *Octopus sinensis* d'Orbigny, 1841 (Cephalopoda: Octopodidae): Valid Species Name for the Commercially Valuable East Asian Common Octopus. *Species Diversity*, *21*(1), 31–42. https://doi.org/10.12782/sd.21.1.031

GAO, Y. M., & NATSUKARI, Y. (1990). *Karyological Studies on Seven Cephalopods* (No. 2). The Malacological Society of Japan. https://doi.org/10.18941/venusjjm.49.2_126

Gates, R. R. (1921). Mutations and Evolution. *Nature*, *107*(2701), 714–715. https://doi.org/10.1038/107714c0

Gilbertson, S. E., Walter, H. C., Gardner, K., Wren, S. N., Vahedi, G., & Weinmann, A. S. (2022). Topologically associating domains are disrupted by evolutionary genome rearrangements forming species-specific enhancer connections in mice and humans. *Cell Reports*, *39*(5), 110769. https://doi.org/10.1016/j.celrep.2022.110769

Gómez-Marín, C., Tena, J. J., Acemel, R. D., López-Mayorga, M., Naranjo, S., De La Calle-Mustienes, E., Maeso, I., Beccari, L., Aneas, I., Vielmas, E., Bovolenta, P., Nobrega, M. A., Carvajal, J., & Gómez-Skarmeta, J. L. (2015). Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proceedings of the National Academy of Sciences*, *112*(24), 7542–7547. https://doi.org/10.1073/pnas.1505463112

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hochner, B. (2012). An Embodied View of Octopus Neurobiology. *Current Biology*, *22*(20), R887–R892. https://doi.org/10.1016/j.cub.2012.09.001

Hofstatter, P. G., Thangavel, G., Lux, T., Neumann, P., Vondrak, T., Novak, P., Zhang, M., Costa, L., Castellani, M., Scott, A., Toegelová, H., Fuchs, J., Mata-Sucre, Y., Dias, Y., Vanzela, A. L. L., Huettel, B., Almeida, C. C. S., Šimková, H., Souza, G., … Marques, A. (2022). Repeat-based holocentromeres influence genome architecture and karyotype evolution. *Cell*, *185*(17), 3153-3168.e18. https://doi.org/10.1016/j.cell.2022.06.045

Hsieh, T.-H. S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., & Rando, O. J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, *162*(1), 108–119. https://doi.org/10.1016/j.cell.2015.05.048

Huang, K., & Rieseberg, L. H. (2020). Frequency, Origins, and Evolutionary Role of Chromosomal Inversions in Plants. *Frontiers in Plant Science*, *11*, 296. https://doi.org/10.3389/fpls.2020.00296

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Huynh, L., & Hormozdiari, F. (2019). TAD fusion score: Discovery and ranking the contribution of deletions to genome structure. *Genome Biology*, *20*(1), 60. https://doi.org/10.1186/s13059-019-1666-7

Jiang, D., Liu, Q., Sun, J., Liu, S., Fan, G., Wang, L., Zhang, Y., Seim, I., An, S., Liu, X., Li,

Q., & Zheng, X. (2022). The gold-ringed octopus (Amphioctopus fangsiao) genome and cerebral single-nucleus transcriptomes provide insights into the evolution of karyotype and neural novelties. *BMC Biology*, *20*(1), 289. https://doi.org/10.1186/s12915-022-01500-2

Kaplan, N., & Dekker, J. (2013). High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nature Biotechnology*, *31*(12), 1143–1147. https://doi.org/10.1038/nbt.2768

Kenny, N. J., Francis, W. R., Rivera-Vicéns, R. E., Juravel, K., de Mendoza, A., Díez-Vives, C., Lister, R., Bezares-Calderón, L. A., Grombacher, L., Roller, M., Barlow, L. D., Camilli, S., Ryan, J. F., Wörheide, G., Hill, A. L., Riesgo, A., & Leys, S. P. (2020). Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge Ephydatia muelleri. *Nature Communications*, *11*(1), 3676. https://doi.org/10.1038/s41467-020-17397-w

Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J. M., Ouellette, S. B., Azhir, A., Kumar, N., Hwang, J., Lee, S., Alver, B. H., Pfister, H., Mirny, L. A., Park, P. J., & Gehlenborg, N. (2018). HiGlass: Web-based visual exploration and analysis of genome interaction maps. *Genome Biology*, *19*(1), 125. https://doi.org/10.1186/s13059-018-1486-1

Kim, B.-M., Kang, S., Ahn, D.-H., Jung, S.-H., Rhee, H., Yoo, J. S., Lee, J.-E., Lee, S., Han, Y.-H., Ryu, K.-B., Cho, S.-J., Park, H., & An, H. S. (2018). The genome of common long-arm octopus Octopus minor. *GigaScience*. https://doi.org/10.1093/gigascience/giy119

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Krefting, J., Andrade-Navarro, M. A., & Ibn-Salem, J. (2018). Evolutionary stability of

topologically associating domains is associated with conserved gene regulation. *BMC Biology*, *16*(1), 87. https://doi.org/10.1186/s12915-018-0556-x

Kruse, K., Hug, C. B., & Vaquerizas, J. M. (2020). FAN-C: A feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biology*, *21*(1), 303. https://doi.org/10.1186/s13059-020-02215-9

Lazar, N. H., Nevonen, K. A., O'Connell, B., McCann, C., O'Neill, R. J., Green, R. E., Meyer, T. J., Okhovat, M., & Carbone, L. (2018). Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Research*, *28*(7), 983–997. https://doi.org/10.1101/gr.233874.117

Leite, T. S., Haimovici, M., Molina, W., & Warnke, K. (2008). Morphological and genetic description of Octopus insularis, a new cryptic species in the Octopus vulgaris complex (Cephalopoda: Octopodidae) from the tropical southwestern Atlantic. *Journal of Molluscan Studies*, *74*(1), 63–74. https://doi.org/10.1093/mollus/eym050

Li, F., Bian, L., Ge, J., Han, F., Liu, Z., Li, X., Liu, Y., Lin, Z., Shi, H., Liu, C., Chang, Q., Lu, B., Zhang, S., Hu, J., Xu, D., Shao, C., & Chen, S. (2020). Chromosome-level genome assembly of the East Asian common octopus (*Octopus sinensis*) using PacBio sequencing and Hi-C technology. *Molecular Ecology Resources*, *20*(6), 1572–1582. https://doi.org/10.1111/1755-0998.13216

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H. (2021). New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, *37*(23), 4572–4574. https://doi.org/10.1093/bioinformatics/btab705

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, M., Wu, B., Zhang, P., Li, Y., Xu, W., Wang, K., Qiu, Q., Zhang, J., Li, J., Zhang, C., Fan, J., Feng, C., & Chen, Z. (2022). Genomes of Two Flying Squid Species Provide Novel Insights into Adaptations of Cephalopods to Pelagic Life. *Genomics, Proteomics & Bioinformatics*, *20*(6), 1053–1065. https://doi.org/10.1016/j.gpb.2022.09.009

Li, Q., Li, H., Huang, W., Xu, Y., Zhou, Q., Wang, S., Ruan, J., Huang, S., & Zhang, Z. (2019). A chromosome-scale genome assembly of cucumber (Cucumis sativus L.). *GigaScience*, *8*(6), giz072. https://doi.org/10.1093/gigascience/giz072

Liao, Y., Zhang, X., Chakraborty, M., & Emerson, J. J. (2021). Topologically associating domains and their role in the evolution of genome structure and function in *Drosophila*. *Genome Research*, *31*(3), 397–410. https://doi.org/10.1101/gr.266130.120

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, *326*(5950), 289–293. https://doi.org/10.1126/science.1181369

Liu, K., Li, H., Li, Y., Wang, J., & Wang, J. (2022). A comparison of topologically associating domain callers based on Hi-C data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1–1. https://doi.org/10.1109/TCBB.2022.3147805

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., … Mundlos, S. (2015). Disruptions of Topological Chromatin Domains

Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, *161*(5), 1012–1025. https://doi.org/10.1016/j.cell.2015.04.004

Lupiáñez, D. G., Spielmann, M., & Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics*, *32*(4), 225–237. https://doi.org/10.1016/j.tig.2016.01.003

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Matharu, N., & Ahanger, S. (2015). Chromatin Insulators and Topological Domains: Adding New Dimensions to 3D Genome Architecture. *Genes*, *6*(3), 790–811. https://doi.org/10.3390/genes6030790

Melo, U. S., Schöpflin, R., Acuna-Hidalgo, R., Mensah, M. A., Fischer-Zirnsak, B., Holtgrewe, M., Klever, M.-K., Türkmen, S., Heinrich, V., Pluym, I. D., Matoso, E., Bernardo de Sousa, S., Louro, P., Hülsemann, W., Cohen, M., Dufke, A., Latos-Bieleńska, A., Vingron, M., Kalscheuer, V., … Mundlos, S. (2020). Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases. *The American Journal of Human Genetics*, *106*(6), 872–884. https://doi.org/10.1016/j.ajhg.2020.04.016

Muller, H. J. (1918). GENETIC VARIABILITY, TWIN HYBRIDS AND CONSTANT HYBRIDS, IN A CASE OF BALANCED LETHAL FACTORS. *Genetics*, *3*(5), 422–499. https://doi.org/10.1093/genetics/3.5.422

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., & Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in

chromosome structure. *Nature*, *502*(7469), 59–64. https://doi.org/10.1038/nature12593

Neel, J. V. (1949). The detection of the genetic carriers of hereditary disease. *American Journal of Human Genetics*, *1*(1), 19–36.

Nishimura, O., Hara, Y., & Kuraku, S. (2017). GVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics*, *33*(22), 3635–3637. https://doi.org/10.1093/bioinformatics/btx445

Nishimura, O., Hara, Y., & Kuraku, S. (2019). Evaluating Genome Assemblies and Gene Models Using gVolante. In M. Kollmar (Ed.), *Gene Prediction* (Vol. 1962, pp. 247–256). Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_15

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., & Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381–385. https://doi.org/10.1038/nature11049

Nossa, C. W., Havlak, P., Yue, J.-X., Lv, J., Vincent, K. Y., Brockmann, H. J., & Putnam, N. H. (2014). Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication. *GigaScience*, *3*(1), 9. https://doi.org/10.1186/2047-217X-3-9

Ong, C.-T., & Corces, V. G. (2014). CTCF: An architectural protein bridging genome topology and function. *Nature Reviews Genetics*, *15*(4), 234–246. https://doi.org/10.1038/nrg3663

Porubsky, D., Sanders, A. D., Höps, W., Hsieh, P., Sulovari, A., Li, R., Mercuri, L., Sorensen, M., Murali, S. C., Gordon, D., Cantsilieris, S., Pollen, A. A., Ventura, M., Antonacci, F., Marschall, T., Korbel, J. O., & Eichler, E. E. (2020). Recurrent inversion toggling

and great ape genome evolution. *Nature Genetics*, *52*(8), 849–858. https://doi.org/10.1038/s41588-020-0646-x

Preger-Ben Noon, E., & Frankel, N. (2022). Can changes in 3D genome architecture create new regulatory landscapes that contribute to phenotypic evolution? *Essays in Biochemistry*, *66*(6), 745–752. https://doi.org/10.1042/EBC20220057

Quinodoz, S. A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J. M., Detmar, E., Lai, M. M., Shishkin, A. A., Bhat, P., Takei, Y., Trinh, V., Aznauryan, E., Russell, P., Cheng, C., Jovanovic, M., Chow, A., Cai, L., McDonel, P., Garber, M., & Guttman, M. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell*, *174*(3), 744-757.e24. https://doi.org/10.1016/j.cell.2018.05.024

Ramani, V., Cusanovich, D. A., Hause, R. J., Ma, W., Qiu, R., Deng, X., Blau, C. A., Disteche, C. M., Noble, W. S., Shendure, J., & Duan, Z. (2016). Mapping 3D genome architecture through in situ DNase Hi-C. *Nature Protocols*, *11*(11), 2104–2121. https://doi.org/10.1038/nprot.2016.126

Raybaut, P. (2009). *Spyder-documentation*.

Rosin, L. F., Crocker, O., Isenhart, R. L., Nguyen, S. C., Xu, Z., & Joyce, E. F. (2019). Chromosome territory formation attenuates the translocation potential of cells. *ELife*, *8*, e49553. https://doi.org/10.7554/eLife.49553

Rowley, M. J., & Corces, V. G. (2018). Organizational principles of 3D genome architecture. *Nature Reviews Genetics*, *19*(12), 789–800. https://doi.org/10.1038/s41576-018-0060-8

Sawh, A. N., & Mango, S. E. (2022). Chromosome organization in 4D: Insights from C. elegans development. *Current Opinion in Genetics & Development*, *75*, 101939. https://doi.org/10.1016/j.gde.2022.101939

Schmidbaur, H., Kawaguchi, A., Clarence, T., Fu, X., Hoang, O. P., Zimmermann, B.,

Ritschard, E. A., Weissenbacher, A., Foster, J. S., Nyholm, S. V., Bates, P. A., Albertin, C. B., Tanaka, E., & Simakov, O. (2022). Emergence of novel cephalopod gene regulation and expression through large-scale genome reorganization. *Nature Communications*, *13*(1), 2172. https://doi.org/10.1038/s41467-022-29694-7

Schultz, D. T., Francis, W. R., McBroome, J. D., Christianson, L. M., Haddock, S. H. D., & Green, R. E. (2021). A chromosome-scale genome assembly and karyotype of the ctenophore *Hormiphora californensis*. *G3 Genes|Genomes|Genetics*, *11*(11), jkab302. https://doi.org/10.1093/g3journal/jkab302

Sefer, E. (2022). A comparison of topologically associating domain callers over mammals at high resolution. *BMC Bioinformatics*, *23*(1), 127. https://doi.org/10.1186/s12859-022-04674-2

Selvaraj, S., R Dixon, J., Bansal, V., & Ren, B. (2013). Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotechnology*, *31*(12), 1111–1118. https://doi.org/10.1038/nbt.2728

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., & Cavalli, G. (2012). Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell*, *148*(3), 458–472. https://doi.org/10.1016/j.cell.2012.01.010

Sharma, S. P., Zuo, T., & Peterson, T. (2021). Transposon-induced inversions activate gene expression in the maize pericarp. *Genetics*, *218*(2), iyab062. https://doi.org/10.1093/genetics/iyab062

Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F., & Zhou, X. J. (2016). TopDom: An efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Research*, *44*(7), e70–e70. https://doi.org/10.1093/nar/gkv1505

Szabo, Q., Bantignies, F., & Cavalli, G. (2019). Principles of genome folding into

topologically associating domains. *Science Advances*, *5*(4), eaaw1668. https://doi.org/10.1126/sciadv.aaw1668

Team, T. P. D. (2023). *pandas-dev/pandas: Pandas* (v2.0.0). Zenodo. https://doi.org/10.5281/ZENODO.3509134

Tena, J. J., & Santos-Pereira, J. M. (2021a). Topologically Associating Domains and Regulatory Landscapes in Development, Evolution and Disease. *Frontiers in Cell and Developmental Biology*, *9*, 702787. https://doi.org/10.3389/fcell.2021.702787

Tena, J. J., & Santos-Pereira, J. M. (2021b). Topologically Associating Domains and Regulatory Landscapes in Development, Evolution and Disease. *Frontiers in Cell and Developmental Biology*, *9*, 702787. https://doi.org/10.3389/fcell.2021.702787

Torosin, N. S., Anand, A., Golla, T. R., Cao, W., & Ellison, C. E. (2020). 3D genome evolution and reorganization in the Drosophila melanogaster species group. *PLOS Genetics*, *16*(12), e1009229. https://doi.org/10.1371/journal.pgen.1009229

Torosin, N. S., Golla, T. R., Lawlor, M. A., Cao, W., & Ellison, C. E. (2022). Mode and Tempo of 3D Genome Evolution in *Drosophila*. *Molecular Biology and Evolution*, *39*(11), msac216. https://doi.org/10.1093/molbev/msac216

Van Bortle, K., Nichols, M. H., Li, L., Ong, C.-T., Takenaka, N., Qin, Z. S., & Corces, V. G. (2014). Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biology*, *15*(5), R82. https://doi.org/10.1186/gb-2014-15-5-r82

Vara, C., Paytuví-Gallart, A., Cuartero, Y., Álvarez-González, L., Marín-Gual, L., Garcia, F., Florit-Sabater, B., Capilla, L., Sanchéz-Guillén, R. A., Sarrate, Z., Aiese Cigliano, R., Sanseverino, W., Searle, J. B., Ventura, J., Marti-Renom, M. A., Le Dily, F., & Ruiz-Herrera, A. (2021). The impact of chromosomal fusions on 3D genome folding and recombination in the germ line. *Nature Communications*, *12*(1), 2981.

https://doi.org/10.1038/s41467-021-23270-1

Wang, J., & Zheng, X. (2017). Comparison of the genetic relationship between nine Cephalopod species based on cluster analysis of karyotype evolutionary distance. *Comparative Cytogenetics*, *11*(3), 477–494. https://doi.org/10.3897/compcytogen.v11i3.12752

Wang, M., Wang, P., Lin, M., Ye, Z., Li, G., Tu, L., Shen, C., Li, J., Yang, Q., & Zhang, X. (2018). Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. *Nature Plants*, *4*(2), 90–97. https://doi.org/10.1038/s41477-017-0096-3

Webster, M., Witkin, K. L., & Cohen-Fix, O. (2009). Sizing up the nucleus: Nuclear shape, size and nuclear-envelope assembly. *Journal of Cell Science*, *122*(10), 1477–1486. https://doi.org/10.1242/jcs.037333

Wilder, A. P., Dudchenko, O., Curry, C., Korody, M., Turbek, S. P., Daly, M., Misuraca, A., Wang, G., Khan, R., Weisz, D., Fronczek, J., Aiden, E. L., Houck, M. L., Shier, D. M., Ryder, O. A., & Steiner, C. C. (2022). A Chromosome-Length Reference Genome for the Endangered Pacific Pocket Mouse Reveals Recent Inbreeding in a Historically Large Population. *Genome Biology and Evolution*, *14*(8), evac122. https://doi.org/10.1093/gbe/evac122

Zarrella, I., Herten, K., Maes, G. E., Tai, S., Yang, M., Seuntjens, E., Ritschard, E. A., Zach, M., Styfhals, R., Sanges, R., Simakov, O., Ponte, G., & Fiorito, G. (2019). The survey and reference assisted assembly of the Octopus vulgaris genome. *Scientific Data*, *6*(1), 13. https://doi.org/10.1038/s41597-019-0017-6

Zarrella, I., Ponte, G., Baldascino, E., & Fiorito, G. (2015). Learning and memory in Octopus vulgaris: A case of biological plasticity. *Current Opinion in Neurobiology*, *35*, 74–79. https://doi.org/10.1016/j.conb.2015.06.012

Zhang, H., Song, L., Wang, X., Cheng, H., Wang, C., Meyer, C. A., Liu, T., Tang, M., Aluru,

S., Yue, F., Liu, X. S., & Li, H. (2021). Fast alignment and preprocessing of chromatin profiles with Chromap. *Nature Communications*, *12*(1), 6566. https://doi.org/10.1038/s41467-021-26865-w

Zhang, J., & Peterson, T. (2004). Transposition of Reversed Ac Element Ends Generates Chromosome Rearrangements in Maize. *Genetics*, *167*(4), 1929–1937. https://doi.org/10.1534/genetics.103.026229

Zhang, J., & Peterson, T. (2005). A Segmental Deletion Series Generated by Sister-Chromatid Transposition of *Ac* Transposable Elements in Maize. *Genetics*, *171*(1), 333–344. https://doi.org/10.1534/genetics.104.035576

Zhang, J., Zhang, F., & Peterson, T. (2006). Transposition of Reversed Ac Element Ends Generates Novel Chimeric Genes in Maize. *PLoS Genetics*, *2*(10), e164. https://doi.org/10.1371/journal.pgen.0020164

Zhang, J., Zuo, T., & Peterson, T. (2013). Generation of Tandem Direct Duplications by Reversed-Ends Transposition of Maize Ac Elements. *PLoS Genetics*, *9*(8), e1003691. https://doi.org/10.1371/journal.pgen.1003691

Zimmermann, B., Robb, S. M. C., Genikhovich, G., Fropf, W. J., Weilguny, L., He, S., Chen, S., Lovegrove-Walsh, J., Hill, E. M., Chen, C.-Y., Ragkousi, K., Praher, D., Fredman, D., Moran, Y., Gibson, M. C., & Technau, U. (2020). *Sea anemone genomes reveal ancestral metazoan chromosomal macrosynteny* [Preprint]. Genomics. https://doi.org/10.1101/2020.10.30.359448

Zufferey, M., Tavernari, D., Oricchio, E., & Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. *Genome Biology*, *19*(1), 217. https://doi.org/10.1186/s13059-018-1596-9