A deep dive into genome assemblies of non-vertebrate animals

Nadège Guiglielmoni[1,2,*], Ramón E. Rivera-Vicéns[3], Romain Koszul[4], and Jean-François Flot[1,5]

[1]Service Evolution Biologique et Ecologie, Université libre de Bruxelles (ULB), 1050 Brussels, Belgium

[2]Institut für Zoologie, Universität zu Köln, 50674 Cologne, Germany

[3]Department of Earth and Environmental Sciences, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, 80333 Munich, Germany

[4]Institut Pasteur, Unité Régulation Spatiale des Génomes, UMR 3525, CNRS, 75015 Paris, France

[5]Interuniversity Institute of Bioinformatics in Brussels – (IB)^2, 1050 Brussels, Belgium

[*]Corresponding author: nguiglie@uni-koeln.de

Abstract

Non-vertebrate species represent about ∼95% of known metazoan (animal) diversity. They remain to this day relatively unexplored genetically, but understanding their genome structure and function is pivotal for expanding our current knowledge of evolution, ecology and biodiversity. Following the continuous improvements and decreasing costs of sequencing technologies, many genome assembly tools have been released, leading to a significant amount of genome projects being completed in recent years. In this review, we examine the current state of genome projects of non-vertebrate animal species. We present an overview of available sequencing technologies, assembly approaches, as well as pre and post-processing steps, genome assembly evaluation methods, and their application to non-vertebrate animal genomes.

keywords: genome assembly, sequencing, non-vertebrate animals

Introduction

The field of genomics is presently thriving, with new genomes of all kind of organisms becoming available every day. For Metazoa, efforts have unsurprisingly focused on human's closest relatives (i.e., vertebrates) so far [1]:

out of 7;894 metazoan assemblies available in the GenBank database (accessed on October 29th, 2021) [2], ∼ 56.9% (4,493) belong to the subphylum Vertebrata. However, from the currently ∼2.1 million described metazoan species, only ∼73,000 (3.5%) belong to vertebrates [3]. The remaining metazoan phyla, hereafter called "non-vertebrate animals", are thus underinvestigated and lack genetic resources.

Non-vertebrate animals are found in nearly all known terrestrial and aquatic ecosystems (both marine and freshwater), and represent the diverse branches of the metazoan tree of life (among which vertebrates are just a twig that originated about 600 millions years ago [4]). Characterizing the genome structure and gene content of non-vertebrate animals is therefore pivotal for expanding our knowledge regarding the evolution, ecology and biodiversity of metazoans.

In recent years, important sequencing efforts have started to tackle the dearth of genomic data for non-vertebrate animals, with a strong focus on arthropods (2,683 assemblies on GenBank). The phylum Arthropoda is very diverse: it consists of more than 1.3 million species, the majority of which belong to the class Insecta (∼1 million species) [5]. Insects have a significant impact on agriculture (e.g. as crop pests) and on the transmission of diseases (e.g. malaria and dengue) [6]. They also play important beneficial and regulatory roles in natural ecosystems, through pollination and decomposition of organic matter [7]. Genome sequencing yields invaluable insights into species that are key in the aforementioned processes. For example, various genome projects have targeted insects such as *Bemisia tabaci*, a common crop pest [8], and the mosquitoes *Aedes aegypti* (vector of yellow fever, dengue and chikungunya) [9] and *Anopheles darlingi* (vector of malaria) [10]. These studies unveiled, among other findings, expansions of genes involved in insecticide resistance. The genomes of these species are so important for human health and food security that many have actually been sequenced multiple times, either because of the availability of newer sequencing methods or to compare different strains (for instance, three versions of the genome of *Aedes aegypti* [11, 12, 9] were successively published). Many phyla with less direct human implications, however, do not even have a single good-quality genome assembly available to date (e.g., chaetognaths).

Other non-vertebrates (and their symbionts) have also shown tremendous importance and relevance with respect to socio-economic impact. Snails, sponges and corals all produce metabolites with biological activities such as anticancer, anti-inflammatory, antibacterial, among others [13, 14, 15]. Terpenoid metabolites have been found in more than 70 gastropod species [16]. In sponges, compounds such as polyketides, terpenoids and alkaloids have also been found in species of the genera *Haliclona*, *Petrosia*, and *Discodemia*, these three genera

being the richest among sponges in terms of bioactive compounds [17]. Thus, genome assemblies are essential to identify and better understand the genes, pathways and sources of these compounds. Among mollusks, several species valued as food resources are studied for their impact in aquaculture [18]. Moreover, non-vertebrates are important model systems to understand processes such as adaptation to climate change, ocean acidification, biomineralization [19, 20, 21, 22]. Various species of corals [23, 24, 25, 26] have been sequenced to study the effects of increasing seawater temperatures and to understand how these species may survive in changing environments.

Some genome projects are motivated by more theoretical questions, to improve species classification and elucidate specific traits. Genome assemblies provide abundant sets of genes to build robust phylogenetic trees, opening the field of phylogenomics [27]. New genome resources bring novel insights into difficult phylogenetic positions: a large analysis based on genomes and transcriptomes confirmed that myxozoans belonged to Cnidaria [28]; the sequence of *Hoilungia hongkongiensis* placed placozoans as a sister group to cnidarians and bilaterians [29].

The lack of non-vertebrate genomic resources may be blamed to the difficulty to collect individuals or extract pure, high-molecular-weight DNA, as well as to their frequently large genomes characterized by high repetitive contents and high heterozygosity. However, sequencing technologies now offer cost-effective solutions and wide applicability to solve some of these problems. Reducing the current unbalance in genomic resources between vertebrates and non-vertebrate animals will increase the precision of future tools and studies. Indeed, genome data are often used as the foundation for different genomic and protein databases. The program BUSCO (Benchmarking Universal Single-Copy Orthologs) [30], used to measure the completeness of a genome assembly, relies on genomic data to build reference gene sets that are used for scoring. It uses hidden Markov models to detect orthologs that are shared by $\geq$90% of the species in a given clade. Thus, results from under-sampled groups could change drastically when more species are added to the gene sets. These could also have major effects in analyses such as phylogenomics, protein families studies and of gene duplication events. Another consequence of the current dearth of genomic resources for non-vertebrate animals is that BLAST [31] searches for these organisms most often recover vertebrate and arthropod hits, even though the target species is distant from these phyla, hampering the identification of sequences from a species lacking a reference or closely related genome.

It is therefore imperative to explore thoroughly the diversity of metazoans, specifically from non-vertebrate animal species. International consortia such as the Global Invertebrate Genomics Alliance (GIGA) [32, 33] have been put in place to overcome some of the aforementioned limitations. Other consortia such as the Earth BioGenome

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

Project [34], the Darwin Tree of Life [35], the Aquatic Symbiosis Genomics Project [36] and the European Reference Genome Atlas [37] are also expected to significantly boost the genomic resources of non-vertebrates in the near future. Undoubtedly, these projects will benefit from the drastic improvements in sequencing technologies over the last years.

Sequencing

Sequencing technologies have dramatically evolved over the last two decades, providing researchers with various options when it comes to tackling a genome project (Table 1). Sanger sequencing, the widely used sequencing method with chain-terminating inhibitors, published in 1977, produces reads around 1,000 basepair (bp) long with an error rate of about 1% [38]. The principle is to synthesize complementary strands of DNA from a single strand with a mixture of regular nucleotides and dideoxynucleotides, the latter stopping the polymerase when incorporated. Four reactions are performed for each type of base, and the resulting oligonucleotides are migrated by electrophoresis to identify the correct base at every position and generate a read. This method laid the foundations for DNA sequencing and was used extensively in several genome assembly projects, which were at that time typically ran by large international consortia: the budding yeast *Saccharomyces cerivisiae* [39] was the first eukaryote sequenced, whereas the nematode *Caenorhabditis elegans* was the first metazoan [40]. Sanger sequencing is a relatively low-throughput method in terms of the number of sequences generated, and is costly as well [41]. Although it is almost not used in genome projects anymore, the technology was pivotal for the generation of the first assembly of the human genome published in 2001, a monumental effort by 20 sequencing centers, to an estimated cost of 300 million US dollars [42].

Second-generation sequencing technologies, initially called next-generation sequencing (NGS), are characterized by a strong increase in sequencing throughputs compared to the Sanger method, with millions of DNA fragments sequenced simultaneously. NGS reads are much smaller than Sanger reads (from 110 bp for the first 454 machine in 2005 up to to 350 bp for MiSeq Illumina machines nowadays), resulting in the need for new analysis algorithms and programs [43]. The arrival of NGS sequencing democratized genome assembly projects, broadening the scope of investigated species beyond well-studied model organisms. Several second-generation sequencing methods have emerged through the years, some of which have since then been discontinued: 454 pyrosequencing [44], Ion Torrent [45], SOLiD [46], and Solexa (for a comparison on the approaches, see [47]). Among these methods, Solexa, subsequently purchased by Illumina [48], became and remains the most widely

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

used approach to this day. This approach consists in amplifying short DNA molecules bound on a flow cell, and sequencing them by sequential addition of fluorescently tagged nucleotides. This protocol generates highly accurate single or paired-end reads with a length up to a few hundred bases. The recent NovaSeq system further increased the output from a single run and abated the cost (up to 3 Terabases per flowcell). Short reads stimulated the whole field of genomics, and led to a large production of assemblies for all sorts of organisms, up to this day (Figure 1). These short-read based assemblies resulted in a tremendous increase of genomic resources, which remained typically quite fragmented (with N50s below 1 Megabase (Mb)).

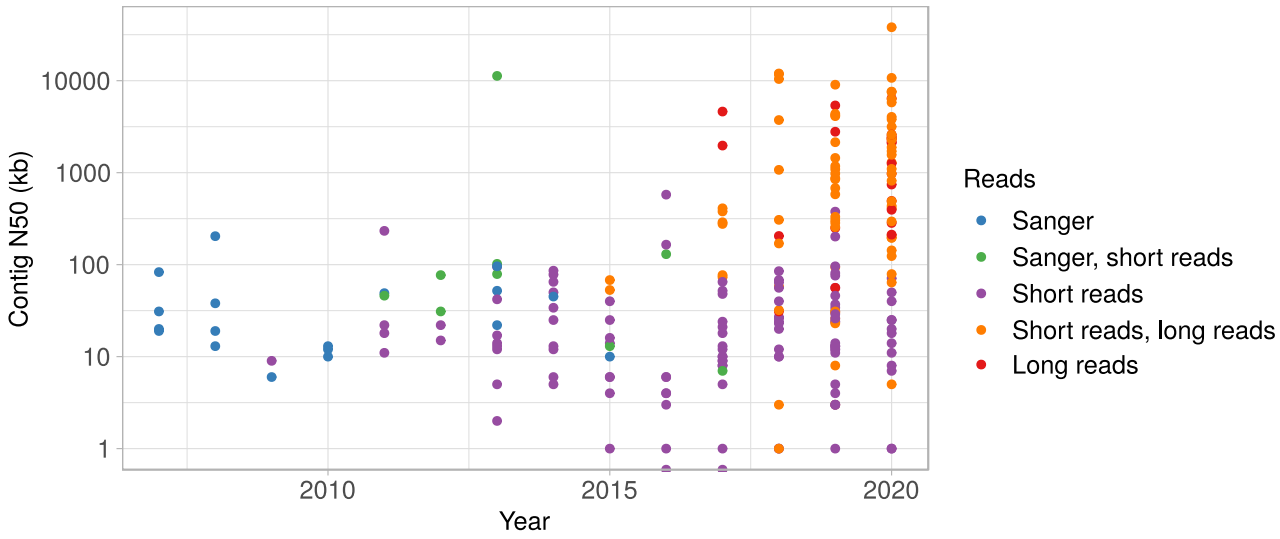

Figure 1: Contig N50 of 237 non-vertebrate animal genome assemblies over time. The N50 represents the contiguity of an assembly and is defined as the length of the largest contig for which at least 50% of the assembly size is contained in contigs equal or greater in length.

Third-generation sequencing has brought a whole new range of sequencing data, with the sequencing of long DNA molecules extending up to hundreds of thousands of bases [49]. The two main players in the field, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (Nanopore), use two different kinds of technologies. PacBio developed Single Molecule Real-Time (SMRT) sequencing, where a complementary strand of DNA is produced from a single strand by addition of fluorescently labeled nucleotides. The fluorescent tag is released and the luminescence is interpreted as a base [50]. The resulting reads have a length around twenty kilobases (kb) and a high error rate, an issue recently addressed by the introduction of an extra step called Circular Consensus Sequencing (CCS). In CCS, the DNA polymerase passes multiple times on the same base on a circularized strand to produce High Fidelity (HiFi) reads that can achieve an accuracy over 99% [51].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

Nanopore sequencing uses a membrane with protein pores, through which an electrical current is flowing. DNA strands are pulled through the pores, with each passing nucleotide generating a distinct disruption signature in the current that can be inferred as a specific base [52]. The firm has specifically oriented its strategy toward a "do it yourself" approach, enabling sequencing in any lab and even directly in the field via a small portable device [53]. Researchers can control how they generate their sequencing data, contribute to protocol development, and develop their own basecalling [54] to increase the yield and improve the quality and length of the reads. Although Nanopore reads still exhibit a high error rate, their length keeps increasing to attain hundreds of kilobases to 1 Mb [55]. The error rate has also been decreasing with the release of new flow cells and the development of more accurate basecallers such as Bonito [56].

Long reads are now routinely included in genome assembly projects and have led to much more contiguous assemblies than short-read only assemblies (Figure 1). A current limitation lies in the amount of DNA required to prepare long-read libraries, and long-read sequencing still remains inaccessible for certain species: whereas Illumina sequencing can handle small DNA amounts, with a poor quality, long-read protocols require high-molecular weight DNA [57]. PacBio and Nanopore sequencing remain difficult when one animal is too small to provide a sufficient amount of DNA, especially when the organism requires extraction protocols that lead to overly fragmented DNA (for example, with skeletons). In addition, secondary metabolites associated to DNA molecules, or branched DNA structures, can also disturb the sequencing reaction.

Genome assembly

A variety of programs have been developed to assemble sequencing reads *de novo*, taking advantage of different sequencing technologies while considering their limitations. Genome assembly aims to correctly reconstruct the original chromosome sequences from short or long, and accurate or error-prone fragments. Assemblers are typically based on one of the following paradigms: greedy, Overlap-Layout-Consensus, de Bruijn graphs.

The assembly problem can be represented as a linear puzzle where the pieces are the reads. Reads match together when they have overlapping sequences. This puzzle could be intuitively solved by iteratively putting together the overlapping pieces that match best: this greedy approach is an efficient heuristic to find the shortest common superstring of the set of reads (i.e., the shortest sequence that includes all the reads as substrings) [123]. Greedy algorithms have been implemented for first-generation sequencing reads, for instance in TIGR [69],

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

Table 1: Sequencing approaches and associated assemblers.

Sequencing	**Length**	**Accuracy**	**Methods**	**Assemblers**
First generation	1 kb	High	Sanger	ARACHNE [58], Atlas [59], CAP3 [60], Celera [61], Euler [62], JAZZ [63], Minimus [64], MIRA [65], phrap [66], Phusion [67], SUTTA [68], TIGR [69]
Second generation	25-300 bp	High	454, IonTorrent, Solexa, SOLiD	ABySS [70, 71], ALLPATHS [72], CABOG [73], BASE [74], Edena [75], Euler-SR [76], EPGA [77], Gossamer [78], IDBA [79], ISEA [80], JR-Assembler [81], LightAssembler [82], Meraculous [83], MIRA [65], Newbler [84], PCAP [85], PERGA [86], Platanus [87], PE-Assembler [88], QSRA [89], Ray [90], Readjoiner [91], SGA [92], SHARGCS [93], SOAPdenovo [94], SOAPdenovo2 [95], SPAdes [96], SparseAssembler [97], SSAKE [98], SUTTA [68], Taipan [99], VCAKE [100], Velvet [101]
Third generation	10-100.000+ kb	Low	PacBio CLR, Nanopore	Canu [102], FALCON [103], Flye [104], HINGE [105], MECAT [106], MECAT2 [106], miniasm [107], NECAT [108], NextDenovo [109], Ra [110], Raven [111], Shasta [112], SMARTdenovo [113], wtdbg [114], wtdbg2 [115]
	20 kb	High	PacBio HiFi	Flye [104], HiCanu [116], hifiasm [117], IPA [118], LJA [119], mdBG [120], MBG [121], Peregrine [122]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

and were further applied in short-read assemblers like PERGA [86], SSAKE [98] and VCAKE [100]. However, they cannot resolve complex, repetitive genomes: for this reason, greedy assemblers are mostly used nowadays to assemble small organelle genomes such as chloroplasts and mitochondria [124].

The Overlap-Layout-Consensus (OLC) paradigm was first described in 1979 by Rodger Staden [125] and is based on an overlap graph (Figure 2). The Overlap step consists in finding overlaps above a certain quality threshold between all the reads and building a directed graph, where the nodes are the reads and the edges represent the overlaps between them. The Layout step removes redundant edges that can be inferred from other edges. Finally, the Consensus step finds the shortest generalized Hamiltonian path through the graph, i.e. returns the shortest path (or set of disconnected paths) that visit each contig of the assembly at least once. The OLC paradigm has thrived with the program Celera [61], which was used to assemble a human genome from a Sanger shotgun dataset [126].

De Bruijn Graphs (DBGs) (Figure 3) are a well studied structure in graph theory, described by Nicolaas Govert de Bruijn in 1946 [127] and before him by Camille Flye Sainte-Marie [128]. DBG-based assemblers require highly accurate reads in which errors are only substitutions, with no indels. They start by indexing all the different sequences of a given k length (k-mers) found in the reads. In node-centric DBGs, the k-mers present in the reads are represented as nodes and are connected in the graph when they have an overlap of a k-1 length. In edge-centric DBGs, the k-mers present in the reads are represented as edges connecting their left and right (k-1)-mers. Once the graph is constructed, DBG assemblers look for a generalized Eulerian (in the case of edge-centric DBGs) or Hamiltonian (in the case of node-centric DBGs) path through the graph, i.e. returns the shortest path (or set of disconnected paths) that visits each k-mer of the assembly at least once. This approach was first used for genome assembly of first-generation sequencing datasets [129] and was quickly implemented in multiple popular short-read assemblers, e.g. ABySS [70, 71], IDBA [79], SOAPdenovo [94] and SOAPdenovo2 [95], SPAdes [96], Velvet [101].

With the advent of third-generation sequencing, OLC assemblers have benefited from a renewed interest whereas DBG-based ones are poorly suited for long, low-accuracy reads, containing many erroneous k-mers. Numerous assemblers have implemented the OLC approach to produce *de novo* assemblies from error-prone long-read datasets: Flye [104], Ra [110], Raven [111], Shasta [112], wtdbg2 [115]. Now that HiFi reads bring a new type of high-accuracy long reads, assemblers have been adapted to better handle these sequences, such as Flye (with adapted parameters), HiCanu [116] and hifiasm [117], and new DBG assemblers adapted for large

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

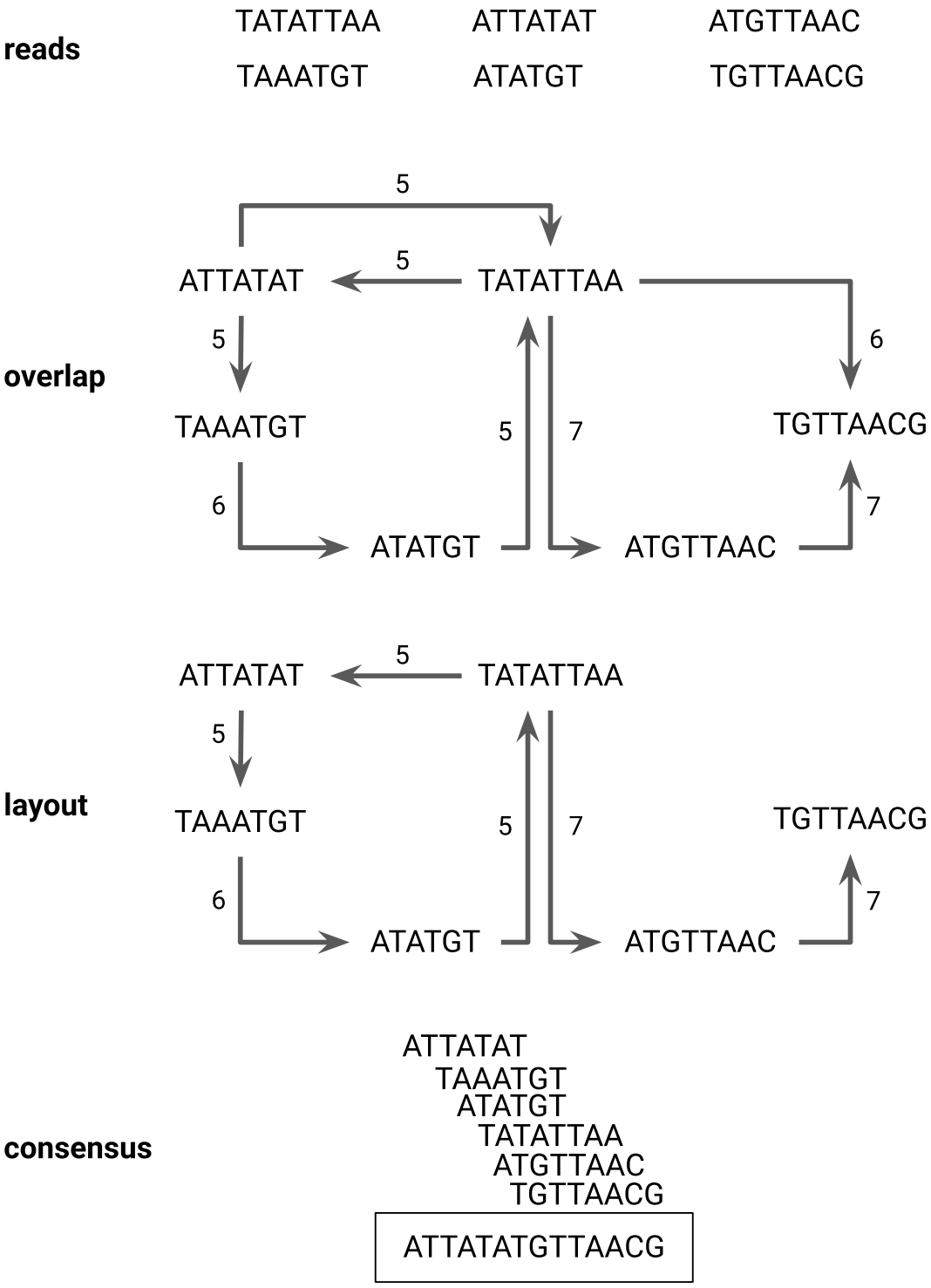

Figure 2: Overview of Overlap-Layout-Consensus assembly. The graph was built with all overlaps of at least 5 bases with a tolerance of 1 mismatch.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

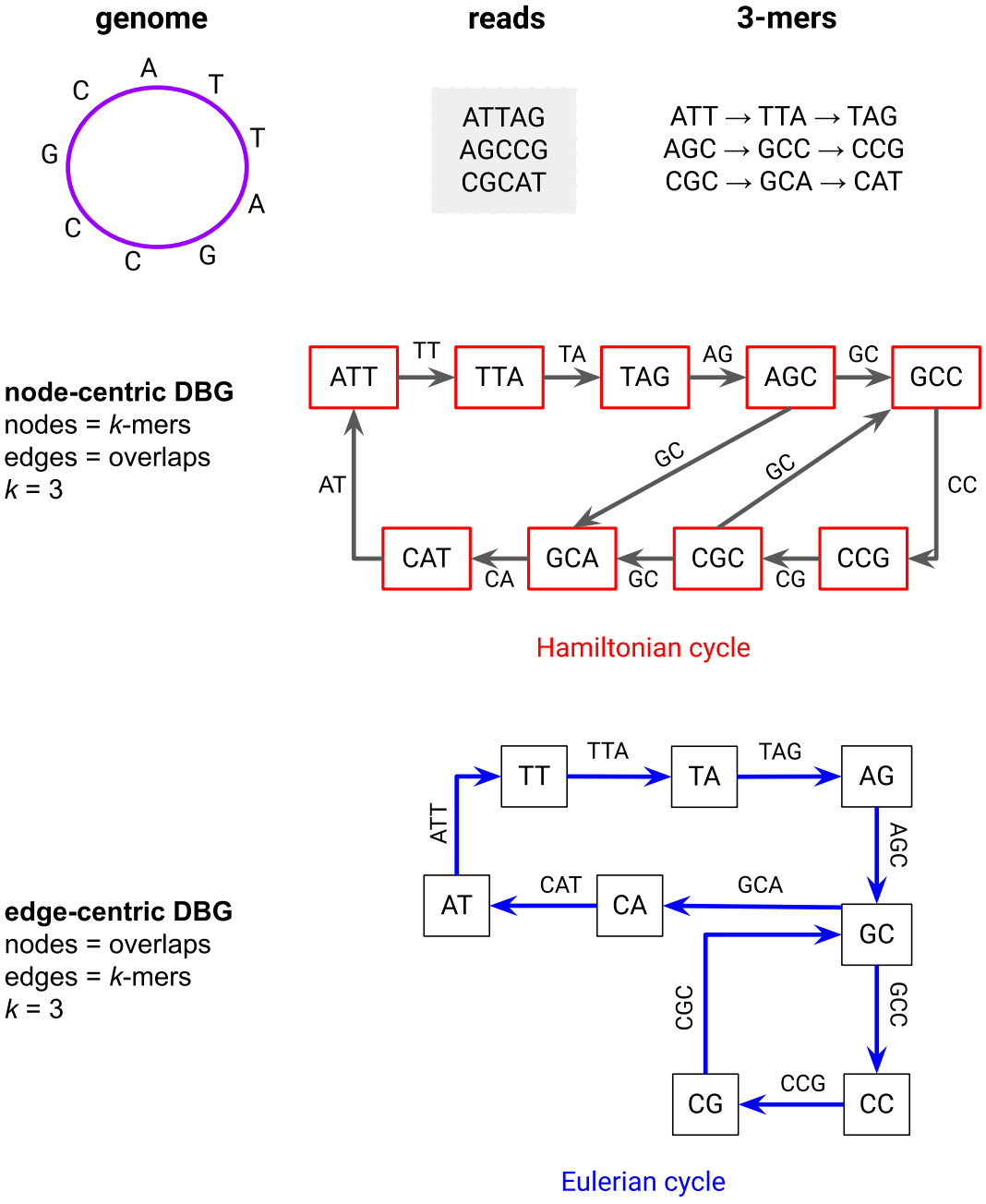

Figure 3: Overview of genome assembly using de Bruijn graphs. A circular genome is assembled based on three reads using node-centric and edge-centric DBGs with $k = 3$. The node-centric DBG is searched for a Hamiltonian cycle (visiting all nodes), and the edge-centric DBG for an Eulerian cycle (visiting all edges). These cycles are represented in blue in the graphs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

k-mer values are now being released [119, 121, 120].

From sequencing reads, assemblers build contiguous sequences called contigs. A perfectly assembled genome should have one contig representing each chromosome, but this is rarely achieved for eukaryotes. Assemblers need to find unambiguous paths in the assembly graph to reconstitute the chromosomes, but they often fail to do so due to the genomic structure: size, heterozygosity, repetitive content. Large genomes require a high amount of sequencing data in order to reach a sufficient depth to represent every locus. Genome sizes have a high variability (Figure 4): in the phylum Cnidaria, some myxozoans have a genome size of only some tens of Megabases (Mb) (*Kudoa iwatai*: 22.5 Mb, *Myxobolus squamalis*: 53.1 Mb, *Henneguya salminicola*: 60.0 Mb [130]), while the hydrozoan *Hydra oligactis* (1.3 Gigabases (Gb)) [131] has a genome size two orders of magnitude larger. Heterozygous regions constitute a major cause for breaks in assemblies of non-model animal genomes, as they generally have higher levels of heterozygosity than model species [132]. Most assemblers try to build a haploid representation of all genomes, even for multiploid (i.e. diploid or polyploid) genomes. To this end, heterozygous regions are collapsed in order to keep a single sequence for every region in the genome. In an assembly graph, these heterozygous regions will appear as bubbles, where one contig (a homozygous region) can be connected to several other contigs (the alternative haplotypes of a heterozygous region). When the assembler is unable to select one path, the homozygous region is not joined with any of the haplotypes, leading to a break in the assembly.

Assembly pre and post-processing

As obtaining high-quality chromosome-level contigs still remains challenging, upstream and downstream tools have been developed in conjunction with assemblers (Table 2). Researchers can test numerous combinations of these tools to devise the pipeline that will yield the best assembly (Figure 5).

Long reads have the advantage over short reads that they result in more contiguous assemblies. Nevertheless, assemblies of PacBio Continuous Long Reads (CLR) or Nanopore reads can have remaining errors due to their low accuracy; while errors in PacBio CLR are random and are compensated with a high coverage, Nanopore reads have systematic errors in homopolymeric regions. Assemblies of error-prone long reads often necessitate additional processes to increase the quality. There are two possible strategies: correct the long reads prior to assembly, and polish the contigs after assembly. Correcting long reads can be done using only the long reads or by adding high-accuracy short reads. Many tools have been developed for both scenarii and have been thoroughly

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

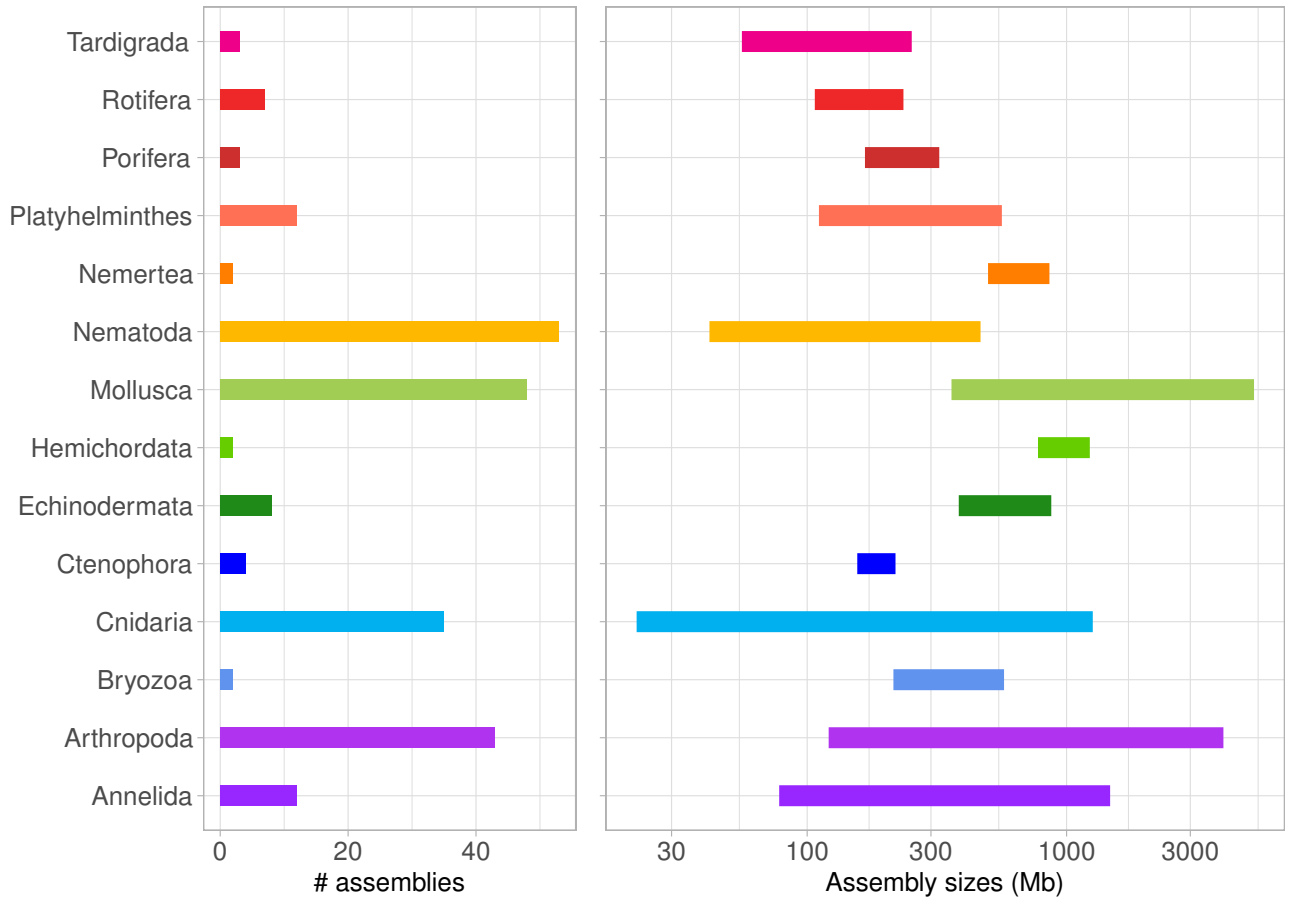

Figure 4: Assembly sizes. The left graph shows the number of assemblies included for each phylum and the right part shows the corresponding assembly-size ranges.

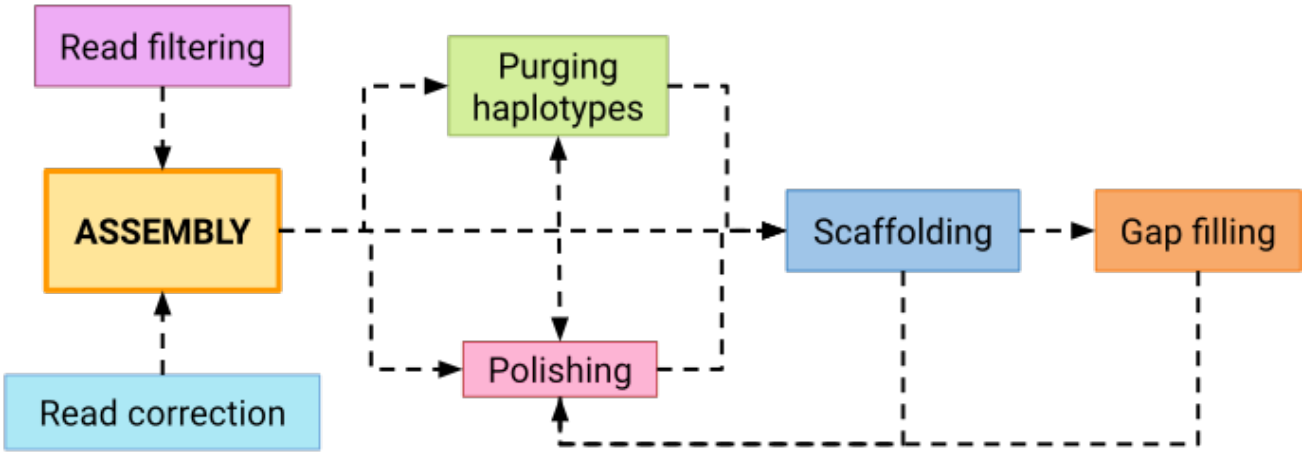

Figure 5: Assembly pipeline, including the assembly and the pre/post-processing steps.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

Table 2: Assembly pre and post-processing tools for haploid assemblies.

Step	**Sequencing data**	**Tools**
Reads filtering	Long reads	Filtlong [133]
Long reads error correction	Short reads	CoLoRMAP [134], Hercules [135], Jabba [136], LoRDEC [137], LoRMA [138], proovread [139]
	Long reads	Canu [102], CONSENT [140], Daccord [141], FLAS [142], HALC [143] NextDenovo [109], MECAT [106], MECAT2 [106], NECAT [108]
Polishing	Short reads	ntEdit [144], Pilon [145], POLCA [146]
	Short & long reads	Apollo [147], HyPo [148], Racon [149]
	Long reads	Arrow [150], CONSENT [140], Medaka [151], NextPolish [152], Nanopolish [153], Quiver [150]
Haplotigs purging	Long reads	HaploMerger2 [154], purge_dups [155], Purge Haplotigs [156]
Scaffolding	Short reads Mate pairs	Bambus [157], BATISCAF [158], BESST [159], BOSS [160], GRASS [161], MIP [162], Opera [163], ScaffMatch [164], ScaffoldScaffolder [165], SCARPA [166], SCOP [167], SLIQ [168], SOPRA [169], SSPACE [170], WiseScaffolder [171]
	Long reads	LINKS [172], LRScaf [173], npScarf [174], PBJelly [175], RAILS [176], SLR [177], SMIS [178], SMSC [179], SSPACE-LongRead [180]
	Genetic maps	ALLMAPS [181]
	Optical maps	AGORA [182], BiSCoT [183], OMGS [184], SewingMachine [185], SOMA [186]
	Linked reads	ARBitR [187], Architect [188], ARCS [189], ARKS [190], fragScaff [191], Scaff10X [192]
	3C/Hi-C	3D-DNA [12], dnaTri [193], GRAAL [194], HiCAssembler [195] instaGRAAL [196], Lachesis [197], SALSA [198], SALSA2 [199]
Gap filling	Short reads	GapFiller [200], GAPPadder [201], Sealer [202]
	Long reads	Cobbler [176], FGAP [203], GMcloser [204], LR_Gapcloser [205], PBJelly [175], PGcloser [206], TGS-GapCloser [207]

reviewed on multiple datasets [208]. When tested on *Caenorhabditis elegans* Nanopore reads, the error rate decreased from 28.93% to less than 1% (using Canu [102], CONSENT [140], FLAS [142], Jabba [136], LORMA [138] or MECAT [106]). Assembling corrected reads is expected to yield contigs with higher quality and contiguity. Alternatively, or additionally, the contigs can be polished to reduce errors, using long reads and/or short reads. Polishing can be a more computationally efficient strategy: the reads are mapped solely to the draft assembly, while correction is usually based on an all-versus-all read mapping.

Assemblers are generally tested on model-organism datasets, and are ill-suited for non-model genomes with variable levels of heterozygosity. They often fail to collapse highly divergent haplotypes, causing artefactually duplicated regions that hinder subsequent analyses [209]. Some long-read assemblers, Ra and wtdbg2, have been identified as less prone to retain uncollapsed haplotypes [210]. Contigs can also be post-processed to remove these duplications with dedicated tools such as HaploMerger2 [154], purge_dups [155] and Purge Haplotigs [156]. HaploMerger2 detects uncollapsed haplotypes based on sequence similarities, while purge_dups and Purge Haplotigs also rely on coverage depth.

To improve the contiguity of an assembly, contigs can be grouped, ordered and oriented into scaffolds. These scaffolds may contain gaps, when the sequence that should connect two contigs cannot be retrieved, represented as a sequence of Ns, and these gaps can be reduced post-scaffolding with gap-filling tools. Chromosome-level scaffolds have become a standard in genome assembly publications: unlike fragmented assemblies, they can be used for synteny analysis, finding rearrangements, and to separate chromosomes from different species. Scaffolding tools were already developed for first-generation sequencing reads (e.g. Celera [61], CAP3 [60], GigAssembler [211]). Since then, several sequencing techniques have been used to scaffold assemblies: mate pairs, long reads, genetic maps, optical mapping, linked reads, and proximity ligation [212]. Mate pairs are short reads with a large insert size (more than several kb), and have been widely used in next-generation assemblies. Among the 237 assemblies we surveyed, 78 included a mate-pair scaffolding step (Figure 6). Both genetic maps [213] and optical maps [214] provide information on the linkage and relative position of a set of markers, spread over the genome, thus they can be used to anchor contigs. Genetic maps were used for the genome assemblies of the flatworm *Schistosoma mansoni* [215], the copepod *Tigriopus japonicus* [216] and the coral *Acropora millepora* [217]. Although existing genetic maps provide precious resources, building one is particularly difficult as it requires breeding [213], making it hardly accessible for wild species, and impossible for asexual species. Markers of optical maps are motifs in the sequence that are labeled and detected by a fluorescent signal. Companies such as Bionano or Nabsys propose this service to scaffold assemblies [218], and this method was included in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

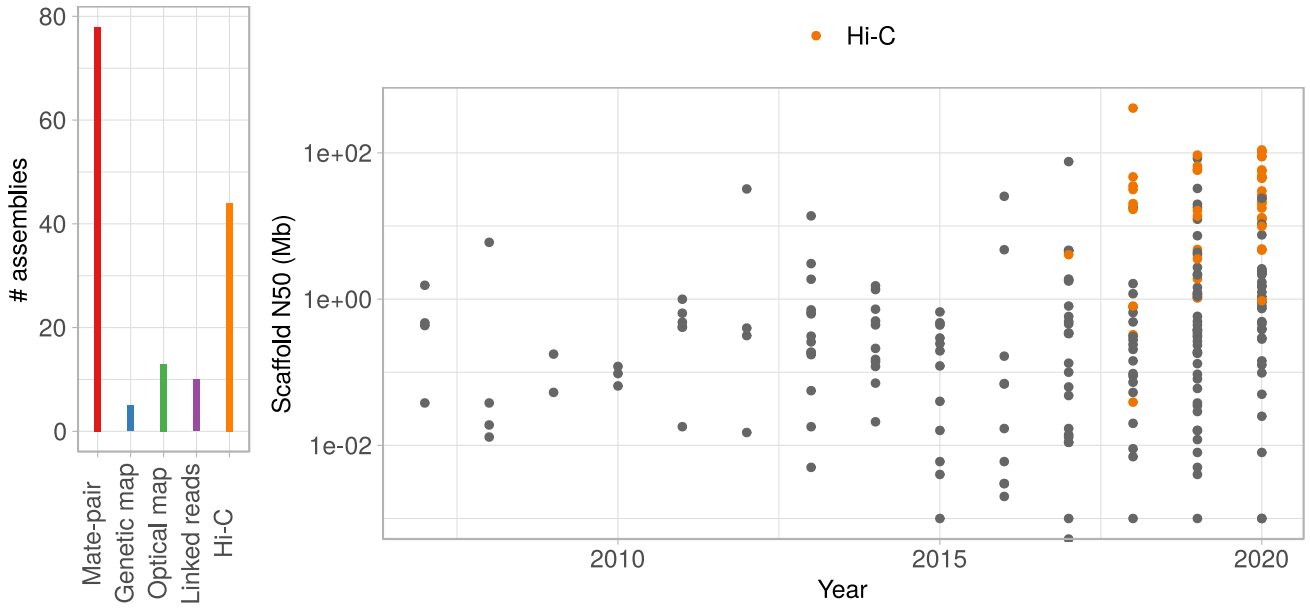

Figure 6: Assemblies scaffolding. Left: number of assemblies that included each scaffolding method. Right: scaffold N50 of non-vertebrate animal genome assemblies over time. The assemblies that included a Hi-C scaffolding step are highlighted in orange; they form a cluster with a scaffold N50 over 1 Mb.

some non-vertebrate genome projects: several nematodes including *Onchocerca volvulus* [219], *Ascaris suum* and *Parascaris univalens* [220], the tapeworms *Echinococcus multilocularis* [221] and *Hymenolepis microstoma* [222], and the chiton *Acanthopleura granulata* [223].

Linked reads and proximity ligation are based on short-read sequencing, preceded by a specific library preparation. For linked reads, also called cloud reads, long fragments of DNA are barcoded and then sequenced. The company 10X Genomics was a leader of this technology, but they chose to discontinue its commercialization in June 2020. Linked reads have been used to scaffold the genomes of the coral *Acropora millepora* [217] and the bee *Lasioglossum albipes* [224]. As linked reads are also shotgun Illumina reads, these reads are sometimes used for assembly (using Architect [188] or Supernova [225]) or polishing, as was done for the mosquito *Anopheles funestus* [226].

Proximity ligation techniques, based on capture of chromosome conformation [227], were not originally developed with genome sequencing applications in mind. Instead, they aimed at investigating the interplay between chromosome 3D organization and DNA processes [228]. A popular genomic derivative of 3C, Hi-C [229] documents the average conformation of the genomes of a population of cells. Briefly, the approach consists in freezing the chromosome folding of each individual cell using chemical fixation by formaldehyde, which generates bonds between proteins and proteins, and proteins and DNA. Then, the genome is cut into fragments using a restriction

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

enzyme, that are then ligated in dilute conditions. As a consequence, fragments that were trapped together by the crosslinking step are more prone to be ligated with each other, rather than with a fragment belonging to a different crosslinked complex. This results in chimeric fragments with respect to the original genome agencement, reflective of their 3D contacts *in vivo*. The relative proportions of ligation events between all restriction fragments of a genome can then be quantified, in theory, through high-throughput sequencing. On average, and because of the polymer nature and physical properties of DNA, the frequency of contacts between a pair of loci reflects either their 1D *cis* disposition along a chromosome, or their *trans* disposition on two independent chromosomes [230, 231]. Hi-C scaffolders have been developed following these principles: some follow a graph approach and use Hi-C links to join contigs (3D-DNA [12], SALSA2 [199]), whereas others exploit Markov Chain Monte Carlo (MCMC) sampling and Bayesian statistics to reorganize DNA segments into the scaffolds most likely to explain the observed interaction frequencies (GRAAL [194] and its later improved version instaGRAAL [196]). Hi-C reads are enriched around restriction sites, which makes them inadequate for *de novo* assembly and polishing. Recent protocols can now use Dnase I instead of restriction enzymes to yield libraries with a uniform distribution, such as Omni-C.

The Hi-C protocol itself is becoming more and more accessible as commercial kits are now available (e.g. Arima Hi-C, Phase Genomics, or Dovetails Genomics). Hi-C scaffolding proved efficient at bringing highly fragmented draft assemblies to chromosome-level scaffolds (Figure 6), and is now included in many genome projects for all sorts of non-vertebrate animals: the arthropods *Varroa destructor* [232] and *Carcinoscorpius rotundicauda* [233], the cnidarians *Xenia* sp. [234] and *Rhopilema esculentum* [235], the echinoderms *Lytechinus variegatus* [236] and *Pisaster ochraceus* [237], the molluscs *Scapharca broughtonii* [238] and *Chrysomallon squamiferum* [239], the nematods *Caenorhabditis remanei* [240] and *Heterodera glycines* [241], the platyhelminthe *Schistosoma haemotabium* [242], the poriferan *Ephydatia muelleri* [243], the rotifer *Adineta vaga* [244], the xenacoelomorph *Hofstenia miamia* [245], and more. A compelling advantage of Hi-C scaffolding over other scaffolding methods is its ability to discriminate different organisms in a draft assembly: DNA from different organisms belong to distinct nuclei, thus they have no 3D interactions. This feature is especially useful for non-vertebrate animals with symbionts, that can hardly be eliminated from the host prior to sequencing, and are often targets for genome assembly as well.

Assembly and pre/post-processing steps are often combined in one tool. Canu, MECAT, MECAT2, NECAT and NextDenovo correct low-accuracy long reads prior to assembly; Flye, Raven and NextDenovo have a polishing step; and many assemblers include a scaffolding step to yield both contigs and scaffolds. Some assemblers

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

propose a hybrid assembly strategy, using both short and long reads, such as HALSR [246], MaSuRCA [247] and WENGAN [248].

Assembly evaluation

A critical step in genome assembly is to estimate the quality of draft assemblies, and choose the best one for subsequent analysis. The first metric to assess is the assembly size and its adequacy with an estimated genome size. The size can be estimated experimentally with flow cytometry or Feulgen densitometry [249], but these methods require a reference species for which the genome size is already well known, exposing them to errors induced by the reference genome size. Reference-free genome size estimation tools are typically k-mer based approaches and use high-accuracy reads (e.g. Illumina, PacBio HiFi). These tools, such as BBtools [250], GenomeScope [251] and KAT [252], build a k-mer spectrum representing the number of k-mers with a certain frequency of occurence. When the sequencing depth is sufficient, the k-mer spectrum should display one or more peaks depending on the ploidy. For a haploid organism, there should be only one peak, whereas a diploid organism should have two peaks. The plot may also show a peak of k-mers with a frequency of occurence close to zero, corresponding to erroneous k-mers. Another recent tool called MGSE [253] estimates genome size based on reads mapping to a highly continuous assembly of the same genome; this method can be used as a post-hoc analysis.

N50 is a popular metric that reflects the contiguity of an assembly: it is defined as the length of the largest contig for which 50% of the assembly size is contained in contigs of equal or greater length. Some tools provide in addition the N75, N90, N99, computed in a similar fashion. The NG50 is a variant of N50 that refers to an estimated genome size instead of the assembly size. The target assembly can further be mapped against a reference assembly to detect misassemblies and break them: the N50 and NG50 of the resulting fragments are called NA50 and NGA50. All these metrics can be computed using QUAST [254]. For genome assemblies of non-model non-vertebrate animals, reference assemblies are seldom available, or they have a poor quality or contiguity that the new assembly aspires to improve. Therefore we will focus on reference-free evaluation methods. Table 3 and Figure 8 present an example of assembly evaluation for the recently published snail *Achatina fulica* [255] and the coral *Xenia* sp. [234].

Another feature to optimize is the completeness of the genome, usually based on orthologs or k-mers. BUSCO [30] searches for orthologs in a user-provided lineage; the current Metazoa lineage (designated as Metazoa odb10)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

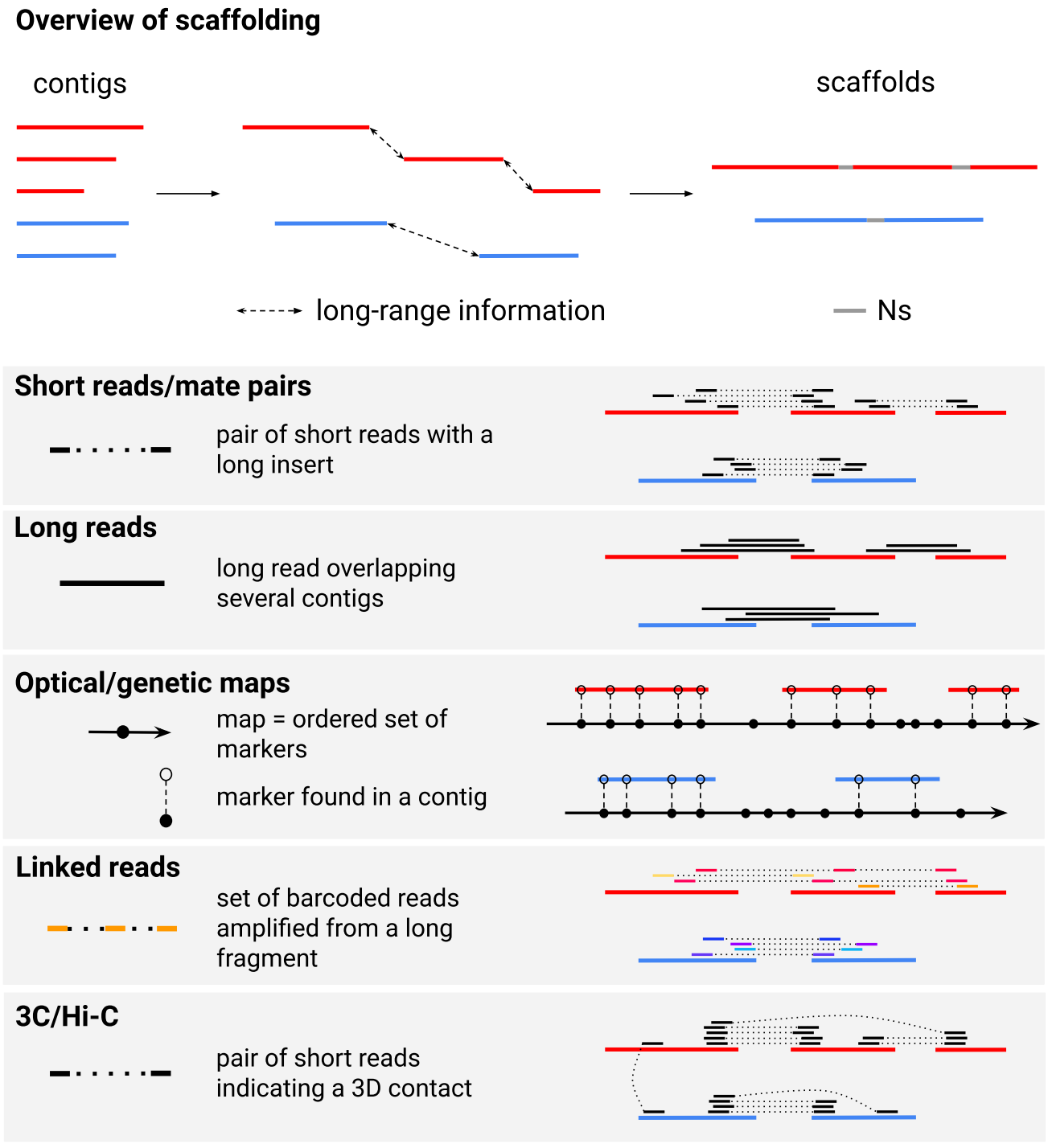

Figure 7: Overview of scaffolding methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

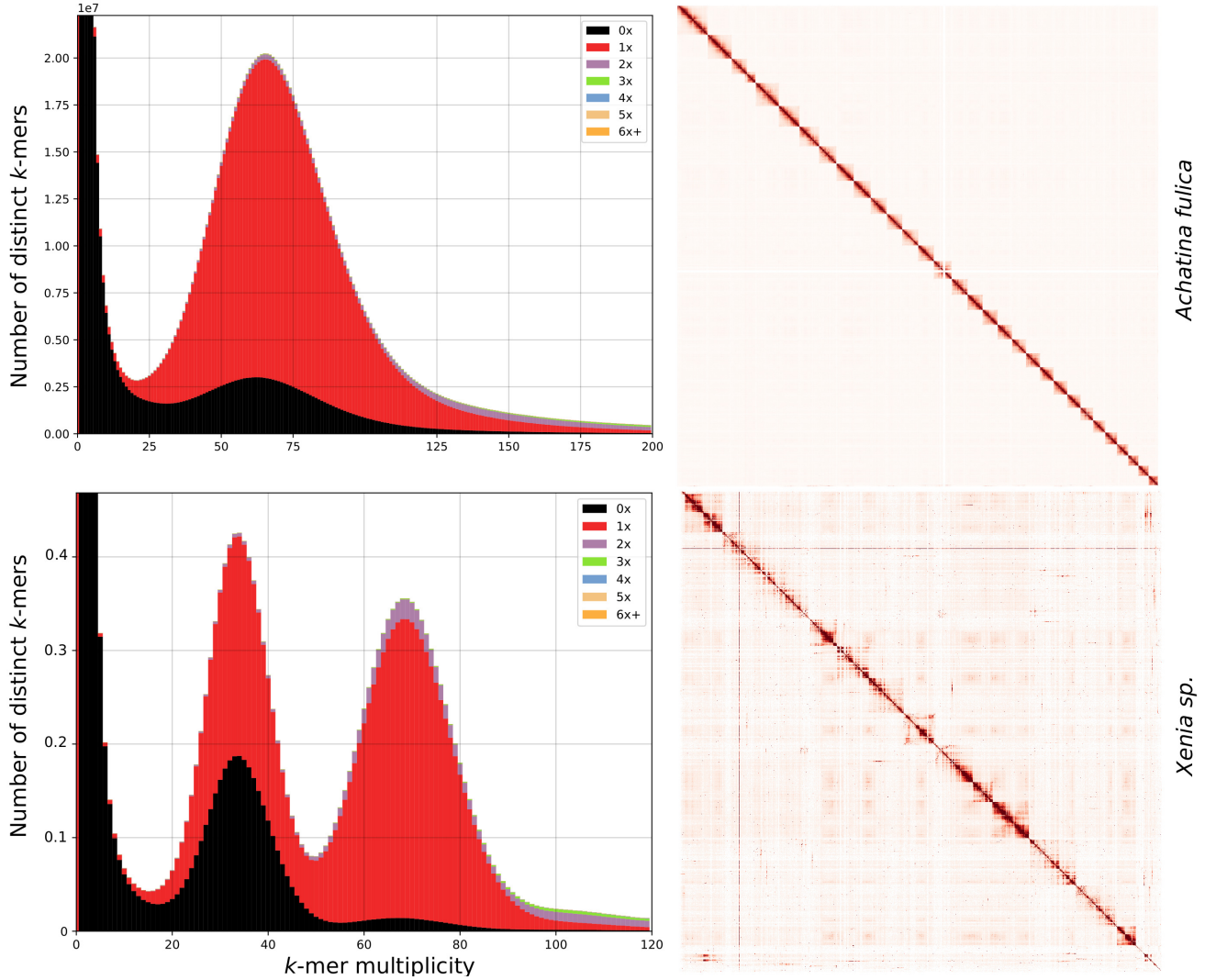

Figure 8: Assembly evaluation of *Achatina fulica* and *Xenia* sp.. Left: KAT comparison of the *k*-mers in the Illumina datasets v. the assembly. Right: Hi-C contact maps, with a binning of 300 for *Achatina fulica*, 30 for *Xenia* sp..

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

Table 3: Assembly evaluation of *Achatina fulica* and *Xenia* sp..

		Achatina fulica	*Xenia* sp.
Basic statistics	Assembly size	1.86 Gb	222.7 Mb
	N50	59.6 Mb	14.8 Mb
	N90	44.1 Mb	6.9 Mb
	Largest scaffold	116.6 Mb	22.5 Mb
	Number of scaffolds	1500	168
	Number of scaffolds larger than 1 Mb	32	17
	N count	3,600,500	194,000
BUSCO completeness	Complete and single-copy BUSCOs	84.4%	86.0%
	Complete and duplicated BUSCOs	3.6%	2.2%
	Fragmented BUSCOs	3.5%	3.5%
	Missing BUSCOs	8.5%	8.3%
Reads mapping	Short reads	96.2%	87.8%
	Long reads	81.62%	99.5%
	Hi-C	70.2%	65.7%

contains 954 features. Assemblies are evaluated based on the proportion of orthologs to these 954 genes that can be retrieved into them; yet, some features are systematically missing in some genomes as they are absent from these species. More specific lineages are available for arthropods, insects, nematodes, vertebrates, mammals, as many assemblies are available for these groups, but other metazoan phyla suffer from their lack of resources. Consequently, BUSCO is most powerful when comparing several draft assemblies for one genome. BUSCO scores provide information on complete single-copy and duplicated features, and the latter can be used to detect improperly duplicated regions in a haploid assembly. However, BUSCO scores are limited to genomic regions and cannot report for non-coding ones.

k-mer completeness scores do not present such limitations: KAT assesses the completeness of a whole assembly based on its representation of *k*-mers from a high-accuracy sequencing dataset. The *k*-mer spectrum should display one or several peaks depending on the ploidy of the genome: one peak for a haploid genome; two peaks for a diploid genome, the first depicting heterozygous *k*-mers, and the second for homozygous *k*-mers. Depending on the ploidy of the genome, every *k*-mer should be represented in the assembly as many times as they actually are in the genome.

Both *Achatina fulica* and *Xenia* sp. have high BUSCO scores (against the lineage Metazoa odb10), yet slightly below 90%, and they have few duplicated BUSCO features. The *k*-mer spectrum of *Achatina fulica* only shows one peak around 70X (Figure 8, top left). These *k*-mers are expected to be represented exactly once, which is the case for the majority; there are almost no *k*-mers that appear twice in the assembly (in purple), but there is a noteworthy amount of missing *k*-mers (in black). For *Xenia* sp., the *k*-mer spectrum has two peaks

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

with a k-mer multiplicity around 35X and 70X (Figure 8, bottem left). The first peak, representing heterozygous k-mers, shows that a portion is represented once in the assembly, while the rest is missing, as expected in a collapsed assembly. The second peak, for homozygous k-mers has a majority of k-mers represented once, and some k-mers either absent or duplicated. These assemblies seem overall properly collapsed and complete.

KAD, for k-mer abundance difference [256], proposes an alternative k-mer-based evaluation. This tool does not compute an overall completeness score, but instead classifies k-mers based on their abundance in the assembly and the sequencing dataset: good k-mers, erroneous k-mers (absent from the dataset), overrepresented k-mers (duplications), and underrepresented k-mers (collapsed repetitions).

Assemblies need to be screened for contaminants, to tell apart the sequences coming from the target and from other species. Contaminants may originate from the environment, the symbiont, or be artificially introduced by the sequencing process. Blobtools [257] and BlobToolKit [258] aim to identify them with GC content, coverage depth and taxonomy assignment using the NCBI TaxID. Discriminating bacteria in metazoan assemblies is usually straightforward based on their distinct GC percentage. The task is more challenging when the target metazoan genome is mixed with other eukaryotes or even metazoans, especially when these species are absent from databases. Chromosome-level assemblies reduce the risk of contamination, as downstream analysis can be run exclusively on sequences that were anchored to the main scaffold. In addition, with Hi-C data, sequences from different species can be separated based on their absence of *trans* interactions. Contamination can lead to false conclusions: for instance, a study on a highly fragmented genome assembly (N50 = 16 kb) of the tardigrade *Hypsibius dujardini* assumed that about 17% of its genome derived from horizontal gene transfers [259], when these sequences were in fact contaminants [260].

When Hi-C data are available, contact maps, i.e. the representation of the paired-end reads from the Hi-C library aligned on the resulting scaffold, procure another evaluation asset to search for misassemblies. The contact map is expected to show heightened frequencies for each chromosome, in a chromosome-level assembly, and these interaction frequencies should decrease with increased distances separating loci on the sequence, based on the distance law. For *Achatina fulica*, 30 chromosome-level scaffolds (out of 31) display relatively consistent and regular contact patterns, representing well individualized entities in the contact map (Figure 8, top right). By contrast, the contact map of *Xenia* sp. does not display such patterns, with multiple *trans* contacts appearing between the scaffolds and most likely corresponding to scaffolding errors.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 November 2021 doi:10.20944/preprints202111.0170.v1

Phasing assemblies

As collapsing multiploid genomes can be difficult for highly divergent regions and frequently causes breaks in the assembly, an intuitive solution would be to phase genomes to retrieve all haplotypes. Phased assemblies represent a whole different challenge as they necessitate to correctly associate alleles, i.e. different versions of a heterozygous region [261]. A first approach, called trio-binning, is to assemble one individual using sequencing data from the individual itself and its parents [262]; yet this method is only adapted when the parents can be identified, and is inapplicable on asexual species. Some tools are able to reconstruct haplotypes from collapsed assemblies using long reads, namely HapCUT2 [263] and WhatsHap [264]. Ideally, genomes should be uncollapsed, as can be done with Bwise [265] and Platanus-Allee [266] using short reads, FALCON-Unzip [103] using PacBio CLR or HiFi. FALCON-Unzip uses the output from the FALCON assembler, that includes both a haploid assembly and alternative haplotigs for heterozygous regions, to associate haplotypes based on long reads. Phased assemblies of low-accuracy long reads are limited, as small heterozygous regions were confused with errors; this led to haplotypes being erroneously collapsed.

HiFi reads have made a disruption in the fields of genomics: they are especially well-suited for phased assemblies thanks to their length and low error rate, and they have already been used to produce phased assemblies of a human [267] and the potato *Solanum tuberosum* [268]. Nevertheless, sequencing HiFi reads can remain inaccessible for non-model organisms as pure DNA is necessary.

Many organisms have already been assembled using low-accuracy long reads and high-accuracy short reads, thus an alternative is to correct long reads with short reads using a tool that conserves haplotypes such as Ratatosk [269]. Phased long-read assemblies can be further polished with adequate programs (e.g. Hapo-G [270]). As Hi-C has already demonstrated its efficiency to scaffold haploid assemblies, the principles were further exploited in ALLHiC [271], GraphUnzip [272] and FALCON-Phase [273] to phase assemblies while increasing their contiguity: as alleles from one haplotype belong to one chromosome, these alleles have higher Hi-C interaction frequencies together than with alleles from alternative haplotypes.

Phasing-specific evaluation methods are still scarce, and publications of phased assembly rely on various datasets to prove their correctness (e.g. parental assemblies [267]). Merqury [274] proposes a k-mer-based approach, inspired by KAT, and computes plots and scores to assess phasing completeness and find haplotype switches. However, similarly to trio-binning, it requires parental data.

Building robust animal genomic databases

We surveyed genome assembly papers from diverse metazoan phyla. Figures 1, 4 and 6 only retained assemblies that were available on GenBank, as we used assembly sizes, contig N50s and scaffold N50s from this source. We also limited these assemblies to those published after the year 2007, as we found that assemblies were seldom available on GenBank before that, and up to the year 2020. Some genomes were not deposited, and were instead available on a personal/lab/university page. This impedes meta-analyses and we are unable to accurately estimate the number of published non-vertebrate animal genome assemblies. The datasets used for the genome assemblies also suffer from this issue, as they are not necessarily publicly available. Efforts are being made to make genome assemblies and datasets findable, accessible, interoperable and reusable (FAIR) [275]. Assembly pipelines are becoming more reproducible thanks to several initiatives using workflow managers, such as the Vertebrate Genome Project assembly pipeline in Galaxy [276].

There were several inconsistencies in genome assembly statistics between the published paper and the assemblies available in the databases. In some cases, the differences were of a few kilobases, generally for the N50. The combination of cheaper sequencing methods, high-accuracy long reads and dynamic consortia have built a momentum in genome assembly promising to escalate the number of assemblies available, and genomic databases should be improved in parallel to better document assembly statistics and strategies. Exhaustive databases with reads, contig-level and scaffold-level assemblies, and also a list of tools used for assembly, could be used to conduct large analyses of these genomes and report on the performance of assembly tools.

Acknowledgements

This project was funded by the Horizon 2020 research and innovation program of the European Union under the Marie Skłodowska-Curie grant agreement No 764840 (ITN IGNITE, www.itn-ignite.eu).

References

[1] Rice, E. S. & Green, R. E. New approaches for genome assembly and scaffolding. *Annual review of animal biosciences* **7**, 17–40 (2019).

[2] National Center for Biotechnology Information. GenBank, `https://www.ncbi.nlm.nih.gov/genbank/` (2021).

[3] International Union for Conservation of Nature. Red List, `www.iucnredlist.org/resources/summary-statistics` (Accessed on May 4th, 2021).

[4] Hedges, S. B. & Kumar, S. *The timetree of life* (OUP Oxford, 2009).

[5] Zhang, Z.-Q. Animal biodiversity: An update of classification and diversity in 2013. In *Animal Biodiversity: An Outline of Higher-level Classification and Survey of Taxonomic Richness (Addenda 2013)*, vol. 3703, 1–82 (Magnolia Press, 2013).

[6] Li, F. *et al.* Insect genomes: progress and challenges. *Insect Molecular Biology* **28**, 739–758 (2019).

[7] Noriega, J. A. *et al.* Research trends in ecosystem services provided by insects. *Basic and Applied Ecology* **26**, 8–23 (2018).

[8] Chen, W. *et al.* The draft genome of whitefly *Bemisia tabaci* MEAM1, a global crop pest, provides novel insights into virus transmission, host adaptation, and insecticide resistance. *BMC Biology* **14**, 1–15 (2016).

[9] Nene, V. *et al.* Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* **316**, 1718–1723 (2007).

[10] Marinotti, O. *et al.* The genome of *Anopheles darlingi*, the main neotropical malaria vector. *Nucleic Acids Research* **41**, 7387–7400 (2013).

[11] Matthews, B. J. *et al.* Improved reference genome of *Aedes aegypti* informs arbovirus vector control. *Nature* **563**, 501–507 (2018).

[12] Dudchenko, O. *et al. De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).

[13] Carroll, A. R., Copp, B. R., Davis, R. A., Keyzers, R. A. & Prinsep, M. R. Marine natural products. *Natural Product Reports* (2021).

[14] Khalifa, S. A. *et al.* Marine natural products: A source of novel anticancer drugs. *Marine Drugs* **17**, 491 (2019).

[15] Ng, T. B., Cheung, R. C. F., Wong, J. H., Bekhit, A. A. & Bekhit, A. E.-D. Antibacterial products of marine organisms. *Applied Microbiology and Biotechnology* **99**, 4145–4173 (2015).

[16] Avila, C. Terpenoids in marine heterobranch molluscs. *Marine Drugs* **18**, 162 (2020).

[17] Han, B.-N. *et al.* Natural products from sponges. In *Symbiotic Microbiomes of Coral Reefs Sponges and Corals*, 329–463 (Springer Netherlands, 2019).

[18] Takeuchi, T. Molluscan genomics: implications for biology and aquaculture. *Current Molecular Biology Reports* **3**, 297–305 (2017).

[19] Prather, C. M. *et al.* Invertebrates, ecosystem services and climate change. *Biological Reviews* **88**, 327–348 (2013).

[20] Gomes-dos Santos, A., Lopes-Lima, M., Castro, L. F. C. & Froufe, E. Molluscan genomics: the road so far and the way forward. *Hydrobiologia* **847**, 1705–1726 (2020).

[21] Conci, N., Vargas, S. & Wörheide, G. The biology and evolution of calcite and aragonite mineralization in octocorallia. *Frontiers in Ecology and Evolution* **9**, 81 (2021).

[22] Clark, M. S. Molecular mechanisms of biomineralization in marine invertebrates. *Journal of Experimental Biology* **223** (2020).

[23] Shinzato, C. *et al.* Using the *Acropora digitifera* genome to understand coral responses to environmental change. *Nature* **476**, 320–323 (2011).

[24] Mao, Y., Economo, E. P. & Satoh, N. The roles of introgression and climate change in the rise to dominance of Acropora corals. *Current Biology* **28**, 3373–3382 (2018).

[25] Fuller, Z. L. *et al.* Population genetics of the coral *Acropora millepora*: Toward genomic prediction of bleaching. *Science* **369** (2020).

[26] Shinzato, C. *et al.* Eighteen coral genomes reveal the evolutionary origin of Acropora strategies to accommodate environmental changes. *Molecular Biology and Evolution* **38**, 16–30 (2021).

[27] Kapli, P., Yang, Z. & Telford, M. J. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics* **21**, 428–444 (2020).

[28] Chang, E. S. *et al.* Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proceedings of the National Academy of Sciences* **112**, 14912–14917 (2015).

[29] Eitel, M. *et al.* Comparative genomics and the nature of placozoan species. *PLoS Biology* **16**, 1–36 (2018).

[30] Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

[31] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

[32] GIGA Community of Scientists. The Global Invertebrate Genomics Alliance (GIGA): developing community resources to study diverse invertebrate genomes. *Journal of Heredity* **105**, 1–18 (2014).

[33] Voolstra, C. R., of Scientists (COS), G. C., Wörheide, G. & Lopez, J. V. Advancing genomics through the Global Invertebrate Genomics Alliance (GIGA). *Invertebrate Systematics* **31**, 1 (2017).

[34] Lewin, H. A. *et al.* Earth biogenome project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences* **115**, 4325–4333 (2018).

[35] Darwin Tree of Life. Darwin Tree of Life, `www.darwintreeoflife.org` (2021).

[36] Aquatic Symbiosis Genomics Project. Aquatic Symbiosis Genomics Project, `www.sanger.ac.uk/collaboration/aquatic-symbiosis-genomics-project` (2021).

[37] European Reference Genome Atlas. European Reference Genome Atlas, `www.erga-biodiversity.eu` (2021).

[38] Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* (1977).

[39] Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546–567 (1996).

[40] *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).

[41] Wajid, B., Sohail, M. U., Ekti, A. R. & Serpedin, E. The A, C, G, and T of genome assembly. *BioMed Research International* **2016**, 1–10 (2016).

[42] International Human Genome Sequencing Consortium and others. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

[43] Pop, M. & Salzberg, S. L. Bioinformatics challenges of new sequencing technology. *Trends in Genetics* **24**, 142–149 (2008).

[44] Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).

[45] Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).

[46] McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research* **19**, 1527–1541 (2009).

[47] Metzker, M. L. Sequencing technologies — the next generation. *Nature Reviews Genetics* **11**, 31–46 (2010).

[48] Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).

[49] Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T. & Sandhu, M. S. Long reads: their purpose and place. *Human Molecular Genetics* **27**, R234–R241 (2018).

[50] Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).

[51] Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* **37** (2019).

[52] Deamer, D., Akeson, M. & Branton, D. Three decades of Nanopore sequencing. *Nature Biotechnology* **34**, 518–524 (2016).

[53] Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* **17**, 1–11 (2016).

[54] Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology* **20**, 1–10 (2019).

[55] Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**, 338–345 (2018).

[56] Bonito, `https://github.com/nanoporetech/bonito`.

[57] Angthong, P. *et al.* Optimization of high molecular weight DNA extraction methods in shrimp for a long-read sequencing platform. *PeerJ* **8**, e10340 (2020).

[58] Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Research* **12**, 177–189 (2002).

[59] Havlak, P. *et al.* The Atlas genome assembly system. *Genome Research* **14**, 721–732 (2004).

[60] Huang, X. & Madan, A. CAP3: a DNA sequence assembly program. *Genome Research* **9**, 868–877 (1999).

[61] Denisov, G. *et al.* Consensus generation and variant detection by Celera Assembler. *Bioinformatics* **24**, 1035–1040 (2008).

[62] Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences* **98**, 9748–9753 (2001).

[63] Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).

[64] Sommer, D. D., Delcher, A. L., Salzberg, S. L. & Pop, M. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **8**, 1–11 (2007).

[65] Chevreux, B., Wetter, T., Suhai, S. *et al.* Genome sequence assembly using trace signals and additional sequence information. In *German Conference on Bioinformatics*, vol. 99, 45–56 (Citeseer, 1999).

[66] Ewing, B. & Green, P. Base-calling of automated sequencer traces using Phred. II. error probabilities. *Genome Research* **8**, 186–194 (1998).

[67] Mullikin, J. C. & Ning, Z. The Phusion assembler. *Genome Research* **13**, 81–90 (2003).

[68] Narzisi, G. & Mishra, B. Scoring-and-unfolding trimmed tree assembler: concepts, constructs and comparisons. *Bioinformatics* **27**, 153–160 (2010).

[69] Sutton, Granger and White, Owen and Adams, Mark D. and Kerlavage, A. R. TIGR Assembler: A new tool for assembling large shotgun projects. *Genome Science and Technology* **1**, 9–19 (1995).

[70] Simpson, J. T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome Research* **19**, 1117–1123 (2009).

[71] Jackman, S. D. *et al.* ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter effect of Bloom filter false positive rate. *Genome Research* **27**, 768–777 (2017).

[72] Butler, J. *et al.* ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Research* **18**, 810–820 (2008).

[73] Miller, J. R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824 (2008).

[74] Liu, B. *et al.* BASE: a practical *de novo* assembler for large genomes using long NGS reads. *BMC Genomics* **17**, 561–569 (2016).

[75] Hernandez, D., François, P., Farinelli, L., Østerås, M. & Schrenzel, J. *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research* **18**, 802–809 (2008).

[76] Chaisson, M. J. & Pevzner, P. A. Short read fragment assembly of bacterial genomes. *Genome Research* **18**, 324–330 (2008).

[77] Luo, J. *et al.* EPGA: *de novo* assembly using the distributions of reads and insert size. *Bioinformatics* **31**, 825–833 (2015).

[78] Conway, T., Wazny, J., Bromage, A., Zobel, J. & Beresford-Smith, B. Gossamer — a resource-efficient *de novo* assembler. *Bioinformatics* **28**, 1937–1938 (2012).

[79] Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA - a practical iterative De Bruijn graph *de novo* assembler. *Research in Computational Molecular Biology* **6044 LNBI**, 426–440 (2010).

[80] Li, M. *et al.* ISEA: Iterative seed-extension algorithm for *de novo* assembly using paired-end information and insert size distribution. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **14**, 916–925 (2016).

[81] Chu, T.-C. *et al.* Assembler for *de novo* assembly of large genomes. *Proceedings of the National Academy of Sciences* **110**, E3417–E3424 (2013).

[82] El-Metwally, S., Zakaria, M. & Hamza, T. LightAssembler: fast and memory-efficient assembly algorithm for high-throughput sequencing reads. *Bioinformatics* **32**, 3215–3223 (2016).

[83] Chapman, J. A. *et al.* Meraculous: *de novo* genome assembly with short paired-end reads. *PloS One* **6**, e23501 (2011).

[84] University of Arizona. Newbler, `https://cals.arizona.edu/swes/maier_lab/kartchner/documentation/index.php/home/docs/newbler` (2012).

[85] Huang, X., Wang, J., Aluru, S., Yang, S.-P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Research* **13**, 2164–2170 (2003).

[86] Zhu, X. *et al.* PERGA: a paired-end read guided *de novo* assembler for extending contigs using SVM and look ahead approach. *PloS One* **9**, e114253 (2014).

[87] Kajitani, R. *et al.* Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research* **24**, 1384–1395 (2014).

[88] Ariyaratne, P. N. & Sung, W.-K. PE-Assembler: *de novo* assembler using short paired-end reads. *Bioinformatics* **27**, 167–174 (2011).

[89] Bryant, D. W., Wong, W.-K. & Mockler, T. C. QSRA – a quality-value guided *de novo* short read assembler. *BMC Bioinformatics* **10**, 1–6 (2009).

[90] Boisvert, S., Laviolette, F. & Corbeil, J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology* **17**, 1519–1533 (2010).

[91] Gonnella, G. & Kurtz, S. Readjoiner: a fast and memory efficient string graph-based sequence assembler. *BMC Bioinformatics* **13**, 1–19 (2012).

[92] Simpson, J. T. & Durbin, R. Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Research* **22**, 549–556 (2012).

[93] Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing. *Genome Research* **17**, 1697–1706 (2007).

[94] Li, R. *et al.* *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20**, 265–272 (2010).

[95] Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* **1**, 2047–217X (2012).

[96] Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**, 455–477 (2012).

[97] Ye, C., Ma, Z. S., Cannon, C. H., Pop, M. & Douglas, W. Y. Exploiting sparseness in *de novo* genome assembly. In *BMC Bioinformatics*, vol. 13 (BioMed Central, 2012).

[98] Warren, R. L., Sutton, G. G., Jones, S. J. & Holt, R. A. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23**, 500–501 (2007).

[99] Schmidt, B., Sinha, R., Beresford-Smith, B. & Puglisi, S. J. A fast hybrid short read fragment assembly algorithm. *Bioinformatics* **25**, 2279–2280 (2009).

[100] Jeck, W. R. *et al.* Extending assembly of short DNA sequences to handle error. *Bioinformatics* **23**, 2942–2944 (2007).

[101] Zerbino, D. R. Using the Velvet *de novo* assembler for short-read sequencing technologies. *Current Protocols in Bioinformatics* **Chapter 11**, 1–12 (2010).

[102] Koren, Sergey and Walenz, Brian P. and Berlin, Konstantin and Miller, Jason R. and Bergman, Nicholas H. and Phillippy, A. M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* (2017).

[103] Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* **13**, 1050–1054 (2016).

[104] Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**, 540–546 (2019).

[105] Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. A. & David, N. T. HINGE: long-read assembly achieves optimal repeat resolution. *Genome Research* **27**, 747–756 (2017).

[106] Xiao, C. L. *et al.* MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. *Nature Methods* **14**, 1072–1074 (2017).

[107] Li, H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).

[108] Chen, Y. *et al.* Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nature Communications* **12**, 1–10 (2021).

[109] NextOmics. NextDenovo, `https://github.com/Nextomics/NextDenovo` (2019).

[110] Vaser, R. & Šikić, M. Yet another *de novo* genome assembler. *International Symposium on Image and Signal Processing and Analysis, ISPA* 147–151 (2019).

[111] Vaser, R. & Šikić, M. Time-and memory-efficient genome assembly with Raven. *Nature Computational Science* **1**, 332–336 (2021).

[112] Shafin, K. *et al.* Nanopore sequencing and the Shasta toolkit enable efficient *de novo* assembly of eleven human genomes. *Nature Biotechnology* **38**, 1044–1053 (2020).

[113] Liu, H., Wu, S., Li, A. & Ruan, J. SMARTdenovo: A *de novo* assembler using long noisy reads. *Preprints* (2020).

[114] Ruan, J. wtdbg, `https://github.com/ruanjue/wtdbg` (2016).

[115] Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nature Methods* **17**, 155–158 (2020).

[116] Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research* **30**, 1291–1305 (2020).

[117] Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods* 1–6 (2021).

[118] Biosciences, P. IPA, `https://github.com/PacificBiosciences/pbbioconda` (2018).

[119] Bankevich, A., Bzikadze, A., Kolmogorov, M. & Pevzner, P. A. Assembling long accurate reads using de Bruijn graphs. *bioRxiv* (2020).

[120] Ekim, B., Berger, B. & Chikhi, R. Minimizer-space de Bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer. *Cell Systems* (2021).

[121] Rautiainen, M. & Marschall, T. MBG: Minimizer-based sparse de Bruijn graph construction. *Bioinformatics* **37**, 2476–2478 (2021).

[122] Chin, C.-S. & Khalak, A. Human genome assembly in 100 minutes. *bioRxiv* (2019).

[123] Tarhio, J. & Ukkonen, E. A greedy approximation algorithm for constructing shortest common superstrings. *Theoretical Computer Science* **57**, 131–145 (1988).

[124] Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Research* **45**, e18–e18 (2017).

[125] Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research* **6**, 2601–2610 (1979).

[126] Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).

[127] de Bruijn, N. G. A combinatorial problem. *Koninklijke Nederlandse Akademie v. Wetenschappen* **49**, 758–764 (1946).

[128] Flye Sainte-Marie, C. 48. *L'Intermédiaire des Mathématiciens* **1**, 107–110 (1894).

[129] Compeau, P. E., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology* **29**, 987–991 (2011).

[130] Yahalomi, D. *et al.* A cnidarian parasite of salmon (Myxozoa: Henneguya) lacks a mitochondrial genome. *Proceedings of the National Academy of Sciences* **117**, 5358–5363 (2020).

[131] Vogg, M. C. *et al.* An evolutionarily-conserved Wnt3/β-catenin/Sp5 feedback loop restricts head organizer activity in Hydra. *Nature Communications* **10**, 1–15 (2019).

[132] Leffler, E. M. *et al.* Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biology* **10**, e1001388 (2012).

[133] Wick, R. R. Filtlong, `https://github.com/rrwick/Filtlong` (2017).

[134] Haghshenas, E., Hach, F., Sahinalp, S. C. & Chauve, C. CoLoRMap: Correcting Long Reads by Mapping short reads. *Bioinformatics* **32**, i545–i551 (2016).

[135] Firtina, C., Bar-Joseph, Z., Alkan, C. & Cicek, A. E. Hercules: a profile HMM-based hybrid error correction algorithm for long reads. *Nucleic Acids Research* **46**, e125–e125 (2018).

[136] Miclotte, G. *et al.* Jabba: hybrid error correction for long sequencing reads. *Algorithms for Molecular Biology* **11**, 1–12 (2016).

[137] Salmela, L. & Rivals, E. LoRDEC: Accurate and efficient long read error correction. *Bioinformatics* **30**, 3506–3514 (2014).

[138] Salmela, L., Walve, R., Rivals, E., Ukkonen, E. & Sahinalp, C. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics* **33**, 799–806 (2017).

[139] Hackl, T., Hedrich, R., Schultz, J. S. & Förster, F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004–3011 (2014).

[140] Morisse, P., Marchet, C., Limasset, A., Lecroq, T. & Lefebvre, A. Scalable long read self-correction and assembly polishing with multiple sequence alignment. *Scientific Reports* **11**, 1–13 (2021).

[141] Tischler, G. & Myers, E. W. Non hybrid long read consensus using local de Bruijn graph assembly. *bioRxiv* (2017).

[142] Bao, E., Xie, F., Song, C. & Song, D. FLAS: fast and high-throughput algorithm for PacBio long-read self-correction. *Bioinformatics* **35** (2019).

[143] Bao, E. & Lan, L. HALC: High throughput algorithm for long read error correction. *BMC Bioinformatics* **18** (2017).

[144] Warren, R. L. *et al.* ntEdit: scalable genome sequence polishing. *Bioinformatics* **35**, 4430–4432 (2019).

[145] Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**, e112963 (2014).

[146] Zimin, A. V. & Salzberg, S. L. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLOS Computational Biology* **16**, 1–8 (2020).

[147] Firtina, C. *et al.* Apollo: a sequencing-technology-independent, scalable, and accurate assembly polishing algorithm. *Bioinformatics* (2020).

[148] Ritu Kundu, Joshua Casey, W.-k. S. HyPo : super fast & accurate polisher for long read assemblies. *bioRxiv* (2019).

[149] Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Research* **27**, 737–746 (2017).

[150] PacificBiosciences. GenomicConsensus, `https://github.com/PacificBiosciences/GenomicConsensus` (2014).

[151] Technologies, O. N. Medaka, `https://github.com/nanoporetech/medaka` (2017).

[152] Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* (2019).

[153] Simpson, J. Nanopolish, `https://github.com/jts/nanopolish` (2014).

[154] Huang, S., Kang, M. & Xu, A. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics* **33**, 2577–2579 (2017).

[155] Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).

[156] Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 1–10 (2018).

[157] Pop, M., Kosack, D. S. & Salzberg, S. L. Hierarchical scaffolding with Bambus. *Genome Research* **14**, 149–159 (2004).

[158] Mandric, I. & Zelikovsky, A. Solving scaffolding problem with repeats. *bioRxiv* (2018).

[159] Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J. & Arvestad, L. BESST - Efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics* **15**, 1–11 (2014).

[160] Luo, J., Wang, J., Zhang, Z., Li, M. & Wu, F.-X. BOSS: a novel scaffolding algorithm based on an optimized scaffold graph. *Bioinformatics* **33**, 169–176 (2017).

[161] Gritsenko, A. A., Nijkamp, J. F., Reinders, M. J. & Ridder, D. d. GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies. *Bioinformatics* **28**, 1429–1437 (2012).

[162] Salmela, L., Mäkinen, V., Välimäki, N., Ylinen, J. & Ukkonen, E. Fast scaffolding with small independent mixed integer programs. *Bioinformatics* **27**, 3259–3265 (2011).

[163] Gao, S., Sung, W.-K. & Nagarajan, N. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *Journal of Computational Biology* **18**, 1681–1691 (2011).

[164] Mandric, I. & Zelikovsky, A. ScaffMatch: scaffolding algorithm based on maximum weight matching. *Bioinformatics* **31**, 2632–2638 (2015).

[165] Bodily, P. M., Fujimoto, M. S., Snell, Q., Ventura, D. & Clement, M. J. ScaffoldScaffolder: solving contig orientation via bidirected to directed graph reduction. *Bioinformatics* **32**, 17–24 (2016).

[166] Donmez, N. & Brudno, M. Scarpa: scaffolding reads with practical algorithms. *Bioinformatics* **29**, 428–434 (2013).

[167] Li, M., Tang, L., Wu, F.-X., Pan, Y. & Wang, J. SCOP: a novel scaffolding algorithm based on contig classification and optimization. *Bioinformatics* **35**, 1142–1150 (2019).

[168] Roy, R. S., Chen, K. C., Sengupta, A. M. & Schliep, A. SLIQ: Simple Linear Inequalities for Efficient Contig Scaffolding. *Journal of Computational Biology* **19**, 1162–1175 (2012).

[169] Dayarian, A., Michael, T. P. & Sengupta, A. M. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. *BMC Bioinformatics* **11**, 1–21 (2010).

[170] Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using sspace. *Bioinformatics* **27**, 578–579 (2011).

[171] Farrant, G. K. *et al.* WiseScaffolder: an algorithm for the semi-automatic scaffolding of next generation sequencing data. *BMC Bioinformatics* **16**, 1–13 (2015).

[172] Warren, R. L. *et al.* LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience* **4**, s13742–015 (2015).

[173] Qin, M. *et al.* LRScaf: improving draft genomes using long noisy reads. *BMC Genomics* **20**, 1–12 (2019).

[174] Cao, M. D. *et al.* Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nature Communications* **8**, 1–10 (2017).

[175] English, A. C. *et al.* Mind the Gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS One* **7** (2012).

[176] Warren, R. L. RAILS and Cobbler: Scaffolding and automated finishing of draft genomes using long DNA sequences. *Journal of Open Source Software* **1**, 116 (2016).

[177] Luo, J. *et al.* SLR: a scaffolding algorithm based on long reads and contig classification. *BMC Bioinformatics* **20** (2019).

[178] Wellcome Sanger Institute. SMIS, `https://www.sanger.ac.uk/tool/smis/` (2015).

[179] Zhu, S., Chen, D. Z. & Emrich, S. J. Single molecule sequencing-guided scaffolding and correction of draft assemblies. *BMC Genomics* **18**, 51–59 (2017).

[180] Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15** (2014).

[181] Tang, H. *et al.* ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biology* **16**, 1–15 (2015).

[182] Lin, H. C. *et al.* AGORA: assembly guided by optical restriction alignment. *BMC Bioinformatics* **13**, 1–14 (2012).

[183] Istace, B., Belser, C. & Aury, J.-M. BiSCoT: improving large eukaryotic genome assemblies with optical maps. *PeerJ* **8** (2020).

[184] Pan, W., Jiang, T. & Lonardi, S. OMGS: optical map-based genome scaffolding. *Journal of Computational Biology* **27**, 519–533 (2020).

[185] Shelton, J. M. *et al.* Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC Genomics* **16** (2015).

[186] Nagarajan, N., Read, T. D. & Pop, M. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics* **24**, 1229–1235 (2008).

[187] Hiltunen, M., Ryberg, M. & Johannesson, H. ARBitR: an overlap-aware genome assembly scaffolder for linked reads. *Bioinformatics* **37**, 2203–2205 (2020).

[188] Kuleshov, V., Snyder, M. P. & Batzoglou, S. Genome assembly from synthetic long read clouds. *Bioinformatics* **32**, i216–i224 (2016).

[189] Yeo, S., Coombe, L., Warren, R. L., Chu, J. & Birol, I. ARCS: scaffolding genome drafts with linked reads. *Bioinformatics* **34**, 725–731 (2018).

[190] Coombe, L. *et al.* ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics* **19** (2018).

[191] Adey, A. *et al.* *In vitro*, long-range sequence information for *de novo* genome assembly via transposase contiguity. *Genome Research* **24**, 2041–2049 (2014).

[192] High Performance Assembly Group at the Wellcome Sanger Institute. Scaff10X, `https://github.com/wtsi-hpag/Scaff10X` (2018).

[193] Kaplan, N. & Dekker, J. High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nature Biotechnology* **31**, 1143–1147 (2013).

[194] Marie-Nelly, H. *et al.* High-quality genome (re)assembly using chromosomal contact data. *Nature Communications* **5** (2014).

[195] Renschler, G. *et al.* Hi-C guided assemblies reveal conserved regulatory topologies on X and autosomes despite extensive genome shuffling. *Genes & Development* **33**, 1591–1612 (2019).

[196] Baudry, L. *et al.* instaGRAAL: chromosome-level quality scaffolding of genomes using a proximity ligation-based scaffolder. *Genome Biology* **21** (2020).

[197] Burton, J. N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology* **31**, 1119–1125 (2013).

[198] Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chin, C.-S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18** (2017).

[199] Ghurye, J. *et al.* Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Computational Biology* **15** (2019).

[200] Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biology* **13** (2012).

[201] Chu, C., Li, X. & Wu, Y. GAPPadder: a sensitive approach for closing gaps on draft genomes with short sequence reads. *BMC Genomics* **20** (2019).

[202] Paulino, D. *et al.* Sealer: a scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics* **16** (2015).

[203] Piro, V. C. *et al.* FGAP: an automated gap closing tool. *BMC Research Notes* **7** (2014).

[204] Kosugi, S., Hirakawa, H. & Tabata, S. GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. *Bioinformatics* **31**, 3733–3741 (2015).

[205] Xu, G.-C. *et al.* LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *GigaScience* **8** (2019).

[206] Lu, P. *et al.* PGcloser: fast parallel gap-closing tool using long-reads or contigs to fill gaps in genomes. *Evolutionary Bioinformatics* **16** (2020).

[207] Xu, M. *et al.* TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* **9** (2020).

[208] Morisse, P., Lecroq, T. & Lefebvre, A. Long-read error correction: a survey and qualitative comparison. *bioRxiv* (2020).

[209] Ko, B. J. *et al.* Widespread false gene gains caused by duplication errors in genome assemblies. *bioRxiv* (2021).

[210] Guiglielmoni, N., Houtain, A., Derzelle, A., Van Doninck, K. & Flot, J.-F. Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. *BMC Bioinformatics* **22** (2021).

[211] Kent, W. J. & Haussler, D. Assembly of the working draft of the human genome with GigAssembler. *Genome Research* **11**, 1541–1548 (2001).

[212] Ghurye, J. & Pop, M. Modern technologies and algorithms for scaffolding assembled genomes. *PLoS Computational Biology* **15** (2019).

[213] Fierst, J. L. Using linkage maps to correct and scaffold *de novo* genome assemblies: methods, challenges, and computational tools. *Frontiers in Genetics* **6**, 220 (2015).

[214] Schwartz, D. C. *et al.* Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* **262**, 110–114 (1993).

[215] Protasio, A. V. *et al.* A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni. *PLoS Neglected Tropical Diseases* **6** (2012).

[216] Jeong, C. B. *et al.* The genome of the harpacticoid copepod *Tigriopus japonicus*: potential for its use in marine molecular ecotoxicology. *Aquatic Toxicology* **222**, 105462 (2020).

[217] Fuller, Z. L. *et al.* Population genetics of the coral *Acropora millepor*a: Toward genomic prediction of bleaching. *Science* **369** (2020).

[218] Yuan, Y., Chung, C. Y.-L. & Chan, T.-F. Advances in optical mapping for genomic research. *Computational and Structural Biotechnology Journal* (2020).

[219] Cotton, J. A. *et al.* The genome of *Onchocerca volvulus*, agent of river blindness. *Nature Microbiology* **2** (2016).

[220] Wang, J. *et al.* Comparative genome analysis of programmed DNA elimination in nematodes. *Genome Research* **27**, 2001–2014 (2017).

[221] Tsai, I. J. *et al.* The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* **496**, 57–63 (2013).

[222] Olson, P. D. *et al.* Complete representation of a tapeworm genome reveals chromosomes capped by centromeres, necessitating a dual role in segregation and protection. *BMC Biology* **18** (2020).

[223] Varney, R. M., Speiser, D. I., McDougall, C., Degnan, B. M. & Kocot, K. M. The iron-responsive genome of the chiton *Acanthopleura granulata*. *Genome Biology and Evolution* **13** (2021).

[224] Kocher, S. D. *et al.* The genetic basis of a social polymorphism in halictid bees. *Nature Communications* **9** (2018).

[225] SuperNova. SuperNova, `https://github.com/10XGenomics/supernova` (2016).

[226] Ghurye, J. *et al.* A chromosome-scale assembly of the major african malaria vector *Anopheles funestus*. *GigaScience* **8** (2019).

[227] Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).

[228] Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics* **14**, 390–403 (2013).

[229] Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–93 (2009).

[230] Flot, J.-F., Marie-Nelly, H. & Koszul, R. Contact genomics: scaffolding and phasing (meta) genomes using chromosome 3D physical signatures. *FEBS Letters* **589**, 2966–2974 (2015).

[231] Oddes, S., Zelig, A. & Kaplan, N. Three invariant Hi-C interaction patterns: applications to genome assembly. *Methods* **142**, 89–99 (2018).

[232] Techer, M. A. *et al.* Divergent evolutionary trajectories following speciation in two ectoparasitic honey bee mites. *Communications Biology* **2** (2019).

[233] Shingate, P. *et al.* Chromosome-level assembly of the horseshoe crab genome provides insights into its genome evolution. *Nature Communications* **11** (2020).

[234] Hu, M., Zheng, X., Fan, C.-M. & Zheng, Y. Lineage dynamics of the endosymbiotic cell type in the soft coral *Xenia*. *Nature* **582**, 534–538 (2020).

[235] Li, Y. *et al.* Chromosome-level reference genome of the jellyfish *Rhopilema esculentum*. *GigaScience* **9** (2020).

[236] Davidson, P. L. *et al.* Chromosomal-level genome assembly of the sea urchin *Lytechinus variegatus* substantially improves functional genomic analyses. *Genome Biology and Evolution* **12**, 1080–1086 (2020).

[237] Ruiz-Ramos, D. V., Schiebelhut, L. M., Hoff, K. J., Wares, J. P. & Dawson, M. N. An initial comparative genomic autopsy of wasting disease in sea stars. *Molecular Ecology* **29**, 1087–1102 (2020).

[238] Bai, C. M. *et al.* Chromosomal-level assembly of the blood clam, *Scapharca (Anadara) broughtonii*, using long sequence reads and Hi-C. *GigaScience* **8** (2019).

[239] Sun, J. *et al.* The scaly-foot snail genome and implications for the origins of biomineralised armour. *Nature Communications* **11** (2020).

[240] Teterina, A. A., Willis, J. H. & Phillips, P. C. Chromosome-level assembly of the *Caenorhabditis remanei* genome reveals conserved patterns of nematode genome organization. *Genetics* **214**, 769–780 (2020).

[241] Lian, Y. *et al.* Chromosome-level reference genome of X12, a highly virulent race of the soybean cyst nematode *Heterodera glycines*. *Molecular Ecology Resources* **19**, 1637–1646 (2019).

[242] Stroehlein, A. J. *et al.* High-quality *Schistosoma haematobium* genome achieved by single-molecule and long-range sequencing. *GigaScience* **8** (2019).

[243] Kenny, N. J. *et al.* Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge *Ephydatia muelleri*. *Nature Communications* **11https://pubmed.ncbi.nlm.nih.gov/32719321/** (2020).

[244] Simion, P. *et al.* Chromosome-level genome assembly reveals homologous chromosomes and recombination in asexual rotifer *Adineta vaga*. *Science Advances* **7** (2021).

[245] Gehrke, A. R. *et al.* Acoel genome reveals the regulatory landscape of whole-body regeneration. *Science* **363** (2019).

[246] Haghshenas, E., Asghari, H., Stoye, J., Chauve, C. & Hach, F. HASLR: Fast hybrid assembly of long reads. *Iscience* **23** (2020).

[247] Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).

[248] Di Genova, A., Buena-Atienza, E., Ossowski, S. & Sagot, M.-F. Efficient hybrid *de novo* assembly of human genomes with WENGAN. *Nature Biotechnology* **39**, 422–430 (2021).

[249] Mulligan, K. L., Hiebert, T. C., Jeffery, N. W. & Gregory, T. R. First estimates of genome size in ribbon worms (phylum Nemertea) using flow cytometry and Feulgen image analysis densitometry. *Canadian Journal of Zoology* **92**, 847–851 (2014).

[250] Joint Genome Institute. BBtools, `https://sourceforge.net/projects/bbmap/` (2013).

[251] Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**, 1432 (2020).

[252] Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a K-mer Analysis Toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2016).

[253] Pucker, B. Mapping-based genome size estimation. *bioRxiv* (2019).

[254] Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

[255] Guo, Y. *et al.* A chromosomal-level genome assembly for the giant African snail *Achatina fulica*. *GigaScience* **8** (2019).

[256] He, C. *et al.* Factorial estimating assembly base errors using k-mer abundance difference (KAD) between short reads and genome assembled sequences. *NAR Genomics and Bioinformatics* **2** (2020).

[257] Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**, 1287 (2017).

[258] Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. BlobToolKit – Interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics* **10**, 1361–1374 (2020).

[259] Boothby, T. C. *et al.* Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 15976–15981 (2015).

[260] Koutsovoulos, G. *et al.* No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proceedings of the National Academy of Sciences* **113**, 5053–5058 (2016).

[261] Zhang, X., Wu, R., Wang, Y., Yu, J. & Tang, H. Unzipping haplotypes in diploid and polyploid genomes. *Computational and Structural Biotechnology Journal* **18**, 66–72 (2020).

[262] Koren, S. *et al.* *De novo* assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology* **36**, 1174–1182 (2018).

[263] Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research* **27**, 801–812 (2017).

[264] Patterson, M. D. *et al.* WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology* **22**, 498–509 (2015).

[265] Limasset, A. *Novel approaches for the exploitation of high throughput sequencing data.* Ph.D. thesis, Université Rennes 1 (2017).

[266] Kajitani, R. *et al.* Platanus-allee is a *de novo* haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nature Communications* **10** (2019).

[267] Porubsky, D. *et al.* Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nature Biotechnology* **39**, 302–308 (2020).

[268] Zhou, Q. *et al.* Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nature Genetics* **52**, 1018–1023 (2020).

[269] Holley, G. *et al.* Ratatosk: hybrid error correction of long reads enables accurate variant calling and assembly. *Genome Biology* **22** (2021).

[270] Aury, J.-M. & Istace, B. Hapo-G, haplotype-aware polishing of genome assemblies. *NAR Genomics and Bioinformatics* **3** (2021).

[271] Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants* **5**, 833–845 (2019).

[272] Faure, R., Guiglielmoni, N. & Flot, J.-F. GraphUnzip: unzipping assembly graphs with long reads and Hi-C. *bioRxiv* (2021).

[273] Kronenberg, Z. N. *et al.* Extended haplotype-phasing of long-read de novo genome assemblies using hi-c. *Nature Communications* **12**, 1–10 (2021).

[274] Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 1–27 (2020).

[275] Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 1–9 (2016).

[276] Lariviere, D. & Ostrovsky, A. VGP assembly pipeline (Galaxy Training Materials), `training.galaxyproject.org/training-material/topics/assembly/tutorials/vgp_genome_assembly/tutorial.html` (2021).