Peer Community Journal

Section: Genomics

RESEARCH ARTICLE

A deep dive into genome assemblies of non-vertebrate animals

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

Volume **2** (2022), article e29

https://doi.org/10.24072/pcjournal.128

Published
2022-05-16

Cite as
Nadège Guiglielmoni, Ramón Rivera-Vicéns, Romain Koszul and Jean-François Flot (2022) *A deep dive into genome assemblies of non-vertebrate animals*, Peer Community Journal, 2: e29.

Correspondence
nguiglie@uni-koeln.de

Peer-review
Peer reviewed and recommended by PCI Genomics,
https://doi.org/10.24072/pci.genomics.100016

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.

[1]Institut für Zoologie, Universität zu Köln, 50674 Cologne, Germany, [2]Service Evolution Biologique et Ecologie, Université libre de Bruxelles (ULB), 1050 Brussels, Belgium, [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

CENTRE MERSENNE

Peer Community Journal is a member of the Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/

e-ISSN 2804-3871

Contents

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], $\sim$ 56.9% (4,493) belong to the subphylum Vertebrata. However, from the currently $\sim$2.1 million described metazoan species, only $\sim$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 ($\sim$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* [9, 11, 12] 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) [13].

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 [14–16]. Terpenoid metabolites have been found in more than 70 gastropod species [17].

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 [18]. 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 [19]. Moreover, non-vertebrates are important model systems to understand processes such as adaptation to climate change, ocean acidification, biomineralization [20–23]. Various species of corals [24–27] 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 [28]. New genome resources bring novel insights into difficult phylogenetic positions: a large analysis based on genomes and transcriptomes confirmed that myxozoans belonged to Cnidaria [29]; the sequence of *Hoilungia hongkongiensis* placed placozoans as a sister group to cnidarians and bilaterians [30]. Genomic studies have also attempted to elucidate the mechanisms underlying asexuality, as sexual reproduction is a character shared by almost all eukaryotes and its strict absence generally leads to rapid extinction due to the accumulation of deleterious mutations [31], yet ancient asexual species are observed in many branches of phylogeny [32–36].

The dearth of non-vertebrate animal genomic resources may be blamed to the difficulty to collect individuals in remote or hardly accessible locations and in accordance with the Nagoya protocol [37]; the scarcity of certain species; non-existing resources to cultivate individuals in laboratories; the lack of protocols to extract pure, high-molecular-weight DNA; 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) [38–40], used to measure the completeness of a genome assembly, relies on reference gene sets that are used for scoring, based on existing assemblies for a group of species. 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 [41] searches for animal species 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. As a result, identifying metazoan contaminants in a fragmented assembly of an animal genome is almost impossible due to similar GC contents and the absence of hits in genomic databases.

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) [42, 43] have been put in place to overcome some of the aforementioned limitations. Other consortia such as the Earth BioGenome Project [44], the Darwin Tree of Life [45], the Aquatic Symbiosis Genomics Project [46] and the European Reference Genome Atlas [47] 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. In this review, we first offer a brief historical overview of sequencing technologies and algorithmic approaches to genome assembly. We then survey software for genome assembly, pre/post-processing steps, assembly evaluation, and phasing assemblies, to help newcomers to the field build their own assembly pipelines and have an overview of past and current tools. Although sequencing methods, algorithms and programs presented in this paper are not restricted to a category of organisms, the challenges and solutions that we describe are specific to non-model non-vertebrate animal genomes.

1. 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% [48]. 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* [49] was the first eukaryote sequenced, whereas the nematode *Caenorhabditis elegans* was the first metazoan [50]. Sanger sequencing is a relatively low-throughput method in terms of the number of sequences generated, and is costly as well [51]. 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 [52].

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 [53]. 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 [54], Ion Torrent [55], SOLiD [56], and Solexa (for a comparison on the approaches, see [57]). Among these methods, Solexa, subsequently purchased by Illumina [58], became and remains the most widely 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)).

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 [59]. 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 [60]. 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%, despite a smaller maximal read length [61].

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 [62]. 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 [63]. Researchers

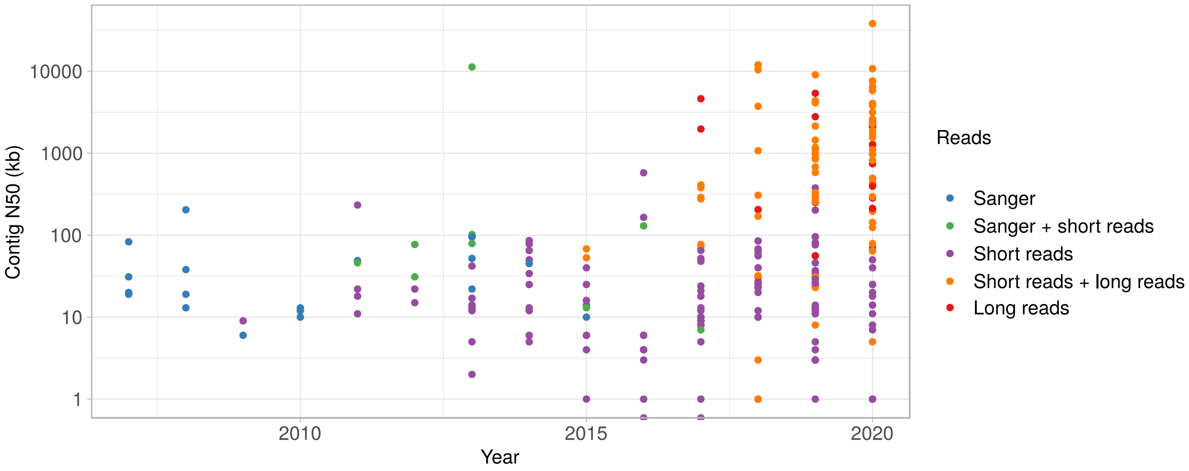

Figure 1 – Contig N50 of 237 non-vertebrate animal genome assemblies over time. The N50 represents the contiguity of an assembly and is defined as the length of the largest contig for which at least 50% of the assembly size is contained in contigs equal or greater in length.

can control how they generate their sequencing data, contribute to protocol development, and develop their own basecalling [64] to increase the yield and improve the quality and length of the reads. Although Nanopore reads still typically exhibit a high error rate, their length keeps increasing to attain hundreds of kilobases to 1 Mb [65]. The error rate has also been decreasing with the development of more accurate basecallers such as Bonito [66] combined with Poreover [67], and the release of the new R10 flow cell which can estimate the length of homopolymeric regions more accurately and produce reads with an error rate below 1% [68].

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 [69]. 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.

2. 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) [135]. Greedy algorithms have been implemented for first-generation sequencing reads, for instance in TIGR [81], and were further applied in short-read assemblers like PERGA [98], SSAKE [110] and VCAKE [112]. 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 [136].

Table 1 – Sequencing approaches and associated assemblers.

First generation 1 kb High accuracy Sanger	ARACHNE [70], Atlas [71], CAP3 [72], Celera [73], Euler [74], JAZZ [75], Minimus [76], MIRA [77], phrap [78], Phusion [79], SUTTA [80], TIGR [81]
Second generation 25-300 bp High accuracy 454, IonTorrent, Solexa, SOLiD	ABySS [82, 83], ALLPATHS [84], BASE [85], CABOG [86], Edena [87], EPGA [88], Euler-SR [89], Gossamer [90], IDBA [91], ISEA [92], JR-Assembler [93], LightAssembler [94], Meraculous [95], MIRA [77], Newbler [96], PCAP [97], PERGA [98], Platanus [99], PE-Assembler [100], QSRA [101], Ray [102], Readjoiner [103], SGA [104], SHARGCS [105], SOAPdenovo [106], SOAPdenovo2 [107], SPAdes [108], SparseAssembler [109], SSAKE [110], SUTTA [80], Taipan [111], VCAKE [112], Velvet [113]
Third generation 10-100,000+ kb PacBio CLR, Nanopore	Canu [114], FALCON [115], Flye [116], HINGE [117], MECAT [118], MECAT2 [118], miniasm [119], NECAT [120], NextDenovo [121], Ra [122], Raven [123], Shasta [124], SMARTdenovo [125], wtdbg [126], wtdbg2 [127]
15-25 kb High accuracy PacBio HiFi,	Flye [116], HiCanu [128], hifiasm [129], IPA [130], LJA [131], mdBG [132], MBG [133], NextDenovo [121], Peregrine [134], Raven [123], wtdbg2 [127]

The Overlap-Layout-Consensus (OLC) paradigm was first described in 1979 by Rodger Staden [137] 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 [73], which was used to assemble a human genome from a Sanger shotgun dataset [138].

De Bruijn Graphs (DBGs) (Figure 3) are a well studied structure in graph theory, described by Nicolaas Govert de Bruijn in 1946 [139] and before him by Camille Flye Sainte-Marie [140]. DBG-based assemblers require highly accurate reads to avoid a large number of erroneous *k*-mers and creating bulges in the assembly graph. 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 [141] and was quickly implemented in multiple popular short-read assemblers, e.g. ABySS [82, 83], IDBA [91], SOAPdenovo [106] and SOAPdenovo2 [107], SPAdes [108], Velvet [113]. The choice of the value *k* greatly affects the output: small *k*-mers lead to complex de Bruijn graphs, while large *k*-mers result in more fragmented assemblies [131]. DBG-based assemblers often use several *k*-mer sizes to combine the paths identified in different graphs.

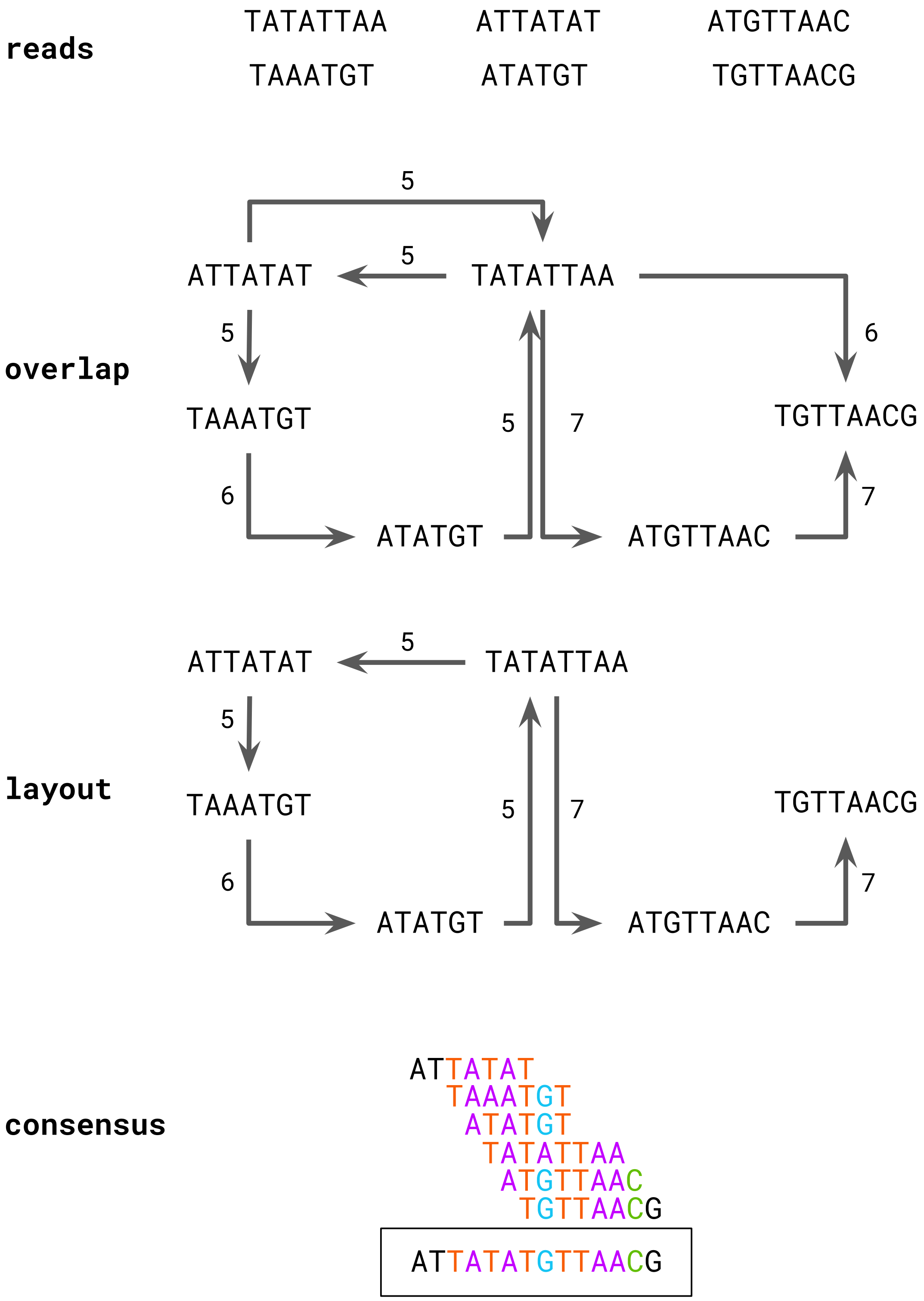

Figure 2 – Overview of Overlap-Layout-Consensus assembly. The graph was built with all overlaps of at least 5 bases with a tolerance of 1 mismatch.

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 [116], Ra [122], Raven [123], Shasta [124], wtdbg2 [127]. 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

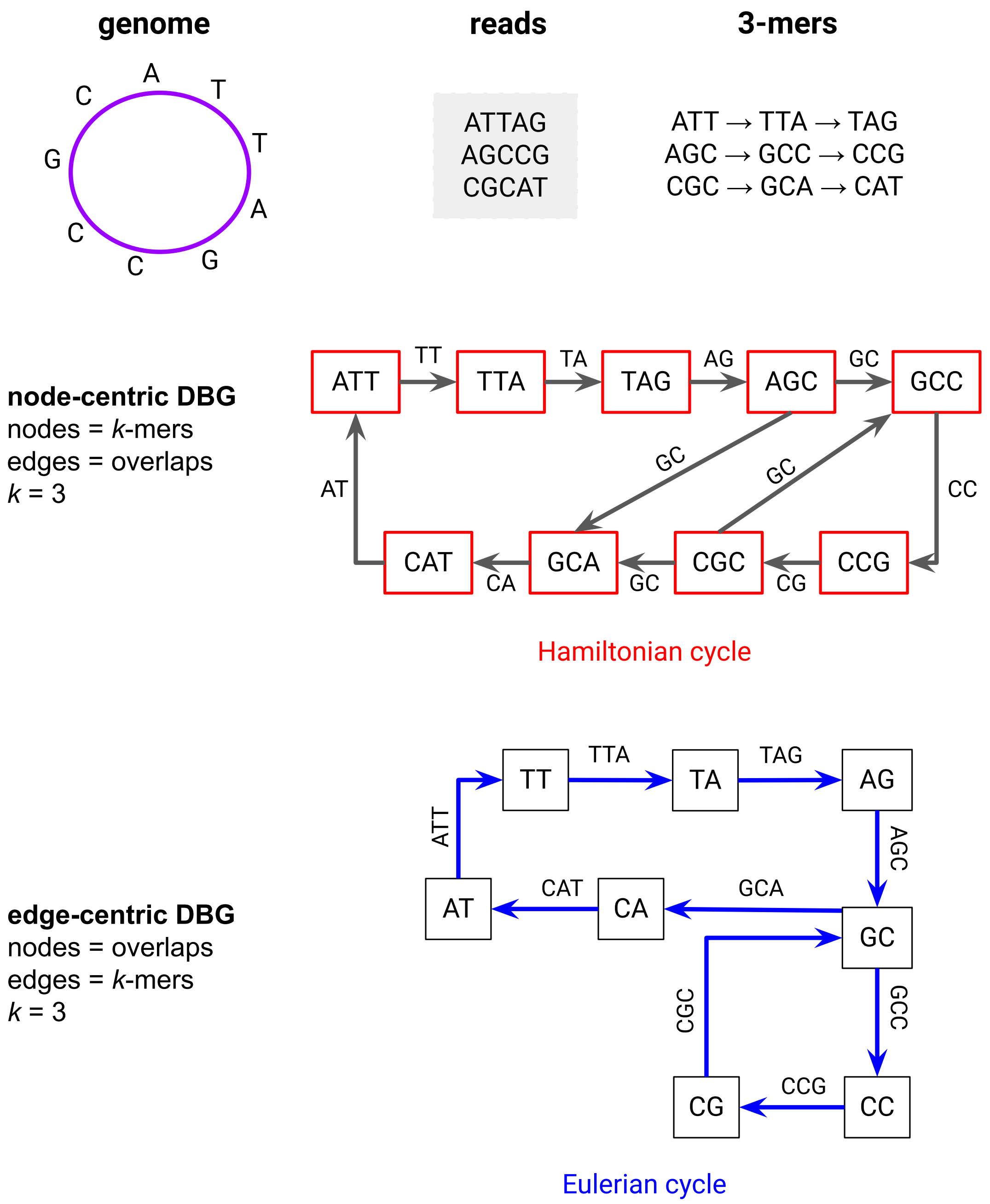

Figure 3 – Overview of genome assembly using de Bruijn graphs. A circular genome is assembled based on three reads using node-centric and edge-centric DBGs with *k* = 3. The node-centric DBG is searched for a Hamiltonian cycle (visiting all nodes), and the edge-centric DBG for an Eulerian cycle (visiting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

parameters), HiCanu [128] and hifiasm [129], and new DBG assemblers adapted for large *k*-mer values are now being released [131–133].

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

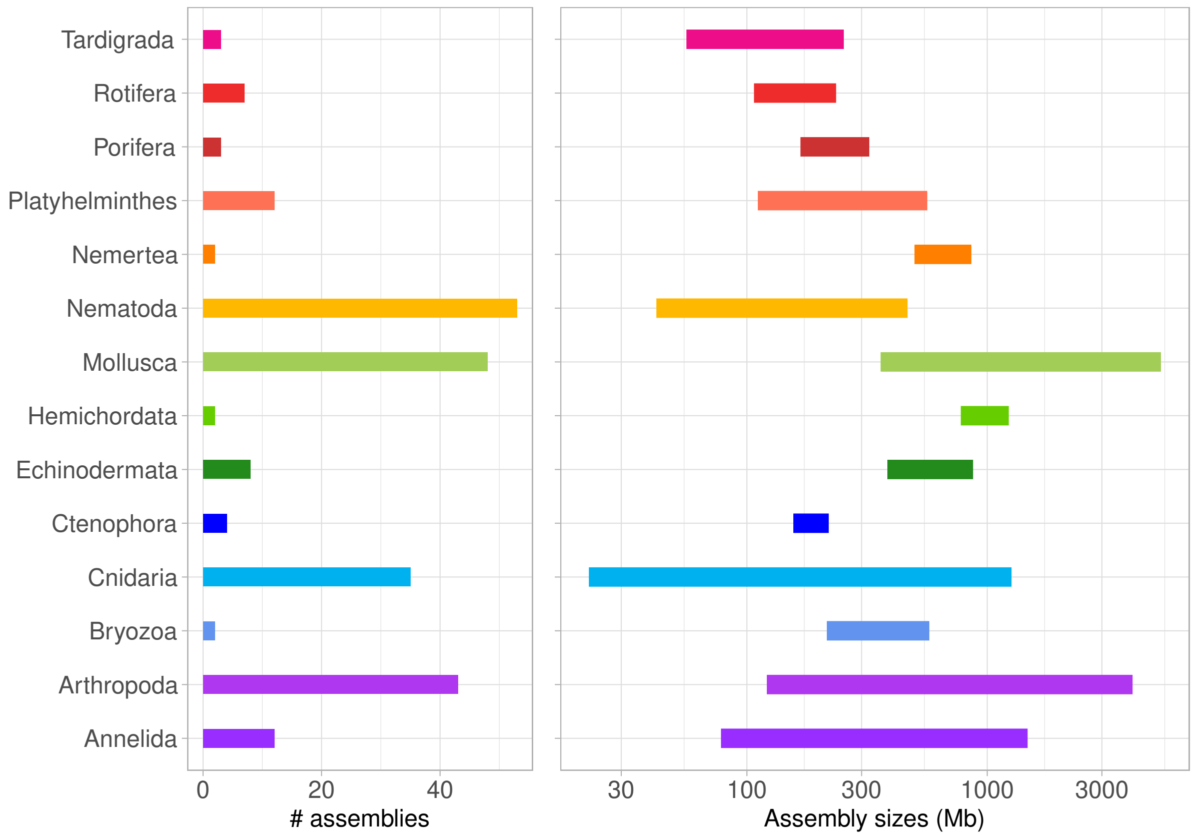

Figure 4 – Assembly sizes. The left graph shows the number of assemblies included for each phylum and the right part shows the corresponding assembly-size ranges.

Megabases (Mb) (*Kudoa iwatai*: 22.5 Mb, *Myxobolus squamalis*: 53.1 Mb, *Henneguya salminicola*: 60.0 Mb [142]), while the hydrozoan *Hydra oligactis* (1.3 Gigabases (Gb)) [143] 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 [144]. 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.

3. 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 [228]. 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

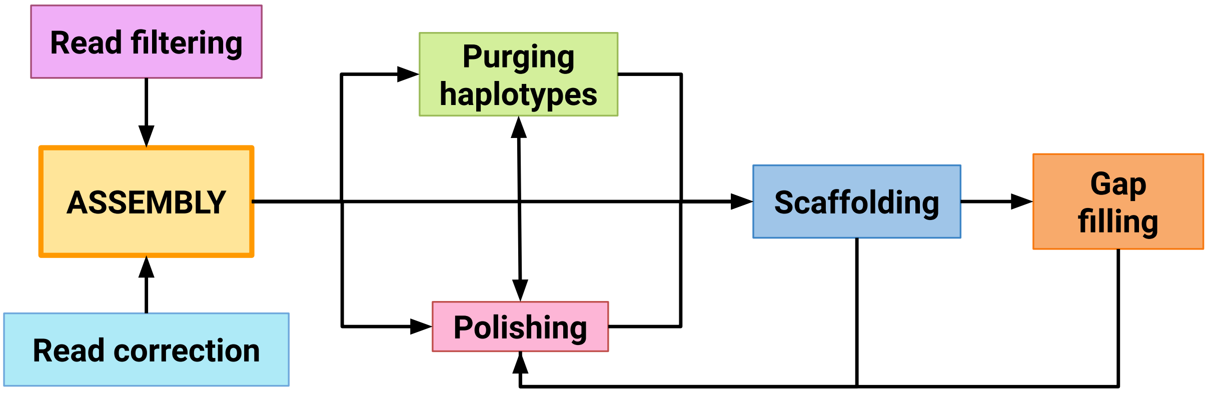

Figure 5 – Assembly pipeline, including the assembly and the pre/post-processing steps.

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

Step	**Data**	**Tools**
Read filtering	Long reads	Filtlong [145]
Long-read error correction	Short reads	CoLoRMAP [146], Hercules [147], HG-CoLoR [148], Jabba [149], LoRDEC [150], LoRMA [151], NaS [152], proovread [153]
	Long reads	Canu [114], CONSENT [154], Daccord [155], FLAS [156], HALC [157], MECAT [118], MECAT2 [118], NECAT [120], NextDenovo [121]
Polishing	Short reads	ntEdit [158], Pilon [159], POLCA [160]
	Short & long reads	Apollo [161], Hapo-G [162], HyPo [163], Racon [164]
	Long reads	Arrow [165], CONSENT [154], Medaka [166] NextPolish [167], Nanopolish [168], Quiver [165]
Haplotig purging	Long reads	HaploMerger2 [169], purge_dups [170], Purge Haplotigs [171]
Scaffolding	Short reads Mate pairs	Bambus [172], BATISCAF [173], BESST [174], BOSS [175], GRASS [176], MIP [177], Opera [178], ScaffMatch [179], ScaffoldScaffolder [180], SCARPA [181], SCOP [182], SLIQ [183], SOPRA [184], SSPACE [185], WiseScaffolder [186]
	Long reads	DENTIST [187], gapless [188], LINKS [189], LRScaf [190], npScarf [191], PBJelly [192], RAILS [193], SLR [194], SMIS [195], SMSC [196], SSPACE-LongRead [197]
	Genetic maps	ALLMAPS [198]
	Optical maps	AGORA [199], BiSCoT [200], OMGS [201], SewingMachine [202], SOMA [203]
	Linked reads	ARBitR [204], Architect [205], ARCS [206], ARKS [207], fragScaff [208], Scaff10X [209]
	3C/Hi-C	3D-DNA [12], dnaTri [210], GRAAL [211], HiCAssembler [212], instaGRAAL [213], Lachesis [214], pin_hic [215], SALSA [216], SALSA2 [217], scaffhic [218], YaHS [219]
Gap filling	Short reads	GapFiller [220], GAPPadder [221], Sealer [222]
	Long reads	Cobbler [193], DENTIST [187], FGAP [223], gapless [188], GMcloser [224], LR_Gapcloser [225], PBJelly [192], PGcloser [226], TGS-GapCloser [227]

and have been thoroughly reviewed on multiple datasets [229]. When tested on *Caenorhabditis elegans* Nanopore reads, the error rate decreased from 28.93% to less than 1% (using Canu [114], CONSENT [154], FLAS [156], Jabba [149], LORMA [151] or MECAT [118]). 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 [230]. Some long-read assemblers, Ra and wtdbg2, have been identified as less prone to retain uncollapsed haplotypes [231]. Contigs can also be post-processed to remove these duplications with dedicated tools such as HaploMerger2 [169], purge_dups [170] and Purge Haplotigs [171]. 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 [73], CAP3 [72], GigAssembler [232]). Since then, several sequencing techniques have been used to scaffold assemblies: mate pairs, long reads, genetic maps, optical mapping, linked reads, and proximity ligation [233]. 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 [234] and optical maps [235] 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* [236], the copepod *Tigriopus japonicus* [237] and the coral *Acropora millepora* [26]. Although existing genetic maps provide precious resources, building one is particularly difficult as it requires breeding [234], 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 [238], and this method was included in some non-vertebrate genome projects: several nematodes including *Onchocerca volvulus* [239], *Ascaris suum* and *Parascaris univalens* [240], the tapeworms *Echinococcus multilocularis* [241] and *Hymenolepis microstoma* [242], and the chiton *Acanthopleura granulata* [243].

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. New linked-read methods are now emerging such as haplotagging [244] and TELL-seq [245], and the latter protocol is able to handle inputs as low as a few nanograms of DNA. Linked reads have been used to scaffold the genomes of the coral *Acropora millepora* [26] and the bee *Lasioglossum albipes* [246]. As linked reads are also shotgun Illumina reads, these reads are sometimes used for assembly (using Architect [205] or Supernova [247]) or polishing, as was done for the mosquito *Anopheles funestus* [248].

Proximity ligation techniques, based on capture of chromosome conformation [249], were not originally developed with genome sequencing applications in mind. Instead, they aimed at investigating the interplay between chromosome 3D organization and DNA processes [250]. A popular genomic derivative of 3C, Hi-C [251] documents the average conformation of the

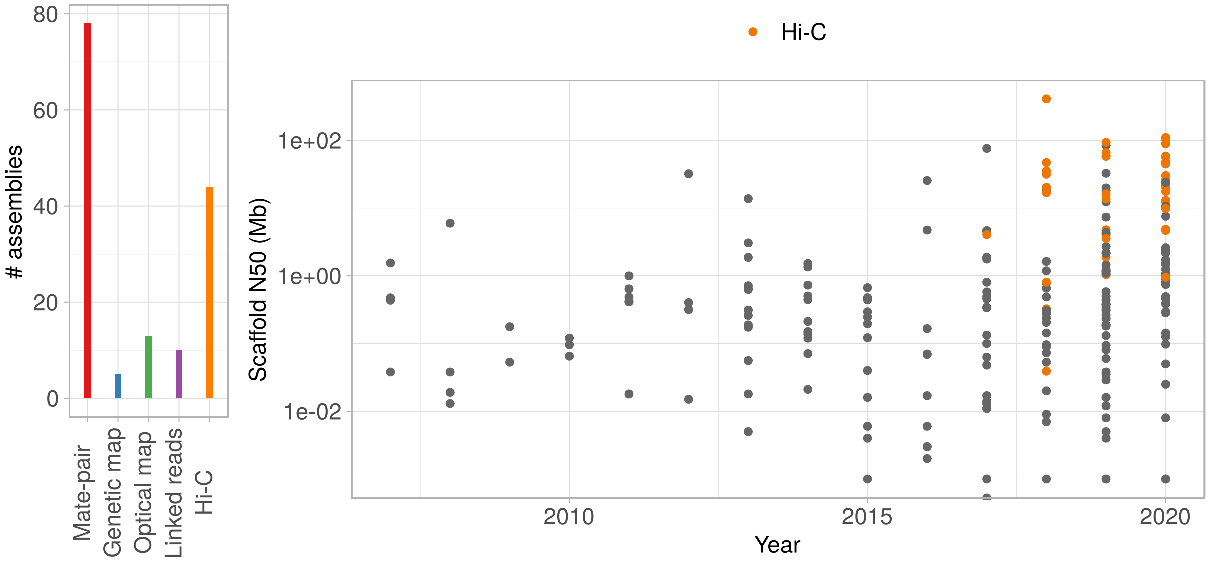

Figure 6 – Assemblies scaffolding. Left: number of assemblies that included each scaffolding method. Right: scaffold N50 of non-vertebrate animal genome assemblies over time. The assemblies that included a Hi-C scaffolding step are highlighted in orange; they form a cluster with a scaffold N50 over 1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

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 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 [252, 253]. 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 [217]), 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 [211] and its later improved version instaGRAAL [213]). These tools are not yet able to estimate the gap size separating two contigs connected into a scaffold, thus they usually insert gaps with an arbitrary length. Most Hi-C protocols use one or several restriction enzymes, leading to an enrichment of Hi-C reads around recognition sites and making 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; these Hi-C reads can be used as single-end reads for short-read assembly.

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), yet they still require a minimum input of about 0.5-1 million cells. 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* [254], *Carcinoscorpius rotundicauda* [255], and *Cataglyphis hispanica* [256], the cnidarians *Xenia* sp. [257] and *Rhopilema esculentum* [258], the echinoderms *Lytechinus variegatus* [259] and *Pisaster ochraceus* [260], the molluscs *Scapharca broughtonii* [261], *Chrysomallon squamiferum* [262], and *Mercenaria mercenaria* [263], the nematods *Caenorhabditis remanei* [264] and *Heterodera glycines* [265], the platyhelminthe *Schistosoma haemotabium* [266], the poriferan *Ephydatia muelleri* [267], the rotifer *Adineta vaga* [36], the xenacoelomorph *Hofstenia miamia* [268], 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.

Pre/post-processing steps are often included in assembly tools: Canu, MECAT, MECAT2, NECAT and NextDenovo correct low-accuracy long reads prior to assembly; Flye, Raven and NextDenovo have a polishing step; and assemblers can include a scaffolding step to yield both contigs and scaffolds. Users can choose however to skip these steps and perform their own pre/post-processing instead, or in addition. Some assemblers propose a hybrid assembly strategy, using both short and long reads, such as HALSR [269], MaSuRCA [270] and WENGAN [271].

4. 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 [272], 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 [273], GenomeScope [274] and KAT [275], 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 [276] estimates genome size based on reads mapping to a highly contiguous 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 (or scaffold) for which 50% of the assembly size is contained in contigs (or scaffolds) 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 [277]. 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* [278] and the coral *Xenia* sp. [257].

Another feature to optimize is the completeness of the genome, usually based on orthologs or *k*-mers. BUSCO [38–40] searches for orthologs in a user-provided lineage; the current Metazoa lineage (designated as Metazoa odb10) 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

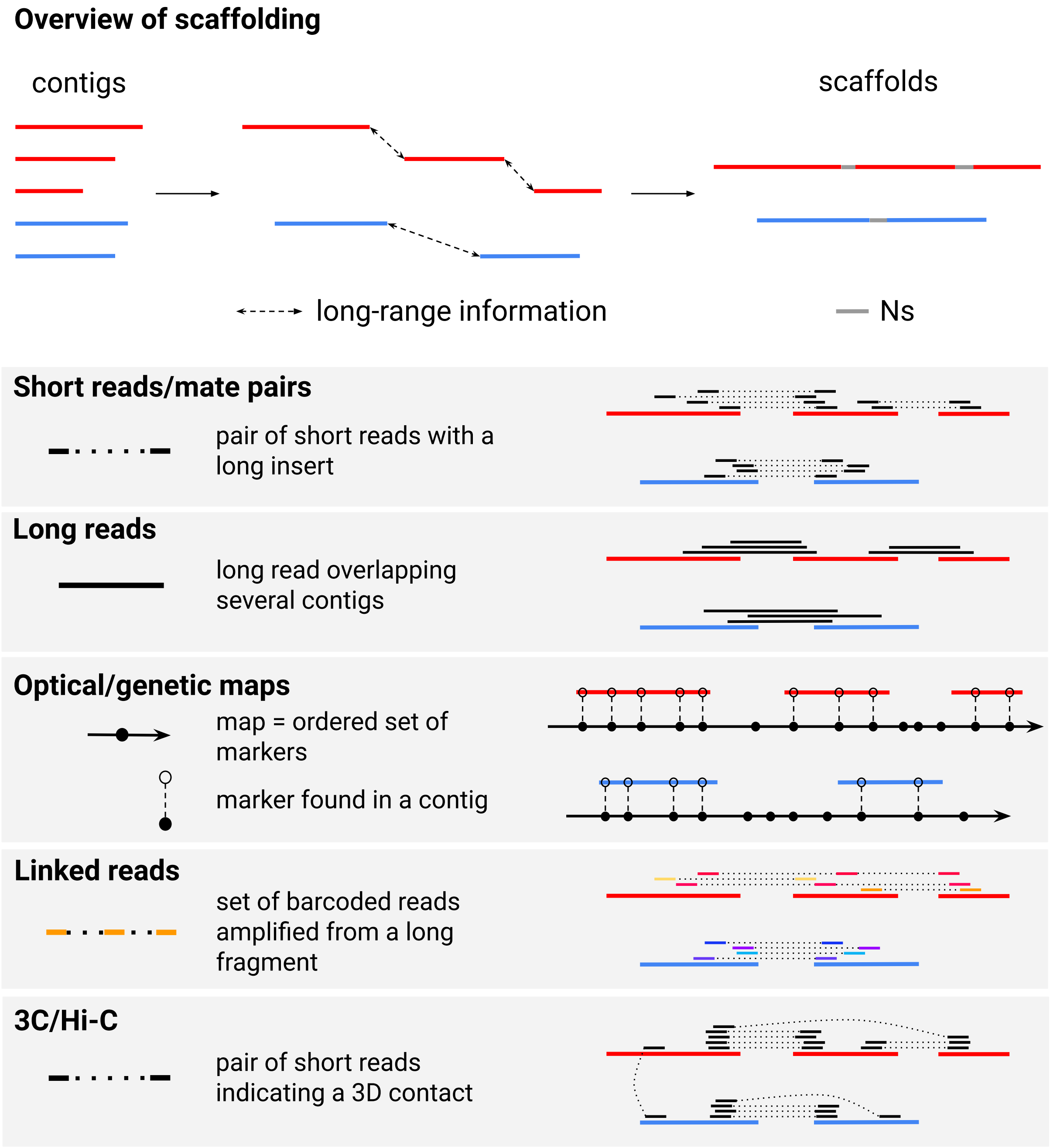
Overview of scaffolding
contigs
scaffolds
long-range information
Ns
Short reads/mate pairs
pair of short reads with a long insert
Long reads
long read overlapping several contigs
Optical/genetic maps
map = ordered set of markers
marker found in a contig
Linked reads
set of barcoded reads amplified from a long fragment
3C/Hi-C
pair of short reads indicating a 3D contact

Figure 7 – Overview of scaffolding methods.

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.

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%

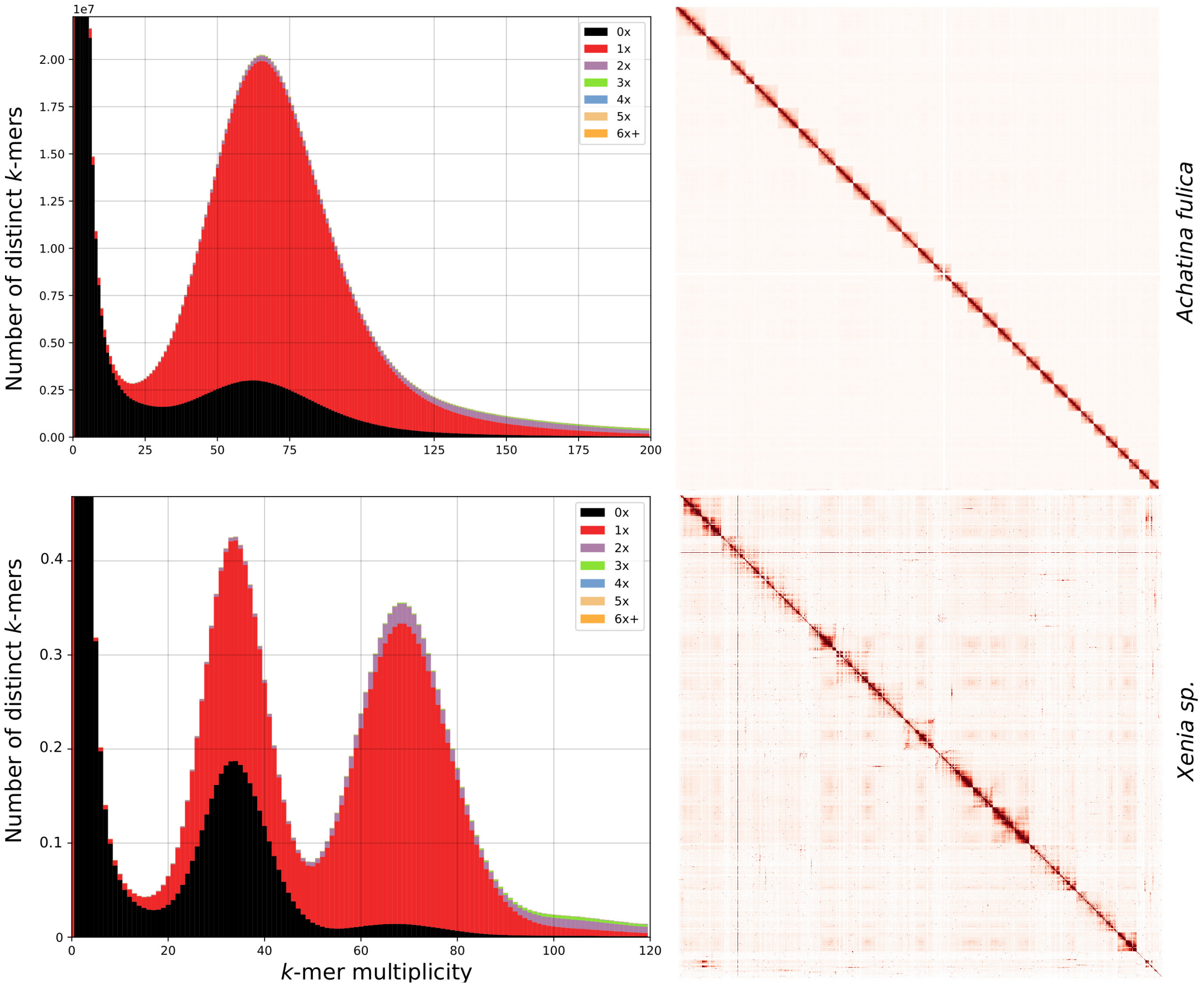

Figure 8 – Assembly evaluation of *Achatina fulica* and *Xenia* sp.. Left: KAT comparison of the *k*-mers in the Illumina datasets v. the assembly. Right: Hi-C contact maps, with a binning of 300 for *Achatina fulica*, 30 for *Xenia* sp..

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 with a *k*-mer multiplicity around 35X and 70X (Figure 8, bottom 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 [279], 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 [280] and BlobToolKit [281] 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 analyses can be run exclusively on sequences that were anchored to the main scaffolds. 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 [282], when these sequences were in fact contaminants [283].

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.

5. 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 [284]. A first approach, called trio-binning, is to assemble one individual using sequencing data from the individual itself and its parents [285]; 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 [286] and WhatsHap [287]. Ideally, genomes should be uncollapsed, as can be done with Bwise [288] and Platanus-Allee [289] using short reads, FALCON-Unzip [115] 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, using hifiasm [129] for instance, thanks to their length and low error rate, and they have already been used to produce phased assemblies of a human [290] and the potato *Solanum tuberosum* [291]. 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 [292]. Phased long-read assemblies can be further polished with adequate programs (e.g. Hapo-G [162]). As Hi-C has demonstrated its efficiency to scaffold haploid assemblies, the principles were further exploited in ALLHiC [293], GraphUnzip [294] and FALCON-Phase [295] 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 assemblies rely on various datasets to prove their correctness (e.g. parental assemblies [290]). Merqury [296] 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.

6. Recommendations

Long reads and Hi-C have become a gold standard for genome assembly and several consortia have adopted this strategy. Ideally, high-accuracy long reads (PacBio HiFi, Nanopore Q20+) should be prefered as they generally yield more contiguous assemblies than low-accuracy long reads, and they improve the resolution of repetitions. HiFi reads also have the advantage that their assembly requires lower computational resources; the computational burden has however shifted to filtering PacBio reads to produce HiFi reads, although this step is usually performed by sequencing providers. More than ten softwares have already been released for or adapted to high-accuracy long reads, and have led to high-quality assemblies, but we can expect that they are not yet able to fully take advantage of these new technologies, and the development of new tools will further elevate the accuracy of *de novo* assemblies. Besides, these reads necessitate an optimisation of high-molecular-weight DNA libraries which is not possible for all non-model species.

Low-accuracy long reads are more accessible, and they have been used to assemble countless reference genomes over the past decade. For low-accuracy PacBio reads, a high coverage depth is sufficient to eliminate errors, due to their random error pattern. Low-accuracy Nanopore reads need to be combined with highly accurate reads to correct or polish their systematic errors. A limiting factor for long-read sequencing is the minimum DNA input. Nanopore reads, necessitate one microgram of high-molecular-weight DNA, and three micrograms are recommended to maximize the output of a flow cell. For PacBio reads, low and ultra-low input protocols are available (for both low- and high-accuracy reads), but they are only adequate for genomes up to 500 Mb. Another factor to weight in when choosing between these reads is their length: with an optimized Nanopore library, reads are typically longer than PacBio reads, and lead to more contiguous assemblies.

When high-molecular-weight DNA cannot be extracted, short reads are the adequate option. The resulting assemblies are more fragmented, yet some short-read assemblers are able to produce good drafts, such as Platanus. These assemblies may have large missing repetitions, thus they are not ideal for analysis of repetitive content and they should be thoroughly assessed in terms of assembly size and completeness.

Hi-C scaffolding has emerged as the most robust method to obtain chromosome-level scaffolds with no contamination. It is applicable as long as fresh or flash-frozen tissue is available for crosslinking, and with a minimum input of a half to one million cells. When these requirements are not fullfilled, linked reads can be used as an alternative (as TELL-seq can use a low input of DNA), or in addition to further reduce assembly errors.

A current issue for non-model species are remaining artefactual duplications in assemblies; these duplications must be identified with BUSCO and *k*-mer analysis tools, and eliminated with haplotig-purging tools prior to scaffolding. However, producing collapsed haploid assemblies is a standard set by genome projects for low-heterozygosity genomes: phasing assemblies may be a better option and a more comprehensive representation of highly heterozygous genomes.

The most crucial step in an assembly pipeline should be the evaluation step. Chromosome-level assemblies are sought for to study structural rearrangements, transposable elements, discard contaminants and compare related species. Genomics consortia have set high standards for quality and contiguity (more than 90% of an assembly anchored to the main scaffolds, BUSCO and *k*-mer completeness superior to 90%), but these goals may not be reached for some difficult species. Imperfect genome assemblies still provide insights into understudied species, as long as their flaws are acknowledged. For instance, fragmented assemblies may be used to identify genes and conduct phylogenomics or population genomics analyses, although the number of genes can be inflated due to their fragmentation [297] and repetitions may be poorly represented. Conclusions should be drawn carefully depending on the quality of the assembly: what would appear as a whole-genome duplication could be the result of large artefactual duplications; contaminants could be erroneously interpreted as horizontal gene transfers.

7. 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) [298]. Assembly pipelines are becoming more reproducible thanks to several initiatives using workflow managers, such as the Vertebrate Genome Project assembly pipeline in Galaxy [299].

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.

Supplementary information

Data presented in Figures 1, 4 and 6 are available in [300]. Tables 1 and 2 are available and will be updated in [301].

Fundings

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) and a complementary fellowship from the David and Alice Van Buuren fund and the Jaumotte-Demoulin foundation.

Acknowledgements

A preprint version of this article has been reviewed and recommended by Peer Community In Genomics (https://doi.org/10.24072/pci.genomics.100016).

Conflict of interest disclosure

The authors of this preprint declare that they have no financial conflict of interest with the content of this article. Romain Koszul is a recommender for PCI Genomics and and Jean-François Flot is a managing board member of PCI Genomics.

References

[1] Edward S Rice and Richard E Green. "New approaches for genome assembly and scaffolding". In: *Annual Review of Animal Biosciences* 7 (2019), pp. 17–40. DOI: 10.1146/annurev-animal-020518-115344.

[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*. 2021.

[4] David A. Morrison. *The Timetree of Life*. Vol. 58. 4. Aug. 2009, pp. 461–462. DOI: 10.1093/sysbio/syp042.

[5] Zhi-Qiang Zhang. "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. Zootaxa, 2013, pp. 1–82. DOI: 10.11646/zootaxa.3703.1.3.

[6] Fei Li et al. "Insect genomes: progress and challenges". In: *Insect Molecular Biology* 28.6 (2019), pp. 739–758. DOI: 10.1111/imb.12599.

[7] Jorge Ari Noriega et al. "Research trends in ecosystem services provided by insects". In: *Basic and Applied Ecology* 26 (2018), pp. 8–23. DOI: 10.1016/j.baae.2017.09.006.

[8] Wenbo Chen 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". In: *BMC Biology* 14.1 (2016), pp. 1–15. DOI: 10.1186/s12915-016-0321-y.

[9] Vishvanath Nene et al. "Genome sequence of *Aedes aegypti*, a major arbovirus vector". In: *Science* 316.5832 (2007), pp. 1718–1723. DOI: 10.1126/science.1138878.

[10] Osvaldo Marinotti et al. "The genome of *Anopheles darlingi*, the main neotropical malaria vector". In: *Nucleic Acids Research* 41.15 (2013), pp. 7387–7400. DOI: 10.1093/nar/gkt484.

[11] Benjamin J. Matthews et al. "Improved reference genome of *Aedes aegypti* informs arbovirus vector control". In: *Nature* 563.7732 (2018), pp. 501–507. DOI: 10.1038/s41586-018-0692-z.

[12] Olga Dudchenko et al. "*De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds". In: *Science* 356.6333 (2017), pp. 92–95. DOI: 10.1126/science.aal3327.

[13] Scott Hotaling, Joanna L. Kelley, and Paul B. Frandsen. "Toward a genome sequence for every animal: Where are we now?" In: *Proceedings of the National Academy of Sciences* 118.52 (2021). DOI: 10.1073/pnas.2109019118.

[14] Anthony R Carroll et al. "Marine natural products". In: *Natural Product Reports* (2021). DOI: 10.1039/C9NP00069K.

[15] Shaden AM Khalifa et al. “Marine natural products: A source of novel anticancer drugs”. In: *Marine Drugs* 17.9 (2019), p. 491. DOI: 10.3390/md17090491.

[16] Tzi Bun Ng et al. “Antibacterial products of marine organisms”. In: *Applied Microbiology and Biotechnology* 99.10 (2015), pp. 4145–4173. DOI: 10.1007/s00253-015-6553-x.

[17] Conxita Avila. “Terpenoids in marine heterobranch molluscs”. In: *Marine Drugs* 18.3 (2020), p. 162. DOI: 10.3390/md18030162.

[18] Bing-Nan Han et al. “Natural Products from Sponges”. In: *Symbiotic Microbiomes of Coral Reefs Sponges and Corals*. Springer Netherlands, 2019, pp. 329–463. DOI: 10.1007/978-94-024-1612-1_15.

[19] Takeshi Takeuchi. “Molluscan genomics: implications for biology and aquaculture”. In: *Current Molecular Biology Reports* 3.4 (2017), pp. 297–305. DOI: 10.1007/s40610-017-0077-3.

[20] Chelse M Prather et al. “Invertebrates, ecosystem services and climate change”. In: *Biological Reviews* 88.2 (2013), pp. 327–348. DOI: 10.1111/brv.12002.

[21] Andre Gomes-dos-Santos et al. “Molluscan genomics: the road so far and the way forward”. In: *Hydrobiologia* 847.7 (2020), pp. 1705–1726. DOI: 10.1007/s10750-019-04111-1.

[22] Nicola Conci, Sergio Vargas, and Gert Wörheide. “The biology and evolution of calcite and aragonite mineralization in octocorallia”. In: *Frontiers in Ecology and Evolution* 9 (2021), p. 81. DOI: 10.3389/fevo.2021.623774.

[23] Melody S Clark. “Molecular mechanisms of biomineralization in marine invertebrates”. In: *Journal of Experimental Biology* 223.11 (2020). DOI: 10.1242/jeb.206961.

[24] Chuya Shinzato et al. “Using the *Acropora digitifera* genome to understand coral responses to environmental change”. In: *Nature* 476.7360 (2011), pp. 320–323. DOI: 10.1038/nature10249.

[25] Yafei Mao, Evan P Economo, and Noriyuki Satoh. “The roles of introgression and climate change in the rise to dominance of Acropora corals”. In: *Current Biology* 28.21 (2018), pp. 3373–3382. DOI: 10.1016/j.cub.2018.08.061.

[26] Zachary L Fuller et al. “Population genetics of the coral *Acropora millepora*: Toward genomic prediction of bleaching”. In: *Science* 369.6501 (2020). DOI: 10.1126/science.aba4674.

[27] Chuya Shinzato et al. “Eighteen coral genomes reveal the evolutionary origin of Acropora strategies to accommodate environmental changes”. In: *Molecular Biology and Evolution* 38.1 (2021), pp. 16–30. DOI: 10.1093/molbev/msaa216.

[28] Paschalia Kapli, Ziheng Yang, and Maximilian J Telford. “Phylogenetic tree building in the genomic age”. In: *Nature Reviews Genetics* 21.7 (2020), pp. 428–444. DOI: 10.1038/s41576-020-0233-0.

[29] E Sally Chang et al. “Genomic insights into the evolutionary origin of Myxozoa within Cnidaria”. In: *Proceedings of the National Academy of Sciences* 112.48 (2015), pp. 14912–14917. DOI: 10.1073/pnas.1511468112.

[30] Michael Eitel et al. “Comparative genomics and the nature of placozoan species”. In: *PLoS Biology* 16 (2018), pp. 1–36. DOI: 10.1371/journal.pbio.2005359.

[31] M. Lynch et al. “The Mutational Meltdown in Asexual Populations”. In: *Journal of Heredity* 84.5 (Sept. 1993), pp. 339–344. DOI: 10.1093/oxfordjournals.jhered.a111354.

[32] John K. Colbourne et al. “The ecoresponsive genome of Daphnia pulex”. In: *Science* 331.6017 (2011), pp. 555–561. DOI: 10.1126/science.1197761.

[33] Zhiqiang Ye et al. “A new reference genome assembly for the microcrustacean *Daphnia pulex*”. In: *G3: Genes, Genomes, Genetics* 7.5 (2017), pp. 1405–1416. DOI: 10.1534/g3.116.038638.

[34] Philipp H. Schiffer et al. “Signatures of the evolution of parthenogenesis and cryptobiosis in the genomes of panagrolaimid nematodes”. In: *iScience* 21 (2019), pp. 587–602. DOI: 10.1016/j.isci.2019.10.039.

[35] Alexander Brandt et al. "Haplotype divergence supports long-term asexuality in the oribatid mite *Oppiella nova*". In: *Proceedings of the National Academy of Sciences* 118.38 (2021). DOI: 10.1073/pnas.2101485118.

[36] Paul Simion et al. "Chromosome-level genome assembly reveals homologous chromosomes and recombination in asexual rotifer *Adineta vaga*". In: *Science Advances* 7.41 (2021). DOI: 10.1126/sciadv.abg4216.

[37] Matthias Buck and Clare Hamilton. "The Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity". In: *Review of European Community & International Environmental Law* 20.1 (2011), pp. 47–61. DOI: 10.1111/j.1467-9388.2011.00703.x.

[38] Felipe A Simão et al. "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs". In: *Bioinformatics* 31.19 (2015), pp. 3210–3212. DOI: 10.1093/bioinformatics/btv351.

[39] Robert M Waterhouse et al. "BUSCO applications from quality assessments to gene prediction and phylogenomics". In: *Molecular Biology and Evolution* 35.3 (2018), pp. 543–548. DOI: 10.1093/molbev/msx319.

[40] Mosè Manni et al. "BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes". In: *Molecular Biology and Evolution* 38.10 (2021), pp. 4647–4654. DOI: 10.1093/molbev/msab199.

[41] Stephen F Altschul et al. "Basic local alignment search tool". In: *Journal of Molecular Biology* 215.3 (1990), pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

[42] GIGA Community of Scientists. "The Global Invertebrate Genomics Alliance (GIGA): developing community resources to study diverse invertebrate genomes". In: *Journal of Heredity* 105.1 (2014), pp. 1–18. DOI: 10.1093/jhered/est084.

[43] Christian R. Voolstra et al. "Advancing genomics through the Global Invertebrate Genomics Alliance (GIGA)". In: *Invertebrate Systematics* 31.1 (2017), p. 1. DOI: 10.1071/is16059.

[44] Harris A Lewin et al. "Earth BioGenome Project: Sequencing life for the future of life". In: *Proceedings of the National Academy of Sciences* 115.17 (2018), pp. 4325–4333. DOI: 10.1073/pnas.1720115115.

[45] Darwin Tree of Life. *Darwin Tree of Life*, *www.darwintreeoflife.org*. 2021.

[46] Aquatic Symbiosis Genomics Project. *Aquatic Symbiosis Genomics Project*, *www.sanger.ac.uk/collaboration/aquatic-symbiosis-genomics-project*. 2021.

[47] Giulio Formenti et al. "The era of reference genomes in conservation genomics". In: *Trends in Ecology & Evolution* (2022). DOI: 10.1016/j.tree.2021.11.008.

[48] Frederick Sanger, Steven Nicklen, and Alan R Coulson. "DNA sequencing with chain-terminating inhibitors." In: *Proceedings of the National Academy of Sciences* 74.12 (1977), pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463.

[49] André Goffeau et al. "Life with 6000 genes". In: *Science* 274.5287 (1996), pp. 546–567. DOI: 10.1126/science.274.5287.546.

[50] *C. elegans* Sequencing Consortium. "Genome sequence of the nematode *C. elegans*: a platform for investigating biology". In: *Science* 282.5396 (1998), pp. 2012–2018. DOI: 10.1126/science.282.5396.2012.

[51] Bilal Wajid et al. "The A, C, G, and T of genome assembly". In: *BioMed Research International* 2016 (May 2016), pp. 1–10. DOI: 10.1155/2016/6329217.

[52] International Human Genome Sequencing Consortium and others. "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822 (2001), pp. 860–921. DOI: 10.1038/35057062.

[53] Mihai Pop and Steven L Salzberg. "Bioinformatics challenges of new sequencing technology". In: *Trends in Genetics* 24.3 (2008), pp. 142–149. DOI: 10.1016/j.tig.2007.12.006.

[54] Marcel Margulies et al. "Genome sequencing in microfabricated high-density picolitre reactors". In: *Nature* 437.7057 (2005), pp. 376–380. DOI: 10.1038/nature03959.

[55] Jonathan M Rothberg et al. "An integrated semiconductor device enabling non-optical genome sequencing". In: *Nature* 475.7356 (2011), pp. 348–352. DOI: 10.1038/nature10242.

[56] Kevin Judd McKernan et al. "Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding". In: *Genome Research* 19.9 (2009), pp. 1527–1541. DOI: 10.1101/gr.091868.109.

[57] Michael L Metzker. "Sequencing technologies — the next generation". In: *Nature Reviews Genetics* 11.1 (2010), pp. 31–46. DOI: 10.1038/nrg2626.

[58] David R. Bentley et al. "Accurate whole human genome sequencing using reversible terminator chemistry". In: *Nature* 456.7218 (2008), pp. 53–59. DOI: 10.1038/nature07517.

[59] Martin O. Pollard et al. "Long reads: their purpose and place". In: *Human Molecular Genetics* 27.R2 (2018), R234–R241. DOI: 10.1093/hmg/ddy177.

[60] John Eid et al. "Real-time DNA sequencing from single polymerase molecules". In: *Science* 323.5910 (2009), pp. 133–138. DOI: 10.1126/science.1162986.

[61] Aaron M Wenger et al. "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome". In: *Nature Biotechnology* 37.October (2019), pp. 1155–1162. DOI: 10.1038/s41587-019-0217-9.

[62] David Deamer, Mark Akeson, and Daniel Branton. "Three decades of Nanopore sequencing". In: *Nature Biotechnology* 34.5 (2016), pp. 518–524. DOI: 10.1038/nbt.3423.

[63] Miten Jain et al. "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community". In: *Genome Biology* 17.1 (2016), pp. 1–11. DOI: 10.1186/s13059-016-1103-0.

[64] Ryan R. Wick, Louise M. Judd, and Kathryn E. Holt. "Performance of neural network basecalling tools for Oxford Nanopore sequencing". In: *Genome Biology* 20.1 (2019), pp. 1–10. DOI: 10.1186/s13059-019-1727-y.

[65] Miten Jain et al. "Nanopore sequencing and assembly of a human genome with ultra-long reads". In: *Nature Biotechnology* 36.4 (2018), pp. 338–345. DOI: 10.1038/nbt.4060.

[66] *Bonito*, *https://github.com/nanoporetech/bonito*.

[67] Jordi Silvestre-Ryan. *Poreover*, *https://github.com/jordisr/poreover*. 2017.

[68] Mantas Sereika et al. "Oxford Nanopore R10.4 long-read sequencing enables near-perfect bacterial genomes from pure cultures and metagenomes without short-read or reference polishing". In: *bioRxiv* (2021). DOI: 10.1101/2021.10.27.466057.

[69] Pacharaporn Angthong et al. "Optimization of high molecular weight DNA extraction methods in shrimp for a long-read sequencing platform". In: *PeerJ* 8 (2020), e10340. DOI: 10.7717/peerj.10340.

[70] Serafim Batzoglou et al. "ARACHNE: a whole-genome shotgun assembler". In: *Genome Research* 12.1 (2002), pp. 177–189. DOI: 10.1101/gr.208902.

[71] Paul Havlak et al. "The Atlas genome assembly system". In: *Genome Research* 14.4 (2004), pp. 721–732. DOI: 10.1101/gr.2264004.

[72] Xiaoqiu Huang and Anup Madan. "CAP3: a DNA sequence assembly program". In: *Genome Research* 9.9 (1999), pp. 868–877. DOI: 10.1101/gr.9.9.868.

[73] Gennady Denisov et al. "Consensus generation and variant detection by Celera Assembler". In: *Bioinformatics* 24.8 (2008), pp. 1035–1040. DOI: 10.1093/bioinformatics/btn074.

[74] Pavel A Pevzner, Haixu Tang, and Michael S Waterman. "An Eulerian path approach to DNA fragment assembly". In: *Proceedings of the National Academy of Sciences* 98.17 (2001), pp. 9748–9753. DOI: 10.1073/pnas.171285098.

[75] Samuel Aparicio et al. "Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*". In: *Science* 297.5585 (2002), pp. 1301–1310. DOI: 10.1126/science.1072104.

[76] Daniel D Sommer et al. "Minimus: a fast, lightweight genome assembler". In: *BMC Bioinformatics* 8.1 (2007), pp. 1–11. DOI: 10.1186/1471-2105-8-64.

[77] Bastien Chevreux, Thomas Wetter, Sándor Suhai, et al. "Genome sequence assembly using trace signals and additional sequence information." In: *German Conference on Bioinformatics*. Vol. 99. Citeseer. 1999, pp. 45–56.

[78] Brent Ewing and Phil Green. "Base-calling of automated sequencer traces using Phred. II. Error probabilities". In: *Genome Research* 8.3 (1998), pp. 186–194. DOI: 10.1101/gr.8.3.186.

[79] James C Mullikin and Zemin Ning. "The Phusion assembler". In: *Genome Research* 13.1 (2003), pp. 81–90. DOI: 10.1101/gr.731003.

[80] Giuseppe Narzisi and Bud Mishra. "Scoring-and-unfolding trimmed tree assembler: concepts, constructs and comparisons". In: *Bioinformatics* 27.2 (Nov. 2010), pp. 153–160. DOI: 10.1093/bioinformatics/btq646.

[81] Granger Sutton et al. "TIGR Assembler: A new tool for assembling large shotgun projects". In: *Genome Science and Technology* 1.1 (1995), pp. 9–19. DOI: 10.1089/gst.1995.1.9.

[82] Jared T. Simpson et al. "ABySS: A parallel assembler for short read sequence data". In: *Genome Research* 19.6 (2009), pp. 1117–1123. DOI: 10.1101/gr.089532.108.

[83] Shaun D Jackman et al. "ABySS 2.0: resource-Efficient Assembly of Large Genomes using a Bloom Filter Effect of Bloom Filter False Positive Rate". In: *Genome Research* 27 (2017), pp. 768–777. DOI: 10.1101/gr.214346.116.

[84] Jonathan Butler et al. "ALLPATHS: *de novo* assembly of whole-genome shotgun microreads". In: *Genome Research* 18.5 (2008), pp. 810–820. DOI: 10.1101/gr.7337908.

[85] Binghang Liu et al. "BASE: a practical *de novo* assembler for large genomes using long NGS reads". In: *BMC Genomics* 17.5 (2016), pp. 561–569. DOI: 10.1186/s12864-016-2829-5.

[86] Jason R. Miller et al. "Aggressive assembly of pyrosequencing reads with mates". In: *Bioinformatics* 24.24 (2008), pp. 2818–2824. DOI: 10.1093/bioinformatics/btn548.

[87] David Hernandez et al. "*De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer". In: *Genome Research* 18.5 (2008), pp. 802–809. DOI: 10.1101/gr.072033.107.

[88] Junwei Luo et al. "EPGA: *de novo* assembly using the distributions of reads and insert size". In: *Bioinformatics* 31.6 (2015), pp. 825–833. DOI: 10.1093/bioinformatics/btu762.

[89] Mark J. Chaisson and Pavel A. Pevzner. "Short read fragment assembly of bacterial genomes". In: *Genome Research* 18.2 (2008), pp. 324–330. DOI: 10.1101/gr.7088808.

[90] Thomas Conway et al. "Gossamer — a resource-efficient *de novo* assembler". In: *Bioinformatics* 28.14 (2012), pp. 1937–1938. DOI: 10.1093/bioinformatics/bts297.

[91] Yu Peng et al. "IDBA - a practical iterative De Bruijn graph *de novo* assembler". In: *Research in Computational Molecular Biology* 6044 LNBI (2010), pp. 426–440. DOI: 10.1007/978-3-642-12683-3_28.

[92] Min Li et al. "ISEA: Iterative seed-extension algorithm for *de novo* assembly using paired-end information and insert size distribution". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 14.4 (2016), pp. 916–925. DOI: 10.1109/TCBB.2016.2550433.

[93] Te-Chin Chu et al. "Assembler for *de novo* assembly of large genomes". In: *Proceedings of the National Academy of Sciences* 110.36 (2013), E3417–E3424. DOI: 10.1073/pnas.1314090110.

[94] Sara El-Metwally, Magdi Zakaria, and Taher Hamza. "LightAssembler: fast and memory-efficient assembly algorithm for high-throughput sequencing reads". In: *Bioinformatics* 32.21 (2016), pp. 3215–3223. DOI: 10.1093/bioinformatics/btw470.

[95] Jarrod A Chapman et al. "Meraculous: *de novo* genome assembly with short paired-end reads". In: *PloS One* 6.8 (2011), e23501. DOI: 10.1371/journal.pone.0023501.

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

[97] Xiaoqiu Huang et al. "PCAP: a whole-genome assembly program". In: *Genome Research* 13.9 (2003), pp. 2164–2170. DOI: 10.1101/gr.1390403.

[98] Xiao Zhu et al. "PERGA: a paired-end read guided *de novo* assembler for extending contigs using SVM and look ahead approach". In: *PloS One* 9.12 (2014), e114253. DOI: 10.1371/journal.pone.0114253.

[99] Rei Kajitani et al. "Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads". In: *Genome Research* 24.8 (2014), pp. 1384–1395. DOI: 10.1101/gr.170720.113.

[100] Pramila Nuwantha Ariyaratne and Wing-Kin Sung. "PE-Assembler: *de novo* assembler using short paired-end reads". In: *Bioinformatics* 27.2 (2011), pp. 167–174. DOI: 10.1093/bioinformatics/btq626.

[101] Douglas W. Bryant, Weng-Keen Wong, and Todd C. Mockler. "QSRA – a quality-value guided *de novo* short read assembler". In: *BMC Bioinformatics* 10.69 (2009), pp. 1–6. DOI: 10.1186/1471-2105-10-69.

[102] Sébastien Boisvert, François Laviolette, and Jacques Corbeil. "Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies". In: *Journal of Computational Biology* 17.11 (2010), pp. 1519–1533. DOI: 10.1089/cmb.2009.0238.

[103] Giorgio Gonnella and Stefan Kurtz. "Readjoiner: a fast and memory efficient string graph-based sequence assembler". In: *BMC Bioinformatics* 13.82 (2012), pp. 1–19. DOI: 10.1186/1471-2105-13-82.

[104] Jared T Simpson and Richard Durbin. "Efficient *de novo* assembly of large genomes using compressed data structures". In: *Genome Research* 22.3 (2012), pp. 549–556. DOI: 10.1101/gr.126953.111.

[105] Juliane C Dohm et al. "SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing". In: *Genome Research* 17.11 (2007), pp. 1697–1706. DOI: 10.1101/gr.6435207.

[106] Ruiqiang Li et al. "*De novo* assembly of human genomes with massively parallel short read sequencing". In: *Genome Research* 20.2 (2010), pp. 265–272. DOI: 10.1101/gr.097261.109.

[107] Ruibang Luo et al. "SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler". In: *GigaScience* 1.18 (2012), pp. 1–6. DOI: 10.1186/2047-217X-1-18.

[108] Anton Bankevich et al. "SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing". In: *Journal of Computational Biology* 19.5 (2012), pp. 455–477. DOI: 10.1089/cmb.2012.0021.

[109] Chengxi Ye et al. "Exploiting sparseness in *de novo* genome assembly". In: *BMC Bioinformatics*. Vol. 13. S1. BioMed Central. 2012. DOI: 10.1186/1471-2105-13-S6-S1.

[110] René L. Warren et al. "Assembling millions of short DNA sequences using SSAKE". In: *Bioinformatics* 23.4 (2007), pp. 500–501. DOI: 10.1093/bioinformatics/btl629.

[111] Bertil Schmidt et al. "A fast hybrid short read fragment assembly algorithm". In: *Bioinformatics* 25.17 (2009), pp. 2279–2280. DOI: 10.1093/bioinformatics/btp374.

[112] William R. Jeck et al. "Extending assembly of short DNA sequences to handle error". In: *Bioinformatics* 23.21 (2007), pp. 2942–2944. DOI: 10.1093/bioinformatics/btm451.

[113] Daniel R. Zerbino. "Using the Velvet *de novo* assembler for short-read sequencing technologies". In: *Current Protocols in Bioinformatics* Chapter 11 (2010), pp. 1–12. DOI: 10.1002/0471250953.bi1105s31.

[114] Sergey Koren et al. "Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation". In: *Genome Research* 27.5 (2017), pp. 722–736. DOI: 10.1101/gr.215087.116.

[115] Chen-Shan Chin et al. "Phased diploid genome assembly with single-molecule real-time sequencing". In: *Nature Methods* 13.12 (2016), pp. 1050–1054. DOI: 10.1038/nmeth.4035.

[116] Mikhail Kolmogorov et al. "Assembly of long, error-prone reads using repeat graphs". In: *Nature Biotechnology* 37.5 (2019), pp. 540–546. DOI: 10.1038/s41587-019-0072-8.

[117] Govinda M Kamath et al. "HINGE: long-read assembly achieves optimal repeat resolution". In: *Genome Research* 27.5 (2017), pp. 747–756. DOI: 10.1101/gr.216465.116.

[118] Chuan Le Xiao et al. "MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads". In: *Nature Methods* 14.11 (2017), pp. 1072–1074. DOI: 10.1038/nmeth.4432.

[119] Heng Li. "Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences". In: *Bioinformatics* 32.14 (2016), pp. 2103–2110. DOI: 10.1093/bioinformatics/btw152.

[120] Ying Chen et al. "Efficient assembly of nanopore reads via highly accurate and intact error correction". In: *Nature Communications* 12.60 (2021), pp. 1–10. DOI: 10.1038/s41467-020-20236-7.

[121] NextOmics. *NextDenovo*, *https://github.com/Nextomics/NextDenovo*. 2019.

[122] Robert Vaser and Mile Šikić. "Yet another *de novo* genome assembler". In: *International Symposium on Image and Signal Processing and Analysis, ISPA* (2019), pp. 147–151. DOI: 10.1109/ISPA.2019.8868909.

[123] Robert Vaser and Mile Šikić. "Time-and memory-efficient genome assembly with Raven". In: *Nature Computational Science* 1.5 (2021), pp. 332–336. DOI: 10.1038/s43588-021-00073-4.

[124] Kishwar Shafin et al. "Nanopore sequencing and the Shasta toolkit enable efficient *de novo* assembly of eleven human genomes". In: *Nature Biotechnology* 38.9 (2020), pp. 1044–1053. DOI: 10.1038/s41587-020-0503-6.

[125] Hailin Liu et al. "SMARTdenovo: a *de novo* assembler using long noisy reads". In: *Gigabyte* 2021 (2021), pp. 1–9. DOI: 10.46471/gigabyte.15.

[126] Jue Ruan. *wtdbg*, *https://github.com/ruanjue/wtdbg*. 2016.

[127] Jue Ruan and Heng Li. "Fast and accurate long-read assembly with wtdbg2". en. In: *Nature Methods* 17.2 (2020), pp. 155–158. DOI: 10.1038/s41592-019-0669-3.

[128] Sergey Nurk et al. "HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads". In: *Genome Research* 30.9 (2020), pp. 1291–1305. DOI: 10.1101/gr.263566.120.

[129] Haoyu Cheng et al. "Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm". In: *Nature Methods* 18.2 (2021), pp. 1–6. DOI: 10.1038/s41592-020-01056-5.

[130] PacificBiosciences. *IPA*, *https://github.com/PacificBiosciences/pbbioconda*. 2018.

[131] Anton Bankevich et al. "LJA: Assembling long and accurate reads using multiplex de Bruijn graphs". In: *bioRxiv* (2021). DOI: 10.1101/2020.12.10.420448.

[132] Barış Ekim, Bonnie Berger, and Rayan Chikhi. "Minimizer-space de Bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer". In: *Cell Systems* 12.10 (2021), pp. 958–968. DOI: 10.1016/j.cels.2021.08.009.

[133] Mikko Rautiainen and Tobias Marschall. "MBG: Minimizer-based sparse de Bruijn graph construction". In: *Bioinformatics* 37.16 (2021), pp. 2476–2478. DOI: 10.1093/bioinformatics/btab004.

[134] Chen-Shan Chin and Asif Khalak. "Human genome assembly in 100 minutes". In: *bioRxiv* (2019). DOI: 10.1101/705616.

[135] Jorma Tarhio and Esko Ukkonen. "A greedy approximation algorithm for constructing shortest common superstrings". In: *Theoretical Computer Science* 57.1 (1988), pp. 131–145. DOI: 10.1016/0304-3975(88)90167-3.

[136] Nicolas Dierckxsens, Patrick Mardulyn, and Guillaume Smits. "NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data". In: *Nucleic Acids Research* 45.4 (2017), e18. DOI: 10.1093/nar/gkw955.

[137] Rodger Staden. "A strategy of DNA sequencing employing computer programs". In: *Nucleic Acids Research* 6.7 (1979), pp. 2601–2610. DOI: 10.1093/nar/6.7.2601.

[138] J Craig Venter et al. "The Sequence of the Human Genome". In: *Science* 291.5507 (2001), pp. 1304–1351. DOI: 10.1126/science.1058040.

[139] Nicolass Govert de Bruijn. "A Combinatorial Problem". In: *Koninklijke Nederlandse Akademie v. Wetenschappen* 49 (1946), pp. 758–764.

[140] Camille Flye Sainte-Marie. "48". In: *L'Intermédiaire des Mathématiciens* 1 (1894), pp. 107–110.

[141] Phillip E.C. Compeau, Pavel A. Pevzner, and Glenn Tesler. "How to apply de Bruijn graphs to genome assembly". In: *Nature Biotechnology* 29.11 (2011), pp. 987–991. DOI: 10.1038/nbt.2023.

[142] Dayana Yahalomi et al. "A cnidarian parasite of salmon (Myxozoa: Henneguya) lacks a mitochondrial genome". In: *Proceedings of the National Academy of Sciences* 117.10 (2020), pp. 5358–5363. DOI: 10.1073/pnas.1909907117.

[143] Matthias C. Vogg et al. "An evolutionarily-conserved Wnt3/β-catenin/Sp5 feedback loop restricts head organizer activity in Hydra". In: *Nature Communications* 10.1 (2019), pp. 1–15. DOI: 10.1038/s41467-018-08242-2.

[144] Ellen M. Leffler et al. "Revisiting an old riddle: what determines genetic diversity levels within species?" In: *PLoS Biology* 10.9 (2012), e1001388. DOI: 10.1371/journal.pbio.1001388.

[145] Ryan R. Wick. *Filtlong*, *https://github.com/rrwick/Filtlong*. 2017.

[146] Ehsan Haghshenas et al. "CoLoRMap: Correcting Long Reads by Mapping short reads". In: *Bioinformatics* 32.17 (2016), pp. i545–i551. DOI: 10.1093/bioinformatics/btw463.

[147] Can Firtina et al. "Hercules: a profile HMM-based hybrid error correction algorithm for long reads". In: *Nucleic Acids Research* 46.21 (2018), e125. DOI: 10.1093/nar/gky724.

[148] Pierre Morisse, Thierry Lecroq, and Arnaud Lefebvre. "Hybrid correction of highly noisy long reads using a variable-order de Bruijn graph". In: *Bioinformatics* 34.24 (June 2018), pp. 4213–4222. DOI: 10.1093/bioinformatics/bty521.

[149] Gilles Miclotte et al. "Jabba: hybrid error correction for long sequencing reads." In: *Algorithms for Molecular Biology* 11.10 (2016), pp. 1–12. DOI: 10.1186/s13015-016-0075-7.

[150] Leena Salmela and Eric Rivals. "LoRDEC: Accurate and efficient long read error correction". In: *Bioinformatics* 30.24 (2014), pp. 3506–3514. DOI: 10.1093/bioinformatics/btu538.

[151] Leena Salmela et al. "Accurate self-correction of errors in long reads using de Bruijn graphs". In: *Bioinformatics* 33.6 (2017), pp. 799–806. DOI: 10.1093/bioinformatics/btw321.

[152] Mohammed-Amin Madoui et al. "Genome assembly using Nanopore-guided long and error-free DNA reads". In: *BMC Genomics* 16.1 (2015), pp. 1–11. DOI: 10.1186/s12864-015-1519-z.

[153] Thomas Hackl et al. "proovread: large-scale high-accuracy PacBio correction through iterative short read consensus." In: *Bioinformatics* 30.21 (2014), pp. 3004–3011. DOI: 10.1093/bioinformatics/btu392.

[154] Pierre Morisse et al. "Scalable long read self-correction and assembly polishing with multiple sequence alignment". In: *Scientific Reports* 11.761 (2021), pp. 1–13. DOI: 10.1038/s41598-020-80757-5.

[155] German Tischler and Eugene W. Myers. "Non Hybrid Long Read Consensus Using Local De Bruijn Graph Assembly". In: *bioRxiv* (2017). DOI: 10.1101/106252.

[156] Ergude Bao et al. "FLAS: fast and high-throughput algorithm for PacBio long-read self-correction." In: *Bioinformatics* 35.20 (2019), pp. 3953–3960. DOI: 10.1093/bioinformatics/btz206.

[157] Ergude Bao and Lingxiao Lan. "HALC: High throughput algorithm for long read error correction". In: *BMC Bioinformatics* 18.204 (2017), pp. 1–12. DOI: 10.1186/s12859-017-1610-3.

[158] René L Warren et al. "ntEdit: scalable genome sequence polishing". In: *Bioinformatics* 35.21 (2019), pp. 4430–4432. DOI: 10.1093/bioinformatics/btz400.

[159] Bruce J Walker et al. "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement". In: *PloS one* 9.11 (2014), e112963. DOI: 10.1371/journal.pone.0112963.

[160] Aleksey V. Zimin and Steven L. Salzberg. "The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies". In: *PLOS Computational Biology* 16.6 (June 2020), pp. 1–8. DOI: 10.1371/journal.pcbi.1007981.

[161] Can Firtina et al. “Apollo: a sequencing-technology-independent, scalable, and accurate assembly polishing algorithm”. In: *Bioinformatics* (2020). DOI: 10.1093/bioinformatics/btaa179.

[162] Jean-Marc Aury and Benjamin Istace. “Hapo-G, haplotype-aware polishing of genome assemblies”. In: *NAR Genomics and Bioinformatics* 3.2 (May 2021). DOI: 10.1093/nargab/lqab034.

[163] Wing-kin Sung Ritu Kundu, Joshua Casey. “HyPo : super fast & accurate polisher for long read assemblies”. In: *bioRxiv* (2019). DOI: 10.1101/2019.12.19.882506.

[164] Robert Vaser et al. “Fast and accurate *de novo* genome assembly from long uncorrected reads”. In: *Genome Research* 27.5 (2017), pp. 737–746. DOI: 10.1101/gr.214270.116.

[165] PacificBiosciences. *GenomicConsensus*, *https://github.com/PacificBiosciences/GenomicConsensus*. 2014.

[166] Oxford Nanopore Technologies. *Medaka*, *https://github.com/nanoporetech/medaka*. 2017.

[167] Jiang Hu et al. “NextPolish: a fast and efficient genome polishing tool for long-read assembly”. In: *Bioinformatics* 36.7 (2019), pp. 2253–2255. DOI: 10.1093/bioinformatics/btz891.

[168] Jared Simpson. *Nanopolish*, *https://github.com/jts/nanopolish*. 2014.

[169] Shengfeng Huang, Mingjing Kang, and Anlong Xu. “HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly”. In: *Bioinformatics* 33.16 (2017), pp. 2577–2579. DOI: 10.1093/bioinformatics/btx220.

[170] Dengfeng Guan et al. “Identifying and removing haplotypic duplication in primary genome assemblies”. In: *Bioinformatics* 36.9 (2020), pp. 2896–2898. DOI: 10.1093/bioinformatics/btaa025.

[171] Michael J Roach, Simon A Schmidt, and Anthony R Borneman. “Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies”. In: *BMC Bioinformatics* 19.1 (2018), pp. 1–10. DOI: 10.1186/s12859-018-2485-7.

[172] Mihai Pop, Daniel S Kosack, and Steven L Salzberg. “Hierarchical scaffolding with Bambus”. In: *Genome Research* 14.1 (2004), pp. 149–159. doi: 10.1101/gr.1536204.

[173] Igor Mandric and Alex Zelikovsky. “Solving scaffolding problem with repeats”. In: *bioRxiv* (2018). doi: 10.1101/330472.

[174] Kristoffer Sahlin et al. “BESST - Efficient scaffolding of large fragmented assemblies”. In: *BMC Bioinformatics* 15.281 (2014), pp. 1–11. doi: 10.1186/1471-2105-15-281.

[175] Junwei Luo et al. “BOSS: a novel scaffolding algorithm based on an optimized scaffold graph”. In: *Bioinformatics* 33.2 (2017), pp. 169–176. doi: 10.1093/bioinformatics/btw597.

[176] Alexey A Gritsenko et al. “GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies”. In: *Bioinformatics* 28.11 (2012), pp. 1429–1437. doi: 10.1093/bioinformatics/bts175.

[177] Leena Salmela et al. “Fast scaffolding with small independent mixed integer programs”. In: *Bioinformatics* 27.23 (2011), pp. 3259–3265. doi: 10.1093/bioinformatics/btr562.

[178] Song Gao, Wing-Kin Sung, and Niranjan Nagarajan. “Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences”. In: *Journal of Computational Biology* 18.11 (2011), pp. 1681–1691. doi: 10.1089/cmb.2011.0170.

[179] Igor Mandric and Alex Zelikovsky. “ScaffMatch: scaffolding algorithm based on maximum weight matching”. In: *Bioinformatics* 31.16 (2015), pp. 2632–2638. doi: 10.1093/bioinformatics/btv211.

[180] Paul M Bodily et al. “ScaffoldScaffolder: solving contig orientation via bidirected to directed graph reduction”. In: *Bioinformatics* 32.1 (2016), pp. 17–24. doi: 10.1093/bioinformatics/btv548.

[181] Nilgun Donmez and Michael Brudno. “SCARPA: scaffolding reads with practical algorithms”. In: *Bioinformatics* 29.4 (2013), pp. 428–434. doi: 10.1093/bioinformatics/bts716.

[182] Min Li et al. “SCOP: a novel scaffolding algorithm based on contig classification and optimization”. In: *Bioinformatics* 35.7 (2019), pp. 1142–1150. doi: 10.1093/bioinformatics/bty773.

[183] Rajat S Roy et al. “SLIQ: Simple Linear Inequalities for Efficient Contig Scaffolding”. In: *Journal of Computational Biology* 19.10 (2012), pp. 1162–1175. doi: 10.1089/cmb.2011.0263.

[184] Adel Dayarian, Todd P Michael, and Anirvan M Sengupta. “SOPRA: Scaffolding algorithm for paired reads via statistical optimization”. In: *BMC Bioinformatics* 11.345 (2010), pp. 1–21. doi: 10.1186/1471-2105-11-345.

[185] Marten Boetzer et al. “Scaffolding pre-assembled contigs using SSPACE”. In: *Bioinformatics* 27.4 (2011), pp. 578–579. doi: 10.1093/bioinformatics/btq683.

[186] Gregory K Farrant et al. “WiseScaffolder: an algorithm for the semi-automatic scaffolding of next generation sequencing data”. In: *BMC Bioinformatics* 16.281 (2015), pp. 1–13. doi: 10.1186/s12859-015-0705-y.

[187] Arne Ludwig et al. “DENTIST — using long reads for closing assembly gaps at high accuracy”. In: *GigaScience* 11 (Jan. 2022). doi: 10.1093/gigascience/giab100.

[188] Stephan Schmeing and Mark D. Robinson. “Gapless provides combined scaffolding, gap filling and assembly correction with long reads”. In: *bioRxiv* (2022). doi: 10.1101/2022.03.08.483466.

[189] René L Warren et al. “LINKS: scalable, alignment-free scaffolding of draft genomes with long reads”. In: *GigaScience* 4.35 (2015). doi: 10.1186/s13742-015-0076-3.

[190] Mao Qin et al. “LRScaf: improving draft genomes using long noisy reads”. In: *BMC Genomics* 20.955 (2019), pp. 1–12. doi: 10.1186/s12864-019-6337-2.

[191] Minh Duc Cao et al. “Scaffolding and completing genome assemblies in real-time with nanopore sequencing”. In: *Nature Communications* 8.14515 (2017), pp. 1–10. doi: 10.1038/ncomms14515.

[192] Adam C English et al. “Mind the Gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology”. In: *PloS One* 7.11 (2012), e47768. doi: 10.1371/journal.pone.0047768.

[193] Rene L Warren. “RAILS and Cobbler: Scaffolding and automated finishing of draft genomes using long DNA sequences”. In: *Journal of Open Source Software* 1.7 (2016), p. 116. doi: 10.21105/joss.00116.

[194] Junwei Luo et al. “SLR: a scaffolding algorithm based on long reads and contig classification”. In: *BMC Bioinformatics* 20.539 (2019), pp. 1–11. doi: 10.1186/s12859-019-3114-9.

[195] Wellcome Sanger Institute. *SMIS*, *https://www.sanger.ac.uk/tool/smis/*. 2015.

[196] Shenglong Zhu, Danny Z Chen, and Scott J Emrich. “Single molecule sequencing-guided scaffolding and correction of draft assemblies”. In: *BMC Genomics* 18.10 (2017), pp. 51–59. doi: 10.1186/s12864-017-4271-8.

[197] Marten Boetzer and Walter Pirovano. “SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information”. In: *BMC Bioinformatics* 15.211 (2014), pp. 1–9. doi: 10.1186/1471-2105-15-211.

[198] Haibao Tang et al. “ALLMAPS: robust scaffold ordering based on multiple maps”. In: *Genome Biology* 16.3 (2015), pp. 1–15. doi: 10.1186/s13059-014-0573-1.

[199] Henry C Lin et al. “AGORA: assembly guided by optical restriction alignment”. In: *BMC Bioinformatics* 13.189 (2012), pp. 1–14. doi: 10.1186/1471-2105-13-189.

[200] Benjamin Istace, Caroline Belser, and Jean-Marc Aury. “BiSCoT: improving large eukaryotic genome assemblies with optical maps”. In: *PeerJ* 8 (2020), e10150. doi: 10.7717/peerj.10150.

[201] Weihua Pan, Tao Jiang, and Stefano Lonardi. “OMGS: optical map-based genome scaffolding”. In: *Journal of Computational Biology* 27.4 (2020), pp. 519–533. doi: 10.1089/cmb.2019.0310.

[202] Jennifer M Shelton et al. “Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool”. In: *BMC Genomics* 16.734 (2015). doi: 10.1186/s12864-015-1911-8.

[203] Niranjan Nagarajan, Timothy D Read, and Mihai Pop. “Scaffolding and validation of bacterial genome assemblies using optical restriction maps”. In: *Bioinformatics* 24.10 (2008), pp. 1229–1235. doi: 10.1093/bioinformatics/btn102.

[204] Markus Hiltunen, Martin Ryberg, and Hanna Johannesson. “ARBitR: an overlap-aware genome assembly scaffolder for linked reads”. In: *Bioinformatics* 37 (2020), pp. 2203–2205. doi: 10.1093/bioinformatics/btaa975.

[205] Volodymyr Kuleshov, Michael P Snyder, and Serafim Batzoglou. “Genome assembly from synthetic long read clouds”. In: *Bioinformatics* 32.12 (2016), pp. i216–i224. doi: 10.1093/bioinformatics/btw267.

[206] Sarah Yeo et al. “ARCS: scaffolding genome drafts with linked reads”. In: *Bioinformatics* 34.5 (2018), pp. 725–731. doi: 10.1093/bioinformatics/btx675.

[207] Lauren Coombe et al. “ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers”. In: *BMC Bioinformatics* 19.234 (2018). doi: 10.1186/s12859-018-2243-x.

[208] Andrew Adey et al. “*In vitro*, long-range sequence information for *de novo* genome assembly via transposase contiguity”. In: *Genome Research* 24.12 (2014), pp. 2041–2049. doi: 10.1101/gr.178319.114.

[209] Wellcome Sanger Institute. *Scaff10X*, *https://github.com/wtsi-hpag/Scaff10X*. 2018.

[210] Noam Kaplan and Job Dekker. “High-throughput genome scaffolding from *in vivo* DNA interaction frequency”. In: *Nature Biotechnology* 31.12 (2013), pp. 1143–1147. doi: 10.1038/nbt.2768.

[211] Hervé Marie-Nelly et al. “High-quality genome (re)assembly using chromosomal contact data”. In: *Nature Communications* 5.5695 (2014). doi: 10.1038/ncomms6695.

[212] Gina Renschler et al. “Hi-C guided assemblies reveal conserved regulatory topologies on X and autosomes despite extensive genome shuffling”. In: *Genes & Development* 33.21-22 (2019), pp. 1591–1612. doi: 10.1101/gad.328971.119.

[213] Lyam Baudry et al. “instaGRAAL: chromosome-level quality scaffolding of genomes using a proximity ligation-based scaffolder”. In: *Genome Biology* 21.148 (2020). doi: 10.1186/s13059-020-02041-z.

[214] Joshua N Burton et al. “Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions”. In: *Nature Biotechnology* 31.12 (2013), pp. 1119–1125. doi: 10.1038/nbt.2727.

[215] Dengfeng Guan et al. “Efficient iterative Hi-C scaffolder based on N-best neighbors”. In: *BMC Bioinformatics* 22.569 (2021), pp. 1–16. doi: 10.1186/s12859-021-04453-5.

[216] Jay Ghurye et al. “Scaffolding of long read assemblies using long range contact information”. In: *BMC Genomics* 18.527 (2017). doi: 10.1186/s12864-017-3879-z.

[217] Jay Ghurye et al. “Integrating Hi-C links with assembly graphs for chromosome-scale assembly”. In: *PLoS Computational Biology* 15.8 (2019). doi: 10.1371/journal.pcbi.1007273.

[218] Zemin Ning. *scaffhic*, *https://github.com/wtsi-hpag/scaffHiC*. 2019.

[219] Chenxi Zhou, Shane A. McCarthy, and Richard Durbin. *YaHS: yet another Hi-C scaffolding tool*. Version v1.1a. 2021. doi: 10.5281/zenodo.5848773.

[220] Marten Boetzer and Walter Pirovano. “Toward almost closed genomes with GapFiller”. In: *Genome Biology* 13.R56 (2012), pp. 1–9. doi: 10.1186/gb-2012-13-6-r56.

[221] Chong Chu, Xin Li, and Yufeng Wu. “GAPPadder: a sensitive approach for closing gaps on draft genomes with short sequence reads”. In: *BMC Genomics* 20.426 (2019). doi: 10.1186/s12864-019-5703-4.

[222] Daniel Paulino et al. “Sealer: a scalable gap-closing application for finishing draft genomes”. In: *BMC Bioinformatics* 16.230 (2015), pp. 1–8. doi: 10.1186/s12859-015-0663-4.

[223] Vitor C Piro et al. "FGAP: an automated gap closing tool". In: *BMC Research Notes* 7.371 (2014). doi: 10.1186/1756-0500-7-371.

[224] Shunichi Kosugi, Hideki Hirakawa, and Satoshi Tabata. "GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments". In: *Bioinformatics* 31.23 (2015), pp. 3733–3741. doi: 10.1093/bioinformatics/btv465.

[225] Gui-Cai Xu et al. "LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly". In: *GigaScience* 8 (2019), pp. 1–14. doi: 10.1093/gigascience/giy157.

[226] Peng Lu et al. "PGcloser: fast parallel gap-closing tool using long-reads or contigs to fill gaps in genomes". In: *Evolutionary Bioinformatics* 16 (2020). doi: 10.1177/1176934320913859.

[227] Mengyang Xu et al. "TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads". In: *GigaScience* 9.9 (2020), pp. 1–11. doi: 10.1093/gigascience/giaa094.

[228] Clara Delahaye and Jacques Nicolas. "Sequencing DNA with nanopores: Troubles and biases". In: *PloS One* 16.10 (2021). doi: 10.1371/journal.pone.0257521.

[229] Pierre Morisse, Thierry Lecroq, and Arnaud Lefebvre. "Long-read error correction: a survey and qualitative comparison." In: *bioRxiv* (2020). doi: 10.1101/2020.03.06.977975.

[230] Byung June Ko et al. "Widespread false gene gains caused by duplication errors in genome assemblies". In: *bioRxiv* (2021). doi: 10.1101/2021.04.09.438957.

[231] Nadège Guiglielmoni et al. "Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms". In: *BMC Bioinformatics* 22.303 (2021), pp. 1–23. doi: 10.1186/s12859-021-04118-3.

[232] W James Kent and David Haussler. "Assembly of the working draft of the human genome with GigAssembler". In: *Genome Research* 11.9 (2001), pp. 1541–1548. doi: 10.1101/gr.183201.

[233] Jay Ghurye and Mihai Pop. "Modern technologies and algorithms for scaffolding assembled genomes". In: *PLoS Computational Biology* 15.6 (2019), pp. 1–20. doi: 10.1371/journal.pcbi.1006994.

[234] Janna L Fierst. "Using linkage maps to correct and scaffold *de novo* genome assemblies: methods, challenges, and computational tools". In: *Frontiers in Genetics* 6 (2015), p. 220. doi: 10.3389/fgene.2015.00220.

[235] David C Schwartz et al. "Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping". In: *Science* 262.5130 (1993), pp. 110–114. doi: 10.1126/science.8211116.

[236] Anna V. Protasio et al. "A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*". In: *PLoS Neglected Tropical Diseases* 6.1 (2012). doi: 10.1371/journal.pntd.0001455.

[237] Chang Bum Jeong et al. "The genome of the harpacticoid copepod *Tigriopus japonicus*: potential for its use in marine molecular ecotoxicology". In: *Aquatic Toxicology* 222 (2020), p. 105462. doi: 10.1016/j.aquatox.2020.105462.

[238] Yuxuan Yuan, Claire Yik-Lok Chung, and Ting-Fung Chan. "Advances in optical mapping for genomic research". In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 2051–2062. doi: 10.1016/j.csbj.2020.07.018.

[239] James A Cotton et al. "The genome of *Onchocerca volvulus*, agent of river blindness". In: *Nature Microbiology* 2.16216 (2016), pp. 1–12. doi: 10.1038/nmicrobiol.2016.216.

[240] Jianbin Wang et al. "Comparative genome analysis of programmed DNA elimination in nematodes". In: *Genome Research* 27.12 (2017), pp. 2001–2014. doi: 10.1101/gr.225730.117.

[241] Isheng J. Tsai et al. "The genomes of four tapeworm species reveal adaptations to parasitism". In: *Nature* 496.7443 (2013), pp. 57–63. doi: 10.1038/nature12031.

[242] Peter D Olson et al. "Complete representation of a tapeworm genome reveals chromosomes capped by centromeres, necessitating a dual role in segregation and protection". In: *BMC Biology* 18.165 (2020), pp. 1–16. doi: 10.1186/s12915-020-00899-w.

[243] Rebecca M Varney et al. "The iron-responsive genome of the chiton *Acanthopleura granulata*". In: *Genome Biology and Evolution* 13.1 (2021). doi: 10.1093/gbe/evaa263.

[244] Joana I Meier et al. "Haplotype tagging reveals parallel formation of hybrid races in two butterfly species". In: *Proceedings of the National Academy of Sciences* 118.25 (2021). doi: 10.1073/pnas.2015005118.

[245] Zhoutao Chen et al. "Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information". In: *Genome Research* 30.6 (2020), pp. 898–909. doi: 10.1101/gr.260380.119.

[246] Sarah D Kocher et al. "The genetic basis of a social polymorphism in halictid bees". In: *Nature Communications* 9.4338 (2018). doi: 10.1038/s41467-018-06824-8.

[247] SuperNova. *SuperNova*, *https://github.com/10XGenomics/supernova*. 2016.

[248] Jay Ghurye et al. "A chromosome-scale assembly of the major African malaria vector *Anopheles funestus*". In: *GigaScience* 8.6 (2019). doi: 10.1093/gigascience/giz063.

[249] Job Dekker et al. "Capturing chromosome conformation". In: *Science* 295.5558 (2002), pp. 1306–1311. doi: 10.1126/science.1067799.

[250] Job Dekker, Marc A Marti-Renom, and Leonid A Mirny. "Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data". In: *Nature Reviews Genetics* 14.6 (2013), pp. 390–403. doi: 10.1038/nrg3454.

[251] Erez Lieberman-Aiden et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". In: *Science* 326.5950 (2009), pp. 289–93. doi: 10.1126/science.1181369.

[252] Jean-François Flot, Hervé Marie-Nelly, and Romain Koszul. "Contact genomics: scaffolding and phasing (meta) genomes using chromosome 3D physical signatures". In: *FEBS Letters* 589.20 (2015), pp. 2966–2974. doi: 10.1016/j.febslet.2015.04.034.

[253] Sivan Oddes, Aviv Zelig, and Noam Kaplan. "Three invariant Hi-C interaction patterns: applications to genome assembly". In: *Methods* 142 (2018), pp. 89–99. doi: 10.1016/j.ymeth.2018.04.013.

[254] Maeva A Techer et al. "Divergent evolutionary trajectories following speciation in two ectoparasitic honey bee mites". In: *Communications Biology* 2.357 (2019). doi: 10.1038/s42003-019-0606-0.

[255] Prashant Shingate et al. "Chromosome-level assembly of the horseshoe crab genome provides insights into its genome evolution". In: *Nature Communications* 11.2322 (2020). doi: 10.1038/s41467-020-16180-1.

[256] Hugo Darras et al. "Chromosome-level genome assembly and annotation of two lineages of the ant *Cataglyphis hispanica*: steppingstones towards genomic studies of hybridogenesis and thermal adaptation in desert ants". In: *bioRxiv* (2022). doi: 10.1101/2022.01.07.475286.

[257] Minjie Hu et al. "Lineage dynamics of the endosymbiotic cell type in the soft coral *Xenia*". In: *Nature* 582.7813 (2020), pp. 534–538. doi: 10.1038/s41586-020-2385-7.

[258] Yunfeng Li et al. "Chromosome-level reference genome of the jellyfish *Rhopilema esculentum*". In: *GigaScience* 9.4 (2020). doi: 10.1093/gigascience/giaa036.

[259] Phillip L Davidson et al. "Chromosomal-level genome assembly of the sea urchin *Lytechinus variegatus* substantially improves functional genomic analyses". In: *Genome Biology and Evolution* 12.7 (2020), pp. 1080–1086. doi: 10.1093/gbe/evaa101.

[260] Dannise V Ruiz-Ramos et al. "An initial comparative genomic autopsy of wasting disease in sea stars". In: *Molecular Ecology* 29.6 (2020), pp. 1087–1102. doi: 10.1111/mec.15386.

[261] Chang Ming Bai et al. "Chromosomal-level assembly of the blood clam, *Scapharca (Anadara) broughtonii*, using long sequence reads and Hi-C". In: *GigaScience* 8.7 (2019). doi: 10.1093/gigascience/giz067.

[262] Jin Sun et al. "The scaly-foot snail genome and implications for the origins of biomineralised armour". In: *Nature Communications* 11.1 (2020). doi: 10.1038/s41467-020-15522-3.

[263] Sarah Farhat et al. “Comparative analysis of the *Mercenaria mercenaria* genome provides insights into the diversity of transposable elements and immune molecules in bivalve mollusks”. In: *BMC Genomics* 23.1 (2022), pp. 1–23. doi: 10.1186/s12864-021-08262-1.

[264] Anastasia A Teterina, John H Willis, and Patrick C Phillips. “Chromosome-level assembly of the *Caenorhabditis remanei* genome reveals conserved patterns of nematode genome organization”. In: *Genetics* 214.4 (2020), pp. 769–780. doi: 10.1534/genetics.119.303018.

[265] Yun Lian et al. “Chromosome-level reference genome of X12, a highly virulent race of the soybean cyst nematode *Heterodera glycines*”. In: *Molecular Ecology Resources* 19.6 (2019), pp. 1637–1646. doi: 10.1111/1755-0998.13068.

[266] Andreas J. Stroehlein et al. “High-quality *Schistosoma haematobium* genome achieved by single-molecule and long-range sequencing”. In: *GigaScience* 8.9 (2019). doi: 10.1093/gigascience/giz108.

[267] Nathan J Kenny et al. “Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge *Ephydatia muelleri*”. In: *Nature Communications* 1 (2020). doi: 10.1038/s41467-020-17397-w.

[268] Andrew R Gehrke et al. “Acoel genome reveals the regulatory landscape of whole-body regeneration”. In: *Science* 363.6432 (2019). doi: 10.1126/science.aau6173.

[269] Ehsan Haghshenas et al. “HASLR: Fast hybrid assembly of long reads”. In: *iScience* 23.8 (2020), p. 101389. doi: 10.1016/j.isci.2020.101389.

[270] Aleksey V. Zimin et al. “The MaSuRCA genome assembler”. In: *Bioinformatics* 29.21 (2013), pp. 2669–2677. doi: 10.1093/bioinformatics/btt476.

[271] Alex Di Genova et al. “Efficient hybrid *de novo* assembly of human genomes with WENGAN”. In: *Nature Biotechnology* 39.4 (2021), pp. 422–430. doi: 10.1038/s41587-020-00747-w.

[272] Kelly L. Mulligan et al. “First estimates of genome size in ribbon worms (phylum Nemertea) using flow cytometry and Feulgen image analysis densitometry”. In: *Canadian Journal of Zoology* 92.10 (2014), pp. 847–851. doi: 10.1139/cjz-2014-0068.

[273] Joint Genome Institute. *BBtools*, *https://sourceforge.net/projects/bbmap/*. 2013.

[274] T. Rhyker Ranallo-Benavidez, Kamil S. Jaron, and Michael C. Schatz. “GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes”. In: *Nature Communications* 11.1 (2020), p. 1432. doi: 10.1038/s41467-020-14998-3.

[275] Daniel Mapleson et al. “KAT: a K-mer Analysis Toolkit to quality control NGS datasets and genome assemblies”. In: *Bioinformatics* 33.4 (2016), pp. 574–576. doi: 10.1093/bioinformatics/btw663.

[276] Boas Pucker. “Mapping-based genome size estimation”. In: *bioRxiv* (2019). doi: 10.1101/607390.

[277] Alexey Gurevich et al. “QUAST: quality assessment tool for genome assemblies”. In: *Bioinformatics* 29.8 (2013), pp. 1072–1075. doi: 10.1093/bioinformatics/btt086.

[278] Yunhai Guo et al. “A chromosomal-level genome assembly for the giant African snail *Achatina fulica*”. In: *GigaScience* 8.10 (2019). doi: 10.1093/gigascience/giz124.

[279] Cheng He et al. “Factorial estimating assembly base errors using *k*-mer abundance difference (KAD) between short reads and genome assembled sequences”. In: *NAR Genomics and Bioinformatics* 2.3 (2020). doi: 10.1093/nargab/lqaa075.

[280] Dominik R Laetsch and Mark L Blaxter. “BlobTools: Interrogation of genome assemblies”. In: *F1000Research* 6.1287 (2017), p. 1287. doi: 10.12688/f1000research.12232.1.

[281] Richard Challis et al. “BlobToolKit – Interactive quality assessment of genome assemblies”. In: *G3: Genes, Genomes, Genetics* 10.4 (2020), pp. 1361–1374. doi: 10.1534/g3.119.400908.

[282] Thomas C. Boothby et al. “Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade”. In: *Proceedings of the National Academy of Sciences of the United States of America* 112.52 (2015), pp. 15976–15981. doi: 10.1073/pnas.1510461112.

[283] Georgios Koutsovoulos et al. “No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*”. In: *Proceedings of the National Academy of Sciences* 113.18 (2016), pp. 5053–5058. doi: 10.1073/pnas.1600338113.

[284] Xingtan Zhang et al. “Unzipping haplotypes in diploid and polyploid genomes”. In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 66–72. doi: 10.1016/j.csbj.2019.11.011.

[285] Sergey Koren et al. “*De novo* assembly of haplotype-resolved genomes with trio binning”. In: *Nature Biotechnology* 36.12 (2018), pp. 1174–1182. doi: 10.1038/nbt.4277.

[286] Peter Edge, Vineet Bafna, and Vikas Bansal. “HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies”. In: *Genome Research* 27.5 (2017), pp. 801–812. doi: 10.1101/gr.213462.116.

[287] Murray D. Patterson et al. “WhatsHap: weighted haplotype assembly for future-generation sequencing reads”. In: *Journal of Computational Biology* 22.6 (2015), pp. 498–509. doi: 10.1089/cmb.2014.0157.

[288] Antoine Limasset. “Novel approaches for the exploitation of high throughput sequencing data”. PhD thesis. Université Rennes 1, 2017.

[289] Rei Kajitani et al. “Platanus-allee is a *de novo* haplotype assembler enabling a comprehensive access to divergent heterozygous regions”. In: *Nature Communications* 10.1702 (2019), pp. 1–15. doi: 10.1038/s41467-019-09575-2.

[290] David Porubsky et al. “Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads”. In: *Nature Biotechnology* 39 (2020), pp. 302–308. doi: 10.1038/s41587-020-0719-5.

[291] Qian Zhou et al. “Haplotype-resolved genome analyses of a heterozygous diploid potato”. In: *Nature Genetics* 52 (2020), pp. 1018–1023. doi: 10.1038/s41588-020-0699-x.

[292] Guillaume Holley et al. “Ratatosk: hybrid error correction of long reads enables accurate variant calling and assembly”. In: *Genome Biology* 22.1 (2021). doi: 10.1186/s13059-020-02244-4.

[293] Xingtan Zhang et al. “Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data”. In: *Nature Plants* 5.8 (2019), pp. 833–845. doi: 10.1038/s41477-019-0487-8.

[294] Roland Faure, Nadège Guiglielmoni, and Jean-François Flot. “GraphUnzip: unzipping assembly graphs with long reads and Hi-C”. In: *bioRxiv* (2021). doi: 10.1101/2021.01.29.428779.

[295] Zev N Kronenberg et al. “Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C”. In: *Nature Communications* 12.1 (2021), pp. 1–10. doi: 10.1038/s41467-020-20536-y.

[296] Arang Rhie et al. “Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies”. In: *Genome Biology* 21.245 (2020), pp. 1–27. doi: 10.1186/s13059-020-02134-9.

[297] Steven L. Salzberg. “Next-generation genome annotation: we still struggle to get it right. ” In: *Genome Biology* 20.92 (2019). doi: 10.1186/s13059-019-1715-2.

[298] Mark D Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3.1 (2016), pp. 1–9. doi: 10.1038/sdata.2016.18.

[299] Delphine Lariviere and Alex Ostrovsky. *VGP assembly pipeline (Galaxy Training Materials)*, *training.galaxyproject.org/training-material/topics/assembly/tutorials/vgp_genome_assembly/tutorial.html*. 2021.

[300] Nadège Guiglielmoni et al. “Supplementary table to "A deep dive into genome assemblies of non-vertebrate animals"”. In: (Apr. 2022). doi: 10.6084/m9.figshare.19672440.v1. url: https://figshare.com/articles/dataset/a_deep_dive_into_genome_assemblies_of_non-vertebrates_tsv/19672440.

[301] Nadège Guiglielmoni. *Genome assembly tools*, *https://github.com/nadegeguiglielmoni/genome_assembly_tools*. 2022.