Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

A deep dive into genome assemblies of non-vertebrate animals

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

[1] Service Evolution Biologique et Ecologie, Université libre de Bruxelles (ULB), 1050 Brussels, Belgium
[2] Institut für Zoologie, Universität zu Köln, 50674 Cologne, Germany
[3] Department of Earth and Environmental Sciences, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, 80333 Munich, Germany
[4] Institut Pasteur, Unité Régulation Spatiale des Génomes, UMR 3525, CNRS, 75015 Paris, France
[5] Interuniversity Institute of Bioinformatics in Brussels – (IB)[2], 1050 Brussels, Belgium

Abstract

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

Keywords: genome assembly; sequencing; non-vertebrate animals

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

Introduction

The field of genomics is presently thriving, with new genomes of all kind of organisms becoming available every day. For Metazoa, efforts have unsurprisingly focused on human's closest relatives (i.e., vertebrates) so far [1]: out of 7,894 metazoan assemblies available in the GenBank database (accessed on October 29th, 2021) [2], ∼ 56.9% (4,493) belong to the subphylum Vertebrata. However, from the currently ∼2.1 million described metazoan species, only ∼73,000 (3.5%) belong to vertebrates [3]. The remaining metazoan phyla, hereafter called "non-vertebrate animals", are thus underinvestigated and lack genetic resources.

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

In recent years, important sequencing efforts have started to tackle the dearth of genomic data for non-vertebrate animals, with a strong focus on arthropods (2,683 assemblies on GenBank). The phylum Arthropoda is very diverse: it consists of more than 1.3 million species, the majority of which belong to the class Insecta (∼1 million species) [5]. Insects have a significant impact on agriculture (e.g. as crop pests) and on the transmission of diseases (e.g. malaria and dengue) [6]. They also play important beneficial and regulatory roles in natural ecosystems, through pollination and decomposition of organic matter [7]. Genome sequencing yields invaluable insights into species that are key in the aforementioned processes. For example, various genome projects have targeted insects such as *Bemisia tabaci*, a common crop pest [8], and the mosquitoes *Aedes aegypti* (vector of yellow fever, dengue and chikungunya) [9] and *Anopheles darlingi* (vector of malaria) [10]. These studies unveiled, among other findings, expansions of genes involved in insecticide resistance. The genomes of these species are so important for human health and food security that many have actually been sequenced multiple times, either because of the availability of newer sequencing methods or to compare different strains (for instance, three versions of the genome of *Aedes aegypti* [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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

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.

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 base-pair (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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

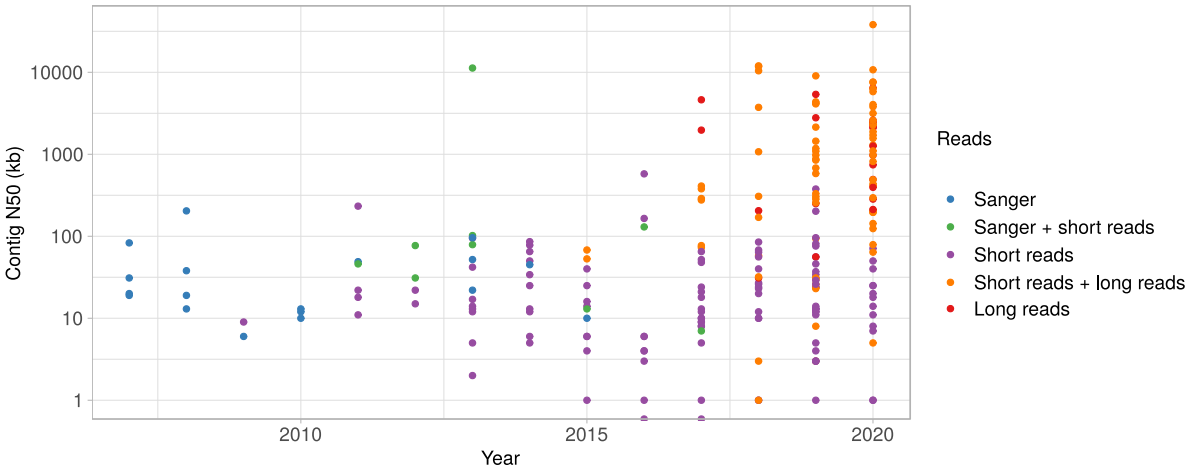

Figure 1. Contig N50 of 237 non-vertebrate animal genome assemblies over time. The N50 represents the contiguity of an assembly and is defined as the length of the largest contig for which at least 50% of the assembly size is contained in contigs equal or greater in length.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

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 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.

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].

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

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]

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.

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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

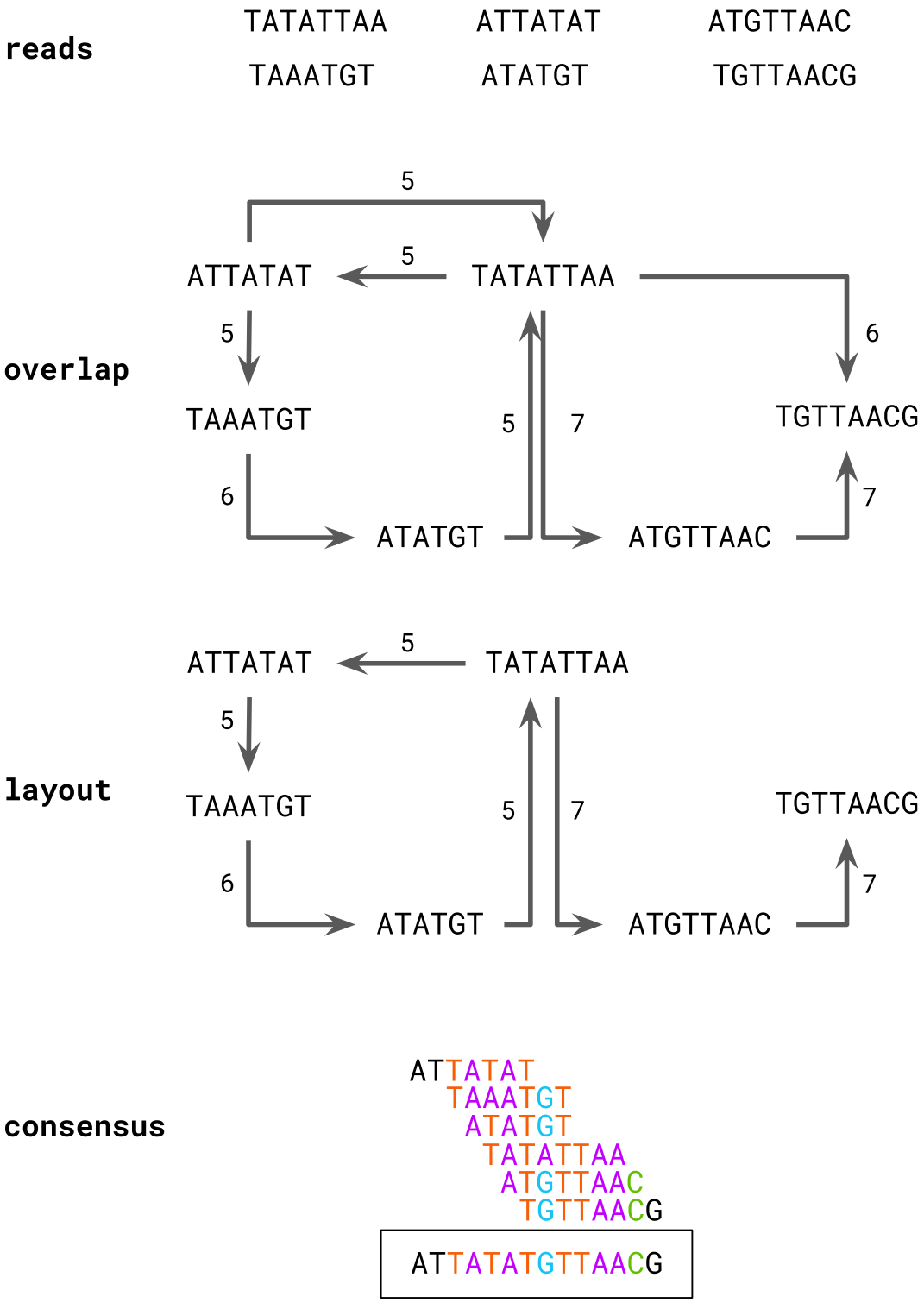

Figure 2. Overview of Overlap-Layout-Consensus assembly. The graph was built with all overlaps of at least 5 bases with a tolerance of 1 mismatch.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

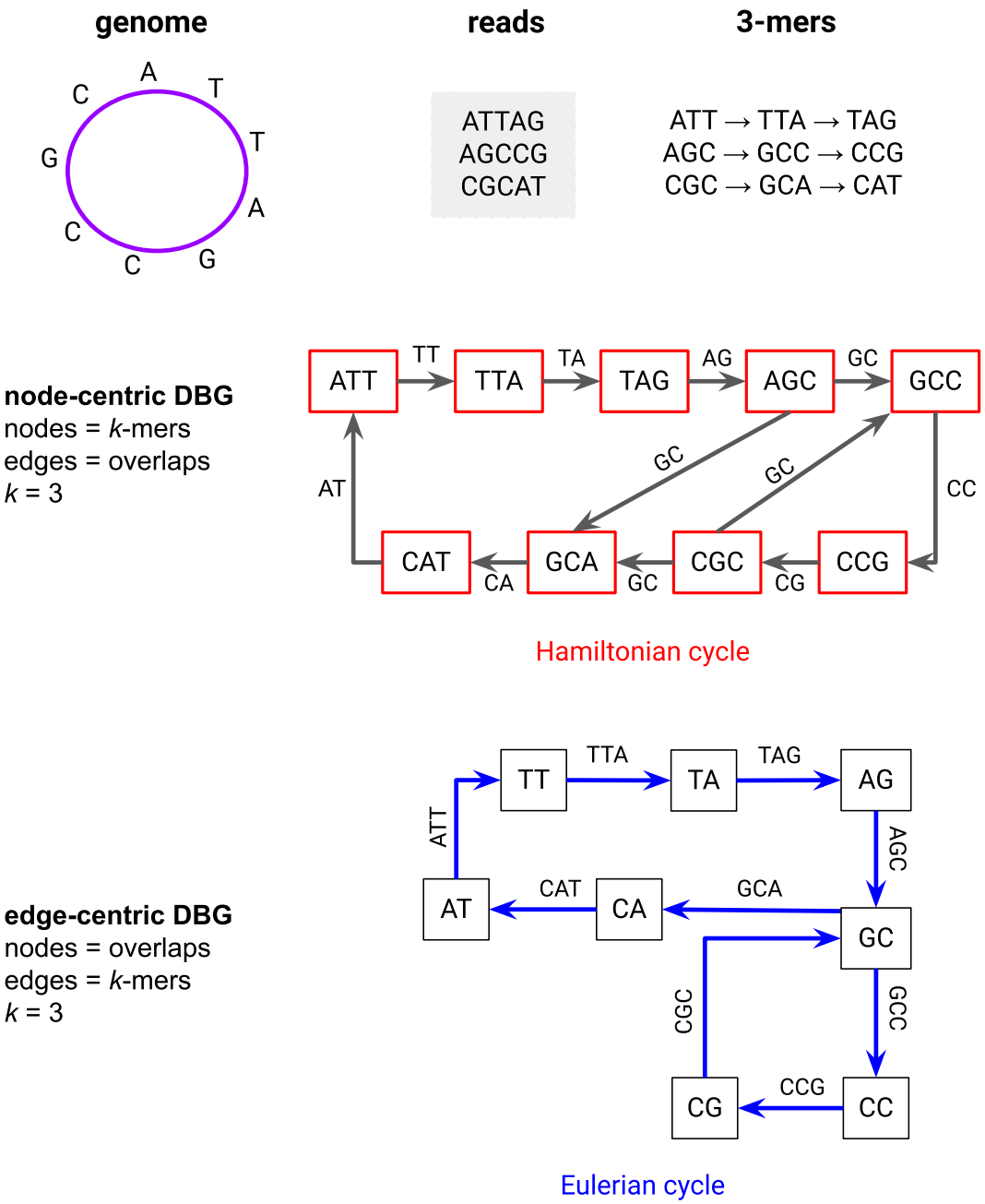

Figure 3. Overview of genome assembly using de Bruijn graphs. A circular genome is assembled based on three reads using node-centric and edge-centric DBGs with $k = 3$. The node-centric DBG is searched for a Hamiltonian cycle (visiting all nodes), and the edge-centric DBG for an Eulerian cycle (visiting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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

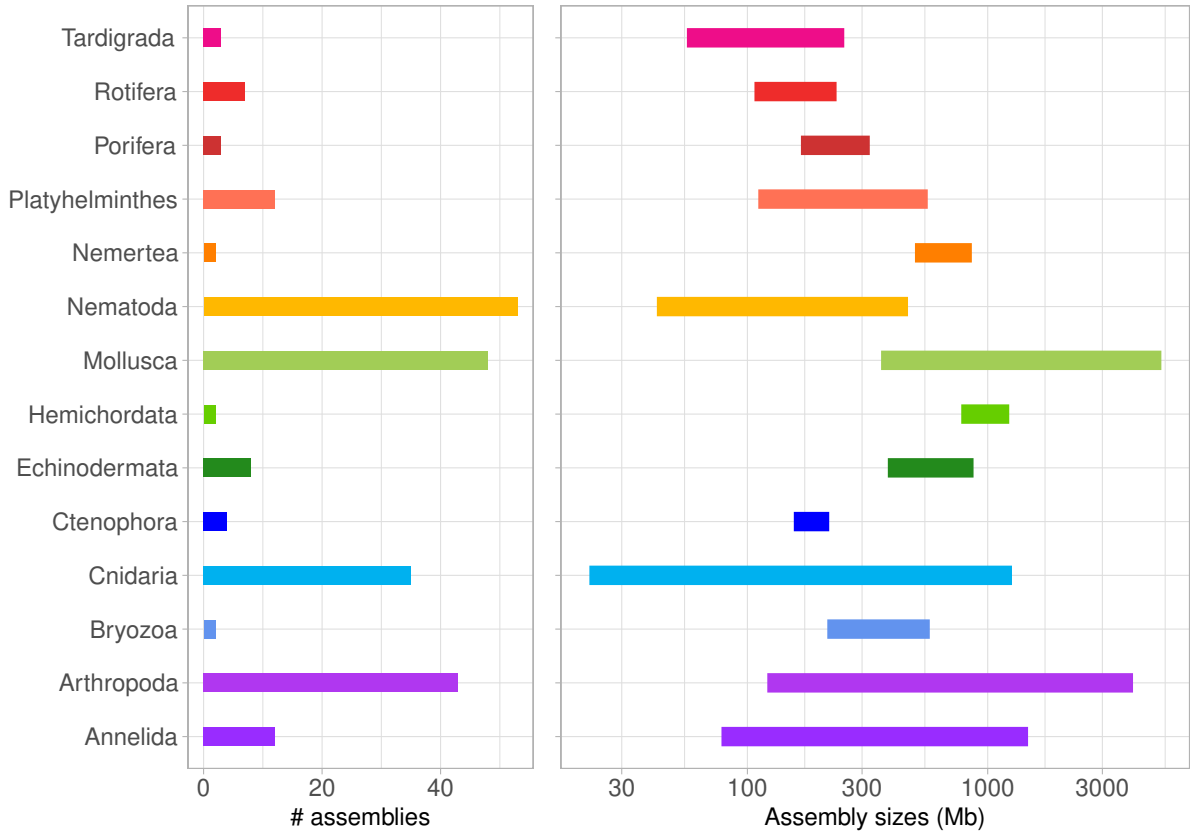

Figure 4. Assembly sizes. The left graph shows the number of assemblies included for each phylum and the right part shows the corresponding assembly-size ranges.

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 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 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

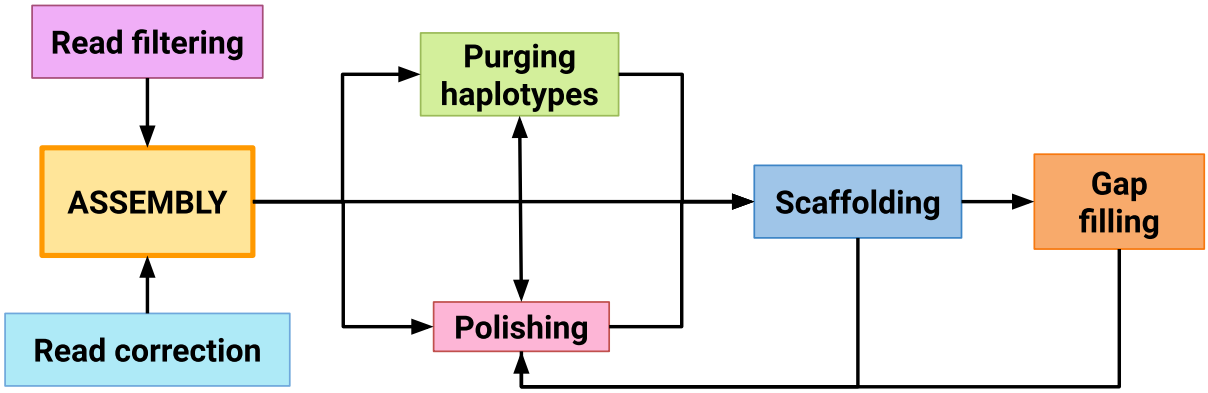

Figure 5. Assembly pipeline, including the assembly and the pre/post-processing steps.

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 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:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

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]

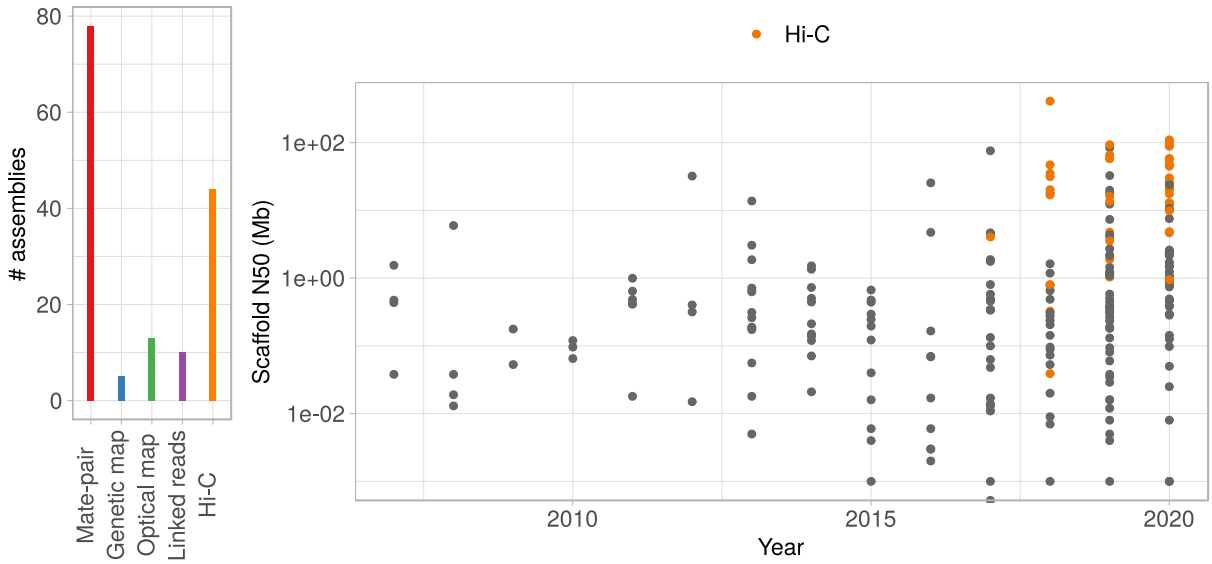

Figure 6. Assemblies scaffolding. Left: number of assemblies that included each scaffolding method. Right: scaffold N50 of non-vertebrate animal genome assemblies over time. The assemblies that included a Hi-C scaffolding step are highlighted in orange; they form a cluster with a scaffold N50 over 1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

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 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

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].

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

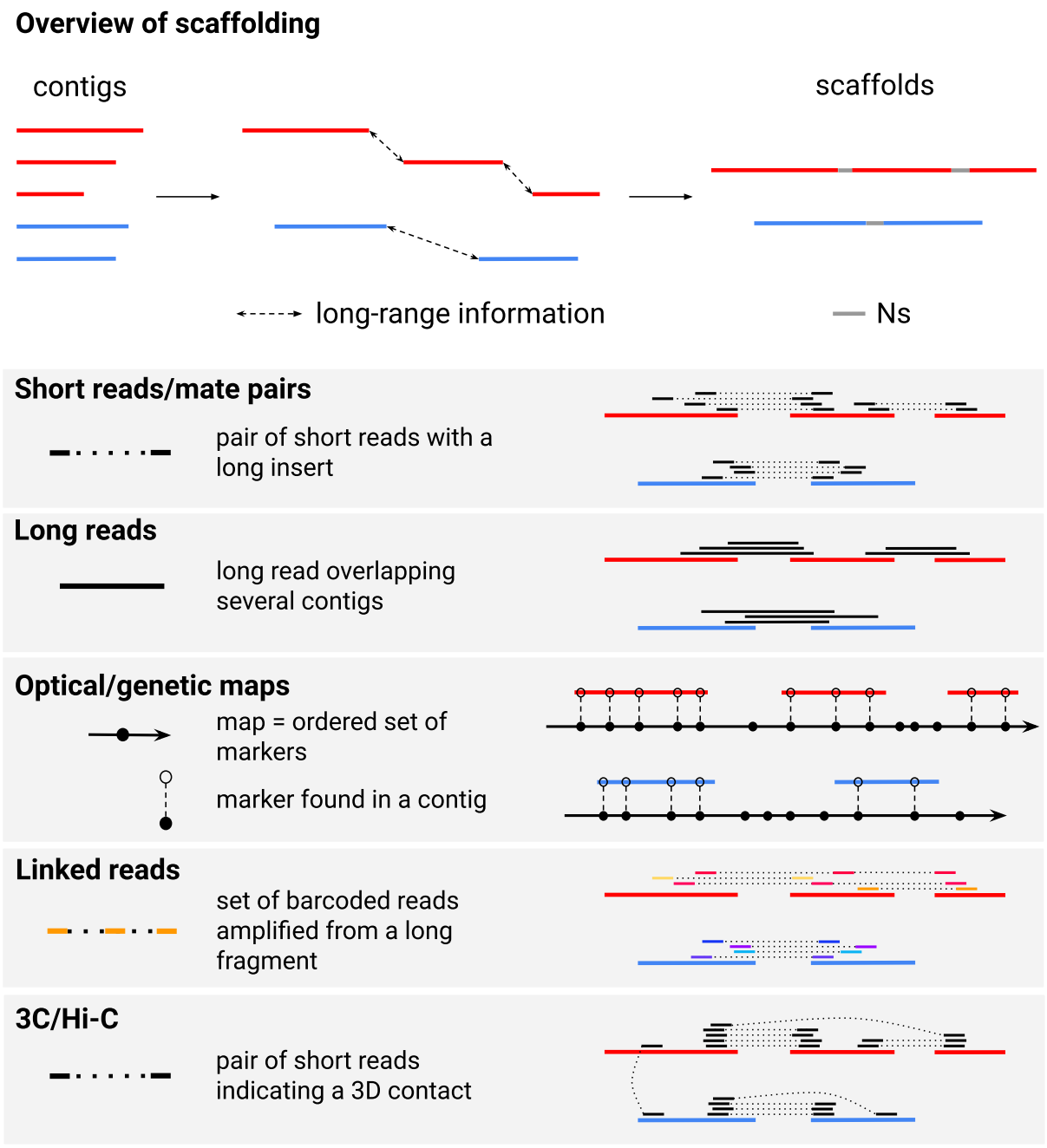

Figure 7. Overview of scaffolding methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

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%

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 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 as-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

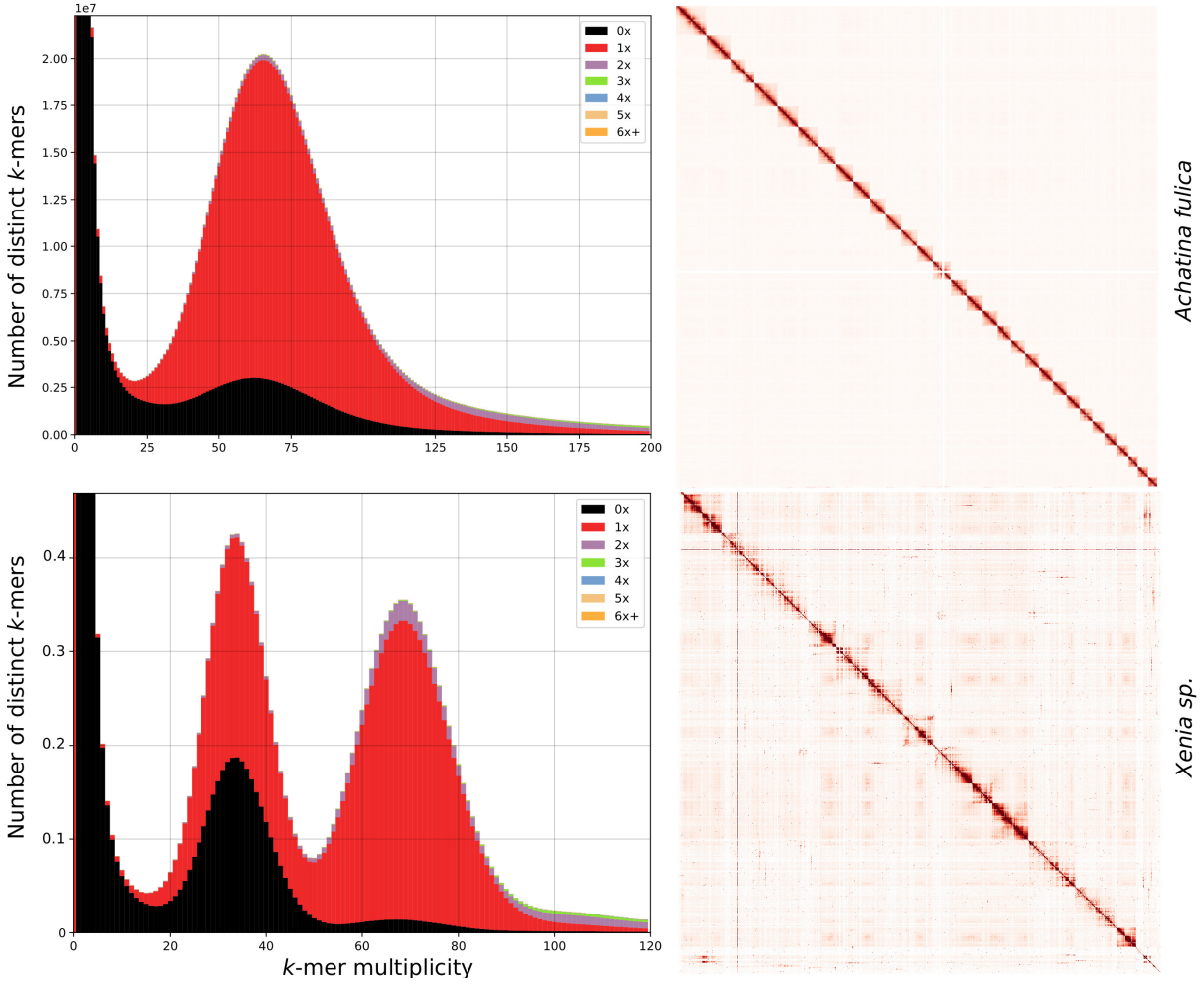

Figure 8. Assembly evaluation of *Achatina fulica* and *Xenia* sp.. Left: KAT comparison of the *k*-mers in the Illumina datasets v. the assembly. Right: Hi-C contact maps, with a binning of 300 for *Achatina fulica*, 30 for *Xenia* sp..

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

sembly based on its representation of *k*-mers from a high-accuracy sequencing dataset. The *k*-mer spectrum should display one or several peaks depending on the ploidy of the genome: one peak for a haploid genome; two peaks for a diploid genome, the first depicting heterozygous *k*-mers, and the second for homozygous *k*-mers. Depending on the ploidy of the genome, every *k*-mer should be represented in the assembly as many times as they actually are in the genome.

Both *Achatina fulica* and *Xenia* sp. have high BUSCO scores (against the lineage Metazoa odb10), yet slightly below 90%, and they have few duplicated BUSCO features. The *k*-mer spectrum of *Achatina fulica* only shows one peak around 70X (Figure 8, top left). These *k*-mers are expected to be represented exactly once, which is the case for the majority; there are almost no *k*-mers that appear twice in the assembly (in purple), but there is a noteworthy amount of missing *k*-mers (in black). For *Xenia* sp., the *k*-mer spectrum has two peaks with a *k*-mer multiplicity around 35X and 70X (Figure 8, bottem left). The first peak, representing heterozygous *k*-mers, shows that a portion is represented once in the assembly, while the rest is missing, as expected in a collapsed assembly. The second peak, for homozygous *k*-mers has a majority of *k*-mers represented once, and some *k*-mers either absent or duplicated. These assemblies seem overall properly collapsed and complete.

KAD, for *k*-mer abundance difference [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.

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 assem-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

blies 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.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 May 2022 doi:10.20944/preprints202111.0170.v3

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.

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

Version 3 of this article has been peer-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] Rice ES and Green RE. New approaches for genome assembly and scaffolding. *Annual Review of Animal Biosciences* 7 (2019), 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. Accessed on May 4th, 2021.

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

[5] Zhang ZQ. 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. Magnolia Press, 2013, pp. 1–82. doi: 10.11646/zootaxa.3703.1.3.

[6] Li F, Zhao X, Li M, He K, Huang C, Zhou Y, Li Z, and Walters JR. Insect genomes: progress and challenges. *Insect Molecular Biology* 28 (2019), 739–758. doi: 10.1111/imb.12599.

[7] Noriega JA, Hortal J, Azcárate FM, Berg MP, Bonada N, Briones MJ, Del Toro I, Goulson D, Ibanez S, Landis DA, et al. Research trends in ecosystem services provided by insects. *Basic and Applied Ecology* 26 (2018), 8–23. doi: 10.1016/j.baae.2017.09.006.

[8] Chen W, Hasegawa DK, Kaur N, Kliot A, Pinheiro PV, Luan J, Stensmyr MC, Zheng Y, Liu W, Sun H, et al. The draft genome of whitefly *Bemisia tabaci* MEAM1, a global crop pest, provides novel insights into virus transmission, host adaptation, and insecticide resistance. *BMC Biology* 14 (2016), 1–15. doi: 10.1186/s12915-016-0321-y.

[9] Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu Z, Loftus B, Xi Z, Megy K, Grabherr M, Ren Q, Zdobnov EM, Lobo NF, Campbell KS, Brown SE, Bonaldo MF, Zhu J, Sinkins SP, Hogenkamp DG, Amedeo P, Arensburger P, Atkinson PW, Bidwell S, Biedler J, Birney E, Bruggner RV, Costas J, Coy MR, Crabtree J, Crawford M, DeBruyn B, DeCaprio D, Eiglmeier K, Eisenstadt E, El-Dorry H, Gelbart WM, Gomes SL, Hammond M, Hannick LI, Hogan JR, Holmes MH, Jaffe D, Johnston JS, Kennedy RC, Koo H, Kravitz S, Kriventseva EV, Kulp D, LaButti K, Lee E, Li S, Lovin DD, Mao C, Mauceli E, Menck CF, Miller JR, Montgomery P, Mori A, Nascimento AL, Naveira HF, Nusbaum C, O'Leary S, Orvis J, Pertea M, Quesneville H, Reidenbach KR, Rogers YH, Roth CW, Schneider JR, Schatz M, Shumway M, Stanke M, Stinson EO, Tubio JM, VanZee JP, Verjovski-Almeida S, Werner D, White O, Wyder S, Zeng Q, Zhao Q, Zhao Y, Hill CA, Raikhel AS, Soares MB, Knudson DL, Lee NH, Galagan J, Salzberg SL, Paulsen IT, Dimopoulos G, Collins FH, Birren B, Fraser-Liggett CM, and Severson DW. Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* 316 (2007), 1718–1723. doi: 10.1126/science.1138878.

[10] Marinotti O, Cerqueira GC, De Almeida LGP, Ferro MIT, Loreto ELdS, Zaha A, Teixeira SM, Wespiser AR, Almeida e Silva A, Schlindwein AD, et al. The genome of *Anopheles darlingi*, the main neotropical malaria vector. *Nucleic Acids Research* 41 (2013), 7387–7400. doi: 10.1093/nar/gkt484.

[11] Matthews BJ, Dudchenko O, Kingan SB, Koren S, Antoshechkin I, Crawford JE, Glassford WJ, Herre M, Redmond SN, Rose NH, Weedall GD, Wu Y, Batra SS, Brito-Sierra CA, Buckingham SD, Campbell CL, Chan S, Cox E, Evans BR, Fansiri T, Filipović I, Fontaine A, Gloria-Soria A, Hall R, Joardar VS, Jones AK, Kay RG, Kodali VK, Lee J, Lycett GJ, Mitchell SN, Muehling J, Murphy MR, Omer AD, Partridge FA, Peluso P, Aiden AP, Ramasamy V, Rašić G, Roy S, Saavedra-Rodriguez K, Sharan S, Sharma A, Smith ML, Turner J, Weakley AM, Zhao Z, Akbari OS, Black WC, Cao H, Darby AC, Hill CA, Johnston JS, Murphy TD, Raikhel AS, Sattelle DB, Sharakhov IV, White BJ, Zhao L, Aiden EL, Mann RS, Lambrechts L, Powell JR, Sharakhova MV, Tu Z, Robertson HM, McBride CS, Hastie AR, Korlach J, Neafsey DE, Phillippy AM, and Vosshall LB. Improved reference genome of *Aedes aegypti* informs arbovirus vector control. *Nature* 563 (2018), 501–507. doi: 10.1038/s41586-018-0692-z.

[12] Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, and Aiden EL. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356 (2017), 92–95. doi: 10.1126/science.aal3327.

[13] Hotaling S, Kelley JL, and Frandsen PB. Toward a genome sequence for every animal: Where are we now? *Proceedings of the National Academy of Sciences* 118 (2021). doi: 10.1073/pnas.2109019118.

[14] Carroll AR, Copp BR, Davis RA, Keyzers RA, and Prinsep MR. Marine natural products. *Natural Product Reports* (2021). doi: 10.1039/C9NP00069K.

[15] Khalifa SA, Elias N, Farag MA, Chen L, Saeed A, Hegazy MEF, Moustafa MS, El-Wahed A, Al-Mousawi SM, Musharraf SG, et al. Marine natural products: A source of novel anticancer drugs. *Marine Drugs* 17 (2019), 491. doi: 10.3390/md17090491.

[16] Ng TB, Cheung RCF, Wong JH, Bekhit AA, and Bekhit AED. Antibacterial products of marine organisms. *Applied Microbiology and Biotechnology* 99 (2015), 4145–4173. doi: 10.1007/s00253-015-6553-x.

[17] Avila C. Terpenoids in marine heterobranch molluscs. *Marine Drugs* 18 (2020), 162. doi: 10.3390/md18030162.

[18] Han BN, Hong LL, Gu BB, Sun YT, Wang J, Liu JT, and Lin HW. 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] Takeuchi T. Molluscan genomics: implications for biology and aquaculture. *Current Molecular Biology Reports* 3 (2017), 297–305. doi: 10.1007/s40610-017-0077-3.

[20] Prather CM, Pelini SL, Laws A, Rivest E, Woltz M, Bloch CP, Del Toro I, Ho CK, Kominoski J, Newbold TS, et al. Invertebrates, ecosystem services and climate change. *Biological Reviews* 88 (2013), 327–348. doi: 10.1111/brv.12002.

[21] Gomes-dos-Santos A, Lopes-Lima M, Castro LFC, and Froufe E. Molluscan genomics: the road so far and the way forward. *Hydrobiologia* 847 (2020), 1705–1726. doi: 10.1007/s10750-019-04111-1.

[22] Conci N, Vargas S, and Wörheide G. The biology and evolution of calcite and aragonite mineralization in octocorallia. *Frontiers in Ecology and Evolution* 9 (2021), 81. doi: 10.3389/fevo.2021.623774.

[23] Clark MS. Molecular mechanisms of biomineralization in marine invertebrates. *Journal of Experimental Biology* 223 (2020). doi: 10.1242/jeb.206961.

[24] Shinzato C, Shoguchi E, Kawashima T, Hamada M, Hisata K, Tanaka M, Fujie M, Fujiwara M, Koyanagi R, Ikuta T, et al. Using the *Acropora digitifera* genome to understand coral responses to environmental change. *Nature* 476 (2011), 320–323. doi: 10.1038/nature10249.

[25] Mao Y, Economo EP, and Satoh N. The roles of introgression and climate change in the rise to dominance of Acropora corals. *Current Biology* 28 (2018), 3373–3382. doi: 10.1016/j.cub.2018.08.061.

[26] Fuller ZL, Mocellin VJ, Morris LA, Cantin N, Shepherd J, Sarre L, Peng J, Liao Y, Pickrell J, Andolfatto P, et al. Population genetics of the coral *Acropora millepora*: Toward genomic prediction of bleaching. *Science* 369 (2020). doi: 10.1126/science.aba4674.

[27] Shinzato C, Khalturin K, Inoue J, Zayasu Y, Kanda M, Kawamitsu M, Yoshioka Y, Yamashita H, Suzuki G, and Satoh N. Eighteen coral genomes reveal the evolutionary origin of Acropora strategies to accommodate environmental changes. *Molecular Biology and Evolution* 38 (2021), 16–30. doi: 10.1093/molbev/msaa216.

[28] Kapli P, Yang Z, and Telford MJ. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics* 21 (2020), 428–444. doi: 10.1038/s41576-020-0233-0.

[29] Chang ES, Neuhof M, Rubinstein ND, Diamant A, Philippe H, Huchon D, and Cartwright P. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proceedings of the National Academy of Sciences* 112 (2015), 14912–14917. doi: 10.1073/pnas.1511468112.

[30] Eitel M, Francis WR, Varoqueaux F, Daraspe J, Osigus HJ, Krebs S, Vargas S, Blum H, Williams GA, Schierwater B, and Wörheide G. Comparative genomics and the nature of placozoan species. *PLoS Biology* 16 (2018), 1–36. doi: 10.1371/journal.pbio.2005359.

[31] Lynch M, Bürger R, Butcher D, and Gabriel W. The Mutational Meltdown in Asexual Populations. *Journal of Heredity* 84 (Sept. 1993), 339–344. doi: 10.1093/oxfordjournals.jhered.a111354.

[32] Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, Bauer DJ, Cáceres CE, Carmel L, Casola C, Choi JH, Detter JC, Dong Q, Dusheyko S, Eads BD, Fröhlich T, Geiler-Samerotte KA, Gerlach D, Hatcher P, Jogdeo S, Krijgsveld J, Kriventseva EV, Kültz D, Laforsch C, Lindquist E, Lopez J, Manak JR, Muller J, Pangilinan J, Patwardhan RP, Pitluck S, Pritham EJ, Rechtsteiner A, Rho M, Rogozin IB, Sakarya O, Salamov A, Schaack S, Shapiro H, Shiga Y, Skalitzky C, Smith Z, Souvorov A, Sung W, Tang Z, Tsuchiya D, Tu H, Vos H, Wang M, Wolf YI, Yamagata H, Yamada T, Ye Y, Shaw JR, Andrews J, Crease TJ, Tang H, Lucas SM, Robertson HM, Bork P, Koonin EV, Zdobnov EM, Grigoriev IV, Lynch M, and Boore JL. The ecoresponsive genome of Daphnia pulex. *Science* 331 (2011), 555–561. doi: 10.1126/science.1197761.

[33] Ye Z, Xu S, Spitze K, Asselman J, Jiang X, Ackerman MS, Lopez J, Harker B, Raborn RT, Thomas WK, Ramsdell J, Pfrender ME, and Lynch M. A new reference genome assembly for the microcrustacean *Daphnia pulex*. *G3: Genes, Genomes, Genetics* 7 (2017), 1405–1416. doi: 10.1534/g3.116.038638.

[34] Schiffer PH, Danchin EG, Burnell AM, Creevey CJ, Wong S, Dix I, O'Mahony G, Culleton BA, Rancurel C, Stier G, Martínez-Salazar EA, Marconi A, Trivedi U, Kroiher M, Thorne MA, Schierenberg E, Wiehe T, and Blaxter M. Signatures of the evolution of parthenogenesis and cryptobiosis in the genomes of panagrolaimid nematodes. *iScience* 21 (2019), 587–602. doi: https://doi.org/10.1016/j.isci.2019.10.039.

[35] Brandt A, Van PT, Bluhm C, Anselmetti Y, Dumas Z, Figuet E, François CM, Galtier N, Heimburger B, Jaron KS, et al. Haplotype divergence supports long-term asexuality in the oribatid mite *Oppiella nova*. *Proceedings of the National Academy of Sciences* 118 (2021). doi: 10.1073/pnas.2101485118.

[36] Simion P, Narayan J, Houtain A, Derzelle A, Baudry L, Nicolas E, Arora R, Cariou M, Cruaud C, Gaudray FR, et al. Chromosome-level genome assembly reveals homologous chromosomes and recombination in asexual rotifer *Adineta vaga*. *Science Advances* 7 (2021). doi: 10.1126/sciadv.abg4216.

[37] Buck M and Hamilton C. 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. *Review of European Community & International Environmental Law* 20 (2011), 47–61. doi: 10.1111/j.1467-9388.2011.00703.x.

[38] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, and Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31 (2015), 3210–3212. doi: 10.1007/978-1-4939-9173-0_14.

[39] Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, and Zdobnov EM. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution* 35 (2018), 543–548. doi: 10.1093/molbev/msx319.

[40] Manni M, Berkeley MR, Seppey M, Simão FA, and Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* 38 (2021), 4647–4654. doi: 10.1093/molbev/msab199.

[41] Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology* 215 (1990), 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. *Journal of Heredity* 105 (2014), 1–18. doi: 10.1093/jhered/est084.

[43] Voolstra CR, Scientists (COS) GC of, Wörheide G, and Lopez JV. Advancing genomics through the Global Invertebrate Genomics Alliance (GIGA). *Invertebrate Systematics* 31 (2017), 1. doi: 10.1071/is16059.

[44] Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, et al. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences* 115 (2018), 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] Formenti G et al. The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution* (2022). doi: https://doi.org/10.1016/j.tree.2021.11.008.

[48] Sanger F, Nicklen S, and Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* 74 (1977), 5463–5467. doi: 10.1073/pnas.74.12.5463.

[49] Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. Life with 6000 genes. *Science* 274 (1996), 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. *Science* 282 (1998), 2012–2018. doi: 10.1126/science.282.5396.2012.

[51] Wajid B, Sohail MU, Ekti AR, and Serpedin E. The A, C, G, and T of genome assembly. *BioMed Research International* 2016 (May 2016), 1–10. doi: 10.1155/2016/6329217.

[52] International Human Genome Sequencing Consortium and others. Initial sequencing and analysis of the human genome. *Nature* 409 (2001), 860–921. doi: 10.1038/35057062.

[53] Pop M and Salzberg SL. Bioinformatics challenges of new sequencing technology. *Trends in Genetics* 24 (2008), 142–149. doi: 10.1016/j.tig.2007.12.006.

[54] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437 (2005), 376–380. doi: 10.1038/nature03959.

[55] Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475 (2011), 348–352. doi: 10.1038/nature10242.

[56] McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, Coleman BE, Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A, MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, De La Vega FM, and Blanchard AP. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research* 19 (2009), 1527–1541. doi: 10.1101/gr.091868.109.

[57] Metzker ML. Sequencing technologies — the next generation. *Nature Reviews Genetics* 11 (2010), 31–46. doi: 10.1038/nrg2626.

[58] Bentley DR et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456 (2008), 53–59. doi: 10.1038/nature07517.

[59] Pollard MO, Gurdasani D, Mentzer AJ, Porter T, and Sandhu MS. Long reads: their purpose and place. *Human Molecular Genetics* 27 (2018), R234–R241. doi: 10.1093/hmg/ddy177.

[60] Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, DeWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, and Turner S. Real-time DNA sequencing from single polymerase molecules. *Science* 323 (2009), 133–138. doi: 10.1126/science.1162986.

[61] Wenger AM, Peluso P, Rowell WJ, Chang Pc, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin Cs, Phillippy AM, Schatz MC, Myers G, Depristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S, Carroll A, Rank DR, and Hunkapiller MW. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* 37 (2019), 1155–1162. doi: 10.1038/s41587-019-0217-9.

[62] Deamer D, Akeson M, and Branton D. Three decades of Nanopore sequencing. *Nature Biotechnology* 34 (2016), 518–524. doi: 10.1038/nbt.3423.

[63] Jain M, Olsen HE, Paten B, and Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* 17 (2016), 1–11. doi: 10.1186/s13059-016-1103-0.

[64] Wick RR, Judd LM, and Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology* 20 (2019), 1–10. doi: 10.1186/s13059-019-1727-y.

[65] Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, and Loose M. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* 36 (2018), 338–345. doi: 10.1038/nbt.4060.

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

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

[68] Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, and Albertsen M. Oxford Nanopore R10.4 long-read sequencing enables near-perfect bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *bioRxiv* (2021). doi: 10.1101/2021.10.27.466057.

[69] Angthong P, Uengwetwanit T, Pootakham W, Sittikankaew K, Sonthirod C, Sangsrakru D, Yoocha T, Nookaew I, Wongsurawat T, Jenjaroenpun P, et al. Optimization of high molecular weight DNA extraction methods in shrimp for a long-read sequencing platform. *PeerJ* 8 (2020), e10340. doi: 10.7717/peerj.10340.

[70] Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, and Lander ES. ARACHNE: a whole-genome shotgun assembler. *Genome Research* 12 (2002), 177–189. doi: 10.1101/gr.208902.

[71] Havlak P, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Weinstock GM, and Gibbs RA. The Atlas genome assembly system. *Genome Research* 14 (2004), 721–732. doi: 10.1101/gr.2264004.

[72] Huang X and Madan A. CAP3: a DNA sequence assembly program. *Genome Research* 9 (1999), 868–877. doi: 10.1101/gr.9.9.868.

[73] Denisov G, Walenz B, Halpern AL, Miller J, Axelrod N, Levy S, and Sutton G. Consensus generation and variant detection by Celera Assembler. *Bioinformatics* 24 (2008), 1035–1040. doi: 10.1093/bioinformatics/btn074.

[74] Pevzner PA, Tang H, and Waterman MS. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences* 98 (2001), 9748–9753. doi: 10.1073/pnas.171285098.

[75] Aparicio S, Chapman J, Stupka E, Putnam N, Chia Jm, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, et al. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297 (2002), 1301–1310. doi: 10.1126/science.1072104.

[76] Sommer DD, Delcher AL, Salzberg SL, and Pop M. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* 8 (2007), 1–11. doi: 10.1186/1471-2105-8-64.

[77] Chevreux B, Wetter T, Suhai S, et al. Genome sequence assembly using trace signals and additional sequence information. In: *German Conference on Bioinformatics*. Vol. 99. Citeseer. 1999, pp. 45–56. doi: 10.1.1.23.7465.

[78] Ewing B and Green P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Research* 8 (1998), 186–194. doi: 10.1101/gr.8.3.186.

[79] Mullikin JC and Ning Z. The Phusion assembler. *Genome Research* 13 (2003), 81–90. doi: 10.1101/gr.731003.

[80] Narzisi G and Mishra B. Scoring-and-unfolding trimmed tree assembler: concepts, constructs and comparisons. *Bioinformatics* 27 (Nov. 2010), 153–160. doi: 10.1093/bioinformatics/btq646.

[81] Sutton G, White O, Adams MD, and Kerlavage AR. TIGR Assembler: A new tool for assembling large shotgun projects. *Genome Science and Technology* 1 (1995), 9–19. doi: 10.1089/gst.1995.1.9.

[82] Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, and Birol I. ABySS: A parallel assembler for short read sequence data. *Genome Research* 19 (2009), 1117–1123. doi: 10.1101/gr.089532.108.

[83] Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, and Birol I. ABySS 2.0: resource-Efficient Assembly of Large Genomes using a Bloom Filter Effect of Bloom Filter False Positive Rate. *Genome Research* 27 (2017), 768–777. doi: 10.1101/gr.214346.116.

[84] Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, and Jaffe DB. ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Research* 18 (2008), 810–820. doi: 10.1101/gr.7337908.

[85] Liu B, Liu CM, Li D, Li Y, Ting HF, Yiu SM, Luo R, and Lam TW. BASE: a practical *de novo* assembler for large genomes using long NGS reads. *BMC Genomics* 17 (2016), 561–569. doi: 10.1186/s12864-016-2829-5.

[86] Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, and Sutton G. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* 24 (2008), 2818–2824. doi: 10.1093/bioinformatics/btn548.

[87] Hernandez D, François P, Farinelli L, Østerås M, and Schrenzel J. *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research* 18 (2008), 802–809. doi: 10.1101/gr.072033.107.

[88] Luo J, Wang J, Zhang Z, Wu FX, Li M, and Pan Y. EPGA: *de novo* assembly using the distributions of reads and insert size. *Bioinformatics* 31 (2015), 825–833. doi: 10.1093/bioinformatics/btu762.

[89] Chaisson MJ and Pevzner PA. Short read fragment assembly of bacterial genomes. *Genome Research* 18 (2008), 324–330. doi: 10.1101/gr.7088808.

[90] Conway T, Wazny J, Bromage A, Zobel J, and Beresford-Smith B. Gossamer — a resource-efficient *de novo* assembler. *Bioinformatics* 28 (2012), 1937–1938. doi: 10.1093/bioinformatics/bts297.

[91] Peng Y, Leung HCM, Yiu SM, and Chin FYL. IDBA - a practical iterative De Bruijn graph *de novo* assembler. *Research in Computational Molecular Biology* 6044 LNBI (2010), 426–440. doi: 10.1007/978-3-642-12683-3_28.

[92] Li M, Liao Z, He Y, Wang J, Luo J, and Pan Y. ISEA: Iterative seed-extension algorithm for *de novo* assembly using paired-end information and insert size distribution. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 14 (2016), 916–925. doi: 10.1109/TCBB.2016.2550433.

[93] Chu TC, Lu CH, Liu T, Lee GC, Li WH, and Shih ACC. Assembler for *de novo* assembly of large genomes. *Proceedings of the National Academy of Sciences* 110 (2013), E3417–E3424. doi: 10.1073/pnas.1314090110.

[94] El-Metwally S, Zakaria M, and Hamza T. LightAssembler: fast and memory-efficient assembly algorithm for high-throughput sequencing reads. *Bioinformatics* 32 (2016), 3215–3223. doi: 10.1093/bioinformatics/btw470.

[95] Chapman JA, Ho I, Sunkara S, Luo S, Schroth GP, and Rokhsar DS. Meraculous: *de novo* genome assembly with short paired-end reads. *PloS One* 6 (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] Huang X, Wang J, Aluru S, Yang SP, and Hillier L. PCAP: a whole-genome assembly program. *Genome Research* 13 (2003), 2164–2170. doi: 10.1101/gr.1390403.

[98] Zhu X, Leung HC, Chin FY, Yiu SM, Quan G, Liu B, and Wang Y. PERGA: a paired-end read guided *de novo* assembler for extending contigs using SVM and look ahead approach. *PloS One* 9 (2014), e114253. doi: 10.1371/journal.pone.0114253.

[99] Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, et al. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research* 24 (2014), 1384–1395. doi: 10.1101/gr.170720.113.

[100] Ariyaratne PN and Sung WK. PE-Assembler: *de novo* assembler using short paired-end reads. *Bioinformatics* 27 (2011), 167–174. doi: 10.1093/bioinformatics/btq626.

[101] Bryant DW, Wong WK, and Mockler TC. QSRA – a quality-value guided *de novo* short read assembler. *BMC Bioinformatics* 10 (2009), 1–6. doi: 10.1186/1471-2105-10-69.

[102] Boisvert S, Laviolette F, and Corbeil J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology* 17 (2010), 1519–1533. doi: 10.1089/cmb.2009.0238.

[103] Gonnella G and Kurtz S. Readjoiner: a fast and memory efficient string graph-based sequence assembler. *BMC Bioinformatics* 13 (2012), 1–19. doi: 10.1186/1471-2105-13-82.

[104] Simpson JT and Durbin R. Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Research* 22 (2012), 549–556. doi: 10.1101/gr.126953.111.

[105] Dohm JC, Lottaz C, Borodina T, and Himmelbauer H. SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing. *Genome Research* 17 (2007), 1697–1706. doi: 10.1101/gr.6435207.

[106] Li R, Zhu H, Ruan J, Qian W, Fang X, Shil Z, Lil Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, and Wang J. *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research* 20 (2010), 265–272. doi: 10.1101/gr.097261.109.

[107] Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* 1 (2012), 1–6. doi: 10.1186/2047-217X-1-18.

[108] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, and Pevzner PA. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19 (2012), 455–477. doi: 10.1089/cmb.2012.0021.

[109] Ye C, Ma ZS, Cannon CH, Pop M, and Douglas WY. 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] Warren RL, Sutton GG, Jones SJ, and Holt RA. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* 23 (2007), 500–501. doi: 10.1093/bioinformatics/btl629.

[111] Schmidt B, Sinha R, Beresford-Smith B, and Puglisi SJ. A fast hybrid short read fragment assembly algorithm. *Bioinformatics* 25 (2009), 2279–2280. doi: 10.1093/bioinformatics/btp374.

[112] Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangl JL, and Jones CD. Extending assembly of short DNA sequences to handle error. *Bioinformatics* 23 (2007), 2942–2944. doi: 10.1093/bioinformatics/btm451.

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

[114] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, and Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Research* 27 (2017), 722–736. doi: 10.1101/gr.215087.116.

[115] Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* 13 (2016), 1050–1054. doi: 10.1038/nmeth.4035.

[116] Kolmogorov M, Yuan J, Lin Y, and Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* 37 (2019), 540–546. doi: 10.1038/s41587-019-0072-8.

[117] Kamath GM, Shomorony I, Xia F, Courtade TA, and David NT. HINGE: long-read assembly achieves optimal repeat resolution. *Genome Research* 27 (2017), 747–756. doi: 10.1101/gr.216465.116.

[118] Xiao CL, Chen Y, Xie SQ, Chen KN, Wang Y, Han Y, Luo F, and Xie Z. MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. *Nature Methods* 14 (2017), 1072–1074. doi: 10.1038/nmeth.4432.

[119] Li H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics* 32 (2016), 2103–2110. doi: 10.1093/bioinformatics/btw152.

[120] Chen Y, Nie F, Xie SQ, Zheng YF, Dai Q, Bray T, Wang YX, Xing JF, Huang ZJ, Wang DP, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nature Communications* 12 (2021), 1–10. doi: 10.1038/s41467-020-20236-7.

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

[122] Vaser R and Šikić M. Yet another *de novo* genome assembler. *International Symposium on Image and Signal Processing and Analysis, ISPA* (2019), 147–151. doi: 10.1109/ISPA.2019.8868909.

[123] Vaser R and Šikić M. Time-and memory-efficient genome assembly with Raven. *Nature Computational Science* 1 (2021), 332–336. doi: 10.1038/s43588-021-00073-4.

[124] Shafin K, Pesout T, Lorig-roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, Sedlazeck FJ, Marschall T, Mayes S, Costa V, Zook JM, Liu KJ, Kilburn D, Sorensen M, Munson KM, Vollger MR, Monlong J, Garrison E, Eichler EE, Salama S, Haussler D, Green RE, Akeson M, Phillippy A, Miga KH, Carnevali P, Jain M, and Paten B. Nanopore sequencing and the Shasta toolkit enable efficient *de novo* assembly of eleven human genomes. *Nature Biotechnology* 38 (2020), 1044–1053. doi: 10.1038/s41587-020-0503-6.

[125] Liu H, Wu S, Li A, and Ruan J. SMARTdenovo: A *de novo* assembler using long noisy reads. *Preprints* (2020). doi: 10.20944/preprints202009.0207.v1.

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

[127] Ruan J and Li H. Fast and accurate long-read assembly with wtdbg2. en. *Nature Methods* 17 (2020), 155–158. doi: 10.1038/s41592-019-0669-3.

[128] Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, and Koren S. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research* 30 (2020), 1291–1305. doi: 10.1101/gr.263566.120.

[129] Cheng H, Concepcion GT, Feng X, Zhang H, and Li H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods* 18 (2021), 1–6. doi: 10.1038/s41592-020-01056-5.

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

[131] Bankevich A, Bzikadze A, Kolmogorov M, Antipov D, and Pevzner PA. LJA: Assembling long and accurate reads using multiplex de Bruijn graphs. *bioRxiv* (2021). doi: 10.1101/2020.12.10.420448.

[132] Ekim B, Berger B, and Chikhi R. Minimizer-space de Bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer. *Cell Systems* 12 (2021), 958–968. doi: 10.1016/j.cels.2021.08.009.

[133] Rautiainen M and Marschall T. MBG: Minimizer-based sparse de Bruijn graph construction. *Bioinformatics* 37 (2021), 2476–2478. doi: 10.1093/bioinformatics/btab004.

[134] Chin CS and Khalak A. Human genome assembly in 100 minutes. *bioRxiv* (2019). doi: 10.1101/705616.

[135] Tarhio J and Ukkonen E. A greedy approximation algorithm for constructing shortest common superstrings. *Theoretical Computer Science* 57 (1988), 131–145. doi: 10.1016/0304-3975(88)90167-3.

[136] Dierckxsens N, Mardulyn P, and Smits G. NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Research* 45 (2017), e18. doi: 10.1093/nar/gkw955.

[137] Staden R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research* 6 (1979), 2601–2610. doi: 10.1093/nar/6.7.2601.

[138] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. The Sequence of the Human Genome. *Science* 291 (2001), 1304–1351. doi: 10.1126/science.1058040.

[139] Bruijn NG de. A Combinatorial Problem. *Koninklijke Nederlandse Akademie v. Wetenschappen* 49 (1946), 758–764.

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

[141] Compeau PE, Pevzner PA, and Tesler G. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology* 29 (2011), 987–991. doi: 10.1038/nbt.2023.

[142] Yahalomi D, Atkinson SD, Neuhof M, Chang ES, Philippe H, Cartwright P, Bartholomew JL, and Huchon D. A cnidarian parasite of salmon (Myxozoa: Henneguya) lacks a mitochondrial genome. *Proceedings of the National Academy of Sciences* 117 (2020), 5358–5363. doi: 10.1073/pnas.1909907117.

[143] Vogg MC, Beccari L, Iglesias Ollé L, Rampon C, Vriz S, Perruchoud C, Wenger Y, and Galliot B. An evolutionarily-conserved Wnt3/β-catenin/Sp5 feedback loop restricts head organizer activity in Hydra. *Nature Communications* 10 (2019), 1–15. doi: 10.1038/s41467-018-08242-2.

[144] Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P, and Przeworski M. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biology* 10 (2012), e1001388. doi: 10.1371/journal.pbio.1001388.

[145] Wick RR. Filtlong, https://github.com/rrwick/Filtlong. 2017.

[146] Haghshenas E, Hach F, Sahinalp SC, and Chauve C. CoLoRMap: Correcting Long Reads by Mapping short reads. *Bioinformatics* 32 (2016), i545–i551. doi: 10.1093/bioinformatics/btw463.

[147] Firtina C, Bar-Joseph Z, Alkan C, and Cicek AE. Hercules: a profile HMM-based hybrid error correction algorithm for long reads. *Nucleic Acids Research* 46 (2018), e125. doi: 10.1093/nar/gky724.

[148] Morisse P, Lecroq T, and Lefebvre A. Hybrid correction of highly noisy long reads using a variable-order de Bruijn graph. *Bioinformatics* 34 (June 2018), 4213–4222. doi: 10.1093/bioinformatics/bty521.

[149] Miclotte G, Heydari M, Demeester P, Rombauts S, Van de Peer Y, Audenaert P, and Fostier J. Jabba: hybrid error correction for long sequencing reads. *Algorithms for Molecular Biology* 11 (2016), 1–12. doi: 10.1186/s13015-016-0075-7.

[150] Salmela L and Rivals E. LoRDEC: Accurate and efficient long read error correction. *Bioinformatics* 30 (2014), 3506–3514. doi: 10.1093/bioinformatics/btu538.

[151] Salmela L, Walve R, Rivals E, Ukkonen E, and Sahinalp C. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics* 33 (2017), 799–806. doi: 10.1093/bioinformatics/btw321.

[152] Madoui MA, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, Lemainque A, Wincker P, and Aury JM. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* 16 (2015), 1–11. doi: 10.1186/s12864-015-1519-z.

[153] Hackl T, Hedrich R, Schultz JS, and Förster F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* 30 (2014), 3004–3011. doi: 10.1093/bioinformatics/btu392.

[154] Morisse P, Marchet C, Limasset A, Lecroq T, and Lefebvre A. Scalable long read self-correction and assembly polishing with multiple sequence alignment. *Scientific Reports* 11 (2021), 1–13. doi: 10.1038/s41598-020-80757-5.

[155] Tischler G and Myers EW. Non Hybrid Long Read Consensus Using Local De Bruijn Graph Assembly. *bioRxiv* (2017). doi: 10.1101/106252.

[156] Bao E, Xie F, Song C, and Song D. FLAS: fast and high-throughput algorithm for PacBio long-read self-correction. *Bioinformatics* 35 (2019), 3953–3960. doi: 10.1093/bioinformatics/btz206.

[157] Bao E and Lan L. HALC: High throughput algorithm for long read error correction. *BMC Bioinformatics* 18 (2017), 1–12. doi: 10.1186/s12859-017-1610-3.

[158] Warren RL, Coombe L, Mohamadi H, Zhang J, Jaquish B, Isabel N, Jones SJ, Bousquet J, Bohlmann J, and Birol I. ntEdit: scalable genome sequence polishing. *Bioinformatics* 35 (2019), 4430–4432. doi: 10.1093/bioinformatics/btz400.

[159] Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* 9 (2014), e112963. doi: 10.1371/journal.pone.0112963.

[160] Zimin AV and Salzberg SL. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLOS Computational Biology* 16 (June 2020), 1–8. doi: 10.1371/journal.pcbi.1007981.

[161] Firtina C, Kim JS, Alser M, Cali DS, Cicek AE, Alkan C, and Mutlu O. Apollo: a sequencing-technology-independent, scalable, and accurate assembly polishing algorithm. *Bioinformatics* (2020). doi: 10.1093/bioinformatics/btaa179.

[162] Aury JM and Istace B. Hapo-G, haplotype-aware polishing of genome assemblies. *NAR Genomics and Bioinformatics* 3 (May 2021). doi: 10.1093/nargab/lqab034.

[163] Ritu Kundu, Joshua Casey WkS. HyPo : super fast & accurate polisher for long read assemblies. *bioRxiv* (2019). doi: 10.1101/2019.12.19.882506.

[164] Vaser R, Sović I, Nagarajan N, and Šikić M. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Research* 27 (2017), 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] Hu J, Fan J, Sun Z, and Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* 36 (2019), 2253–2255. doi: 10.1093/bioinformatics/btz891.

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

[169] Huang S, Kang M, and Xu A. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics* 33 (2017), 2577–2579. doi: 10.1093/bioinformatics/btx220.

[170] Guan D, McCarthy SA, Wood J, Howe K, Wang Y, and Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* 36 (2020), 2896–2898. doi: 10.1093/bioinformatics/btaa025.

[171] Roach MJ, Schmidt SA, and Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* 19 (2018), 1–10. doi: 10.1186/s12859-018-2485-7.

[172] Pop M, Kosack DS, and Salzberg SL. Hierarchical scaffolding with Bambus. *Genome Research* 14 (2004), 149–159. doi: 10.1101/gr.1536204.

[173] Mandric I and Zelikovsky A. Solving scaffolding problem with repeats. *bioRxiv* (2018). doi: 10.1101/330472.

[174] Sahlin K, Vezzi F, Nystedt B, Lundeberg J, and Arvestad L. BESST - Efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics* 15 (2014), 1–11. doi: 10.1186/1471-2105-15-281.

[175] Luo J, Wang J, Zhang Z, Li M, and Wu FX. BOSS: a novel scaffolding algorithm based on an optimized scaffold graph. *Bioinformatics* 33 (2017), 169–176. doi: 10.1093/bioinformatics/btw597.

[176] Gritsenko AA, Nijkamp JF, Reinders MJ, and Ridder Dd. GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies. *Bioinformatics* 28 (2012), 1429–1437. doi: 10.1093/bioinformatics/bts175.

[177] Salmela L, Mäkinen V, Välimäki N, Ylinen J, and Ukkonen E. Fast scaffolding with small independent mixed integer programs. *Bioinformatics* 27 (2011), 3259–3265. doi: 10.1093/bioinformatics/btr562.

[178] Gao S, Sung WK, and Nagarajan N. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *Journal of Computational Biology* 18 (2011), 1681–1691. doi: 10.1089/cmb.2011.0170.

[179] Mandric I and Zelikovsky A. ScaffMatch: scaffolding algorithm based on maximum weight matching. *Bioinformatics* 31 (2015), 2632–2638. doi: 10.1093/bioinformatics/btv211.

[180] Bodily PM, Fujimoto MS, Snell Q, Ventura D, and Clement MJ. ScaffoldScaffolder: solving contig orientation via bidirected to directed graph reduction. *Bioinformatics* 32 (2016), 17–24. doi: 10.1093/bioinformatics/btv548.

[181] Donmez N and Brudno M. SCARPA: scaffolding reads with practical algorithms. *Bioinformatics* 29 (2013), 428–434. doi: 10.1093/bioinformatics/bts716.

[182] Li M, Tang L, Wu FX, Pan Y, and Wang J. SCOP: a novel scaffolding algorithm based on contig classification and optimization. *Bioinformatics* 35 (2019), 1142–1150. doi: 10.1093/bioinformatics/bty773.

[183] Roy RS, Chen KC, Sengupta AM, and Schliep A. SLIQ: Simple Linear Inequalities for Efficient Contig Scaffolding. *Journal of Computational Biology* 19 (2012), 1162–1175. doi: 10.1089/cmb.2011.0263.

[184] Dayarian A, Michael TP, and Sengupta AM. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. *BMC Bioinformatics* 11 (2010), 1–21. doi: 10.1186/1471-2105-11-345.

[185] Boetzer M, Henkel CV, Jansen HJ, Butler D, and Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27 (2011), 578–579. doi: 10.1093/bioinformatics/btq683.

[186] Farrant GK, Hoebeke M, Partensky F, Andres G, Corre E, and Garczarek L. WiseScaffolder: an algorithm for the semi-automatic scaffolding of next generation sequencing data. *BMC Bioinformatics* 16 (2015), 1–13. doi: 10.1186/s12859-015-0705-y.

[187] Ludwig A, Pippel M, Myers G, and Hiller M. DENTIST — using long reads for closing assembly gaps at high accuracy. *GigaScience* 11 (Jan. 2022). doi: 10.1093/gigascience/giab100.

[188] Schmeing S and Robinson MD. Gapless provides combined scaffolding, gap filling and assembly correction with long reads. *bioRxiv* (2022). doi: 10.1101/2022.03.08.483466.

[189] Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJ, and Birol I. LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience* 4 (2015). doi: 10.1186/s13742-015-0076-3.

[190] Qin M, Wu S, Li A, Zhao F, Feng H, Ding L, and Ruan J. LRScaf: improving draft genomes using long noisy reads. *BMC Genomics* 20 (2019), 1–12. doi: 10.1186/s12864-019-6337-2.

[191] Cao MD, Nguyen SH, Ganesamoorthy D, Elliott AG, Cooper MA, and Coin LJ. Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nature Communications* 8 (2017), 1–10. doi: 10.1038/ncomms14515.

[192] English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, et al. Mind the Gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS One* 7 (2012), e47768. doi: 10.1371/journal.pone.0047768.

[193] Warren RL. RAILS and Cobbler: Scaffolding and automated finishing of draft genomes using long DNA sequences. *Journal of Open Source Software* 1 (2016), 116. doi: 10.21105/joss.00116.

[194] Luo J, Lyu M, Chen R, Zhang X, Luo H, and Yan C. SLR: a scaffolding algorithm based on long reads and contig classification. *BMC Bioinformatics* 20 (2019), 1–11. doi: 10.1186/s12859-019-3114-9.

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

[196] Zhu S, Chen DZ, and Emrich SJ. Single molecule sequencing-guided scaffolding and correction of draft assemblies. *BMC Genomics* 18 (2017), 51–59. doi: 10.1186/s12864-017-4271-8.

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

[198] Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, Schnable PS, Lyons E, and Lu J. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biology* 16 (2015), 1–15. doi: 10.1186/s13059-014-0573-1.

[199] Lin HC, Goldstein S, Mendelowitz L, Zhou S, Wetzel J, Schwartz DC, and Pop M. AGORA: assembly guided by optical restriction alignment. *BMC Bioinformatics* 13 (2012), 1–14. doi: 10.1186/1471-2105-13-189.

[200] Istace B, Belser C, and Aury JM. BiSCoT: improving large eukaryotic genome assemblies with optical maps. *PeerJ* 8 (2020), e10150. doi: 10.7717/peerj.10150.

[201] Pan W, Jiang T, and Lonardi S. OMGS: optical map-based genome scaffolding. *Journal of Computational Biology* 27 (2020), 519–533. doi: 10.1089/cmb.2019.0310.

[202] Shelton JM, Coleman MC, Herndon N, Lu N, Lam ET, Anantharaman T, Sheth P, and Brown SJ. Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC Genomics* 16 (2015). doi: 10.1186/s12864-015-1911-8.

[203] Nagarajan N, Read TD, and Pop M. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics* 24 (2008), 1229–1235. doi: 10.1093/bioinformatics/btn102.

[204] Hiltunen M, Ryberg M, and Johannesson H. ARBitR: an overlap-aware genome assembly scaffolder for linked reads. *Bioinformatics* 37 (2020), 2203–2205. doi: 10.1093/bioinformatics/btaa975.

[205] Kuleshov V, Snyder MP, and Batzoglou S. Genome assembly from synthetic long read clouds. *Bioinformatics* 32 (2016), i216–i224. doi: 10.1093/bioinformatics/btw267.

[206] Yeo S, Coombe L, Warren RL, Chu J, and Birol I. ARCS: scaffolding genome drafts with linked reads. *Bioinformatics* 34 (2018), 725–731. doi: 10.1093/bioinformatics/btx675.

[207] Coombe L, Zhang J, Vandervalk BP, Chu J, Jackman SD, Birol I, and Warren RL. ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics* 19 (2018). doi: 10.1186/s12859-018-2243-x.

[208] Adey A, Kitzman JO, Burton JN, Daza R, Kumar A, Christiansen L, Ronaghi M, Amini S, Gunderson KL, Steemers FJ, et al. *In vitro*, long-range sequence information for *de novo* genome assembly via transposase contiguity. *Genome Research* 24 (2014), 2041–2049. doi: 10.1101/gr.178319.114.

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

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

[211] Marie-Nelly H, Marbouty M, Cournac A, Flot JF, Liti G, Parodi DP, Syan S, Guillén N, Margeot A, Zimmer C, et al. High-quality genome (re)assembly using chromosomal contact data. *Nature Communications* 5 (2014). doi: 10.1038/ncomms6695.

[212] Renschler G, Richard G, Valsecchi CIK, Toscano S, Arrigoni L, Ramırez F, and Akhtar A. Hi-C guided assemblies reveal conserved regulatory topologies on X and autosomes despite extensive genome shuffling. *Genes & Development* 33 (2019), 1591–1612. doi: 10.1101/gad.328971.119.

[213] Baudry L, Guiglielmoni N, Marie-Nelly H, Cormier A, Marbouty M, Avia K, Mie YL, Godfroy O, Sterck L, Cock JM, et al. instaGRAAL: chromosome-level quality scaffolding of genomes using a proximity ligation-based scaffolder. *Genome Biology* 21 (2020). doi: 10.1186/s13059-020-02041-z.

[214] Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, and Shendure J. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology* 31 (2013), 1119–1125. doi: 10.1038/nbt.2727.

[215] Guan D, McCarthy SA, Ning Z, Wang G, Wang Y, and Durbin R. Efficient iterative Hi-C scaffolder based on N-best neighbors. *BMC Bioinformatics* 22 (2021), 1–16. doi: 10.1186/s12859-021-04453-5.

[216] Ghurye J, Pop M, Koren S, Bickhart D, and Chin CS. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* 18 (2017). doi: 10.1186/s12864-017-3879-z.

[217] Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, and Koren S. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Computational Biology* 15 (2019). doi: 10.1371/journal.pcbi.1007273.

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

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

[220] Boetzer M and Pirovano W. Toward almost closed genomes with GapFiller. *Genome Biology* 13 (2012), 1–9. doi: 10.1186/gb-2012-13-6-r56.

[221] Chu C, Li X, and Wu Y. GAPPadder: a sensitive approach for closing gaps on draft genomes with short sequence reads. *BMC Genomics* 20 (2019). doi: 10.1186/s12864-019-5703-4.

[222] Paulino D, Warren RL, Vandervalk BP, Raymond A, Jackman SD, and Birol I. Sealer: a scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics* 16 (2015), 1–8. doi: 10.1186/s12859-015-0663-4.

[223] Piro VC, Faoro H, Weiss VA, Steffens MB, Pedrosa FO, Souza EM, and Raittz RT. FGAP: an automated gap closing tool. *BMC Research Notes* 7 (2014). doi: 10.1186/1756-0500-7-371.

[224] Kosugi S, Hirakawa H, and Tabata S. GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. *Bioinformatics* 31 (2015), 3733–3741. doi: 10.1093/bioinformatics/btv465.

[225] Xu GC, Xu TJ, Zhu R, Zhang Y, Li SQ, Wang HW, and Li JT. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *GigaScience* 8 (2019), 1–14. doi: 10.1093/gigascience/giy157.

[226] Lu P, Jin J, Li Z, Xu Y, Hu D, Liu J, and Cao P. PGcloser: fast parallel gap-closing tool using long-reads or contigs to fill gaps in genomes. *Evolutionary Bioinformatics* 16 (2020). doi: 10.1177/1176934320913859.

[227] Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, Fan G, Liu X, Xu X, Deng L, et al. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* 9 (2020), 1–11. doi: 10.1093/gigascience/giaa094.

[228] Delahaye C and Nicolas J. Sequencing DNA with nanopores: Troubles and biases. *PloS One* 16 (2021). doi: 10.1371/journal.pone.0257521.

[229] Morisse P, Lecroq T, and Lefebvre A. Long-read error correction: a survey and qualitative comparison. *bioRxiv* (2020). doi: doi.org/10.1101/2020.03.06.977975.

[230] Ko BJ, Lee C, Kim J, Rhie A, Yoo D, Howe K, Wood J, Cho S, Brown S, Formenti G, et al. Widespread false gene gains caused by duplication errors in genome assemblies. *bioRxiv* (2021). doi: 10.1101/2021.04.09.438957.

[231] Guiglielmoni N, Houtain A, Derzelle A, Van Doninck K, and Flot JF. Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. *BMC Bioinformatics* 22 (2021), 1–23. doi: 10.1186/s12859-021-04118-3.

[232] Kent WJ and Haussler D. Assembly of the working draft of the human genome with GigAssembler. *Genome Research* 11 (2001), 1541–1548. doi: 10.1101/gr.183201.

[233] Ghurye J and Pop M. Modern technologies and algorithms for scaffolding assembled genomes. *PLoS Computational Biology* 15 (2019), 1–20. doi: 10.1371/journal.pcbi.1006994.

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

[235] Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, and Wang YK. Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* 262 (1993), 110–114. doi: 10.1126/science.8211116.

[236] Protasio AV, Tsai IJ, Babbage A, Nichol S, Hunt M, Aslett MA, Silva N de, Velarde GS, Anderson TJ, Clark RC, Davidson C, Dillon GP, Holroyd NE, LoVerde PT, Lloyd C, McQuillan J, Oliveira G, Otto TD, Parker-Manuel SJ, Quail MA, Wilson RA, Zerlotini A, Dunne DW, and Berriman M. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS Neglected Tropical Diseases* 6 (2012). doi: 10.1371/journal.pntd.0001455.

[237] Jeong CB, Lee BY, Choi BS, Kim MS, Park JC, Kim DH, Wang M, Park HG, and Lee JS. The genome of the harpacticoid copepod *Tigriopus japonicus*: potential for its use in marine molecular ecotoxicology. *Aquatic Toxicology* 222 (2020), 105462. doi: 10.1016/j.aquatox.2020.105462.

[238] Yuan Y, Chung CYL, and Chan TF. Advances in optical mapping for genomic research. *Computational and Structural Biotechnology Journal* 18 (2020), 2051–2062. doi: 10.1016/j.csbj.2020.07.018.

[239] Cotton JA, Bennuru S, Grote A, Harsha B, Tracey A, Beech R, Doyle SR, Dunn M, Hotopp JCD, Holroyd N, et al. The genome of *Onchocerca volvulus*, agent of river blindness. *Nature Microbiology* 2 (2016), 1–12. doi: 10.1038/nmicrobiol.2016.216.

[240] Wang J, Gao S, Mostovoy Y, Kang Y, Zagoskin M, Sun Y, Zhang B, White LK, Easton A, Nutman TB, Kwok PY, Hu S, Nielsen MK, and Davis RE. Comparative genome analysis of programmed DNA elimination in nematodes. *Genome Research* 27 (2017), 2001–2014. doi: 10.1101/gr.225730.117.

[241] Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL, Tracey A, Bobes RJ, Fragoso G, Sciutto E, Aslett M, Beasley H, Bennett HM, Cai J, Camicia F, Clark R, Cucher M, De Silva N, Day TA, Deplazes P, Estrada K, Fernández C, Holland PW, Hou J, Hu S, Huckvale T, Hung SS, Kamenetzky L, Keane JA, Kiss F, Koziol U, Lambert O, Liu K, Luo X, Luo Y, MacChiaroli N, Nichol S, Paps J, Parkinson J, Pouchkina-Stantcheva N, Riddiford N, Rosenzvit M, Salinas G, Wasmuth JD, Zamanian M, Zheng Y, Cai X, Soberon X, Olson PD, Laclette JP, Brehm K, Berriman M, Morett E, Portillo T, Jose MV, Carrero JC, Larralde C, Morales-Montor J, Limon-Lason J, Cevallos MA, Gonzalez V, Ochoa-Leyva A, Landa A, Jimenez L, and Valdes V. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* 496 (2013), 57–63. doi: 10.1038/nature12031.

[242] Olson PD, Tracey A, Baillie A, James K, Doyle SR, Buddenborg SK, Rodgers FH, Holroyd N, and Berriman M. Complete representation of a tapeworm genome reveals chromosomes capped by centromeres, necessitating a dual role in segregation and protection. *BMC Biology* 18 (2020), 1–16. doi: 10.1186/s12915-020-00899-w.

[243] Varney RM, Speiser DI, McDougall C, Degnan BM, and Kocot KM. The iron-responsive genome of the chiton *Acanthopleura granulata*. *Genome Biology and Evolution* 13 (2021). doi: 10.1093/gbe/evaa263.

[244] Meier JI, Salazar PA, Kučka M, Davies RW, Dréau A, Aldás I, Power OB, Nadeau NJ, Bridle JR, Rolian C, et al. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proceedings of the National Academy of Sciences* 118 (2021). doi: 10.1073/pnas.2015005118.

[245] Chen Z, Pham L, Wu TC, Mo G, Xia Y, Chang PL, Porter D, Phan T, Che H, Tran H, 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. *Genome Research* 30 (2020), 898–909. doi: 10.1101/gr.260380.119.

[246] Kocher SD, Mallarino R, Rubin BE, Douglas WY, Hoekstra HE, and Pierce NE. The genetic basis of a social polymorphism in halictid bees. *Nature Communications* 9 (2018). doi: 10.1038/s41467-018-06824-8.

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

[248] Ghurye J, Koren S, Small ST, Redmond S, Howell P, Phillippy AM, and Besansky NJ. A chromosome-scale assembly of the major African malaria vector *Anopheles funestus*. *GigaScience* 8 (2019). doi: 10.1093/gigascience/giz063.

[249] Dekker J, Rippe K, Dekker M, and Kleckner N. Capturing chromosome conformation. *Science* 295 (2002), 1306–1311. doi: 10.1126/science.1067799.

[250] Dekker J, Marti-Renom MA, and Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics* 14 (2013), 390–403. doi: 10.1038/nrg3454.

[251] Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, and Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326 (2009), 289–93. doi: 10.1126/science.1181369.

[252] Flot JF, Marie-Nelly H, and Koszul R. Contact genomics: scaffolding and phasing (meta) genomes using chromosome 3D physical signatures. *FEBS Letters* 589 (2015), 2966–2974. doi: 10.1016/j.febslet.2015.04.034.

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

[254] Techer MA, Rane RV, Grau ML, Roberts JM, Sullivan ST, Liachko I, Childers AK, Evans JD, and Mikheyev AS. Divergent evolutionary trajectories following speciation in two ectoparasitic honey bee mites. *Communications Biology* 2 (2019). doi: 10.1038/s42003-019-0606-0.

[255] Shingate P, Ravi V, Prasad A, Tay BH, Garg KM, Chattopadhyay B, Yap LM, Rheindt FE, and Venkatesh B. Chromosome-level assembly of the horseshoe crab genome provides insights into its genome evolution. *Nature Communications* 11 (2020). doi: 10.1038/s41467-020-16180-1.

[256] Darras H, Souza Araujo N de, Baudry L, Guiglielmoni N, Lorite P, Marbouty M, Rodriguez F, Arkhipova I, Koszul R, Flot JF, and Aron S. 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. *bioRxiv* (2022). doi: 10.1101/2022.01.07.475286.

[257] Hu M, Zheng X, Fan CM, and Zheng Y. Lineage dynamics of the endosymbiotic cell type in the soft coral *Xenia*. *Nature* 582 (2020), 534–538. doi: 10.1038/s41586-020-2385-7.

[258] Li Y, Gao L, Pan Y, Tian M, Li Y, He C, Dong Y, Sun Y, and Zhou Z. Chromosome-level reference genome of the jellyfish *Rhopilema esculentum*. *GigaScience* 9 (2020). doi: 10.1093/gigascience/giaa036.

[259] Davidson PL, Guo H, Wang L, Berrio A, Zhang H, Soborowski AL, Mcclay DR, Fan G, and Wray GA. Chromosomal-level genome assembly of the sea urchin *Lytechinus variegatus* substantially improves functional genomic analyses. *Genome Biology and Evolution* 12 (2020), 1080–1086. doi: 10.1093/gbe/evaa101.

[260] Ruiz-Ramos DV, Schiebelhut LM, Hoff KJ, Wares JP, and Dawson MN. An initial comparative genomic autopsy of wasting disease in sea stars. *Molecular Ecology* 29 (2020), 1087–1102. doi: 10.1111/mec.15386.

[261] Bai CM, Xin LS, Rosani U, Wu B, Wang QC, Duan XK, Liu ZH, and Wang CM. Chromosomal-level assembly of the blood clam, *Scapharca (Anadara) broughtonii*, using long sequence reads and Hi-C. *GigaScience* 8 (2019). doi: 10.1093/gigascience/giz067.

[262] Sun J, Chen C, Miyamoto N, Li R, Sigwart JD, Xu T, Sun Y, Wong WC, Ip JC, Zhang W, Lan Y, Bissessur D, Watsuji To, Watanabe HK, Takaki Y, Ikeo K, Fujii N, Yoshitake K, Qiu JW, Takai K, and Qian PY. The scaly-foot snail genome and implications for the origins of biomineralised armour. *Nature Communications* 11 (2020). doi: 10.1038/s41467-020-15522-3.

[263] Farhat S, Bonnivard E, Pales Espinosa E, Tanguy A, Boutet I, Guiglielmoni N, Flot JF, and Allam B. Comparative analysis of the *Mercenaria mercenaria* genome provides insights into the diversity of transposable elements and immune molecules in bivalve mollusks. *BMC Genomics* 23 (2022), 1–23. doi: 10.1186/s12864-021-08262-1.

[264] Teterina AA, Willis JH, and Phillips PC. Chromosome-level assembly of the *Caenorhabditis remanei* genome reveals conserved patterns of nematode genome organization. *Genetics* 214 (2020), 769–780. doi: 10.1534/genetics.119.303018.

[265] Lian Y, Wei H, Wang J, Lei C, Li H, Li J, Wu Y, Wang S, Zhang H, Wang T, et al. Chromosome-level reference genome of X12, a highly virulent race of the soybean cyst nematode *Heterodera glycines*. *Molecular Ecology Resources* 19 (2019), 1637–1646. doi: 10.1111/1755-0998.13068.

[266] Stroehlein AJ, Korhonen PK, Chong TM, Lim YL, Chan KG, Webster B, Rollinson D, Brindley PJ, Gasser RB, and Young ND. High-quality *Schistosoma haematobium* genome achieved by single-molecule and long-range sequencing. *GigaScience* 8 (2019). doi: 10.1093/gigascience/giz108.

[267] Kenny NJ, Francis WR, Rivera-Vicéns RE, Juravel K, Mendoza A de, Dıez-Vives C, Lister R, Bezares-Calderón LA, Grombacher L, Roller M, et al. Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge *Ephydatia muelleri*. *Nature Communications* (2020). doi: 10.1038/s41467-020-17397-w.

[268] Gehrke AR, Neverett E, Luo YJ, Brandt A, Ricci L, Hulett RE, Gompers A, Ruby JG, Rokhsar DS, Reddien PW, et al. Acoel genome reveals the regulatory landscape of whole-body regeneration. *Science* 363 (2019). doi: 10.1126/science.aau6173.

[269] Haghshenas E, Asghari H, Stoye J, Chauve C, and Hach F. HASLR: Fast hybrid assembly of long reads. *iScience* 23 (2020), 101389. doi: 10.1016/j.isci.2020.101389.

[270] Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, and Yorke JA. The MaSuRCA genome assembler. *Bioinformatics* 29 (2013), 2669–2677. doi: 10.1093/bioinformatics/btt476.

[271] Di Genova A, Buena-Atienza E, Ossowski S, and Sagot MF. Efficient hybrid *de novo* assembly of human genomes with WENGAN. *Nature Biotechnology* 39 (2021), 422–430. doi: 10.1038/s41587-020-00747-w.

[272] Mulligan KL, Hiebert TC, Jeffery NW, and Gregory TR. First estimates of genome size in ribbon worms (phylum Nemertea) using flow cytometry and Feulgen image analysis densitometry. *Canadian Journal of Zoology* 92 (2014), 847–851. doi: 10.1139/cjz-2014-0068.

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

[274] Ranallo-Benavidez TR, Jaron KS, and Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11 (2020), 1432. doi: 10.1038/s41467-020-14998-3.

[275] Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, and Clavijo BJ. KAT: a K-mer Analysis Toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* 33 (2016), 574–576. doi: 10.1093/bioinformatics/btw663.

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

[277] Gurevich A, Saveliev V, Vyahhi N, and Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29 (2013), 1072–1075. doi: 10.1093/bioinformatics/btt086.

[278] Guo Y, Zhang Y, Liu Q, Huang Y, Mao G, Yue Z, Abe EM, Li J, Wu Z, Li S, Zhou X, Hu W, and Xiao N. A chromosomal-level genome assembly for the giant African snail *Achatina fulica*. *GigaScience* 8 (2019). doi: 10.1093/gigascience/giz124.

[279] He C, Lin G, Wei H, Tang H, White FF, Valent B, and Liu S. Factorial estimating assembly base errors using *k*-mer abundance difference (KAD) between short reads and genome assembled sequences. *NAR Genomics and Bioinformatics* 2 (2020). doi: 10.1093/nargab/lqaa075.

[280] Laetsch DR and Blaxter ML. BlobTools: Interrogation of genome assemblies. *F1000Research* 6 (2017), 1287. doi: 10.12688/f1000research.12232.1.

[281] Challis R, Richards E, Rajan J, Cochrane G, and Blaxter M. BlobToolKit – Interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics* 10 (2020), 1361–1374. doi: 10.1534/g3.119.400908.

[282] Boothby TC, Tenlen JR, Smith FW, Wang JR, Patanella KA, Nishimura EO, Tintori SC, Li Q, Jones CD, Yandell M, Messina DN, Glasscock J, and Goldstein B. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences of the United States of America* 112 (2015), 15976–15981. doi: 10.1073/pnas.1510461112.

[283] Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, Maroon H, Thomas F, Aboobaker AA, and Blaxter M. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proceedings of the National Academy of Sciences* 113 (2016), 5053–5058. doi: 10.1073/pnas.1600338113.

[284] Zhang X, Wu R, Wang Y, Yu J, and Tang H. Unzipping haplotypes in diploid and polyploid genomes. *Computational and Structural Biotechnology Journal* 18 (2020), 66–72. doi: 10.1016/j.csbj.2019.11.011.

[285] Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TP, and Phillippy AM. *De novo* assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology* 36 (2018), 1174–1182. doi: 10.1038/nbt.4277.

[286] Edge P, Bafna V, and Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research* 27 (2017), 801–812. doi: 10.1101/gr.213462.116.

[287] Patterson MD, Marschall T, Pisanti N, Van Iersel L, Stougie L, Klau GW, and Schönhuth A. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology* 22 (2015), 498–509. doi: 10.1089/cmb.2014.0157.

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

[289] Kajitani R, Yoshimura D, Okuno M, Minakuchi Y, Kagoshima H, Fujiyama A, Kubokawa K, Kohara Y, Toyoda A, and Itoh T. Platanus-allee is a *de novo* haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nature Communications* 10 (2019), 1–15. doi: 10.1038/s41467-019-09575-2.

[290] Porubsky D, Ebert P, Audano PA, Vollger MR, Harvey WT, Marijon P, Ebler J, Munson KM, Sorensen M, Sulovari A, et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nature Biotechnology* 39 (2020), 302–308. doi: 10.1038/s41587-020-0719-5.

[291] Zhou Q, Tang D, Huang W, Yang Z, Zhang Y, Hamilton JP, Visser RG, Bachem CW, Buell CR, Zhang Z, et al. Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nature Genetics* 52 (2020), 1018–1023. doi: 10.1038/s41588-020-0699-x.

[292] Holley G, Beyter D, Ingimundardottir H, Møller PL, Kristmundsdottir S, Eggertsson HP, and Halldorsson BV. Ratatosk: hybrid error correction of long reads enables accurate variant calling and assembly. *Genome Biology* 22 (2021). doi: 10.1186/s13059-020-02244-4.

[293] Zhang X, Zhang S, Zhao Q, Ming R, and Tang H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants* 5 (2019), 833–845. doi: 10.1038/s41477-019-0487-8.

[294] Faure R, Guiglielmoni N, and Flot JF. GraphUnzip: unzipping assembly graphs with long reads and Hi-C. *bioRxiv* (2021). doi: 10.1101/2021.01.29.428779.

[295] Kronenberg ZN, Rhie A, Koren S, Concepcion GT, Peluso P, Munson KM, Porubsky D, Kuhn K, Mueller KA, Low WY, et al. Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nature Communications* 12 (2021), 1–10. doi: 10.1038/s41467-020-20536-y.

[296] Rhie A, Walenz BP, Koren S, and Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* 21 (2020), 1–27. doi: 10.1186/s13059-020-02134-9.

[297] Salzberg SL. Next-generation genome annotation: we still struggle to get it right. *Genome Biology* 20 (2019). doi: 10.1186/s13059-019-1715-2.

[298] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, Silva Santos LB da, Bourne PE, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (2016), 1–9. doi: 10.1038/sdata.2016.18.

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

[300] Guiglielmoni N, Rivera-Vicens RE, Koszul R, and Flot JF. Supplementary table to "A deep dive into genome assemblies of non-vertebrate animals" (Apr. 2022). doi: 10.6084/m9.figshare.19672440.v1.

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