A deep dive into genome assemblies of non-vertebrate animals

Nadège Guiglielmoni, Ramón Rivera-Vicéns, Romain Koszul and Jean-François Flot (2022)
A deep dive into genome assemblies of non-vertebrate animals, Peer Community Journal, 2: e29.
https://doi.org/10.24072/pcjournal.128

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.
Introduction

The field of genomics is presently thriving, with new genomes of all kinds 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 belonged to Cnidaria [29]; the sequence of *Hoilungia hongkongiensis* placed placozoans as a sister group to cnidarians and bilaterians [30]. Genomic studies have also attempted to elucidate the mechanisms underlying asexuality, as sexual reproduction is a character shared by almost all eukaryotes and its strict absence generally leads to rapid extinction due to the accumulation of deleterious mutations [31], yet ancient asexual species are observed in many branches of phylogeny [32–36].

The dearth of non-vertebrate animal genomic resources may be blamed to the difficulty to collect individuals in remote or hardly accessible locations and in accordance with the Nagoya protocol [37]: the scarcity of certain species; non-existing resources to cultivate individuals in laboratories; the lack of protocols to extract pure, high-molecular-weight DNA; their frequently large genomes characterized by high repetitive contents and high heterozygosity. However, sequencing technologies now offer cost-effective solutions and wide applicability to solve some of these problems. Reducing the current unbalance in genomic resources between vertebrates and non-vertebrate animals will increase the precision of future tools and studies. Indeed, genome data are often used as the foundation for different genomic and protein databases. The program BUSCO (Benchmarking Universal Single-Copy Orthologs) [38–40], used to measure the completeness of a genome assembly, relies on reference gene sets that are used for scoring, based on existing assemblies for a group of species. Thus, results from under-sampled groups could change drastically when more species are added to the gene sets. These could also have major effects in analyses such as phylogenomics, protein families studies and of gene duplication events. Another consequence of the current dearth of genomic resources for non-vertebrate animals is that BLAST [41] searches for animal species most often recover vertebrate and arthropod hits, even though the target species is distant from these phyla, hampering the identification of sequences from a species lacking a reference or closely related genome. As a result, identifying metazoan contaminants in a fragmented assembly of an animal genome is almost impossible due to similar GC contents and the absence of hits in genomic databases.

It is therefore imperative to explore thoroughly the diversity of metazoans, specifically from non-vertebrate animal species. International consortia such as the Global Invertebrate Genomics Alliance (GIGA) [42, 43] have been put in place to overcome some of the aforementioned limitations. Other consortia such as the Earth BioGenome Project [44], the Darwin Tree of Life [45], the Aquatic Symbiosis Genomics Project [46] and the European Reference Genome Atlas [47] are also expected to significantly boost the genomic resources of non-vertebrates in the near future. Undoubtedly, these projects will benefit from the drastic improvements in sequencing technologies over the last years. In this review, we first offer a brief historical overview of sequencing technologies and algorithmic approaches to genome assembly. We then survey software for genome assembly, pre/post-processing steps, assembly evaluation, and phasing assemblies, to help newcomers to the field build their own assembly pipelines and have an overview of past and current tools. Although sequencing methods, algorithms and programs presented in this paper are not restricted to a category of organisms, the challenges and solutions that we describe are specific to non-model non-vertebrate animal genomes.
1. Sequencing

Sequencing technologies have dramatically evolved over the last two decades, providing researchers with various options when it comes to tackling a genome project (Table 1). Sanger sequencing, the widely used sequencing method with chain-terminating inhibitors, published in 1977, produces reads around 1,000 basepair (bp) long with an error rate of about 1% [48]. The principle is to synthesize complementary strands of DNA from a single strand with a mixture of regular nucleotides and dideoxynucleotides, the latter stopping the polymerase when incorporated. Four reactions are performed for each type of base, and the resulting oligonucleotides are migrated by electrophoresis to identify the correct base at every position and generate a read. This method laid the foundations for DNA sequencing and was used extensively in several genome assembly projects, which were at that time typically ran by large international consortia: the budding yeast *Saccharomyces cerevisiae* [49] was the first eukaryote sequenced, whereas the nematode *Caenorhabditis elegans* was the first metazoan [50]. Sanger sequencing is a relatively low-throughput method in terms of the number of sequences generated, and is costly as well [51]. Although it is almost not used in genome projects anymore, the technology was pivotal for the generation of the first assembly of the human genome published in 2001, a monumental effort by 20 sequencing centers, to an estimated cost of 300 million US dollars [52].

Second-generation sequencing technologies, initially called next-generation sequencing (NGS), are characterized by a strong increase in sequencing throughputs compared to the Sanger method, with millions of DNA fragments sequenced simultaneously. NGS reads are much smaller than Sanger reads (from 110 bp for the first 454 machine in 2005 up to to 350 bp for MiSeq Illumina machines nowadays), resulting in the need for new analysis algorithms and programs [53]. The arrival of NGS sequencing democratized genome assembly projects, broadening the scope of investigated species beyond well-studied model organisms. Several second-generation sequencing methods have emerged through the years, some of which have since then been discontinued: 454 pyrosequencing [54], Ion Torrent [55], SOLiD [56], and Solexa (for a comparison on the approaches, see [57]). Among these methods, Solexa, subsequently purchased by Illumina [58], became and remains the most widely used approach to this day. This approach consists in amplifying short DNA molecules bound on a flow cell, and sequencing them by sequential addition of fluorescently tagged nucleotides. This protocol generates highly accurate single or paired-end reads with a length up to a few hundred bases. The recent NovaSeq system further increased the output from a single run and abated the cost (up to 3 Terabases per flowcell). Short reads stimulated the whole field of genomics, and led to a large production of assemblies for all sorts of organisms, up to this day (Figure 1). These short-read based assemblies resulted in a tremendous increase of genomic resources, which remained typically quite fragmented (with N50s below 1 Megabase (Mb)).

Third-generation sequencing has brought a whole new range of sequencing data, with the sequencing of long DNA molecules extending up to hundreds of thousands of bases [59]. The two main players in the field, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (Nanopore), use two different kinds of technologies. PacBio developed Single Molecule Real-Time (SMRT) sequencing, where a complementary strand of DNA is produced from a single strand by addition of fluorescently labeled nucleotides. The fluorescent tag is released and the luminescence is interpreted as a base [60]. The resulting reads have a length around twenty kilobases (kb) and a high error rate, an issue recently addressed by the introduction of an extra step called Circular Consensus Sequencing (CCS). In CCS, the DNA polymerase passes multiple times on the same base on a circularized strand to produce High Fidelity (HiFi) reads that can achieve an accuracy over 99%, despite a smaller maximal read length [61].

Nanopore sequencing uses a membrane with protein pores, through which an electrical current is flowing. DNA strands are pulled through the pores, with each passing nucleotide generating a distinct disruption signature in the current that can be inferred as a specific base [62]. The firm has specifically oriented its strategy toward a “do it yourself” approach, enabling sequencing in any lab and even directly in the field via a small portable device [63]. Researchers...
Figure 1 – Contig N50 of 237 non-vertebrate animal genome assemblies over time. The N50 represents the contiguity of an assembly and is defined as the length of the largest contig for which at least 50% of the assembly size is contained in contigs equal or greater in length.

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

Long reads are now routinely included in genome assembly projects and have led to much more contiguous assemblies than short-read only assemblies (Figure 1). A current limitation lies in the amount of DNA required to prepare long-read libraries, and long-read sequencing still remains inaccessible for certain species: whereas Illumina sequencing can handle small DNA amounts, with a poor quality, long-read protocols require high-molecular-weight DNA [69]. PacBio and Nanopore sequencing remain difficult when one animal is too small to provide a sufficient amount of DNA, especially when the organism requires extraction protocols that lead to overly fragmented DNA (for example, with skeletons). In addition, secondary metabolites associated to DNA molecules, or branched DNA structures, can also disturb the sequencing reaction.

2. Genome assembly

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

The assembly problem can be represented as a linear puzzle where the pieces are the reads. Reads match together when they have overlapping sequences. This puzzle could be intuitively solved by iteratively putting together the overlapping pieces that match best: this greedy approach is an efficient heuristic to find the shortest common superstring of the set of reads (i.e., the shortest sequence that includes all the reads as substrings) [135]. Greedy algorithms have been implemented for first-generation sequencing reads, for instance in TIGR [81], and were further applied in short-read assemblers like PERGA [98], SSAKE [110] and VCAKE [112]. However, they cannot resolve complex, repetitive genomes: for this reason, greedy assemblers are mostly used nowadays to assemble small organelle genomes such as chloroplasts and mitochondria [136].
### Table 1 – Sequencing approaches and associated assemblers.

<table>
<thead>
<tr>
<th>First generation</th>
<th>ARACHNE [70], Atlas [71], CAP3 [72], Celera [73], Euler [74], JAZZ [75], Minimus [76], MIRA [77], phrap [78], Phusion [79], SUTTA [80], TIGR [81]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 kb</strong></td>
<td></td>
</tr>
<tr>
<td><strong>High accuracy</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Sanger</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Second generation</strong></td>
<td>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], QSR [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]</td>
</tr>
<tr>
<td><strong>25-300 bp</strong></td>
<td></td>
</tr>
<tr>
<td><strong>High accuracy</strong></td>
<td></td>
</tr>
<tr>
<td><strong>454, IonTorrent, Solexa, SOLiD</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Third generation</strong></td>
<td>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]</td>
</tr>
<tr>
<td><strong>10-100,000+ kb</strong></td>
<td></td>
</tr>
<tr>
<td><strong>PacBio CLR, Nanopore</strong></td>
<td></td>
</tr>
<tr>
<td><strong>15-25 kb</strong></td>
<td>Flye [116], HiCanu [128], hifiasm [129], IPA [130], LJA [131], mdbG [132], MBG [133], NextDenovo [121], Peregrine [134], Raven [123], wtdbg2 [127]</td>
</tr>
</tbody>
</table>
With the advent of third-generation sequencing, OLC assemblers have benefited from a renewed interest whereas DBG-based ones are poorly suited for long, low-accuracy reads, containing many erroneous k-mers. Numerous assemblers have implemented the OLC approach to produce de novo assemblies from error-prone long-read datasets: Flye [116], Ra [122], Raven [123], Shasta [124], wtdbg2 [127]. Now that HiFi reads bring a new type of high-accuracy long reads, assemblers have been adapted to better handle these sequences, such as Flye (with adapted...
Figure 3 – Overview of genome assembly using de Bruijn graphs. A circular genome is assembled based on three reads using node-centric and edge-centric DBGs with $k = 3$. The node-centric DBG is searched for a Hamiltonian cycle (visiting all nodes), and the edge-centric DBG for an Eulerian cycle (visiting all edges). These cycles are represented in blue in the graphs.

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.

3. Assembly pre and post-processing

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

Long reads have the advantage over short reads that they result in more contiguous assemblies. Nevertheless, assemblies of PacBio Continuous Long Reads (CLR) or Nanopore reads can have remaining errors due to their low accuracy; while errors in PacBio CLR are random and are compensated with a high coverage, Nanopore reads have systematic errors in homopolymeric regions [228]. Assemblies of error-prone long reads often necessitate additional processes to increase the quality. There are two possible strategies: correct the long reads prior to assembly, and polish the contigs after assembly. Correcting long reads can be done using only the long reads or by adding high-accuracy short reads. Many tools have been developed for both scenarii
### Table 2 – Assembly pre and post-processing tools for haploid assemblies.

<table>
<thead>
<tr>
<th>Step</th>
<th>Data</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read filtering</td>
<td>Long reads</td>
<td>Filtlong [145]</td>
</tr>
<tr>
<td>Long-read error correction</td>
<td>Short reads</td>
<td>CoLoRMAP [146], Hercules [147], HG-CoLoR [148], Jabba [149], LoRDEC [150], LoRMA [151], NaS [152], proovread [153]</td>
</tr>
<tr>
<td></td>
<td>Long reads</td>
<td>Canu [114], CONSENT [154], Daccord [155], FLAS [156], HALC [157], MECAT [118], MECAT2 [118], NECAT [120], NextDenovo [121]</td>
</tr>
<tr>
<td>Polishing</td>
<td>Short reads</td>
<td>ntEdit [158], Pilon [159], POLCA [160]</td>
</tr>
<tr>
<td></td>
<td>Short &amp; long reads</td>
<td>Apollo [161], Hapo-G [162], HyPo [163], Racon [164]</td>
</tr>
<tr>
<td></td>
<td>Long reads</td>
<td>Arrow [165], CONSENT [154], Medaka [166], NextPolish [167], Nanopolish [168], Quiver [165]</td>
</tr>
<tr>
<td>Haplotig purging</td>
<td>Long reads</td>
<td>HaploMerger2 [169], purge_dups [170], Purge Haplotigs [171]</td>
</tr>
<tr>
<td>Scaffolding</td>
<td>Mate pairs</td>
<td>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], SSSPACE [185], WiseScaffold [186]</td>
</tr>
<tr>
<td></td>
<td>Long reads</td>
<td>DENTIST [187], gapless [188], LINKS [189], LRScaf [190], npScarf [191], PBJelly [192], RAILS [193], SLR [194], SMIS [195], SMSC [196], SSSPACE-LongRead [197]</td>
</tr>
<tr>
<td>Genetic maps</td>
<td></td>
<td>ALLMAPS [198]</td>
</tr>
<tr>
<td>Optical maps</td>
<td></td>
<td>AGORA [199], BiSCoT [200], OMGS [201], SewingMachine [202], SOMA [203]</td>
</tr>
<tr>
<td>Linked reads</td>
<td></td>
<td>ARBitR [204], Architect [205], ARCS [206], ARKS [207], fragScaFF [208], Scaff10X [209]</td>
</tr>
<tr>
<td>3C/Hi-C</td>
<td></td>
<td>3D-DNA [12], dnaAfri [210], GRAAL [211], HiAssembler [212], instaGRAAL [213], Lachesis [214], pin_hic [215], SALSA [216], SALSA2 [217], scaffhic [218], YahS [219]</td>
</tr>
<tr>
<td>Gap filling</td>
<td>Short reads</td>
<td>GapFiller [220], GAPPadder [221], Sealer [222]</td>
</tr>
<tr>
<td></td>
<td>Long reads</td>
<td>Cobbler [193], DENTIST [187], FGAP [223], gapless [188], GMcloser [224], LR_Gapcloser [225], PBJelly [192], PGcloser [226], TGS-GapClosor [227]</td>
</tr>
</tbody>
</table>

---

**Figure 5** – Assembly pipeline, including the assembly and the pre/post-processing steps.
and have been thoroughly reviewed on multiple datasets [229]. When tested on Caenorhabditis elegans Nanopore reads, the error rate decreased from 28.93% to less than 1% (using Canu [114], CONSENT [154], FLAS [156], Jabba [149], LORMA [151] or MECAT [118]). Assembling corrected reads is expected to yield contigs with higher quality and contiguity. Alternatively, or additionally, the contigs can be polished to reduce errors, using long reads and/or short reads. Polishing can be a more computationally efficient strategy: the reads are mapped solely to the draft assembly, while correction is usually based on an all-versus-all read mapping.

Assemblers are generally tested on model-organism datasets, and are ill-suited for non-model genomes with variable levels of heterozygosity. They often fail to collapse highly divergent haplotypes, causing artefactually duplicated regions that hinder subsequent analyses [230]. Some long-read assemblers, Ra and wtdbg2, have been identified as less prone to retain uncollapsed haplotypes [231]. Contigs can also be post-processed to remove these duplications with dedicated tools such as HaploMerger2 [169], purge_dups [170] and Purge Haplotigs [171]. HaploMerger2 detects uncollapsed haplotypes based on sequence similarities, while purge_dups and Purge Haplotigs also rely on coverage depth.

To improve the contiguity of an assembly, contigs can be grouped, ordered and oriented into scaffolds. These scaffolds may contain gaps, when the sequence that should connect two contigs cannot be retrieved, represented as a sequence of Ns, and these gaps can be reduced post-scaffolding with gap-filling tools. Chromosome-level scaffolds have become a standard in genome assembly publications: unlike fragmented assemblies, they can be used for synteny analysis, finding rearrangements, and to separate chromosomes from different species. Scaffolding tools were already developed for first-generation sequencing reads (e.g. Celera [73], CAP3 [72], GigAssembler [232]). Since then, several sequencing techniques have been used to scaffold assemblies: mate pairs, long reads, genetic maps, optical mapping, linked reads, and proximity ligation [233]. Mate pairs are short reads with a large insert size (more than several kb), and have been widely used in next-generation assemblies. Among the 237 assemblies we surveyed, 78 included a mate-pair scaffolding step (Figure 6). Both genetic maps [234] and optical maps [235] provide information on the linkage and relative position of a set of markers, spread over the genome, thus they can be used to anchor contigs. Genetic maps were used for the genome assemblies of the flatworm Schistosoma mansoni [236], the copepod Tigriopus japonicus [237] and the coral Acropora millepora [26]. Although existing genetic maps provide precious resources, building one is particularly difficult as it requires breeding [234], making it hardly accessible for wild species, and impossible for asexual species. Markers of optical maps are motifs in the sequence that are labeled and detected by a fluorescent signal. Companies such as Bionano or Nabsys propose this service to scaffold assemblies [238], and this method was included in some non-vertebrate genome projects: several nematodes including Onchocerca volvulus [239], Ascaris suum and Parascaris univalens [240], the tapeworms Echinococcus multilocularis [241] and Hymenolepis microstoma [242], and the chiton Acanthopleura granulata [243].

Linked reads and proximity ligation are based on short-read sequencing, preceded by a specific library preparation. For linked reads, also called cloud reads, long fragments of DNA are barcoded and then sequenced. The company 10X Genomics was a leader of this technology, but they chose to discontinue its commercialization in June 2020. New linked-read methods are now emerging such as haplotagging [244] and TELL-seq [245], and the latter protocol is able to handle inputs as low as a few nanograms of DNA. Linked reads have been used to scaffold the genomes of the coral Acropora millepora [26] and the bee Lasiosglossum 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 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 Dovetail 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], Carcinocorpus 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 broughtoni [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 xenoacoelomorph Hofstenia miamia [268], and more. A compelling advantage of Hi-C scaffolding over other scaffolding methods is its ability to discriminate different organisms in a draft assembly: DNA from different organisms belong to distinct nuclei, thus they have no 3D interactions. This feature is especially useful for non-vertebrate animals with symbionts, that can hardly be eliminated from the host prior to sequencing, and are often targets for genome assembly as well.

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

4. Assembly evaluation

A critical step in genome assembly is to estimate the quality of draft assemblies, and choose the best one for subsequent analysis. The first metric to assess is the assembly size and its adequacy with an estimated genome size. The size can be estimated experimentally with flow cytometry or Feulgen densitometry [272], but these methods require a reference species for which the genome size is already well known, exposing them to errors induced by the reference genome size. Reference-free genome size estimation tools are typically k-mer based approaches and use high-accuracy reads (e.g. Illumina, PacBio HiFi). These tools, such as BBtools [273], GenomeScope [274] and KAT [275], build a k-mer spectrum representing the number of k-mers with a certain frequency of occurrence. 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 occurrence close to zero, corresponding to erroneous k-mers. Another recent tool called MGSE [276] estimates genome size based on reads mapping to a highly contiguous assembly of the same genome; this method can be used as a post-hoc analysis.

N50 is a popular metric that reflects the contiguity of an assembly: it is defined as the length of the largest contig (or scaffold) for which 50% of the assembly size is contained in contigs (or scaffolds) of equal or greater length. Some tools provide in addition the N75, N90, N99, computed in a similar fashion. The NG50 is a variant of N50 that refers to an estimated genome size instead of the assembly size. The target assembly can further be mapped against a reference assembly to detect misassemblies and break them: the N50 and NG50 of the resulting fragments are called NA50 and NGA50. All these metrics can be computed using QUAST [277]. For genome assemblies of non-model non-vertebrate animals, reference assemblies are seldom available, or they have a poor quality or contiguity that the new assembly aspires to improve. Therefore we will focus on reference-free evaluation methods. Table 3 and Figure 8 present an example of assembly evaluation for the recently published snail Achatina fulica [278] and the coral Xenia sp. [257].

Another feature to optimize is the completeness of the genome, usually based on orthologs or k-mers. BUSCO [38–40] searches for orthologs in a user-provided lineage; the current Metazoa lineage (designated as Metazoa odb10) contains 954 features. Assemblies are evaluated based on the proportion of orthologs to these 954 genes that can be retrieved into them; yet, some features are systematically missing in some genomes as they are absent from these species. More specific lineages are available for arthropods, insects, nematodes, vertebrates, mammals, as many assemblies are available for these groups, but other metazoan phyla suffer from their lack of resources. Consequently, BUSCO is most powerful when comparing several draft assemblies for one genome. BUSCO scores provide information on complete single-copy and duplicated
Overview of scaffolding

contigs

→

long-range information

Ns

scaffolds

Short reads/mate pairs

pair of short reads with a long insert

Long reads

long read overlapping several contigs

Optical/genetic maps

map = ordered set of markers

marker found in a contig

Linked reads

set of barcoded reads amplified from a long fragment

3C/Hi-C

pair of short reads indicating a 3D contact

features, and the latter can be used to detect improperly duplicated regions in a haploid assembly. However, BUSCO scores are limited to genomic regions and cannot report for non-coding ones.

$k$-mer completeness scores do not present such limitations: KAT assesses the completeness of a whole assembly based on its representation of $k$-mers from a high-accuracy sequencing dataset. The $k$-mer spectrum should display one or several peaks depending on the ploidy of the genome: one peak for a haploid genome; two peaks for a diploid genome, the first depicting heterozygous $k$-mers, and the second for homozygous $k$-mers. Depending on the ploidy of the genome, every $k$-mer should be represented in the assembly as many times as they actually are in the genome.
Table 3 – Assembly evaluation of *Achatina fulica* and *Xenia* sp..

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

Figure 8 – Assembly evaluation of *Achatina fulica* and *Xenia* sp.. Left: KAT comparison of the k-mers in the Illumina datasets v. the assembly. Right: Hi-C contact maps, with a binning of 300 for *Achatina fulica*, 30 for *Xenia* sp..
Both *Achatina fulica* and *Xenia* sp. have high BUSCO scores (against the lineage Metazoa odb10), yet slightly below 90%, and they have few duplicated BUSCO features. The k-mer spectrum of *Achatina fulica* only shows one peak around 70X (Figure 8, top left). These k-mers are expected to be represented exactly once, which is the case for the majority; there are almost no k-mers that appear twice in the assembly (in purple), but there is a noteworthy amount of missing k-mers (in black). For *Xenia* sp., the k-mer spectrum has two peaks with a k-mer multiplicity around 35X and 70X (Figure 8, bottom left). The first peak, representing heterozygous k-mers, shows that a portion is represented once in the assembly, while the rest is missing, as expected in a collapsed assembly. The second peak, for homozygous k-mers has a majority of k-mers represented once, and some k-mers either absent or duplicated. These assemblies seem overall properly collapsed and complete.

KAD, for k-mer abundance difference [279], proposes an alternative k-mer-based evaluation. This tool does not compute an overall completeness score, but instead classifies k-mers based on their abundance in the assembly and the sequencing dataset: good k-mers, erroneous k-mers (absent from the dataset), overrepresented k-mers (duplications), and underrepresented k-mers (collapsed repetitions).

Assemblies need to be screened for contaminants, to tell apart the sequences coming from the target and from other species. Contaminants may originate from the environment, the symbiont, or be artificially introduced by the sequencing process. Blobtools [280] and BlobToolKit [281] aim to identify them with GC content, coverage depth and taxonomy assignment using the NCBI TaxID. Discriminating bacteria in metazoan assemblies is usually straightforward based on their distinct GC percentage. The task is more challenging when the target metazoan genome is mixed with other eukaryotes or even metazoans, especially when these species are absent from databases. Chromosome-level assemblies reduce the risk of contamination, as downstream analyses can be run exclusively on sequences that were anchored to the main scaffolds. In addition, with Hi-C data, sequences from different species can be separated based on their absence of trans interactions. Contamination can lead to false conclusions: for instance, a study on a highly fragmented genome assembly (N50 = 16 kb) of the tardigrade *Hypsibius dujardini* assumed that about 17% of its genome derived from horizontal gene transfers [282], when these sequences were in fact contaminants [283].

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

5. Phasing assemblies

As collapsing multiploid genomes can be difficult for highly divergent regions and frequently causes breaks in the assembly, an intuitive solution would be to phase genomes to retrieve all haplotypes. Phased assemblies represent a whole different challenge as they necessitate to correctly associate alleles, i.e. different versions of a heterozygous region [284]. A first approach, called trio-binning, is to assemble one individual using sequencing data from the individual itself and its parents [285]; yet this method is only adapted when the parents can be identified, and is inapplicable on asexual species. Some tools are able to reconstruct haplotypes from collapsed assemblies using long reads, namely HapCUT2 [286] and WhatsHap [287]. Ideally, genomes should be uncollapsed, as can be done with Bwise [288] and Platanus-Allee [289] using short reads, FALCON-Unzip [115] using PacBio CLR or HiFi. FALCON-Unzip uses the output from the
FALCON assembler, that includes both a haploid assembly and alternative haplotigs for heterozygous regions, to associate haplotypes based on long reads. Phased assemblies of low-accuracy long reads are limited, as small heterozygous regions were confused with errors; this led to haplotypes being erroneously collapsed.

HiFi reads have made a disruption in the fields of genomics: they are especially well-suited for phased assemblies, using hifiasm [129] for instance, thanks to their length and low error rate, and they have already been used to produce phased assemblies of a human [290] and the potato Solanum tuberosum [291]. Nevertheless, sequencing HiFi reads can remain inaccessible for non-model organisms as pure DNA is necessary.

Many organisms have already been assembled using low-accuracy long reads and high-accuracy short reads, thus an alternative is to correct long reads with short reads using a tool that conserves haplotypes such as Ratatosk [292]. Phased long-read assemblies can be further polished with adequate programs (e.g. Hapo-G [162]). As Hi-C has demonstrated its efficiency to scaffold haploid assemblies, the principles were further exploited in ALLHiC [293], GraphUnzip [294] and FALCON-Phase [295] to phase assemblies while increasing their contiguity: as alleles from one haplotype belong to one chromosome, these alleles have higher Hi-C interaction frequencies together than with alleles from alternative haplotypes.

Phasing-specific evaluation methods are still scarce, and publications of phased assemblies rely on various datasets to prove their correctness (e.g. parental assemblies [290]). Merqury [296] proposes a \(k\)-mer-based approach, inspired by KAT, and computes plots and scores to assess phasing completeness and find haplotype switches. However, similarly to trio-binning, it requires parental data.

### 6. Recommendations

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

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

When high-molecular-weight DNA cannot be extracted, short reads are the adequate option. The resulting assemblies are more fragmented, yet some short-read assemblers are able to produce good drafts, such as Platanus. These assemblies may have large missing repetitions, thus they are not ideal for analysis of repetitive content and they should be thoroughly assessed in terms of assembly size and completeness.
Hi-C scaffolding has emerged as the most robust method to obtain chromosome-level scaffolds with no contamination. It is applicable as long as fresh or flash-frozen tissue is available for crosslinking, and with a minimum input of a half to one million cells. When these requirements are not fulfilled, linked reads can be used as an alternative (as TELL-seq can use a low input of DNA), or in addition to further reduce assembly errors.

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

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

7. Building robust animal genomic databases

We surveyed genome assembly papers from diverse metazoan phyla. Figures 1, 4 and 6 only retained assemblies that were available on GenBank, as we used assembly sizes, contig N50s and scaffold N50s from this source. We also limited these assemblies to those published after the year 2007, as we found that assemblies were seldom available on GenBank before that, and up to the year 2020. Some genomes were not deposited, and were instead available on a personal/lab/university page. This impedes meta-analyses and we are unable to accurately estimate the number of published non-vertebrate animal genome assemblies. The datasets used for the genome assemblies also suffer from this issue, as they are not necessarily publicly available. Efforts are being made to make genome assemblies and datasets findable, accessible, interoperable and reusable (FAIR) [298]. Assembly pipelines are becoming more reproducible thanks to several initiatives using workflow managers, such as the Vertebrate Genome Project assembly pipeline in Galaxy [299].

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

Supplementary information

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

This project was funded by the Horizon 2020 research and innovation program of the European Union under the Marie Skłodowska-Curie grant agreement No 764840 (ITN IGNITE, www.itn-ignite.eu) and a complementary fellowship from the David and Alice Van Buuren fund and the Jaumotte-Demoulin foundation.

Acknowledgements

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

Conflict of interest disclosure

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

References


