Mini Review

Long walk to genomics: History and current approaches to genome sequencing and assembly

Alice Maria Giani, Guido Roberto Gallo, Luca Gianfranceschi, Giulio Formenti

Department of Surgery, Weill Cornell Medical College, New York, NY, USA
Department of Biosciences, University of Milan, Milan, Italy
The Rockefeller University, New York, NY, USA

Article info

Article history:
Received 29 August 2019
Received in revised form 3 November 2019
Accepted 6 November 2019
Available online 17 November 2019

Keywords:
Genome assembly
Sequencing
Reference
Bioinformatics
Next-generation
Third-generation

Abstract

Genomes represent the starting point of genetic studies. Since the discovery of DNA structure, scientists have devoted great efforts to determine their sequence in an exact way. In this review we provide a comprehensive historical background of the improvements in DNA sequencing technologies that have accompanied the major milestones in genome sequencing and assembly, ranging from early sequencing methods to Next-Generation Sequencing platforms. We then focus on the advantages and challenges of the current technologies and approaches, collectively known as Third Generation Sequencing. As these technical advancements have been accompanied by progress in analytical methods, we also review the bioinformatic tools currently employed in de novo genome assembly, as well as some applications of Third Generation Sequencing technologies and high-quality reference genomes.

1. Introduction

A genome is the complete genetic information of an organism or a cell. Single or double stranded nucleic acids store this information in a linear or in a circular sequence. To precisely determine this sequence, progressively more efficient technologies characterized by increased accuracy, throughput and sequencing speed have been developed. Nonetheless, sequencers can generate sequences, known as reads, comprised only in defined ranges of lengths, usually far shorter than the size of the genomes investigated. The complete genome sequence has to be deduced from the overlaps of these shorter fragments, a process defined as de novo genome assembly. Historically, mostly due to time and cost constraints, only an individual per species was addressed, and its sequence generally represents the ‘reference’ genome for the species. These reference genomes can guide resequencing efforts in the same species, acting as a template for read mapping. They can be annotated to understand gene function or used to design gene manipulation experiments. Sequences from different species can be aligned and compared to study molecular evolution. Due to the impact of reference genomes in all these downstream applications, it is paramount that their sequence is as much complete and error-free as possible. In the last 50 years of the XX century, available sequencing technologies allowed to focus mostly on relatively small genomes. Since the new millennium, novel platforms, known as Next Generation Sequencing (NGS), have been developed to address larger genomes, in a process called Whole Genome Sequencing (WGS). In the two decades following the advent of WGS, NGS has become increasingly more efficient and affordable. In parallel, new sequencing technologies have emerged that promise to revolutionize the field and generate genomes of higher quality.

https://doi.org/10.1016/j.csbj.2019.11.002

© 2019 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Abbreviations: BAC, Bacterial Artificial Chromosome; bp, base pair; ddNTP, 2,3-dideoxynucleoside triphosphate; dNTPs, deoxynucleoside triphosphates; HapMap, haplotype map; HGP, Human Genome Project; HMW, high molecular weight; NGS, Next Generation Sequencing; OLC, Overlap-Layout-Consensus; QV, Quality Value (QV); SBS, Sequencing by Synthesis; SMRT, Single Molecule Real-Time; SNPs, Single Nucleotide Polymorphisms; SRA, Short Read Archive; SV, Structural Variant; TGS, Third Generation Sequencing; WGS, Whole Genome Sequencing; ZMW, Zero-Mode Waveguide.

Corresponding author.
E-mail addresses: amg4001@med.cornell.edu (A.M. Giani), guido.gallo@guest.unimi.it (G.R. Gallo), luca.gianfranceschi@unimi.it (L. Gianfranceschi), gformenti@rockefeller.edu (G. Formenti).

First author.
1 ORCID: 0000-0001-8805-7826.
2 ORCID: 0000-0001-5980-0326.
3 ORCID: 0000-0002-7644-5968.
4 ORCID: 0000-0002-7554-5991.

https://doi.org/10.1016/j.csbj.2019.11.002

2001-0370© 2019 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
2. Historical background

2.1. Sequencing nucleic acids in the XX century

In 1953 Watson and Crick published their seminal paper unraveling the double helix structure of DNA [1]. In their work they noted “so far as is known the sequence of bases along the chain is irregular” and “the sequence of bases on a single chain does not appear restricted in any way”, two features that entail a role of DNA in the storage of genetic information, and highlight the importance of determining the exact sequence of bases along the chain. Year 1953 also marked the first sequencing of a biological molecule. Thanks to a refined partition chromatography method, Sanger was able to sequence the two chains of insulin protein [2,3]. In his approach, proteins were randomly fragmented, the fragments were individually read, and sequences from each fragment were overlapped to yield a complete, consensus sequence. Proteins were sequenced before nucleic acids, but many principles remained the same, paving the way to modern DNA sequencing. When in 1965 it came to nucleic acids, first it was the turn of the 76 bases long alanine tRNA from Saccharomyces cerevisiae, as methods to cleave RNA fragments were available since the 1940s [4]. As for the insulin, RNA was first fragmented with RNase A and RNase T1, the pieces were separated by chromatography, and the degradation products were used to deduce its sequence. At the end of the 1960s, RNA sequencing was still ahead of DNA sequencing. In 1968, the 12 bases long complementary extremities of phage λ-cos-site were the first DNA molecules sequenced [5]. The same year, a team including Sanger determined the 120 base pair (bp) long 5s rRNA using 32P-labeled RNA and a paper fractionation-based approach [6]. In 1972, Fiers sequenced the first gene, the 510 bp of the coat protein gene from the RNA virus phage MS2 [7]. In 1973, Gilbert and Maxam reported the 24 bases of Escherichia coli lactose-repressor binding site by copying its DNA into RNA [8]. In 1976, phage MS2 was the first organism to have its 3,569 bp RNA genome completely sequenced [9] (Fig. 1). In the mid 1970s, DNA sequencing was about to flourish. In 1975, Sanger determined the 5368 bp genome of phage ϕX174, the first DNA genome to be sequenced [12] (Fig. 1). The same year, he developed a new method that could decode fragments of approximately four hundred bases in a day [13]. The ‘Sanger’ method relies on four separate polymerization reactions performed using tritium-radiolabeled primers, where each reaction is supplied with small amounts of one chain-terminating 2,3-dideoxynucleoside triphosphate (ddNTP) to produce fragments of different lengths [13]. When the DNA polymerase incorporates a ddNTP at the 3'-end of the growing DNA strand, it lacks a 3'-hydroxyl group and chain elongation is terminated [14]. Similarly to the ‘plus and minus’ approach, the sequence is deduced by comparing the size of the fragments. The same year, Maxam and Gilbert proposed an alternative, purely chemical, approach [15]. While extremely popular for years, with the improvement of the chain-termination method, this approach fell out of favor due to its technical complexity, use of hazardous chemicals, and difficulties to be scaled up.

2.2. Shotgun sequencing

To hasten the sequencing process, in 1979 Staden proposed the idea of ‘shotgun sequencing’, where bacterial vectors are used to clone random fragments of a long DNA molecule. Fragments are then sequenced in parallel and reads are assembled using their overlaps [16], allowing to sequence larger genomes in shorter times. In 1981, Messing developed the first shotgun sequencing protocol using the single-stranded M13 phage vector [17].
in 2007 [31] and Watson in 2008 [32]. In 2011, six-year-old individuals to have their genome fully re-sequenced were Venter. Ome resequencing started to become more affordable. The first projects as the International HapMap Project [30]. Human genotyping became popular, with millions of Single Nucleotide Polymorphisms (SNPs) from hundreds of individuals identified through RNA sequencing (RNAseq) [27], chromatin immunoprecipitation followed by sequencing (ChIP–seq) [28], and genome-wide epigenetic landscape determination [29]. Human genotyping became popular, with millions of Single Nucleotide Polymorphisms (SNPs) from hundreds of individuals identified by projects as the International HapMap Project [30]. Human genome resequencing started to become more affordable. The first individuals to have their genome fully re-sequenced were Venter in 2007 [31] and Watson in 2008 [32]. In 2011, six-year-old Nicholas Volker was reported as the first patient saved by DNA sequencing, as his one in 1 billion genetic mutation of XIAP gene was succeeded (Fig. 1), as testified by the increasing availability of data in dedicated repositories [20]. GenBank, the US National Institute of Health (NIH) sequence database, was founded in 1982 with about half a million bases, and by the end of the decade it included over 40 million bases [20]. This growth rate has never ceased, with almost 10-fold increases every 5 years. “Sequences, Sequences, and Sequences” was the title of a Sanger’s article in 1988, suggestive of the general excitement [21]. In the 1980s and 1990s, technological breakthroughs combined with the adoption of industrial processes increased throughput and reduced sequencing errors (Box 1.1). These advancements were further accelerated by the Human Genome Project (HGP), which aimed to produce genetic maps, physical maps, and finally the complete sequence of human chromosomes (Box 2). Moreover, the competition with Celera, a private company that also decided to assemble the human genome, stimulated both initiatives to obtain results rapidly [22,23] (Box 2). With the HGP, companies realized that sequencing could be a profitable business and the market competition gave birth to a plethora of sequencing technologies, collectively referred to as Next-Generation Sequencing (NGS) (Box 1.2).

2.3. DNA sequencing in the third millennium

From a technical standpoint, the new generation of sequencing approaches resulted from a combination of advancements in microfabrication, high-resolution imaging and computational power [24] (Box 1.2). All NGS approaches rely on a ‘library’ preparation using native or amplified DNA. In a classical protocol, after DNA fragmentation and fragment size selection, adapters are ligated to the ends of each fragment. This is generally followed by a step of DNA amplification. The resulting library is loaded on a flow cell and sequenced in Massive Parallel Sequencing reactions [25]. Each reaction usually involves the stepwise, polymerase-mediated incorporation of fluorescently labeled deoxynucleotides in an elongating DNA chain immobilized on a surface (Box 1.2). Results from the blossom of high-throughput sequencing technologies have been innumerable. The power of the new methods was proved by the cost-effective and rapid resequencing of many milestone genomes. During this period of excitement, genome drafts for most eukaryotic model species were produced (Fig. 1). NGS also allowed extremely novel applications, including whole-exome sequencing [26], high-throughput RNA sequencing (RNAseq) [27], chromatin immunoprecipitation followed by sequencing (ChIP–seq) [28], and genome-wide epigenetic landscape determination [29]. Human genotyping became popular, with millions of Single Nucleotide Polymorphisms (SNPs) from hundreds of individuals identified by projects as the International HapMap Project [30]. Human genome resequencing started to become more affordable. The first individuals to have their genome fully re-sequenced were Venter in 2007 [31] and Watson in 2008 [32]. In 2011, six-year-old Nicholas Volker was reported as the first patient saved by DNA sequencing, as his one in 1 billion genetic mutation of XIAP gene resulted to be treatable with cord transplant [33]. Efforts in human genome sequencing were resumed by the 1000 Genomes Project (2008–15) [34], and more large-scale human genome

Box 1. The industry of DNA sequencing

Box 1.1. Automation of Sanger Sequencing

In the early 1980s, DNA sequencing was performed with the original Sanger method, involving four sequencing reactions in four separate tubes using radiolabeled primers. In 1985, Smith and Hood synthesized fluorescent DNA primers [117] and used them for the partial automation of Sanger sequencing [118]. In this method, known as dye terminator sequencing, four dyes distinguished by their fluorescent emission spectra are covalently attached to the oligonucleotide primers and employed in distinct dyeoxy extension reactions. The products are then run together on a polyacrylamide gel and Fluorescent Energy Resonance Transfer is used to read the sequence. In 1987, Prober described dye terminator sequencing, a novel approach based on four chain-terminating dyeoxy nucleotides, each carrying a succinyl fluorescein dye with a different emission spectra [119]. The distinct fluorescence signals allowed reactions to be performed in a single tube. In 1986, Applied Biosystems, a company founded in 1981 focused mostly on protein sequencing, switched to nucleic acids and released the first commercially available four-color fluorescence automated DNA sequencer (ABI 370A) based on these advancements. A scanning system allowed multiple samples to be run simultaneously, with automatic computer-based base calling. The machine was able to handle 32 samples per run, quadrupling the throughput. The next year this approach was refined by introducing an optimized T7 DNA polymerase [120]. This enzyme was highly processive, lacked proofreading 3’ to 5’ exonuclease activity and efficiently used nucleotide analogs, allowing to generate around 1000 bases per day [121]. Two years later, the introduction of the thermostable Taq DNA polymerase and of cycle sequencing greatly reduced template requirements and facilitated miniaturization. Now the sequencing reaction could be performed at 72 °C and repeated multiple times in the same tube, achieving linear amplification of the template [122]. The technological advancements of the 1980s were further reinforced and extended in the 1990s. Major milestones included the introduction of fluorescent boron-dipyrromethene dyes (bodipy) more effective than conventional dyes [123], magnetic bead-based DNA purification methods to simplify the automation of pre-sequencing steps [124], and capillary electrophoresis, which eliminated the pouring and loading of gels, while also simplifying the identification and interpretation of the fluorescent signal [125]. These rapid advancements were further prompted by the HGP race (Box 2). In 1995, Applied Biosystems replaced ABI 370A with another slab-gel electrophoresis system, ABI Prism 377, which could deal with up to 96 lanes at a time. One year later was released ABI Prism 310, the first capillary DNA sequencer, and in 1998 the high-throughput ABI Prism 3700. In 2001, it was the turn of the 16-capillary ABI Prism 3100, and in 2002 that of ABI 3730xl provided with 48 to 96 capillaries, where sequences included QVs. By 1998, routine DNA sequencing of 1000 bases was achieved in less than one hour [126]. During the great period of excitement of the 80s and 90s, even a completely new method for DNA sequencing by stepwise dNTP incorporation was developed [127,128]. Known as “pyrosequencing”, this approach relies on measuring the enzymatic luminometric signal generated by pyrophosphate release during DNA polymerization [129]. This strategy has several advantages over Sanger’s approach, including the use of natural nucleotides and the possibility of observing nucleotide synthesis in real time [130].
Box 1.2. A $1000 genome

In the 1990s, Applied Biosystems was the undisputed market leader. In contrast with this earlier monopoly, in the 2000s many companies started to offer different technologies with progressively higher sequencing throughput at lower costs. These included 454, Solexa, Illumina, Agencourt, Complete Genomics, Applied Biosystems and Ion Torrent. Following the enthusiasm generated by the HGP (Box 2), the ultimate goal became making the cost of sequencing a whole human genome more affordable, achieving a price below $1000 per genome. The first generation of NGS sequencers was based on pyrosequencing and initially produced by Pyrosequencing AB. Through a series of acquisition, this technology was later commercialized by 454 Life Sciences, which in 2003 introduced on the market the first NGS DNA sequencer, the GS20. The GS20 used single-molecule template synthesis of small, bead-bound DNA fragments in a water-in-oil emulsion clonal PCR (emPCR) [131], and dNTP incorporation detection by CCD sensors beneath the surface of about one million microwells [132]. The system produced approximately 400–500 bp-long reads, had 99% accuracy and could sequence up to 25 million bp in a single 4-hour run at less than one-sixth the cost of conventional methods. In 2003, Solexa started to develop a new sequencing approach, later known as Sequencing by Synthesis (SBS) [133]. Next year Solexa acquired from Manteia the colony sequencing technology, also known as bridge amplification [134]. In bridge amplification, tightly clustered copies of individual molecules, or “polonies” [135], are produced on a surface from an immobilized template library [136]. The resulting stronger signal enhanced the accuracy of base calling, while reducing the cost of the system optics. In 2005, Solexa added to its method an engineered DNA polymerase [137] and the recently developed reversible terminators that can be washed away, allowing for continuous extension steps [138,139]. The resulting platform was able to image and determine each single nucleotide added to all the DNA fragments placed on the surface of a flow cell. Genome Analyzer, the first Solexa commercial sequencer, was launched in 2006. In contrast to GS20, Solexa sequencer had a higher throughput of 1 Gbp in a single run, but sequencing reads were shorter, only 35 bp long [140]. However, a great advantage was that these represented paired end reads (i.e. both DNA strands were simultaneously sequenced), allowing to size the gap between relatively distant sequences in a DNA fragment. In 2007, Solexa was acquired by Illumina, a company founded in 1998 that by 2005 had started over 2000 NGS gene expression services. The 2007 was also the year of SOLiD from Applied Biosystems. This technology was based on a ligation strategy, relying on the specificity of DNA ligases to ligate fluorescent oligonucleotides to templates in a sequence-dependent manner [141]. This method was later reported to have some issues in sequencing palindromic sequences [142] and was consequently abandoned. A new approach based on proton detection in semiconductors was released in 2011 by Ion Torrent [143]. This approach generated 100 bp long reads by measuring the pH variations induced by hydrogen ions released during nucleotide addition in DNA synthesis, using a semiconductor sensor. Parallel measurements of multiple templates were carried out in microwell plates where each of the four nucleotides was added in succession. The process of reading each nucleotide could occur in a few seconds, without imaging and therefore at a considerably lower cost. Based on an advancement of this technology, in 2012 Ion Torrent released a more powerful machine called Ion Proton, which the company claimed could have allowed to sequence a human genome in a single day for the price of $1000. However, similarly to SOLiD, Ion Torrent platforms often mismeasured the length of homopolymers [144] and were therefore abandoned. Due to the limitations of the other sequencing platforms, the SBS approach commercialized by Illumina ultimately prevailed. By 2014, the company appeared to have reached a position of near monopoly [145], holding 70% of the market for DNA sequencers and accounting for more than 90% of all DNA data produced [146]. The same year, the company announced the HiSeq X Ten, claiming that forty of these machines would have been able to sequence more genomes in one year than had been produced by all other sequencers to date, allowing large-scale WGS for $1000 per genome [147]. This goal has been exceeded in 2017, when Illumina released NovaSeq, a sequencer that can output over 3 Gbp in a single run. However, these short read-based genomes are partially incomplete and represent only draft assemblies, that can be improved through complementary sequencing and scaffolding approaches (Section 3). Consequently, the new frontier is becoming to make hybrid genome assembly also more affordable.

Box 2. The reference human genome

In 1988, the US Department of Energy (DOE) and NIH received funding and formalized an agreement to “coordinate research and technical activities related to the human genome” [148]. Watson was appointed to lead the NIH component and the official kick-off was in 1990 with the publication of a joint research plan for the next five years. The deadline for project completion was set to 2005, however Watson resigned in 1992 due to the persistent confrontations with the NIH Director [149]. The following year Collins was appointed in his place and traced a new five-year plan for the HGP [150]. By 1994 the HGP team published the first linkage map of the human genome [151]. Despite the successes with smaller genomes, in 1998, formally halfway through the schedule for completing the HGP, only 50 Mb of human sequence had been determined [152]. The HGP team had developed an accurate but laborious strategy named “hierarchical shotgun” sequencing. The pipeline involved cloning large fragments of human genome into Bacterial Artificial Chromosomes (BACs). BACs were then fragmented, size-selected and subcloned. DNA from individual clones was used as a template for automated Sanger sequencing [22]. Nonetheless, HGP leaders announced the intent to complete the project in 2003, two years ahead of previous projections [153]. However, a private effort led by Venter’s company Celera decided to rival the public HGP, accelerating the events. Counter to HGP leaders, the team at Celera believed direct shotgun sequencing without BAC intermediates to be the fastest and most effective approach also to sequence larger genomes [154]. Building upon the experience gained at The Institute for Genomic Research (TIGR) with bacteria, Celera team rapidly sequenced the 175 Mbp of Drosophila melanogaster genome using an internally-developed WGS strategy and a new genome assembler based on a more effective OLC approach [155] (Box3). The success with D. melanogaster prompted Venter to turn to the human genome. In March 1999, HGP leaders announced the successful completion of the pilot phase and at the beginning of 2000 the HGP team deposited two of the three Gbp of the human genome [156]. Progress was fast also at Celera, and in less than two years they generated their draft of the human genome. The race culminated in June 2000 with a joint announcement. The HGP officially released the sequence in February 2001 [22] and Celera published its genome assembly one day later [23]. At that time, both the HGP and Celera teams had published only working drafts of the human genome. The most complete assembly, from the HGP, represented about 90% of the human genome, with only 25% in curated form [22].
3. Genome sequencing and assembly in the Third Generation Sequencing era

Starting from the 2010s, radically new technologies opened the era of Third-Generation Sequencing (TGS). TGS definition may vary, but it is generally given to technologies capable of sequencing single DNA molecules without amplification. Nowadays, these technologies allow to produce reads far longer than NGS, each spanning several to hundreds kbps. In the context of genome assembly, the availability of long reads constitutes a great advantage, as the difficulty of detecting overlaps between NGS short reads reduces the ability of generating long continuous consensus sequences [36], impacting the overall quality of the assemblies [37–39]. This issue is further exacerbated by the presence of repeated and low complexity DNA regions [40]. As more complex genomes are addressed, the impact of repeat elements increases [41]. In these cases, the final assembly continuity may be reduced even adopting refined NGS-based approaches, such as 10X Genomics linked reads, where barcodes are associated with individual high molecular weight (HMW) DNA molecules prior to NGS library preparation and later used to group short reads sharing the same barcode as originating from the same HMW DNA fragment [42,43]. Extra-large genome sizes such as those of amphibians [44], where low-complexity regions account for most of the ‘extra’ DNA [45], constitute a practically insurmountable barrier for NGS-based WGS (Fig. 1). TGS enables to address these shortcomings to generate genome assemblies of unprecedented quality [46–48] (Fig. 1).

3.1. Single-molecule sequencing

The first single-molecule sequencing technology, based on fluorescence detection and Sequencing by Synthesis (SBS), was developed by Quake and commercialized in 2009 by Helicos BioSciences [49,50]. It worked similar to Illumina sequencers, but without any bridge amplification, thereby avoiding DNA amplification-associated biases. However, it was slow, expensive and produced relatively short reads, around 35 bp long. Two single-molecule approaches were developed in the last decade that overcame these disadvantages. The first approach, Single Molecule Real-Time (SMRT) sequencing was developed by Craighead, Korlach, Turner and Webb and was further refined and commercialized by Pacific Biosciences (PacBio) since 2011. The second approach, Nanopore sequencing, was first hypothesized in the 1990s, but developed into a technology only decades later [51] and commercialized by Oxford Nanopore Technologies (ONT) since 2005. SMRT sequencing normally generates reads with average median length of approximately 10–20 kbp, but also several reads longer than 50 kbp. During library preparation, specific hairpin adapters are ligated to both ends of double-stranded DNA molecules to circularize them, forming a structure known as SMRTbell [52]. SMRTbell libraries are then loaded onto a SMRT cell containing hundreds of thousands Zero-Mode Waveguide (ZMW) nanowell arrays [53]. Each ZMW contains a polymerase immobilized at the bottom that can bind a primer complementary to the hairpin adapters of the SMRTbell and start the replication process [54]. Individual sequencing reactions take place within each ZMW, corresponding to the smallest available volume for light detection (20 zeptoliters each), and consist in the incorporation of the four different fluorescent-labeled nucleotides in succession using the SMRTbell DNA as complementary template. Each incorporation event produces a specific emission signal that is recorded for real-time base-identification by an ultrahigh resolution camera coupled with a computer [54]. A ‘movie’ of light pulses, corresponding to a sequence of bases, is named Continuous Long Read (CLR). The individual base-calling accuracy of these reads is far less accurate than Illumina short reads, with ~1 error every 10 nucleotides [55] compared to the reported ~1 error every 1000 nucleotides using NGS [56]. However, in contrast to NGS, there appears to be little or no sequencing bias [57] and error profiles seem to be purely random, allowing high-quality in the final consensus sequence [58]. Interestingly, the circular nature of DNA molecules in SMRTbells allows to read the same sequence multiple times. Over the years, PacBio has further developed this concept to increase base-calling accuracy, culminating in 2019 with the release of a new method named HiFi (‘High Fidelity’) long-reads, which can generate Circular Consensus Sequences (CCSs) approximately 10–20 kbp-long, potentially as accurate as Illumina short reads [59]. The advantages of SMRT sequencing over NGS have come at the price of higher per base sequencing costs. In 2019 PacBio has released Sequel II, a revised platform based on new chemicals and equipped with a novel 8 million ZMW nanowells SMRT cell, theoretically able to generate a throughput of 160 Gb per SMRT Cell, with a concomitant drop of up to 8-fold in sequencing costs.

Nowadays, the direct competitor of SMRT sequencing for both quality and price is Nanopore sequencing. Nanopore sequencing uses genetically modified bacterial nanopores inserted into an artificial lipid bilayer, placed in individual microwells tens of micrometers wide and arrayed on a sensor chip [60–62]. As each single DNA strand travels through a channel, it disrupts the current running through the pore, and the change is measured by a semiconductor sensor. Since each base disrupts the electric field in a slightly different way, the recorded current changes can be translated into a DNA sequence. The final result is similar to that of SMRT sequencing, with long reads of low base-calling quality. Read length is in the order of 10–20 kbp, and the longest molecule sequenced to date was 2.3 Mbp long [63]. This places ONT reads as the longest DNA reads generated so far, allowing to sequence through repeats where even PacBio reads may fail. Furthermore, since ONT technology relies on the detection of an electrical rather than an optical signal, Nanopore devices can be as small as a USB stick. In 2016, such portability allowed to sequence Ebola virus at field sites in West Africa in less than 60 min [64]. However, ONT reads appeared to have sequencing biases difficult to be corrected [65]. In 2019, Nanopore has released into early access R10, a new nanopore with a double sensor for improved base calling [66].
3.2. Advantages of Third Generation Sequencing over Next Generation Sequencing

TGS has been shown to considerably improve the quality of genome assemblies [58,67]. Ultralong read lengths and low sequencing-context bias (GC-content or low-complexity) enable to assess many regions missing from NGS-based assemblies [39,68], providing a more uniform coverage along the genome and increased structural accuracy of the resulting assemblies. Moreover, in diploid (or polyploid) genomes TGS allows the generation of long phased blocks of haplotypes (haplotigs), where the paternal and maternal contributions to a homologous region of the chromosome are reported separately [69]. Avoiding chimeric genomes, where paternal and maternal haplotypes are mixed or even collapsed, facilitates accurate mapping of sequencing reads for structural variants (SV) detection, including length variations in highly repeated motifs, other large insertion and deletion events (indels), duplications, inversions and translocations [70,71] (Fig. 2). Indeed, while SNPs were long regarded as the most relevant type of genetic variation, it is now clear that SVs (>50 bp) also play a key biological role [72,73]. SVs comprise a large portion of bp accounting for interspecies genotypic differences [74], implying that an important fraction of genomic variation has so far escaped detection by NGS. This could explain much of the “missing heritability” problem [75]. Accordingly, evidence for the importance of SVs in determining both simple and complex phenotypic traits is constantly growing [76,77]. Finally, another advantage of TGS over NGS is that it provides the simultaneous capability of characterizing a variety of epigenetic marks along with DNA sequencing [78,79]. This is achieved by measuring polymerase kinetics in Pacbio [80] and ionic current signals in Nanopore [81].

3.3. Genome assembly using Third Generation Sequencing

The ultralong, noisy reads produced by TGS required the adaptation of assembly methods. At present, the two most popular softwares for de novo genome assembly of TGS data are FALCON and Canu. FALCON was developed directly by PacBio and published in 2013 [82]. It relies on a Hierarchical Genome Assembly Process (HGAP) where the longest reads are initially selected and act as ‘seed’ reads for the alignment of shorter reads. Reads are then ‘corrected’ to account for the high base calling error rate of TGS using the consensus sequence derived from the alignment. These ‘pre-assembled’ reads are then overlapped to each other using de Bruijn graphs to generate the final contigs (Fig. 3 and Box 3). FALCON is diploid-aware and can generate alternative contigs representing the second haplotype out of the assembly graph. Three years later PacBio released FALCON-Unzip, an improved diploid-aware assembler that can take advantage of the heterozygous SNPs identified in the initial assembly to provide highly phased divergent haplotypes [83]. Today Canu is a very popular alternative. It was derived from the Celera Assembler in 2015 [84]. The latter was initially adapted to PacBio sequencing [85], demonstrating finished bacterial genomes [46] before being dismissed. Similarly to the Celera Assembler, Canu is an Overlap-Layout-Consensus (OLC) based assembler [86], specialized in the assembly of TGS long reads.

**Fig. 2.** Structural variations. Schematic representation of the five major types of SVs.

**Fig. 3.** Chromosome-level scaffolding of de novo genome assemblies. Schematic illustration of an hybrid de novo genome assembly approach where linkage information obtained from optical and Hi-C maps is used to properly position contigs along the chromosomes. First, sequencing reads are assembled together to form contigs. Then, contigs are aligned to consensus BioNano optical maps, a process that enables to accurately order and orient the contigs with respect to each others and assign each contig to a specific chromosome. Here, BioNano optical maps are represented by black lines and the green dots indicate the labeled sequence motifs. Further ordering of the resulting scaffolds is made possible by Hi-C maps, that take into account the interactions between contiguous genomic loci. In the resulting chromosome-level scaffold colored segments represent genomic regions of known sequence while the remaining gaps between the scaffolds of unknown sequence are depicted as gray lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Box 3. Bioinformatics

The continuous advancements experienced by sequencing technologies in the last 70 years, and the resulting exponential increase in the amount of sequencing data available, have been paralleled by improvements in data analysis tools. In the 1980s, the advent of information technology laid the foundations of bioinformatics. Processor computing allowed to automate the general principles of overlapping sequences by similarity using dedicated computer programs [16,161], the first genome assemblers. In 1980, to describe the data obtained after assembly of shotgun sequencing reads, Staden coined the word ‘contig’. In his words, “a contig is a set of gel readings that are related to one another by overlap of their sequences. [...] The gel readings in a contig can be summed to form a contiguous consensus sequence and the length of this sequence is the length of the contig” [161]. This is probably the first outline of the general principle of Overlap-Layout-Consensus (OLC) used to infer the original sequence from a subset of sequences. Since brute force alignment (i.e. generate all possible alignments and pick the one most likely to be true) was proven nearly impossible to solve even for short sequences, several overlap algorithms were implemented, including the Needleman–Wunsch algorithm for global alignments [162] and the ‘edit’ or Levenshtein distance [163]. Others were newly developed, such as those for local alignment [164]. Starting from 1985, the National Center for Biotechnology Information (NCBI) released its Basic Local Alignment Search Tool (BLAST) [165], which considerably speeded up the process of sequence alignment. The increased throughput of sequencers combined with the shotgun approach allowed to tackle more complex genomes. Venter’s team at TIGR developed a new software, named TIGR Assembler, that generated the whole genome of *H. Influenzae* out of approximately 24,000 DNA fragments using 30 CPU hours with half a gigabyte of RAM [166]. As the complexity of genomes was progressively understood, more refined tools were made available: RepeatMasker in 1996, to deal with repeated genomic regions [167], and GENSCAN in 1997, to predict gene structures [168]. In 1998, Green developed Phred, an algorithm that is still widely used today, along with two associated softwares for phred output analysis, phrap and consed [169]. Phred assigns a Quality Value (QV) score to each base called using the formula $QV = -10 \times \log_{10}(P_{\text{error}})$ where $P_{\text{error}}$ is the probability of erroneous base-calling (e.g. QV20 implies 99% accuracy while QV30 implies 99.9% accuracy). Phred introduced reliable quality metrics for base calling, providing support to automated read analyses, particularly in repeated sequences [170].

The challenges associated with the human genome (Box 2) further fostered research in the field of genome assembly to an unprecedented scale. This is especially true for the effort led by Venter, that aimed at reconstructing the sequence of the human genome without BAC intermediates. After the release of the human genome assemblies, in 2002 two new online repositories were created in addition to Genbank to host human genome data, the University of California Santa Cruz (UCSC) Genome Browser [171] and Ensembl [172]. The unprecedented flood of sequencing data was barely accompanied by advancements in bioinformatic tools to store, manage, analyze, process, and visualize them. A new series of integrated open-source software and algorithms were released, including a radically new assembly approach based on de Bruijn graphs (DBG) [173]. Contrary to OLC, DBG assembly is counterintuitive, as it first chop the reads into shorter k-mers and then uses the k-mers to form a DBG from which the genome sequence is derived [174].

The first software using this approach was EULER in 2001 [175]. Other major milestones included short-read and vertebrate-specific aligner BLAT [176]; the platform for integrated genome analysis Galaxy [177]; the NCBI Short Read Archive (SRA) to store NGS raw data [2006]; the assembly algorithms ALL-PATHS [178], Velvet [179] and SOAPdenovo [180], all based on de Bruijn graphs; more efficient read alignment algorithms as Bowtie [181] and BWA [182]; integrated tools for read data management as SAMtools [183]; general algorithms for variant discovery as GATK [184] and Freebayes [185].

(3.4. Scaffolding of genome assemblies)

For large genomes, even long reads fail to generate end-to-end chromosome sequences, requiring linkage information to orient and order the contigs, a process known as scaffolding. Scaffolds comprise multiple contigs in their putative order on chromosomes, separated by genomic regions of undetermined sequence (represented by runs of Ns). One leading single-molecule DNA technology relevant for genome scaffolding is Bionano optical mapping, commercialized by Bionano Genomics (BioNano). The method is based on DNA labeling rather than sequencing and aims at generating genetic optical maps. Optical maps were first described for yeast in the 1990s [96]. In the early approach, DNA molecules were stretched and fixed in agarose gel, resident restriction endonucleases were activated by Mg++ addition, and the differential fluorescence patterns between cleaved and intact sites were assessed using a microscope. In 2003, the technique was refined using restriction enzymes directed at specific recognition sites. T7 exonuclease was used to open single-strand gaps at those sites and gaps were filled with nucleotides labeled with fluorescent tags
The current technology relies on the isolation of DNA fragments >150 Kbp and staining of the filaments using restriction enzymes recognizing specific 6–7 bp sequence motifs. DNA molecules are then stretched into nanoscale channels (NanoChannels) through progressively smaller sievers and imaged with a high-resolution camera [98]. Information from individual DNA molecules is built into consensus maps (cmaps) using the software BioNano Solve. The distances between motif-specific patterns represent the optical maps [98]. The linkage information provided by these maps is crucial to improve the process of de novo genome assembly [98,99], potentially identifying and resolving misassemblies [100–102]. Draft assemblies generated using either NGS or TGS sequencing data alone are matched with consensus optical maps. The overlaps allow to orient and order the contigs, resolve chimeric joints, and estimate the length of gaps between adjacent contigs. This produces longer scaffolds, potentially spanning entire chromosomes [43,103] (Fig. 3). Given the improved contiguity of hybrid genome assemblies generated using sequencing data in combination with optical mapping, this approach has been recently adopted by many de novo human [43,104,105], non-human [44,106,107], and plant [108] genome assembly projects (Fig. 1). In addition to its relevance for genome assembly, optical mapping allows the detection of SVs exceeding the length of individual NGS or long reads, by comparing the results to a reference genome [109]. Moreover, optical mapping was demonstrated to efficiently identify genome-wide methylation patterns, through methylation-sensitive restriction enzymes [110]. Until 2018, two enzymes were available for the staining of the DNA molecules: Nt.BspQI with recognition sequence CCTCTCN^ and Nb.BsSI with recognition sequence CACGA^C. These were both nicking enzymes that labeled DNA molecules introducing the fluorescent tag via a nick on one of the two DNA strands. This approach is collectively referred as Nick, Label, Repair and Stain (NLRS). In 2018, BioNano introduced the Direct Label and Stain (DLS) approach based on the enzyme DLE-1, with recognition sequence CTTAAG. This enzyme can label the DNA without nicking it, considerably increasing the average length of the molecules that are stained and visualized. Currently, this is the only non-nicking enzyme available, however BioNano claims to release several more over the next few years.

Another popular scaffolding approach is Hi-C, a Chromosome Conformation Capture (3C) based technique that generates chromatin proximity information for all vs all regions of the genome [111]. This spatial information can be used to arrange contigs and scaffolds in a linear, potentially chromosome-level, sequence [112,186][186]. In Hi-C sample preparation, cells are embedded in a matrix and the layers surrounding chromosomes are progressively washed away, thus preserving chromosome folding information. The DNA is then treated with a restriction enzyme and ligated to form new bonds with different, but spatially close, DNA molecules. The resulting library is sequenced with illumina short reads and the obtained ‘map’ of DNA interactions is used to assign contigs and scaffolds to chromosomes [113,114] (Fig. 3). While Hi-C currently appears to be the only approach enabling chromosome-level scaffolding in large genomes, it is usually less conservative than BioNano scaffolding. The unpredictable folding of chromatin resulting in long-range interactions of distant regions of chromosomes can lead to misassembly errors such as artificial inversions, scaffold misplacement within the same chromosome, or scaffold misassignment to different chromosomes [112]. These errors are better corrected combining different techniques and therefore several projects have started to adopt complementary scaffolding strategies [115,116] (Fig. 1), even potentially allowing telomere-to-telomere assembly of entire chromosomes without gaps [48].

4. Summary and outlook

The way scientists generate genome assemblies has been and is constantly evolving. Through the constant refinement of existing technologies and the introduction of radically new DNA sequencing approaches and bioinformatic tools, the quality of the assemblies has never ceased to improve. The high-throughput capacity introduced by NGS and the enhanced quality offered by TGS have finally made also complex genomes accessible for whole-genome investigation. Human genetics research, including population genomics, genetic diseases mapping and diagnostics, personalized medicine programs, cancer research and prenatal testing have already benefited from the advancements in genome sequencing and assembly of the last decade. Similarly, these approaches are being increasingly employed in non-model organisms to understand ecological and evolutionary processes. The commitment to reference genome sequencing and assembly has now scaled up from single-species projects to multiple-species coordinated efforts, and projects aimed to produce high-quality genomes for most organisms using a combination of NGS and TGS approaches are currently underway.

Acknowledgments

We apologize to those whose work we were unable to discuss and cite in this review due to space constraints. We thank Erich Jarvis for critical reading of the manuscript and inputs. We also thank the two anonymous reviewers for providing insightful comments that helped improve and clarify this manuscript.

Declaration of interest

G.F. had one travel sponsored by BioNano Genomics. A.M.G., G.R.G. and L.G. declare no conflicts of interest.

References

[1] Watson JD, Crick FHC. Molecular structure of nucleic acids. Nature 1953;171:737–8.
[2] Sanger F, Thompson EOP. The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from enzymic hydrolysates. Biochem J 1953;53:366–74.
[3] Sanger F, Thompson EOP. The amino-acid sequence in the glycyl chain of insulin. II. The identification of lower peptides from partial hydrolysates. Biochem J 1953;53:353–66.
[4] Holley RW, Appar J, Everett GA, Madison JT, Marquisee M, et al. Structure of a ribonucleic acid. Science 1965;147:1462–5.
[5] Wu R, Kaiser AD. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. J Mol Biol 1968;35:523–37.
[6] Brownlee GG, Sanger F, Barrett BG. The sequence of 5 s ribosomal ribonucleic acid. J Mol Biol 1968;34:379–412.
[7] Min Jou W, Haegeman G, Ysebaert M, Fiers W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature 1972;237:82–8.
[8] Gilbert W, Maxam A. The nucleotide sequence of the lac operon. Proc Natl Acad Sci USA 1973;70:3581–4.
[9] Fiers W, Contreras R, Duerrnick F, Haegeman G, Iserentant D, et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 1976;260:500–7.
[10] Sanger F, Coulson AR. A rapid method for determining sequences by DNA primed synthesis with DNA polymerase. J Mol Biol 1975;94:441–8.
[11] Maniatis T, Jeffrey A, van de Sande H. Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis. Biochemistry 1975;14:3787–94.
[12] Sanger F, Air GM, Barrett BG, Brown NL, Coulson AR, et al. Nucleotide sequence of bacteriophage 6074 DNA. Nature 1977;265:687–95.
[13] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977;74:5463–7.
[14] Atkinson MR, Deutscher MP, Kornberg A, Russell AF, Moffatt JG. Enzymatic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2,3- dideoxyribonucleotide. Biochemistry 1969;8:4897–904.
[15] Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci USA 1977;74:560–4.
Rand AC, Banerjee O, Fang G, Feng Z, Wong WH, et al. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. Genome Res 2013;23:129–41.

Raj AC, Jain M, Eisenberg JM, Musselman-Brown A, Olsen HE, et al. Mapping DNA methylation with high-throughput nanopore sequencing. Nat Methods 2017;14:411–3.

Alderfer DH, Marks P, Klammer AA, Drake J, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 2013;10:563–9.

Chin C, Peluso P, Sedlacek FJ, Nattestad M, Conception GT, et al. Phased diploid genome assemblies with single-molecule real-time sequencing. Nat Methods 2016;13:1050–4.

Berlin K, Koren S, Chin C, Drake J, Landolin JM, et al. Assembling large genomes with single-molecule sequencing and locality sensitive hashing. Nat Biotechnol 2017;35:333–8.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 2012;30:993–700.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, et al. Canu: scalable and accurate long-read assembly via iterative k-mer weighting and repeat separation. Genome Res 2017;27:722–36.

Koren S, Rhee A, Walenz BF, Dilthey AT, Bickhart DM, et al. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol 2018;36:1174–82.

Liu Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, et al. Assembly of long error-prone reads using de Bruijn graphs. Proc Natl Acad Sci USA 2016;113:7045–50.

Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long reads. Bioinformatics 2016;32:2103–10.

Kamath GM, Shomorony I, Xia F, Courtaud T, Tse DN, HINEC: long-read analysis achieves high resolution of human repeat regions. Genome Res 2017;27:747–57.

Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. BioRxiv 2015:150972.

Kolmogorov M, Yuan J, Lin Y, Pevez PA, Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 2019;37:540–6.

Chin C-S, Khakta A. Human Genome Assembly in 100 Minutes. BioRxiv 2019:105616.

Shafin K, Pesout T, Long-Roach T, Haukness M, Olsen HE, et al. Efficient de novo assembly of eleven human genomes using PromethION sequencing and a novel nanopore toolkit. bioRxiv 2019:95611.

Dominguez Del Angel V, Hjerde E, Capella-Gutierrez S, Notredame C, Dominguez Del Angel V, Hjerde E, Capella-Gutierrez S, Notredame C, et al. Assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 2017;356:92–5.

Lee LG, Connell CR, Woo SL, Cheng RD, McArdle BF, et al. The draft genome of tropical fruit durian (Durozium zibethinus). Nat Genet 2017;49:1633–41.

Van den Wyngaarden MG, Hooger M, Nyquist SK, Seo H, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 2017:356:92–5.

Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, et al. The draft genome of tropical fruit durian (Durozium zibethinus). Nat Genet 2017;49:1633–41.

Wallberg A, Kuniks P, Tuts JU, Mosbech M-B, Childers AK, et al. A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. BMC Genomics 2019;20:275.

Smith LM, Fung S, Hunakpiller MW, Hunakpiller TJ, Howard O. De novo genome assembly and annotation of a large diploid lineage within the genus Tetrathyrea. Science 2018;361:689–93.

Hastie AR, Dong L, Li H. Accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 2017;27:722–36.

Schwartz DC, Li X, Hernandez LD, Kidd KF, et al. Complete genome sequence of the malaria mosquito, Anopheles gambiae. Nat Genet 2007;39:21–30.

Lee LG, Connell CR, Woo SL, Cheng RD, McArdle BF, et al. The draft genome of tropical fruit durian (Durozium zibethinus). Nat Genet 2017;49:1633–41.

Van den Wyngaarden MG, Hooger M, Nyquist SK, Seo H, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 2017;356:92–5.
