REVIEW

doi:10.1038/nature24286

# DNA sequencing at 40: past, present and future

Jay Shendure[1,2], Shankar Balasubramanian[3,4], George M. Church[5], Walter Gilbert[6], Jane Rogers[7], Jeffery A. Schloss[8] & Robert H. Waterston[1]

**This review commemorates the 40th anniversary of DNA sequencing, a period in which we have already witnessed multiple technological revolutions and a growth in scale from a few kilobases to the first human genome, and now to millions of human and a myriad of other genomes. DNA sequencing has been extensively and creatively repurposed, including as a 'counter' for a vast range of molecular phenomena. We predict that in the long view of history, the impact of DNA sequencing will be on a par with that of the microscope.**

DNA sequencing has two intertwined histories—that of the underlying technologies and that of the breadth of problems for which it has proven useful. Here we first review major developments in the history of DNA sequencing technologies (Fig. 1). Next we consider the trajectory of DNA sequencing applications (Fig. 2). Finally, we discuss the future of DNA sequencing.

## History of DNA sequencing technologies

The development of DNA sequencing technologies has a rich history, with multiple paradigm shifts occurring within a few decades. Below, we review early efforts to sequence biopolymers, the invention of electrophoretic methods for DNA sequencing and their scaling to the Human Genome Project, and the emergence of second (massively parallel) and third (real-time, single-molecule) generation DNA sequencing. Some key technical milestones are also summarized in Box 1.

### Early sequencing

Fred Sanger devoted his scientific life to the determination of primary sequence, believing that knowledge of the specific chemical structure of biological molecules was necessary for a deeper understanding[1]. Ironically, given the state of sequencing technology for each biopolymer today, proteins and RNA came first.

The first protein sequence, of insulin, was determined in the early 1950s by Sanger, who fragmented its two chains, deciphered each fragment and overlapped the fragments to yield a complete sequence. His work showed unequivocally that proteins had defined patterns of amino acid residues[2]. The later development of Edman degradation, a repeated elimination of an N-terminal residue from the peptide chain, made protein sequencing easier[3]. Although these methods were cumbersome, many proteins had been sequenced by the late 1960s, and it became clear that each protein's sequence varied across species and between individuals.

In the 1960s, RNA sequencing was tackled by this same general process: an RNA species was first fragmented with RNases, next the pieces were separated by chromatography and electrophoresis, then individual fragments were deciphered by sequential exonuclease digestion, and finally the sequence was deduced from the overlaps. The first RNA sequence, of alanine tRNA, required five people working three years with one gram of pure material (isolated from 140 kg of yeast) to determine 76 nucleotides[4]. This process was greatly simplified by 'fingerprinting' techniques, which included the separation of radioactively labelled RNA fragments and visualization in two dimensions, with the resulting positions diagnostic of their size and sequence[5].

### The invention of DNA sequencing

Early attempts to sequence DNA were cumbersome. In 1968, Wu reported the use of primer extension methods to determine 12 bases of the cohesive ends of bacteriophage lambda[6]. In 1973, Gilbert and Maxam reported 24 bases of the lactose-repressor binding site, by copying it into RNA and sequencing those fragments. This took two years: one base per month[7].

The development, in around 1976, of two methods that could decode hundreds of bases in an afternoon transformed the field. Both methods—the chain terminator procedure developed by Sanger and Coulson, and the chemical cleavage procedure developed by Maxam and Gilbert—used distances along a DNA molecule from a radioactive label to positions occupied by each base to determine nucleotide order. Sanger's method involved four extensions of a labelled primer by DNA polymerase, each with trace amounts of one chain-terminating nucleotide, to produce fragments of different lengths[8]. Gilbert's method took a terminally labelled DNA-restriction fragment, and, in four reactions, used chemicals to create base-specific partial cleavages[9]. For both methods, the sizes of fragments present in each base-specific reaction were measured by electrophoresis on polyacrylamide slab gels[10], which enabled separation of the DNA fragments by size with single-base resolution. The gels, with one lane per base, were put onto X-ray film, producing a ladder image from which the sequence could be read off immediately, going up the four lanes by size to infer the order of bases.

These methods came into immediate use. Shotgun sequencing—sequencing of random clones followed by sequence assembly based on the overlaps—was suggested by Staden in 1979[11], greatly facilitated by Messing's development of the single-stranded M13 phage cloning vector around 1980[12], and used to assemble genomes *de novo*, such as bacteriophage lambda as early as 1982[13]. By 1987, automated, fluorescence-based Sanger sequencing machines, developed by Smith, Hood and Applied Biosystems[14,15], could generate around 1,000 bases per day. Sequence data grew exponentially, approximating Moore's law and motivating the creation of central data repositories (such as GenBank) that, through search tools (such as BLAST[16]), amplified the value of each sequence and engendered a spirit of data sharing. By 1982, over half a million bases had been deposited in GenBank; by 1986, nearly 10 million bases (GenBank and WGS Statistics; https://www.ncbi.nlm.nih.gov/genbank/statistics/).

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Howard Hughes Medical Institute, Seattle, Washington, USA. [3]Department of Chemistry, University of Cambridge, Cambridge, UK. [4]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [5]The Wyss Institute & Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [6]Department of Molecular and Cellular Biology, Harvard University, Cambridge Massachusetts, USA. [7]International Wheat Genome Sequencing Consortium, Little Eversden, Cambridge, UK. [8]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA.

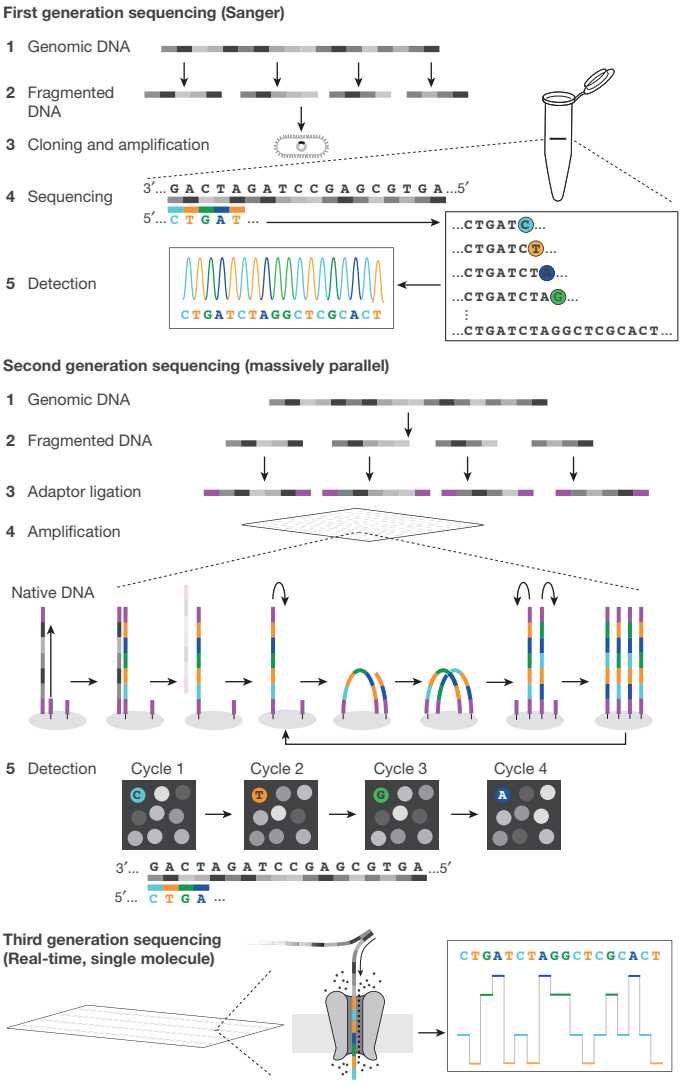


**Figure 1 | DNA sequencing technologies.** Schematic examples of first, second and third generation sequencing are shown. Second generation sequencing is also referred to as next-generation sequencing (NGS) in the text.

## Scaling to the human genome

For the 'hierarchical shotgun' strategy that emerged as the workhorse of the Human Genome Project (HGP), large fragments of the human genome were cloned into bacterial artificial chromosomes (BACs). DNA from each BAC was fragmented, size-selected and sub-cloned. Individual clones were picked and grown, and the DNA was isolated. The purified DNA was used as a template for automated Sanger sequencing, the signal was extracted from laser-scanned images of the gels, and bases were called to finally produce the sequence. The fact that this process involved many independent steps, each of which had to work well, led sceptics to doubt it could ever be made efficient enough to sequence the human genome at any reasonable cost.

Indeed, as efforts to sequence larger genomes took shape, it became clear that the scale and efficiency of each step needed to be vastly increased. This was achieved in fits and spurts in the 1990s. Noteworthy improvements included: (1) a switch from dye-labelled primers to dye-labelled terminators, which allowed one rather than four sequencing reactions[17]; (2) a mutant T7 DNA polymerase that more readily incorporated dye-labelled terminators[18]; (3) linear amplification reactions, which greatly reduced template requirements and facilitated miniaturization[19]; (4) a magnetic bead-based DNA purification method that simplified automation of presequencing steps[20]; (5) methods enabling sequencing of double-stranded DNA, which enabled the use of plasmid clones and therefore paired-end sequencing; (6) capillary electrophoresis, which eliminated the pouring and loading of gels, while also simplifying the extraction and interpretation of the fluorescent signal[21]; (7) adoption of industrial processes to maximize efficiencies and minimize errors (for example, automation, quality control, standard operating procedures, and so on).

Wet laboratory protocols were only half the challenge. Substantial effort was invested into the development of software to track clones, and into the interpretation and assembly of sequence data. For example, manual editing of the sequence reads was replaced by the development of phred, which introduced reliable quality metrics for base calls and helped sort out closely related repeat sequences[22]. Individual reads were then assembled from overlaps in a quality-aware fashion to generate long, continuous stretches of sequence. As more complex genomes were tackled, repetitive sequences were increasingly confounding. Even after deep shotgun sequencing of a BAC, some sequences were not represented, resulting in discontinuities that had to be tackled with other methods. Paired-end sequencing[23] helped to link contigs into gapped scaffolds that could be followed up by directed sequencing to close gaps. Some problems were only resolved by eye; scientists who were trained 'finishers' assessed the quality and signed off on the assembled sequence of individual clones[22].

Although the process remained stable in its outlines, rapid-fire improvements led to steady declines in the cost throughout the 1990s, while parallel advances in computing increasingly replaced human decision making. By 2001, a small number of academic genome centres were operating automated production lines generating up to 10 million bases per day. Software for genome assembly matured both inside and outside of the HGP, with tools, such as phrap, the TIGR assembler and the Celera assembler, able to handle genomes of increasing complexity[22,24,25]. A yearly doubling in capacity enabled the successful completion of high-quality genomes beginning with *Haemophilus influenza* (around 2 Mb, 1995) followed quickly by *Saccharomyces cerevisiae* (around 12 Mb, 1996) and *Caenorhabditis elegans* (around 100 Mb, 1998)[26–28]. The HGP's human genome, which is 30 times the size of *C. elegans* and with much more repetitive content, came first as a draft (2001) and then as a finished sequence (2004)[29,30]. The HGP was paralleled by a private effort to sequence a human genome by Craig Venter and Celera (2001)[31] with a whole-genome shotgun strategy piloted on *Drosophila melanogaster* (around 175 Mb; 2000)[32]. The strategic contrasts between these projects are further discussed below.

By 2004, instruments were churning out 600–700 bp at a cost of US$1 per read, but creating additional improvements was an increasingly marginal exercise. Furthermore, with the completion of the HGP, the future of large-scale DNA sequencing was unclear.

## Massively parallel DNA sequencing

Throughout the 1980s and 1990s, several groups explored alternatives to electrophoretic sequencing. Although these efforts did not pay off until after the HGP, within a decade of its completion, 'massively parallel' or 'next-generation' DNA sequencing (NGS) almost completely superseded Sanger sequencing. NGS technologies sharply depart from electrophoretic sequencing in several ways, but the key change is multiplexing. Instead of one tube per reaction, a complex library of DNA templates is densely immobilized onto a two-dimensional surface, with all templates accessible to a single reagent volume. Rather than bacterial cloning, *in vitro* amplification generates copies of each template to be sequenced. Finally, instead of measuring fragment lengths, sequencing comprises cycles of biochemistry (for example, polymerase-mediated incorporation of fluorescently labelled nucleotides) and imaging, also known as 'sequencing-by-synthesis' (SBS).

Although amplification is not strictly necessary (for example, single-molecule SBS[33–35]), the dense multiplexing of NGS, with millions to billions of immobilized templates, was largely enabled by clonal *in vitro* amplification. The simplest approach, termed 'polonies' or 'bridge amplification', involves amplifying a complex template library with primers immobilized on a surface, such that copies of each template remain tightly clustered[36–39]. Alternatively, clonal PCR can be performed in an emulsion,

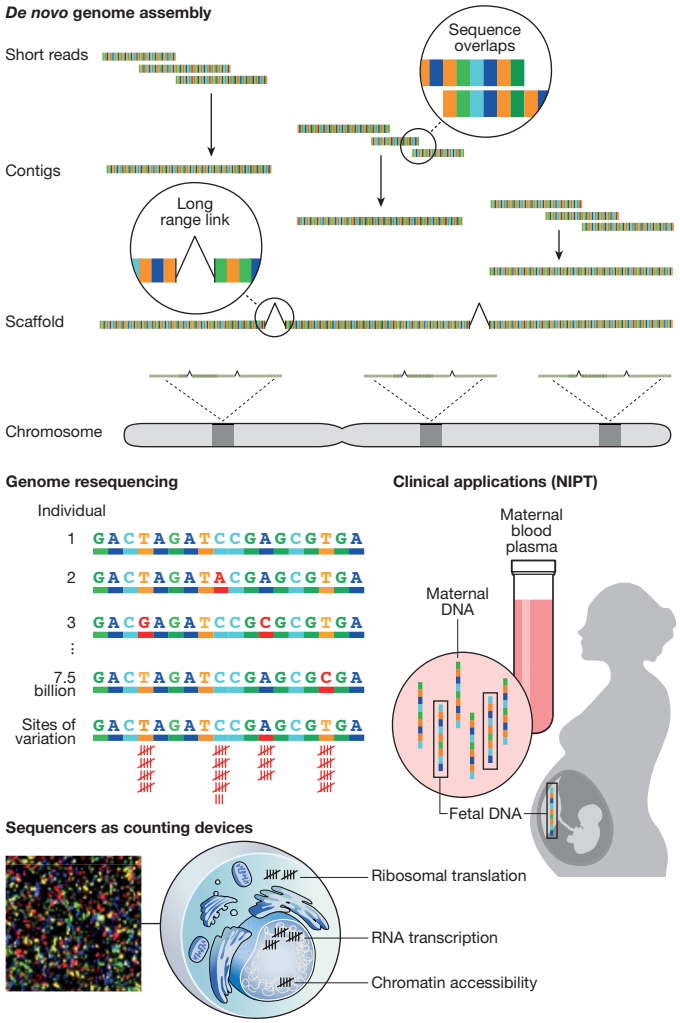


**Figure 2 | DNA sequencing applications.** Major categories of the application of DNA sequencing include *de novo* genome assembly, individual genome resequencing, clinical applications such as non-invasive prenatal testing, and using sequencers as counting devices for a broad range of biochemical or molecular phenomena.

such that copies of each template are immobilized on beads that are then arrayed on a surface for sequencing[40–42]. A third approach involves rolling circle amplification in solution to generate clonal 'nanoballs' that are arrayed and sequenced[43].

For SBS, there were three main strategies. The pyrosequencing approach of Ronaghi and Nyrèn involves discrete, step-wise addition of each deoxynucleotide (dNTP). Incorporation of dNTPs releases pyrophosphate, which powers the generation of light by firefly luciferase[44]. With an analogous approach, natural dNTP incorporations can be detected with an ion-sensitive field effect transistor[45,46]. A second strategy uses the specificity of DNA ligases to attach fluorescent oligonucleotides to templates in a sequence-dependent manner[41,43,47,48]. A third approach, which has proven the most durable, involves the stepwise, polymerase-mediated incorporation of fluorescently labelled deoxynucleotides[33,34,49]. Critical to the success of polymerase-mediated SBS, was the development of reversibly terminating, reversibly fluorescent dNTPs, and a suitably engineered polymerase[50], such that each template incorporates one and only one dNTP on each cycle. After imaging to determine which of four colours was incorporated by each template on the surface, both blocking and fluorescent groups are removed to set up the next extension[51–53]; this general approach was used by Solexa, founded by Balasubramanian and Klenerman in 1998.

The first integrated NGS platforms came in 2005, with resequencing of *Escherichia coli* by Shendure, Porreca, Mitra and Church[41], *de novo* assembly of *Mycoplasma genitalium* by Margulies, Rothberg and 454 (ref. 40), and resequencing of phiX174 and a human BAC by Solexa[54]. These studies demonstrated how useful even very short reads are, given a reference genome to which to map them. Within three years, human genome resequencing would become practical on the Solexa platform with 35-bp paired reads[55].

In 2005, 454 released the first commercial NGS instrument. In the wake of the HGP, large-scale sequencing was still the provenance of a few genome centres. With 454 and other competing instruments that followed closely after, individual laboratories could instantly access the capacity of an entire HGP-era genome centre. This 'democratization' of sequencing capacity had a profound impact on the culture and composition of the genomics field, with new methods, results, genomes and other innovations arising from all corners.

In contrast to the monopoly of Applied Biosystems during the HGP, several companies, including 454 (acquired by Roche), Solexa (acquired by Illumina), Agencourt[47,48] (acquired by Applied Biosystems), Helicos[34,35] (founded by Quake), Complete Genomics[43] (founded by Drmanac) and Ion Torrent[46] (founded by Rothberg), intensely competed on NGS, resulting in a rapidly changing landscape with new instruments that were flashily introduced at the annual Advances in Genome Biology and Technology (AGBT) meeting in Marco Island, Florida. Between 2007 and 2012, the raw, per-base cost of DNA sequencing plummeted by four orders of magnitude[56].

Since 2012, the pace of improvement has slowed, as has the competition. The 454, SOLiD and Helicos platforms are no longer being developed, and the Illumina platform is dominant (although Complete Genomics[43] remains a potential competitor). Nonetheless, it is astonishing to consider how far we have come since the inception of NGS in 2005. Read lengths, although still shorter than Sanger sequencing, are in the low hundreds of bases, and mostly over 99.9% accurate. Over a billion independent reads, totalling a terabase of sequence, can be generated in two days by one graduate student on one instrument (Illumina NovaSeq) for a few thousand dollars. This exceeds the approximately 23 gigabases that were generated for the HGP's draft human genome by a factor of 40.

## Real-time, single-molecule sequencing

Nearly all of the aforementioned platforms require template amplification. However, the downsides of amplification include copying errors, sequence-dependent biases and information loss (for example, methylation), not to mention added time and complexity. In an ideal world, sequencing would be native, accurate and without read-length limitations. To reach this goal, stretching back to the 1980s, a handful of groups explored even more radical approaches than NGS. Many of these were dead ends, but at least two approaches were not, as these have recently given rise to real-time, single-molecule sequencing platforms that threaten to upend the field once again.

A first approach, initiated by Webb and Craighead and further developed by Korlach, Turner and Pacific Biosciences (PacBio), is to optically observe polymerase-mediated synthesis in real time[57,58]. A zero mode waveguide, a hole less than half the wavelength of light, limits fluorescent excitation to a tiny volume within which a single polymerase and its template reside. Therefore, only fluorescently labelled nucleotides incorporated into the growing DNA chain emit signals of sufficient duration to be 'called'. The engineered polymerase is highly processive; reads over 10 kb are typical, with some reads approaching 100 kb. The throughput of PacBio is still over an order of magnitude less than the highest-throughput NGS platforms, such as Illumina, but not so far from where NGS platforms were a few years ago. Error rates are very high (around 10%) but randomly distributed. PacBio's combination of minimal bias (for example, tolerance of extreme GC content), random errors, long reads and redundant coverage can result in *de novo* assemblies of unparalleled quality with respect to accuracy and contiguity, for many species exceeding what would be possible even with efforts similar to the HGP.

A second approach is nanopore sequencing. This concept, which was first hypothesized in the 1980s[59–61], is based on the idea that patterns in

**BOX 1**

The milestones listed below correspond to key developments in the evolution of sequencing technologies. This is a large topic, and we apologize for any omissions.

**Technical milestones**

1953: Sequencing of insulin protein[2]
1965: Sequencing of alanine tRNA[4]
1968: Sequencing of cohesive ends of phage lambda DNA[6]
1977: Maxam–Gilbert sequencing[9]
1977: Sanger sequencing[8]
1981: Messing's M13 phage vector[12]
1986–1987: Fluorescent detection in electrophoretic sequencing[14,15,17]
1987: Sequenase[18]
1988: Early example of sequencing by stepwise dNTP incorporation[139]
1990: Paired-end sequencing[23]
1992: Bodipy dyes[140]
1993: *In vitro* RNA colonies[37]
1996: Pyrosequencing[44]
1999: *In vitro* DNA colonies in gels[38]
2000: Massively parallel signature sequencing by ligation[47]
2003: Emulsion PCR to generate *in vitro* DNA colonies on beads[42]
2003: Single-molecule massively parallel sequencing-by-synthesis[33,34]
2003: Zero-mode waveguides for single-molecule analysis[57]
2003: Sequencing by synthesis of *in vitro* DNA colonies in gels[49]
2005: Four-colour reversible terminators[51–53]
2005: Sequencing by ligation of *in vitro* DNA colonies on beads[41]
2007: Large-scale targeted sequence capture[93–96]
2010: Direct detection of DNA methylation during single-molecule sequencing[65]
2010: Single-base resolution electron tunnelling through a solid-state detector[141]
2011: Semiconductor sequencing by proton detection[142]
2012: Reduction to practice of nanopore sequencing[143,144]
2012: Single-stranded library preparation method for ancient DNA[145]

the flow of ions, which occur when a single-stranded DNA molecule passes through a narrow channel, will reveal the primary sequence of the strand. Decades of work were required to go from concept to reality. Firstly, electric field-driven transport of DNA through a nanometre-scale pore is so fast that the number of ions per nucleotide is insufficient to yield an adequate signal. Solutions have eventually been developed to these and other challenges, including interposing an enzyme as a 'ratchet', identifying and engineering improved nucleopore proteins, and better analytics of the resulting signals[62]. These advances recently culminated in successful nanopore sequencing, in both academia[63] and industry, most prominently by Oxford Nanopore Technologies (ONT), founded by Bayley in 2005. Sequence read lengths of ONT are on par with or exceed the reads generated by PacBio; with the longest obtained reads presently at 900 kilobases (ref. 64). A major differentiator from other sequencing technologies is the extreme portability of nanopore devices, which can be as small as a memory (USB) stick, because they rely on the detection of electronic, rather than optical, signals. Important challenges remain (for example, errors may not be randomly distributed), but progress is rapid.

Nucleic-acid sequencing would ideally also capture DNA modifications. Indeed, both PacBio and nanopore sequencing have demonstrated the detection of native covalent modifications, such as methylation[64,65]. Single-molecule methods also open up the intriguing possibility of directly sequencing RNA[66,67] or even proteins[68–71].

Since 1977, DNA-sequencing technology has evolved at a fast pace and the landscape continues to change shift under our feet. Although Illumina is presently the dominant supplier of sequencing instruments, the commercial market is no longer monolithic and other technologies may successfully occupy important niches (for example, PacBio for *de novo* assembly and ONT for portable sequencing). Neither NGS nor single-molecule methods have fully plateaued in cost and throughput, and there are additional concepts that are still in development, which are not discussed here (for example, solid-state pores and electron microscopy)[70,71]. Not all will work out, but as the above examples make clear, transformative sequencing technologies can take decades to mature.

## Applications of DNA sequencing

The range and scope of DNA sequencing applications has also expanded over the past few decades, shaped in part by the evolving constraints of sequencing technologies. Below we review key areas of application including *de novo* genome assembly, individual genome resequencing, sequencing in the clinic and the transformation of sequencers into molecular counting devices. Some key milestones for the generation of reference genomes and development of applications and software are summarized in Box 2.

### *De novo* genome assembly

For its first 25 years, the primary purpose of DNA sequencing was the partial or complete sequencing of genomes. Indeed, the inception of Sanger sequencing in 1977 included the first genome (phiX174; 5.4 kb), essentially assembled by hand[72]. However, DNA sequencing was only one of several technologies that enabled assembly of larger genomes. If the DNA sequence was random, arbitrarily large genomes could be assembled to completion solely based on fragment overlaps. However, it is not random, and the combination of repetitive sequences and technical biases makes it impossible to obtain high-quality assemblies of large genomes from kilobase-scale reads alone. Additional 'contiguity information' is required.

For the HGP[29,30], these additional sources of contiguity information included the following. (1) Genetic maps, which were based on the segregation of genetic polymorphisms through pedigrees, that provided orthogonal information about the order of sequences locally and at the scale of chromosomes. (2) Physical maps, for which BACs were cloned, restriction-enzyme 'fingerprinted' to identify overlaps and ordered into a 'tiling path' that spanned the genome. Clones were individually shotgun sequenced and assembled, thereby isolating different repeat copies from one another, and then further ordered and assembled. (3) Paired-end sequencing, introduced by Ansorge in 1990[23], comprises sequencing into both ends of a DNA fragment of approximately known length, effectively linking those end-sequences. Depending on the cloning method, the spanned length could range from a few kilobases to a few hundred kilobases. Sequence coverage at 8–10-fold redundancy, coupled to these sources of contiguity information, enabled not only genome assembly, but also improved quality to about 1 error per 100,000 bases for most of the genome. Additional, focused experiments were performed to fill the gaps or clarify ambiguities.

The Celera effort went straight to paired-end sequencing, eschewing physical maps as an intermediate[31]. An important advance was the transition from greedy algorithms, such as phrap and the TIGR assembler, to the Celera assembler's graph-based approach (overlap–layout–consensus)[22,24,25]. Although Celera had a reasonable strategy for a draft genome, because of the pervasiveness of repetitive sequences, it did not, by itself, result in a high-quality reference, such as the one produced by the HGP's clone-based approach. The current human reference genome descends from the HGP's 2004 product[30], with continuous work by the Genome Reference Consortium to further improve it, including regular releases of reference genome updates[73].

With the advent of NGS in 2005, the number of *de novo* assemblies increased vastly. The seemingly disastrous combination of short reads and repetitive genomes was overcome by new assembly algorithms based on de Bruijn graphs (for example, EULER and Velvet)[74,75]. Nonetheless, particularly when applied to larger genomes and when compared to the genomes of the HGP, their quality was, on average, quite poor. Although shorter read lengths are partly to blame, this is usually overstated. Instead, a principal reason for the poorer quality was the paucity of contiguity

## BOX 2

The milestones listed below correspond to key developments in the availability of new reference genomes, new sequencing-related computational tools and the applications of DNA sequencing in new ways or to new areas. These are large topics, and we apologize for any omissions.

**Genome milestones**

1977: Bacteriophage ΦX174 (ref. 72)
1982: Bacteriophage lambda[13]
1995: *Haemophilus influenzae*[26]
1996: *Saccharomyces cerevisiae*[27]
1998: *Caenorhabditis elegans*[28]
2000: *Drosophila melanogaster*[32]
2000: *Arabidopsis thaliana*[146]
2001: *Homo sapiens*[29–31]
2002: *Mus musculus*[147]
2004: *Rattus norvegicus*[148]
2005: *Pan troglodytes*[149]
2005: *Oryza sativa*[150]
2007: *Cyanidioschyzon merolae*[126]
2009: *Zea mays*[151]
2010: Neanderthal[88]
2012: Denisovan[145]
2013: The HeLa cell line[152,153]
2013: *Danio rerio*[154]
2017: *Xenopus laevis*[155]

**Computational milestones**

1981: Smith–Waterman[156]
1982: GenBank (https://www.ncbi.nlm.nih.gov/genbank/statistics/)
1990: BLAST[16]
1995: TIGR assembler[24]
1996: RepeatMasker
1997: GENSCAN[157]
1998: phred, phrap, consed[22]
2000: Celera assembler[25]
2001: Bioconductor
2001: EULER[74]
2002: BLAT[158]
2002: UCSC Genome Browser[159]
2002: Ensembl[160]
2005: Galaxy[161]
2007: NCBI Short Read Archive
2008: ALLPATHS[162]
2008: Velvet[75]
2009: Bowtie[83]
2009: BWA[82]
2009: SAMtools[84]
2009: BreakDancer[163]
2009: Pindel[164]
2009: TopHat[115]
2010: SOAPdenovo[165]
2010: GATK[85]
2010: Cufflinks[116]
2011: Integrated Genomics Viewer[166]
2013: HGAP/Quiver[167]
2017: Canu[81]

**Application milestones**

1977: Genome sequencing[72]
1982: Shotgun sequencing[13]
1983, 1991: Expressed sequence tags[107,108]
1995: Serial analysis of gene expression[109]
1998: Large-scale human SNP discovery[168]
2004: Metagenome assembly[122]
2005: Bacterial genome resequencing with NGS[40,41]
2007: Chromatin immunoprecipitation followed by sequencing (ChIP–seq) using NGS[117]
2007–2008: Human genome and cancer genome resequencing using NGS[55,90–92]
2008: RNA-seq using NGS[110–114]
2008: Chromatin accessibility using NGS[118]
2009: Exome resequencing using NGS[97]
2009: Ribosome profiling using NGS[119]
2010: Completion of Phase I of the 1000 Genomes Project[98]
2010: *De novo* assembly of a large genome from short reads[169]
2011: Haplotype-resolved human genome resequencing using NGS[170,171]
2016: Human genome *de novo* assembly with PacBio[172]
2017: Human genome *de novo* assembly with nanopore[64]

methods to complement NGS. Paired-end sequencing was possible with NGS, but *in vitro* library methods were more restricted with respect to the distances that could be spanned. Furthermore, the field lacked 'massively parallel' equivalents of genetic and physical maps.

This 'dark' period notwithstanding, there are good reasons to be optimistic about the future of *de novo* assembly. Firstly, *in vitro* methods that subsample high molecular weight (HMW) genomic fragments, analogous to hierarchical shotgun sequencing, have recently been developed[76,77]. Secondly, methods, such as Hi-C (genome-wide chromosome conformation capture) and optical mapping, provide scalable, cost-effective means of scaffolding genomes into chromosome-scale assemblies[78–80]. Finally, the read lengths of PacBio and ONT sequencing have risen to hundreds of kilobases, and are now more limited by the preparation of HMW DNA than by the sequencing itself. The absence of cloning or amplification steps in single-molecule sequencing pays off, as shown by high-quality PacBio *de novo* assemblies of bacterial genomes with extreme GC content. Long reads have resulted in a re-emergence of strategies used by the Celera assembler, improved to deal with the high error rates and multiple platforms[81]. By combining long reads and even longer-range contiguity information (for example, subsampling HMW DNA, chromatin proximity, optical maps and so on), *de novo* genome assemblies of the quality of the original human reference genome using 'post-Sanger' approaches are finally within sight[73,80].

## Genome resequencing

After the HGP, a clear next step was to catalogue genetic variation among humans, that is, 'resequencing'. Because Sanger sequencing costs remained high, resequencing was primarily used to discover common variants, which were then cost-effectively genotyped with microarrays to facilitate genome-wide association studies. The rallying cry for changing this was the 'US$1,000 human genome', the ambitious goal of the resequencing of individuals at a cost nearly one-million-fold below that of assembling the first human genome. The US$1,000 genome concept was discussed as early as 2001 (at the University of California, Santa Cruz Human Genome Symposium (http://genomesymposium.ucsc.edu)), when NGS strategies barely existed, and was formalized a few years later by the Revolutionary DNA Sequencing Technologies program of the National Human Genome Research Institute (NHGRI). The commitment of US$220 million in funding to over 40 academic and 27 commercial entities has helped to drive much of the technological development described above, including direct or indirect support of nearly every successful commercial platform.

Resequencing, that is, mapping sequence reads to a reference genome to identify genetic variants, is a very different task than genome assembly. New algorithms, such as Bowtie and Burrows–Wheeler Aligner (BWA), borrowed concepts from data-compression techniques to enable millions of reads to be efficiently mapped to the reference genome[82,83]. Redundant coverage (for example, 30-fold) is necessary to identify heterozygous

variants as well as to distinguish sequencing errors from bona fide variants. Popular packages, initially SAMtools and later GATK, adapted the confidence framework of phred to NGS bases, reads and variants[84,85]. Short reads, particularly when paired, can be uniquely mapped to most of the human genome. But most is not all, and a problem of short-read resequencing is that variants in repetitive regions and structural variants are routinely missed. The extent of this shortcoming is quantified by recent studies that resequence human genomes with PacBio[86]. A second aspect of incompleteness relates to phase relationships between variants in a diploid genome, that is, haplotypes[87]. Fortunately, haplotypes are recovered by many of the same methods that enable contiguity for *de novo* NGS assemblies (and ideally, even *de novo* assemblies would be haplotype-resolved)[77]. Although still not broadly used, these methods are becoming increasingly scalable.

The HGP's human genome was constructed from a mosaic of DNA donors, but mostly derives from one individual, from Buffalo, New York, who had roughly equal parts European and African ancestry[88]. The first individual to have their whole genome resequenced was Craig Venter in 2007, one of the subjects of the Celera genome, which was supplemented with additional data[89]. This was quickly followed in 2008 by the genome of Jim Watson on 454 (ref. 90), and then the genomes of two anonymous individuals[55,91] and the germline and tumour genome of a patient[92] on Solexa/Illumina, and five individuals on Complete Genomics[43]. In this period, whole-genome sequencing (WGS) remained too expensive for most groups to scale, motivating the development of targeted capture methods[93–96] and then whole-exome sequencing (WES), that is, selective sequencing of the 1–2% of the genome that is encodes proteins[97].

As costs approached US$1,000 for WGS[56] and a few hundred dollars for WES, the pace at which individual humans are resequenced has accelerated. The 1000 Genomes Project, launched in 2008, released low-coverage WGS of a few hundred individuals in 2010 and a few thousand individuals in 2015[98,99]. The Exome Sequencing Project released over 6,500 exomes in 2013[100]. The recently released Genome Aggregation Database (http://gnomad.broadinstitute.org/) includes more than 120,000 exomes and over 15,000 genomes. The Genomics England (https://www.genomicsengland.co.uk/), GenomeAsia100K (http://www.genomeasia100k.com/) and NHLBI TOPMed (Trans-Omics for Precision Medicine, https://www.nhlbiwgs.org/) projects each aim to complete WGS on more than 100,000 individuals within the next year or two. Given that these projects represent a fraction of all sequencing being conducted, it is plausible that the genomes of over one million humans have already been resequenced by WES or WGS.

### Clinical applications of sequencing

Our ability to sequence human genomes has vastly outpaced our ability to interpret genetic variation, which partly explains why clinical medicine has yet to wholeheartedly embrace WGS. Nonetheless, there are some areas in which DNA sequencing is already proving clinically useful, three of which we highlight here.

The most unexpected area of the clinical impact of DNA sequencing has been non-invasive prenatal testing (NIPT, see Fig. 2). Pioneering studies by Lo and Quake in 2008 have demonstrated that the simple counting of DNA fragments released into the maternal circulation by a fetus during pregnancy can detect chromosomal aneuploidies[101,102]. Screening tests that were based on this strategy were adopted faster than any molecular test in history, and several million pregnant women around the world have already benefited from low-pass WGS for NIPT.

An early application of WES was to rapidly discover new genes for, and to diagnose patients affected by, Mendelian disorders[97,103]. This was quickly followed by the discovery that substantial proportions of neurodevelopmental disorders are attributable to *de novo* mutations in coding sequences[104]. WES is increasingly used as a primary tool for diagnosing Mendelian and neurodevelopmental disorders, particularly in paediatric populations, with the rate of diagnosis of patients with suspected Mendelian disease by WES at 25% and rising[105].

Our understanding of cancer, fundamentally a disease of the genome, is gradually being transformed by DNA sequencing. Large-scale resequencing has laid bare the extraordinary genetic heterogeneity of cancers, effectively defining a molecular taxonomy[106]. DNA sequencing is impacting clinical cancer care by: (1) suggesting targeted therapies, based on the mutations present in an individual cancer; (2) enabling non-invasive diagnosis or monitoring by sequencing of tumour-released circulating cells or cell-free DNA; (3) identifying cancer-specific, protein-altering mutations that may serve as neoantigens for 'personal vaccines'. Although, the success stories in each of these areas are still few and far between, relative to the overall burden of cancer, progress is clearly being made.

### Sequencers as a molecule counting device

While 'expressed sequenced tags'[107] were considered a shortcut to gene discovery as early as 1983[108], it was SAGE (serial analysis of gene expression; 1995) that introduced the idea of sequencing to 'digitally quantify' gene expression[109]. SAGE concatenated cDNA-derived tags for Sanger sequencing, with tags that are just long enough to map to a gene. As early as 2000, Brenner and Lynx Therapeutics demonstrated 'massively parallel signature sequencing' of cDNA tags, an important forerunner of NGS[47]. However, this concept was not widely adopted until the development of RNA sequencing (RNA-seq) by five groups in 2008. RNA-seq uses NGS to quantify and characterize the transcriptome by shotgun sequencing of either full-length or 3′ ends of cDNA[110–114]. RNA-seq has marked advantages over microarrays, the most notable of which is that transcript counts lead to straightforward statistics relative to analogue, hybridization-based signals, facilitated by new software packages, such as TopHat and Cufflinks[115,116].

Also around 2008, small laboratories that were early adopters of NGS developed 'digital quantification' methods for transcription-factor binding[117], chromatin accessibility[118] and translation[119]. In the following decade, hundreds of protocols were developed that facilitate the use of DNA sequencing as a 'molecule counter' for the characterization of a remarkable range of biochemical or molecular phenomena, including transcription, translation, DNA replication, the secondary structure of RNA, chromosome conformation, nucleic-acid modifications, post-translational modifications, nucleic acid–protein interactions and protein–protein interactions. These are catalogued in other reviews and resources (ref. 120 and http://enseqlopedia.com/).

The use of sequencers as molecule-counting devices was immediately immensely popular, and probably had a larger role than assembly or resequencing in driving the widespread adoption of NGS in biomedical research. DNA sequencers are increasingly to the molecular biologist what a microscope is to the cellular biologist—a basic and essential tool for making measurements. In the long run, this may prove to be greatest impact of DNA sequencing.

### Metagenome sequencing

Shotgun sequencing of complex communities of microorganisms[121–123], for example, metagenome sequencing of environmental or human microbiomes, has emerged as a field of its own, bringing with it unique challenges with respect to assembly, resequencing and counting. Other reviews have recently covered this topic[124,125].

## The future of DNA sequencing

In the long view of scientific history, DNA sequencing remains a young technology. Here, we briefly consider its future in a few existing or emerging areas.

### Genome diversity

A 100% complete genome, that is, the telomere-to-telomere sequence for each chromosome with no gaps or ambiguities, has been achieved for possibly only one eukaryote so far[126]. As sequencing technologies continue to evolve, we are optimistic that we will resolve challenging regions of additional genomes (for example, centromeres). There are

millions of living species on earth (and far more extinct species), each with a genome waiting to be sequenced, as well as countless microbiomes and metagenomes. A comprehensive view of genomic diversity may prove useful in surprising ways, for example, for protein structure determination[127].

### Population-scale resequencing

We are approaching the milestone where approximately 0.1% of living humans will have had their genomes resequenced to some degree, while resequencing of the genomes of our ancestors and other hominins is reshaping our understanding of human history[88]. The number of *de novo* point mutations occurring in recent generations vastly exceeds the number of nucleotides in the human genome. Eventually, aggregating tens of millions of genomes may enable a nucleotide-level footprint of the human genome (that is, observing all heterozygous variants compatible with life). DNA sequencing also is increasingly useful for forensics, without necessarily requiring a sample from the identified individual[128].

### Developmental biology

We each develop from a single cell into a highly organized mass of trillions of cells. However, our understanding of development remains coarse. Recent technologies enable scalable, sequencing-based profiling of single cells. Although popular approaches are *ex vivo* (for example, single-cell RNA-seq), a more radical approach is to perform RNA or protein sequencing *in situ*, thereby retaining the spatial context[129,130]. Other emerging strategies use *in vivo* genome editing to track cell-lineage relationships[131] or transport barcodes to catalogue neuronal connections[132]. Editing of DNA can potentially be used to record biological events more generally, for example, to monitor gene expression[133] or calcium[134].

### Real-time, portable sensors

Nanopore sequencers currently have a mass of 70 g and yield data within 30 min of sample application. One can imagine disseminated networks of nanopore sequencers enabling 'universal monitoring' of nucleic acids, in environmental settings and in everyday human life, for example, fine-grained tracking of our air, food and body, potentially streaming data from millions of devices and integrating with GPS and audio-visual data.

### Unconventional uses

DNA-sequencing technologies will probably prove useful in additional, surprising ways. For example, NGS has recently been used to recover large amounts of data encoded in synthetic DNA[135]. Nanopores may find uses beyond sequencing, for example, for monitoring analyte binding[136], chemical nanomachines[137] or protein folding/unfolding[138].

## DNA sequencing as the new microscope

It has been about 400 years since the invention of light microscopy, a technology which continues to be used and to evolve. By comparison, it has been only 40 years since the invention of DNA sequencing; the technologies for which are likely to also continue to develop in the coming decades and centuries. On the basis of how quickly it has transformed biomedical research, and is beginning to transform clinical medicine, we predict that DNA sequencing will have a longevity and impact on par with or exceeding that of the microscope.

**Received 13 July; accepted 21 September 2017.**
**Published online 11 October 2017.**

1. Sanger, F. Sequences, sequences, and sequences. *Annu. Rev. Biochem.* **57,** 1–28 (1988).
2. Sanger, F. Nobel lecture: the chemistry of insulin. https://www.nobelprize.org/nobel_prizes/chemistry/laureates/1958/sanger-lecture.html (2017).
3. Edman, P. Method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.* **4,** 283–293 (1950).
4. Holley, R. W. *et al.* Structure of a ribonucleic acid. *Science* **147,** 1462–1465 (1965).
5. Sanger, F., Brownlee, G. G. & Barrell, B. G. A two-dimensional fractionation procedure for radioactive nucleotides. *J. Mol. Biol.* **13,** 373–398 (1965).
6. Wu, R. & Kaiser, A. D. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.* **35,** 523–537 (1968).
7. Gilbert, W. & Maxam, A. The nucleotide sequence of the lac operator. *Proc. Natl Acad. Sci. USA* **70,** 3581–3584 (1973).
8. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74,** 5463–5467 (1977).
**Refs 8, 9: The seminal papers by Sanger, Nicklen & Coulson and Maxam & Gilbert describing the first widely adopted methods for DNA sequencing.**
9. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA* **74,** 560–564 (1977).
10. Maniatis, T., Jeffrey, A. & van deSande, H. Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis. *Biochemistry* **14,** 3787–3794 (1975).
11. Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* **6,** 2601–2610 (1979).
12. Messing, J., Crea, R. & Seeburg, P. H. A system for shotgun DNA sequencing. *Nucleic Acids Res.* **9,** 309–321 (1981).
13. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.* **162,** 729–773 (1982).
14. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321,** 674–679 (1986).
15. Connell, C. *et al.* Automated DNA sequence analysis. *Biotechniques* **5,** 342–348 (1987).
16. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
17. Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238,** 336–341 (1987).
18. Tabor, S. & Richardson, C. C. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proc. Natl Acad. Sci. USA* **84,** 4767–4771 (1987).
19. Craxton, M. Linear amplification sequencing, a powerful method for sequencing DNA. *Methods* **3,** 20–26 (1991).
20. DeAngelis, M. M., Wang, D. G. & Hawkins, T. L. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res.* **23,** 4742–4743 (1995).
21. Zhang, J. *et al.* Use of non-cross-linked polyacrylamide for four-color DNA sequencing by capillary electrophoresis separation of fragments up to 640 bases in length in two hours. *Anal. Chem.* **67,** 4589–4593 (1995).
22. Green, P. phred, phrap, consed. http://www.phrap.org/phredphrapconsed.html (2017).
**phred introduced quantitative, reliable metrics for base quality, substituting human judgement with computers, a process that occurred repeatedly over the course of the HGP.**
23. Edwards, A. *et al.* Automated DNA sequencing of the human HPRT locus. *Genomics* **6,** 593–608 (1990).
24. Sutton, G. G., White, O., Adams, M. D. & Kerlavage, A. R. TIGR assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol.* **1,** 9–19 (1995).
25. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287,** 2196–2204 (2000).
**The Celera assembler introduced an overlap–layout–consensus approach to deal with the problems posed by repeats and the millions of reads needed to produce a reliable assembly.**
26. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269,** 496–512 (1995).
27. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546–567 (1996).
28. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).
29. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
**Refs 29–31: The HGP and Celera produced draft sequences of the human genome with the HGP later publishing a more complete, relatively error-free reference.**
30. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
31. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
32. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
33. Balasubramanian, S., Klenerman, D. & Barnes, C. Arrayed polynucleotides and their use in genome analysis. Patent US20030022207 (2003).
34. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proc. Natl Acad. Sci. USA* **100,** 3960–3964 (2003).
35. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320,** 106–109 (2008).
36. Adams, C. P. & Kron, S. J. Method for performing amplification of nucleic acid with two primers bound to a single solid support. Patent US5641658 (1997).
37. Chetverina, H. V. & Chetverin, A. B. Cloning of RNA molecules *in vitro*. *Nucleic Acids Res.* **21,** 2349–2353 (1993).
38. Mitra, R. D. & Church, G. M. *In situ* localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res.* **27,** e34–e39 (1999).

39. Adessi, C. *et al.* Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.* **28,** e87 (2000).
40. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).
**Refs 40, 41: These papers described the first integrated systems for next-generation DNA sequencing.**
41. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309,** 1728–1732 (2005).
42. Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl Acad. Sci. USA* **100,** 8817–8822 (2003).
43. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78–81 (2010).
44. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242,** 84–89 (1996).
45. Toumazou, C. & Purushothaman, S. Sensing apparatus and method. Patent US7686929 (2004).
46. Rothberg, J. M., Hinz, W., Johnson, K. L. & Bustillo, J. Apparatus for measuring analytes using large scale FET arrays. Patent EP2639579 (2016).
47. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18,** 630–634 (2000).
48. McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19,** 1527–1541 (2009).
49. Mitra, R. D., Shendure, J., Olejnik, J., Edyta-Krzymanska-Olejnik, & Church, G. M. Fluorescent *in situ* sequencing on polymerase colonies. *Anal. Biochem.* **320,** 55–65 (2003).
50. Ost, T. B. *et al.* Improved polymerases. Patent WO2006120433 (2006).
51. Ruparel, H. *et al.* Design and synthesis of a 3′-*O*-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. *Proc. Natl Acad. Sci. USA* **102,** 5932–5937 (2005).
52. Seo, T. S. *et al.* Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc. Natl Acad. Sci. USA* **102,** 5926–5931 (2005).
53. Barnes, C., Balasubramanian, S., Liu, X., Swerdlow, H. & Milton, J. Labelled nucleotides. Patent US7057026 (2006).
54. Smith, T. J. Cloned single molecule sequencing with reversible terminator chemistry. Genome Sequencing and Analysis Conference (2015).
55. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
**Advances in sequencing-by-synthesis culminated in the Solexa, later Illumina, platform, which quickly became, and remains today, the most widely used sequencing instrument.**
56. Wetterstrand, K. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). http://www.genome.gov/sequencingcostsdata (2017).
57. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299,** 682–686 (2003).
**One of earliest real time observations of DNA synthesis in single molecules, using fluorescently labelled nucleotides and a DNA polymerase anchored in zero-mode waveguides, which with further development led to the PacBio platform.**
58. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).
59. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34,** 518–524 (2016).
60. Bayley, H. Nanopore sequencing: from imagination to reality. *Clin. Chem.* **61,** 25–31 (2015).
61. Church, G., Deamer, D. W., Branton, D., Baldarelli, R. & Kasianowicz, J. Characterization of individual polymer molecules based on monomer-interface interactions. Patent US5795782 (1998).
**The concept of ssDNA modulating an electronic signal while moving through a membrane pore led eventually to practical nanopore sequencing.**
62. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26,** 1146–1153 (2008).
63. Laszlo, A. H. *et al.* Decoding long nanopore sequencing reads of natural DNA. *Nat. Biotechnol.* **32,** 829–833 (2014).
64. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. Preprint at https://www.biorxiv.org/content/early/2017/04/20/128835 (2017).
65. Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7,** 461–465 (2010).
66. Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R. & Akeson, M. Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing. Preprint at https://www.biorxiv.org/content/early/2017/04/29/132274 (2017).
67. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. Preprint at https://www.biorxiv.org/content/early/2016/08/12/068809 (2016).
68. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat. Biotechnol.* **31,** 247–250 (2013).
69. Zhao, Y. *et al.* Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nat. Nanotechnol.* **9,** 466–473 (2014).
70. Wilson, J., Sloman, L., He, Z. & Aksimentiev, A. Graphene nanopores for protein sequencing. *Adv. Funct. Mater.* **26,** 4830–4838 (2016).
71. Di Ventra, M. & Taniguchi, M. Decoding DNA, RNA and peptides with quantum tunnelling. *Nat. Nanotechnol.* **11,** 117–126 (2016).
72. Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265,** 687–695 (1977).
73. Schneider, V. A. *et al.* Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27,** 849–864 (2017).
74. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl Acad. Sci. USA* **98,** 9748–9753 (2001).
75. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
76. Adey, A. *et al. In vitro*, long-range sequence information for *de novo* genome assembly via transposase contiguity. *Genome Res.* **24,** 2041–2049 (2014).
77. Mostovoy, Y. *et al.* A hybrid approach for *de novo* human genome sequence assembly and phasing. *Nat. Methods* **13,** 587–590 (2016).
78. Burton, J. N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31,** 1119–1125 (2013).
79. Kaplan, N. & Dekker, J. High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nat. Biotechnol.* **31,** 1143–1147 (2013).
80. Bickhart, D. M. *et al.* Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. *Nat. Genet.* **49,** 643–650 (2017).
81. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27,** 722–736 (2017).
82. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
83. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
84. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
85. McKenna, A. *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
86. Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517,** 608–611 (2015).
87. Snyder, M. W., Adey, A., Kitzman, J. O. & Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet.* **16,** 344–358 (2015).
88. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
89. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
90. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
91. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60–65 (2008).
92. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).
93. Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4,** 903–905 (2007).
94. Okou, D. T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* **4,** 907–909 (2007).
95. Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4,** 931–936 (2007).
96. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* **39,** 1522–1527 (2007).
97. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–276 (2009).
**Refs 97, 103, 106: Targeting all coding sequences or the exome, by PCR and later by exome capture, facilitated the direct discovery of cancer driver genes and Mendelian disease genes.**
98. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
99. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
100. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493,** 216–220 (2013).
101. Chiu, R. W. *et al.* Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc. Natl Acad. Sci. USA* **105,** 20458–20463 (2008).
102. Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L. & Quake, S. R. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc. Natl Acad. Sci. USA* **105,** 16266–16271 (2008).
103. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl Acad. Sci. USA* **106,** 19096–19101 (2009).
104. Vissers, L. E. L. M., Gilissen, C. & Veltman, J. A. Genetic studies in intellectual disability and related disorders. *Nat. Rev. Genet.* **17,** 9–18 (2016).
105. Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *J. Am. Med. Assoc.* **312,** 1870–1879 (2014).
106. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–1113 (2007).
107. Adams, M. D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252,** 1651–1656 (1991).

108. Putney, S. D., Herlihy, W. C. & Schimmel, P. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302,** 718–721 (1983).
109. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270,** 484–487 (1995).
**The SAGE method captures 3′ tags from mRNAs, therefore introducing the idea of using a DNA sequencer to count molecules, an idea that has exploded with the later introduction of RNA-seq, chromatin immunoprecipitation followed by sequencing (ChIP–seq) and so on.**
110. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5,** 613–619 (2008).
111. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133,** 523–536 (2008).
112. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* **5,** 621–628 (2008).
113. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349 (2008).
114. Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453,** 1239–1243 (2008).
115. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25,** 1105–1111 (2009).
116. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
117. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316,** 1497–1502 (2007).
118. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132,** 311–322 (2008).
119. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).
120. Shendure, J. & Lieberman Aiden, E. The expanding scope of DNA sequencing. *Nat. Biotechnol.* **30,** 1084–1094 (2012).
121. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
122. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428,** 37–43 (2004).
123. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).
124. Blaser, M., Bork, P., Fraser, C., Knight, R. & Wang, J. The microbiome explored: recent insights and future challenges. *Nat. Rev. Microbiol.* **11,** 213–217 (2013).
125. Shokralla, S., Spall, J. L., Gibson, J. F. & Hajibabaei, M. Next-generation sequencing technologies for environmental DNA research. *Mol. Ecol.* **21,** 1794–1805 (2012).
126. Nozaki, H. *et al.* A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga *Cyanidioschyzon merolae*. *BMC Biol.* **5,** 28 (2007).
127. Ovchinnikov, S. *et al.* Protein structure determination using metagenome sequence data. *Science* **355,** 294–298 (2017).
128. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science* **339,** 321–324 (2013).
129. Larsson, C. *et al. In situ* genotyping individual DNA molecules by target-primed rolling-circle amplification of padlock probes. *Nat. Methods* **1,** 227–232 (2004).
130. Lee, J. H. *et al.* Highly multiplexed subcellular RNA sequencing *in situ*. *Science* **343,** 1360–1363 (2014).
131. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353,** aaf7907 (2016).
132. Peikon, I. D. *et al.* Using high-throughput barcode sequencing to efficiently map connectomes. *Nucleic Acids Res.* **45,** e115 (2017).
133. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* **353,** aaf1175 (2016).
134. Zamft, B. M. *et al.* Measuring cation dependent DNA polymerase fidelity landscapes by deep sequencing. *PLoS ONE* **7,** e43876 (2012).
135. Organick, L. *et al.* Scaling up DNA data storage and random access retrieval. Preprint at https://www.biorxiv.org/content/early/2017/03/07/114553 (2017).
136. Harrington, L., Alexander, L. T., Knapp, S. & Bayley, H. Pim kinase inhibitors evaluated with a single-molecule engineered nanopore sensor. *Angew. Chem. Int. Edn Engl.* **54,** 8154–8159 (2015).
137. Pulcu, G. S., Mikhailova, E., Choi, L. S. & Bayley, H. Continuous observation of the stochastic motion of an individual small-molecule walker. *Nat. Nanotechnol.* **10,** 76–83 (2015).
138. Rodriguez-Larrea, D. & Bayley, H. Protein co-translocational unfolding depends on the direction of pulling. *Nat. Commun.* **5,** 4841 (2014).
139. Hyman, E. D. A new method of sequencing DNA. *Anal. Biochem.* **174,** 423–436 (1988).
140. Lee, L. G. *et al.* DNA sequencing with dye-labeled terminators and T7 DNA polymerase: effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments. *Nucleic Acids Res.* **20,** 2471–2483 (1992).
141. Huang, S. *et al.* Identifying single bases in a DNA oligomer with electron tunnelling. *Nat. Nanotechnol.* **5,** 868–873 (2010).
142. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475,** 348–352 (2011).
143. Manrao, E. A. *et al.* Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat. Biotechnol.* **30,** 349–353 (2012).
144. Cherf, G. M. *et al.* Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nat. Biotechnol.* **30,** 344–348 (2012).
145. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
146. *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
147. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
148. Gibbs, R. A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
149. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
150. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436,** 793–800 (2005).
151. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326,** 1112–1115 (2009).
152. Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500,** 207–211 (2013).
153. Landry, J. J. *et al.* The genomic and transcriptomic landscape of a HeLa cell line. *G3 (Bethesda)* **3,** 1213–1224 (2013).
154. Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496,** 498–503 (2013).
155. Session, A. M. *et al.* Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* **538,** 336–343 (2016).
156. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).
157. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
158. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
159. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
160. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30,** 38–41 (2002).
161. Giardine, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* **15,** 1451–1455 (2005).
162. Butler, J. *et al.* ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.* **18,** 810–820 (2008).
163. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6,** 677–681 (2009).
164. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).
165. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–272 (2010).
166. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).
167. Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10,** 563–569 (2013).
168. Wang, D. G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280,** 1077–1082 (1998).
169. Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463,** 311–317 (2010).
170. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29,** 59–63 (2011).
171. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.* **29,** 51–57 (2011).
172. Seo, J. S. *et al. De novo* assembly and phasing of a Korean human genome. *Nature* **538,** 243–247 (2016).
**BLAST and GenBank (GenBank and WGS Statistics; https://www.ncbi.nlm.nih.gov/genbank/statistics/) were essential tools for sharing and searching sequencing data, vastly amplifying the value of each sequence to the field.**


**Acknowledgements** This is a large topic to cover in a single review. We apologize to colleagues whose work we were unable to discuss or failed to cite owing to space constraints. We thank L. Starita, C. Trapnell and A. McKenna for suggestions, and T. Tolpa and M. Gillies for extensive assistance with preparing the manuscript.

**Author Contributions** All authors contributed to the writing of this review.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.S. (shendure@uw.edu).

**Reviewer Information** *Nature* thanks M. Gerstein, S. L. Salzberg and the other anonymous reviewer(s) for their contribution to the peer review of this work.

CORRECTION

https://doi.org/10.1038/s41586-019-1120-8

# Publisher Correction: DNA sequencing at 40: past, present and future

Jay Shendure, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss & Robert H. Waterston

Correction to: *Nature* https://doi.org/10.1038/nature24286, published online 11 October 2017.

In this Review, the year of publication of reference 54 should be 2005, not 2015. In Box 2, "1982: GenBank (https://www.ncbi.nlm.nih.gov/genbank/statistics/)" should read "1982: Genbank/ENA/DDBJ" and "2007: NCBI Short Read Archive" should read "2007: NCBI and ENA Short Read Archives"; this is because the launches of these American, European and Japanese databases were coordinated. These errors have not been corrected.