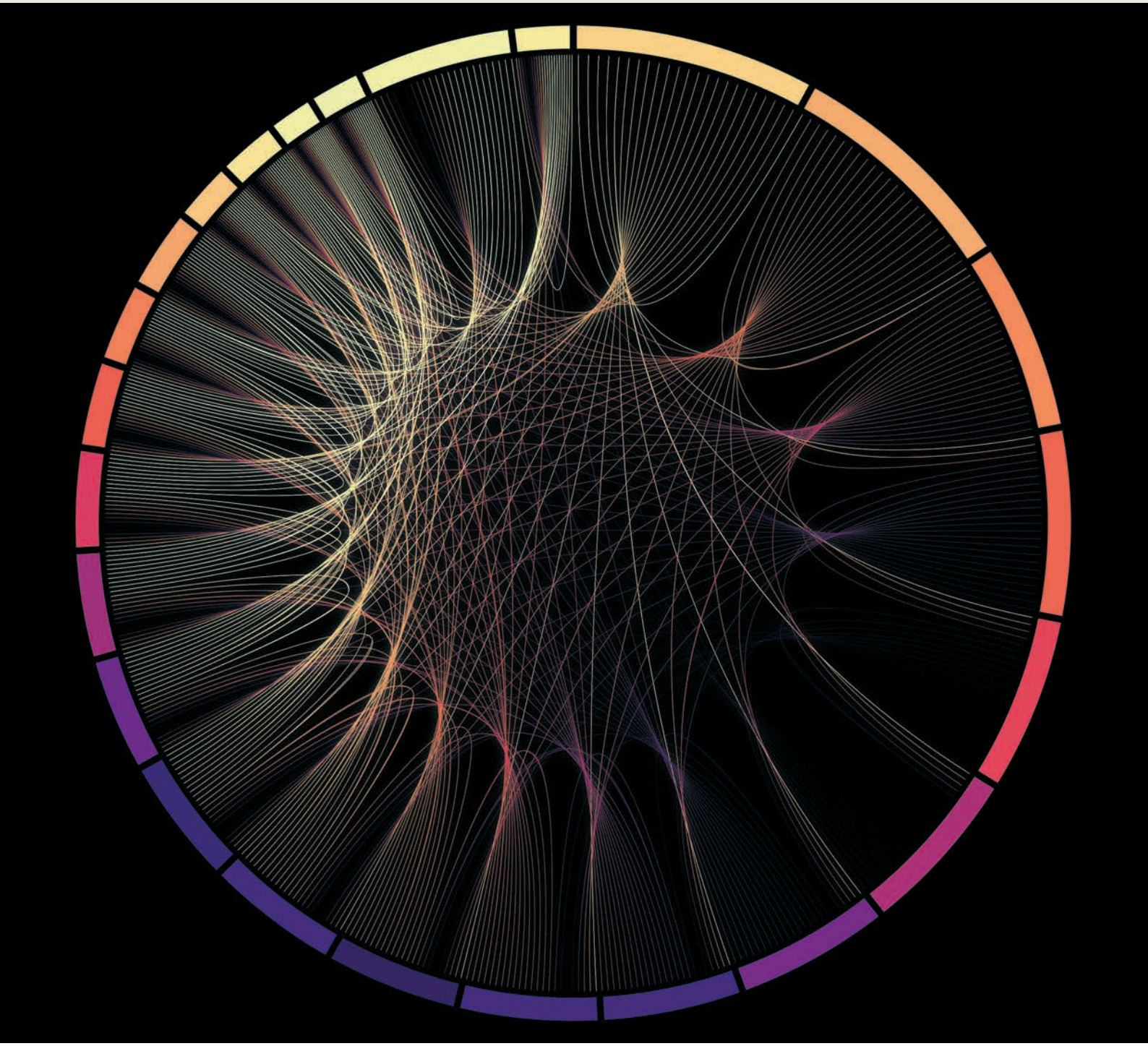
www.nature.com/collections/genomic-sequencing-milestones February 2021

nature milestones

Genomic sequencing



Produced by:
Nature, Nature Genetics and Nature Reviews Genetics

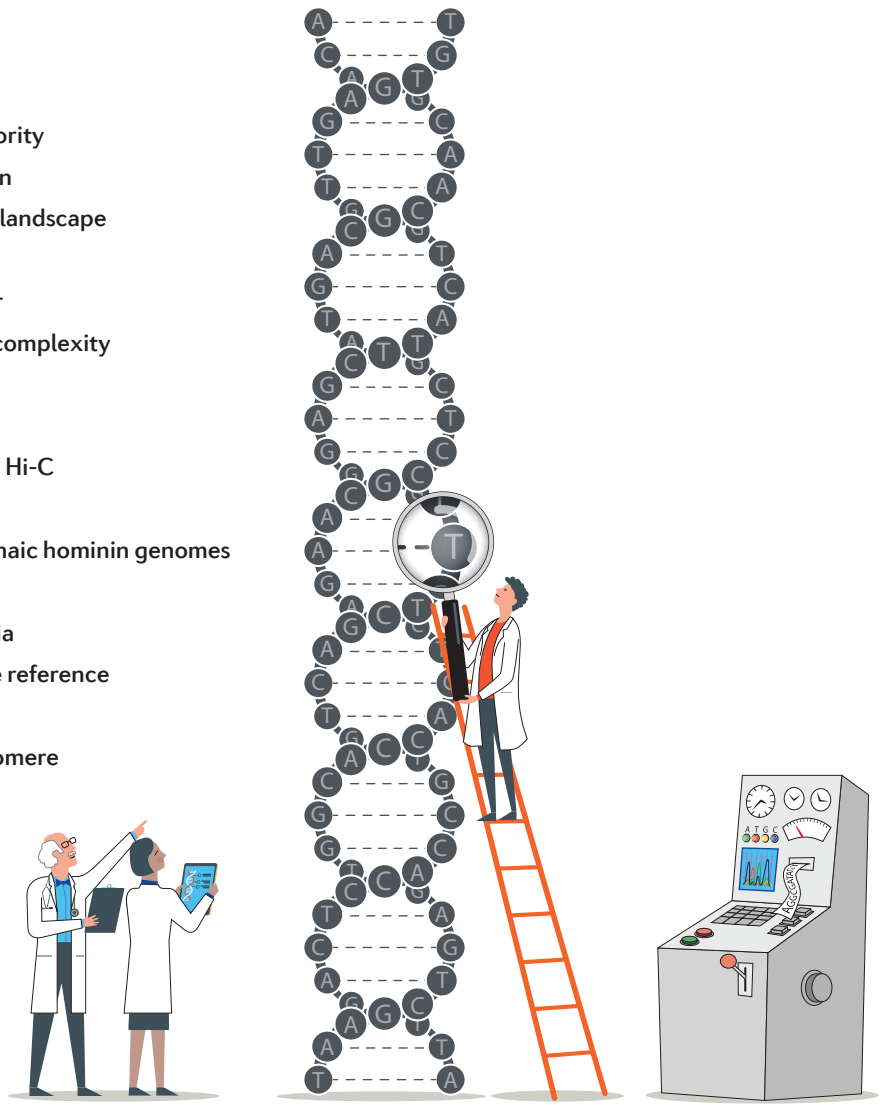
With support from:
illumina®

nature milestones

Genomic sequencing

MILESTONES

S3 Foreword
S4 Timeline
S5 The Human Genome Project
S6 Sequencing the unculturable majority
S7 Sequencing — the next generation
S8 ChIP–seq captures the chromatin landscape
S9 The dawn of personal genomes
S10 A sequencing revolution in cancer
S11 Transcriptomes — a new layer of complexity
S12 Long reads become a reality
S13 Exploring whole exomes
S14 Probing nuclear architecture with Hi-C
S15 Sequencing one cell at a time
S16 Waking the dead: sequencing archaic hominin genomes
S17 Cataloguing a public genome
S18 Our most elemental encyclopaedia
S19 Pan-genomes: moving beyond the reference
S20 Genomes go platinum
S21 Filling in the gaps telomere to telomere



Credit: S.Fenwick / Springer Nature Limited

CITING THE MILESTONES

Nature Milestones in Genomic Sequencing includes Milestone articles written by our editors and an online Collection of previously published material. To cite the full project, please use *Nature Milestones: Genomic Sequencing* https://www.nature.com/collections/genomic-sequencing-milestones (2021). Should you wish to cite any of the individual Milestones, please list Author, A. Title. *Nature Milestones: Genomic Sequencing* <Article URL> (2021). For example, Milestone 1 is Barranco, C. The Human Genome Project. *Nature Milestones: Genomic Sequencing* https://www.nature.com/articles/d42859-020-00101-9 (2021). To cite articles from the Collection, please use the original citation, which can be found online.

VISIT THE SUPPLEMENT ONLINE

The *Nature Milestones in Genomic Sequencing* supplement can be found at www.nature.com/collections/genomic-sequencing-milestones

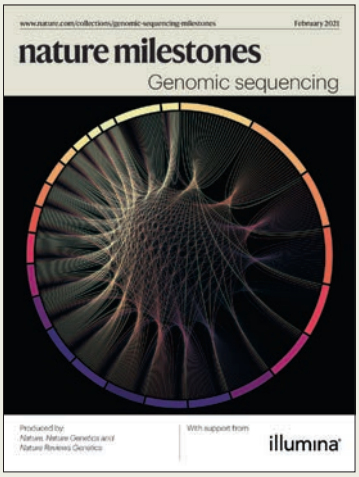
CONTRIBUTING JOURNALS

Communications Biology, *Genome Biology*, *Nature*, *Nature Communications*, *Nature Genetics*, *Nature Protocols*, *Nature Reviews Gastroenterology & Hepatology*, *Nature Reviews Genetics*, *Nature Reviews Methods Primers*

SUBSCRIPTIONS AND CUSTOMER SERVICES

Springer Nature, Subscriptions, Cromwell Place, Hampshire International Business Park, Lime Tree Way, Basingstoke, Hampshire RG24 8YJ, UK
Tel: +44 (0) 1256 329242
subscriptions@springernature.com

CUSTOMER SERVICES: www.nature.com/help



► **COVER:** Design by Chris Ryan.

EDITORIAL OFFICES
LONDON
Springer Nature
The Campus, 4 Crinan Street, London N1 9XW, UK
Tel: +44 (0)20 7833 4000
COORDINATING EDITORS: Linda Koch, Catherine Potenski, Michelle Trenkmann
COPY EDITOR: Rebecca Hill
PRODUCTION: Simon Fenwick, Susan Gray, Nick Bruni
DESIGN & WEB DEVELOPMENT: Chris Ryan
MARKETING: Helen Burgess
PUBLISHER: Rebecca Jones
VP, PUBLISHING: Richard Hughes
EDITOR-IN-CHIEF, NATURE PUBLICATIONS: Magdalena Skipper
SPONSORSHIP: David Bagshaw, Heather Penn, Claudia Danci

PRODUCED WITH SUPPORT FROM:
Illumina

SPRINGER NATURE

As anyone who has ever assembled a piece of furniture can attest to, instruction manuals are only useful if you can read them. Similarly, knowing what makes each of us unique by sourcing the information contained in our genomes requires the ability to read the order of As, Cs, Gs and Ts that constitute our DNA. This feat is made possible by DNA sequencing technologies. In this *Nature Milestones in Genomic Sequencing*, we chart the history of these extraordinary technologies and their continuously expanding applications over the past two decades.

Since its emergence, genomic sequencing has become one of the most influential tools in biomedical research. The potential of sequencing technologies was quickly recognized, with half of the 1980 Nobel Prize in Chemistry being awarded to Walter Gilbert and Frederick Sanger "for their contributions concerning the determination of base sequences in nucleic acids", a mere 3 years after the development of Maxam–Gilbert sequencing and Sanger sequencing in 1977. In the late 1980s, automated Sanger sequencing machines could sequence approximately 1,000 bases per day. With continuous methodological advances and computational developments occurring in parallel, the 1990s saw DNA sequencing applied to large bacterial genomes and the first unicellular and multicellular eukaryotic genomes.

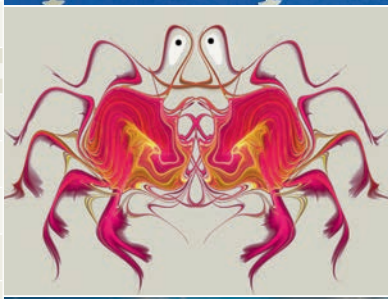
Sanger sequencing dominated the research landscape until the early twenty-first century and led to exceptional achievements, including the completion of a high-quality, reference sequence of the human genome under the auspices of the Human Genome Project (HGP), which is where we have chosen to start our milestones (MILESTONE 1). The field took off in earnest with the development and commercialization of high-throughput, massively parallel or next-generation sequencing, which democratized sequencing by offering individual laboratories access to the technology. In this *Nature Milestones in Genomic Sequencing* timeline, we want to highlight methodological and computational advances and projects that have propelled the field forwards, culminating in an entire, virtually gap-free human chromosome, assembled telomere to telomere (MILESTONE 17). We also shine a light on research areas revolutionized by the application of sequencing technologies over the past 20 years.

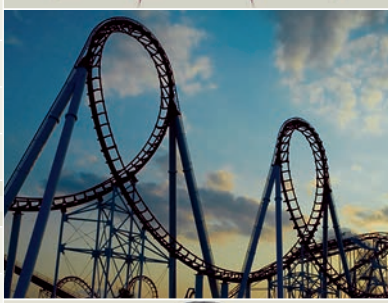
Science is a team effort. We appreciate that each milestone that we have selected stands on a mountain of preceding work and that the histories of technological advances, applications and discoveries are interwoven. We apologize in advance for any missed contributions and extend our gratitude to all the researchers who have advised on this project or agreed to be interviewed. Finally, we would like to acknowledge financial support from Illumina. As always, responsibility for the editorial content remains with Springer Nature.

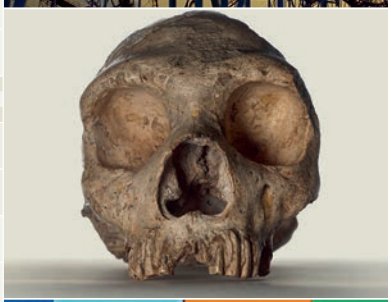
Linda Koch, Chief Editor, *Nature Reviews Genetics*
Catherine Potenski, Chief Editor, *Nature Genetics*
Michelle Trenkmann, Senior Editor, *Nature*

MILESTONES IN GENOMIC SEQUENCING

2000	Release of the first full genomic sequence for *Drosophila melanogaster*
	First plant genome, *Arabidopsis thaliana*, sequenced
2001	Publication of the first draft sequence and analysis of the human genome (MILESTONE 1)
2002	Draft sequence of the *Mus musculus* genome released
	Genome browsers as a new form of access to genome sequences and their annotations
2004	Metagenomics comes of age (MILESTONE 2)
2005	The International HapMap Project releases its initial phase I map
	Publication of the first draft genome sequence of a non-human primate (*Pan troglodytes*)
	Sequencing of the first crop genome (*Oryza sativa*)
	Development of the first next-generation sequencing technologies (MILESTONE 3)
2007	ChIP–seq maps DNA–protein binding (MILESTONE 4)
2008	DNase-seq identifies open chromatin regions
	A new sequencing approach is applied to human genomes (MILESTONE 5)
	Sequencing of a cancer genome reveals disease-associated mutations (MILESTONE 6)
	Genome assembly tools developed for reconstructing whole genomes from short reads
	Transcriptomics adds a new layer of complexity (MILESTONE 7)
	Sequencing of cell-free fetal DNA for non-invasive prenatal genetic testing
2009	New computational tools to meet the needs of expanding genomics applications
	Long-read sequencing technologies begin to emerge (MILESTONE 8)
	First application of whole-exome sequencing to monogenic disease (MILESTONE 9)
	Whole-genome bisulfite sequencing maps DNA methylation patterns genome-wide
	Ribosome profiling allows direct measurement of in vivo protein synthesis
	Hi-C reveals the 3D architecture of the genome (MILESTONE 10)
	Single-cell sequencing adds a new perspective on cellular heterogeneity (MILESTONE 11)
2010	Reconstruction of a large, complex genome from short-read sequences
	Emerging technologies provide new approaches to sequence DNA
	Sequencing of ancient DNA (MILESTONE 12)
2012	Large-scale sequencing studies catalogue human genetic variation (MILESTONE 13)
	ENCODE characterizes the functional genome (MILESTONE 14)
2013	Release of the zebrafish genome sequence
	Development of ATAC-seq for multimodal profiling of the epigenome
2014	Pan-genomes capture genetic variation from many representatives of a species (MILESTONE 15)
2015	A roadmap of the human epigenome
2016	A microfluidics-based sequencing approach generates linked-reads
2017	Combining multiple sequencing technologies yields a 'platinum' genome (MILESTONE 16)
2020	First gapless, telomere-to-telomere assembly of a human chromosome (MILESTONE 17)







MILESTONES

Credit: Rawpixel Ltd / Alamy Stock Photo

MILESTONE 1

The Human Genome Project

The signature aim of the Human Genome Project (HGP), which was launched in 1990, was to sequence the 3 billion bases of the human genome. Additional goals included the generation of physical and genetic maps of the human genome, as well as mapping and sequencing of key model organisms used in biomedical research.

In 1998, the HGP formally implemented the Bermuda Principles, specifically the following: automatic release of sequence assemblies >1 kb, preferably within 24 h; immediate publication of finished annotated sequences; and making the entire sequence freely available in the public domain for both research and development in order to maximize its benefits to society. In exchange for the immediate online release of HGP-funded sequence data, research groups from the USA, UK, Japan, France, Germany and China conducting the sequencing retained the right to be the first to describe their complete datasets and to analyse their findings in peer-reviewed publications.

By insisting on the Bermuda Principles, the HGP sought to undermine the efforts of parties aiming to patent or commercialize human genomic sequences, which could restrict subsequent research efforts. In 1998, it was announced that a new company, later renamed Celera Genomics, would 'race' the publicly funded HGP to complete the sequencing of the human genome. Celera Genomics also intended to sell subscriptions to its database, release data quarterly, and obtain patents on genes and related technologies.

This new presence threatened the survival of the HGP (which by early 1998 had sequenced only a small fraction of the human genome), but after US President Clinton and UK Prime Minister Tony Blair jointly declared on 14 March 2000 that the human genome sequence "should be made freely available to scientists everywhere", the HGP and Celera Genomics brokered a deal leading to the simultaneous publication in February 2001 of two articles (by Venter et al. in *Science* and the International Human Genome Sequencing Consortium in *Nature*) describing the draft human genome sequence. The sequence included 26,588 protein-coding transcripts for which there was strong corroborating evidence and an additional ~12,000 computationally derived genes with mouse homologues or other weak evidence.

The HGP used a hierarchical shotgun sequencing approach, in which the genome was broken into ~150-kb segments and cloned into bacterial artificial chromosomes, before being matched to a genome-wide physical map comprising >96% of the euchromatic part of the human genome (~94% of the entire human genome). Selected bacterial artificial chromosomes were sequenced and finally reassembled to generate the draft sequence. By contrast, Celera Genomics used both HGP and their own private data in their whole-genome shotgun sequencing approach, which fragmented the genome into ~500-bp segments and subjected them to pairwise end sequencing (in which a given fragment is sequenced from both ends to produce a 'mate pair') to reconstruct the original sequence.

In 2003, a group of ~40 professionals working in genomics publicly declared their support for the free and unrestricted use of genome-sequencing data by the scientific community before formal publication. This declaration, known as the Fort Lauderdale Agreement, enshrined the collective responsibility of funding agencies, resource producers and users to maintain and expand a communal trove of genomic data. These principles were later implemented as policy by several funding agencies, notably the US National Institutes of Health (NIH), which today still mandates rapid data-sharing in its grant requirements. Collectively, these initiatives can be considered the forerunners of open-access publishing in biomedicine.

> "Today, the HGP remains notable for... revenue and paradigm shifts [it] generated"

In 2004, the work of the HGP culminated in publication of a highly accurate (~1 error per 100,000 bases) human genome sequence that included ~99% of the euchromatic genome. The current version of the human reference genome, GRCh38.p13, comprises 3.27 billion nucleotides and 19,116 nuclear protein-coding genes.

Today, the HGP remains notable for an estimated US$800 billion of revenue and paradigm shifts generated by this publicly funded 'big science' project. Offering a first view into the entire human genome, the HGP acted as a gateway to an era of high-throughput digital biology, ushering in rapid technological and computational developments and team-oriented research, the fruits of which continue to be felt across the clinical and life sciences.

Caroline Barranco,
Nature Reviews Cross-Journal Team

ORIGINAL ARTICLES International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001) | Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001)
FURTHER READING International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004) | Maxson Jones, K., Ankeny, R. A. & Cook-Deegan, R. The Bermuda Triangle: the pragmatics, policies, and principles for data sharing in the history of the human genome project. *J. Hist. Biol.* **51**, 693–805 (2018)

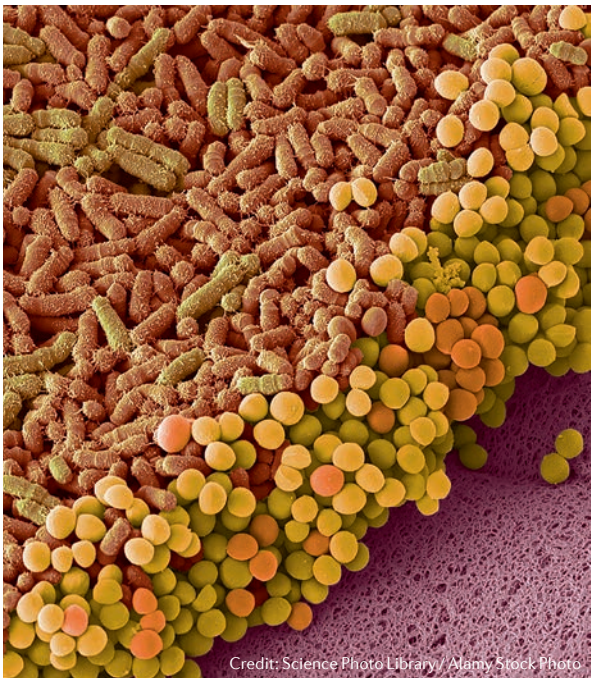
MILESTONE 2

Sequencing the unculturable majority

Before the twenty-first century, the study of microorganisms typically required the ability to cultivate them in isolation. The advent of metagenomics — exemplified by two key studies published in 2004 —provided approaches that enabled unbiased and culture-independent analysis of DNA directly from communities in the environment, revolutionizing the study of complex microbial communities.

Metagenomics was a term first used in 1998 by Handelsman et al. in describing the study of the "collective genomes of soil microflora". Yet, over the 100 years before this point, many microbiologists had noted that the number of microbial cells they could count microscopically was not aligned with the number of colonies that they could grow on plates. This phenomenon, termed the 'great plate count anomaly' by Staley and Konopka in 1985, led to the realization that only ~1% of microbial diversity could be accessed through standard cultivation approaches. This prompted the question: how do we study the remaining 99%?

Pioneering work starting in the late 1970s on the 16S ribosomal RNA (rRNA) gene led to various PCR-based and other molecular tools to enable quantitative and qualitative analyses of microbial identity and diversity from environmental DNA. These advances facilitated early cloning approaches that sought to access new genes from unculturable microorganisms. For example, in the first targeted attempt at metagenomic sequencing led by Edward DeLong in 1996, Stein et al. recovered a 40 kb genome fragment from an uncultivable archaeon from a marine picoplankton assemblage using PCR amplification of the 16S rRNA gene to identify clones containing archaeal DNA. However, PCR-based studies are inherently biased, and the ultimate goal was to access the full genetic potential of microbial genomes from the environment — a daunting task considering that most microbial communities comprise hundreds to thousands of species.



Credit: Science Photo Library / Alamy Stock Photo

The first genome-resolved metagenomics study came in 2004 in a study by Jill Banfield and colleagues. Tyson et al. reported the successful reconstruction of multiple genomes from a DNA sample taken from a biofilm in an acid mine drainage system. The near-complete genomes of a bacterial and an archaeon species, plus the partial genomes from a further three microorganisms, were recovered from environmental DNA using random shotgun sequencing, in which total DNA is fragmented, cloned and sequenced. This study was aided by the low species richness of the sampled biofilm and the low intraspecies genomic variation of the microorganisms, which facilitated assembly of the sequencing reads. Analysis of the recovered genomes from the unculturable iron-oxidizing microorganisms in the acid mine enabled characterization of metabolic pathways and provided insights into survival strategies used by extremophiles.

A second large-scale metagenomics project in 2004 provided the first whole-genome shotgun sequencing study of oceanic microbial populations. Craig Venter and colleagues examined the Sargasso Sea using pooled environmental DNA that was filtered and extracted from seawater samples. From 1.66 million sequencing reads, 265 Mb of sequence data were generated, which led to the identification of >1.2 million previously unknown microbial genes. Data binning and phylogenetic analyses to predict the origin of the sequences led to estimates that the data derived from 1,800 genomic species, including 148 previously unknown 'species' (phylotypes). This study demonstrated that metagenomic sequencing could assess the taxonomical composition of complex microbial communities in an unbiased manner and confirmed previous vast underestimates of microbial biodiversity.

> "Many discoveries... have their origins in our ability to access the genomes of the unculturable majority"

These two studies revealed exciting new opportunities for metagenomics, with the only real limitation being the cost of sequencing. Since 2004, the dramatic drop in the cost of sequencing following the emergence of next-generation sequencing has led to widespread adoption of metagenomics approaches at a scale that was previously thought unachievable.

Mining the huge wealth of metagenomic data has improved our understanding of microbial biodiversity, ecology and evolution and led to valuable gains in biotechnology and medicine. Many discoveries of new antibiotics, anti-cancer drugs and biosynthetic pathways of biomedical, agricultural and industrial importance have their origins in our ability to access the genomes of the unculturable majority.

Iain Dickson, *Nature Reviews Gastroenterology & Hepatology*

ORIGINAL ARTICLES Tyson, G. W. et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004) | Venter, J. C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004)

FURTHER READING Handelsman, J. et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, R245–R249 (1998) | Staley, J. T. & Konopka, A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* **39**, 321–346 (1985) | Stein, J. L. et al. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J. Bacteriol.* **178**, 591–599 (1996)

MILESTONE 3

Sequencing — the next generation

Credit: Zoran Obradovic / Alamy Stock Photo

Following successful efforts to sequence bacterial, animal and human genomes throughout the 1990s using Sanger sequencing, investigators began looking for cheaper and faster approaches. Sanger sequencing was limited in speed and cost owing to its reliance on dideoxynucleotide (ddNTP) 'terminators' and the need to use electrophoresis, limiting sequencing to a single DNA fragment at a time.

In 1998, Pal Nyrén's laboratory described a sequencing-by-synthesis method known as pyrosequencing. This method is based on the measurement of pyrophosphate — which is released following the addition of a nucleotide to a growing DNA strand by DNA polymerase — using a two-enzyme, luciferase-based system. By sequentially adding different deoxynucleotides (dNTPs), the sequence of a template DNA molecule can be inferred by detecting light released at the site of nucleotide incorporation, allowing real-time sequencing and avoiding lengthy electrophoresis.

In parallel with these advances, researchers were investigating how to increase the throughput of preparing templates for sequencing, specifically cloning and amplifying DNA. In 1999, Mitra and Church described a method for amplifying DNA by performing PCR in a polyacrylamide film, producing an array of PCR colonies, or 'polonies', consisting of thousands of clonal amplification products localized with their respective templates. These arrays set the stage for systems capable of interrogating many clonal populations in parallel. A study by Brenner et al. in 2000 described a method for interrogating gene expression by cloning cDNA derived from *Saccharomyces cerevisiae* transcripts onto microbeads, distributing these beads in an array on a flow cell and sequencing the attached cDNA. This study used a sequencing-by-ligation method involving rounds of restriction enzymes and the addition of a library of fluorescent adaptors. In 2003, Dressman et al. improved on this bead system, attaching biotinylated PCR primers to streptavidin-coated beads before dispersing the beads in a microemulsion and carrying out PCR. This emulsion PCR (ePCR) method allowed spatial isolation of each template to a single bubble and enabled clonal amplification of the template on each bead.

In 2005, two ground-breaking studies built on these discoveries and described high-throughput methods for rapidly and cheaply sequencing a whole bacterial genome. Jay Shendure, Greg Porreca and colleagues working in George Church's laboratory described a workflow based on ePCR and sequencing by ligation to sequence a tryptophan-deficient derivative of the *Escherichia coli* strain MG1655. Clonal amplification was achieved with the ePCR system of Dressman et al., and the beads then packed to the surface of a glass slide for sequencing. A library of ePCR-amplified DNA fragments was sequenced using a strategy involving fluorophore-tagged degenerate nonamers, the ligation of which to the template DNA and subsequent fluorescent signal depended on the complementarity of a single base. Mapping of reads to a MG1655 reference genome showed a high degree of accuracy. The estimated cost of US$0.11 per kb of sequence generated was one-ninth of the cost of electrophoretic sequencing methods at that time. The high accuracy, low cost and speed of this method pointed to its potential as a high-throughput technique for resequencing organisms to interrogate genetic variation.

> "making sequencing high-throughput and affordable"

The second paper, by a group of researchers at 454 Life Sciences led by Jonathan Rothberg, introduced a high-throughput sequencing platform based on ePCR and pyrosequencing. Amplification and sequencing of the fragmented *Mycobacterium genitalium* genome was performed in wells containing a single fragment, bead and picolitre reaction volumes on a fibre-optic slide over which dNTPs were added in waves. The blinking light pattern produced as rounds of dNTPs passed over the beads was used to determine the sequence of the bacterial genome.

These massively parallel sequencing techniques set the stage for rapid, low-cost sequencing of genomes. Soon after publication of the 2005 studies, the first high-throughput sequencing machines were marketed, with products based on a range of massively parallel sequencing technologies (MILESTONE 5). By making sequencing high-throughput and affordable, these technologies ushered in a new age of next-generation sequencing.

Joseph Willson,
Nature Reviews Cross-Journal Team

ORIGINAL ARTICLES Shendure, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005) | Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005)

FURTHER READING Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363–365 (1998) | Mitra, R. D. & Church, G. M. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res.* **27**, e34 (1999) | Brenner, S. et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000) | Dressman, D. et al. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl Acad. Sci. USA* **100**, 8817–8822 (2003)

Credit: blickwinkel / Alamy Stock Photo

MILESTONE 4

ChIP–seq captures the chromatin landscape

The fundamental instructions for cellular function are encoded in DNA, which tightly associates with histones to form a complex structure called chromatin. This ensures the stability, replication and proper interpretation of the genetic code.

Chromatin is physically linked to and regulated by a multitude of proteins (for example, transcription factors) and RNAs, ensuring that genes are correctly expressed or silenced in the appropriate cellular context. To understand how chromatin-bound proteins affect gene expression and, ultimately, cell behaviour, it is essential to characterize how, where and when these regulatory proteins bind to chromatin.

Chromatin immunoprecipitation, a technique that goes back to the 1980s, allows the identification of DNA regions that are bound by a protein of interest. First, cells are treated with one or more crosslinking agents (usually formaldehyde) so that covalent bonds are formed between DNA and associated proteins, thus preserving structural and regulatory interactions. Then, the chromatin is fragmented, and an antibody that recognizes a specific protein is used to capture its DNA-bound fragments. Finally, the induced covalent bonds are broken, and the DNA is purified for further analysis.

Initially, these immunoprecipitated DNA fragments were hybridized to microarray platforms (as in the 'ChIP–chip' method) to gain a genome-scale perspective. However, this assay has substantial limitations, including restricted resolution and genome coverage and noisy signals.

In 2007, capitalizing on the rise of next-generation sequencing technologies, ChIP–seq was born, and a series of papers drafted the first genome-wide landscapes of protein–DNA interactions at high resolution. Early studies focused on histone post-translational modifications (Barski et al. and Mikkelsen et al.) and transcription factors (Johnson et al. and Robertson et al.).

Large initiatives such as the ENCODE (MILESTONE 14) or Roadmap Epigenomics Mapping consortia were among the pioneers that leveraged ChIP–seq to characterize the epigenomic profiles of a variety of broadly used cell lines and primary cell types and tissues. To this day, these maps constitute reference datasets for the research community.

> "ChIP–seq … provided the basis for a notable leap in our understanding of chromatin biology"

However, ChIP–seq conducted on heterogeneous cell populations can mask phenomena unique to or more prevalent in certain subpopulations. Technological advances have enabled the development of single-cell ChIP–seq (Rotem et al. and Grosselin et al.). Although these techniques represent exciting steps forwards, they remain technically challenging. Recent and promising alternatives to ChIP–seq include CUT&RUN (Skene and Henikoff) and CUT&Tag sequencing (Kaya-Okur et al.). The advantages over ChIP–seq include not requiring crosslinking and providing a high signal-to-noise ratio at lower sequencing depth. Nevertheless, ChIP–seq continues to be a standard method widely used in transcriptional and epigenetic studies.

The rapid and widespread adoption of ChIP–seq by researchers across fields, such as development, evolution and cancer, provided the basis for a notable leap in our understanding of chromatin biology. By carefully studying patterns of histone modifications and transcription factor binding with improved resolution and with the aid of sophisticated computational tools, scientists deepened their knowledge about basic gene regulatory mechanisms and the role of non-coding genetic variants in disease.

Despite the current popularity of ChIP–seq, available transcription factor and histone mark ChIP–seq data in different cell contexts remain sparse. Data integration and imputation methods using published ChIP–seq are likely to contribute significantly to the ongoing quest to decipher gene regulatory mechanisms in physiological processes and diseases.

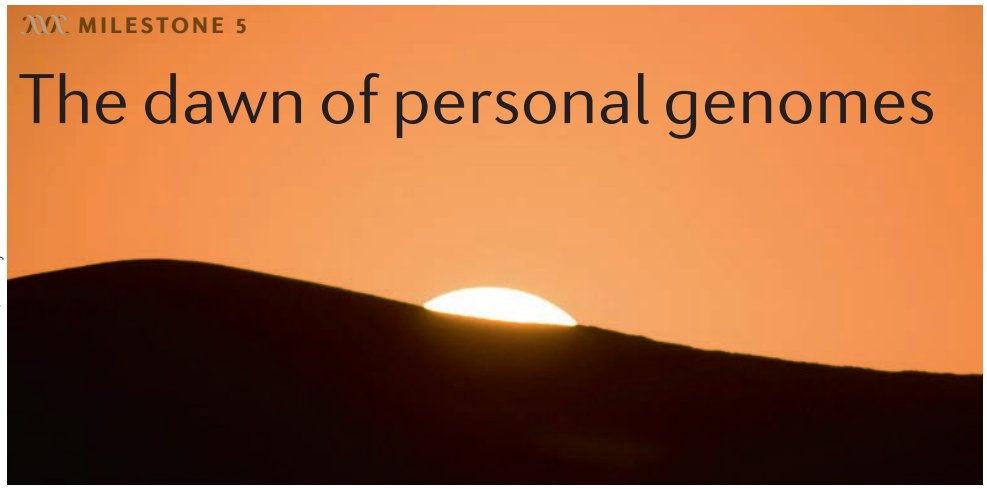
The biological insights obtained in the last decade thanks to the use of ChIP–seq data have been and continue to be transformational.

Tiago Faial, *Nature Genetics*

ORIGINAL ARTICLES Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007) | Mikkelsen, T. S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007) | Johnson, D. S. et al. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007) | Robertson, G. et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007)
FURTHER READING Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015) | Grosselin, K. et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* **51**, 1060–1066 (2019) | Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6**, e21856 (2017) | Kaya-Okur, H. S. et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019)

MILESTONE 5

The dawn of personal genomes



Credit: Redmond Durrell / Alamy Stock Photo

The Human Genome Project showed how large-scale international collaboration could generate a resource of enduring scientific value: a human genome reference sequence (MILESTONE 1). But with a cost of approximately US$300 million and requiring several years, technological advances would be required for whole-genome sequencing to become more widely achievable. The year 2008 saw the publication of pioneering applications of next-generation sequencing technologies to generate genomes of human individuals at a fraction of the cost and time of Sanger sequencing.

In *Nature*, Bentley et al. and Wang et al. reported the genomes of an African individual and an Asian individual, respectively. Bentley et al. introduced a novel massively parallel reversible terminator approach, which was adopted by Wang et al. as well as in a separate study by Ley and colleagues (MILESTONE 6). Originally known as Solexa sequencing, this technique remains the mainstay of short-read Illumina sequencing approaches to the present day.

Whereas Sanger sequencing determines the sequence of a single DNA clone per capillary channel, the major advancement of next-generation approaches was to enable the sequencing of millions of DNA clones simultaneously. To achieve this parallelization, single DNA molecules were immobilized on a flow-cell surface, followed by localized amplification to generate a focal cluster of identical DNA molecules from each starting molecule. A camera then images the flow cell as the DNA clones undergo sequential rounds of controlled, stepwise single-nucleotide addition, with each of the four possible added nucleotides labelled with a different coloured fluorophore that is removed after each cycle. For each clonal spot on the flow cell, the sequential colour changes reveal the DNA sequence.

Although the read lengths of ~35 bases achieved at the time were short relative to Sanger sequencing (and indeed relative to current-day improved short-read methods and long-read methods), the high depth of coverage (>30×) allowed reads to be overlapped and mapped to the existing reference genome, thus generating genome sequences that were near complete except for challenging repetitive regions. Each genome sequence was produced for less than US$500,000 in a few weeks. This progress represented a major step change in affordability and set the stage for time and cost reductions that continued until recently.

New human genomes enable analyses of genetic variation between individuals, and sequencing is particularly powerful for identifying novel variation relative to genotyping microarrays, which require the variants to be already known and predesigned onto the genotyping chip. Bentley et al. and Wang et al. identified several million single-nucleotide variants (SNVs) relative to the human reference genome, many of which were novel.

Despite the short reads, both studies identified thousands of larger structural variants. This was possible owing to the paired-end nature of the sequencing whereby the ends of DNA molecules are sequenced but an intervening genomic region of known approximate length remains unsequenced. Structural variation can be identified when the two sequenced ends map to the reference genome at unexpected distances or orientations to each other. Both studies highlighted the prevalence of polymorphisms in transposable element insertion sites as a major source of human genetic variation.

> “major groundwork for future larger-scale human sequencing projects”

These studies laid major groundwork for future larger-scale human sequencing projects (MILESTONE 13), not just by demonstrating the feasibility and resource value of human genome sequences but also in testing the suitability of associated bioinformatic tools and design considerations such as necessary sequencing depths per individual.

Such studies emerged at a time of great interest in possibilities for personal genomics. As the human reference genome is a composite of genomes from several anonymous donors, it does not represent any single individual. Opening up genome sequencing to individuals allows participants to be informed about not just their own ancestral make-up but also any potential disease-associated genetic variants that may inform risk of future diseases for themselves or their family members.

Lower-depth (~7×) personal genomes of two notable named scientists were revealed at a similar time by Levy et al. using Sanger sequencing and Wheeler et al. using the (now-discontinued) 454 pyrosequencing system. Collectively, these four personal genomes identified several genetic variants of potential medical relevance for the individuals.

Overall, these studies paved the way for the widespread sequencing that we see today in population genomics projects and in clinical applications.

Darren J. Burgess,
Nature Reviews Genetics

ORIGINAL ARTICLES Bentley, D. R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008) | Wang, J. et al. The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008)
FURTHER READING Levy, S. et al. The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007) | Wheeler, D. A. et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008) | Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016)

MILESTONE 6

A sequencing revolution in cancer

By the early 2000s, the first draft of the human genome had been published (MILESTONE 1), and a small number in the cancer research community quickly recognized the promise of sequencing to transform research across the translational pipeline. The technique proved its clinical potential in 2004, when three publications showed that *EGFR* mutations were associated with sensitivity to EGFR inhibitors, demonstrating that sequencing cancer genes could guide precision medicine strategies. Others recognized that genomics — still a fledgling field — could be harnessed to provide answers to some fundamental questions about cancer, such as what is the repertoire of mutations in a tumour?

In 2006, Sjöblom et al. reported results of Sanger-based sequencing of 13,023 genes in 11 breast and 11 colorectal cancers. A main finding was that the repertoire of cancer-associated genes was much larger and more diverse than had been expected and that the mutational spectra between different tissues were surprisingly distinct. This landmark analysis showed that interrogating the cancer genome could identify disease-associated mutations. But it used a labour-intensive, targeted sequencing approach, which was likely to miss important coding and non-coding variants. The field was ripe for an unbiased, genome-wide approach.

Reporting in *Nature* in 2008, Ley et al. presented the first whole-genome sequence of a cytogenetically normal acute myeloid leukaemia (AML) sample using the Solexa/Illumina platform. Two samples taken from a woman aged in her 50s were sequenced; a tumour sample (sequenced at >30-fold coverage) and a normal skin sample. The normal sample allowed the team to filter out somatic mutations from germline ones, a crucial step towards distilling the genetic lesions that were driving the disease. A total of 2,647,695 single-nucleotide

Credit: Sabena Jane Blackbird / Alamy Stock Photo

variants (SNVs) were discovered, of which 2,584,418 were also present in the skin genome. Further filtering for variants occurring in the coding sequences of annotated genes whittled the list down to eight novel non-synonymous mutations and two previously reported ones. The novel mutations occurred in genes with unknown roles in cancer, and their biological functions provided intriguing insight into the pathways disrupted in AML, including small-molecule and drug transport, cell–cell interactions, self-renewal and cell signalling. All mutations, except for a mutation in *FLT3*, were present in every tumour cell at both initial presentation and relapse, suggesting that the disease was driven by a single dominant clone and that the *FLT3* mutation had occurred most recently. Of note, sequencing in an independent cohort of 187 AML samples revealed no recurrent somatic mutations in these genes, providing the first hints that driver mutations are not always indigenous to a tumour type.

The ground-breaking work by Ley et al. confirmed what many had suspected: unbiased, genome-wide sequencing can unmask novel genes that likely contribute to tumorigenesis and offers a palette of potentially druggable targets.

And thus began a sequencing revolution in cancer. A parallel explosion in bioinformatics allowed scientists to make sense of swathes of sequencing data and identify an increasingly complex catalogue of mutations, from SNVs to complex structural rearrangements. The genomic landscapes of different cancer types revealed what Ley et al. had intimated: no two cancer genomes are alike.

> “the first whole-genome sequence of a cytogenetically normal acute myeloid leukaemia”

In 2009, sequencing of the whole genome and transcriptome of a metastatic breast cancer sample offered insights into the evolution of the cancer genome during disease progression. Through its genome, a tumour’s history was laid bare, allowing researchers to identify not only the repertoire of driver mutations in a tumour but also the order in which they occurred.

The field continues to advance at an astonishing pace. As sequencing costs plummeted, small groups coagulated into large, international consortia, reporting on integrated, pan-cancer analyses of thousands of tumours. Sequencing-based assays can now identify disease-specific drivers, mutational signatures, tumour mutational burden and neo-antigens, offering tremendous promise to guide personalized patient care. Ever more powerful analytical tools are allowing us to interrogate these genetic features and infer their biological significance. While the fruits of these innovations are yet to dominate the treatment landscape, it is clear that the best is yet to come.

Safia Danovi,
Nature Genetics

ORIGINAL ARTICLE Ley, T. J. et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008)
FURTHER READING Lynch, T. J. et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **350**, 2129–2139 (2004) | Paez, J. G. et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497–1500 (2004) | Pao, W. et al. EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc. Natl Acad. Sci. USA* **101**, 13306–13311 (2004) | Sjöblom, T. et al. The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006) | Campbell, P. J. et al. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020) | Shah, S. P. et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009)

MILESTONE 7

Transcriptomes — a new layer of complexity

Characterizing genes has been a primary objective for researchers in genetics and genomics. However, direct measurement of the gene product, transcribed RNA, provides a functional read-out of the genome that is essential for our understanding of development and disease.

The early 2000s saw both hybridization-based and sequencing-based approaches being used for the transcriptome-wide quantification of RNA. However, microarray technologies were limited in their ability to detect very lowly expressed genes (sensitivity) and differentiate between genes with sequence homology (specificity). An additional limitation is that only known genes and exons are incorporated into the array, and de novo discovery is not possible. By contrast, sequence-based approaches such as serial analysis of gene expression (SAGE) or massively parallel signature sequencing (MPSS) relied on expensive Sanger sequencing and were limited in their ability to detect all transcript isoforms.

The emergence of next-generation DNA sequencing (MILESTONES 3, 5) enabled the development of high-throughput sequencing of whole transcriptomes, known as RNA sequencing (RNA-seq), reported in 2008 in a series of milestone publications across different species. A typical RNA-seq experiment works by isolating messenger RNA (mRNA) with a poly(A) tail, reverse-transcribing the RNA into complementary DNA (cDNA), sequencing this cDNA using a next-generation sequencing instrument and mapping the resulting reads to the reference genome. RNA-seq reveals an overall picture of which parts of the genome are transcribed and enables accurate RNA quantification at higher resolution and greater dynamic range than previous methods because of its digital read-out, allowing the detection of lowly expressed transcripts. It also permits new genes, exons and transcript isoforms to be identified.

> “The ability to discover new genes or transcripts, the high throughput and the scalability of the technology are major advantages of RNA-seq”

A study by Nagalakshmi et al. demonstrated the power of RNA-seq by sequencing the yeast transcriptome. The majority of reads mapped to known yeast genes, but many reads mapped to regions of the genome that had previously not been known to be transcribed. This study demonstrated the capacity of RNA-seq to capture gene boundaries at unprecedented resolution, including variability in 3′ untranslated regions, which are important for transcript stability and localization. Additionally, it revealed the complexity of the eukaryotic coding genome, uncovering many overlapping genes and alternative transcription start sites.

A contemporaneous study by Lister et al. showed how RNA-seq can be integrated with other genomic data sources to extract functional information about the genome. Here, the researchers combined bisulfite sequencing with RNA-seq in *Arabidopsis thaliana* to study the influence of DNA methylation on the transcriptome.

The ability to discover new genes or transcripts, the high throughput and the scalability of the technology are major advantages of RNA-seq over previous gene expression profiling methodologies. In a study on fission yeast, Wilhelm et al. were able to detect transcriptomes at high resolution over time and under different experimental conditions, capturing the dynamic nature of transcription.

The impact of RNA-seq is far-reaching, and its full potential has yet to be realized. By helping to define the functional genome and quantify RNA over time and under changing conditions, RNA-seq has had a profound impact on research in fields across genetics, biology and medicine, in which it has become a staple research tool.

Margot Brandt, *Nature Communications*

Credit: Y H Lim / Alamy Stock Photo

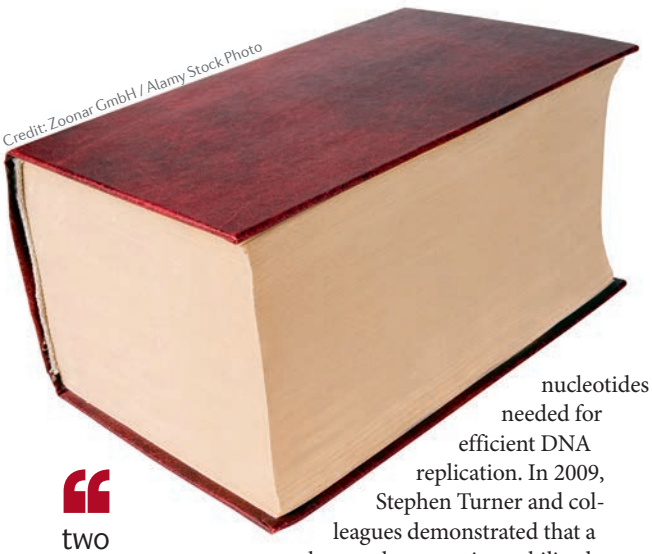
ORIGINAL ARTICLES Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008) | Lister, R. et al. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008) | Wilhelm, B. T. et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single nucleotide resolution. *Nature* **453**, 1239–1243 (2008)
FURTHER READING Cloonan, N. et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008) | Mortazavi, A. et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008) | Velculescu, V. E. et al. Serial analysis of gene expression. *Science* **270**, 484–487 (1995) | Brenner, S. et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000) | Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009)

MILESTONE 8

Long reads become a reality

Despite the high accuracy of short-read sequencing methods, inherent limitations prevent the analysis of some very interesting parts of the genome. In 2009, two fundamentally different sequencing technologies — single-molecule real-time sequencing (SMRT-seq) by Pacific Biosciences (PacBio) and nanopore sequencing by Oxford Nanopore Technologies (ONT) — began to emerge, changing the genomics landscape by introducing multi-kilobase-scale reads.

Instrumental to the development of SMRT-seq were zero-mode waveguides (ZMWs), which modify wave properties of light to allow emitted fluorescence to be observed in a very small volume. This permits single-molecule detection of a high concentration of fluorophore-labelled nucleotides needed for efficient DNA replication. In 2009, Stephen Turner and colleagues demonstrated that a phage polymerase immobilized on a ZMW could be used to amplify DNA and read its sequence in real time. Using deoxynucleotides linked with distinct fluorophores via the terminal phosphate results in cleavage and diffusion of the fluorescent signal following phosphodiester bond formation, such that a brief pulse in fluorescence is detected only while the nucleotide resides in the active site of the enzyme. The emission spectra of the fluorophores permit different nucleotides to be distinguished, while the interval between successive pulses provides information on the kinetics of incorporation. Multiple linked copies of a circular template are generated owing to the high processivity and strand-displacement activity of the polymerase — a feature that greatly improved the accuracy of PacBio sequencing in its later iterations.



Credit: Zoonar GmbH / Alamy Stock Photo

> two fundamentally different sequencing technologies … [changed] the genomics landscape by introducing multi-kilobase-scale reads

Work leading to the launch of the other major player in long-read sequencing, the MinION device by ONT, had been ongoing for several decades. The concept of using an electric potential to drive a DNA molecule through a protein pore embedded in a lipid bilayer and inferring the sequence from the subtle changes in detected current as nucleotides temporarily blocked the protein channel was first proposed in 1989.

In 2009, Hagan Bayley's group showed that an engineered version of the bacterial channel protein α-haemolysin can be used for ordered continuous detection of cleaved nucleotides that were electrophoresed through the pore. A few years later, the groups of Mark Akeson and Jens Gundlach demonstrated that preloading a polymerase on the DNA allowed it to pass through the pore slowly enough for changes in current to be detected, solving a key challenge for the practical implementation of nanopore sequencing. In 2014, the pocket-sized MinION device was offered to a group of early adopters through the MinION Access Programme (MAP), which soon led to the first publication highlighting the device's performance and computational tools for analysis of the resulting data.

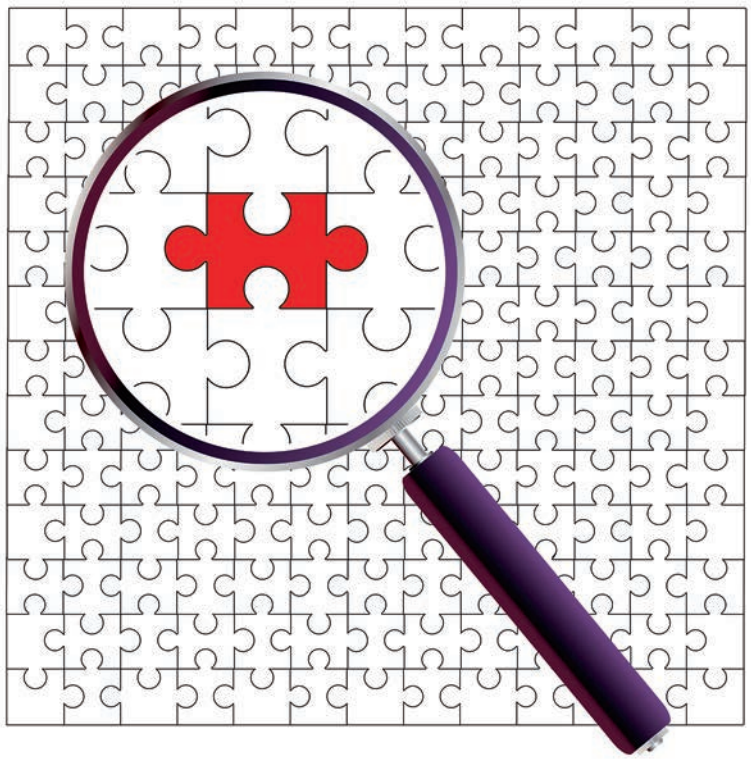
An exciting feature of long-read sequencing is the use of native DNA, which allows modifications to be detected. In SMRT-seq, modified bases affect the kinetics of the polymerase, permitting their identification, but accurate detection relies on high coverage. In nanopore sequencing, modified bases can be detected as electrical current changes; in 2017, two studies provided the first demonstration of genome-wide DNA methylation profiling by the MinION. The ability to detect modified nucleotides is not limited to DNA, as demonstrated by a study showing that nanopore sequencing can be extended to profile the human transcriptome, providing information on splicing, polyadenylation, haplotypes and RNA modifications.

Sequencing long continuous stretches of DNA or RNA, aided by advances in computational algorithms to decode raw signals and assemble genomes de novo, has opened numerous possibilities. Following improvements in library preparation methods and MinION chemistry, in 2018 the first human genome assembled de novo from ultra-long (>100 kb) nanopore reads was published. This was an improvement over previous drafts, providing longer contigs and insight into regions such as the major histocompatibility complex (MHC) locus and telomere repeats. This effort has continued; the Telomere-to-telomere (T2T) consortium recently gave us the first gapless assembly of a human chromosome (MILESTONE 17), a feat achieved by a combination of the latest developments in nanopore and PacBio sequencing and assembly algorithms capable of harnessing the power of long reads.

Ivanka Kamenova, *Nature Protocols*

ORIGINAL ARTICLES Eid, J. et al. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009) | Clarke, J. et al. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009)

FURTHER READING Flusberg, B. A. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465 (2010) | 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) | 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) | Jain, M. et al. Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* **12**, 351–356 (2015) | Rand, A. C. et al. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* **14**, 411–413 (2017) | Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017) | Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018) | Workman, R. E. et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019)



Credit: Panther Media GmbH / Alamy Stock Photo

> The ability to interrogate all protein-coding regions at high sequencing depth in a cost-efficient way has dramatically accelerated the pace of human disease gene discovery

MILESTONE 9

Exploring whole exomes

Historically, pinpointing the genetic basis of a monogenic disease relied on the ability to first delineate a specific chromosomal region within the vast genomic search space where the causal variation resides. The reason for this bottleneck was the lack of a suitable technology to allow researchers to efficiently interrogate the DNA sequences of all protein-coding genes simultaneously. Instead, gene hunters applied tools such as linkage analysis in large pedigrees to identify genetic markers that co-segregated with the disease phenotype, thereby reducing their search space to a tractable number of candidate genes within a defined genomic interval. The development of whole-exome sequencing was a key breakthrough that removed this bottleneck, allowing researchers to identify disease-causing mutations without any prior knowledge of the chromosomal location or biological role of the causal gene.

The technological advance that laid the essential groundwork for whole-exome sequencing was the adaptation of microarrays to perform targeted capture of exon sequences from genomic DNA before high-throughput sequencing. This capture step yields an enrichment of exons relative to other regions of the genome, allowing for high-depth sequence coverage of protein-coding segments at a fraction of the cost of sequencing entire genomes.

In 2009, Sarah Ng and colleagues at the University of Washington reported the first successful application of whole-exome sequencing to monogenic disease. In this landmark proof-of-principle study, the authors performed targeted capture and massively parallel sequencing of whole exomes from eight control samples and four individuals with Freeman–Sheldon syndrome, a rare autosomal dominant disorder known to be caused by mutations in *MYH3*. After filtering out known common variants as well as variants observed among the eight control samples, *MYH3* emerged from their analysis as the only gene with at least one non-synonymous coding variant or splice-site disruption in all four affected individuals. Thus, this study established both a working technology and an analytical framework to examine whole exomes for disease-causing mutations in a robust and cost-effective way.

To explore the potential of this strategy to elucidate the genetic basis of a monogenic disorder of unknown cause, Ng and colleagues next applied their whole-exome sequencing workflow to Miller syndrome, a recessive disorder that had been refractory to standard genetic approaches. By performing whole-exome sequencing of four affected individuals from three independent families and filtering out variants observed in population samples, they identified a single gene, *DHODH*, that was disrupted by rare recessive variants in all affected individuals. To validate this discovery, they analysed *DHODH* using Sanger sequencing in three additional individuals with Miller syndrome and again found biallelic *DHODH* mutations in all three cases, illustrating the power of whole-exome sequencing for monogenic disease gene discovery. In a third study published later that year, Ng and colleagues successfully used whole-exome sequencing to identify mutations in *MLL2* (also known as *KMT2D*) as a major cause of Kabuki syndrome, an autosomal dominant developmental disorder.

Although these early applications of whole-exome sequencing focused on monogenic diseases, this technology has had a transformative impact on many areas of human disease genetics, including cancer genomics and common diseases with complex genetic inheritance. The ability to interrogate all protein-coding regions at high sequencing depth in a cost-efficient way has dramatically accelerated the pace of human disease gene discovery, particularly for rare diseases, and is poised to yield important insights into the genetics of more common diseases as whole-exome sequencing is applied at scale to very large population samples.

Kyle Vogan, *Nature Genetics*

ORIGINAL ARTICLE Ng, S. B. et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009)
FURTHER READING Albert, T. J. et al. Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905 (2007) | Okou, D. T. et al. Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* **4**, 907–909 (2007) | Porreca, G. J. et al. Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007) | Hodges, E. et al. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007) | Ng, S. B. et al. Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010) | Ng, S. B. et al. Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790–793 (2010)

MILESTONE 10

Probing nuclear architecture with Hi-C

Hi-C revolutionized our view of the arrangement of DNA in the nucleus. Before, most insights into nuclear packing came from light microscopy, for which specific loci were tagged with fluorescent proteins to measure their spatial relationship. Here, the major limitation is that the resolution is too low to probe interactions at fine-grained base pair level. Analysis is limited to only a few loci at a time, and cell line construction can be laborious, so the process is low-throughput.

In 2002, Dekker et al. published the chromatin conformation capture (3C) technique, where DNA is crosslinked to its supporting proteins to fix its conformation in space, then digested with a restriction enzyme, leaving sticky ends that can be re-ligated. Any two sequences close in 3D space, but not necessarily close on the linear DNA molecule, can be joined. After crosslink reversal and DNA fragmenting, the hybrid DNA molecules created by ligation of distal sequences can be identified by PCR using primers that bind to each fragment. However, this low-throughput, targeted approach does not enable the discovery of novel interactions.

> “a real whole-genome view of higher-order chromatin architecture”

Credit: Marcio Silva / Alamy Stock Photo

Modifications to this method were made to improve the number of interactions that could be measured, leading to chromatin conformation capture on a chip (4C) and chromatin conformation capture carbon copied (5C). In 2009, Lieberman-Aiden et al. published the Hi-C approach, which used high-throughput short-read sequencing to sequence all hybrid DNA fragments after the ligation step. It was thus possible to identify nearly all fragments of the genome that are physically adjacent to all others, giving a real whole-genome view of higher-order chromatin architecture. For any given nucleus in the sample, the crosslinking provides a snapshot of the interactions at that time. In a bulk sample, the number of reads for a given ligation product increases the more often those two fragments are physically (or spatially) proximal to each other.

The first Hi-C study identified the existence of two nuclear compartments. Sequences in one compartment are more likely to interact with each other, and interactions between the compartments are rare. The compartments were arbitrarily labelled A and B, with the A compartment containing mostly open chromatin with expressed genes and the B compartment containing more densely packed heterochromatin.

Subsequent higher-resolution studies revealed the existence of topologically associating domains (TADs) in mammalian cells and *Drosophila*. TADs are observed at a finer scale than A and B compartments, and intra-TAD sequences interact more frequently than inter-TAD sequences. Domain boundaries are marked by various features including binding sites for CTCF and cohesin proteins.

Technical advances revealed finer-scale structure; sub-TADs and loops are nested within TADs and differ between differentiated and undifferentiated cells, with some changes observed to be correlated with gene expression changes.

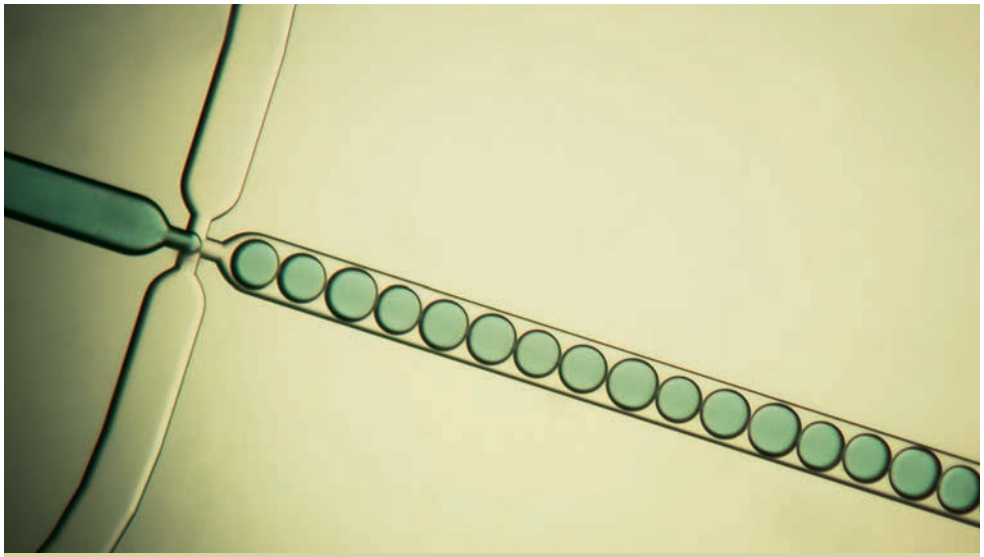
Hi-C represents an averaged view of interaction frequency across a population of crosslinked cells; the extent to which a given interaction occurred in individual cells was not clear. The first single-cell Hi-C study found that single cells have similar domain structure to that seen in bulk Hi-C but that there is stochastic variation in interdomain contacts between cells. Further single-cell studies gave a more dynamic view of chromosome movements and showed that cells could be grouped according to cell cycle stage.

Hi-C has also been used in chromosome scaffolding pipelines for genome assembly to order sequencing reads (MILESTONE 16).

Since the first Hi-C study, our knowledge of how chromosomes are arranged in the nucleus has improved immeasurably, but there is still much to learn. In particular, the functional consequences of the various levels of organization, and how they relate to transcription, are still to be determined. Further refinements and improvements to the Hi-C method will undoubtedly contribute to this understanding in the coming years.

Andrew Cosgrove, *Genome Biology*

ORIGINAL ARTICLE Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009)
FURTHER READING Dekker, J. et al. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002) | Zhao, Z. et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38**, 1341–1347 (2006) | Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006) | Dostie, J. et al. Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006) | Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012) | Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012) | Sexton, T. et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012) | Phillips-Cremins, J. E. et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013) | Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014) | Nagano, T. et al. Single cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013) | Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017)



Credit: Greenshoots Communications / Alamy Stock Photo

MILESTONE 11

Sequencing one cell at a time

Bulk analysis of tissue-based gene expression results in averages that mask contributions of individual cell types. Single-cell genomic approaches are needed to understand developmental processes, homeostatic function and disease initiation and progression. The development of single-cell techniques has enabled analysis of cell states, rare cell types, developmental trajectories, lineage relationships, diseased states and tumour heterogeneity.

Laying the groundwork for these studies, in 2009 Tang et al. performed the first whole-transcriptome analysis of a single mouse blastomere, adapting an amplification method to generate longer cDNAs in an efficient, unbiased manner. This method could detect the expression of 75% more genes than traditional microarrays, while finding new splice junctions.

Islam et al. were among the pioneers who piloted scaled-up single-cell RNA sequencing (scRNA-seq). Sequencing 85 cells of two distinct types, they used single-cell expression profiles to generate 2D maps that clustered cells based on similarities in expression. They could visualize distinct cell populations without a priori knowledge of markers previously used to classify these cell types.

Subsequently, Ramsköld et al. developed Smart-seq to enable single-cell analysis of more cells with enhanced read coverage. Testing the method on circulating tumour cells from melanomas, they were able to identify potential biomarkers for melanoma circulating tumour cells based on distinct gene expression patterns. Compared with the approach of Tang et al., the method showed improved identification of single-nucleotide polymorphisms and alternative transcript isoforms, while quantifying sensitivity and accuracy.

To increase scalability by another three orders of magnitude, allowing for single-cell analysis of entire tissues, the technology had to transition to high-throughput functionality. Two groups dedicated to this goal both turned to microfluidics.

In 2015, Macosko et al. introduced Drop-seq for profiling thousands of cells by encapsulating them into nanolitre-sized droplets, each of which is linked to a unique barcode that connects each mRNA transcript to its cell of origin. Sequencing 44,808 mouse retinal cells identified 39 distinct cell clusters, creating a gene expression atlas with known and newly discovered retinal cell types.

The same year, Klein et al. developed inDrop, which also separates cells into individual droplets that contain hydrogels carrying barcoded primers to connect each cell to its transcriptome. They sequenced >10,000 mouse embryonic stem cells and tracked the onset of differentiation following leukaemia inhibitory factor withdrawal, revealing the population dynamics of differentiating embryonic stem cells. With this technical advance enabling the rapid and efficient sequencing of >10,000 cells, huge datasets were generated — this called for the development of computational techniques to keep pace. Previous strategies needed to be modified to integrate data from different experiments, species, time points or treatment groups.

> “single-cell techniques [have] enabled analysis of cell states, … diseased states and tumour heterogeneity”

Researchers were keen to apply these high-throughput methods to the investigation of different tissues at the single-cell level. Haber et al. profiled individual murine small intestine epithelial cells directly after harvest or following organoid culture. Cells clustered into 15 distinct subpopulations, revealing unexpected diversity and two previously unknown subtypes of Tuft cell. This group also analysed regional differences between the duodenum, jejunum and ileum and cellular response to bacterial and helminth infections.

Another example to illustrate potential applications of scRNA-seq is in the context of characterizing single cells from first-trimester placentas to study the early maternal–fetal interface, as conducted by Vento-Tormo and colleagues. By analysing this specialized tissue, they identified subsets of perivascular and stromal cells, as well as natural killer cells with unique immunomodulatory profiles.

Since the first whole-transcriptome analysis of a single cell in 2009, the community has improved technologies from both engineering and computational perspectives to enable unprecedented insight into biological processes in development, health and disease.

Aline Lückgen,
Nature Communications

ORIGINAL ARTICLE Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009)
FURTHER READING Islam, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011) | Ramsköld, D. et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012) | Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015) | Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015) | Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017) | Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018)

MILESTONE 12

Waking the dead: sequencing archaic hominin genomes

> Ancient DNA research has been limited only by the technology, and never by a lack of interesting questions to be asked

The advent of PCR in the 1980s made ancient DNA (aDNA) sequencing a reality, but early attempts to sequence human aDNA were frustrated by sample contamination and degradation. In 2010, the first draft sequence of a Neanderthal genome heralded a revolution in palaeogenomics, advancing our understanding of the relationships between extinct and extant hominin lineages and how modern humans spread throughout the world.

Ancient DNA research has been limited only by the technology, and never by a lack of interesting questions to be asked. The first aDNA studies using soft tissues from museum specimens were hampered by depurination and fragmentation in the sequenced DNA. By the late 1980s, it was possible to extract DNA from ancient bone, but the limited throughput of the Sanger sequencing technology, and the absence of human reference genomes for comparisons or filtering, made the detection of genuine nuclear aDNA sequences challenging. While the field moved forwards with studies of plants and non-human animals, hominins were somewhat neglected, until next-generation sequencing revolutionized the genomics field as a whole.

Credit: The Natural History Museum / Alamy Stock Photo

A major step forwards took place in February 2010, when Rasmussen and colleagues published the first ancient human genome sequence for an extinct Palaeo-Eskimo, quickly followed by the first Neanderthal genome sequences in April. In their 2010 study, Green et al. generated libraries from three Neanderthal bones from Croatia, dating to >38,000 years before present, fine-tuned these to screen out contamination from microorganisms and modern humans, and sequenced them with a combination of 454 and Illumina technologies, combining the three individuals into a 1.3× coverage genome. Comparisons with the human and chimpanzee genomes allowed the identification of Neanderthal sequences, leading to extensive new inferences about hominin molecular evolution, adaptation and — perhaps most controversially — gene flow between hominin groups. The Neanderthal genome shared more genetic variants with present-day Europeans and Asians than with Africans, suggesting some gene flow after the divergence of these lineages of modern humans. Genomic segments with high similarity to Neanderthal DNA were detected in present-day non-African genomes, providing direct evidence for this introgression (and allowing estimation of the time when it occurred).

To tackle the problem of limited endogenous aDNA quantities, Meyer et al. developed a single-stranded DNA library preparation method for a Denisovan sample in 2012. Their approach substantially increased the number of ancient molecules that could be incorporated into the DNA sequencing libraries, thereby yielding enough DNA sequence to obtain the first high-quality ancient genome, with 30× coverage of a single individual. This study provided further evidence for hominin admixture. Serendipitously, a more recent sequencing study revealed the genome of the offspring of a Neanderthal and a Denisovan.

DNA capture technologies have revolutionized our understanding of human disease and their introduction into the palaeogenomics field enabled the study of polymorphisms present in tens or hundreds of ancient genomes. In the first large-scale study by Haak et al. in 2015, capture was used for the analysis of 394,577 polymorphisms in 69 European individuals dating from 8,000–3,000 years ago, allowing the authors to make conclusions about population movements and turnover during the Neolithic period and the spread of Indo-European languages into Europe.

The field of ancient DNA has illuminated aspects of history that fascinate people from all walks of life. With rapid technological advances and many questions already tackled, the limiting factor may now become the availability of suitable samples, in itself a potentially controversial topic for many reasons.

Rebecca F. Furlong,
Nature Communications

ORIGINAL ARTICLES Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010) | Meyer, M. et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012) | Rasmussen, M. et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010)

FURTHER READING Slon, V. et al. The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* **561**, 113–116 (2018) | Haak, W. et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015) | Sankararaman, S. et al. The date of interbreeding between Neandertals and modern humans. *PLoS Genet.* **8**, e1002947 (2012)

Credit: Cosmo Condina / Alamy Stock Photo

MILESTONE 13

Cataloguing a public genome

Within 10 years of the first human genome being sequenced, the cost of next-generation sequencing was reducing rapidly. Combined with advances in sequencing technology, this meant that exploring the genetic variation in large numbers of people was becoming increasingly tangible.

The 1000 Genomes Project was launched in 2007 with the aim of forming a public reference database to catalogue human genetic variation across major human population groups. In 2012, Gil McVean and colleagues presented phase one, the genomes of 1,092 individuals across 14 populations (including Europe, East Asia, sub-Saharan Africa and the Americas). They combined targeted deep exome sequencing, low-coverage whole-genome sequencing and dense single-nucleotide polymorphism (SNP) genotyping to establish one of the first high-quality resources of its kind. A key feature of this resource included a haplotype map consisting of 38 million SNPs. Low-frequency variants in the non-coding genome were also characterized, showing the ways in which purifying selection acts at functionally relevant sites in the genome. It highlighted the need for rare variants to be interpreted in the context of local genetic background and produced the first 'null expectation', that is, the number of rare, low-frequency and common variants one would expect to find in individuals across different populations.

The methods, technologies and systems developed to form the dataset provided a framework for the generation of large-scale genetic catalogues.

By the time the project reached its final phase in 2015, the dataset represented more than 2,500 individuals from 26 human populations and had contributed to or validated 80% of the variants in the dbSNP catalogue. In the final iteration, Adam Auton and colleagues painted a picture of what a typical human genome looks like, that is, one that consists of around 4–5 million sites that differ from the human reference genome and with variation that differs substantially among population groups.

More recent catalogues of human genetic variation have included the Exome Aggregation Consortium (ExAC) and the Genome Aggregation Database (gnomAD). Released in 2016, the ExAC database was the aggregation of exome sequencing data from more than 60,000 individuals of diverse ancestries (European, African, South Asian, East Asian and Latino ancestries). ExAC has now been replaced by gnomAD, which contains 125,748 exomes and 76,156 genomes in its current release, representing the largest catalogue of human variation.

> “The aggregation of population-level genetic datasets has given us unprecedented access to the depths of the genome”

This catalogue includes high-quality loss-of-function (LOF) annotations, allowing researchers to zoom in on per-base variation and gene-level selection. LOF variation can be used as an in vivo model for human gene inactivation and thus an indicator of a gene's intolerance to inactivation. By comparing expected versus observed variation, Konrad Karczewski, Monkol Lek and colleagues created metrics to quantify the sensitivity of individual genes to variation. These gene-level metrics, as well as estimates of allele frequencies, have been invaluable for the human clinical genetics community when filtering or classifying candidate pathogenic variants in rare disease analyses.

The aggregation of population-level genetic datasets has given us unprecedented access to the depths of the genome. Moreover, the public access of these databases set a precedent for related fields to continue in this spirit of data sharing.

While efforts have been made to include diverse population groups, many of these datasets are nearly devoid of populations from regions such as the Middle East, Oceania and large proportions of Africa. There is still a long road ahead to improve the representation of understudied population groups, making initiatives such as GenomeAsia 100K and H3Africa crucial. We also expect to see more initiatives linking genotype and phenotype information, such as the *All of Us*, 100,000 Genomes and TOPMed programmes. Each of these initiatives are already showing potential to fine-tune our understanding of the impact of genetic variation on human health and disease.

Ingrid Knarston,
Nature Communication

ORIGINAL ARTICLES McVean, G. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012) | Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016) | Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020)
FURTHER READING Durbin, R. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010) | Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015)

MILESTONE 14

Our most elemental encyclopaedia

Long before the first draft of the human genome was completed, it was well established that the DNA sequence alone is insufficient to control the molecular processes of life. Such information needs to be interpreted and regulated, and protein interactions, chromatin structure and chemical modifications were known to play a key role in this process. At the time, such regulatory processes had only been defined for a limited number of genes. The availability of the blueprint of human genetic information (MILESTONE 1), together with exciting new technologies to analyse gene expression and protein–DNA interactions (MILESTONES 4, 7) formed the foundation for the birth of the Encyclopedia of DNA Elements (ENCODE) project.

For the pilot phase of the project, 35 groups around the world produced and analysed 200 datasets, largely based on microarrays, that comprehensively characterized the functional elements of 30 Mb of the human genome — roughly 1% of the total sequence. Intriguing insights of this initial exploration were that the genome is pervasively transcribed and that many of the newly characterized transcripts did not encode any protein and emerged from regions that were thought to be transcriptionally silent. Epigenetic elements associated with active transcription or DNA replication were identified and sorted into large-scale domains. In addition, functional elements were related to evolutionary constraints and genetic variation, which are critical to understanding both the conservation and adaptability of regulatory processes.

The second version of the encyclopaedia, now applying next-generation sequencing technologies to the entire genome, defined not only a set of 20,687 protein-coding genes but also how their expression is controlled in 147 different cell types. About 80.4% of the genome was associated with at least one biochemical event (that is, RNA- or chromatin-related), and 95% was within reach of a DNA–protein interaction. Technologies to probe long-range physical interactions between distinct chromosomes revealed a plethora of promoter–enhancer interactions that are critical for gene activation.

ENCODE 2 was a landmark for the understanding of molecular genetics and a major feat in data standardization, analysis and integration. Predictive models of gene expression were developed based on epigenetic marks or transcription factor binding patterns. Machine learning methods were trained to cluster functional regions across the genome that were afterwards associated with concrete biological processes, such as immune response or neural activity. There was enough statistical power to quantify the impact of negative selection on genomic regulation, and the integration with large databases of annotated variants enabled the exploration of individual functional alterations, with implications for diseases such as cancer. Meanwhile, modENCODE (fly and worm) and mouse ENCODE began the mapping of functional elements in quintessential model organisms.

> “ENCODE 2 was a landmark for the understanding of molecular genetics and a major feat in data standardization, analysis and integration”

Despite the unprecedented resource that the 2012 release of ENCODE represented, the work to characterize the entire functional genome was far from complete. Most of the data had been generated in cell lines, which cannot fully recapitulate the profiles of primary tissues, and most transcription factors had not been assigned to their corresponding genomic elements. The latest version of the encyclopaedia is a collection of 5,992 experiments; it considerably expanded the landscape of regulatory elements both in humans and in the mouse. The ENCODE 3 release also provided maps of cell type-specific 3D chromatin interactions and RNA-binding proteins in human samples, as well as comprehensive profiles of epigenetic changes throughout mouse fetal development.

In the future, we can expect the ENCODE project to expand towards even more comprehensive and functionally tested biochemical profiles, likely incorporating the information from individual genomes and single-cell multi-omics, ensuring that ENCODE remains a crucial reference for our understanding of human biology, evolution and disease.

Ilse Valtierra, *Nature Communications*

Credit: Oleksandra Korobova / Getty images

ORIGINAL ARTICLES Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012) | Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007) | Moore, J. E. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020)

Credit: Art of Food / Alamy Stock Photo

> a single reference genome for a species does not adequately represent the genetic diversity across the population

MILESTONE 15

Pan-genomes: moving beyond the reference

The pan-genome represents the entire set of genes within a species, consisting of a core genome — containing sequences shared between all individuals of the species — and the 'dispensable' genome. The idea of a pan-genome was first conceived for bacterial species in 2005, when the genomes of six strains of *Streptococcus agalactiae* were sequenced, revealing a core genome containing 80% of *S. agalactiae* genes. Since then, there have been efforts to elucidate the pan-genome of many species beyond bacteria. Assembling and studying pan-genomes has shown that relying on a single reference genome for a species can have an adverse effect on our understanding of the genomic basis of diverse traits. For example, many agronomically important genes in plant species are most often found in the dispensable genome.

Putting together a pan-genome for a genome more complex than those of bacterial species was facilitated by improvements in genome sequencing technologies, particularly long-read sequencing. Larger genomes contain higher proportions of repetitive sequences (up to 50% of the human genome and up to 90% of plant genomes consist of repetitive DNA), which are more difficult to analyse using short reads.

The first plant pan-genome was published in 2014, in a study by Li et al. The authors sequenced seven accessions of *Glycine soja*, a wild relative of cultivated soybean (*Glycine max*). Cultivated soybean has lost much of its genetic diversity through domestication, and so *G. soja* represents a potential source of new genes for soybean improvement. The seven accessions used represent 87% of the genetic diversity found in *G. soja*. Performing de novo assembly of the genomes rather than resequencing, the authors found that approximately 80% of the pan-genome is present in all seven accessions, representing the core genome of this species. However, the dispensable genome of *G. soja* contains more than 51% of gene families. Ultimately, this study concluded that having a single reference genome for a species does not adequately represent the genetic diversity across the population.

Subsequent plant pan-genomes have equally shown the importance of looking at the entire gene repertoire in a species. A study of 54 lines of the grass *Brachypodium distachyon* yielded a pan-genome containing twice the number of genes found in any single individual. Many of the genes found in the dispensable genome are involved in functions such as biotic stress response and development. Indeed, disease resistance genes are among the 4,873 genes in the tomato dispensable genome. As climate change and decreases in arable land worsen, these pools of genetic diversity in the dispensable genome represent a promising avenue for introducing beneficial genes into important crop species.

The inadequacy of single reference genomes is not reserved to plants. A study by Sherman et al., published in 2018, sequenced a dataset of 910 human individuals of African descent, working towards assembling a human pan-genome. The authors estimated that up to 10% of the sequences in the total genome are missing from the reference, many of which fall within protein-coding genes. Having a human pan-genome — or the pan-genome of a subset of the human population — allows the discovery of variants that are missing from the reference genome but may be associated with specific phenotypes. Attempts to create a human pan-genome are relatively rare compared with other species, although efforts are underway to capture global diversity.

Major challenges remain. The studies by Li et al. and Sherman et al. were conducted using short-read sequencing. This requires increased sequencing coverage to ensure sufficient coverage to identify variants with confidence. The complexity of human and plant genomes makes assembly of deep sequencing reads time-consuming and computationally expensive. For example, no computational tool exists that is powerful enough to assemble a pan-genome representing all human sequence variation. Advances in sequencing technology and computational tools to assemble genomes should facilitate the construction and study of pan-genomes from humans, plants and many other species.

Dominique Morneau,
Nature Reviews Methods Primers

ORIGINAL ARTICLES Li, Y.-h. et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052 (2014) | Gordon, S. P. et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017) | Sherman, R. M. et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019)

FURTHER READING Tettelin, H. et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl Acad. Sci. USA* **102**, 13950–13955 (2005) | Li, R. et al. Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010) | Golicz, A. A. et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* **7**, 13390 (2016) | Montenegro, J. D. et al. The pangenome of hexaploid bread wheat. *Plant J.* **90**, 1007–1013 (2017) | Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018) | Gao, L. et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**, 1044–1051 (2019) | Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell.* **182**, 145–161 (2020) | Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176 (2020)

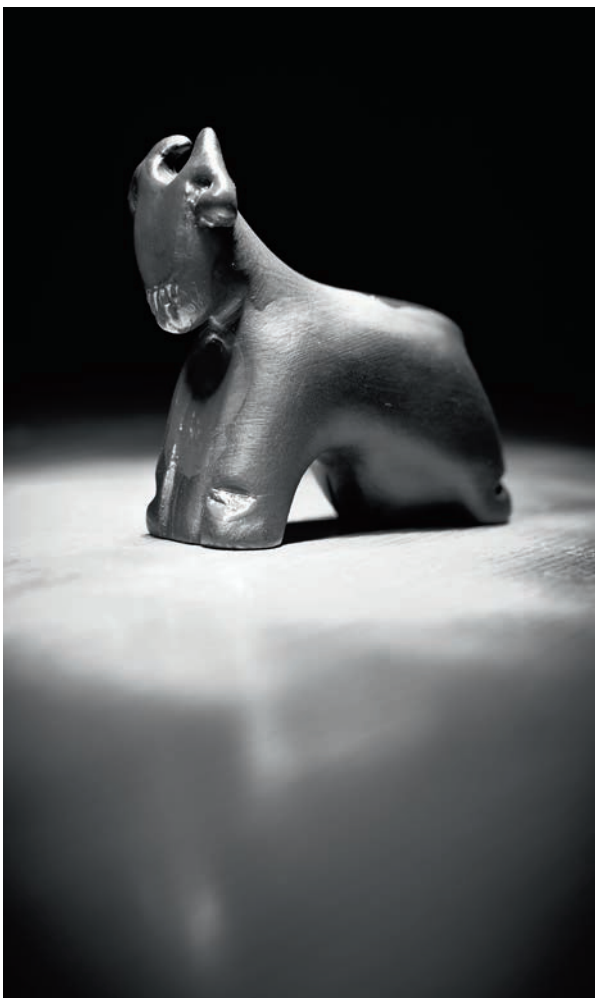
MILESTONE 16

Genomes go platinum

Sequencing a vertebrate genome had almost become routine by 2017, but, with very few exceptions, assemblies of most diploid genomes remained highly fragmented and incomplete. The domestic goat genome ARS1 created a new standard for de novo assemblies of complex genomes.

This was not the first de novo goat genome assembly, but it was by far the most complete. The number of contigs, sequences without gaps, was reduced from more than 330,000 in its predecessor CHIR_1.0 to fewer than 31,000 in ARS1; the number of scaffolds, sequences with gaps, was reduced from 77,431 to approximately 30,000, including all autosomes and the X chromosome. The success of the ARS1 goat genome project hinged on the development and improvement of multiple technologies that had been used in recent genome assemblies, albeit never at once: high-throughput short read DNA sequencing (MILESTONE 5), PacBio long-read sequencing (MILESTONE 8), optical mapping, and Hi-C chromatin interaction data (MILESTONE 10).

> “The domestic goat genome … created a new standard for de novo assemblies of complex genomes”



Credit: Ian M Butterfield (Concepts) / Alamy Stock Photo

Although low-cost high-throughput sequencing provided a way to quickly generate huge amounts of sequence data, these were not sufficient for assembling genomes from scratch. Long-read sequencing promised great advancements in genome completeness but came with the cost of high error rates. In 2012, Koren and colleagues took advantage of the complementary properties of these approaches to develop a strategy for polishing out the errors in the long reads using highly accurate short reads. Around the same time, scaffolding methods were hitting their stride. Optical mapping, a restriction enzyme-based method originally developed in 1993, was merged with microfluidics to produce high-resolution physical maps that could be used to guide genome assembly. This technology was used by Dong et al. to generate the first domestic goat genome (CHIR_1.0) in 2013, as well as by Seo et al. in 2016 to obtain the most contiguous diploid human genome at the time. In parallel, Shendure and colleagues repurposed the intrachromosomal contacts generated in Hi-C experiments to inform the order and orientation of reads to generate chromosome-length scaffolds.

Bickhart et al. showed that these techniques could be combined to capitalize on each of their respective strengths. Optical mapping can correct orientation errors introduced by Hi-C mapping and assembly errors in PacBio contigs, which in turn are consolidated into longer scaffolds by Hi-C. Both scaffolding methods benefit from very long contigs, made possible by the PacBio reads (with sequencing errors corrected by short reads). Together, these four ingredients comprise the recipe for a platinum genome.

This approach was noteworthy for another reason: its price tag. Compared with shotgun-assembled genomes, the goat genome was ~3 times more expensive, but compared with existing reference genomes the savings were substantial. This also opened the door for using a fairly small number of reference taxa as anchors for genome assemblies of closely related species, which could then be sequenced using the more affordable shotgun method.

Today, a true platinum genome requires the ingredients listed above, with the potential addition of haplotype phasing and removal of associated false duplications. Trio-binning, that is the use of parental genomes to resolve haplotypes in diploid individuals, has been proposed as the most effective phasing approach, and it is already being implemented in some of the output of the Vertebrate Genomes Project — an effort aimed at producing complete, error-free and phased diploid genomes for every known vertebrate species on Earth.

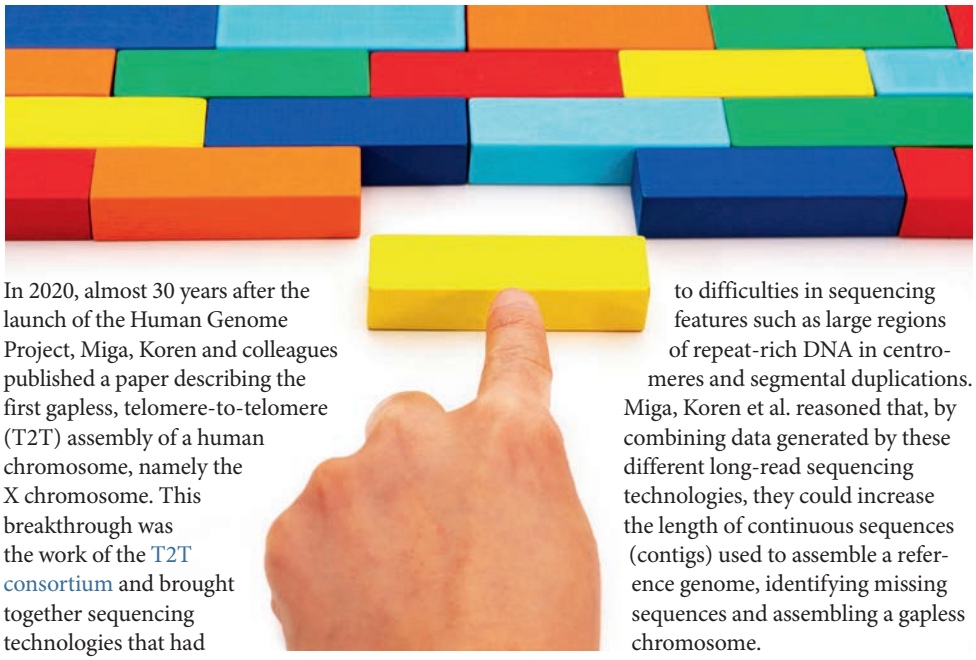
The drive towards telomere-to-telomere genomes (MILESTONE 17) is taking place at breakneck speed, but the humble goat serves as an important signpost marking the start of the platinum era.

Brooke LaFlamme,
Communications Biology

ORIGINAL ARTICLE 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)
FURTHER READING Koren, S. et al. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012) | Dong, Y. et al. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat. Biotechnol.* **31**, 135–141 (2013) | Seo, J.-S. et al. *De novo* assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016) | Burton, J. N. et al. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013) | Koren, S. et al. *De novo* assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018) | Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.22.110833 (2020) | Hastie, A. R. et al. Rapid genome mapping in nanochannel arrays for highly complete and accurate *de novo* sequence assembly of the complex *Aegilops tauschii* genome. *PLoS ONE* **8**, e55864 (2013)

MILESTONE 17

Filling in the gaps telomere to telomere



Credit: PORNCHAI SODA / Alamy Stock Photo

In 2020, almost 30 years after the launch of the Human Genome Project, Miga, Koren and colleagues published a paper describing the first gapless, telomere-to-telomere (T2T) assembly of a human chromosome, namely the X chromosome. This breakthrough was the work of the T2T consortium and brought together sequencing technologies that had been developed in the preceding 6 years.

In 2015, Chaisson et al. showed that long-read sequencing technology from Pacific Biosciences (PacBio) could be used to sequence a human genome, specifically that of the complete hydatidiform mole (CHM) cell line CHM1. As CHM cells have a duplicated paternal (but no maternal) genome, bypassing the need to assemble both haplotypes of a diploid genome, they became a key reference genome. Later that year, Berlin, Koren et al. reported the first de novo assembly of a human genome based on PacBio sequencing long reads alone. Then, in 2018, Jain et al. revealed that ultra-long-read nanopore sequencing (from Oxford Nanopore Technologies) could also be used to assemble a human genome de novo (MILESTONE 8). Finally, in 2019, Wenger, Peluso et al. introduced PacBio high-fidelity (HiFi) sequencing, which was 99.8% accurate in sequencing the human genome reference standard HG002 over average read lengths of 13.5 kb.

Although these technological advancements were reported to have closed gaps in the GRCh37 or GRCh38 version of the human reference genome, no chromosome had been sequenced in full owing to difficulties in sequencing features such as large regions of repeat-rich DNA in centromeres and segmental duplications. Miga, Koren et al. reasoned that, by combining data generated by these different long-read sequencing technologies, they could increase the length of continuous sequences (contigs) used to assemble a reference genome, identifying missing sequences and assembling a gapless chromosome.

> “The sequencing of the first two complete chromosomes … suggested that it was technically possible to complete the human genome sequence”

Consequently, they sequenced 155 Gb of DNA from CHM13 cells with nanopore sequencing, using the genome assembly tool Canu to combine these ultra-long reads with data previously generated by PacBio sequencing. Nanopore sequencing, PacBio sequencing and linked-read Illumina sequencing were used to polish their assembly of the CHM13 genome, a 2.94-Gb assembly with a median consensus accuracy of ~99.99% and in which 50% of the genome was within contigs of ≥70 Mb. The presence of 41 of 46 telomeres at contig ends suggested that CHM13 was a more complete reference genome than GRCh38.

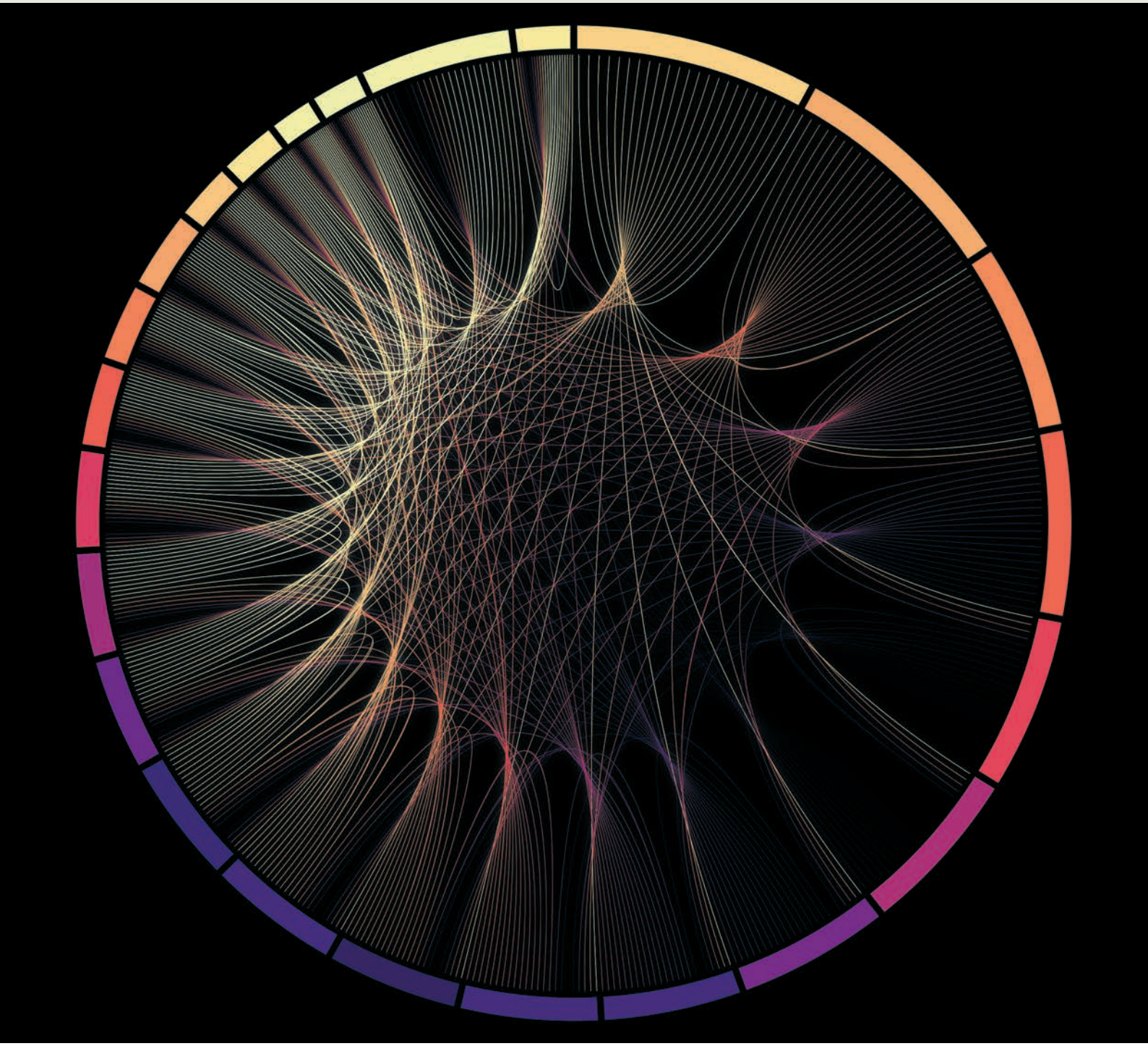
Indeed, Miga, Koren et al. noted that the X chromosome in their CHM13 assembly was broken in just three places. To fill in these gaps, they first mapped ultra-long reads against the assembly, manually identifying reads that joined breaks between contigs; this approach resolved two breaks resulting from segmental duplications. These findings were validated by mapping independent long-read PacBio HiFi data from CHM13 to the X chromosome. To resolve the third break, which was at the centromere, the researchers uniquely tiled ultra-long reads across the repeat-rich centromeric α-satellite array on the X chromosome, confirming the results with long-read PacBio HiFi data and benchmarking and improving the centromere assembly using an automated satellite assembly method (CentroFlye) and evaluation tools (TandemTools). After polishing, the gapless X chromosome assembly was ≥99.9% accurate and had resolved 29 reference gaps. By precisely mapping long-read data to the finished chromosome, the researchers also produced the first comprehensive, T2T profile of DNA methylation, enhancing our picture of epigenetic regulation over repeat-rich regions.

Sequencing of the X chromosome led the way to the T2T assembly of the first autosome, chromosome 8 from CHM13 cells, as announced by Logsdon et al. later in 2020. Combining nanopore, PacBio and PacBio HiFi sequencing, this work closed up five gaps in chromosome 8 and produced an assembly with an accuracy of >99.99%.

The sequencing of the first two complete chromosomes, 20 years after the release of the first draft human genome (MILESTONE 1), suggested that it was technically possible to complete the human genome sequence. Indeed, in September 2020, the T2T consortium announced that they had filled in all of the gaps, obtaining complete sequences for all the chromosomes in CHM13 cells (apart from the five ribosomal DNA arrays) and thus, outstandingly, a v1.0 assembly of a complete human genome.

Katharine H. Wrighton,
Nature Reviews Cross-Journal Team

ORIGINAL ARTICLE Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020)
FURTHER READING Berlin, K. et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015) | Chaisson, M. J. P. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015) | Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018) | Logsdon, G. A. et al. The structure, function, and evolution of a complete human chromosome 8. Preprint at *bioRxiv* https://doi.org/10.1101/2020.09.08.285395 (2020) | Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019) | Jain, M. et al. Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323 (2018) | 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) | 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)



SPRINGER NATURE