Beyond Popular Science

David H. Silver

BEYOND POPULAR SCIENCE

David H. Silver

https://www.openbookpublishers.com

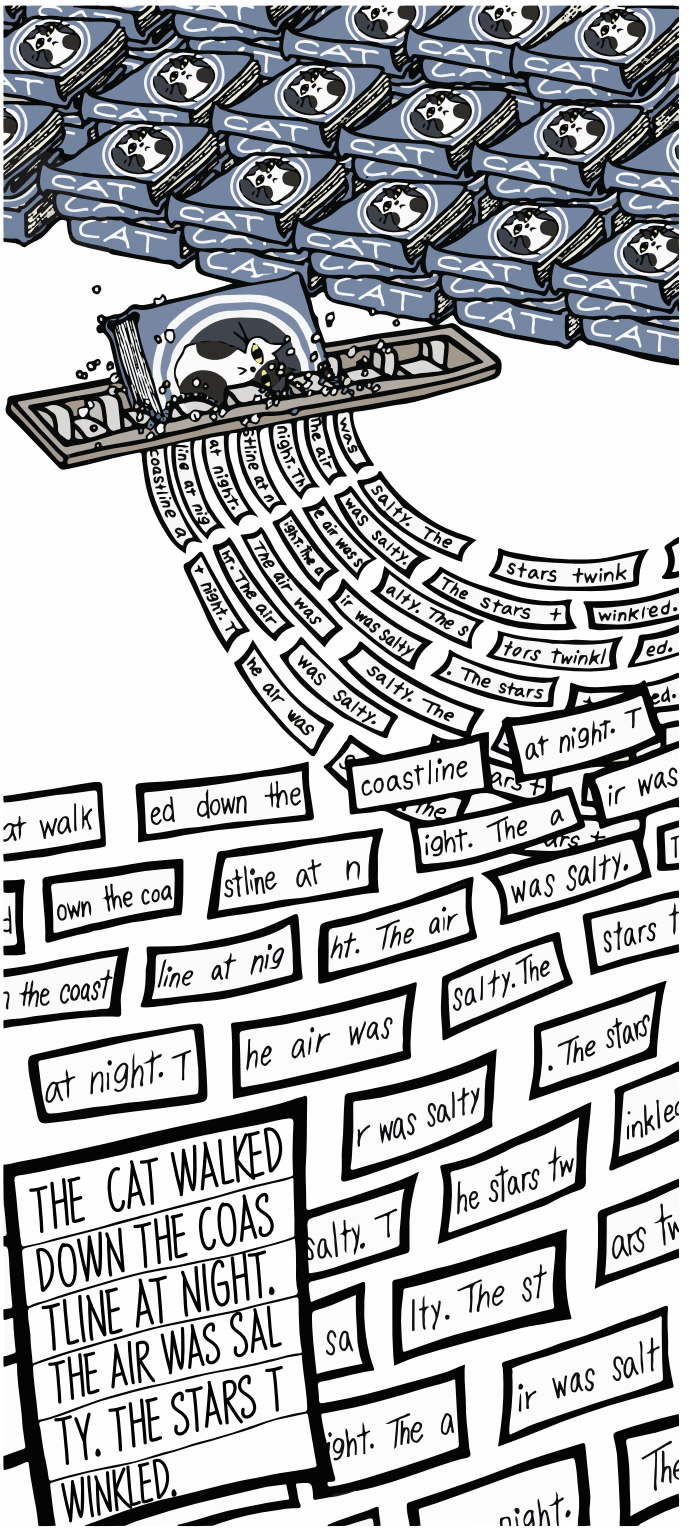
All external links were active at the time of publication unless otherwise stated and have been archived via the Internet Archive Wayback Machine at https://archive.org/web

Digital material and resources associated with this volume are available at https://doi.org/10.11647/OBP.0526#resources

ISBN Paperback:	978-1-80511-877-0
ISBN Hardback:	978-1-80511-878-7
ISBN Digital (PDF):	978-1-80511-879-4
ISBN HTML:	978-1-80511-881-7
ISBN Digital ebook (epub):	978-1-80511-880-0
DOI:	10.11647/OBP.0526

Cover image by Enny Silver and David H. Silver
Cover design by Jeevanjot Kaur Nagpal

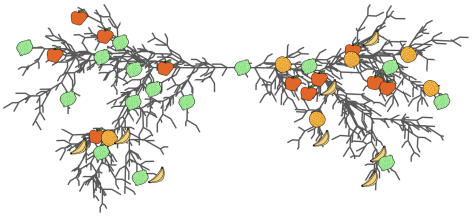
Slices of Life



Sequencing as Text Reconstruction: The image parallels DNA sequencing to reassembling a book from shredded fragments from hundreds of identical copies. The books are pulverised into overlapping sentence fragments—analogous to sequencing short reads from long DNA strands. By aligning overlapping sequences, the original string is reconstructed. As with genome assembly, redundancy, overlaps, and statistical inference allow one to piece together each chromosome from fragmented reads.

Slices of Life

DNA sequencing has evolved from Sanger's chain-termination method through the next-generation revolution of 454 pyrosequencing, Ion Torrent, and Illumina platforms to modern nanopore technologies. The computational challenge of genome assembly uses sophisticated algorithms such as de Bruijn graphs to reconstruct complete genomes from millions of short fragments, while paired-end chemistry and long-read technologies help resolve repetitive regions that have long frustrated genomic reconstruction efforts.



DNA Sequencing Evolution ∘ Sanger ddNTP Method ∘ Illumina Reversible Terminators ∘ Bridge Amplification ∘ PacBio SMRT Technology ∘ Oxford Nanopore ∘ De Bruijn Graph Assembly ∘ K-mer Overlap ∘ Contigs vs Scaffolds ∘ N50 Statistic ∘ Long-Read Revolution

"Would I have invented PCR if I hadn't taken LSD? I seriously doubt it. I could sit on a DNA molecule and watch the polymers go by. I learned that partly on psychedelic drugs."

— Kary Mullis, 1998

Slices of Life

Frederick Sanger's 1977 chain-termination method emerged from years of frustration with earlier approaches to reading DNA. His insight—using modified nucleotides to randomly terminate DNA synthesis—provided the first practical way to determine base sequences. While Allan Maxam and Walter Gilbert simultaneously developed a chemical cleavage method, Sanger's approach proved more robust and became the foundation for three decades of genomic research.

The Human Genome Project launched in 1990 as biology's moonshot—a publicly funded effort to read all 3.2 billion letters of human DNA. Francis Collins led the international consortium, methodically mapping and sequencing chromosomes piece by piece. Then in 1998, Craig Venter announced that his company, Celera Genomics, would sequence the human genome in just three years using a 'whole genome shotgun' approach (Venter et al., *Science* 280, 1998, pp. 1540–1542)—fragmenting the entire genome at once and using computational power to reassemble it.

The race was on. The public project, with its careful clone-by-clone strategy, suddenly faced a nimble competitor unconstrained by academic collaboration requirements. Venter's team used hundreds of automated sequencers running 24/7, while the public consortium scrambled to accelerate their timeline. Both sides published draft sequences simultaneously in February 2001—a diplomatic resolution to a bitter competition that had featured Congressional hearings, patent disputes, and public acrimony. The project cost approximately $3 billion and required a decade of work.

Sanger sequencing's limitations—high cost and low throughput—motivated a new generation of technologies. 454 Life Sciences introduced pyrosequencing in 2005, detecting DNA synthesis through light emission and reading millions of fragments simultaneously. This began the 'next-generation' era, where parallelisation replaced precision.

Illumina emerged as the dominant platform after acquiring Solexa technology in 2007. Their reversible terminator chemistry solved pyrosequencing's homopolymer problems while maintaining massive throughput. The cost per genome plummeted from millions to thousands of dollars, democratising genomic research.

The push for longer reads drove development of single-molecule technologies. Pacific Biosciences (PacBio) spent a decade perfecting zero-mode waveguides—zeptolitre observation chambers that could watch individual DNA polymerase enzymes at work. Oxford Nanopore took a different path, threading DNA through protein pores and reading the sequence from electrical current fluctuations. When they released the MinION in 2014—a sequencer the size of a USB stick—it showcased the progress the technology has made from room-sized machines of the genome project era.

The computational challenge evolved in parallel. Early assembly algorithms handled thousands of Sanger reads; modern de Bruijn graph methods process billions of short reads. Long-read assemblers now tackle the ultimate challenge: reconstructing complete chromosomes from end to end. Sequencing costs have fallen faster than Moore's

 https://doi.org/10.11647/OBP.0526.40

> Law—from roughly dollars per base in 1990 to well under a dollar per megabase today, and under a thousand dollars per human genome on leading platforms.

DNA (deoxyribonucleic acid) is the molecule that stores genetic information in all living organisms. It consists of two complementary strands twisted into a double helix, where each strand (Watson & Crick, 1953) is a linear sequence of four chemical bases: adenine (A), cytosine (C), guanine (G), and thymine (T). The sequence of these bases encodes the instructions for building and maintaining an organism.

The central dogma of biology describes how genetic information flows (Crick, 1970). DNA is transcribed into RNA, which is translated into proteins. Each three-base sequence (codon) in DNA specifies one amino acid in the resulting protein. A single change in the DNA sequence can alter the protein's structure and function, causing disease or evolutionary adaptation. Understanding DNA sequences is a key component in understanding the molecular basis of life.

Consider a short DNA sequence: ATGCGATCGATCG. This 12-base fragment contains four codons: ATG, CGA, TCG, ATC. The codon ATG typically signals 'start translation,' while the others specify specific amino acids. If the first base changes from A to T, creating TTGCGATCGATCG, the first codon becomes TTG, which codes for a different amino acid, potentially altering the resulting protein's function.

DNA sequencing is the process of determining the exact order of bases in a DNA molecule. This requires overcoming a scale mismatch: individual bases measure roughly one nanometre, while human chromosomes stretch millions of bases long which makes it practically impossible to read the entire sequence in one pass.

The solution involves fragmenting DNA into manageable pieces, reading each fragment separately, then computationally reconstructing the original sequence. This creates two challenges. The first is the biochemical problem of reading individual fragments and the second is the algorithmic problem of assembling them correctly. Each generation of sequencing technology has approached these challenges in different ways.

Frederick Sanger solved the reading problem through controlled interruption of DNA synthesis. DNA polymerase builds new strands (Sanger, Nicklen, & Coulson, 1977) by adding nucleotides complementary to a template; a 3'-hydroxyl group on each nucleotide enables the next to attach. Sanger introduced dideoxynucleotides (ddNTPs) that lack this hydroxyl group, terminating synthesis when incorporated.

This process is probabilistic. In a mix containing a DNA template, primers, polymerase, all four normal dNTPs, and a small amount of one ddNTP type (e.g., ddATP), the polymerase occasionally incorporates a ddATP, terminating the strand. Across millions of template copies, this generates a collection of fragments of different lengths, each ending at a different A position.

Running four parallel reactions—one for each ddNTP type—produces four fragment collections. Gel electrophoresis separates these by size: DNA fragments migrate through a polymer matrix under an electric field, with smaller fragments moving faster. After separation, each gel lane shows a ladder of bands. Reading from shortest to longest

fragment across all four lanes reveals the sequence. If the shortest fragment appears in the G lane, the first base is G. If the next shortest is in the A lane, the second base is A. Going through all four lanes reveals the sequence.

Sanger sequencing powered the Human Genome Project but had limitations. Each reaction produced only 500–1000 readable bases. Preparing samples, running gels, and reading results consumed hours per reaction. Radioactive or fluorescent labelling added complexity and cost. The throughput ceiling meant that sequencing a human genome required years of work and hundreds of millions of dollars.

Next-generation platforms achieved a breakthrough with parallelization, performing millions of reactions simultaneously on a single surface. Early systems distributed single DNA fragments into millions of microscopic wells and flowed one nucleotide type at a time (first all As, wash; then all Cs, wash). Detection chemistry varied. 454 Life Sciences used pyrosequencing, where nucleotide incorporation releases pyrophosphate (PPi), which an enzyme cascade converts into a light signal via luciferase. Ion Torrent used semiconductor sequencing, where incorporation releases a hydrogen ion, and millions of ISFET (ion-sensitive field-effect transistor) sensors detect the resulting pH change as a voltage signal. Both methods suffered from the same core limitation. Because nucleotides were added sequentially, homopolymer runs like AAAA caused all four bases to incorporate at once, producing a signal four times stronger. Distinguishing a 4× signal from a 5× signal was error-prone, limiting accuracy.

Illumina took a different path, solving the homopolymer problem through reversible termination. Their innovation (Bentley et al., 2008) combined three key elements: surface-bound amplification, chemically cleavable terminators, and four-colour imaging.

The process begins with bridge amplification. DNA fragments attach to a glass surface coated with two types of oligonucleotide primers. Each fragment bends to hybridise with a nearby complementary primer, forming a bridge. Polymerase extends the primer, creating a complementary strand anchored at both ends. Denaturation releases the original strand, and the process repeats. After 35 cycles, each original molecule generates a tight cluster of ∼1,000 identical copies, all within a few hundred nanometres—small enough to act as a single sequencing unit but bright enough for fluorescence detection.

Illumina's sequencing chemistry uses nucleotides engineered with two modifications: a fluorescent dye unique to each base (A, C, G, T) and a chemical block on the 3'-OH that prevents further extension. Unlike Sanger's permanent terminators, these blocks can be cleaved chemically.

Each sequencing cycle follows four steps: add all four labelled terminators simultaneously, wait for incorporation, image in four colours, then cleave both dye and terminator. Because only one base can be added per cycle (due to the 3'-block), homopolymers read accurately—AAAA requires four separate cycles, each adding one A. This solved 454's limitation.

Illumina's paired-end innovation provided long-range information. Sequence both ends of a DNA fragment, keeping track that they came from the same molecule. If fragments are 500 bases long but you only read 150 bases from each end, you know those two 150-base

sequences sit 200 bases apart in the genome. These distance constraints prove essential for genome assembly.

Assembling a 3-billion-base human genome from 20 million 150-base fragments is computationally demanding. Early overlap-layout-consensus algorithms, which compare all read pairs to find overlaps, were feasible for thousands of Sanger reads but fail for millions of short reads where all-pairs comparison is prohibitive.

De Bruijn graphs, a 1946 mathematical structure, provided a scalable solution. Instead, of connecting reads, they connect k-mers—all possible k-letter substrings. A sequence traces a path through a graph where each unique k-mer is a node and edges connect k-mers overlapping by k-1 bases. The scalability arises because a genome of length G contains at most $G - k + 1$ distinct k-mers, regardless of sequencing depth. Finding Eulerian paths that traverse each edge once is tractable even for graphs with billions of nodes.

Consider the sequence ATCGATCG and extract all 3-mers: ATC, TCG, CGA, GAT, ATC, TCG. Build a graph where each unique k-mer is a node, and edges connect k-mers that overlap by k-1 bases. The sequence ATCGATCG traces a path through this graph: ATC→TCG→CGA→GAT→ATC→TCG (compare to the superpermutation problem in Chapter 39).

Repeats in the genome, such as transposable elements, complicate assembly. When a repeat is longer than a read, it creates ambiguity in the assembly graph, resulting in multiple valid paths. Paired-end constraints resolve some ambiguities, but short reads cannot span long repeats, requiring the development of long-read technologies.

Pacific Biosciences (PacBio) developed single-molecule real-time (SMRT) sequencing, observing individual DNA polymerase enzymes. The primary challenge was detecting single fluorescent nucleotides against the background of unincorporated ones. The solution was zero-mode waveguides (ZMWs): 70 nanometre holes in an aluminium film that confine laser illumination to a 20 zeptolitre volume. A polymerase at the bottom of each ZMW holds an incorporating nucleotide for milliseconds, long enough to generate a detectable fluorescent flash distinct from the transient signals of freely diffusing nucleotides. This method generates reads exceeding 10,000 bases. Though noisy, with error rates of 10–15%, these long reads are effective at spanning genomic repeats.

Oxford Nanopore technology uses no enzymes or fluorescence. It passes a single DNA strand through a protein nanopore (Kasianowicz, Brandin, Branton, & Deamer, 1996) embedded in a membrane. An applied voltage drives both the DNA and an ionic current. As the DNA translocates, nucleotides in the pore's 1.4 nanometre constriction modulate the current. The narrowest region spans approximately five bases, so the signal reflects a 5-mer. Signal processing algorithms, and recently, neural networks, decode the complex current modulations into a DNA sequence, achieving >95% accuracy and read lengths that can exceed one million bases.

Long reads simplified assembly graphs, as most repeats become trivial to span. Modern projects often use a hybrid approach: Illumina provides an accurate short-read backbone, while PacBio or Nanopore provides a long-read scaffold to resolve repeats and structural variants.

On Assembly Statistics

Evaluating a genome assembly requires understanding its output format. Assemblies consist of fragments at two organisational levels. A **contig** is a contiguous stretch of sequence assembled from overlapping reads—an unbroken text. A **scaffold** links multiple contigs that are ordered and oriented but separated by gaps of estimated size. Consider recovering a book's text from shredded copies: contigs are complete pages reconstructed from overlapping fragments, scaffolds are chapters where page order is known but some pages remain missing.

The median contig length—the middle value in a sorted list—is uninformative because assemblies contain thousands of short contigs and few long ones. The **N50** statistic measures contiguity differently. Sort contigs from longest to shortest, then sum their lengths sequentially. The N50 is the length of the contig that brings this cumulative sum to 50% of the total assembly size. For contigs of lengths 10, 9, 8, 7, 1, 1, 1, 1, and 1 kb (total 39 kb), the cumulative sum surpasses 50% after adding the first three contigs ($10 + 9 + 8 = 27 > 39/2 = 19.5$ kb). The N50 is 8 kb—the length of that third contig. The median is 1 kb.

The scientific literature sometimes misreports N50 as 'median contig length (N50).' The N50 is a length-weighted metric, not the simple median of contig lengths. Describing N50 as a 'weighted median' is correct if one creates an expanded list where each contig appears once for each base it contains, then takes that list's median.

My first first-author paper addressed a bottleneck of this era: when paired-end reads were short, their two ends often failed to overlap, leaving assemblers unable to merge them into longer fragments. ELOPER (Silver et al., *Bioinformatics*, 2013) detected pairs whose ends overlapped simultaneously at both positions, effectively doubling usable read length while also providing long-range scaffolding, thus improving assembly quality. Within a few years, advances in sequencing technology made reads long enough to render the problem—and the tool—unnecessary. I also helped assemble the genome of *Acrobeloides nanus*, a nematode whose gene regulatory program revealed conserved phylum-specific expression patterns across deep evolutionary distances (Schiffer et al., *PNAS*, 2018)—a collaboration with a dear friend in Edinburgh, completed as part of my honeymoon trip with my wife Enny.

That work was conducted under Itai Yanai at the Technion, where most of my biology-related research took place. Yanai has since moved to NYU, where he is Professor of Biochemistry and Molecular Pharmacology and founding director of the Institute for Computational Medicine. A pioneer of single-cell transcriptomics and co-author of *The Society of Genes*, he is among the most effective communicators of science to graduate students and young researchers. His teaching, curiosity, and attitude toward what science is shaped much of the sensibility behind this book.

Illumina Sequencing and De Bruijn Assembly

Sequencing by Synthesis

Illumina uses reversible terminators with cleavable fluorescent labels. Each cycle:

$$\begin{aligned}&\text{DNA}_n + \text{dNTP-3'-block-fluor}\\ &\quad\xrightarrow{\text{pol}} \text{DNA}_{n+1}\\ &\text{Imaging} \rightarrow \text{Base identification}\\ &\text{Chemical cleavage} \rightarrow \text{3'-OH restoration.}\end{aligned}$$

Bridge amplification creates clonal clusters ($\sim 10^3$ copies) on flow cell surface. Fluorescence signal $S \propto N_{\text{mol}}$ enables base calling with error rate $\varepsilon \approx 0.1\%$.

De Bruijn Graph Construction

For read set $\mathcal{R}$ with k-mer length k:

$$\begin{aligned}V &= \{w \in \Sigma^{k-1} : w \text{ is prefix or suffix}\\ &\qquad \text{of some k-mer in } \mathcal{R}\}\\ E &= \{w \in \Sigma^{k} : w \text{ appears in } \mathcal{R}\},\end{aligned}$$

where each k-mer edge connects its $(k-1)$-mer prefix to its $(k-1)$-mer suffix.

Each read of length L contributes $L-k+1$ k-mer edges, compressing redundant sequence information into a compact graph structure.

Eulerian Path Assembly

Assembly seeks Eulerian path through G:

$$\begin{aligned}\text{Path} &= e_1 e_2 \ldots e_m \text{ where}\\ &\qquad \forall i : \text{tail}(e_i) = \text{head}(e_{i+1})\\ \text{Genome} &= e_1[1..k] + e_2[k] + e_3[k]\\ &\qquad + \ldots + e_m[k],\end{aligned}$$

where $e_i[k]$ denotes the last base of k-mer e_i.

For an Eulerian path to exist, the underlying graph over nonzero-degree vertices must be weakly connected, with all vertices balanced (in-degree = out-degree), or exactly two semi-balanced vertices: one with out-degree = in-degree + 1 and one with in-degree = out-degree + 1.

Coverage and k-mer Selection

Expected k-mer coverage:

$$C_k = C_{\text{read}} \cdot \frac{L-k+1}{L},$$

where $C_{\text{read}} = NL/G$ (reads $\times$ length / genome).

Optimal k balances a tradeoff: smaller k yields more connections and higher coverage but introduces ambiguity, while larger k reduces repeat ambiguity at the cost of lower coverage and potential gaps.

Typically $k \in [50, 250]$ for Illumina data.

Graph Complexity

Real graphs contain three main types of structural features: **bubbles** (parallel paths created by SNPs or errors), **tips** (dead ends from coverage gaps), and **repeats** (creating branching and convergence). Error correction typically removes k-mers with coverage below a threshold.

Paired-End Constraints

Insert size $d \sim \mathcal{N}(\mu, \sigma^2)$ provides scaffolding:

$$|p(r_1, r_2) - \mu| < 3\sigma,$$

where $p(r_1, r_2)$ is genomic distance between read pairs.

References:

Bentley et al. (2008). *Nature* 456:53–59.
Pevzner et al. (2001). *PNAS* 98:9748–9753.