The History of Sequencing and Modern Approaches

Julian Catchen

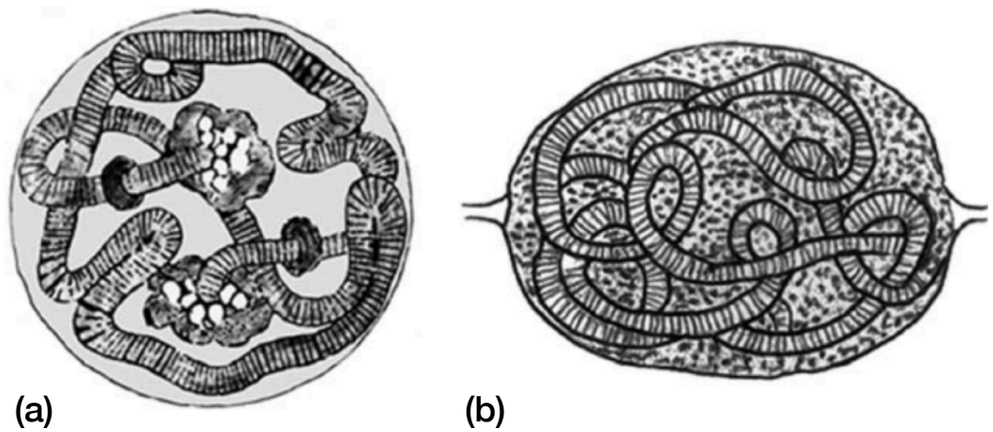
jcatchen@illinois.edu
@jcatchen

Department of Animal Biology

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

What was known about the *hereditary material* in the 1940s?

- ***Chromosomes*** (polytene) could be seen under a light microscope



(a) (b)

First drawings of polytene chromosome made by Balbiani in (a) 1881 and (b) 1890. (a) Salivary gland cells of *Chironomus plumosus* and (b) *Loxophyllum meleagris*.

Zhimulev and Koryakov. (2009) Polytene Chromosomes. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons.

- chromosomes contained multiple ***loci*** responsible for a hereditary traits
- ***recombination*** of chromosomes was understood

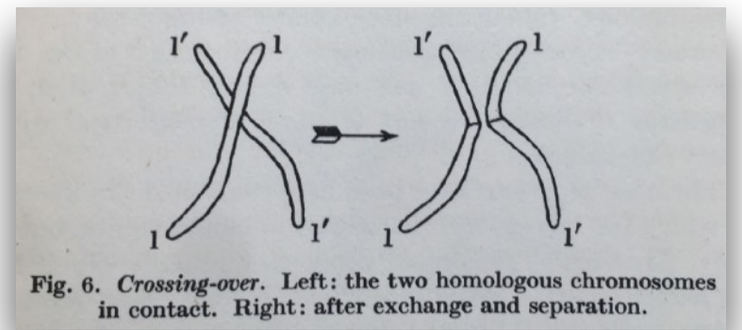
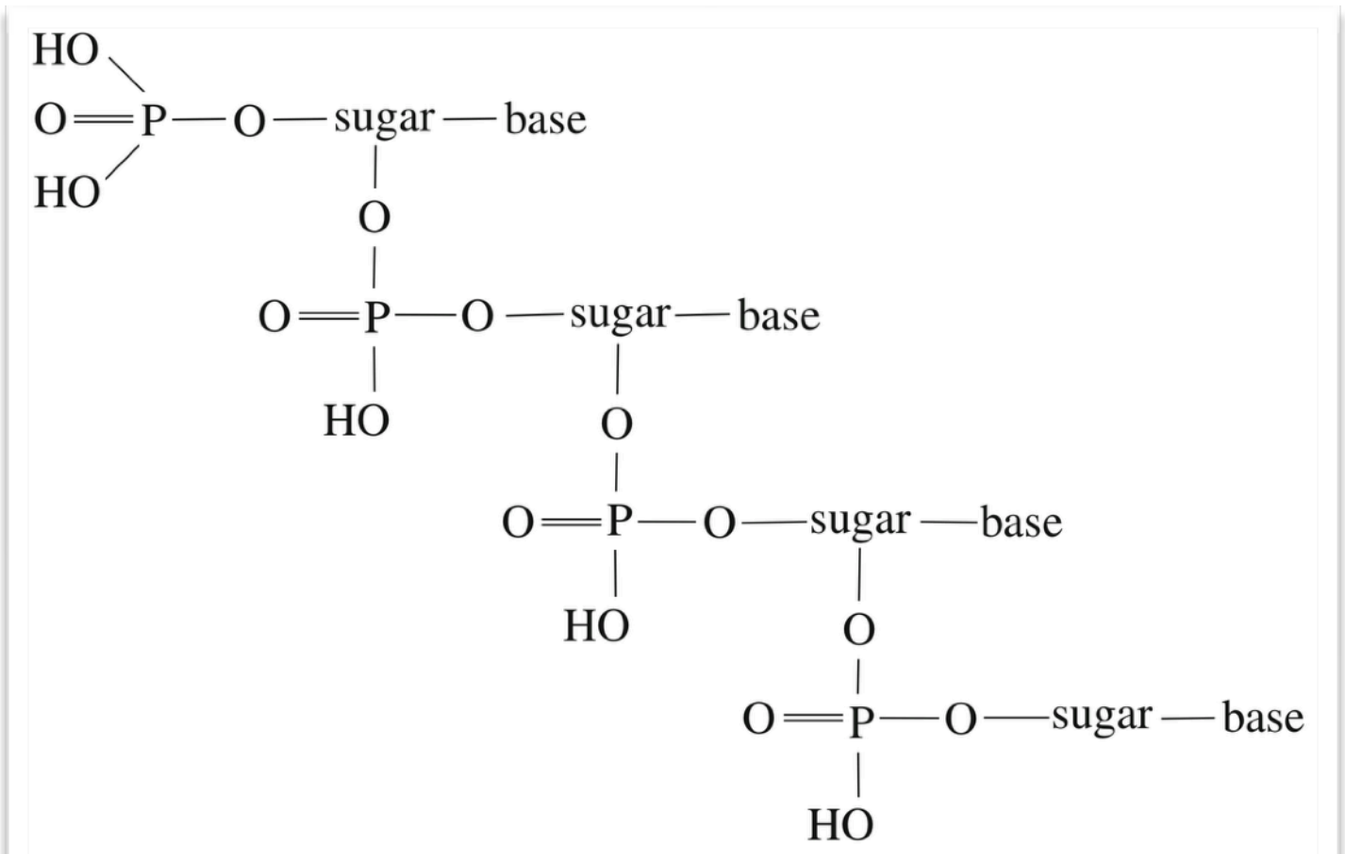


Fig. 6. *Crossing-over*. Left: the two homologous chromosomes in contact. Right: after exchange and separation.

Schrödinger. (1944) What is Life?.

What was known about the *hereditary material* in the 1940s?

- The **Tetranucleotide hypothesis**:
  - DNA was a sequence of four repeating nucleotides that provided structural support to chromosomes
  - Genes were proteins contained within each locus of the chromosome



Hargittai. (2009) The tetranucleotide hypothesis: a centennial. Structural Chemistry, 20:753–756.

WHAT IS LIFE?

The Physical Aspect of the Living Cell

BY

ERWIN SCHRÖDINGER

SENIOR PROFESSOR AT THE DUBLIN INSTITUTE FOR ADVANCED STUDIES

Based on Lectures delivered under the auspices of the Institute at Trinity College, Dublin, in February 1943

Schrödinger: What is Life? (1944)

Chromosomes: *The Hereditary Code-script*

"How can we, from the point of view of statistical physics, reconcile the facts that the gene structure seems to involve only a comparatively small number of atoms (of the order of 1,000 and possibly much less), and that value nevertheless it displays a most regular and lawful activity - with a durability or permanence that borders upon the miraculous?"

"The gene has been kept at a temperature around 98°F during all that time. How are we to understand that it has remained unperturbed by the disordering tendency of the heat motion for centuries?"

"***These material structures can only be molecules.***"

Schrödinger: What is Life? (1944)

"We shall assume the structure of a gene to be that of a huge molecule, capable only of discontinuous change, which consists in a rearrangement of the atoms and leads to an ***isomeric molecule****. The rearrangement may affect only a small region of the gene, and a* ***vast number of different rearrangements may be possible****. The energy thresholds, separating the actual configuration from any possible isomeric ones, have to be high enough (compared with the average heat energy of an atom) to make the change-over a rare event.* ***These rare events we shall identify with spontaneous mutations****."*

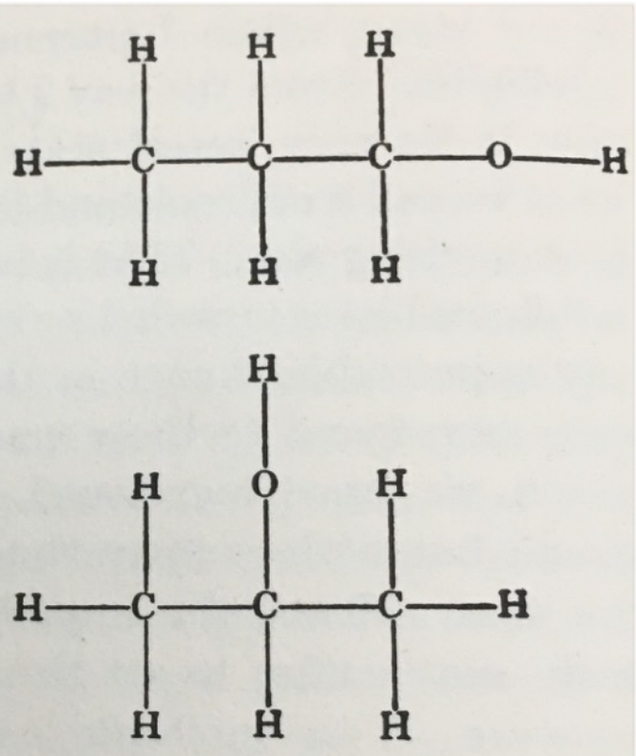


Fig. 11. The two isomeres of propyl-alcohol.

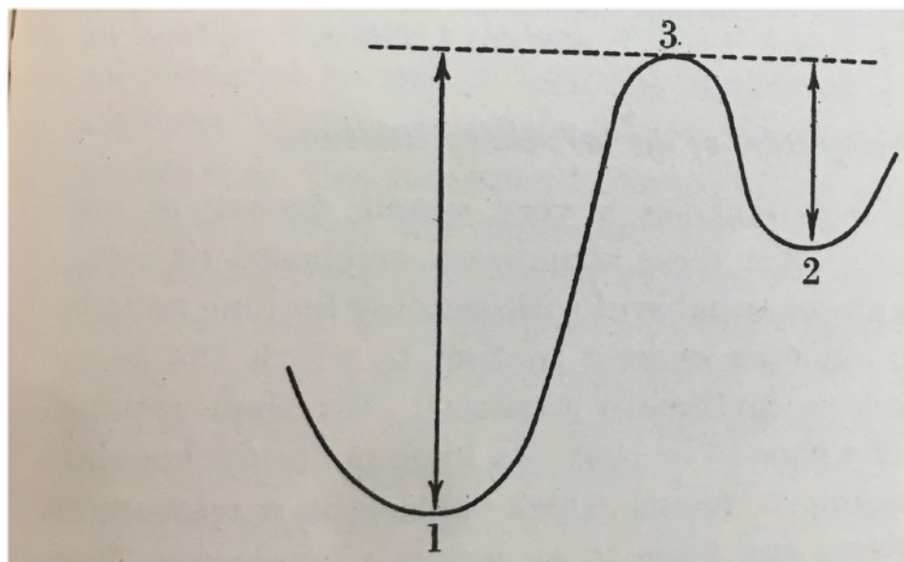


Fig. 12. Energy threshold (3) between the isomeric levels (1) and (2). The arrows indicate the minimum energies required for transition.

Schrödinger: What is Life? (1944)

*"[T]hink of the **Morse code**. The two different signs of dot and dash in well-ordered groups of not more than four allow thirty different specifications. Now, if you allowed yourself the use of a third sign, in addition to dot and dash, and used groups of not more than ten, you could form 88,572 different 'letters'."*

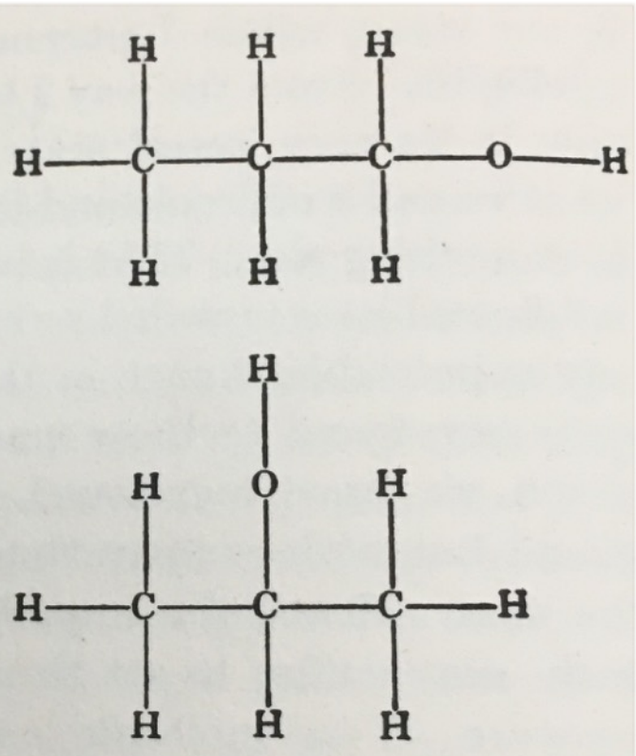


Fig. 11. The two isomeres of propyl-alcohol.

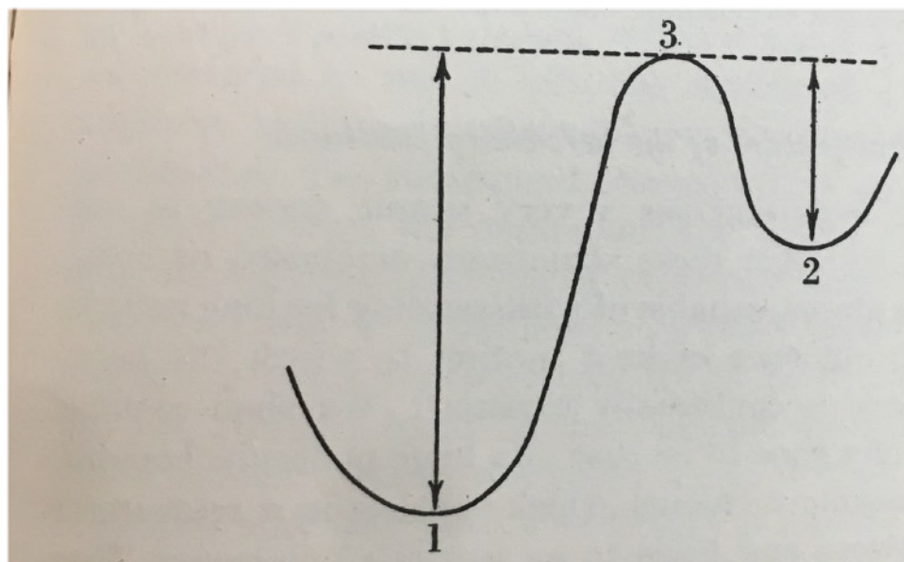
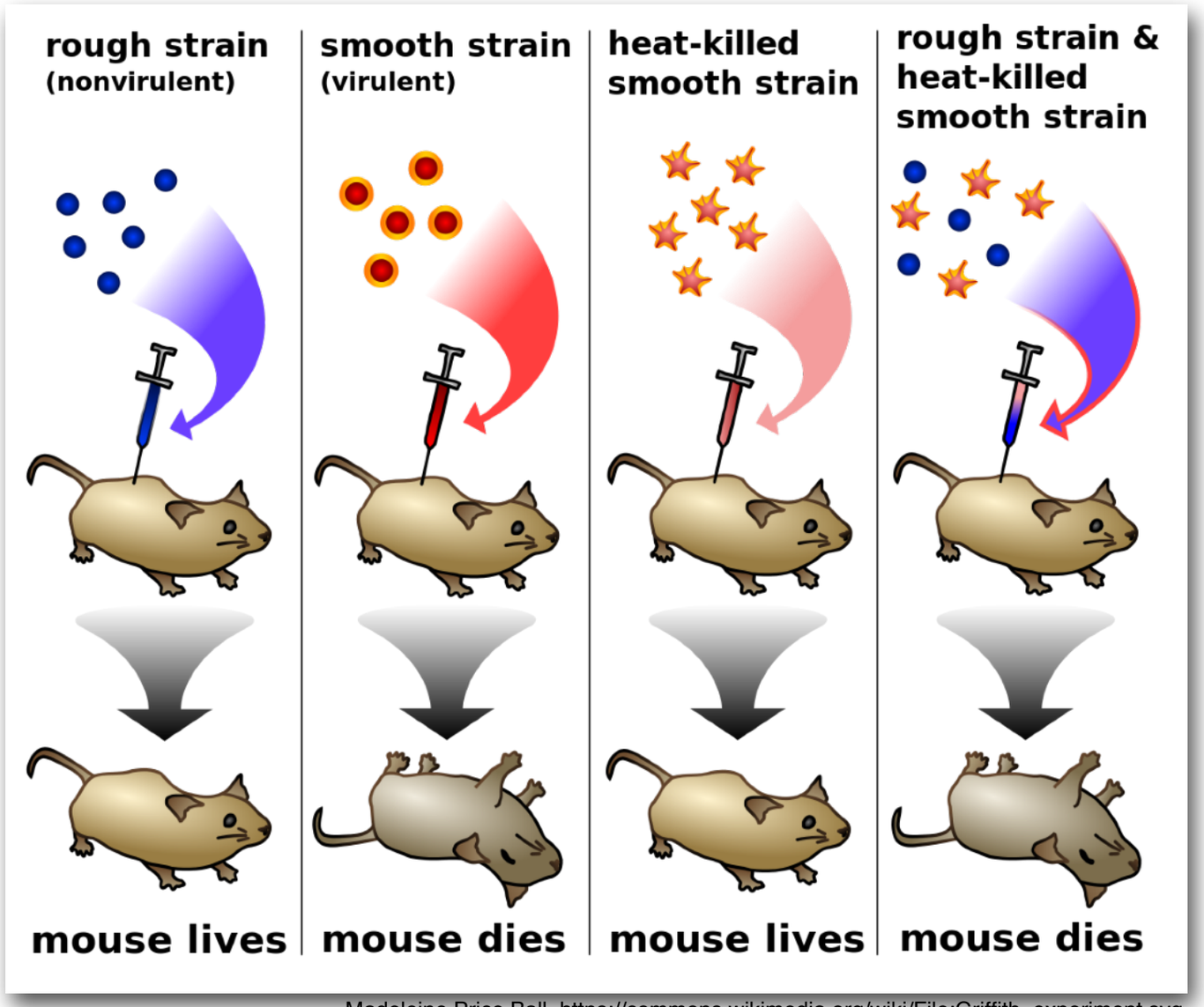


Fig. 12. Energy threshold (3) between the isomeric levels (1) and (2). The arrows indicate the minimum energies required for transition.

1944: DNA is responsible for the *Transforming Principle*

Griffith's experiment (1928):

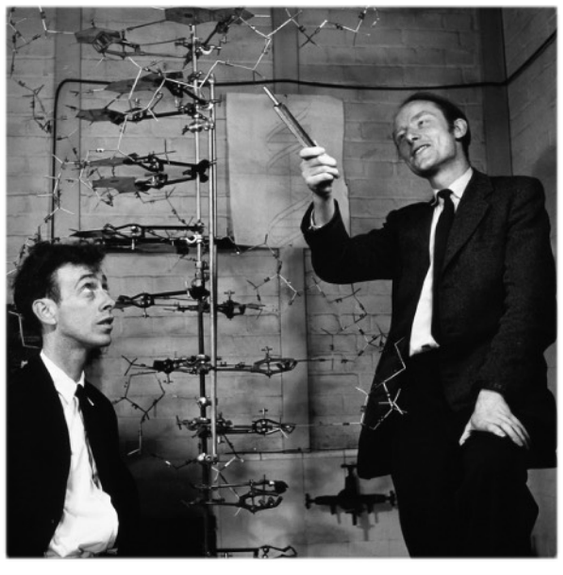


Madeleine Price Ball, https://commons.wikimedia.org/wiki/File:Griffith_experiment.svg

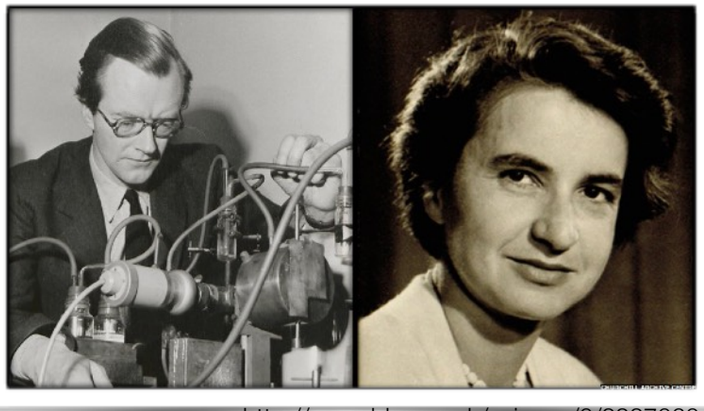
Avery, MacLeod, and McCarty (1944) showed that DNA was the transforming substance.

1953: the structure of DNA is elucidated

- Watson and Crick solve the 3-dimensional structure of DNA by model building
- Relied on the crystallographic data of Franklin and Watkins
- Intuitively provided a model of how DNA can be read and replicated



http://www.thehistoryblog.com/archives/25193



http://www.bbc.co.uk/science/0/22270604

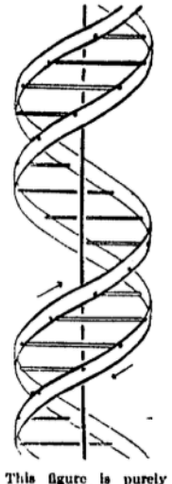
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate–sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the *z*-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical *z*-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

Steps on the road to sequencing

- **1949-1955:** Sanger and colleagues sequence the first protein molecule: insulin
- **1959:** Sinsheimer purifies the first homeogenous DNA
  - Bacteriophage ΦX174 aka PhiX, 5000bp circular chromosome
- **1965:** First alanine tRNA molecule sequenced
  - It required five people working three years with one gram of pure material (isolated from 140 kg of yeast) to determine 76 nucleotides
- **1970:** First discovery of type II restriction enzymes
  - Cleaved DNA at specific 4-6bp sequence - first general method to cleave DNA
- **1971:** First DNA sequence determined (12bp!) - phage lambda
- **1972:** sequencing of the first gene from RNA by Walter Fiers (*E. coli* lac operon)
- **1976:** sequencing of the first complete genome by Fiers (Bacteriophage MS2 which infects *E.coli*)
- **1975:** Sanger first publishes his plus/minus method of sequencing

Hutchison. (2007) DNA sequencing: bench to bedside and beyond. Nucl. Acids Res. 35(18): 6227-6237

Sanger '*Plus and Minus*' Sequencing

- Used *E. coli* DNA polymerase I and DNA polymerase from bacteriophage T4 to synthesize sequence
- Round 1: Generate variable length molecules by synthesis in poor (variable) conditions.
- Round 2: split reactions into 'plus' and 'minus' vessels
  - *Plus* vessel gets a single nucleotide
  - *minus* vessels get the other three nucleotides
- Run the eight *plus* and *minus* reactions on the same gel in adjacent lanes
- Molecules that were 1bp different could be identified in the adjacent gel lanes.
- ***'Plus and minus'* sequencing is unable to handle homopolymers runs**

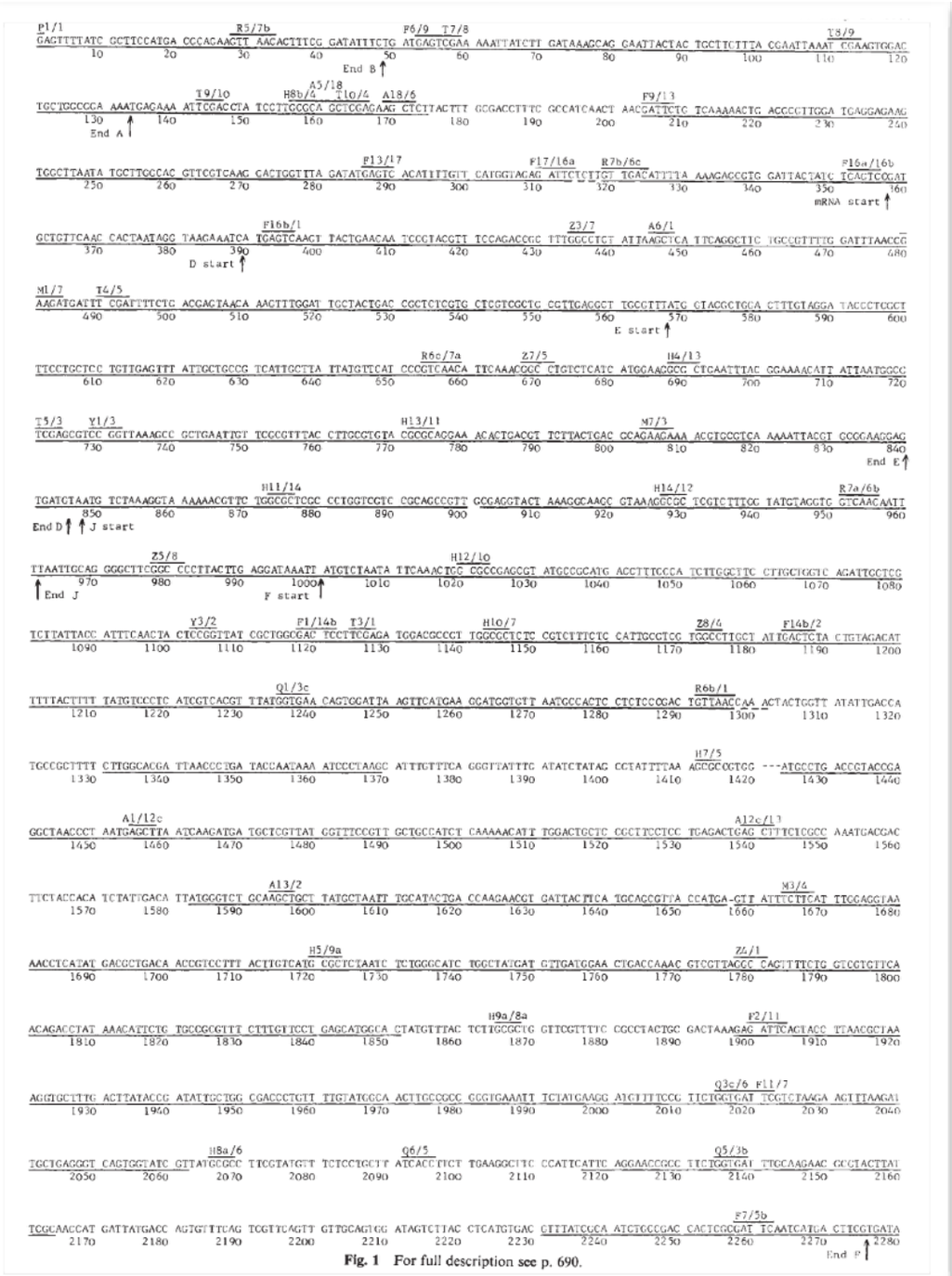
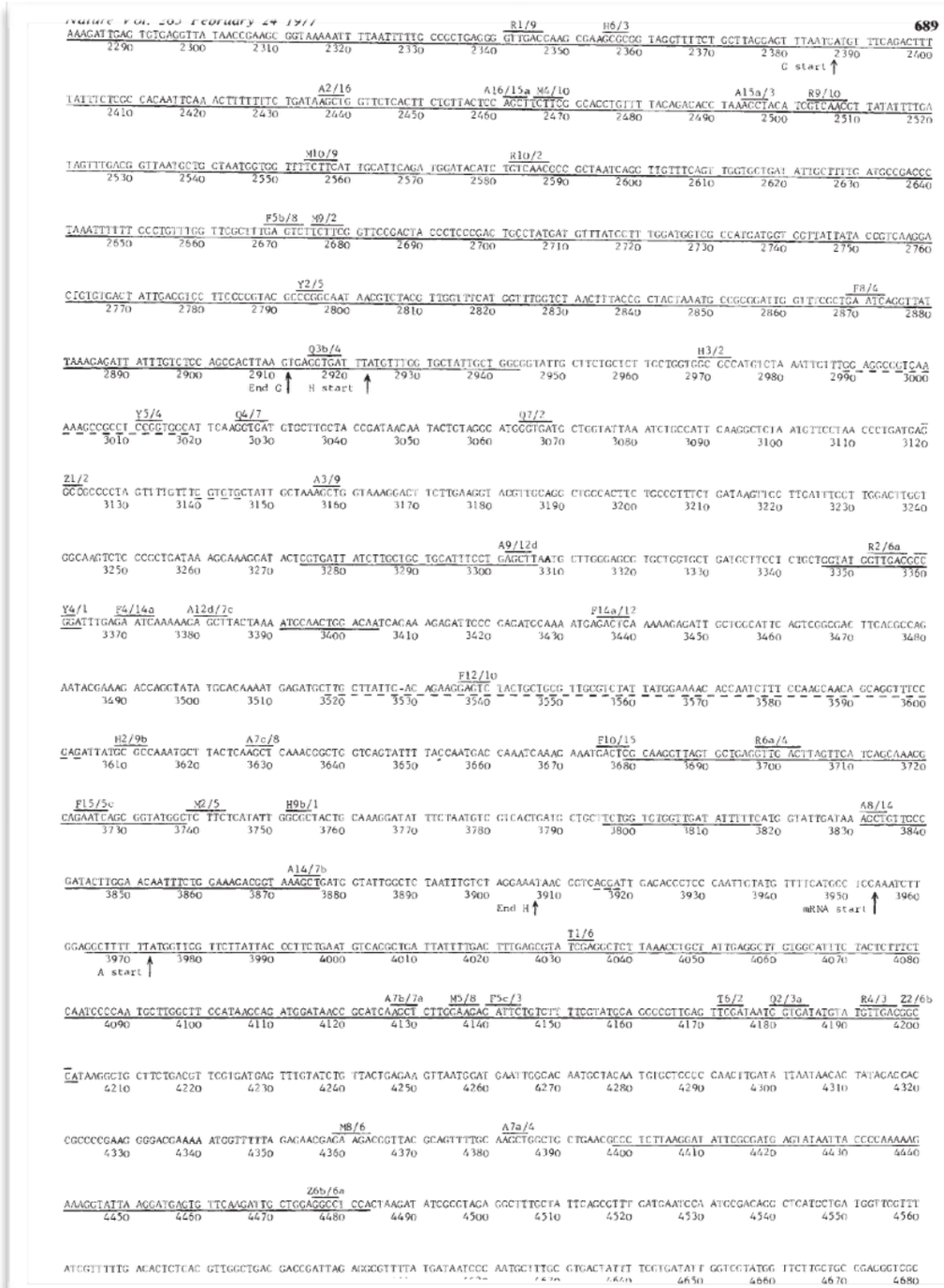
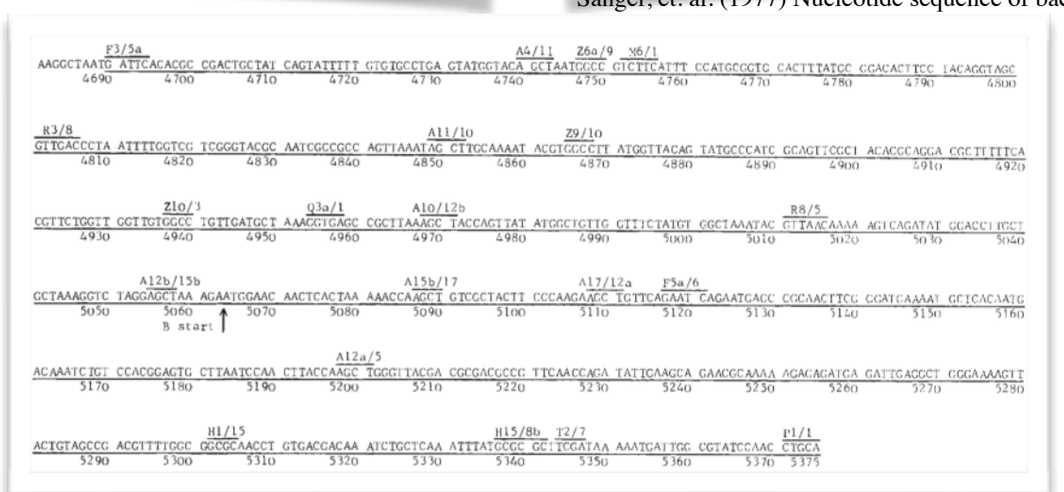


Fig. 1 For full description see p. 690.



Sanger, et. al. (1977) Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature* 265(5596): 687-695.



ΦX Genome 1977

15 Years from structure to sequencing

Why did it take so long?

- The chemical properties of different DNA molecules were so similar that it appeared difficult to separate them
- The chain length of naturally occurring DNA molecules is much greater than for proteins and made complete sequencing seem unapproachable.
- The 20 amino acid residues found in proteins have widely varying properties that had proven useful in the separation of peptides.
  - ✦ Only four bases in DNA made sequencing a more difficult problem for DNA than for protein.
- No base-specific DNAases were known.
  - ✦ Protein sequencing had depended upon proteases that cleave adjacent to certain amino acids
- DNA was considered boring compared to proteins

Hutchison. (2007) DNA sequencing: bench to bedside and beyond. Nucl. Acids Res. 35(18): 6227-6237

Maxam-Gilbert Sequencing

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 2, pp. 560–564, February 1977
Biochemistry

A new method for sequencing DNA

(DNA chemistry/dimethyl sulfate cleavage/hydrazine/piperidine)

ALLAN M. MAXAM AND WALTER GILBERT

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138

Contributed by Walter Gilbert, December 9, 1976

ABSTRACT **DNA can be sequenced by a chemical procedure that breaks a terminally labeled DNA molecule partially at each repetition of a base. The lengths of the labeled fragments then identify the positions of that base. We describe reactions that cleave DNA preferentially at guanines, at adenines, at cytosines and thymines equally, and at cytosines alone. When the products of these four reactions are resolved by size, by electrophoresis on a polyacrylamide gel, the DNA sequence can be read from the pattern of radioactive bands. The technique will permit sequencing of at least 100 bases from the point of labeling.**

We have developed a new technique for sequencing DNA molecules. The procedure determines the nucleotide sequence of a terminally labeled DNA molecule by breaking it at adenine, guanine, cytosine, or thymine with chemical agents. Partial cleavage at each base produces a nested set of radioactive fragments extending from the labeled end to each of the positions of that base. Polyacrylamide gel electrophoresis resolves these single-stranded fragments; their sizes reveal *in order* the points of breakage. The autoradiograph of a gel produced from four different chemical cleavages, each specific for a base in a sense we will describe, then shows a pattern of bands from which the sequence can be read directly. The method is limited

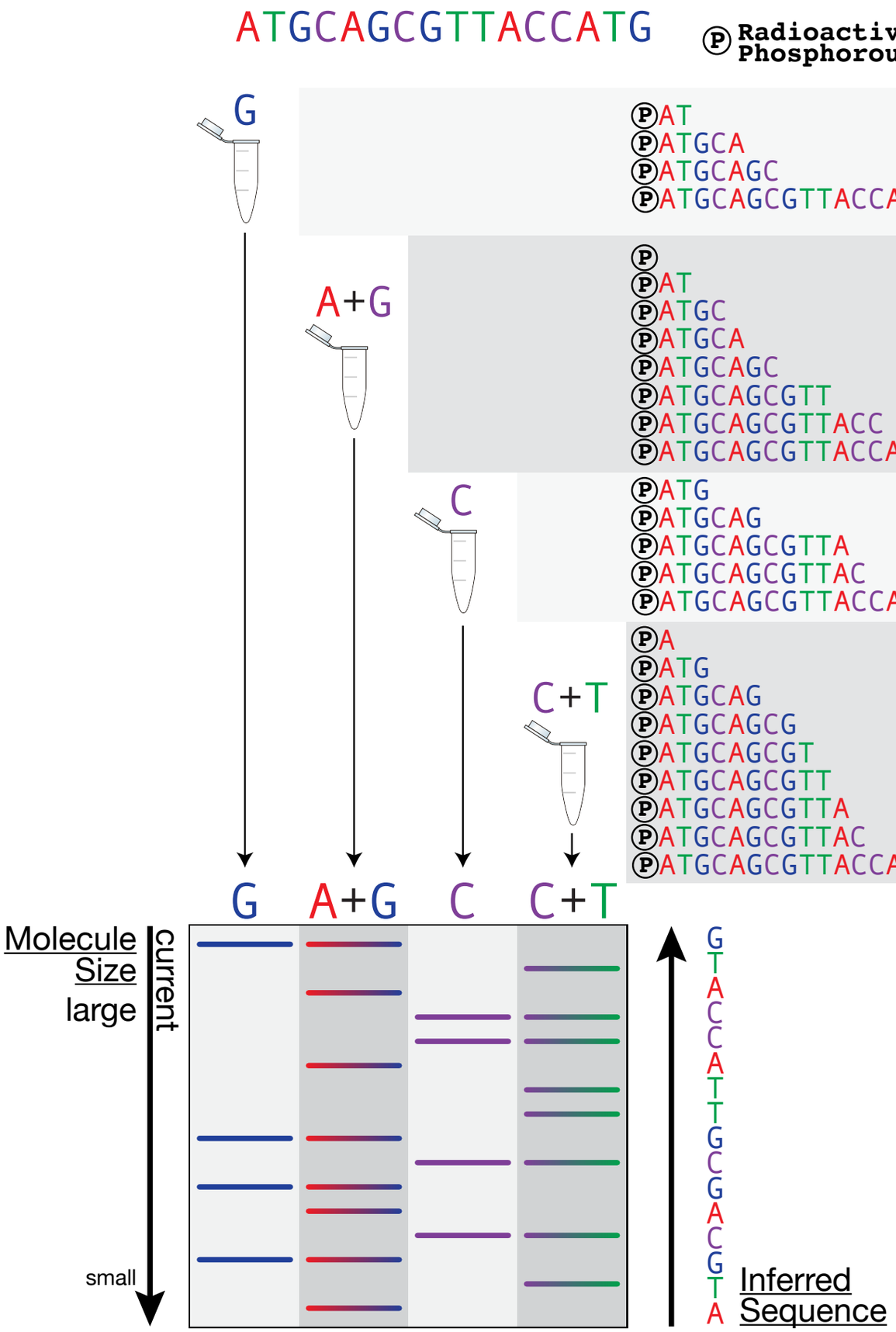
THE SPECIFIC CHEMISTRY

A Guanine/Adenine Cleavage (2). Dimethyl sulfate methylates the guanines in DNA at the N7 position and the adenines at the N3 (3). The glycosidic bond of a methylated purine is unstable (3, 4) and breaks easily on heating at neutral pH, leaving the sugar free. Treatment with 0.1 M alkali at 90° then will cleave the sugar from the neighboring phosphate groups. When the resulting end-labeled fragments are resolved on a polyacrylamide gel, the autoradiograph contains a pattern of dark and light bands. The dark bands arise from breakage at guanines, which methylate 5-fold faster than adenines (3).

This strong guanine/weak adenine pattern contains almost half the information necessary for sequencing; however, ambiguities can arise in the interpretation of this pattern because the intensity of isolated bands is not easy to assess. To determine the bases we compare the information contained in this column of the gel with that in a parallel column in which the breakage at the guanines is suppressed, leaving the adenines apparently enhanced.

An Adenine-Enhanced Cleavage. The glycosidic bond of methylated adenosine is less stable than that of methylated guanosine (4); thus, gentle treatment with dilute acid releases

Maxam-Gilbert Sequencing



Sequencing by Chain Breakage

- **Thousands of** copies of the genomic sequence in each reaction vessel
- Chemicals **cleave** at particular nucleotides
- Chemical concentration is controlled to create, on average, one break per molecule
- Sequencing occurs **simultaneously** at all bases — one per molecule
- Highly toxic and dangerous reactions

Sanger Sequencing

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

F. SANGER, S. NICKLEN, AND A. R. COULSON

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977

ABSTRACT **A new method for determining nucleotide sequences in DNA is described. It is similar to the "plus and minus" method [Sanger, F. & Coulson, A. R. (1975) *J. Mol. Biol.* 94, 441–448] but makes use of the 2′,3′-dideoxy and arabinonucleoside analogues of the normal deoxynucleoside triphosphates, which act as specific chain-terminating inhibitors of DNA polymerase. The technique has been applied to the DNA of bacteriophage ϕX174 and is more rapid and more accurate than either the plus or the minus method.**

The "plus and minus" method (1) is a relatively rapid and simple technique that has made possible the determination of the sequence of the genome of bacteriophage ϕX174 (2). It depends on the use of DNA polymerase to transcribe specific regions of the DNA under controlled conditions. Although the method is considerably more rapid and simple than other available techniques, neither the "plus" nor the "minus" method is completely accurate, and in order to establish a sequence both must be used together, and sometimes confirmatory data are necessary. W. M. Barnes (*J. Mol. Biol.*, in press) has recently developed a third method, involving ribo-substitution, which has certain advantages over the plus and minus method, but this has not yet been extensively exploited.

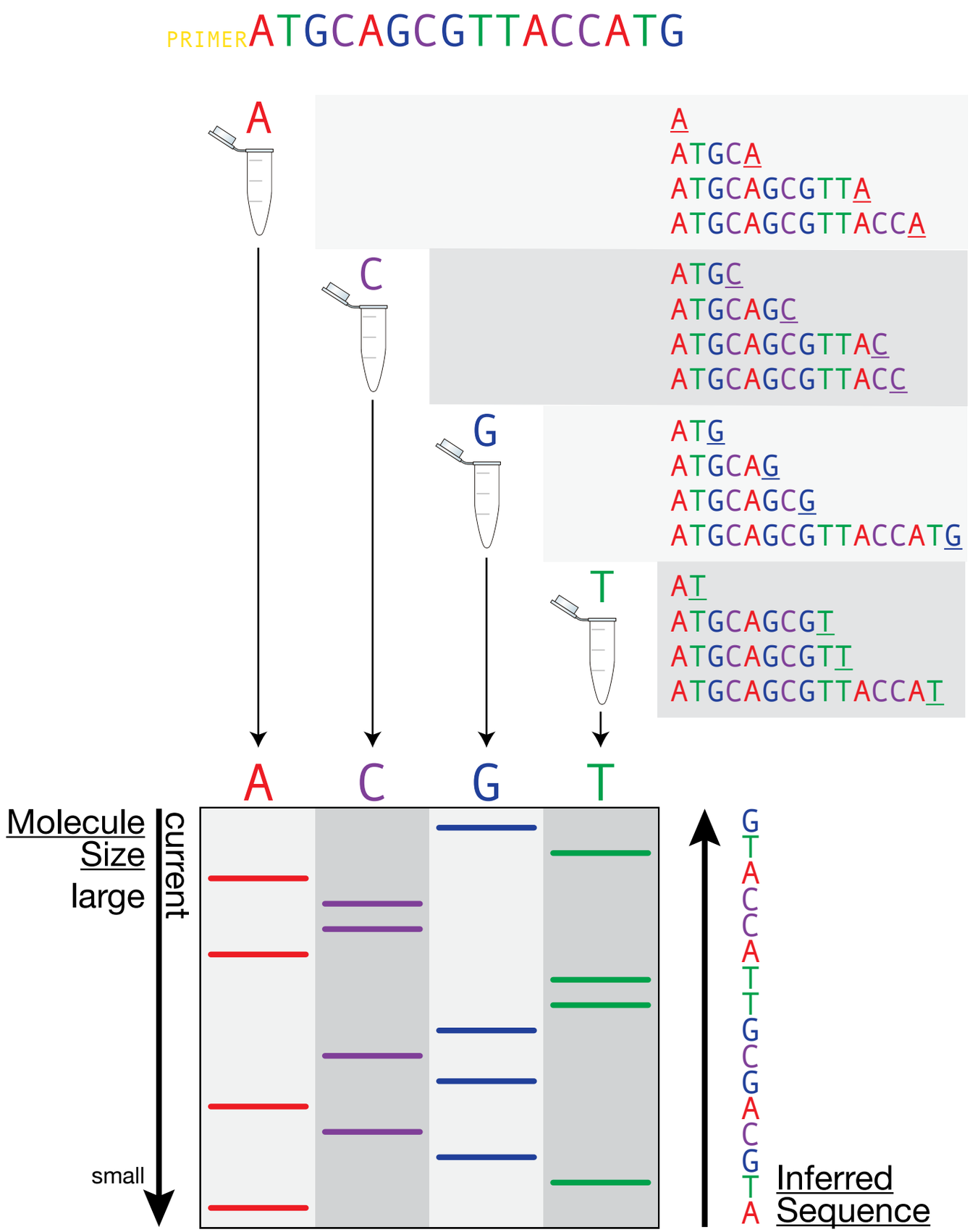
Another rapid and simple method that depends on specific chemical degradation of the DNA has recently been described by Maxam and Gilbert (3), and this has also been used extensively for DNA sequencing. It has the advantage over the plus and minus method that it can be applied to double-stranded DNA, but it requires a strand separation or equivalent fractionation of each restriction enzyme fragment studied, which makes it somewhat more laborious.

a stereoisomer of ribose in which the 3′-hydroxyl group is oriented in *trans* position with respect to the 2′-hydroxyl group. The arabinosyl (ara) nucleotides act as chain terminating inhibitors of *Escherichia coli* DNA polymerase I in a manner comparable to ddT (4), although synthesized chains ending in 3′ araC can be further extended by some mammalian DNA polymerases (5). In order to obtain a suitable pattern of bands from which an extensive sequence can be read it is necessary to have a ratio of terminating triphosphate to normal triphosphate such that only partial incorporation of the terminator occurs. For the dideoxy derivatives this ratio is about 100, and for the arabinosyl derivatives about 5000.

METHODS

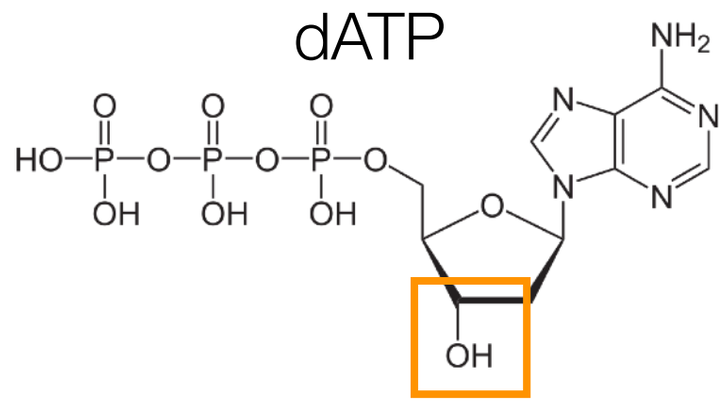
Preparation of the Triphosphate Analogues. The preparation of ddTTP has been described (6, 7), and the material is now commercially available. ddA has been prepared by McCarthy *et al.* (8). We essentially followed their procedure and used the methods of Tener (9) and of Hoard and Ott (10) to convert it to the triphosphate, which was then purified on DEAE-Sephadex, using a 0.1–1.0 M gradient of triethylamine carbonate at pH 8.4. The preparation of ddGTP and ddCTP has not been described previously; however we applied the same method as that used for ddATP and obtained solutions having the requisite terminating activities. The yields were very low and this can hardly be regarded as adequate chemical characterization. However, there can be little doubt that the activity was due to the dideoxy derivatives.

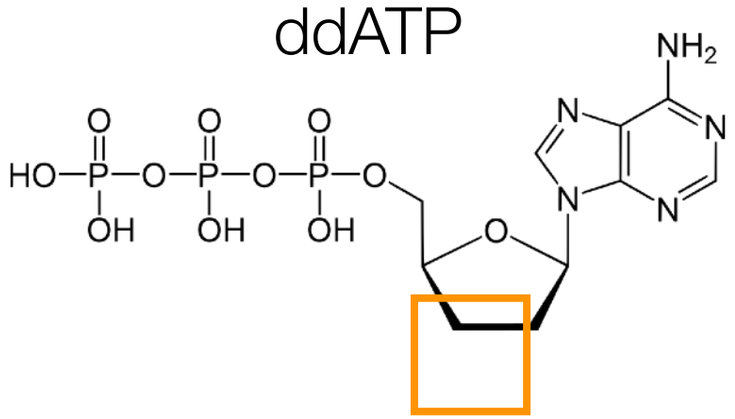
Sanger Sequencing



Sequencing by Synthesis (SBS)

- Reactions are **catalyzed** with a polymerase
- **Thousands of** copies of the genomic sequence in each reaction vessel
- Sequencing occurs **simultaneously** at all bases — one per molecule
- a **tiny fraction** of ddNTPs are mixed proportionately with dNTPs

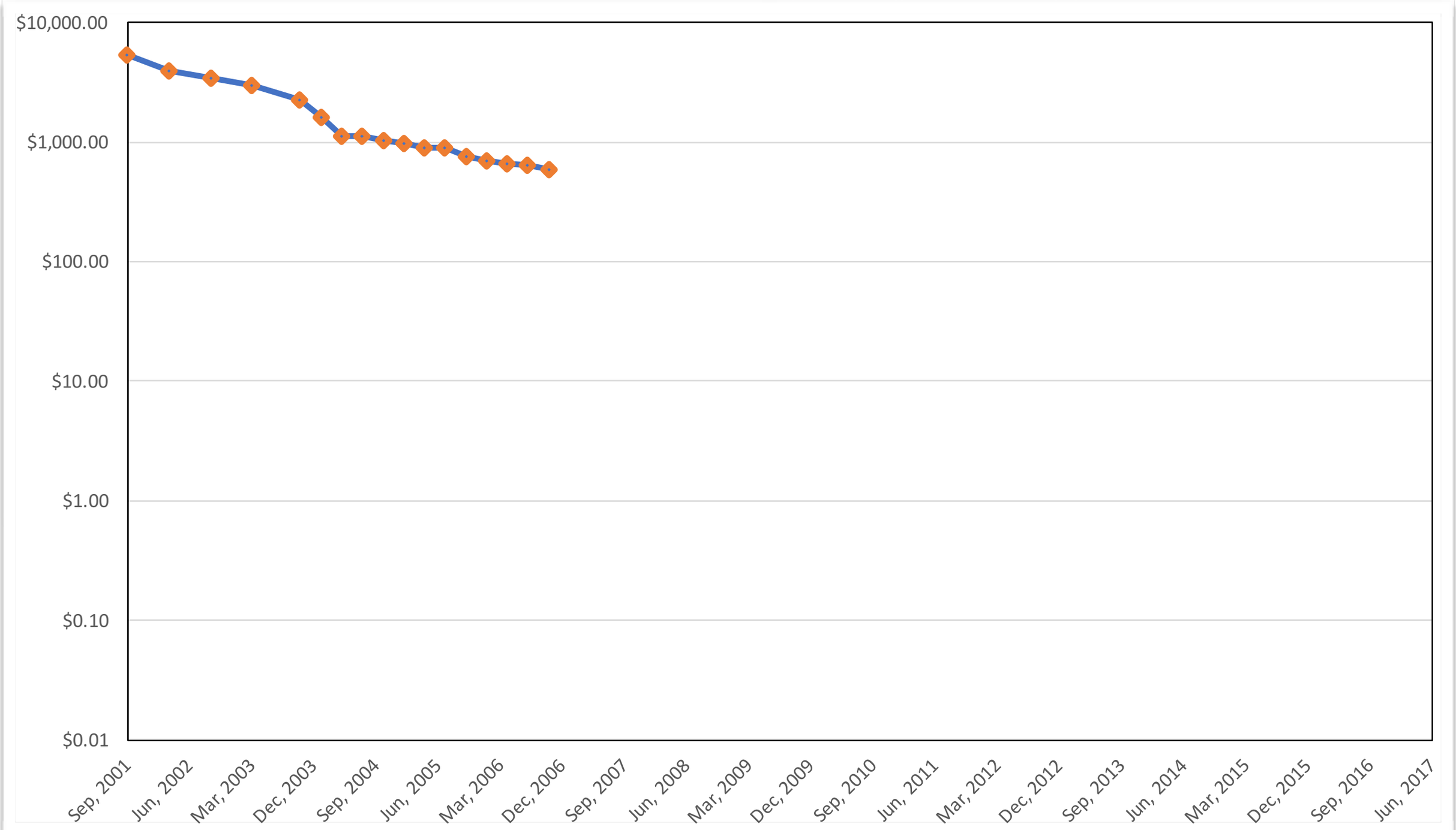




Automating Sanger Sequencing

- Moved from radio labeled DNA to fluorescent-labeled DNA
- Moved from a single fluorescent label to four different colored labels — enabled move from a four lane gel to a single lane gel
- Gels were replaced by capillaries
- PHRED scores were introduced so computers could determine florescence
- The ABI PRISM 377 could sequence 96 molecules per run, two runs per day.

Sequencing Costs after Sanger



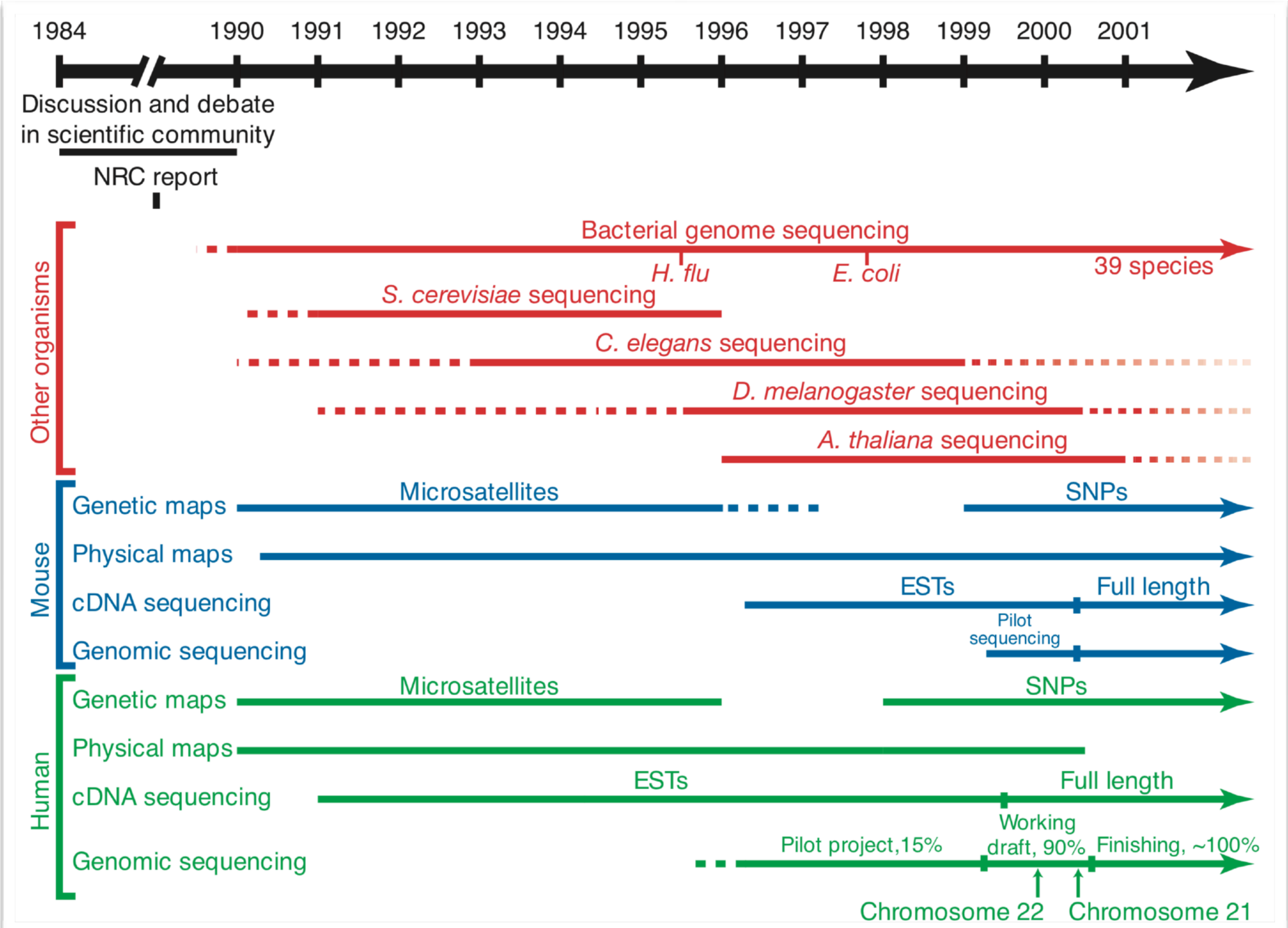
https://www.genome.gov/sequencingcostsdata/

Human Genome Project

Human Genome Project

- One of the largest scientific endeavors
  - Target accuracy 1:10,000 bases
  - Started in 1990 by U.S. DOE and NIH
  - $3Billion and 15 years
- Goal was to identify 25K genes and 3 billion bases
- Used the Sanger sequencing method
- Draft assembly done in 2000, complete genome by 2003, last chromosome published in 2006

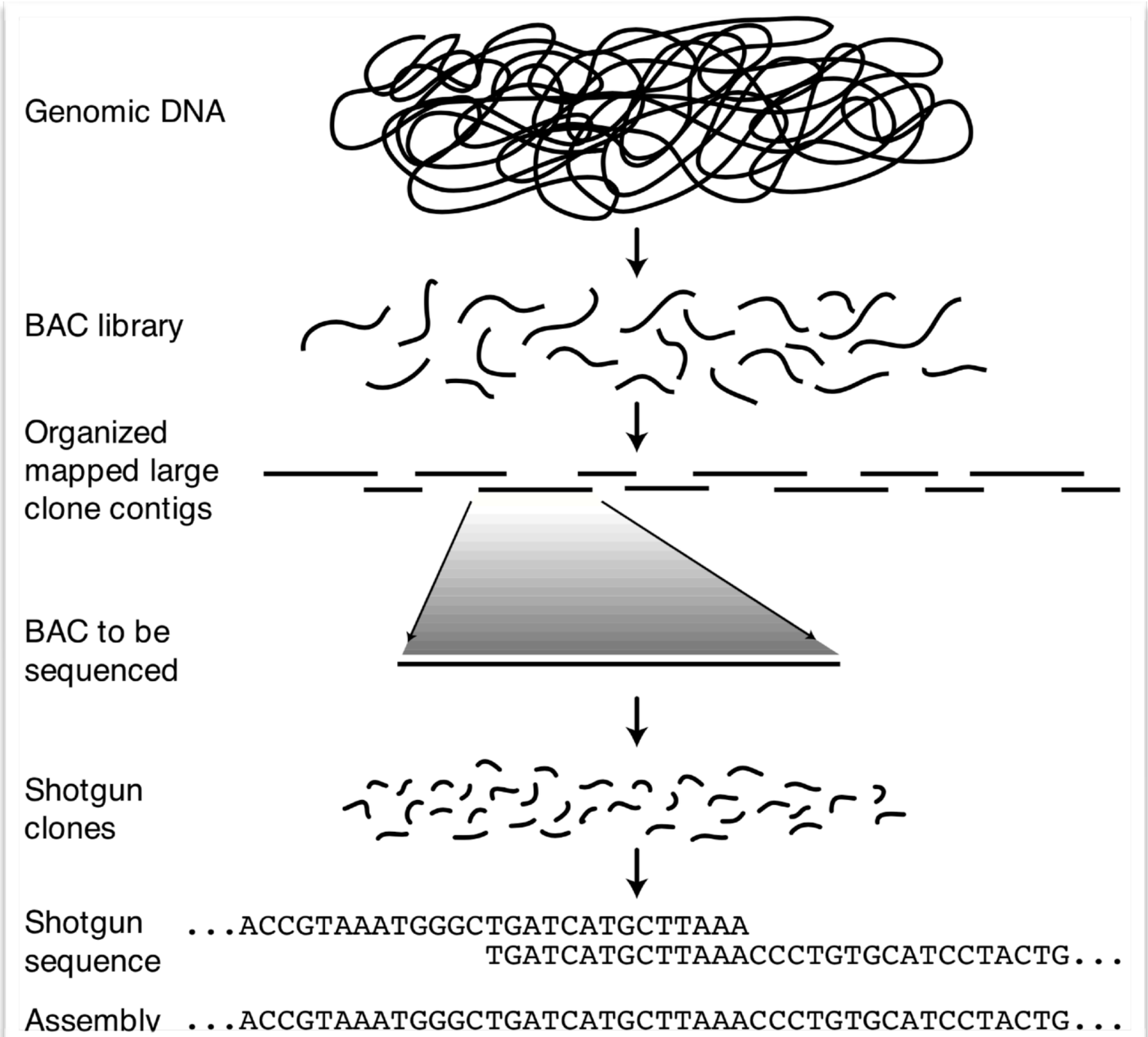
Human Genome Project



International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature. 409:860–921

Human Genome Project

Hierarchical shotgun sequencing



International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature. 409:860–921

Human Genome Project
Celera Attempts Privatization

- Started in Sept 1999, goal was to do in $300M and 3 years what the public project was doing for $3B and 15 years!
- Whole-genome shotgun sequencing
  - At that time had only been used for bacterial genomes of 6Mb
- Celera sought patent protection on 200 to 300 genes
  - Eventually filed preliminary ("place-holder") patent applications on 6,500 whole or partial genes

Human Genome Project

Celera Attempts Privatization

The New York Times

Scientist's Plan: Map All DNA Within 3 Years

By **NICHOLAS WADE** MAY 10, 1998

A pioneer in genetic sequencing and a private company are joining forces with the aim of deciphering the entire DNA, or genome, of humans within three years, far faster and cheaper than the Federal Government is planning.

If successful, the venture would outstrip and to some extent make redundant the Government's $3 billion program to sequence the human genome by 2005.

Despite a host of new questions, the charting of the full human genome would offer enormous medical and scientific benefits.

The principals have high credibility in the world of genome sequencing. They are Dr. J. Craig Venter, president of the nonprofit Institute for Genomic Sciences in Rockville, Md., and Michael W. Hunkapiller, president and technical maestro of the Applied Biosystems division of the Perkin-Elmer Corporation of Norwalk, Conn.

The director of the Federal human genome project at the National Institutes of Health, Dr. Francis Collins, first heard of the new company's plan on Friday, as did the director of the N.I.H., Dr Harold Varmus. Both said that the plan, if successful, would enable them to reach a desired goal sooner.

Human Genome Project
Celera Attempts Privatization

The New York Times

Scientist's Plan: Map All DNA Within 3 Years

By NICHOLAS WADE MAY 10, 1998

A pioneer in genetic sequencing and a private company are joining forces with the aim of deciphering the entire DNA, or genome, of humans within three years, far faster and cheaper than the Federal Government is planning.

If successful, the venture would outstrip and to some extent make redundant the Government's $3 billion program to sequence the human genome by 2005.

Despite a host of new questions, the charting of the full human genome would offer enormous medical and scientific benefits.

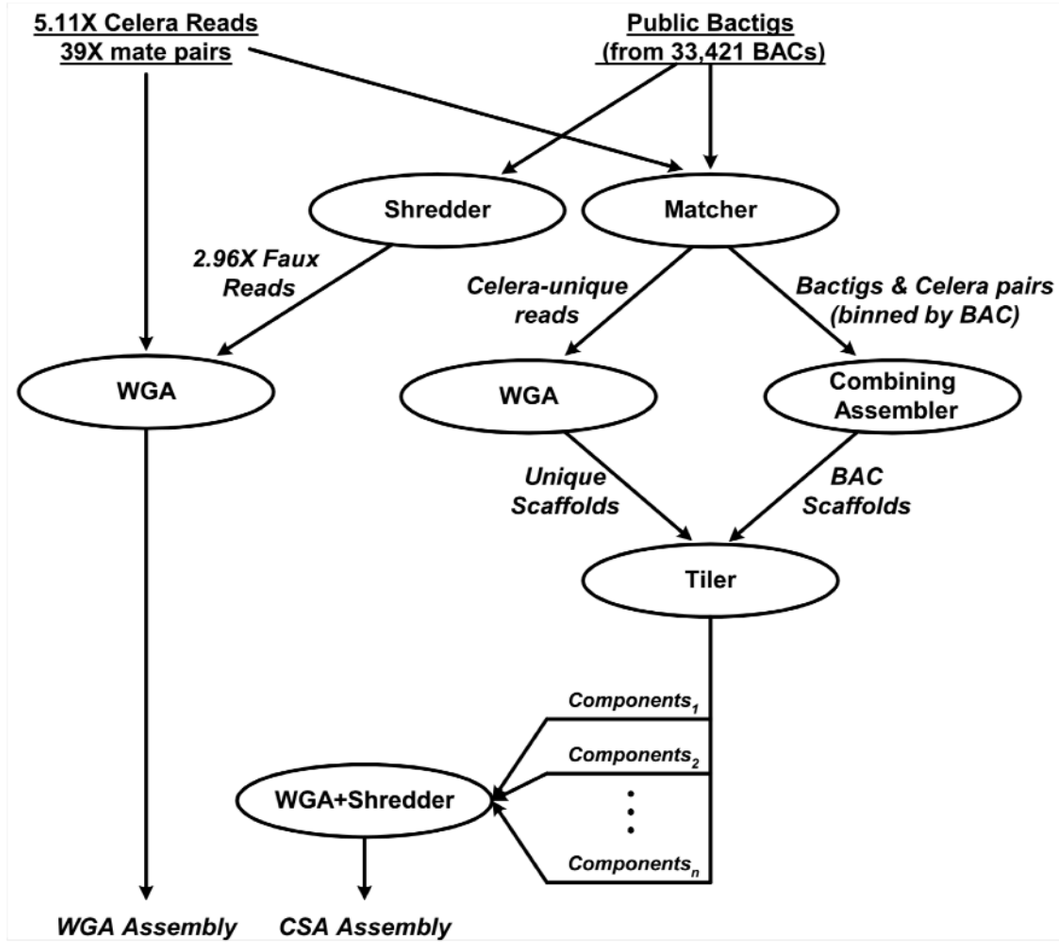
The principals have high credibility in the world of genome sequencing. They are Dr. J. Craig Venter, president of the nonprofit Institute for Genomic Sciences in Rockville, Md., and Michael W. Hunkapiller, president and technical maestro of the Applied Biosystems division of the Perkin-Elmer Corporation of Norwalk, Conn.

The director of the Federal human genome project at the National Institutes of Health, Dr. Francis Collins, first heard of the new company's plan on Friday, as did the director of the N.I.H., Dr Harold Varmus. Both said that the plan, if successful, would enable them to reach a desired goal sooner.

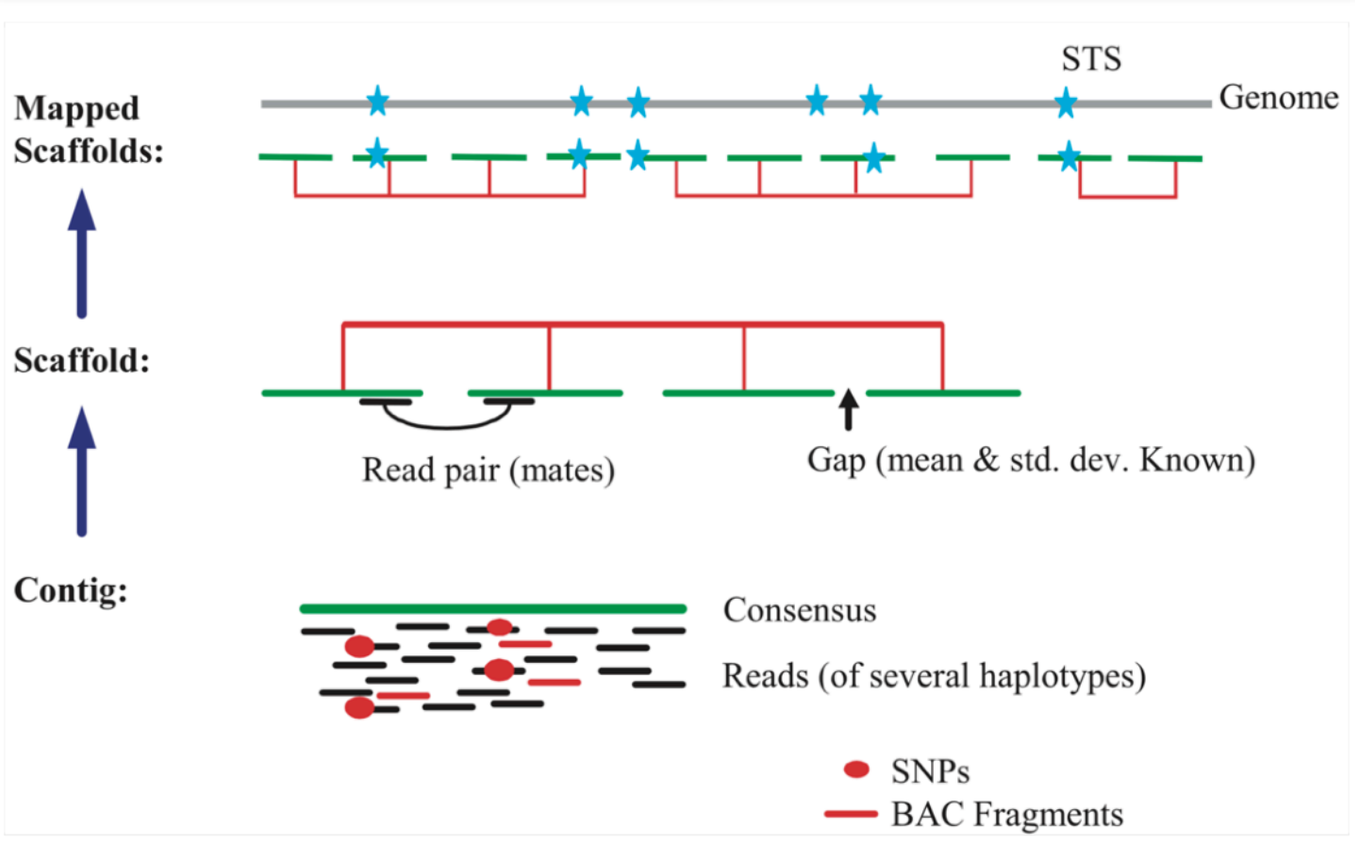
- [T]he venture would outstrip and…make redundant the Government's $3 billion program to sequence the human genome
- The director of the Federal human genome project at the National Institutes of Health, Dr. Francis Collins, first heard of the new company's plan on Friday.
  - Dr. Collins expressed confidence that they could persuade Congress to accept the need for this change in focus.
- [I]t would make possible almost overnight many developments that had been expected to unfold over the next decade.
- The possession of the entire human genome by a single private company could become an issue of public concern.
- The new venture was conceived only a few months ago. The two men concluded in January that it would be possible to sequence the three billion letters of human DNA within three years.
- Congress…might ask why it should continue to finance the human genome project…if the new company is going to finish first.
- ''It's not impossible at all that [Venter] could succeed''
- Dr. Venter and his new colleagues plan not just to sequence the human genome but to construct a ''definitive'' data base that will integrate medical and other information with the basic DNA sequence.
- The new company's data base seems likely to rival or supersede Genbank, the data bank operated by the National Institutes of Health.

Human Genome Project

Used data from Public Project

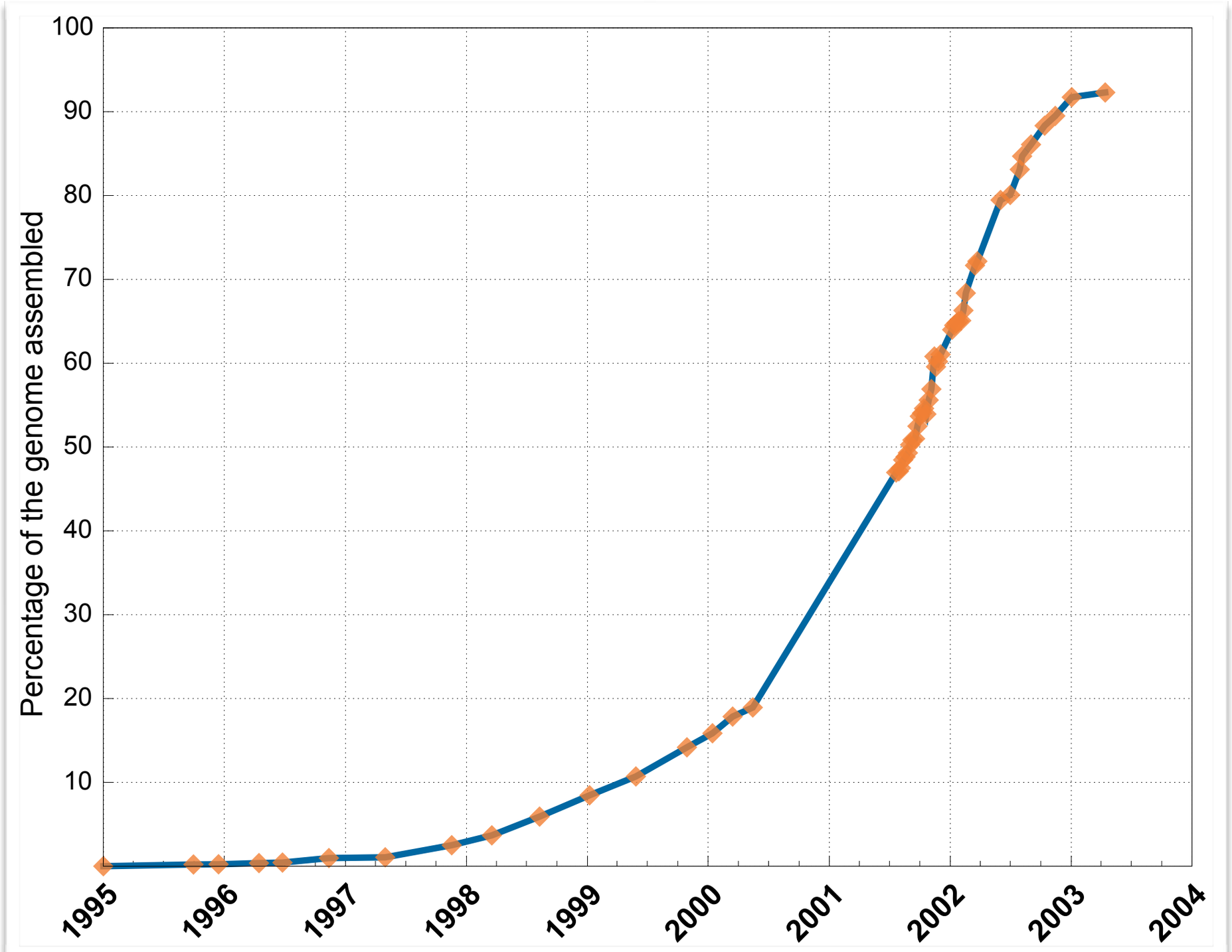


Pure Shotgun Sequencing



Venter, et al. (2001) The sequence of the human genome. Science. 291:1304–1351

Human Genome Project

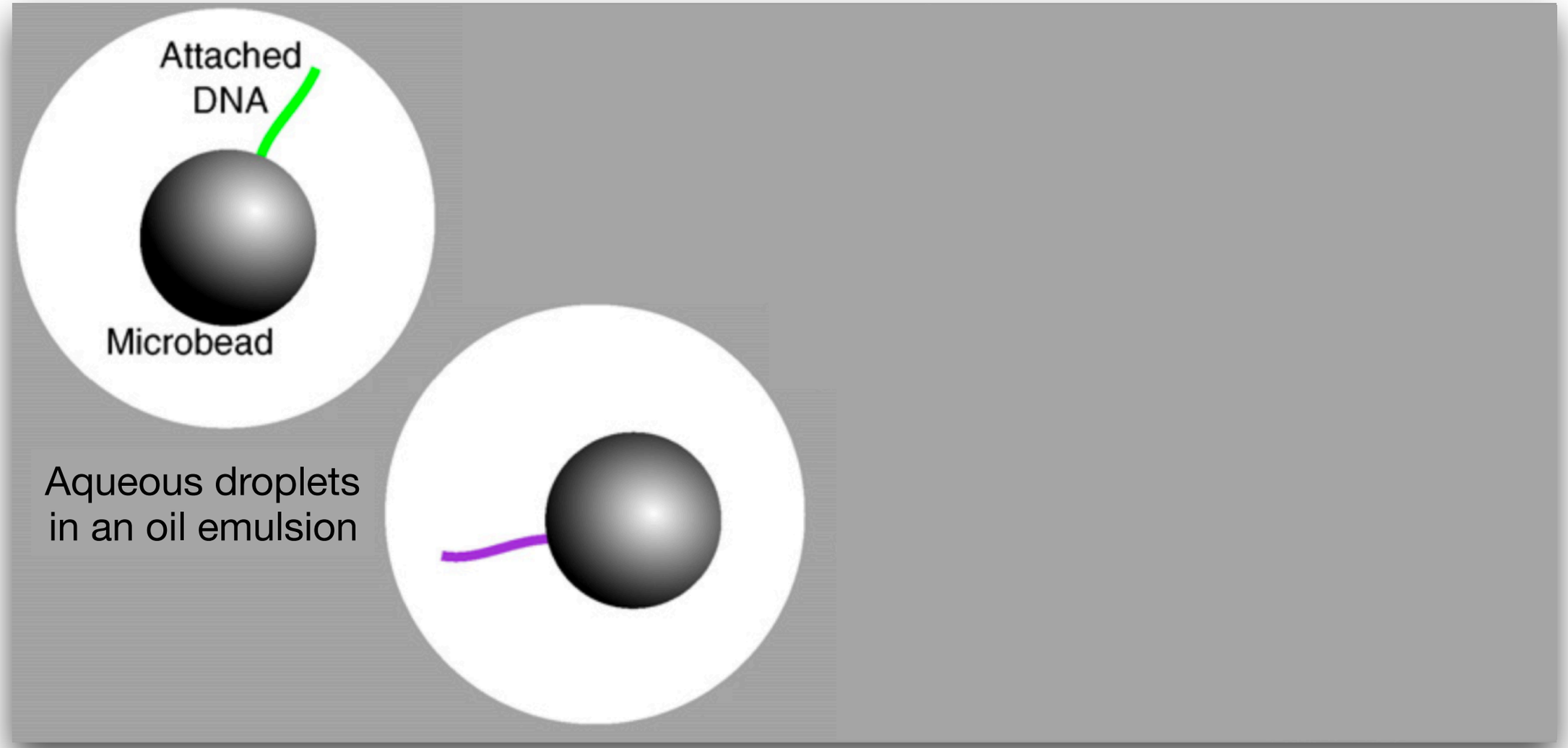


Jared Roach, http://www.strategicgenomics.com/Genome/index.htm

Second Generation Sequencing

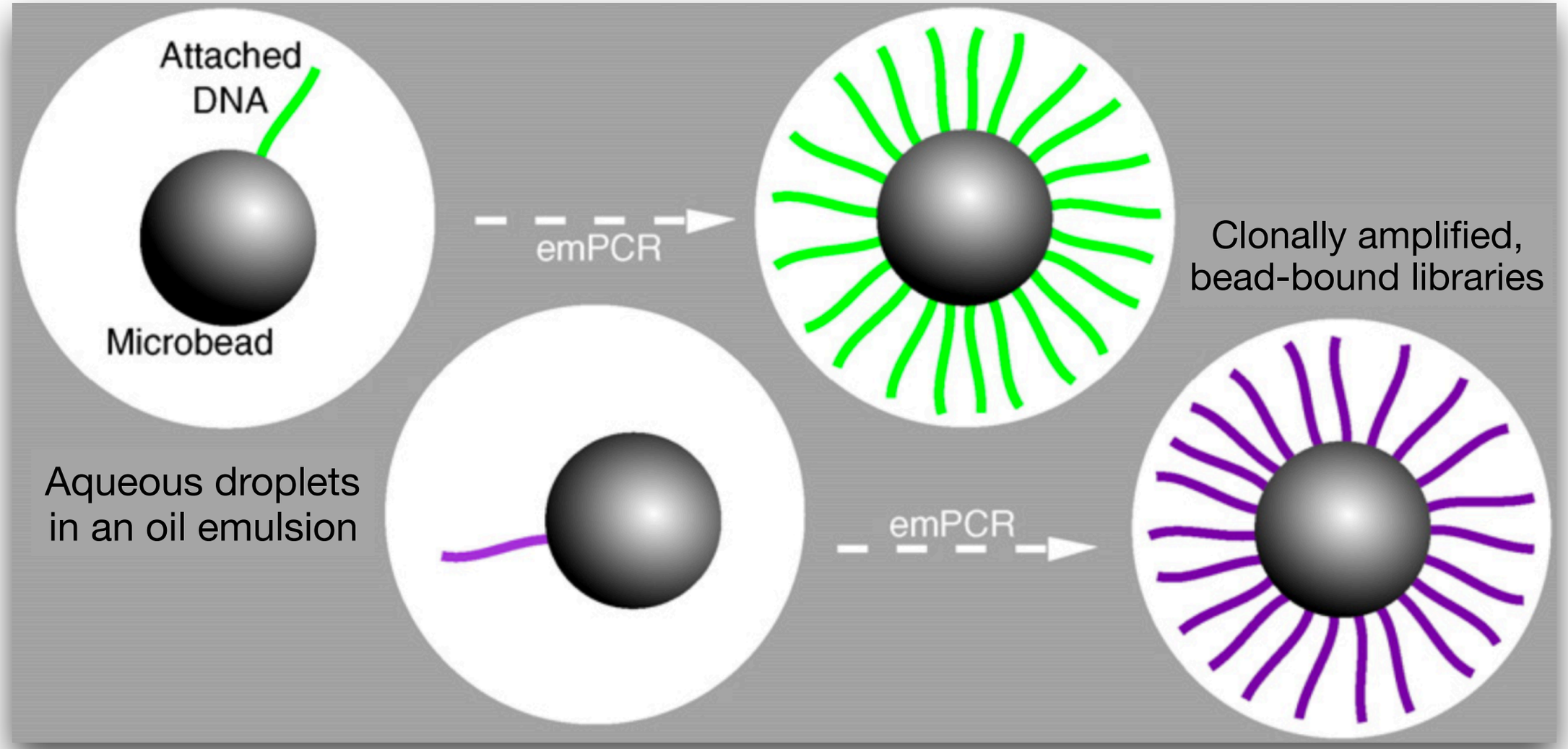
- Performing huge number of reactions on a micrometer scale
    - Enabled by improvements in microfabrication and high resolution imaging
- Sequencing is done on clusters of molecules
    - enabled with various methods of PCR amplification
        - requires the construction of libraries with adaptors
- Typically biological samples are highly multiplexed
- Read lengths are a fraction of Sanger, but data volume is massive
- Two remaining major methods after heavy competition

Second Generation Sequencing: Pyrosequencing



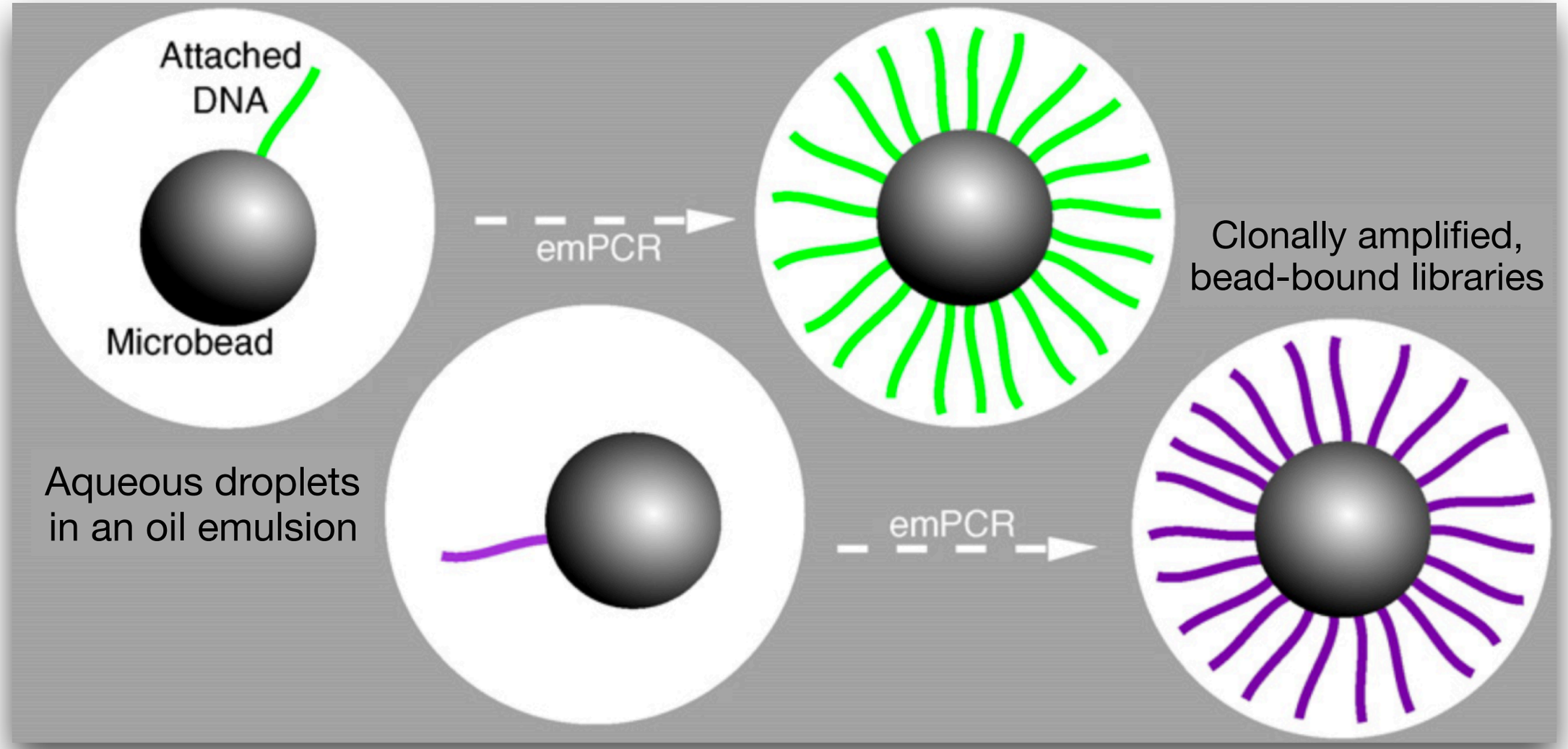
Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1–8.

Second Generation Sequencing: Pyrosequencing

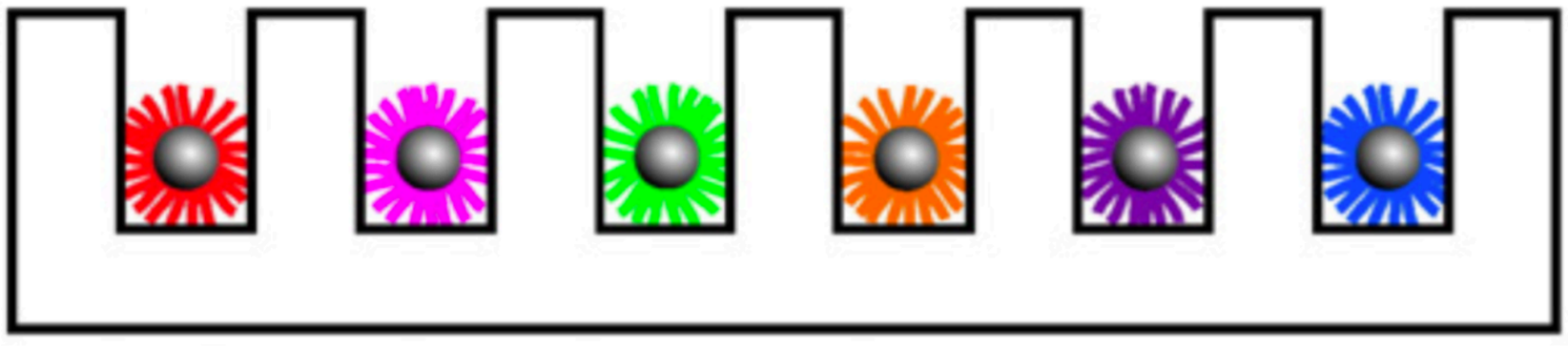


Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1–8.

Second Generation Sequencing: Pyrosequencing

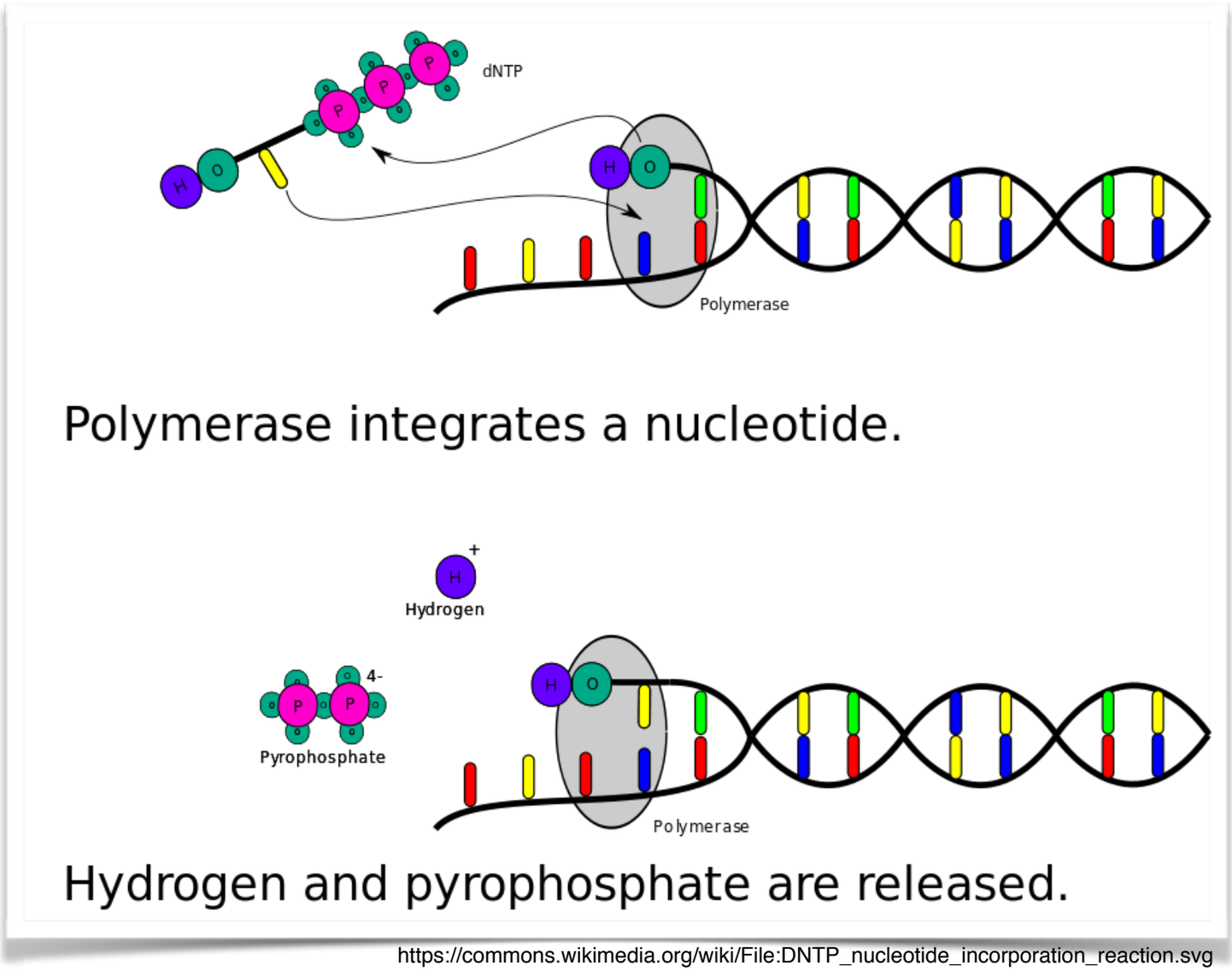


Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1–8.



Picotitre wells containing DNA-bearing microbeads

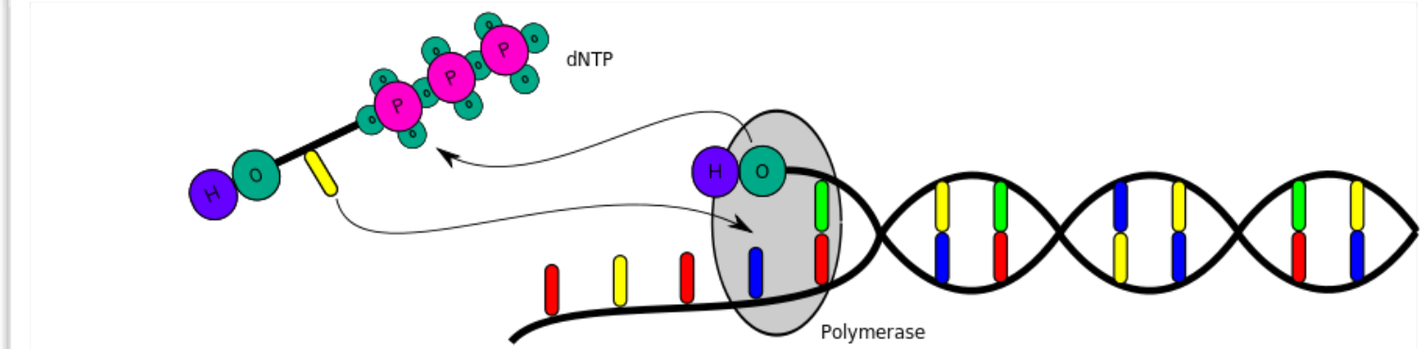
Second Generation Sequencing: Pyrosequencing



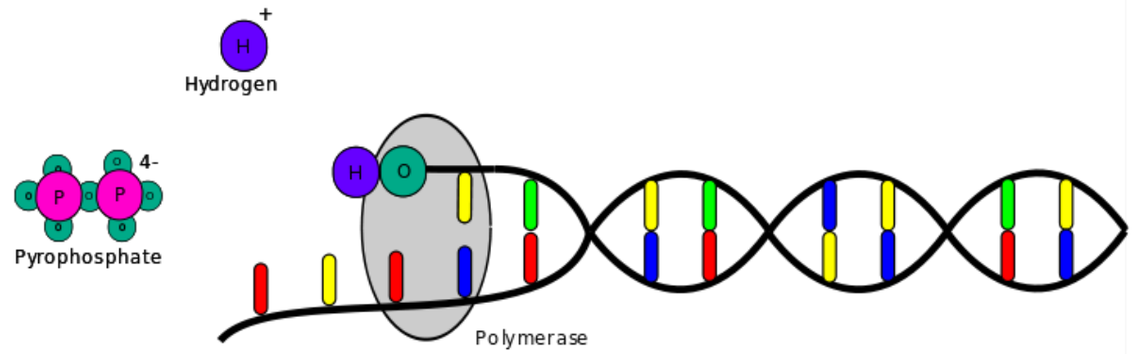
https://commons.wikimedia.org/wiki/File:DNTP_nucleotide_incorporation_reaction.svg

- Sequencing By Synthesis (like Sanger)
- ATP sulfurylase is used to convert pyrophosphate into ATP, which is then used as the substrate for luciferase, thus producing light proportional to the amount of pyrophosphate.
- Produces reads 300-500bp in length
- Its major flaw: **can’t handle homopolymer runs**

Second Generation Sequencing: Pyrosequencing



Polymerase integrates a nucleotide.

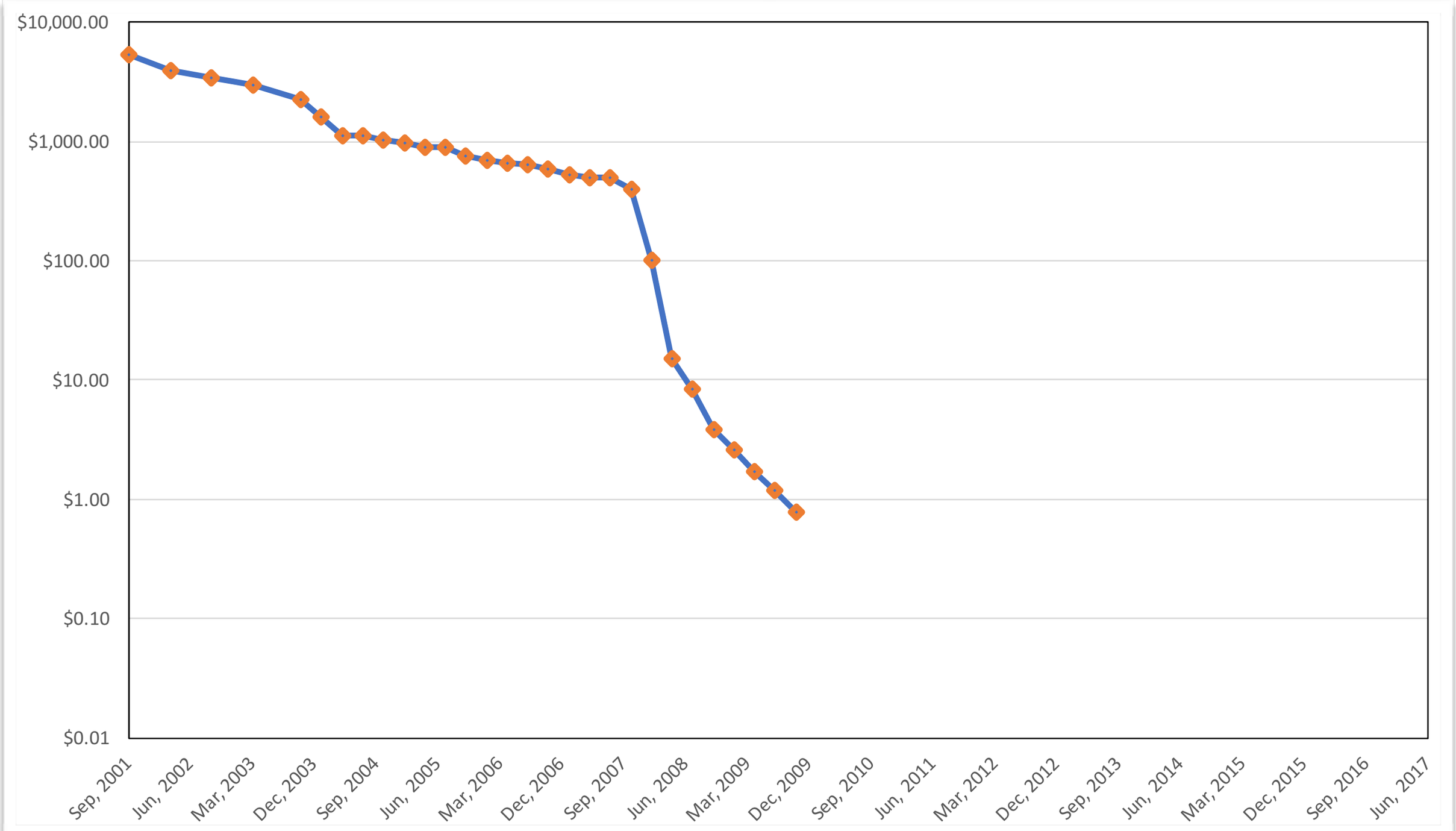


Hydrogen and pyrophosphate are released.

https://commons.wikimedia.org/wiki/File:DNTP_nucleotide_incorporation_reaction.svg

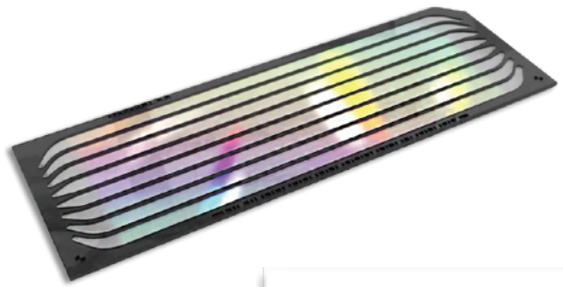
- Ion Torrent
  - Uses a silicon chip to host beads, does not use optics to detect florescence
  - The chip, like a pH meter, detects when a Hydrogen atom is released after the incorporation of a nucleotide
  - Each nucleotides are washed over the chip every 15 seconds, cycling.

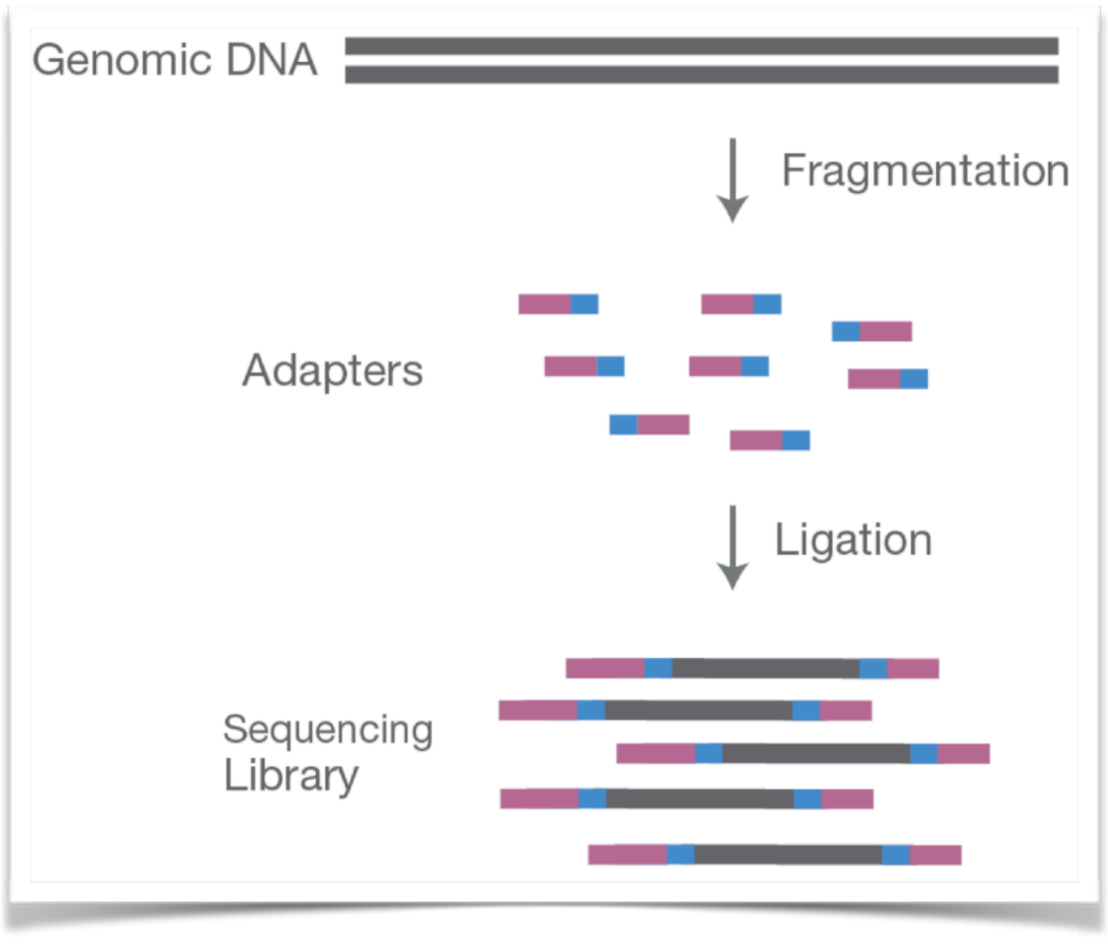
Sequencing Costs after Pyrosequencing

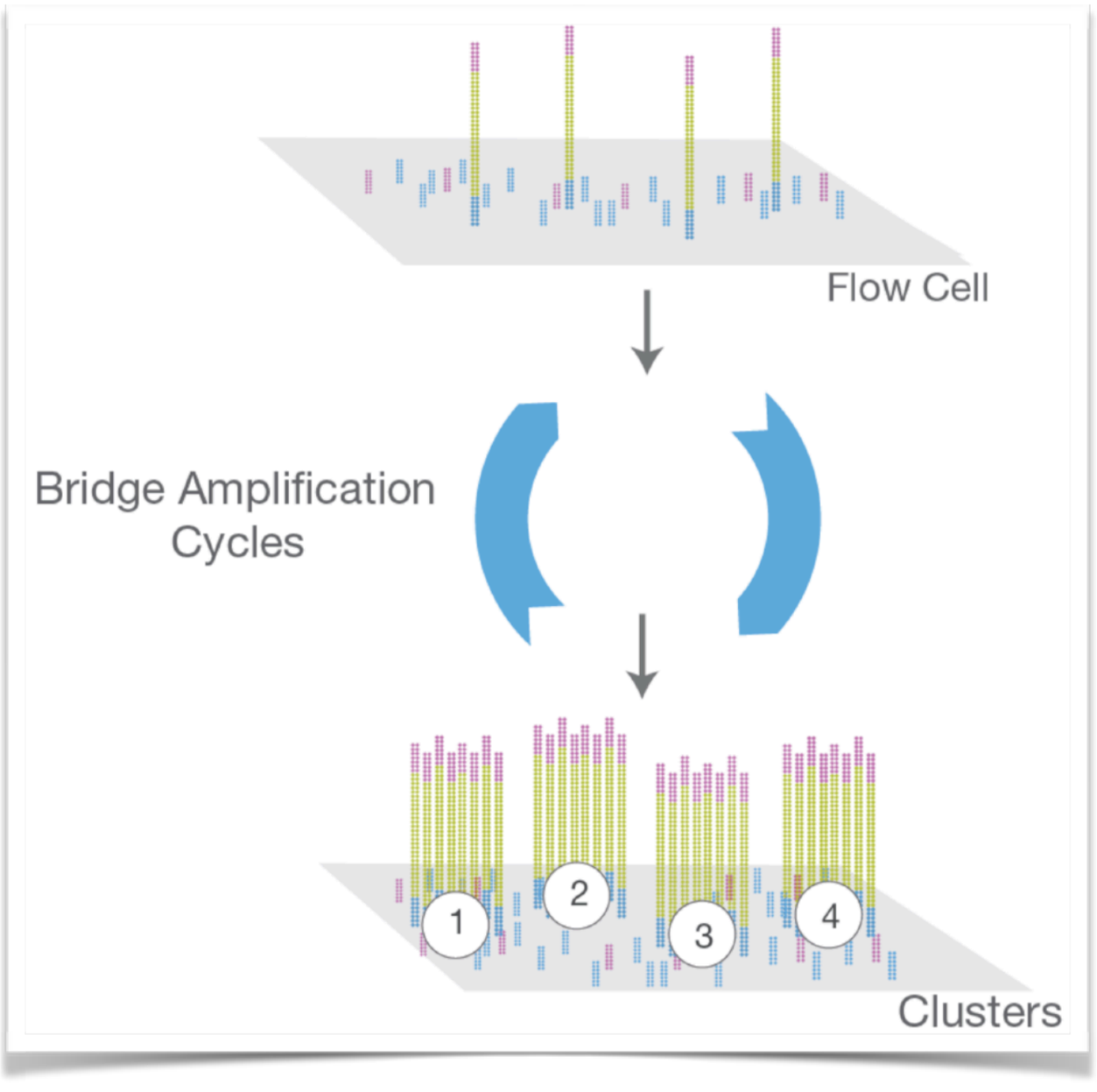


https://www.genome.gov/sequencingcostsdata/

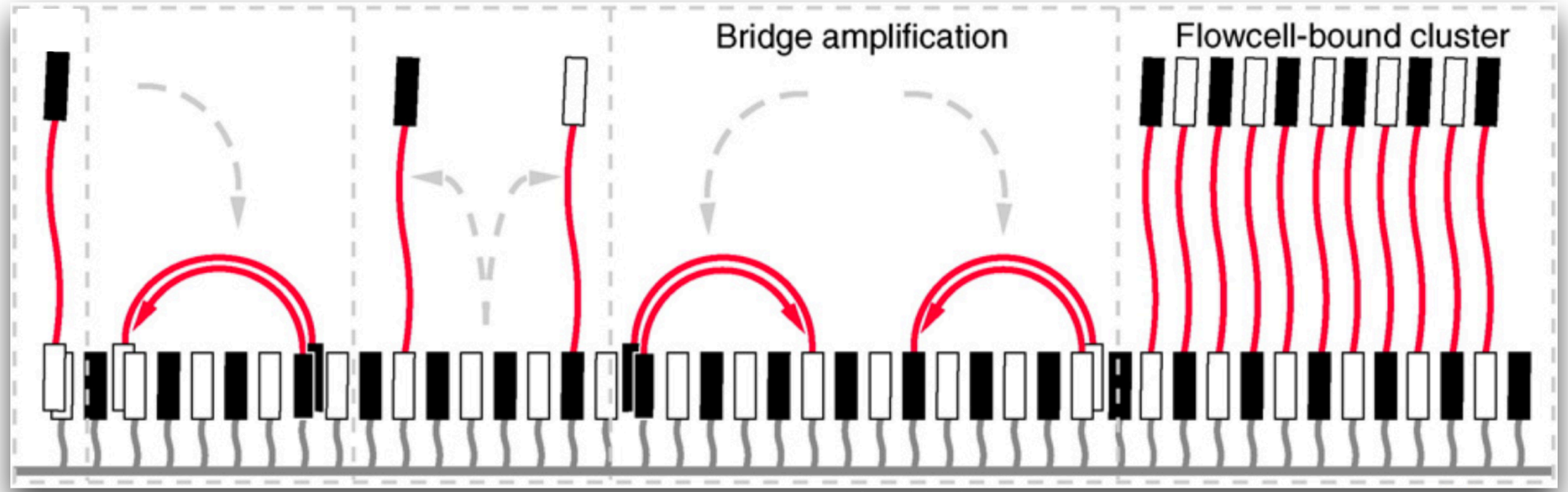
Solexa (Illumina) Sequencing





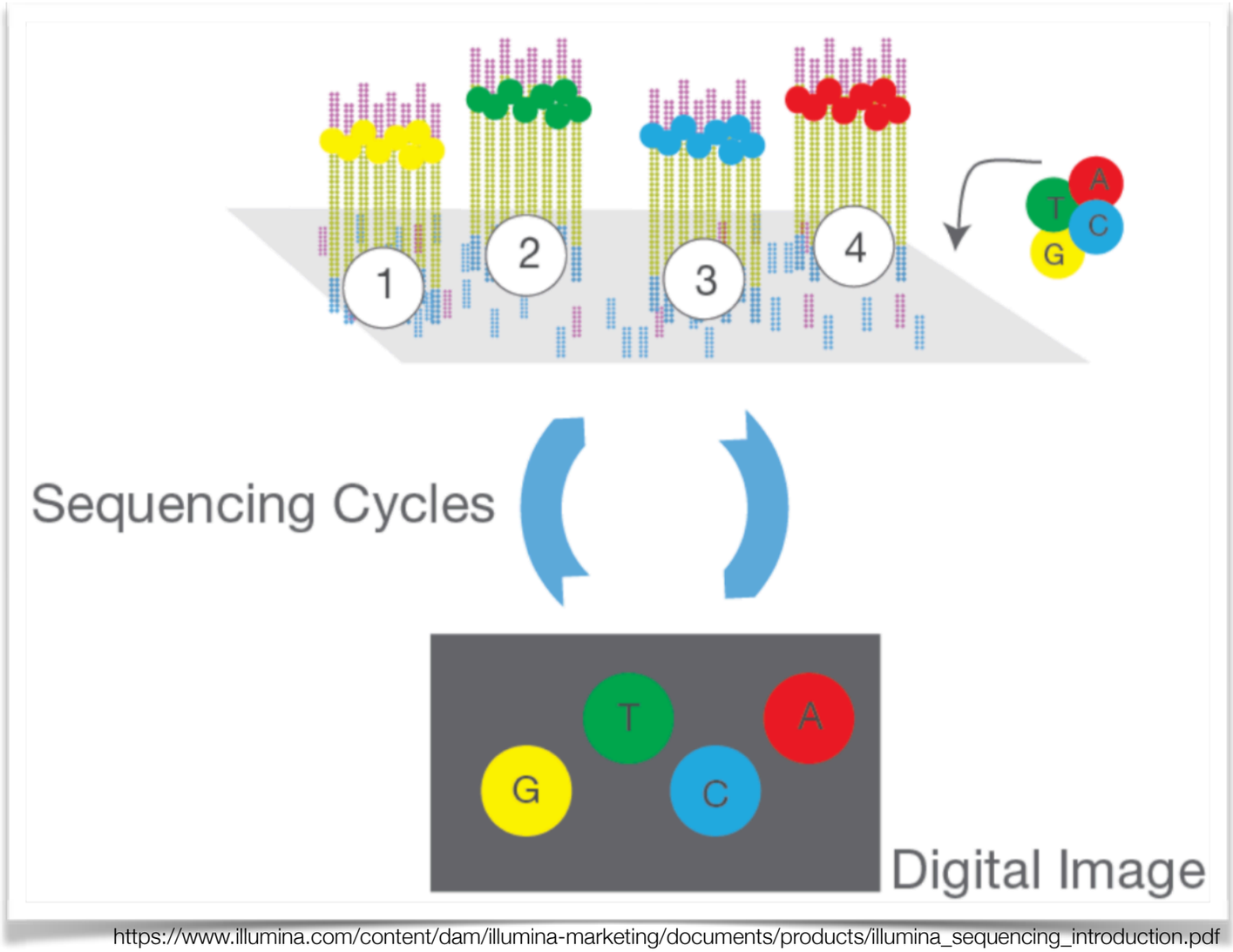


Solexa (Illumina) Sequencing



Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1–8.

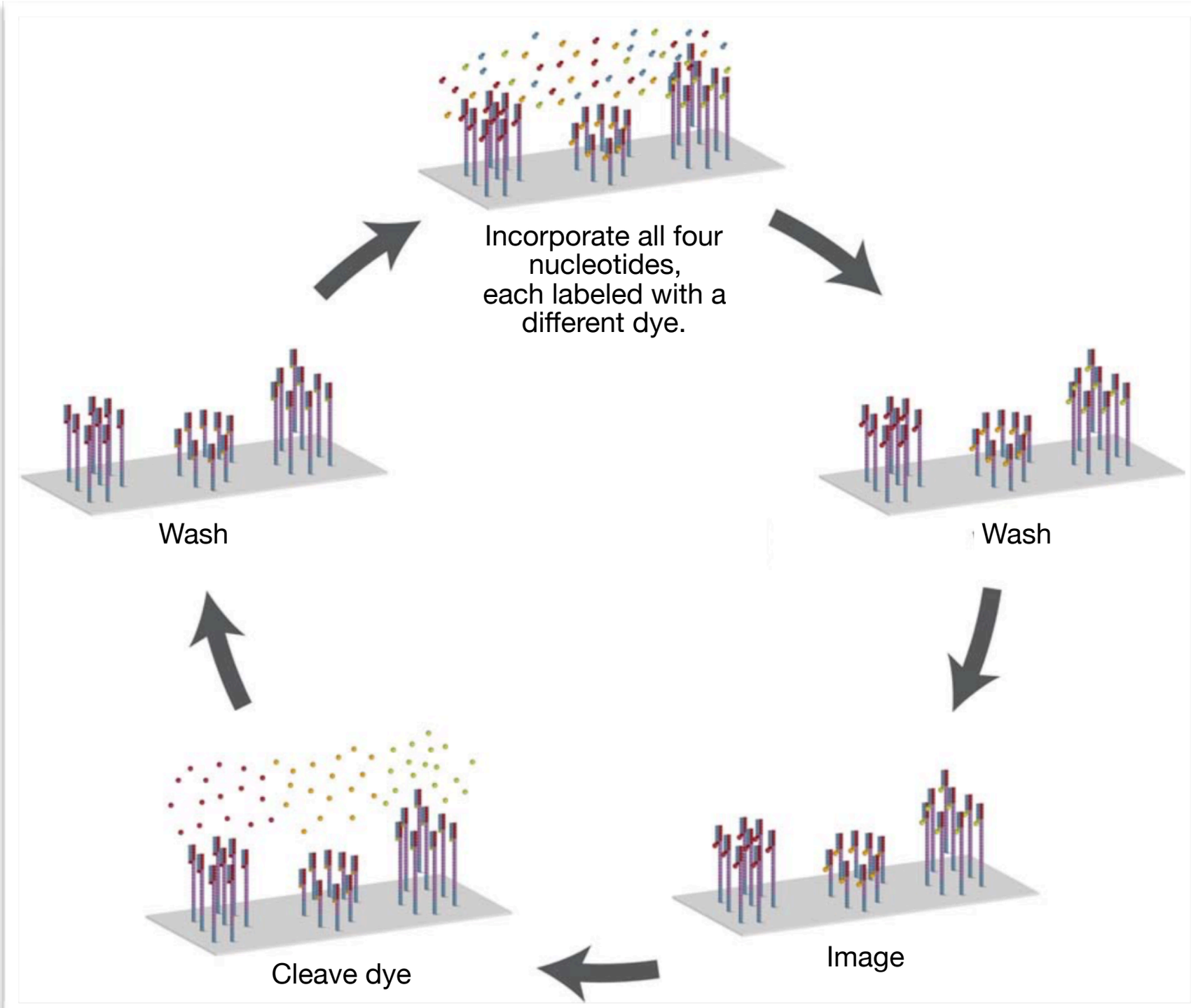
Solexa (Illumina) Sequencing



https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

- Sequencing By Synthesis (like Sanger)
- Uses fluorescent ‘reversible-terminator’ dNTPs
  - cannot immediately bind further nucleotides as the fluorophore occupies the 3′ hydroxyl position
- Fluorophore is cleaved away to allow polymerisation to continue
  - Allows the sequencing to occur in a synchronous manner
- A digital camera captures the fluorescence
- Reads up to 150bp in length (300bp on the MiSeq)
- Can do paired-end sequencing
- 400 million reads per run (25 million on MiSeq)

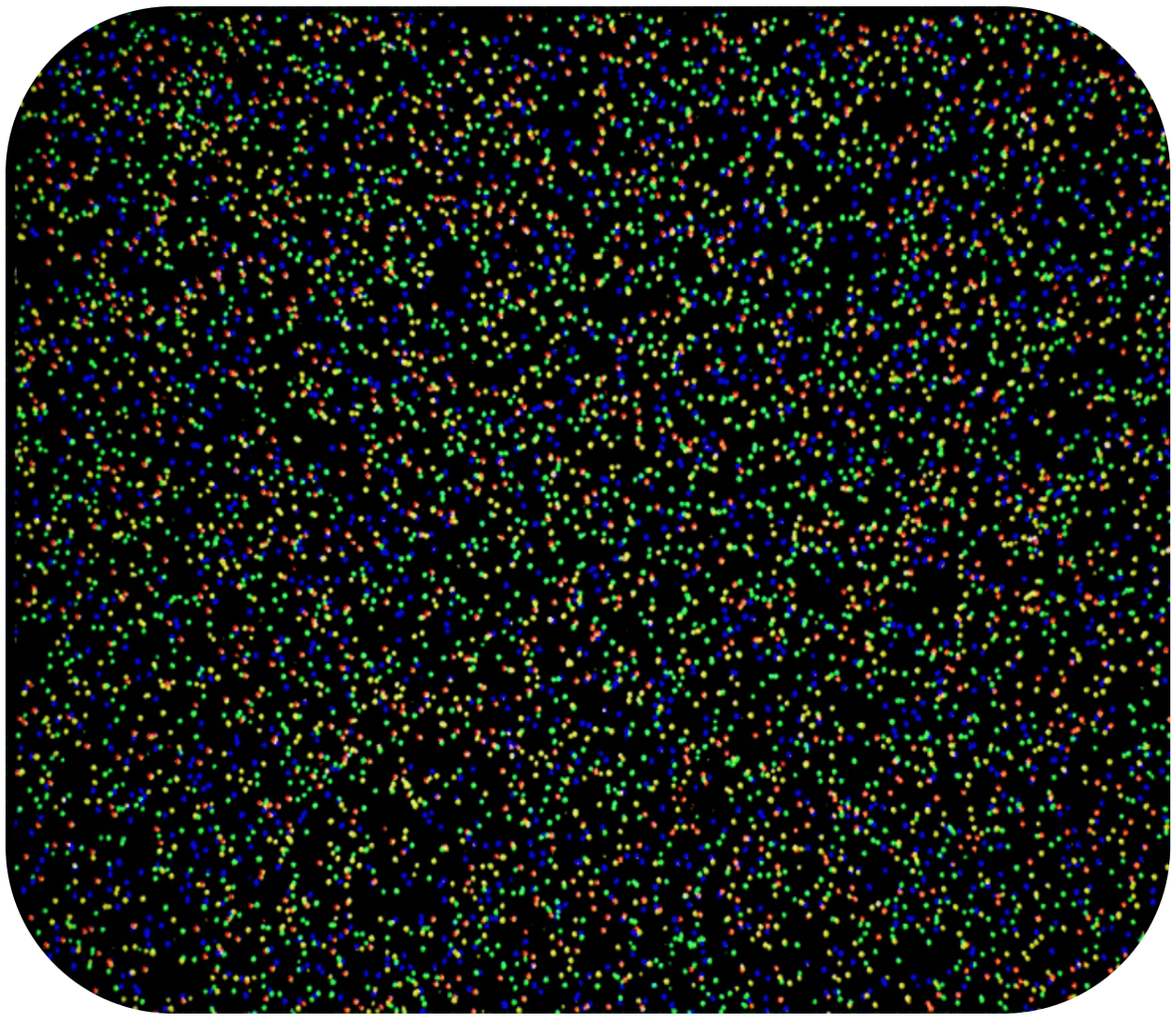
Solexa (Illumina) Sequencing



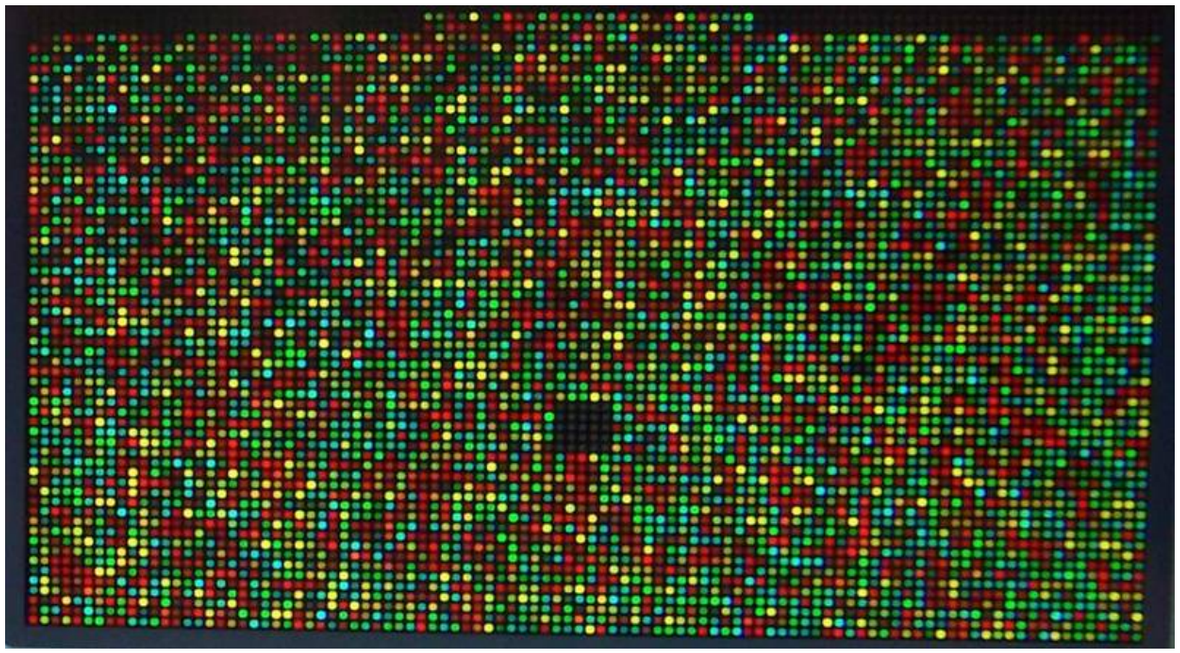
Schadt, et al. (2010) A window into third generation sequencing. Human Molecular Genetics. 19(2).

Solexa (Illumina) Sequencing

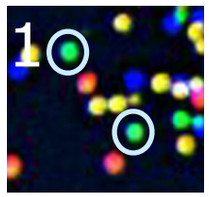
Flow Cell



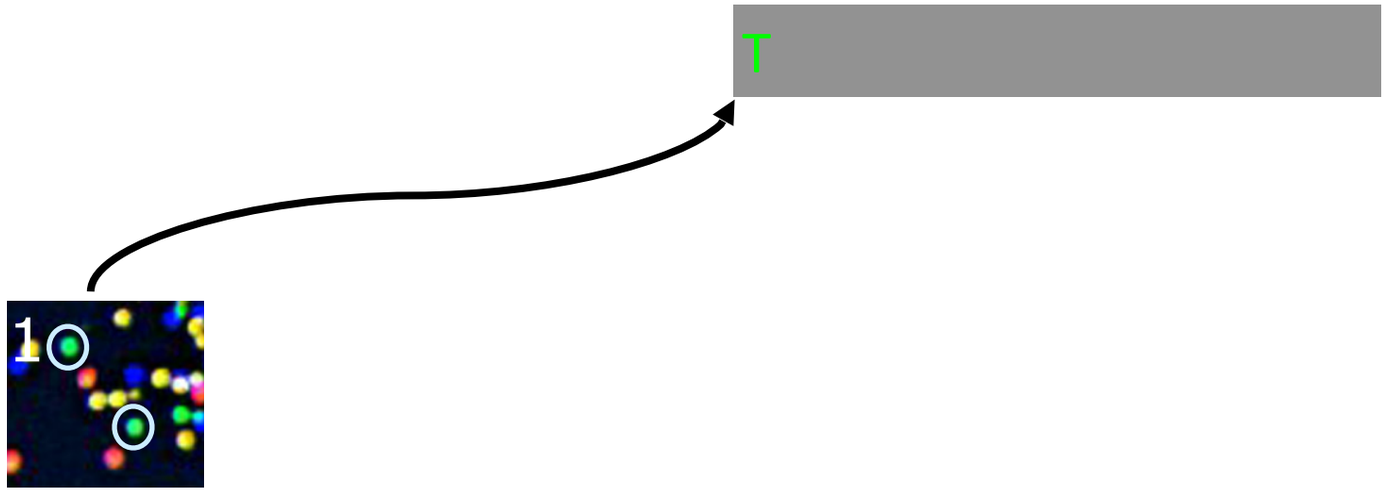
Patterned Flow Cell



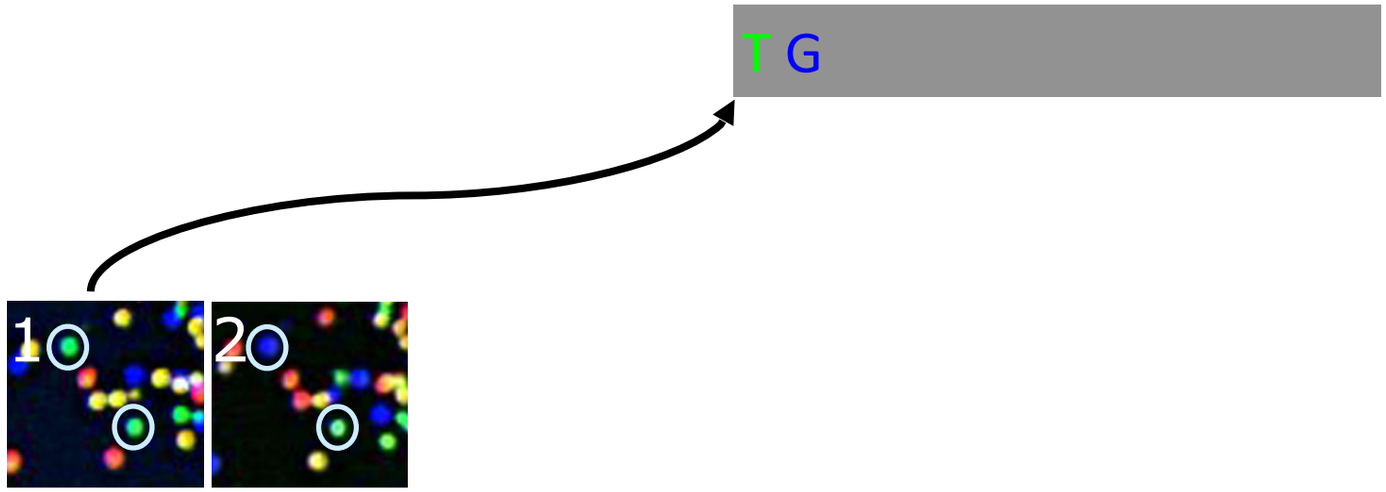
Solexa (Illumina) Sequencing



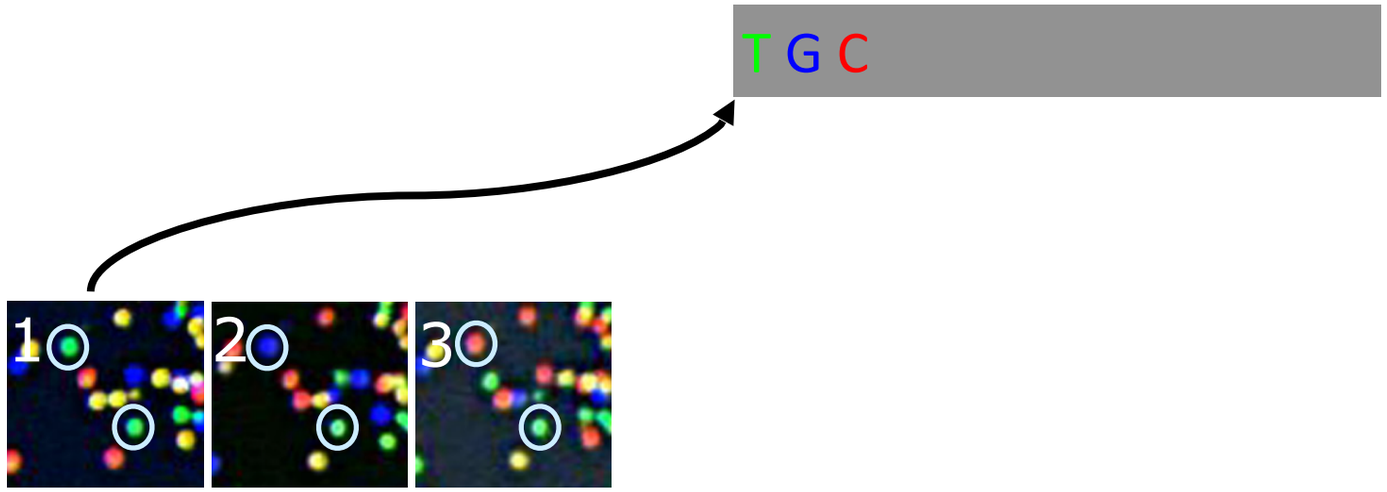
Solexa (Illumina) Sequencing



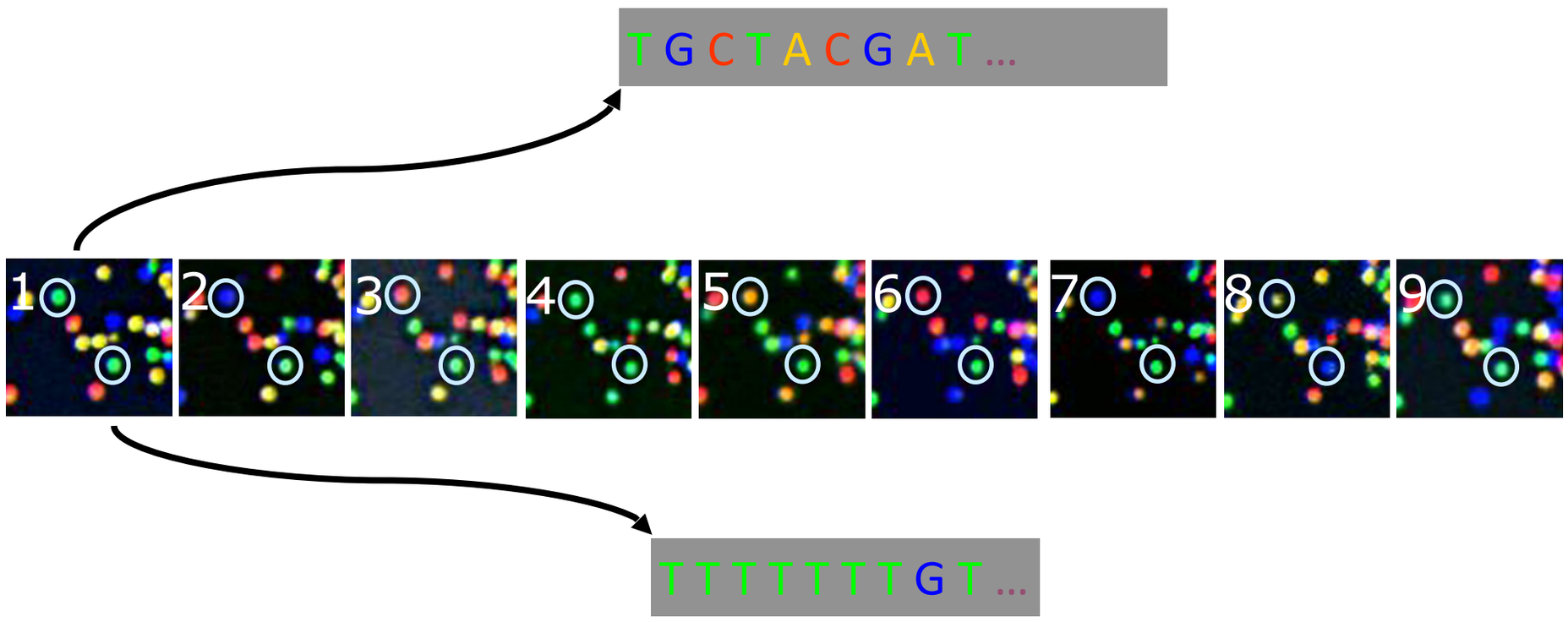
Solexa (Illumina) Sequencing



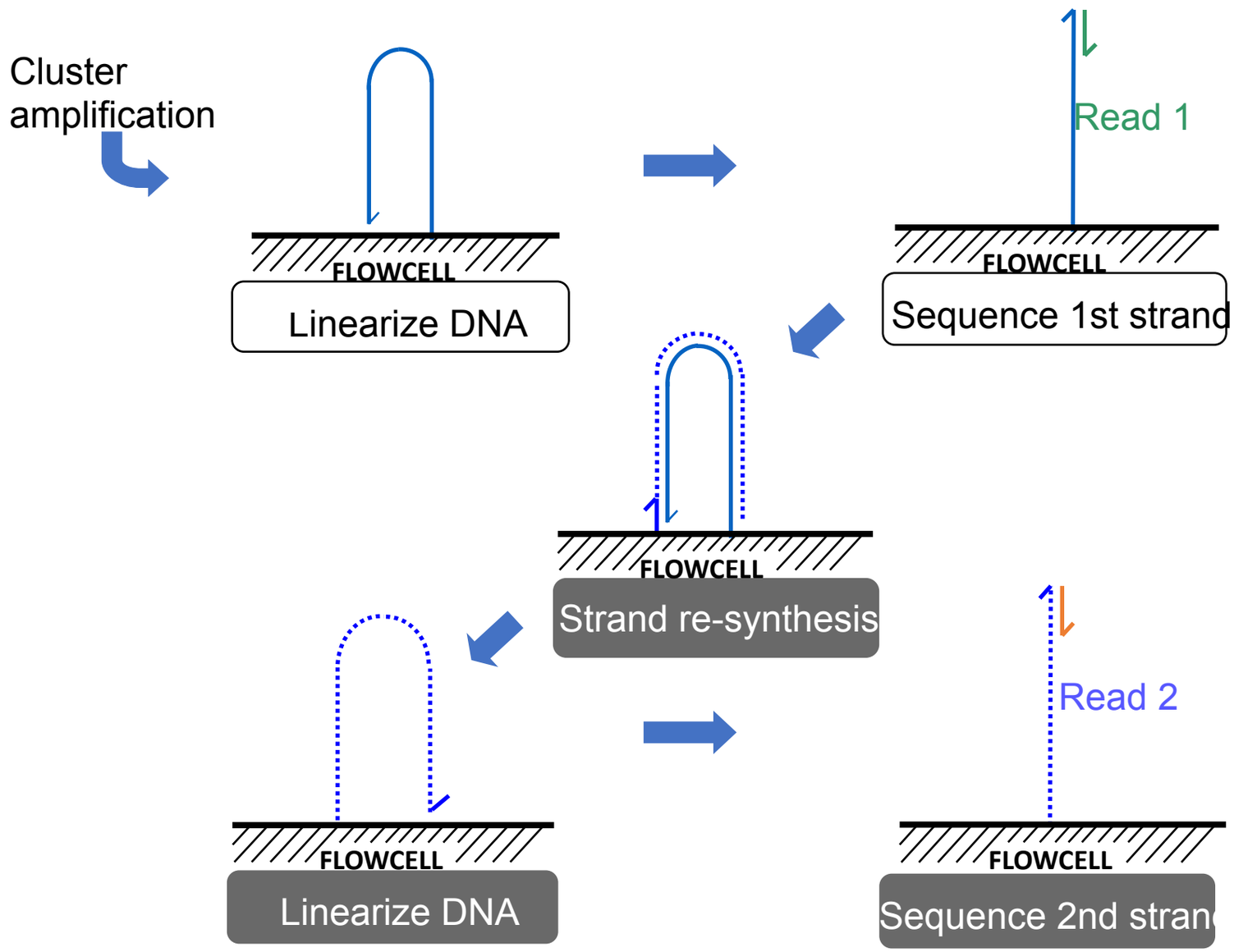
Solexa (Illumina) Sequencing



Solexa (Illumina) Sequencing

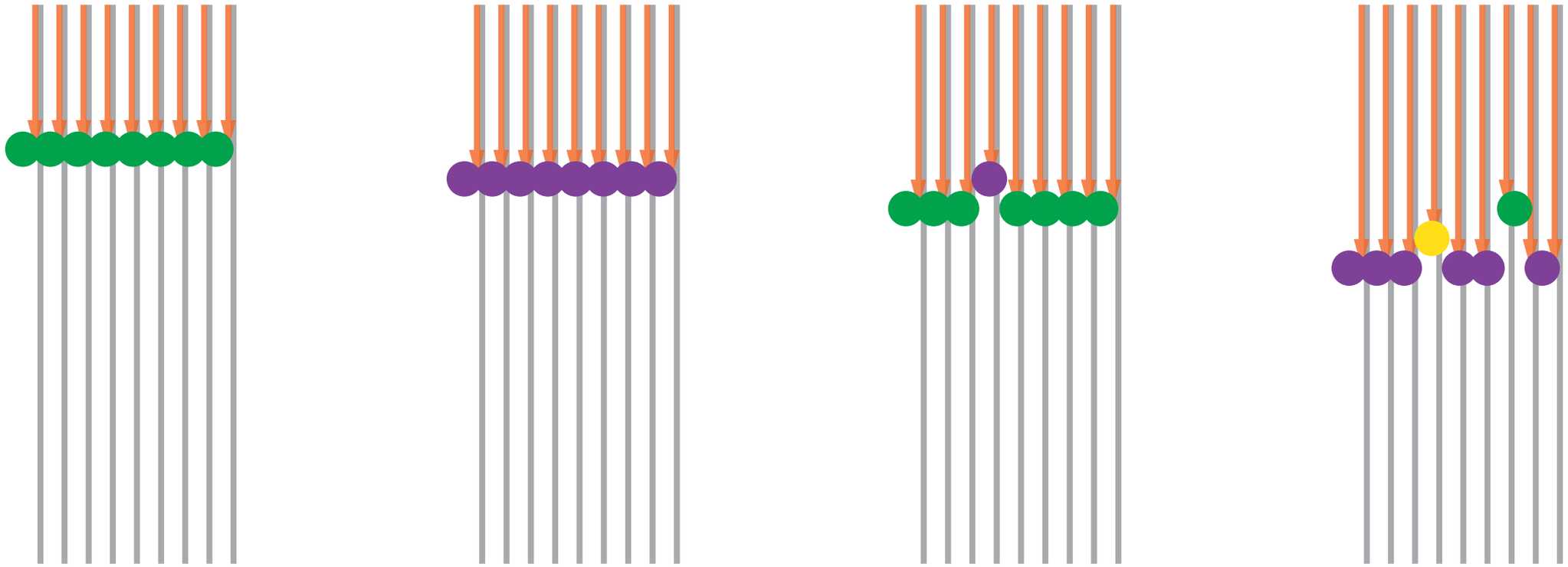


Solexa (Illumina) Paired-end Sequencing

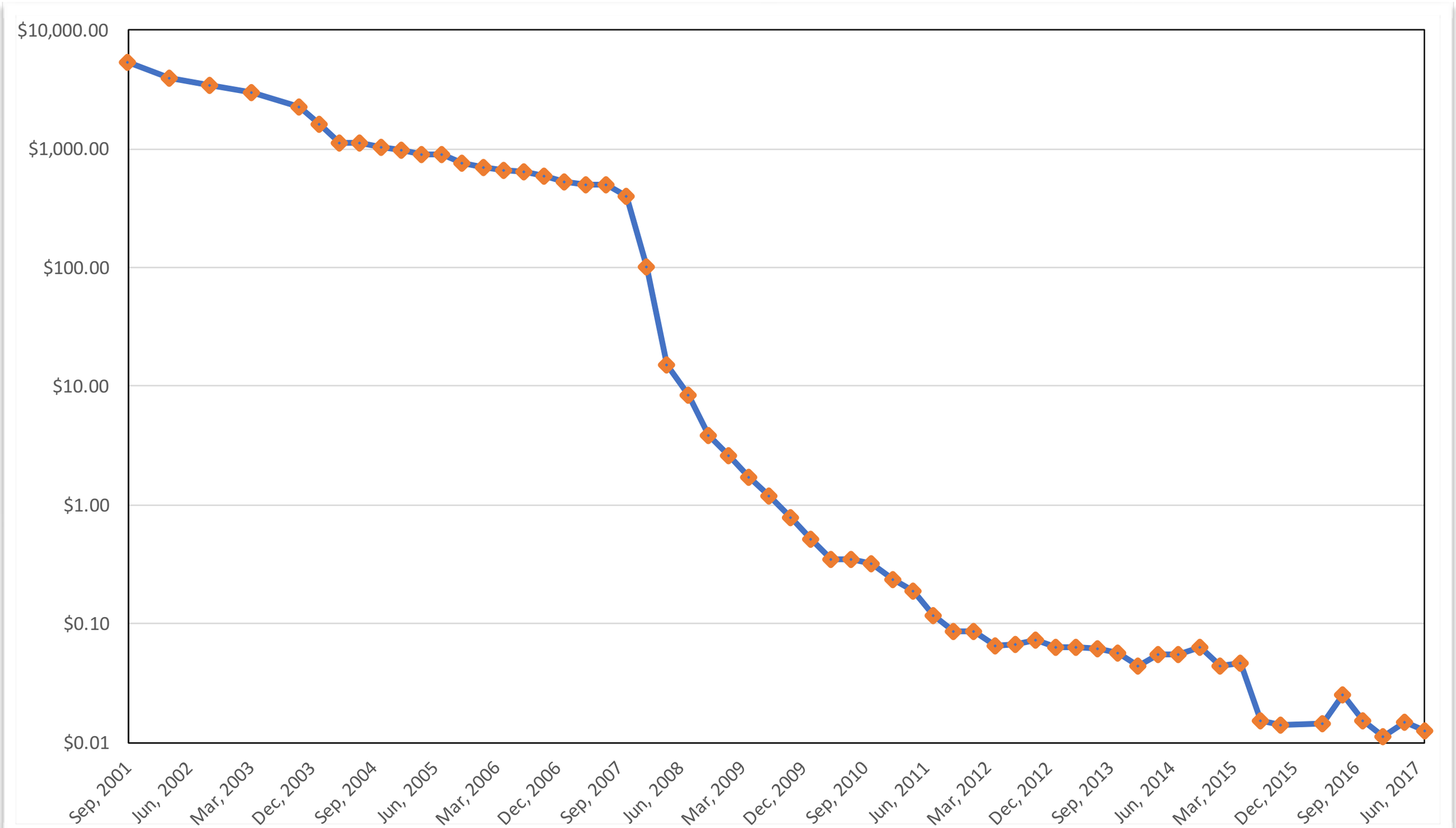


Solexa (Illumina) Sequencing

Error model: clusters come out of phase on the flow cell



Sequencing Costs after Illumina



https://www.genome.gov/sequencingcostsdata/

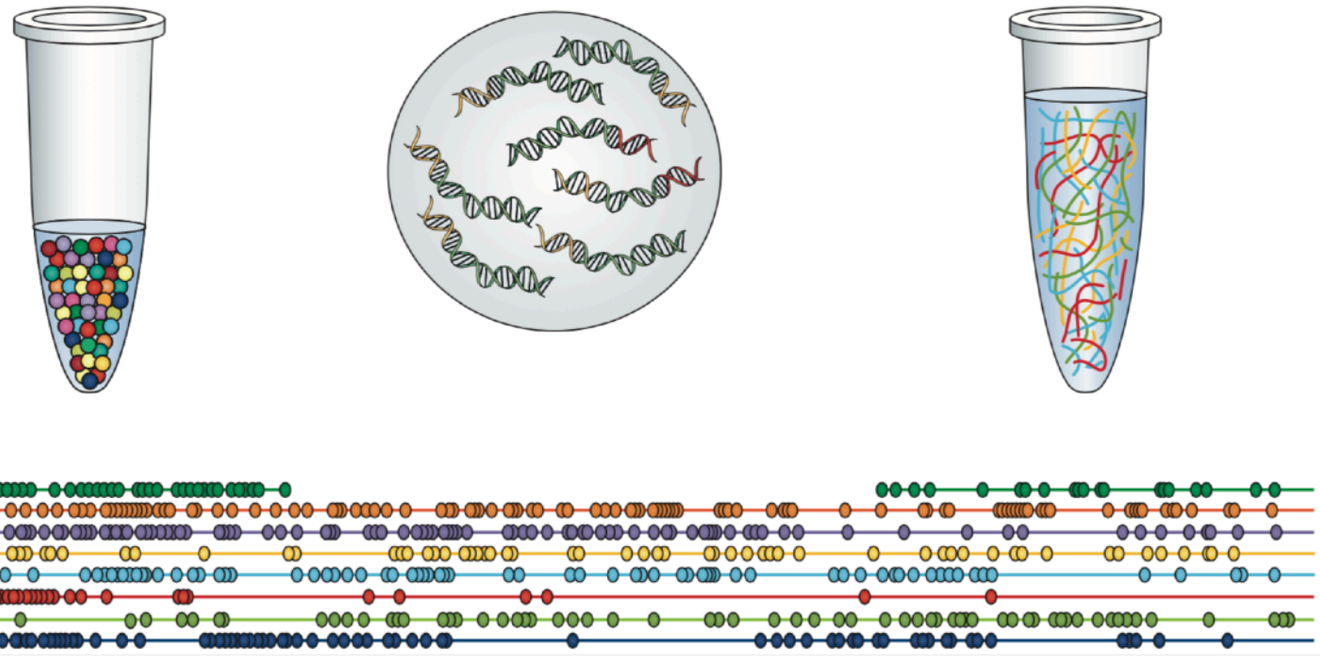
Solexa (Illumina) Sequencing + 10X Genomics

Emulsion PCR
Arbitrarily long DNA is mixed with beads loaded with barcoded primers, enzyme and dNTPs

GEMs
Each micelle has 1 barcode out of 750,000

Amplification
Long fragments are amplified such that the product is a barcoded fragment ~350 bp

Pooling
The emulsion is broken and DNA is pooled, then it undergoes a standard library preparation



Goodwin, et al. (2016) Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 17:333–351.

Third Generation Sequencing

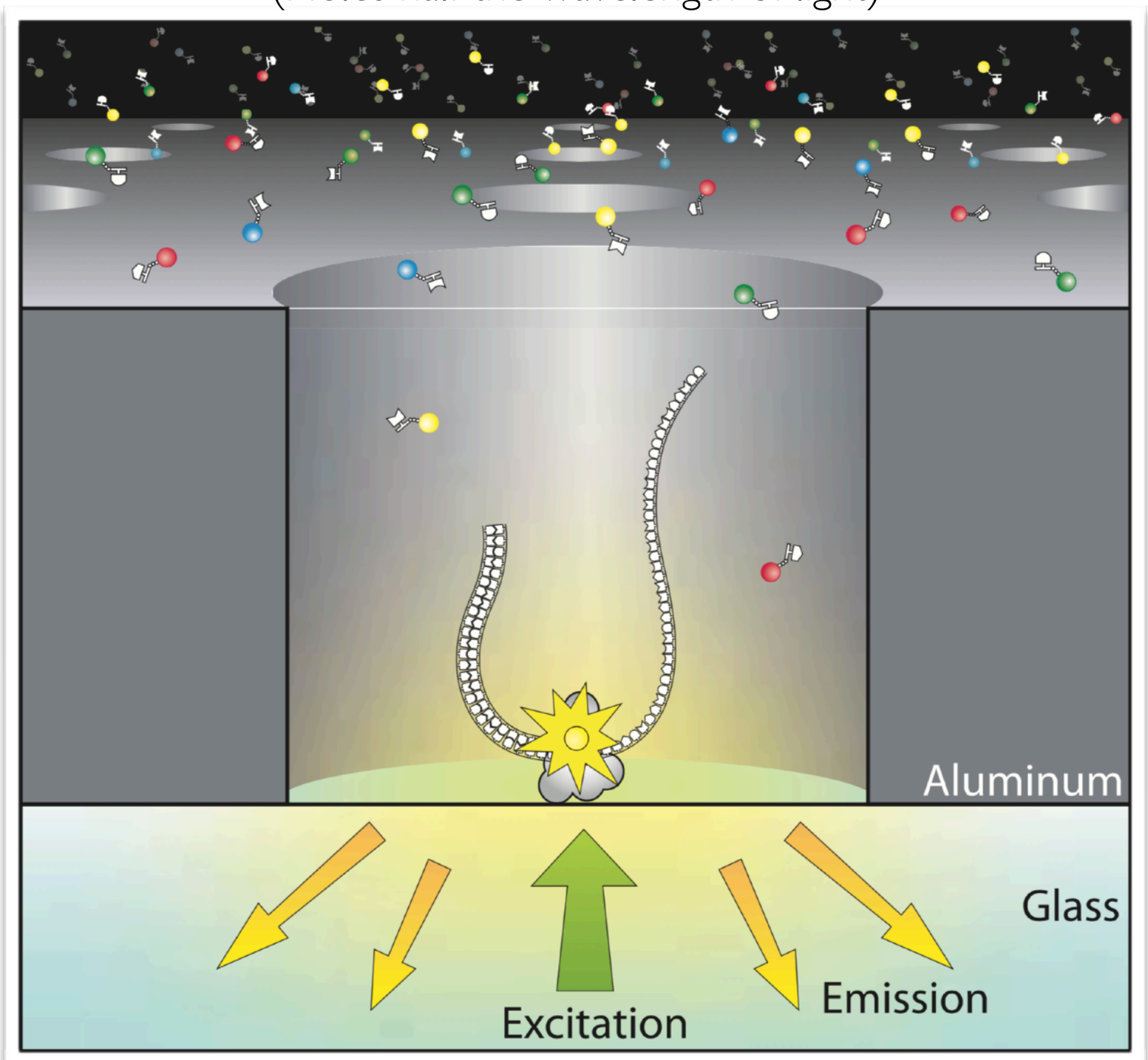
- Those capable of sequencing single molecules
- With no requirement for DNA amplification.

PacBio Single Molecule Sequencing

Zero Mode Wave Guide

(Holes half the wavelength of light)

Use DNA polymerase as a "*real-time sequencing engine*"



Eid, et al. (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 233.

PacBio Single Molecule Sequencing

Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations

M. J. Levene,[1] J. Korlach,[1,2] S. W. Turner,[1]* M. Foquet,[1] H. G. Craighead,[1] W. W. Webb[1]†

Optical approaches for observing the dynamics of single molecules have required pico- to nanomolar concentrations of fluorophore in order to isolate individual molecules. However, many biologically relevant processes occur at micromolar ligand concentrations, necessitating a reduction in the conventional observation volume by three orders of magnitude. We show that arrays of zero-mode waveguides consisting of subwavelength holes in a metal film provide a simple and highly parallel means for studying single-molecule dynamics at micromolar concentrations with microsecond temporal resolution. We present observations of DNA polymerase activity as an example of the effectiveness of zero-mode waveguides for performing single-molecule experiments at high concentrations.

Data from a single molecule can reveal information about kinetic processes not normally accessible by ensemble measurements, such as variances in kinetic rates, memory effects, and lifetimes of transient intermediates (*1*, *2*). Single-molecule approaches to drug screening, mRNA expression profiling, single-nucleotide polymorphism detection and DNA sequencing may also have many advantages over current techniques (*3*–*5*). Common approaches to studying single molecules include fluorescence correlation spectroscopy (FCS) (*6*, *7*) and direct observation of sparse molecules using diffraction-limited optics (*8*, *9*). These approaches provide observation volumes on the order of 0.2 fl and therefore require pico- to nanomolar concentrations of fluorophore in order to isolate individual molecules in solution (*10*). However, many enzymes naturally work at much higher ligand concentrations, and their Michaelis constants are often in the micro- to millimolar range (*11*). Low concentrations of ligand can influence the mechanistic pathway of enzyme kinetics (e.g., by allosteric control or conformational relaxation) and alter the partitioning between multiple catalytic pathways, thus affecting turnover cycle histories and distributions (*12*). Working at biologically more relevant, micromolar concentrations requires reducing the observation volume by over three orders of magnitude.

In addition to requiring low concentrations of ligand, the temporal resolution of conventional approaches to single-molecule kinetics is often limited by the time it takes for molecules to diffuse out of the observation volume, usually on the order of several hundred microseconds. The temporal resolution and the upper limit of practicable concentrations would be greatly improved by reducing the effective observation volume.

Previous approaches include total internal reflection illumination, which can reduce the observation volume by a factor of 10 (*13*), and near-field scanning optical microscopy (NSOM), which achieves observation volumes with lateral dimensions on the order of 50 nm by illumination through a small aperture, usually terminating a tapered optical fiber (*14*). NSOM has been used to observe single molecules on a surface (*15*), but it suffers from unreliable probe manufacture and its complexity is not easily amenable to highly parallel implementations.

Concurrent to the development of single-molecule analytical techniques has been rapid progress in nanobiotechnology and efforts at

[1]Applied and Engineering Physics, [2]Graduate Program in Biochemistry, Molecular, and Cell Biology, Cornell University, Clark Hall, Ithaca, NY 14853, USA.

*Present address: Nanofluidics Incorporated, 17 Sheraton Drive, Ithaca, NY 14850, USA.
†To whom correspondence should be addressed. E-mail: www2@cornell.edu

PacBio Single Molecule Sequencing

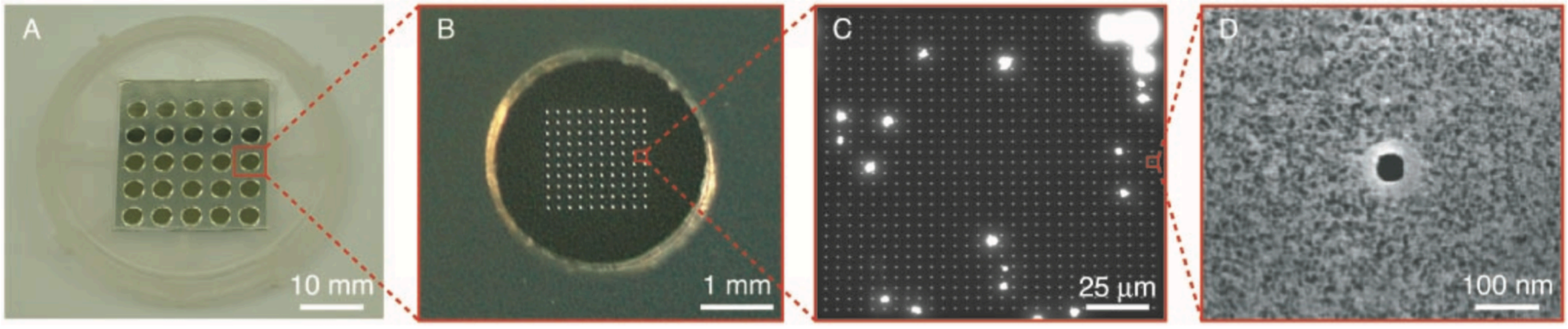
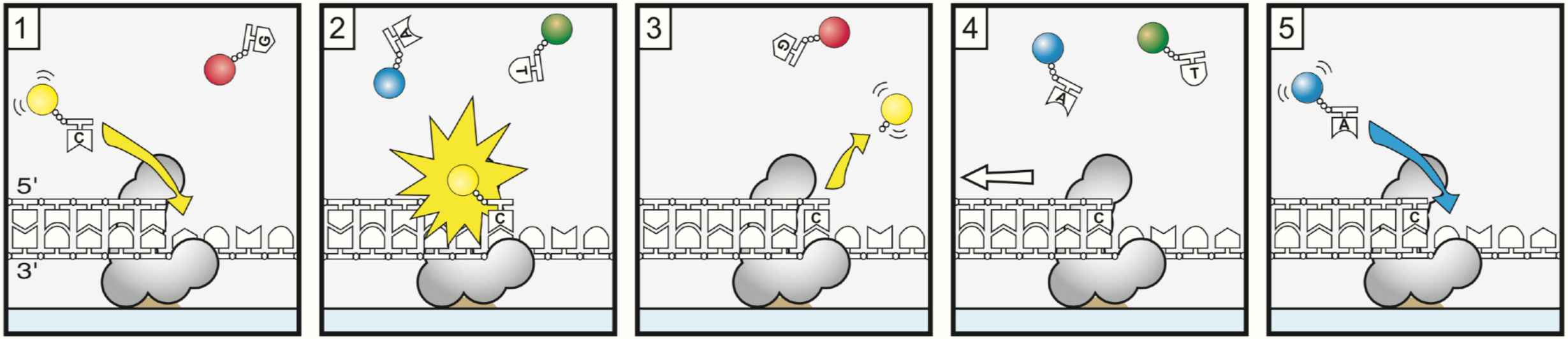
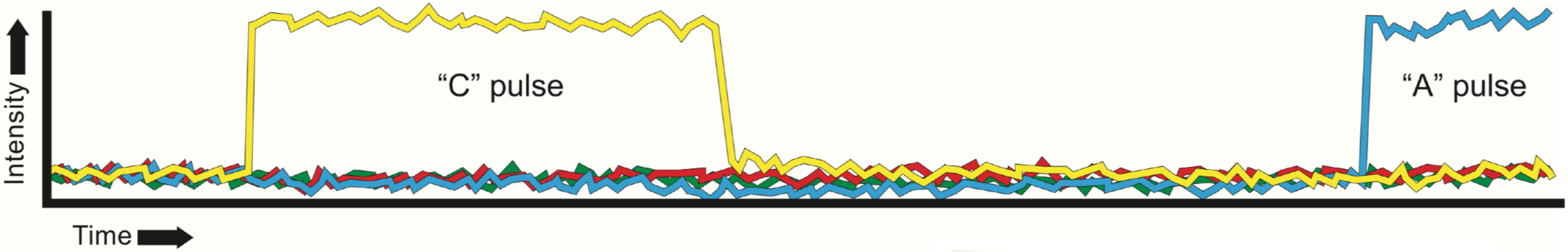
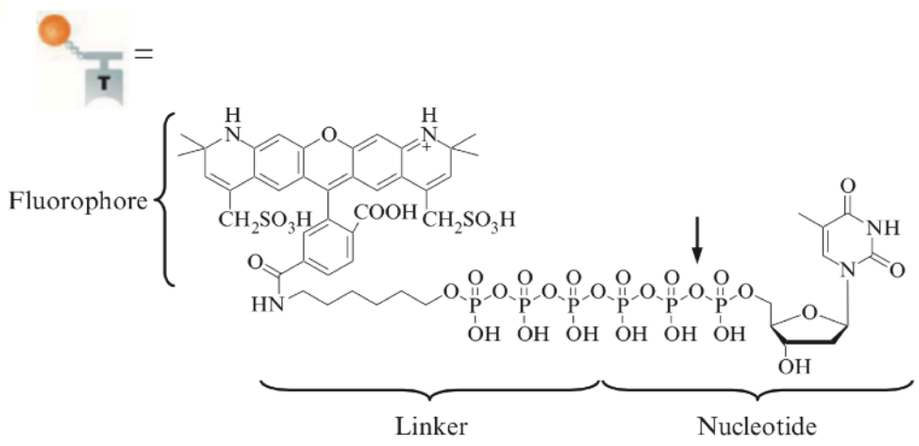


Fig. 3. A fused silica coverslip with zero-mode waveguides arrays. (**A**) The coverslip, with overlying gasket to isolate arrays for individual experiments. Successive increases in scale are shown in (**B**) to (**D**). A scanning electron microscope image of an individual waveguide is shown in (D). The bright spots in (C) correspond to defects in the metal film. The large bright pattern in the upper right corner is a coded orientation marker.

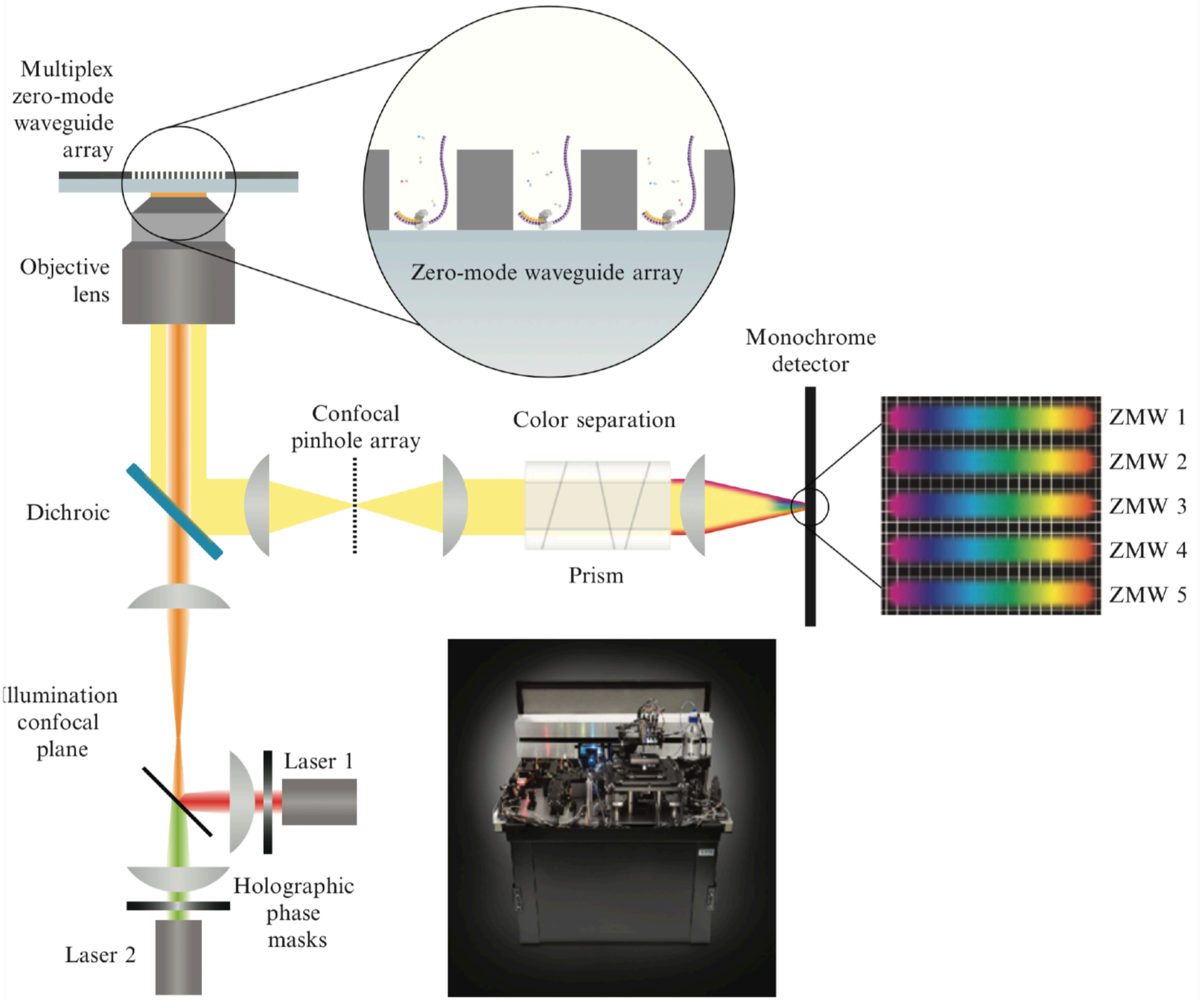
PacBio Single Molecule Sequencing





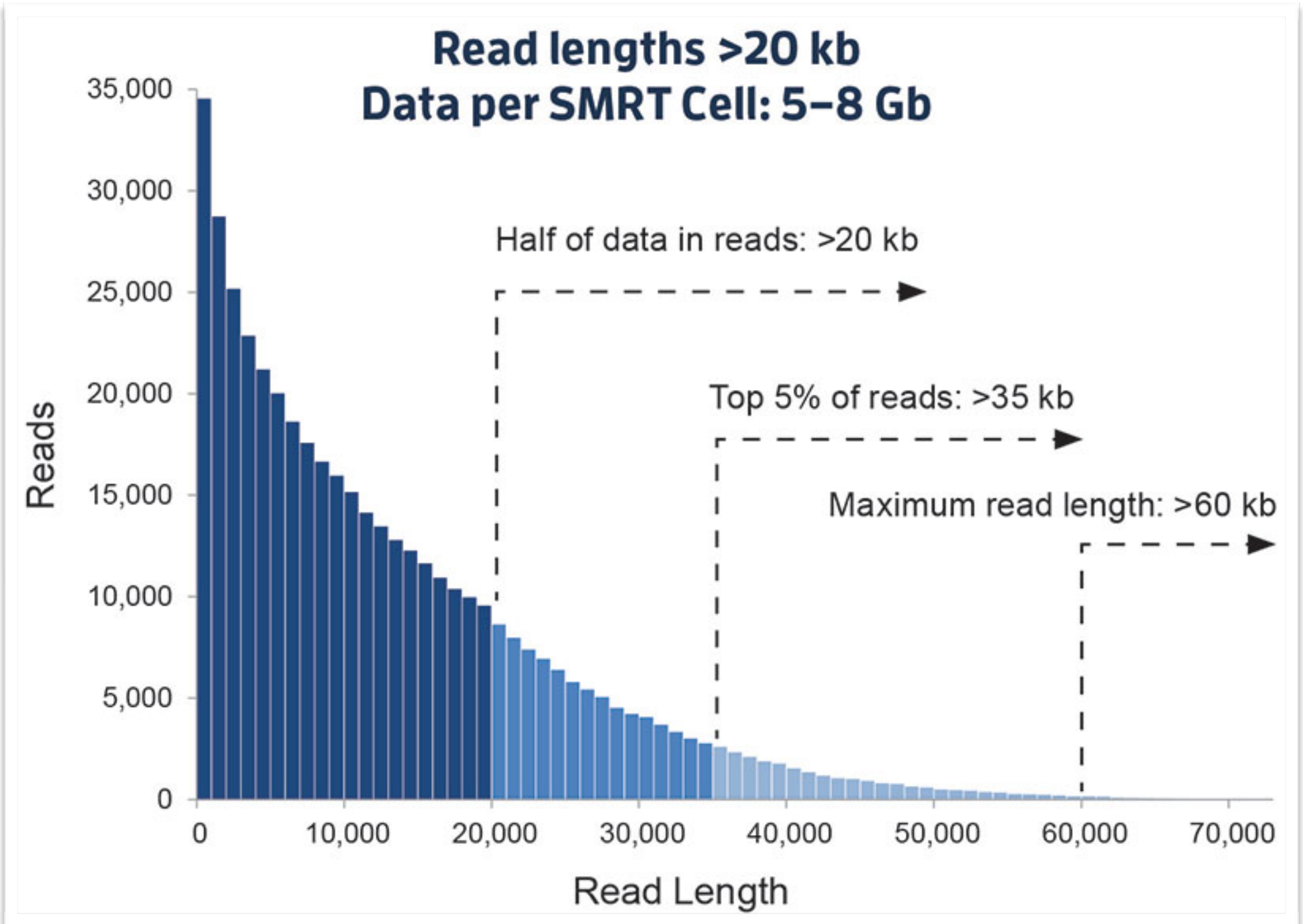


PacBio Single Molecule Sequencing



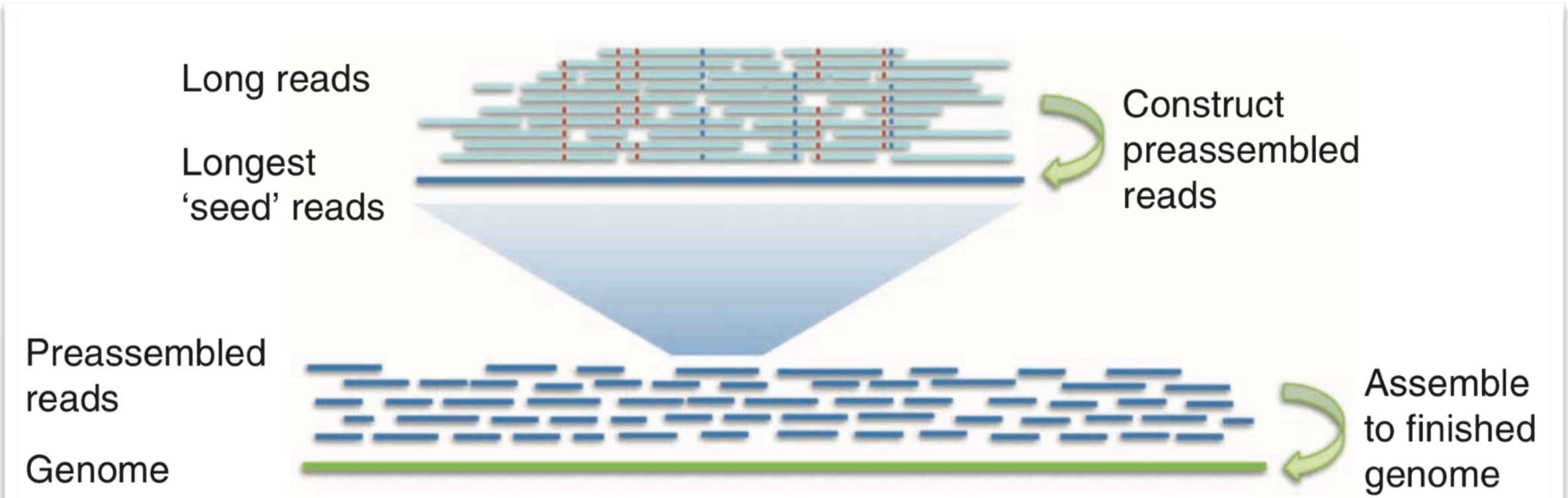
Korlach, et al. (2010) Real-Time DNA Sequencing from Single Polymerase Molecules. Methods in Enzymology. 472.

PacBio Single Molecule Sequencing



http://www.pacb.com/smrt-science/smrt-sequencing/read-lengths/

PacBio Single Molecule Sequencing



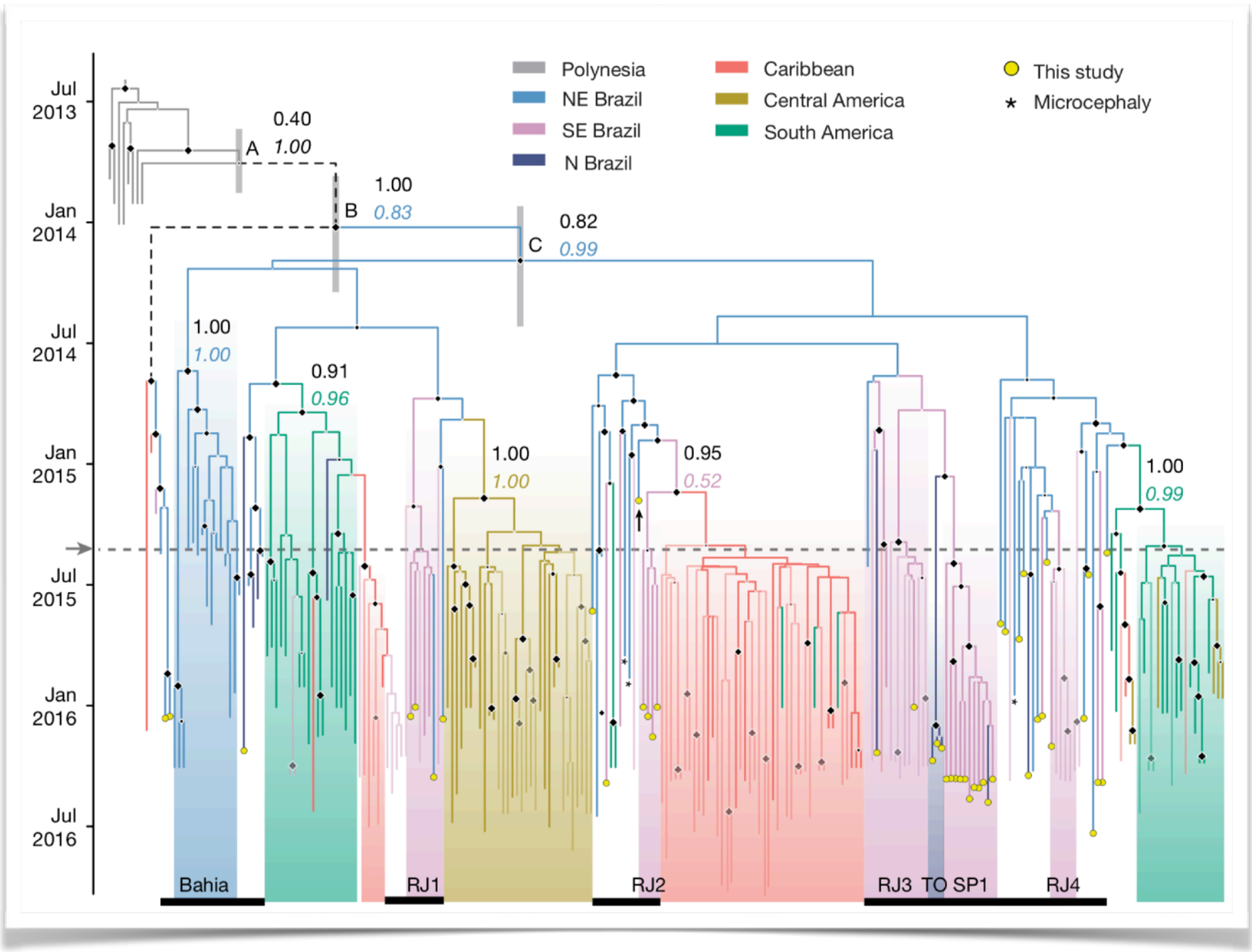
Chin, et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods.

- Error model: indels from polymerase slippage
- Errors are independent and random
- can be corrected by piling-up multiple reads.

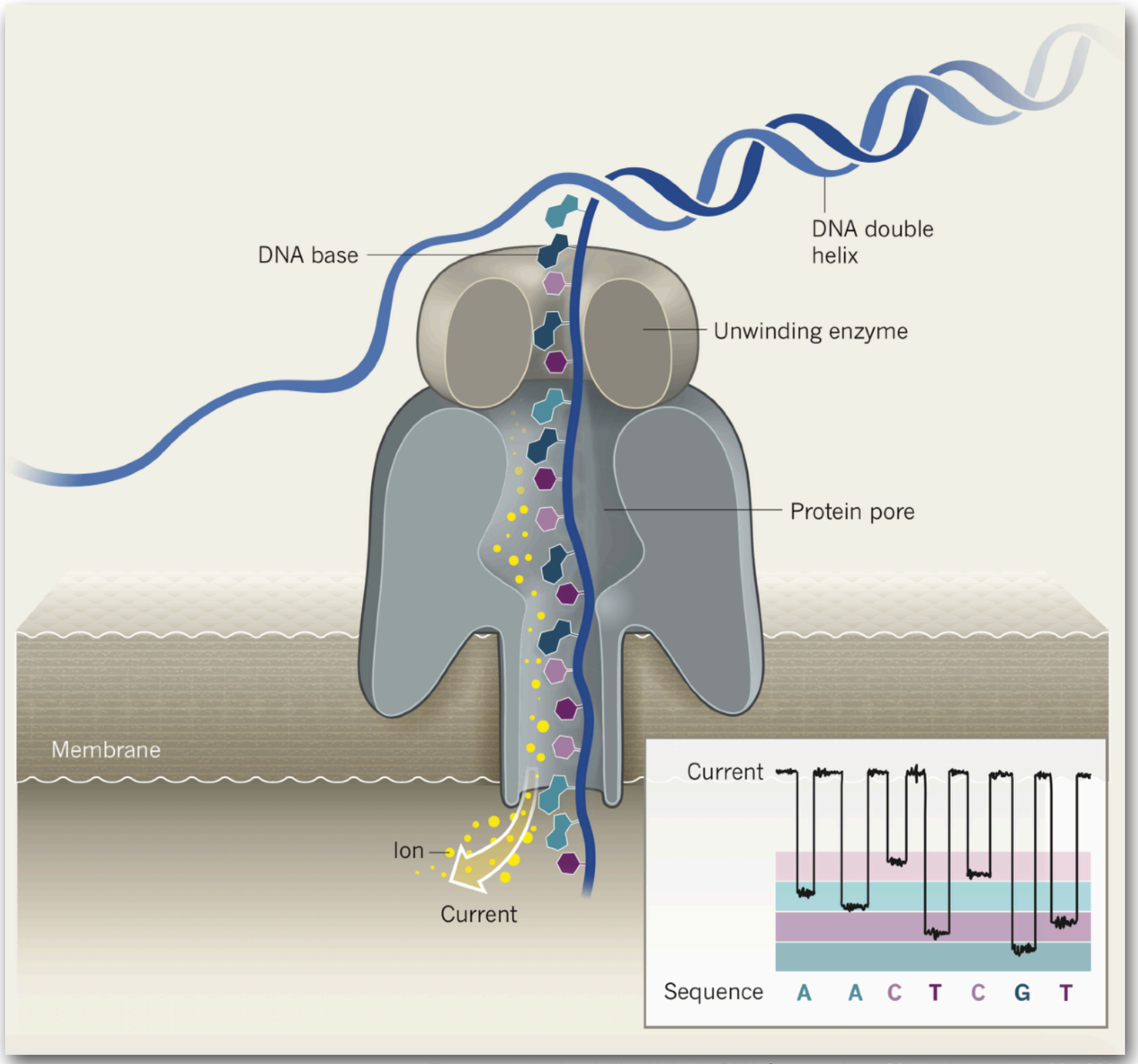
Nanopore Sequencing

Establishment and cryptic transmission of Zika virus in Brazil and the Americas

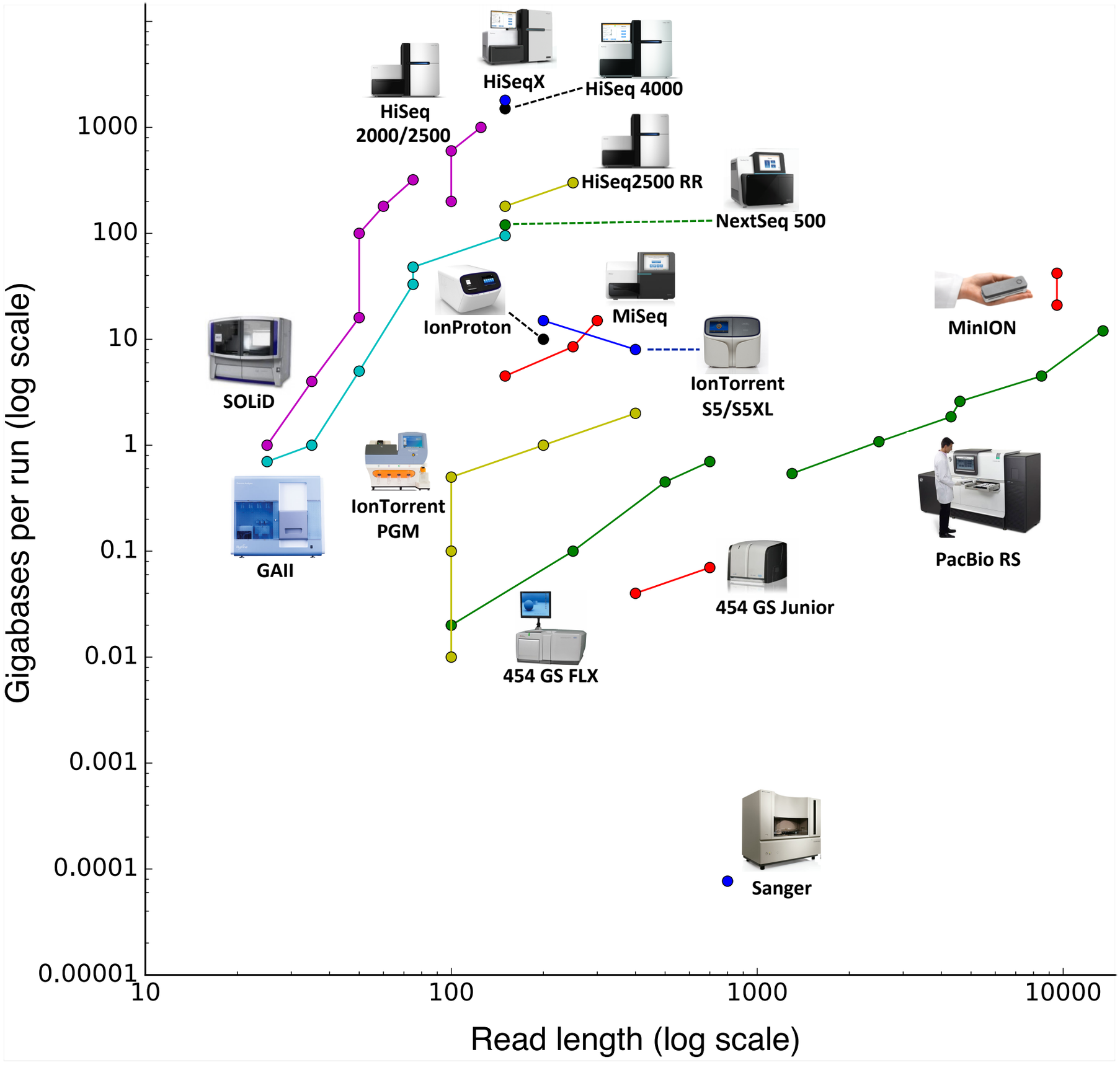
N. R. Faria[1,2]*, J. Quick[3]*, I. M. Claro[4]*, J. Thézé[1]*, J. G. de Jesus[5]*, M. Giovanetti[5,6]*, M. U. G. Kraemer[1,7,8]*, S. C. Hill[1]*, A. Black[9,10]*, A. C. da Costa[4], L. C. Franco[2], S. P. Silva[2], C.-H. Wu[11], J. Raghwani[1], S. Cauchemez[12,13], L. du Plessis[1], M. P. Verotti[14], W. K. de Oliveira[15,16], E. H. Carmo[17], G. E. Coelho[18,19], A. C. F. S. Santelli[18,20], L. C. Vinhal[18], C. M. Henriques[17],



Nanopore Sequencing



Eisenstein, An Ace in the Hole for DNA Sequencing. (2017) Nature, 550:285-288



Lex Nederbragt, Developments in High-Throughput Sequencing, http://dx.doi.org/10.6084/m9.figshare.100940